BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

SES OLAYI TANIMA VE AKUSTİK SAHNE GERİ GETİRİMİ

AHMET MELİH BAŞBUĞ

YÜKSEK LİSANS TEZİ
2019

SES OLAYI TANIMA VE AKUSTİK SAHNE GERİ GETİRİMİ

SOUND EVENT RECOGNITION AND ACOUSTIC SCENES RETRIEVAL

AHMET MELİH BAŞBUĞ

Başkent Üniversitesi
Lisansüstü Eğitim Öğretim ve Sınav Yönetmeliğinin
BİLGİSAYAR Mühendisliği Anabilim Dalı İçin Öngördüğü
YÜKSEK LİSANS TEZİ
olarak hazırlanmıştır.

2019

“SES OLAYI TANIMA VE AKUSTİK SAHNE GERİ GETİRİMİ” başlıklı bu çalışma, jürimiz tarafından, 10/09/2019 tarihinde, **BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI 'nda YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Başkan	: Doç. Dr. Ahmet Burak CAN
Üye (Danışman)	: Dr. Öğr. Üyesi Mustafa SERT
Üye	: Dr. Öğr. Üyesi Emre SÜMER

ONAY

. .../..../.....

Prof. Dr. Ömer Faruk ELALDI

Fen Bilimleri Enstitüsü Müdürü

BAŞKENT ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ
YÜKSEK LİSANS / DOKTORA TEZ ÇALIŞMASI ORİJİNALLİK RAPORU

Tarih: 25 / 09 / 2019

Öğrencinin Adı, Soyadı : Ahmet Melih BAŞBUĞ

Öğrencinin Numarası : 21510209

Anabilim Dalı : Bilgisayar Mühendisliği A.B.D.

Programı : Bilgisayar Mühendisliği Tezli Y.L.

Danışmanın Adı, Soyadı : Mustafa SERT

Tez Başlığı : Ses Olayı Tanıma ve Akustik Sahne Geri Getirimi

Yukarıda başlığı belirtilen Yüksek Lisans/Doktora tez çalışmamın; Giriş, Ana Bölümler ve Sonuç Bölümünden oluşan, toplam 58 sayfalık kısmına ilişkin, 25 / 09 / 2019 tarihinde şahsım/tez danışmanım tarafından Turnitin adlı intihal tespit programından aşağıda belirtilen filtrelemeler uygulanarak alınmış olan orijinallik raporuna göre, tezimin benzerlik oranı % 14'tür.

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 Esasları"nı 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

… / … / 2019

Dr. Öğr. Üyesi Mustafa SERT

TEŞEKKÜR

Bu çalışmanın gerçekleştirilmesinde, yüksek lisans eğitimim boyunca bana inanan ve değerli bilgilerini benimle paylaşan, motivasyonumu yüksek tutan ve ayrıca desteğini hiçbir zaman benden esirgemeyen tez danışmanım Dr. Öğr. Üyesi Mustafa SERT'e çalışmanın sonuca ulaştırılmasında ve karşılaşılan güçlüklerin aşılmasında her zaman yardımcı ve yol gösterici olduğu için içten teşekkürlerimi sunarım.

Beni bu günlere gelmem için yetiştiren, zor günlerde arkamda duran sevgili Aileme ve değerli desteklerini hiçbir zaman esirgemediklerinden dolayı minnettarım.

ÖZ

SES OLAYI TANIMA VE AKUSTİK SAHNE GERİ GETİRİMİ

Ahmet Melih BAŞBUĞ

Başkent Üniversitesi Fen Bilimleri Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

Çevresel ses olarak tanımlanan ses olayları içerisinde birçok önemli bilgiler bulunabilir. Bu çözümlenmemiş ses sinyallerinin otomatik sistemler tarafından anlamlı verilere dönüştürülmesi önemlidir. Bunun için otomatik sistemlerde ses tanıma, sınıflandırma ve geri getirimi gibi işlemlerin performanslı olması istenmektedir. Çalışma alanları bakımından; savunma sanayi, güvenlik sistemleri, çokluortam arama motorları ve nesnelerin interneti gibi popüler alanlarında bu geliştirilen sistemler kullanılabilir. Bu sinyallerin belirli bir karakteristik özellikleri bulunmaması ve ardı ardına veya örtüşen arka plan seslerine sahip olması bu problemi zorlaştıran nedenler olarak sayılabilir. Bu çalışmada; sayısal ses kayıtlarından anlamsal bilgi çıkarımı (ses olayı ve akustik sahne) ve bu bilgilerin kullanımı ile ses kayıtlarının geri getirimi problemleri ele alınmıştır. Çalışma kapsamında, başarıma katkıda bulunabileceği düşünüldüğünden ses sinyallerinden çeşitli öznitelik çıkarım yöntemleri denenmiştir. Ayrıca çeşitli derin sinir ağlar ile geliştirilmiş öğrenme modelleri incelenmiştir. Tarafımızca bilindiği kadarıyla daha önce akustik sahne sınıflandırma probleminde uygulanmamış, imgesel tanımlama problemlerinde başarılı olan uzamsal piramit veri birleştirme (SPP) yöntemi ilk defa akustik sahne sınıflandırma probleminde uygulanmıştır. Bu uygulamada, spektrogram öznitelikleri kullanılması ile başarımına katkıda bulunulduğu görülmüştür. Tanıma ve sınıflandırma çalışmalarından sonra çevresel ses kayıtlarının geri getirimi yöntemi üzerine çalışılmıştır. Sınıflandırma modelinin eklenmesi ile etkili bir örnekle sorgulama modeli geliştirilmiştir. Geliştirilen yöntem ile etiket bazlı arama sistemlerine kıyaslanacak sonuçlar elde edilmiştir.

ANAHTAR SÖZCÜKLER: Ses Olay Tanıma, Akustik Sahne Sınıflandırma, Akustik Sahne Geri Getirimi, Evrişimsel Sinir Ağları (CNN), Yinelemeli Sinir Ağları (RNN), Uzun Kısa Süreli Bellek (LSTM), Uzamsal Piramit Veri Birleştirme (SPP), Spektrogram, Logaritmik Mel Enerjileri, MFCC.

Danışman: Dr. Öğr. Üyesi Mustafa SERT, Başkent Üniversitesi, Bilgisayar Mühendisliği Bölümü.

ABSTRACT

SOUND EVENT RECOGNITION AND ACOUSTIC SCENE RETRIEVAL

Ahmet Melih BAŞBUĞ

Başkent Üniversitesi Institute of Science and Engineering

Computer Engineering Department

The signal of sound events, which defined in environmental sounds, may contain a lot of important information. In the computer systems, audio signals need to perform some processes such as the conversion into the meaningful data, classification and recovery of signals. The necessity of these processes is increasing day by day. It can be used in popular work areas like defense industry, security systems, multimedia search engines and internet of objects. It could be very difficult problem because sound events have no specific characteristic. Moreover in their background, there could be consecutive or overlapping sounds. In this study; we examine and develop performances of sound event identification and acoustic scenes classification. Since it is thought that it can contribute to success of study, various feature extraction methods have been tried and various deep neural network models have been used. To the best of our knowledge, method of the Spatial Pyramid Pooling (SPP), which was successful in imagery identification problems, was first applied to the acoustic scenes classification problem. In our experiments, it has been shown that it contributes to the success on spectrogram features. Moreover, in this study, we added to develop an effective Query-by-Example sound information retrieval system using acoustically and semantically similarities. We investigated; the result of effective acoustic similarity model could be compared against the result of Query-by-Keyword systems.

KEYWORDS: Sound Event Recognition, Acoustic Scene Classification, Acoustic Scene Retrieval, Convolutional Neural Networks (CNNs), Recurrent Neural Network (RNN), Long Short Term Memory (LSTM), Spatial Pyramid Pooling (SPP), Spectrogram, Log Mel Energies, MFCC.

Advisor: Assist. Prof. Dr. Mustafa SERT, Baskent University, Computer Engineering Department.

İÇİNDEKİLER LİSTESİ

ŞEKİLLER LİSTESİ

Sayfa

ÇİZELGELER LİSTESİ

Sayfa

SİMGELER VE KISALTMALAR LİSTESİ

DNN	Derin Sinir Ağları (Deep Neural Network)
RNN	Yinelemeli Sinir Ağları (Recurrent Neural Network)
CNN	Evreşimsel Sinir Ağları (Convolutional Neural Network)
MLP	Çok Katmanlı Algılayıcı (Multi Layer Perceptron)
QbE	Örnek ile Sorgulama (Query by Example)
SPP	Uzamsal Pramit Havuzlama (Spatial Pyramid Pooling)
MFCC	Mel Frekans Katsayıları (Mel Frequency Cepstral Coefficients)
F1	F1 Değerlendirme (F-measure)
ER	Hata Oranı (error-rate)
GMM	Gaussian Karışım Modeli (Gaussian Mixture Model)
HMM	Saklı Markov Modeli (Hidden Markov Model)
LSTM	Uzun Kısa Süreli Bellek (Long-Short Term Memory)
GRU	Geçitli Tekrarlayan Üniteli (Gated Recurrent Unit)
CRNN	Evrişimsel Yinelemeli Sinir Ağları (Convolutional Recurrent Neural Network)
KL	Kullback-Leibler ıraksama (KL-divergence)
DFT	Ayrık Fourier Dönüşümü (Discrete Fourier Transform)
Hz	Hertz
sn	saniye
ms	milisaniye
BP	Geriye Doğru Hesaplama (Back Propogation)
BoW	Kelime Torbası (Bag of Works)
GM	Global Maksimum Havuzlama (Global Max Pooling)
GA	Global Ortalama Havuzlama (Global Average Pooling)
FC	Yoğun Katman (Fully Connected Layer)
Max pooling	Maksimum Havuzlama (Max Pooling)
ASC	Akustik Sahne Sınıflandırma (Acoustic Scene Classification)
mAP	Hassasiyet Ölçütü (Mean Average Precision)
P@k	Ortalama Hassasiyet Ölçütü (k adet veride hassasiyet değeri)
AP	Ortalama Hassasiyet (Average Precision)

1 GİRİŞ

Ses sinyalleri; cisimlerin etkileşimi sonucu oluşan sesin iletilmesi ve saklanması amacıyla elektromanyetik enerjiye dönüştürülmüş bir elektriksel formdur. Sayısal ortamlara sesin tam olarak aktarılması; sürekli ses sinyallerinin belli zaman aralıklarında örneklenmesi sonucu ile olabilmektedir. Son yıllarda; ses sinyal verilerinin makine öğrenim algoritmaları ile sınıflandırılması ve çokluortam veri tabanları ile ilişkilendirerek içerik tabanlı otomatik arama sistemlerinin geliştirilmesi araştırmacılar tarafından ele alınan güncel konular arasındadır. Çokluortam verilerinin artması ile birlikte kapasite ihtiyaçlarının gün geçtikçe arttığı görülmektedir. Bu verilerin uygun kapasitede muhafaza edilebilmesi, kullanılabilmesi için çeşitli donanım ihtiyaçları ortaya çıkmıştır. Bu ihtiyaçlar, ses sinyallerini makine öğrenme ve sinyal analizi alanlarında önemli bir araştırma konusu haline getirmektedir.

Ses verileri; konuşma, müzik ve çevresel ses olarak belirtilen ses olayları gibi çeşitli işitilebilen seslerin alt kategorilerine ayrılabilmektedir. Ses olayları tez çalışması kapsamında ele alınacak konudur. Ses olayları, müzik ve konuşma verilerinden farklı olarak çevredeki nesnelerin titreşimi sonucunda oluşan ses sinyalleridir. Konuşma sesleri; insanların vokal yolu ile ürettikleri dilsel içerikler içeren seslerdir. Bu sesler karakteristik özellikleri ve spektral dağılımları diğer ses türlerine göre farklılık göstermektedir. Müzik ise melodi, ritim gibi tekrarlanan sabit kalıp yapılarına sahiptir. Çevresel sesler bünyesinde yapısal olmayan birçok karakteristik özellik ve iç içe geçmiş birçok sesleri barındırması ile üzerinde birçok bilgi bulundurabilmektedir. Zorlu bir problem olarak görülmesi ile birçok çalışmalara konu olmuştur. Bu ses sinyalleri; ortam, faaliyet, durum gibi akustik sahneler olabildiği gibi, bu akustik sahnelerde gerçekleşen kaynağın ürettiği titreşim de olabilir. Çevresel seslerin sınıflandırılması ve tanımlanması sayesinde ses kaynağı hakkında çeşitli bilgiler edinilebilir. Örnek olarak; araç sesinden ses kaynağının trafik ortamına ait olduğu, adım seslerinden kaldırımda yürüme ve çay bardağında kaşığın çay karıştırma olayında çıkardığı sesler gibi birçok sahne ve olay bilgisi ses sinyallerindeki içeriklerden çıkarım yapılarak tanımlanabilmektedir.

Günümüz internet çağında, iletişimin yoğun yaşandığı, veri saklama ve veri paylaşımı gibi işlemlerin çoğalması ile birlikte, çokluortam verilerinin muhafaza edilip işlenmesi için çeşitli gelişmiş akıllı sistemlere ihtiyaç duyulmaktadır. Bu akıllı sistemler sayesinde doğru sonuç elde edebilecek hızlı arama, sınıflandırma ve veri geri getirimi gibi birçok sistemlerin ihtiyaçları karşılanabilir. Kontrolsüz ortam özellikleri ve örtüşen çevresel seslerin değişken karakteristik çeşitliliği nedeniyle, bu seslerin bilişim alanında makine öğrenme yöntemleriyle otomatik tanımlanıp anlamlı bilgilerin çıkarımı yapabilmesi güç bir problemdir. Akademik alanda bu problem üzerinde birçok güncel araştırma konuları yer almaktadır. Ses olayları ile ilgili içerik tabanlı çokluortam geri-getirimi [1], çokluortam veri tabanlarında içerik temelli indeksleme, mobil cihaz [2], sağlık alanında göze çarpmayan izleme, gözetleme ve tespit sistemleri [3], robot ve akıllı şehir gibi çeşitli alanlarda uygulama ve çalışmalar yapılmasından ötürü literatürde yüksek önem potansiyeline sahiptir. Bu nedenle, ses sinyallerinin otomatik sistemler tarafından performanslı bir şekilde tanımlanabilmesi ve sınıflandırılabilmesi önemlidir. Bilişim alanında otomatik olarak ses olay analizi, sınıflandırma ve öğrenme modeli üzerinden içerik taraması yapılarak ses veri getirimi gibi kavramsal gereksinimlere ihtiyaç duyulmaktadır.

1.1 Ses Olayı ve Akustik Sahne Sınıflandırma

Ses olayları; bir akustik sahnede gerçekleşen faaliyetlerin tanımlanması ve anlaşılması için çok iyi bir tanımlayıcı olarak kabul edilen belirleyici bir etikettir. Bilinen olayların akustik veya etiket anlamı ile çıkarım yapılıp diğer gerçekleşen olaylar ile ilişkilendirilmesi sağlanabilir. Ayrıca gerçekleşen bir olayın arka planında oluşan seslerden gerçekleştiği sahneyi anlama konusunda etkili bir yardımcı bileşendir. Bu konu ile ilgili örnek verilecek olunursa; yanan ocakta ateş sesi, bıçak ile kesim sesleri, yağ dökme sesi ile bir restoranın yemekhanesinin mesai saatinde olduğu ve çalışıldığı bilgisi edinilebilir. Buradaki akustik sahne (yemekhane) ve karakteristik ses olayları (ateş sesi, kesim sesi, yağ dökülme sesi) ile tanımlanabilir.

Akustik sahneler; çevredeki nesnelerin titreşimi sonucunda oluşan ses sinyallerinden sesin bulunduğu yerin etiketi (“park”, “ev”, “ofis”), durumu (“toplantı”, “trafik”) veya yapılan faaliyetler (“yemek pişirme”, “yürüme”, “koşma”) gibi anlamların karşılığı olarak gelmektedir. Bu ses sinyallerinden bu ve bunun gibi anlamların önceden kategorize edilmesi ile diğer kategorize edilmeyen ses sinyalleri hakkında

bir anlam çıkarabilme işlemi akustik sahne sınıflandırma problemi olarak tanımlanabilir.

Ses olayı tanıma ve akustik sahne sınıflandırma problemlerinde, otomatik sınıflandırma sistemleri sayesinde bulunulan çevreye dair birçok önemli bilgiler edinilebilmektedir. İnsan; duyu organı olan kulağının aracılığı ile; araç sesinden trafik olayı olduğunu, köpek sesini, oynayan çocukların seslerinden çocuk parkının sesini ve adım seslerinden kaldırımda yürüme gibi olayların gerçekleştiğini algılamada oldukça yeteneklidir. Bu eşsiz yeteneğin, bilişim alanında makine öğrenme yöntemleriyle, ses olayları sonucunda gerçekleşen bir sahne hakkından veri çıkarımını otomatik olarak yapabilme becerisinin kazanılması ve bu sınıflandırma tahmininin doğruluğunun geliştirilmesi gibi konular güncel araştırma konuları olmuştur. Ses olayları içerisinde birden fazla sesin aynı anda oluştuğu veya bir olayın gerçekleştiği sırada başka bir ses olayının meydana gelmesi üzerine bu akustik sahnelerin otomatik yöntemler ile kategorize edilmesi önemlidir. Potansiyel kullanım alanları nedeniyle, son yıllarda bu konu üzerindeki araştırmalar hızla artmaktadır.

1.2 Ses Sahne Geri Getirimi

Teknolojinin gelişmesi ve verinin internet ortamında çoğalması sonucu, bilgiye erişim ile ilgili problemler ortaya çıkmış, bu konu üzerinde ise çeşitli araştırma ve geliştirme yapılmıştır. Günümüzde metin içeren verilerin erişimiyle ilgili problemlerinin çözümü tam olarak sağlanamamışken, çokluortam veri tabanlarında bulunan işitsel bilgiye erişimdeki sorunların çözümü beklemektedir. Bilgisayar sistemlerinin daha çok veri arayabilme, istenilen veriye hızlı ve kolayca erişebilme yetenekleri kazanması, arama motorlarının geliştirilmesi konusunda büyük önem verilmesine neden olmuştur. Ayrıca geri getirim için geliştirilen sistemler ile son yıllarda çokluortam veri tabanlarının önemlilik derecesi artmıştır. İstenilen veriye erişebilmek için verilerin tutulması kadar o veriye erişiminin kolay olabilmesi için etiketlenmesi, indekslenmesi, sınıflandırılması, yapılandırılması gibi işlemleri yapılabilmesi de önemlidir.

Ses sinyallerinde veri geri getirimi, çokluortam veri tabanlarından ihtiyaç duyulan bilgileri elde etmek için ses sinyalleri üzerinde çalışmaktadır. Bu tür sistemler

kullanıcılara veri tabanlarından istenilen verilerin geri getiriminde tatmin edici sonuç vermelidir. Çokluortam veri tabanlarında etiketlenmemiş, öznel olarak etiketlenmiş ya da hatalı etiketlenmiş ses sinyalleri olabilir. Bu sebeple geri getirim sistemleri, son kullanıcılara sağladığı hizmette aksaklık olabilmektedir. Bu tür sorunların çözülme isteği bu problemin araştırmacılar tarafından güncel olmasını sağlamaktadır. Araştırmacılar, daha kaliteli ve tatmin edici sonuçlar elde edebilen içerik tabanlı arama sistemleri üzerinde çalışmaya yönlendiği görülmektedir. İçeriğe dayalı bilgiler ile arama motorlarının başarım etkinliğini üst seviyeye taşınabilir. Günümüzde birçok arama motorları ve veri geri getirim sistemleri, içerik tabanlı verilerin getirimi üzerine çalışmaktadır.

Ses bilgisi alma uygulamalarında işitsel benzerlik sorguda gönderilen ses verisi ile benzer sesleri getirme işlemlerinde önemli bir yer tutmaktadır. Otomatik konuşma tanıma, müzik bilgisi alma, ses segmentasyonu ve çevresel ses alma uygulamaları ses verisinin geri getirim problemlerindeki ana başlıklar altında sayılabilmektedir. İçerik tabanlı ses geri getirimi sistemlerinin temel amacı, ses arama motorlarında işitsel benzerliği kullanarak algısal olarak benzer ses içeriğinin tanımlanmasıdır. Müzik sesleri üzerinde bulunan bilgilerin erişiminde işitsel benzerliği ritim, tını, akor, vb. içeriklerden incelenebilmektedir. Aynı şekilde, konuşma tanımlama problemlerinde kullanılan işitsel benzerliği; tonaliteyi, perdeyi, frekans sıklıkları gibi özellikler ile incelenebilir. Fakat çevresel sesler için; ne tür bir benzerlik sisteminin arandığının bilinememesi bu problemi daha zahmetli bir problem olarak tanımlamaktadır. Dahası, ses olayları, kontrolsüz ortam özelliklerine sahip olabilir ve sesler birbiri ile örtüşebilir. Bahsedilen bu zorlaştıran nedenlere örnek verecek olursak; çocuklar tarafından parkta oynarken çıkardıkları “çığlık” sesleri ve ağaçlarda “kuş ötmesi” seslerinin örtüşmesi ile sistemin sadece “kuş sesi” olarak tanım yapması, sistemin çalışmasında hata olarak tanımlanabilir. Çevresel seslerdeki benzer belirsizlik ve zorluklara bakıldığında; müzik ve konuşma sesleri ile karşılaştırıldığında bu tür seslerden çok farklı kategoride olduğu görülmektedir.

İçerik Tabanlı aramalar ile daha etkili dizin oluşturma, kaliteli sorgu sonuçları ve hatalı etiketlenmelerde ortaya çıkan sorunların çözümü üzerinde çalışılabilir. Sorgulanan ses verisi, içerik tabanlı aramalarda akustik ve semantik benzerlik sistemlerinin ortak çalıştırılması son dönemlerde popüler konulardan sayılmaktadır.

1.3 Problem Tanımı

Çokluortam (resim, ses, video ve metin) içeriğine sahip sistemlerden anlamsal bilgilerin çıkarılması uzun zamandır zorlu ve popüler bir araştırma alanı olmuştur. Ayrıca bu verilerin boyutlarının zamanla büyümesi analiz, sınıflandırma ve geri getirim maliyetlerini de büyük ölçüde artırmıştır. Bunun dışında insan sesi ve çalgı aleti dışında kalmış kaynakların ürettiği sesler olan ses olayları; müzik ve konuşma tanımlama problemi araştırmalarına kıyasla üzerinde çalışılması yetersiz kalmış bir konudur. Müzik bilgisinde çıkarım yapılabilmesi için şarkı zaman içindeki ritmi ve bu eserde kullanılan enstrümanın yardımı ile karakteristik özellikler bulunurken, konuşma seslerinde ise sesin karakteristik spektral dağılımından fonetik yapısına kadar birçok değişik karakteristik özellikte tanımlayıcılara sahiptir [4]. Çevresel seslerin müzik ve konuşma seslerine nazaran sesin kaynağı dışında ayırt ediciliğini sağlayacak bir tanımlayıcı bulunmasının zorluğundan dolayı üzerinde çalışılması zorlu bir görev olmaktadır.

Çokluortam verilerinin bilgisayar ortamlarında son yıllarda gerçekleşen donanımsal geliştirmeler sayesinde büyük sistemlerde eğitim işlemine alınması oldukça kolaydır. Son yıllarda popüler olan derin sinir ağları (DNN [5]), yinelemeli sinir ağları (RNN [6]) ve konvolüsyonel sinir ağları yöntemleri ile geliştirilen popüler derin öğrenme mimarilerinin (AlexNet [7], VGG [8], ResNet [9] vb.) kullanımı ile birlikte çokluortam verileri üzerinde eğitim aşamasında başarılı sonuçlar elde edildiği görülmektedir. Büyük veri kümelerinin bu mimariler ile eğitimi için paralel mimarisine sahip sistemlere erişimin kolay olduğu bu günlerde daha etkili ve başarılı bu öğrenim çalışmaları gerçekleştirilmiştir.

Özetle, bu tez çalışmasında çevresel ses kliplerinde bulunan ses olaylarının tanımlanabilmesi ve akustik sahnelerin sınıflandırılabilmesi için, çeşitli öznitelik çıkarım ve sınıflandırma tekniklerinin performansa katkısı incelenecektir. Ayrıca eğitim maliyetini düşürecek mimarilerin geliştirilmesi amaçlanmaktadır. Geliştirilen başarılı mimari ile ses sahne geri getirimi probleminde kullanılarak performans katkısı gözlemlenecektir.

1.4 Tezin Amacı ve Kapsamı

Teknolojinin hızla gelişmesi ve bilgiye internet üzerinden erişilmeye çalışıldığı bu dönemde hedeflenen bilgilere hızlı erişim ile ilgili problemler ortaya çıkmış, birçok araştırmacı bu konular hakkında çeşitli çalışmalar yapmıştır. Konuşma ve müzik sesleri üzerinde geliştirilmiş birçok uygulama ve araştırma bulunurken çevresel sesler üzerinde yapılmış çalışmaların azlığı bu tez üzerinde çalışılması konusunda ana etken olmuştur. Çevresel ses sinyallerinde arka planda bulunan birçok kaynaktan oluşan ses sinyalleri birbiri ile örtüşebilmesi sonucu bilgisayar ortamlarında otomatik öğrenme sistemleri içinde zorlaşan bir problem haline gelmektedir. Bu tip ses sinyallerinde içeriklerinde arama yapmak, sezimlemek ve daha erişilebilir hale getirmek için doğru bir şekilde sınıflandırılması gerekmektedir.

Bu tez kapsamında; çevresel ses sinyallerinin bilgisayar ortamlarında otomatik olarak sınıflandırılması, tanımlanması ve geri getirimi üzerinde çalışılmıştır. Sinyal işleme ve makine öğrenme konularında çalışan araştırmacıların çevresel sesler üzerindeki çalışmalar ile müzik ve konuşma ses sinyalleri üzerindeki çalışmaları kıyaslandığında yeteri sayıda olmadığı görülmektedir. Bu sebeple bu zorlu problem ile ilgili çözümler üretmek ve konu hakkında gelecekte referans olacak çalışmalar yapılması amaçlanmıştır. Çalışmamızda amaç, çevresel seslerin daha başarılı bir şekilde tanımlanabilmesi işleminin gerçekleşmesidir. Sınıflandırma ve tespit işlemleri için kullanılan yöntemlerin performansları karşılaştırmalı olarak çalışmamızda sunulmuştur. Ayrıca akustik sahne sınıflandırma problemlerinde bilgimiz dâhilinde daha önce kullanılmamış, görsel tanımlama problemlerindeki performans başarısı elde etmiş olan Spatial Pyramid Pooling (SPP) yöntemi kullanımı ile ilgili çalışmamızdan bahsedilecektir. Ses olay tanıma problemi ile ilgili çalışmalarımızın ardından bir sonraki çalışma olan geri getirim (retrieval) sistemlerinin çevresel ses sinyalleri içinde bulunan akustik sahneler üzerindeki performansı irdelenecektir. Zamanla büyüyen çokluortam verilerinden ses sinyal dosyalarının sayısal ve kapasite olarak artması karşısında, son kullanıcılarda bu artışa paralel olarak arama motorlarının performansının artmasını talep etmektedir. Bu talep karşısında ses verilerinin çokluortam veri tabanlarında bulunan diğer ses dosyalarının arasından kolayca arama ve hemen erişilebilmesi önemli bir konudur. Bu verilerin bilgisayar ortamlarındaki dizinleme performansının arttırılması ve ilgili

arama motorlarının geliştirilmesi amacı ile çevresel ses verilerinin sistemler tarafından tanımlanabilmesi sonrası geri getirim işleminin yapılabilmesi üzerinde çalışılmıştır. Bu çalışma kapsamında anahtar kelime ile aramadan ziyade örnekle sorgulama (QBE) çalışmamız örnek bir ses verisi ile arama yapabilen sistemlerin geliştirilmesi noktasına odaklanmıştır.

Son yıllarda gelişen teknolojilerin kullanımı ile güncel, uygulanabilir başarılı bir ses olayı tanıma ve geri getirim sistemi ortaya koyabilme ve ayrıca sonrasında bu konu ile ilgili gelecek çalışmalara referans olabilecek bir eser bırakmak bu tez çalışmasında en büyük hedefimizdir.

1.5 Araştırma Soruları

- Ses olay sinyalleri üzerinde farklı öznitelik çıkarım yöntemlerinin kullanımı eğitim başarısını nasıl etkiler?
- Ses sinyallerinin öznitelik çıkarım aşamasında pencere boyutunu kısaltıp daha ayrıntılı öznitelikler elde edilebilir. Bu öznitelikler ile geliştireceğimiz mimarilerde kullanımı sonucu performans katkısı nasıl olabilir?
- Yinelemeli sinir ağları algoritmalarına eklenen LSTM hafıza hücrelerinin kullanımı ile ses sinyallerinde çeşitli zamansal bilgilerin çıkarımı yapılabilir. Bu algoritma kullanımıyla geliştirdiğimiz mimarimiz ile sınıflandırma başarımı elde edilebilir miyiz?
- İmgesel sınıflandırma algoritmalarında başarılı olan SPP yönteminin akustik sahne sınıflandırılma problemlerinde kullanımı sonucu performans katkısı nasıl olacaktır?
- Ses kayıtlarındaki ardı ardına veya üst üste gelen sinyallerden sıralı bilgiler yakalamasında GRU, LSTM gibi yineleme sinir ağları algoritmalarının kullanımı efektif bir sınıflandırma avantajı sağlayabilir mi?
- Sorgu olarak ses sinyali gönderdiğimiz bir sistemde benzer ses sinyallerinin geri getirimi geliştirdiğimiz başarılı akustik sahne sınıflandırma mimarisi ile sağlanabilir mi? Daha efektif bir akustik benzerlik modeli geliştirilmesi için anlamsal benzerlik modeli ile birlikte kullanımı akustik sahne geri getirimi başarımını nasıl etkilenecektir?

1.6 Tez Organizasyonu

Bu tez çalışması altı bölümden oluşmaktadır. Diğer bölümlerin organizasyonu şöyledir; Bölüm 2'de konu ile ilgili bugüne kadar yapılmış ilgili çalışmalar yer almaktadır. Bölüm 3'de tez çalışması boyunca kullanılan genel tanımlamalardan bahsedilecektir. Bölüm 4, ses olay ve akustik tanıma ve Bölüm 5'de ise ses sahne geri getirimi çalışmaları anlatılmaktadır. Son olarak Bölüm 6'da ise sonuçlar ve gelecek çalışmalar sunulmaktadır.

1.7 Katkılar

Bu tez çalışmasındaki amaç, çevresel ses kategorisinde bulunan ses olayı ve akustik sahne verileri üzerinde farklı öznitelik çıkarım teknikleri ve derin öğrenme mimarileri kullanımı ile ses olayı tanıma ve akustik sahne sınıflandırma başarımının arttırılmasını sağlamaktır. Ayrıca akustik sahne sınıflandırma için geliştirilen etkin öğrenim modeli ile çokluortam sistemlerde akustik sahne geri getirimi için etkili bir içerik tabanlı arama sistemlerinin geliştirilmesini sağlamaktır. Sınıflandırma performansına ek olarak eğitim maliyeti performansı da göz önünde bulundurulmuştur. Çalışmalar sırası ile TUT Sound Event 2017 [10], TUT Urban Acoustic Scenes 2018 [11] ve TAU Urban Acoustic Scenes 2019 [12] veri kümeleri üzerinde gerçekleştirilmiştir.

Bu çalışmanın katkıları aşağıdaki maddelerde içermektedir:

- Ses olay veri kümeleri üzerinde çeşitli öznitelik çıkarım tekniklerinin kullanımı ile eğitim başarımının gözlemlenmesi
- Eğitim aşamasında güncel ve popüler derin sinir ağları mimarilerinin eğitim başarımına etkisi incelenmesi
- SPPnet gibi görsel sınıflandırma problemlerinde kullanılan havuzlama katmanının incelenip bu problemde kullanılarak eğitim başarımının incelenmesi
- Eğitimde geliştirilen mimarinin ses sahne geri getirimi problemindeki başarımının incelenmesi

Bu tezde aşağıda sunulan yayın çalışmaları yapılmıştır:

- Basbug, Ahmet-M., Sert, M. Acoustic Scene Classification Using Spatial Pyramid Pooling With Convolutional Neural Networks, The 13th IEEE International Conference on Semantic Computing (ICSC2019), 30 Ocak – 1 Şubat, Newport Beach, California, USA, s.128-131, 2019.
- Basbug, Ahmet-M., Sert, M. Analysis of Deep Neural Network Models for Acoustic Scene Classification,, IEEE 27th Signal Processing and Communications Applications Conference (SIU2019), 26-28 Nisan, Sivas, Turkey, s.128-131, 2019.

2 İLGİLİ ÇALIŞMALAR

Ses sinyalinden ses olay tanıma ve akustik sahne sınıflandırma konusunda yapılan çalışmalar ağırlıklı olarak makine öğrenme problemi ile sınıflandırıcı mimarisinin oluşturulması ile ilgili olmasının yanı sıra ses verisinden öznitelik çıkarımı konularına da yoğunlaşmaktadır. Ayrıca akustik sahne ve ses olay geri getirim problemlerinde ise araştırmacılar son yıllarda sinyal işleme, anlamsal veri çıkarımı, makine öğrenme gibi alanlarda çalıştıkları görülmektedir. Bu konular içinde günümüze kadar bu alanlarda yapılan bazı çalışmalar aşağıda özetlenmiştir. Bu bölümde sırası ile ses olay tanıma, akustik sahne sınıflandırma ve ses sahne geri getirimi ilgili alt bölümlerde anlatılacaktır.

2.1 Ses Olay Tanıma

Ses olay tanıma problemlerinde yapılan çalışmalardan ilk olarak bahsedilecek araştırma; Piczak [13] tarafından yapılan, derin öğrenme tabanlı önerdiği yöntemin üzerinde çalıştığı çevresel sesler içeren kısa ses kliplerinin otomatik sınıflandırılması üzerinedir. Bu çalışmada, farklı veri kümeleri üzerinde çıkarılan Mel Frekans Kepstrum Katsayılarından (MFCC) yararlanan Piczak, derin öğrenme algoritması olarak da CNN mimarisini kullanmıştır. Çalışmasında karmaşık olmayan kısa ses kayıtları içeren bir veri kümesi kullanmış olmasına rağmen güncel çalışmalar ile kıyaslanabilecek sonuçlar elde ettiği görülmüştür. Ayrıca bu çalışmasıyla CNN mimarisinin ses olay tanıma problemlerindeki sınıflandırma başarımının yüksek olduğu gözlemlenmiştir.

Gorin vd. [14], ses olay tanıma problemine CNN mimarisini uygulamıştır. Ayrıca, CNN mimarilerinin eğitim esnasında büyük miktarda veriye ihtiyaç duyduğunu savunan araştırmacılar veri kümesi üzerinde dönüşümler uygulayarak eğitim için ek kaynaklar üretilebileceğini göstermişlerdir. Bu yapay veri büyütme ile birlikte oluşturulan iki katmanlı CNN modelinin eğitim sürecini işleyen araştırmacılar F1 değerlendirme sonucu %38,1 oranı ve 0,84 hata oranı elde etmişlerdir. Araştırmacılar çalışmalarındaki dezavantaj olarak veri üzerindeki kısa ses olayları göstererek, bu kısa olayları tespit etmenin zor bir problem olduğunu belirtmişlerdir. Sonuç olarak bu çalışma ile veri çoğaltma teknikleri ile CNN tabanlı yöntemlerin başarımlarının artırılabileceği gösterilmiştir.

Diğer bir çalışmada, Schröder vd. [15] Gauss Mixture Model (GMM), GMM-Saklı Markov Modeli (HMM) ve melez bir derin sinir ağı öğrenim modeli olan DNN-HMM sistemini geliştirmiştir. Çalışmada, öznitelik olarak MFCC, *Gabor Süzgeç Kümesi* ve *Non-negatif Matrix Factorization* kullanılması tasarlanmıştır. Araştırmacılar bu öznitelikler ile geliştirilen GMM, GMM - HMM melez modeli ve DNN-HMM melez modelleri üzerinde problemi çözmeye çalışmışlardır. GMM-HMM modeli üzerinde GSK özniteliklerinin kullanılması sonucu elde edilen sonuçlar diğer kullanılan modellere kıyasla en iyi sonuç olarak görülmektedir. Fakat araştırmada *Gabor Süzgeç Kümesi* özniteliklerinin derin öğrenme algoritmalarında kullanılmaması kıyaslama açısından bir eksiklik olarak görülebilir.

Birçok örüntü tanıma probleminde olduğu gibi, kullanılan öznitelikler ses olay tanıma probleminde de önemlidir. Bu konuyu ele alan bir çalışmada, uzamsal ve harmonik ses öznitelikleri ses olay tanıma probleminde kullanılmıştır. Adavanne vd. [16] çalışmasında, insan kulağını model alan çift kanallı sesler oluşturulması ile ses olay sezimi performansı arttığı gözlemlenmiştir. Çalışmada öznitelik olarak logaritmik mel-bandı enerjisi, harmonik öznitelikler ve probleme özgü olarak tasarlanan varış zaman farkı (Time Difference of Arrival - TDoA) öznitelikleri kullanılmıştır. Eğitim modeli için ise iki katmanlı 32 birimli Long-Short Term Memory (LSTM) mimarisi tasarlanmıştır. DCASE veri kümesi üzerindeki deneylerde, LSTM ile Mel özniteliklerinin ev içi ses olaylarının seziminde başarılı olduğunu, aynı modelin Mel ve TDoA özniteliklerinin birlikte kullanımında ise ev dışı (çevresel) ses olaylarının seziminde daha başarılı oldukları görülmektedir. Adavanne vd. [16], bir başka çalışmasında çift kanallı seslerden düşük seviyeli logaritmik mel-bandı enerjisi, otomatik ilinti ve genelleştirilmiş karşılıklı ilinti olmak üzere 3 farklı özniteliklerinin çıkarımı yapılarak evreşimli çift yönlü yinelemeli sinir ağları mimarisi oluşturulmuştur. Burada CNN ve RNN yapısı birleştirilerek oluşturulan öğrenme ağ modeli ile çok kanallı ses özniteliklerinin tek-kanallı ses özniteliklerine kıyasla daha başarılı olduğu ifade edilmektedir. Ayrıca, çok katmanlı öğrenme ağ modeli çok kanallı seslerdeki ses olaylarını tanımada daha performanslı bir yapı olduğu savunulmaktadır.

Li vd. [17], ses sinyallerinden ses olay sınıflandırma problemi için DNN ile çıkarılan derin ses öznitelikleri ile geliştirdikleri LSTM-RNN yöntemi ile otomatik ses

sınıflandırma başarımını gözlemlemişlerdir. Bu çalışmada öznitelik çıkarımı adımlarında elde ettikleri derin ses özniteliklerinin daha efektif bir şekilde karakterize edilebileceği savunulmuştur. Öte yandan, LSTM-RNN modelinin ses sinyalindeki zamansal olarak ardı ardına gelen veya üst üste gelen sinyallerden sıralı bilgiler yakalamasıyla efektif bir sınıflandırma avantajından bahsedilmiştir. Zhou [18], çalışmasında insan kulağının çok sesli iç içe geçmiş ses olaylarını başarılı şekilde ayrıştırabilmesinden esinlenerek ses olay tanıma sistemi üzerine çalışmıştır. Bu sistemde logaritmik *mel* enerji özniteliklerini LSTM yapısını modellemiştir. Ayrıca ses veri özniteliklerini farklı füzyon stratejileriyle üç farklı kanaldan oluşacak şekilde genişletilmesinin geliştirdiği modelde kullanımı sonucu daha performanslı bir yapı geliştirdiğini göstermiştir. Çalışmasının sonucunda performans ve hata payı kayda değer şekilde artırdığı gözlemlenmiştir.

Adavanne vd. [19], araştırma konusu çevresel ses kategorisinde yer alan ses olayı ile ilgili çalışmasında elde edilen kuş seslerinin tanımlanması problemini ele almışlardır. Bu çalışmada, evreşimli çift yönlü yinelemeli sinir ağları modelini görünmeyen veriler üzerinde güçlü bir öğrenim modeli olmasına yönelik tasarlanmıştır. Eğitim verisi üzerindeki aşırı uyum probleminin önlenmesi için bırakma oranı (dropout) ve erken durma (early stopping) parametreleri üzerinde çalışılmıştır. Eğitim modeli için logaritmik *mel* enerjileri içeren özniteliklerin yanı sıra baskın frekans (*dominant frequencies*) özniltelikleri kullanılmıştır. Ayrıca iki öznitelik verilerinin birleşimi de eğitim modeline gönderilerek analizi sağlanmıştır. CNN algoritmasının yüksek seviyede zamansal ve spektral değişimlerden etkilenmeyen öznitelik çıkarımı ile RNN algoritmasının yüksek performansta sınıflandırma yapabilme yeteneğinin ortak bir çalışmada kullanımı sonucunda başarılı sonuç elde edilmesi üzerine söz konusu çalışma araştırmacıların dikkatini çekmiştir. Doğal ortamlarda meydana gelen ses olaylarının frekans içerikleri ve zamansal yapısındaki farklılıklarına dikkat çeken Çakır vd. [20]; CNN ile bu değişmeyen yerel spektral ve zamansal varyasyonları elde ederek ses sinyalindeki uzun vadeli geçici bağlamları sınıflandırmada değerlendirmek istemiştir. Bu motivasyon ile; günlük ses olaylarından oluşan dört farklı veri seti üzerinde CNN ve RNN kombinasyonunu sağlayarak CRNN tabanlı bir yöntemi ses olay sezimi problemine uygulamış ve CNN, RNN, GMM yöntemlerine kıyasla akustik model tanımlama uygulamalarında

başarılı sonuçlar elde etmiştir. Ayrıca bu bileşim ile işbirliğinde kullanılan her bir modelinin bireysel zayıf yönlerinin üstesinden gelebildiğini görülmüştür.

Han vd. [21], derin öğrenme algoritmalarının kullanımının ses sinyalleri ile ilgili araştırmalarda uyumlu bir şekilde geliştirilmesinin araştırma problemlerine olumlu katkı göstereceğini düşünmektedirler. Bu sebeple çeşitli önişleme yöntemlerinin yanı sıra uzamsal bilgiler içeren ses kayıtlarından en iyi şekilde yararlanmak için öğrenme ağ yapısı geliştirilmiştir. Çalışmada önerilen ağ mimarisi ve ön işleme yöntemleri öğrenme karakteristiğini geliştirdiği gibi kullanılan topluluk modeli ile birlikte hata oranının düştüğü gözlemlenmiştir.

Adavanne vd. [22], bir başka çalışmada, derin öğrenme algoritmalarından olan CNN mimarisi ile farklı çift kanallı ses öznitelikleri kullanılarak ses olay tanımadaki başarımları incelenmiştir. Üç katmanlı 128 filtreli 3x3 konvolüsyonel katmanları içeren mimariye ek olarak 2 katman 32 birimli Çift-yönlü GRU modellerini ekleyerek üç farklı stereofonik öznitelikler üzerinde çalışmak üzere bir mimari tasarlamışlardır. Veri kümesinden elde ettiği öznitelikleri, ayrı ayrı geliştirdiği çok kanallı ağ mimarisinde yapılan deneyler sonucu çift kanallı seslerin, tek kanallı seslerden daha iyi performans verdiği göstermektedir.

2.2 Akustik Sahne Tanıma

Bu alanda yapılan çalışmalardan bahsedilecek olunursa; Bae vd. [23], derin sinir ağlarının zamansal bilgileri tam olarak kullanamaması nedeni ile iki ayrı alt ağlar ve bir üst ağlardan oluşan bir eğitim modeli kombinasyonunu tasarlayarak sıralı bilgilerin otomatik sınıflandırılmasını araştırmışlardır. Bu kombinasyon CNN mimarisi zamansal spektrogram yerleşimini öğrenmesi ve LSTM mimarisi sıralı bilgileri ardışık ses özelliklerinden temin etmesi sağlanmıştır. Bu kombinasyonun sağladığı avantaj ile konvansiyonel DNN, CNN ve LSTM mimarilerine karşı daha yüksek başarı elde edilmiştir. Valenti vd. [24] ise, bu kısa çevresel ses dizilerinin akustik sahne sınıflandırması problemini incelemiş, öznitelik olarak çıkardığı log-mel spektrogram değerlerini CNN öğrenme mimarisinde eğitimini yapmıştır. Eğitim sırasında sistem doğrulama ile eğitim performansının yükselmesi hedeflemiştir. Belirli şartlar altında genelleme performansını izlemeden eğitilmesi sonucunda doğruluk iyileştirilmesi elde edilmiş, böylece eğitim verisindeki eksikliği nedeniyle

genelleme performansının darboğaz yapmasını önledikleri görülmektedir. Wei vd. [25], ses kayıtlarını akustik sahne ve olaylara göre sınıflandırmak için *MFCC*, *Smile6k* ve *Smile983* özniteliklerini derin sinir ağları algoritmalarıyla eğitimini incelemiş, RNN algoritmasıyla oluşturulan zamansal modellerin daha üstün performans gösterdiğini gözlemlemiştir. Fakat *Smile6k* ile elde edilen büyük veri seti ile DNN modelleri, zamansal modellerden daha üstün performans elde etmiştir.

Kukanov vd. [26], akustik sahne sınıflandırma problemi için evrişimsel yinelemeli sinir ağları algoritmasının sınıflandırmadaki başarısını incelemişlerdir. Çalışmada temel aldığı sisteme kıyasla %11 civarında daha yüksek doğruluk başarımı elde edildiği gözlemlenmiştir. Modelinde evreşimli katmanlarla ilgili *mel* özniteliklerinden faydalı çıkarımları yapmakta ve dengesiz ses bozulmalarını azaltırken, zamansal bağlam değişikliklerini öğrenebilmesi için ekledikleri Gated Recurrent Unit (GRU) katmanları sayesinde başarılı sonuçlar elde edinilebileceği gözlemlenmiştir. Yine evreşimli katmanlarının yüksek seviyede öznitelik çıkarabilme özelliğini kullanan Jallet vd. [27], akustik sahnelerin uzun vadeli zamansal bağlamını modelleyebilmek için kapılı tekrarlayan katmanlar kullanmıştır. Uyguladıkları evrişimsel yinelemeli sinir ağları mimarisinde GRU katmanları, verilmiş işlem ile ilgili eski çerçevelerdeki özniteliklerin ipuçlarından öğrenme aşamasında yararlanabilmektedir. Böylece ses olaylarındaki akustik sahnelerde meydana gelen çeşitli ses olaylarından bilgi toplayarak, öğrenme için uzun vadeli geçici modelleme oluşturabildikleri ve bu yöntem ile birlikte öğrenimde doğruluk başarımının yükseldiği gözlemlenmiştir.

Kong v.d. [28] görsel işleme ve sınıflandırma problemlerinde kullanılan *AlexNetish* ve *VGGish* içeren CNN mimarisi ile sınıflandırma performansını karşılaştırmıştır. Derin yapıda olması ve iyi bir performans gösteren mimari ile kaydedildiği ortamı karakterize eden ses kayıtlarının akustik sahne sınıflandırma probleminde başarılı bir şekilde çalıştığı gözlemlenmiştir.

2.3. Ses Olay ve Sahne Geri Getirimi

Çokluortam sistemlerde; içerik bazlı ses arama ve geri getirim yöntemleri üzerine son yıllarda müzik, konuşma tanımlama, çevresel sesler gibi alanlarda birçok araştırma yapılmıştır [29][30][31]. Konuşma ve müzik uygulamalarında kullanılan ses sinyali ile veri geri getirimi uygulamaları dışında çevresel seslerin bulunduğu

ses sinyallerinden veri çıkarımı ve geri getirimi çalışmaları ve uygulamaları az da olsa araştırmacılar tarafından üzerinde çalışılmakta olan konulardan birisidir. Son yıllarda ev güvenlik uygulamaları, savunma sanayi, dinleme uygulamaları ve video geri getirimi uygulamalarında bu problem ilgi odağı olmuştur [32][33].

Mesaros vd. [34], ses olay sinyal verilerini içeren veri tabanında ses ve etiket arasındaki ilişkilendirme sorunu üzerinde çalışmıştır. Ses verileri üzerindeki kullanıcılar tarafından öznel olarak etiketlenmesi sonucu ses olaylarından oluşan dosyalar üzerindeki etiketlenmiş bilgilerin çok çeşitlilik sorunu ortaya çıkmış ve sistemdeki bu dağınıklık sonucu bilgiye erişimin zor olacağına değinmişlerdir. Sesteş, eş anlamlı sözcükler ve çoğul kelimeler ile etiketlenen bu ses verileri ile otomatik sistemlerin geri getirim işlemleri için kullanacağı indekslerin düzenlenmesi konusunda ses verilerinin bu veri tabanı sistemi üzerinde otomatik olarak objektif bir şekilde yeniden etiketleme olasılığı üzerinde çalışılmıştır. Bu sorunun çözümü sırasında ses verilerinin MFCC ile öznitelik çıkarımından yararlanan Mesaros; öznitelik vektörlerini GMM algoritması ile geliştirdiği yöntem süreçlerinden geçirerek işitsel benzerliğini araştırmıştır. İşitsel benzerliği için sorgulanan ve veri tabanında bulunan ses verilerinden GMM algoritması ile elde ettiği değerlendirme sonuçlarını simetrik bir Kullback-Leibler ıraksama (KL-divergence) algoritmaları ile sesler arası yakınlık değerlerini hesaplamıştır. Bu çalışması ile akustik olarak benzer ses olay örneklerinin etiketlerinin anlamsal benzerliklerini değerlendirmiş ve sonraki çalışmalarında geliştirme için zemin hazırlamıştır.

Bir başka çalışmasında Mesaros vd. [35], çokluortam veri tabanlarında kullanılabilecek çevresel seslerden oluşan ses olay kayıtları üzerinde indeksleme ve veri geri getirme sisteminin etkili bir şekilde geliştirilebilmesi üzerine çalışmalar yapmıştır. Bu çalışmasında geliştirdiği yeni yaklaşım ile anlamsal ve sesin akustik benzerliklerinin birleşiminin QbE geri getirim sistemi üzerinde geliştirmiştir. Bu geliştirdiği yöntem ile ses bazlı geri getirim yöntemlerinden daha başarılı bir geri getirim performansı elde etmesi ile birlikte; örnek veri üzerinde etiket bazlı geri getirim yöntemlerinden akustik olarak daha yakın bir başarım elde edildiği gözlemlenmiştir. Çalışmaları kapsamında örnekleme ve test aşamalarında ses verileri üzerinde *20 ms* lik pencere boyutu ve %50 atlama oranı parametrelerinin kullanımı ile *MFCC* öznitelik çıkarımları yaparak öznitelik vektörel dizilerini elde

etmiştir. Elde ettiği öznitelik değerlerini GMM algoritması ile etkileşimi sonucu çıkan değerlendirme sonuçlarını önceki çalışmasında olduğu gibi simetrik bir *KL-divergence* algoritması ile sesler arası yakınlık değerlerini sonuçlarını değerlendirmiştir. Böylece çalışma boyunca hedeflediği ses verileri üzerindeki etiketlenmelerdeki çok çeşitlilik ve hatalı-eksik etiketlenme gibi sorunlar üzerinde iyileştirmeler elde etmiştir. Çalışmasının geliştirme aşamasını semantik ve akustik benzerlikler olarak iki ayrı kategoride geliştiren Mesaros, etiketler arası uyarlanan anlamsal benzerlik aşamasında WordNet [36] taksonomisi kullanmıştır. Bu taksonomi sayesinde etiket bazlı eşleşme üzerinde geliştirme sağlanmaya çalışılmıştır. Önerdiği yöntemde semantik kısmın, elde edilen verilerinin içerik bazlı ses arama sistemini ses içeriği numune olarak kullandığı ses verisi ile ilişkili olmayan sistem tarafından çıktı olarak getirilmekte olan seslerin elenmesi için kullanmaktadır. Örnek ses ile ilişkili olmayan ses verilerinin elenmesi ile akustik olarak işitsel benzerliğinin artması performans kazancı elde edilmiştir. Wang vd. [37] ise, insan beynindeki ezberleme süreçlerinden esinlenerek, geleneksel modellere karşı veri geri getirim sistemlerinde daha iyi performans gösterecek ve bu sistemlerdeki gürültü sorununa karşı kuvvetli bir yapı öneren model geliştirmişlerdir. Geliştirdikleri modelde insan hafıza sistemini model alarak; üç aşamalı bir ezberleme süreci tasarlanmıştır. Bu ezberleme süreci; kodlama, ezberleme ve hatırlama olarak belirlenerek öğrenme modeli için geliştirdiği bir derin sinir ağı modeli üzerine inşa edilmeye çalışılmıştır. Sonuç olarak çalışmasında; önerilen evrimsel modelin bu araştırma problemleri üzerinde daha iyi performans gösterdiği gözlemlenmiştir.

3 TEMEL TANIM VE KAVRAMLAR

Çalışmamız işitilebilen sesler içinde bulunan çevresel sesler ve bu seslerin elektromanyetik enerjiye dönüştürülerek elde edilen sayısal verilerinin işlenmesi etrafında şekillenmektedir. Aşağıda araştırmamız boyunca kullandığımız temel tanım, kavram, yöntem ve kullanılan veri kümeleri ile ilgili bilgiler yer almaktadır.

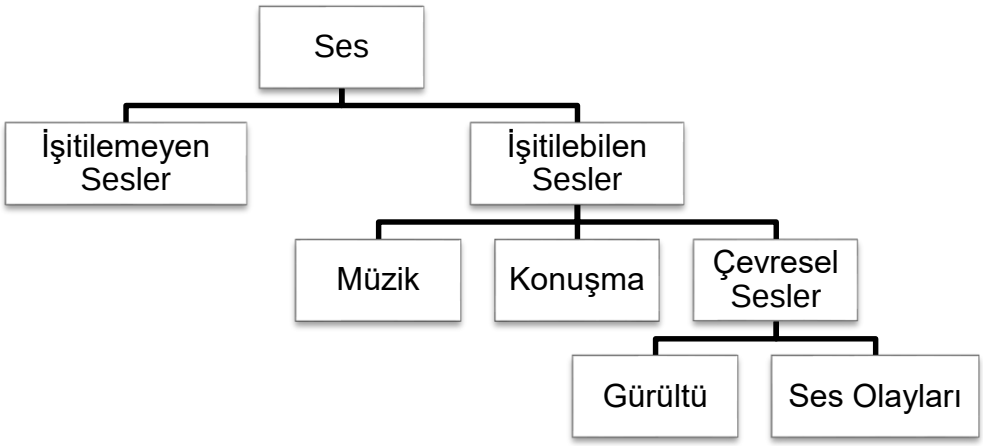
3.1 Ses

Ses, havada veya başka bir ortamda dolaşan ve canlıların duyu organlarına vardığında duyulabilen titreşimlerdir. Ses ortamdaki sıkıştırma dalgaları vasıtasıyla katı, sıvı ve gaz içinde hareket eden boyuna mekanik dalgalardır [38]. Herhangi bir engel ile karşılaşılmadığı sürece, ses dalgaları kaynaktan dışa doğru bütün yönlerde yayılmaktadırlar. 20 *Hz* İle 20000 *Hz* frekans değerleri arası insan kulağı için işitilebilen ses olarak tanımlanmaktadır. Ses sinyalleri, bir sesi iletmek veya saklamak için sesin elektromanyetik enerjiye çevrilmiş bir elektriksel formudur. Doğadaki sesleri bilgisayar ortamlarına aktarılması diğer tüm sinyallerde olduğu gibi örnekleme yapılarak sağlanmaktadır.

3.1.1 Çevresel Sesler, Ses Olayı ve Akustik Sahneler

Günlük hayatta insan kulağı tarafından işitilebilen sesler müzik, konuşma ve diğer çevresel sesler olarak alt kategorilerde incelenebilir. Müzik sesleri; belli notalar ve nakarat bilgilerinden meydana gelerek seslerin melodik olarak kullanılması ile oluşturulan sanatsal seslerdir. Konuşma sesleri ise; insanların duygu ve düşüncelerini sözlü olarak anlatması eylemi sonucu oluşan seslerdir. Belli bir ton ve dil bilgisi bulunmaktadır. Bu kategorilerin dışında kalan ses olayları ise ortam ve zamana göre çeşitlilik gösterebilmektedir. Çevredeki diğer kaynaklardan elde edilen sesler çevresel sesler olarak tanımlanabilir. Bu seslerin bilgisayar sistemlerinde otomatik olarak tanımlanabilmesi, kontrolsüz ortamlarda oluşmaları ve bu ortamların özelliklerini içeren sesler sebebiyle oldukça zorlu bir işlemdir. Bu zorluğa ek olarak akustik ortamlarda oluşan ses olayları, birbiri ile çakışarak ses tanımlama görevini zorlaştırabilmektedir. İnsan kulağı, beynindeki karmaşık işlemleri kullanarak belirli bir akustik ortamda arka planda oluşan işitilebilen ses ve ses olaylarını ayırt etme ve sınıflandırma konusunda büyük bir yeteneği mevcuttur. Ortamda bir konuşma

olması veya müzik sesi olması, arka planda gerçekleşen seslerin insan kulağı tarafından ayırt edilmesinde hiçbir engel teşkil etmemektedir. Bu yeteneğin bilgisayar ortamlarında modellenebilmesi ise araştırmacılar tarafından güncel araştırma konusudur. Bu eşsiz yeteneği modelleyerek otomatik ses tanıma ve geri getirimi sistemleri üzerinde araştırmalarını sürdürmektedirler.



Şekil 3.1. Ses sınıfları.

Çevresel Ses: Belirli bir ses kaynağının ürettiği işitilebilir seslerdir. Çevresel sesler, birçok kaynaktan yapısal olmayan sesleri içermektedir. Bu sayede diğer ses kategorileri; müzik sesi ve konuşma seslerinden farklı olarak ifade edilmektedir. Akan ırmağın su sesi, öten bir kuş sesi, trafik, şehir gürültüsü vb. örnekler verilebilir. Bir konuşma sesinde; ses dosyasının ön planındaki sesler tanımlama için kullanılırken, müzik seslerinde ise kaynak olan enstrümanların ürettiği belli bir ritim, akor ve tını gibi özellikler ile ilgilenilmektedir. Çevresel sesler ise arka plandaki kaynaklardan oluşan seslerle ilgilenilmektedir.

Ses Olayları: Çevresel sesler içinde bulunan ses olayları; bir bölgede gerçekleşen bir olayı tanımlamak için kullanılan bir etikettir. Bu etiket ile arka planda gerçekleşen olayı anlamada ve bu olayı diğer olaylar ile ilişkilendirmede kullanılabilir. Örnek olarak; Araç korna seslerinden trafikte yoğunluk olduğu, kasiyer ve ortamdaki kalabalığın sesleri ile bir alışveriş merkezinde alışveriş yapıldığı gibi bilgiler edinilebilmektedir. Bu gibi birçok farklı ortam karakteristiğinden meydana gelen sesler bir akustik sahne çatısı altında gerçekleşmektedir.

Akustik Sahne: Bir veya birden çok ses kaynağından oluşan ses olaylarının birleşiminden elde edinilen ses klibinin mantıksal bir parçası olarak tanımlanan akustik sahneler birçok farklı ortam karakteristiklerini barındırabilmektedir. Bir akustik sahne içerisinde birden çok ses olayı bulunabilir veya birden çok akustik sahne içerisinde aynı ses olayını barındırabilir [39]. Örnek olarak verdiğimiz araç korna sesleri ve araba motoru sesleri ile bir trafik veya araç yolu ortamı akustik sahne olarak tanımlanabilir.

3.2 Ses Öznitelik Çıkarımı

Ses sinyallerinde eğitim modelleri için sinyalin karakteristiğini yansıtacak şekilde basit ve anlamlı veriler elde edilebilmesi gereklidir. Bu doğrultuda karmaşık sinyallerden öznitelik çıkarımı adımları ile ses sinyalleri üzerinde anlamsal bilgilerin çıkarımı sonucu öznitelik vektörü olarak tanımlanan veri tanımlayıcıları değer kümesi elde edilebilmektedir. Ses özniteliğinin çıkarılmasının amacı, kaynağı tanımlarken sinyalde bulunan akustik özelliklerden ödün vermeden yüklü miktardaki karmaşık çokluortam verisini özetlemektir. Böylece sayısal ortamlarda verimli bir şekilde ses tanımlama işlemleri yapılabilir. Literatürde çok sayıda çalışmada çeşitli öznitelik çıkarma yöntemleri kullanılmış ve öznitelik üzerinde birçok araştırmalar yapılmıştır. Bu tez kapsamında çalışmalarda sesin karakteristiğini elde edebilmek için *MFCC*, logaritmik *mel* enerjileri ve spektrogram öznitelik temsillerinden yararlanılmıştır. Ayrıca birçok farklı parametre seçenekleri ve çeşitli katsayılar ile bu öznitelikler üzerinde deneyler yapılmıştır.

3.2.1 Mel Frekans Kepstrum Katsayıları (MFCC)

Mel Frekans Kepstrum Kaysayıları insan kulağının algılama şeklini model alan araştırmacılar tarafından birçok çalışmada kullanılmış ve başarılı sonuçlar elde edilmiş bir öznitelik çıkarım yöntemidir. Analog ses dalgalarını dijitalleştirerek ses özellik vektörüne dönüştürme işlemidir. *MFCC* katsayıları ses sinyalinin kısa süreli güç spektrumunu temsil eden *MFCC*, kepstrumların önemli noktalarını baz almaktadır [40].

Şekil 3.2'de gösterilen MFCC öznitelik çıkarım adımlarından bahsedilecek olunursa; ilk olarak öznitelik çıkarımına gönderilen ses sinyali ön vurgu adımında yüksek

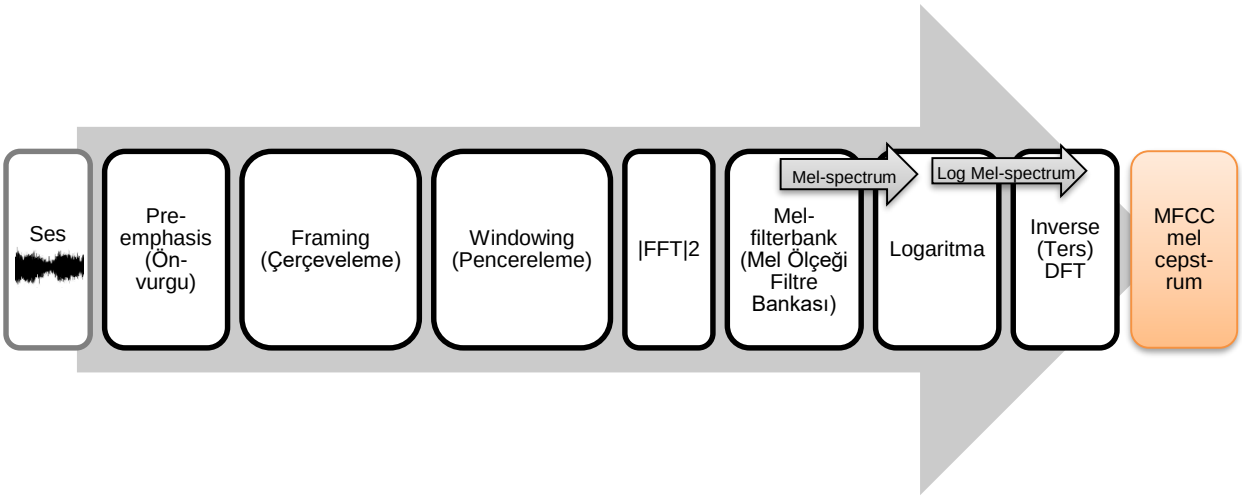
frekans bileşenleri düşük frekans bileşenlerine göre küçük genlik oluşacağından ötürü sinyal filtreleme işlemini yaparak enerji arttırır. Daha sonra çerçeveleme (framing) adımı sırasında ses sinyalleri kararlılık sağlanabilmesi için kısa zaman aralıklarına bölünür. Böylece sürekli olan sinyal bölünerek ayrık bir yapılara dönüştürülerek pencereleme işlemine gönderilir. Burada Fourier analizinde kullanılan özel bir Fourier dönüşümü olan Ayrık Zamanlı Fourier Dönüşümü (DFT) hesaplanması için çerçeveleme işleminde çerçeve başı ve sonundaki süreksizlikleri ortadan kaldırır. Kaiser, dikdörtgen, Barlett ve Hamming gibi birçok pencereleme (windowing) fonksiyonu mevcuttur. Bu çalışmada en yaygın kullanılan Hamming fonksiyonu denklemi şu şekilde tanımlanmaktadır [41].

$$w[n] = \begin{cases} 0{,}54 + 0{,}46\cos(\frac{2\pi n}{N-1}), & |n| \leq \frac{N-1}{2} \\ 0 \quad , & diğer\ yerlerde \end{cases} \qquad (3.1)$$

Pencereleme işlemi sonrasında Hızlı Fourier Dönüşümü (FFT) uygulanır ve sesin dalga boyuna göre dağılımı yani genlik spektrumu elde edilir. Seste oluşan yüksek ve alçak tonlar gösterilir. Bu işlem ile zaman uzayından frekans uzayına geçilmektedir. *FFT* işlemi sonrasında elde edilen sonuca insan kulağını karakterize eden *mel* ölçüsü uygulanır. Bunun sebebi; insan kulağının her sinyale aynı hassasiyet gösterememesidir. Örneğin 1000 *Hz* üstü sinyaller için insan kulağı az hassasiyet göstermektedir. Denklemi aşağıdaki belirtilen *mel* ölçüsü uygulanması sonucunda girdi frekansı *f*, *mel* frekansına çevrilmektedir.

$$Mel\ f = 2595\ \log_{10}(1 + \frac{f}{7000}) \qquad (3.2)$$

Logaritmik enerjilerinin hesaplanması için mel-filterbank çıktısının karesinin logaritması alınmaktadır. Böylece frekans tahminlerini girdideki küçük değişimlere karşı daha az duyarlı hale getirilmektedir. Ayrık Fourier Dönüşümünün tersi (Ters DFT) alınarak frekans uzayından zaman uzayına dönüştürülür ve MFCC katsayıları elde edilir.



Şekil 3.2. MFCC öznitelik çıkarım adımları.

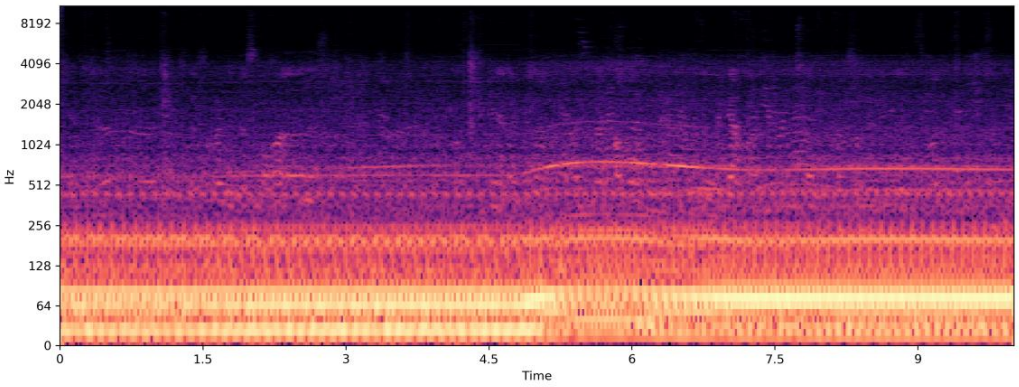
3.2.2 Log-Mel

Logaritmik *mel* ölçeği; birbirinden eş uzaklıktaki dinleyiciler tarafından ses perdelerine karar veren kavramsal ölçektir. Bu ölçek,1 *kHz* altındaki frekanslara doğrusal ölçekli ve 1 *kHz* üstündeki frekanslara ise logaritmik ölçekli olarak tanımlanmıştır. Frekans değerini (*f)*; *Mel* (m) tabanına çevirmek için kullanılan denklem aşağıdaki şekilde tanımlanmaktadır [42]:

$$M = \begin{cases} f, & f < 1000 \\ fc\left(1 + \log_{10}\left(\frac{f}{fc}\right)\right), & f \geq 1000 \end{cases} \quad (3.3)$$

3.2.3 Spektrogram

Spektrogram; ses sinyalinin frekans spektrumunun zamansal değişkenliğinin görsel bir gösterimidir [43][44]. Bir başka deyişle; her zaman diliminde bir ses sinyalinin frekans tayfının hesaplanıp zaman-frekans eksenli bir grafik üzerinde görsel temsilidir. Spektrogram gösterimlerinde dikey eksen frekans değerini gösterirken, yatay eksende zaman bilgisi gösterilmektedir. Sinyal belli parçalara ayırarak her bir parçanın spektrumu hesaplanmak üzere işleme tabi tutulur. Bu farklı spektrumlar daha sonra iki boyutlu bir görüntü oluşturmak için yan yana dikey çizgiler olacak şekilde konularak Şekil 3.3'deki gibi gösterilir. Spektrogramlar sinüslerin birbiri ardına yığılmış bir şekilde göstererek ses parçalarının frekans yapısını oldukça basit bir yapıya indirmektedir.



Şekil 3.3 Frekans (Hz) ve saniye bazınca zaman (time) aralığı gösteren spektrogram.

3.3 Sınıflandırıcılar

Bu bölümde; test aşamalarında kullanılması tercih edilen Çok Katmanlı Algılayıcılar, Yinelemeli Sinir Ağları, Konvolüsyonel Sinir Ağları gibi sınıflandırıcılar anlatılacaktır.

3.3.1 Çok Katmanlı Algılayıcılar (MLP)

Çok katmanlı algılayıcılar, en az üç düğüm katmanından oluşan ileri beslemeli yapay sinir ağlarıdır. Giriş, bir veya birden fazla gizli katman ve son olarak çıkış katmanından oluşmaktadır. Giriş düğümleri dışında, her düğüm doğrusal olmayan bir etkinleştirme işlevi kullanan en az bir nörona sahiptir. Her katmandaki nöronlar; bir önceki ve bir sonraki katmanlarda bulunan nöronlara yönlü olarak bir veya birden fazla bağlantı sağlamaktadır.

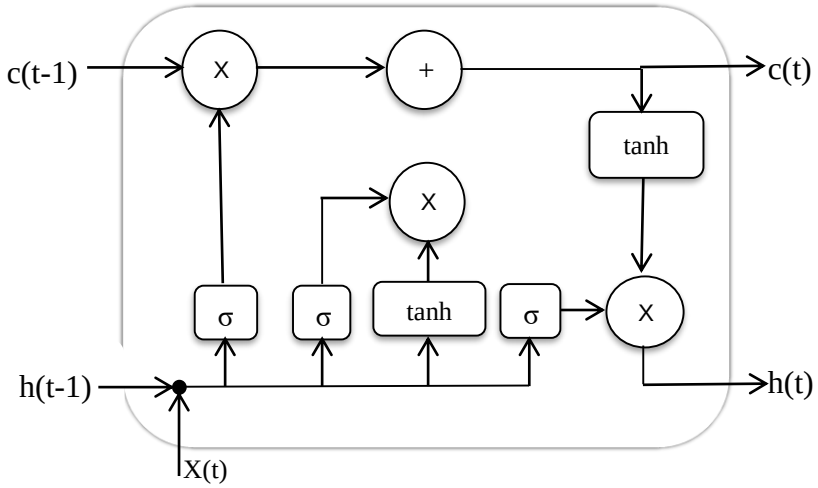
Burada giriş katmanı gelen verileri alarak belirli ağırlık işlemlerine tabi tutup bir sonraki gizli katmana verileri aktarır. Ardından o gizli katman, varsa kendisinden sonraki gizli katmana gelen verileri aktarmaya başlar. Böylece her katman çıkışı bir sonraki katmanın girişi olmaktadır. Ara katman sayısı en az bir olmak üzere probleme göre değişir ve ihtiyaca göre ayarlanır. Ayrıca katmanlardaki nöron sayıları probleme göre belirlenmektedir. Çıkış katmanı ise önceki katmanlardan gelen verileri işleyerek ağın çıkışını belirler. Aktivasyon fonksiyonu; *Sigmoid*, *tang*, *lineer*, *threshold* ve *hard limiter* fonksiyonları gibi popüler fonksiyonlardan biri olabilmektedir. Bu sistemlerde öğrenme metodu genel olarak ileri doğru hesaplama

ve geri doğru hesaplama (back propogation) olmak üzere iki aşamadan oluşur. İlk aşamada girdi verileri ileri doğru gerekli aktivasyon işlemlerini yaparak çıktı katmanına kadar işlenir. İkinci aşamada ise geriye doğru hesaplama yaparak hata ağırlık değerlerine dağıtıp her yenilemede hata payı azaltılması ile güçlenmesi beklenir.

3.3.2. Yinelemeli Sinir Ağları (RNN)

Tez çalışması kapsamında yinelemeli sinir ağları öğrenme modelinden Uzun Kısa Süreli Bellek Yinelemeli Sinir Ağı (LSTM) ve Geçitli Tekrarlayan Üniteli Yinelemeli Sinir Ağı (GRU) yöntemleri kullanılması tercih edilmiştir.

Çalışmada kullanılan RNN algoritmalarından biri olan LSTM yapısı; uzun vadeli bağımlılıkları öğrenebilen bir yinelemeli sinir ağı algoritması olarak tanımlanır. Standart yinelemeli sinir ağı algoritmalarında oluşan uzun vadeli bilgileri hatırlayan ve bu sürede bağımlılık sorununun önüne geçmek için tasarlanmıştır. RNN yapısında temel problemlerden biri zaman içinde geriye doğru olan bağımlılıktan gelmektedir. Eğitim sırasında oluşturulan öğrenme ağı karmaşık hale gelmesi, ağın geriye doğru ağırlık değerlerinde güncelleme yapılması sonucunda sıfır veya sıfıra yakın değerler olmasından dolayı güncelleme olamamakta ve eğitim durabilmektedir. RNN yapısındaki bu geriye doğru bağlılık sorununa, LSTM yapısında bir hafıza hücresi RNN yapısına eşlik etmesi ile çözüm olarak sunulmuştur. Bu hafıza hücresi ile önceki zamandan gelen bilgi alınabilmekte ve bir sonrakine aktarılabilmektedir [45][46].



Şekil 3.4. LSTM birim yapısı.

LSTM ağında bulunan bu birimler uzun veya kısa zaman periyotlarını hatırlar. Bu birimler içinde hatırlatılması için tutulan değerler hiçbir şekilde herhangi bir etkileşime girmez veya değişim yaşayıp kaybolmaz. Şekil 3.4'de LSTM biriminin yapısı örnek olarak gösterilmektedir. Şekilde gösterilen LSTM birim yapısında; girdi olarak *X(t)* o anki mevcut girdi değeri, *h(t-1)* önceki gizli durum ve *c(t-1)* ise önceki hafıza durumu değerlerini almaktadır. Çıktı olarak; *h(t)* mevcut gizli durum ve *c(t)* mevcut hafıza durumu üretmektedir.

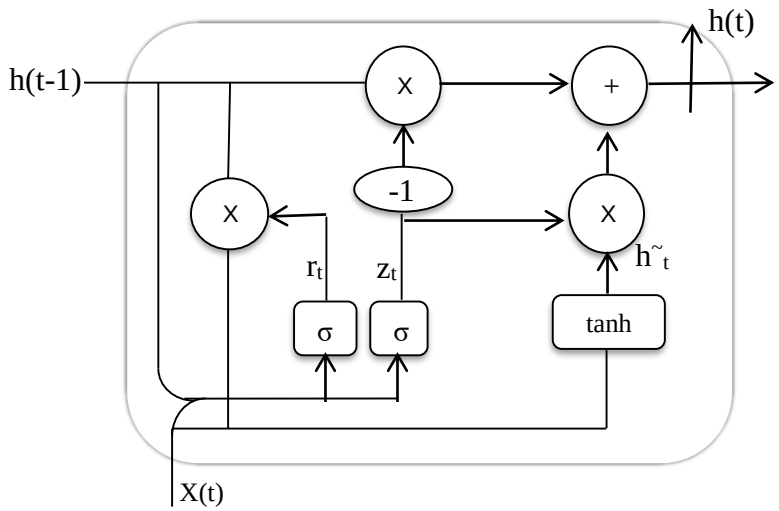
Bir başka RNN modellerinden biri olan GRU, Cho vd. [45] tarafından 2014 yılında standart bir yinelemeli sinir ağı modelinde oluşan; kaybolan gradyan problemini çözmeyi amaçlamak üzere geliştirilmiştir. LSTM birimlerine benzer bir şekilde tasarlandığı ve yaklaşık aynı performansta çalıştığı için bu yapının bir varyasyonu olarak düşünülmektedir. LSTM biriminden farklı olarak modelde bir sonraki zaman adımlarına veri aktaracak güncelleme kapısı (update gate) ve modelden geçmiş verilerin ne kadarını geçip ne kadarının unutulacağına karar veren sıfırlama kapısı (reset gate) bulunmaktadır. Bu kapılar ile veriler depolanabilir ve filtrelenebilir. Çıktı olarak hangi verilerin aktarılması gerektiğine karar veren güncelleme (*u*) ve sıfırlama (*r*) vektörlerinin denklemi aşağıdaki gibidir [47]:

$$z_t = \sigma(W^{(z)}x_t + U^{(z)}h_{t-1}) \tag{3.4}$$

$$r_t = \sigma(W^{(r)}x_t + U^{(r)}h_{t-1}) \tag{3.5}$$

$$h_t^{\sim} = tanh(W^{(h)}x_t + U^{(z)}h_{t-1} \odot r_t) \tag{3.6}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot h_t^{\sim} \tag{3.7}$$



Şekil 3.5. GRU birim yapısı.

Şekil 3.5'de gösterilen GRU birim yapısından bahsedilecek olunursa; girdi olarak *X(t)* o anki mevcut girdi değeri ve *h(t-1)* önceki gizli durum değerlerini almaktadır. GRU birimi çıktı olarak; *h(t)* mevcut gizli durumu ve çıktıyı üretmektedir. Şekilde güncelleme kapısı *r* ve sıfırlama kapısı ise *z* olarak belirtilmiştir.

3.3.3 Evrişimsel Sinir Ağları (CNN)

Makine öğrenme tekniklerinden ileri yönlü bir sinir ağı olan evrişimsel sinir ağları bir MLP türüdür. İlk olarak 1980'lerin başında LeCun vd. [48] tarafından LeNet-5 mimarisi ile tanıtıldı. Bu mimari bir giriş katmanı, birkaç konvolüsyonel katmanlardan, havuzlama (pooling) ve çıktı katmanlarından oluşmaktadır. Konvolüsyonel katmanlar seviyesine göre kendisinden önceki katmanlardan aldığı girdiler üzerinde işlemler yaparak özelliklerin çıkarılmasında rol oynamaktadırlar. İlk konvolüsyonel katmanı ile en düşük özellik temin edilir iken, konvolüsyonel katmanları eklendikçe, yani seviyesi yükseldikçe, daha yüksek seviyeli özellik çıkarımlarını sağlamaktadır. Havuzlama katmanı ise; konvolüsyonel katmanlar ile üretilen çıktıların sadeleşmesi için kullanılmaktadır. Mimarinin çıktı katmanı; Tam Bağlı katmanlara veya Sigmoid katmanı gibi sınıflandırıcı katmanlarına bağlanabilmektedir. Girdi katmanında veriler; görüntü, ses, video gibi çokluortam verileri olabileceği için son yıllarda birçok sinyal işleme alanında çalışan araştırmacılar tarafından performanslı kullanımı olduğundan dolayı tercih edilmektedir [49][50]. CNN mimarilerinin eğitim süreci boyunca öğrenme ağırlıklarını

hata payları ile güncellenmesini ayarlayan bir geri yayılım algoritması kullanılarak hata payını azaltacak bir strateji uygulanmaktadır.

3.4. SPP (Spatial Pyramid Pooling)

Bilgisayarlı görü ve akıllı sitemlerinde Uzamsal Piramit Veri Birleştirme (SPP); resim tanıma, sınıflandırma ve tespit problemleri için geliştirilmiş bir alternatif veri birleştirme yöntemidir. Karşılaşılan birçok probleme karşı başarılı sonuçlar elde edilmiştir [51]. SPP yapısı, sözcük çantası (BoW – Bag of Works) modelinin bir uzantısı olarak geliştirilmiştir. SPP yapısı CNN yapısında son prevalanstan önceki katmana yerleştirilecek şekilde; girdi boyutundan bağımsız olarak sabit uzunluklu çıkışlara sahip olma ve sıralı pencere havuzlaması yerine çok seviyeli havuz yapısını kullanmak üzere veri birleştirme yöntemlerinde alternatif bir yöntem olması için geliştirilmiştir. Son yıllarda, bilgisayarlı görü problemlerinde, CNN ile geliştirilen öğrenme modelleri ses olay ve akustik sahne sınıflandırma problemlerinde üstün sonuçlar ortaya koymuştur [49][50]. Bilgimiz doğrultusunda, SPP yönteminin CNN algoritmaları ile ilk kullanımı He vd. TPAMI yayınında görülmektedir [52].

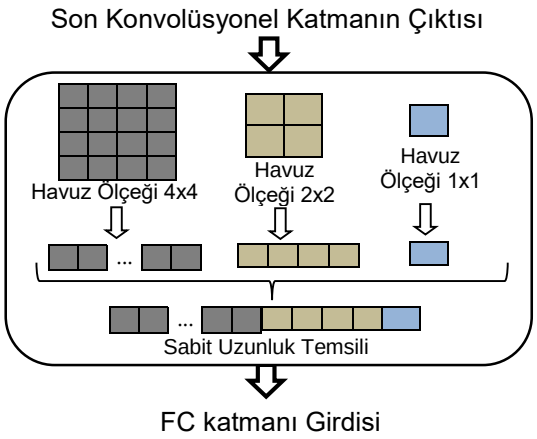
Bir havuzlama stratejisi olan SPP; görüntüyü daha alt ham bölgelere böler ve bu alt bölgeler içinde bölgelerin yerel özellikleri üzerinde bir veri birleştirme (havuzlama) işlemini yapmaktadır. SPP yöntemi ile iki boyutlu gelişigüzel boyutlanmış görsel girdiler üzerinde boyutuna ve ölçeğine duyarlı olacak şekilde işleme tabi tutarak sabit bir boy gösterimi haline getirme işleminde oldukça başarılıdır. *Softmax* sınıflandırması ve tam-bağlı yoğun katmanlara nazaran, CNN katmanları bu gelişigüzel boyutlu girdileri kabul ederek çalışabilmektedirler. Fakat sabit bir boyut çıktı verdikleri sırada gelişigüzel girdiler üzerinde boyut ve ölçek bozukluğuna yol açılabilmektedir. CNN'leri kullanırken genel uygulama, bu gerekliliği sağlamak için girdilerin çarpıtılması veya kırpılmasına ihtiyaç duyacaktır. Bu işlemler sonucunda kırpılmış veya çarpık bölgeler istenen nesneyi veya olayı içermeyebilir. Aynı zamanda bazı geometrik bozulmalar içerebilir. Bu sebeplerden ötürü gelişigüzel boyuta sahip görsel girdiler üzerinde SPP yöntemi, bu tür problemlere karşı güçlü bir yöntem olarak geliştirilmiştir.

Herhangi bir boyutta veya ölçekte girdileri kabul eden SPP havuzlama yöntemi; CNN çıktı olarak veren *feature maps* matrisini bölümler ve çok seviyeli uzamsal

kutuları kullanarak yerel özellikleri havuzlar. Bunun sonucunda ise sabit uzunluklu çıktılar üretmektedir. Herhangi bir *(m x m)* boyutunda çıktı matrisi ve her piramit seviyelerinde *(n x n)* seviyeli kutulara sahip *l* seviyeli piramitten; her *l* çıktıları tamamen bağlı yoğun katmanın girişini oluşturmak için bir araya getirilecektir (Şekil 3.6). Filtrelerin boyutu *(size)* ve sırası *(stride)* sırasıyla şu şekilde hesaplanmaktadır [53]:

$$size = \lceil m/n \rceil \qquad (3.8)$$

$$stride = \lfloor m/n \rfloor \qquad (3.9)$$



Şekil 3.6. SPP Katmanı.

Örnek olarak Şekil 3.6'dan bahsedilecek olunursa; 3 seviyeli bir piramit havuzu (örneğin; 1x1, 2x2, 3x3) kullanıldığı zaman verilen herhangi bir giriş için tam bağlı yoğun katmanı veya *softmax* sınıflandırma katmanı için 14 birim uzunluklu sabit uzunlukta bir girdi boyutu (1x1 + 2x2 + 3x3 havuzlamalar ile) elde ederiz. CNN mimarisinde SPP yönteminin bir diğer önemli özelliği de çoklu seviyeli havuzlamanın nesne deformasyonlarına karşı sağlamlığı olarak bildirilmektedir [51].

Birçok video ve resim gibi görsel sınıflandırma problemlerinde kullanılması tercih edilen SPP yöntemi daha önce hiçbir ses olay sınıflandırma ve akustik sahne sınıflandırma probleminde kullanılmaması motivasyonu ile bu çalışmamızda uygulanmıştır.

4 SES OLAY VE AKUSTİK SAHNE TANIMA

Çevresel ses kayıtları içerisinde bulunan ses olayları, çevremizdeki nesnelerin titreşimleri sonucunda oluşan müzik, konuşma ve gürültü dışında kalan sesler olarak tanımlanmaktadır. Ses (akustik) olayları, bulunulan çevreye dair önemli bilgiler içerebilir. Bu sesler birbirinden bağımsız ya da zamanda iç içe geçmiş biçimde bulunabilir. İnsan kulağı; araç sesi, parkta oynayan çocukların sesleri, çay bardağı kaşığının çıkardığı sesler, adım atma sürecinde oluşan ses ve bunun gibi ses olaylarını tanımada oldukça yeteneklidir. Bu eşsiz yeteneği bilişim alanında makine öğrenme yöntemleriyle aktarılması, otomatik sistemler tarafından ses olaylarından anlamlı bilgi çıkarabilme becerisinin kazanılması ve bu işlemler sırasında sistemin performansının akademik alanda güncel araştırma konuları olarak sayılmaktadır. İçerik-tabanlı çokluortam geri getirimi [54], mobil cihaz uygulamaları [55], robot ve akıllı şehirler gibi uygulamalarda ele aldığımız araştırma konusu önemli bir potansiyele sahiptir. Bu nedenle, çevresel seslerin otomatik yöntemlerle önceden belirlenen ses olayı sınıflarına atanması (etiketlenmesi) önemlidir. Diğer yandan, birden fazla sesin aynı anda oluştuğu ve/veya çevresel nedenlerden dolayı çarpıtılmış ses olaylarının otomatik yöntemlerle tanımlanması zorlu bir problemdir. Bu problemin çözümü için son yıllarda çeşitli sinir ağı tabanlı öğrenme modelleri ile çeşitli öznitelik çıkarım yöntemleri geliştirilmiş ve kullanılmıştır [14][22][56]. Ayrıca bu çalışmalar sonucu öğrenim modellerinin başarımı önemli ölçüde arttırıldığı görülmüştür.

Çalışmamızın bu bölümünde söz konusu problem kapsamında otomatik sınıflandırma eğitim modelleri geliştirilmesi ve performanslarının analizi üzerinde durulacaktır. Sınıflandırma eğitimi için MLP, LSTM, GRU, CNN gibi önemli derin sinir ağı algoritmaları kullanılmıştır. Ayrıca çoğunlukla imgesel sınıflandırma problemlerinde kullanılması için geliştirilmiş SPP yöntemini ilk defa çevresel sesler kategorisinde ses olaylarında akustik sahne sınıflandırma problemi üzerindeki kullanarak performans izlenimi yapılmıştır. Tüm ses sinyali sınıflandırma problemlerinde sıkça kullanılan *MFCC*, *log mel* enerjileri ve zamanla değişen ses sinyalinin frekans ve genlik bilgisini temsili olan spektrogramlar eğitim modellerinde girdi olarak kullanılması tercih edilmiştir. Görsel gösterim olan spektrogram ile SPP havuzlama yöntemi kullanımını inceleyerek imgesel sınıflandırma problemlerindeki

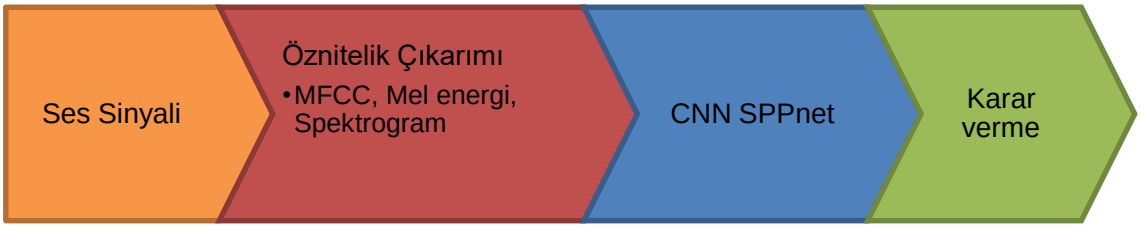
başarımı çevresel ses kayıtları içeren sinyal problemlerinde de görülebilmesi üzerine çeşitli araştırmalar ve deneyler yapılmıştır.

Tez kapsamında yapılan çalışmaların ilkinden bahsedilecek olunursa; ses olay tanıma problemi için ilk olarak, popüler çok katmanlı sinir ağları mimarisi olan MLP kullanımı ile performans gözlemlenmesi yapılmıştır. Ayrıca ses olaylarının zamansal karakteristiklerinin farklı olması ve ses olaylarının ardışık özelliklerini modelleyebilmek amacıyla ikinci mimari olarak uzun kısa süreli bellek hafıza olan LSTM sınıflandırma modelleri tasarlanarak başarımı incelenmiştir. Yinelemeli sinir ağlarında temel problemlerden birinin zaman içinde geriye doğru olan bağımlılıktan olduğu belirtilmektedir. Eğitim sırasında oluşturulan öğrenme ağının karmaşık hale gelmesi, ağın geriye doğru ağırlık değerlerinde güncelleme yapılması sonucunda sıfır veya sıfıra yakın değerler olmasından dolayı güncelleme olamamakta ve eğitim durabilmektedir. RNN mimari yapısındaki bu geriye doğru bağlılık, LSTM yapısı sayesinde bir hafıza hücresi RNN yapısına eşlik etmesi ile bu temel problemlerin giderilmesi ve bunun ses olay tanıma problemimiz karşısında performans kazanımının gözlemlemesi için mimarimize eklenmiştir.

Sistemin incelenmesi ve öğrenme performansının geliştirilmesi için öğrenme adımları Detection and Classification of Acoustic Scenes and Events (DCASE) topluluğu tarafından sağlanan TUT Sound Event 2017 veriseti üzerinde çalışılmıştır [57]. Burada LSTM algoritması ile ses olay kayıtları içeren ses sinyalleri üzerinde tüm zamansal verileri işleyebilen bir sistem geliştirerek performansı izlenilmiştir. Ayrıca ses olayı tanıma performansını artırmak amacıyla, öznitelik çıkarım süresi parametresinin, model aktivasyon işlevinin ve veri birleştirme katmanlarının başarıma olan etkileri incelenmiştir. Öznitelik çıkarım süresi boyunca MFCC ve mel ölçeği öznitelik çıkarım yöntemleri kullanılarak üzerinde yapılan parametre değişiklikleri ile sonuçlar araştırılmıştır.

Çalışmamız kapsamında ikinci çalışma olarak çevresel sesler kategorisi içinde bulunan akustik sahne sınıflandırma problemi üzerinde araştırmalar yapılmıştır. Bu çalışma kapsamında IEEE tarafından düzenlenen ICSC konferansına tarafımızca hazırlanan bildirinin [67] yayını yayınlanmıştır. Bu çalışmada bilgisayar görü literatüründe geliştirilmiş ve ölçek bağımsız görsel öğeler üzerinde herhangi bir kırpma/boyutlama işlemleri yapılmadan obje deformasyonunu engelleyerek, veri

boyutunu azaltabilen SPP algoritması mimaride kullanılmıştır. SPP algoritması CNN mimarisinin kullanımı ile görüntü tanıma/sınıflandırma ve tespit problemlerine başarıyla uygulandığı görülmektedir [51][52]. CNN algoritmasının ise bilgisayar görü problemlerinde olduğu gibi son yıllarda sinyal işleme problemlerinde ses olay ve sahne tanıma görevlerinde üstün sonuçlar ortaya koymuştur [24][43][49][50][58]. Bildiğimiz kadarıyla, SPP havuzlama yönteminin CNN mimarilerinde ilk kullanımı He vd. TPAMI yayınında [51] görülmektedir. Araştırmalarımıza göre daha önce Akustik Sahne Sınıflandırması (ASC) problemi için düşünülmediği tespit edilmiştir. Bu motivasyon ile bir veri birleştirme yöntemi olan SPP mimari yapısının bu problem için geliştirilecek eğitim mimarisinde konvolüsyonel katmanlarının son katmanındaki veri birleştirme yöntemi yerine kullanılarak sınıflandırma başarım performansı artışı gerçekleşip gerçekleşmeyeceği gözlemlenmesi hedeflenmiştir. Son konvolüsyonel katmandaki maksimum veri birleştirme yöntemi yerine kullanılan SPP, ayrıca ardında hiçbir birleştirme katmanı kullanılmadan yoğun katman (Fully Connected - FC) bağlantısı sağlanarak sınıflandırma çıktı katmanına bağlantısı sağlanmıştır (Şekil 4.1).



Şekil 4.1. ASC için geliştirilen CNN-SPP mimari eğitim yolu.

Bu çalışma kapsamında öznitelik çıkarım yöntemlerinden *MFCC* , *mel* enerjileri kullanılmıştır. Bu öznitelik çıkarımlarına ek olarak SPP yönteminin bilgisayar görü sistemlerindeki başarısından ötürü genlik bilgisinin görsel temsili olan spektrogram kullanılarak geliştirdiğimiz eğitim modeli üzerinde çalışılması ve performans başarımının gözlemlenmesi düşünülmüştür. Ayrıca öznitelik çıkarım yöntemlerinde pencere ve atlama zaman parametreleri üzerinde değişiklikler yapılarak analiz sırasında incelenmiştir. Eğitim modelimizin aktivasyon parametresinde ise Leaky ReLU [59][60][61] adapte edilmiştir.

Tez kapsamında bir diğer sınıflandırma algoritmasının geliştirilmesi için ASC probleminde kontrolsüz ortam özellikleri ve örtüşen çevresel ses sinyallerinin zamansal bilgilerin çıkarımı üzerinde çalışılmıştır. Bu problem kapsamında sinyallerde bulunan zamansal içeriklerden çeşitli bilgiler elde edilmesi için RNN algoritmalarından LSTM ve GRU mimarilerinin kullanılması tasarlanmış ve performans karşılaştırmaları yapılmıştır. Çevresel sesler içindeki akustik sahnelerin bilgisayar ortamlarında otomatik sınıflandırılması performansı izlenilmiştir. Sonuç olarak konvolüsyonel katmanlar ile bütünleşebilen AlexNetish ve VGGish kaynaklı CNN [28] yöntemleri üzerinde LSTM ve GRU mimarileri ile ayrı ayrı denemeler yapılmış, GRU mimarisi ile performans kazanımı gözlemlenmiştir.

4.1 Deneysel Çalışmalar

Çalışma boyunca araştırmalar için açık kaynak kodlu yazılım olan Python programlama dili ve bu programlama diline bağlı kütüphanelerin kullanımı tercih edilmiştir. Bu kütüphanelerden bazıları; derin öğrenme algoritmaları ve öznitelik çıkarımı için kullanılan Tensorflow [62], Keras [63] ve Theano [64] kütüphaneleridir. DCASE topluluğu tarafından ses olay tanıma ve akustik sahne sınıflandırma üzerinde çalışan araştırmacılar için geliştirilen; çeşitli öznitelik çıkarma ve derin öğrenme algoritmaları fonksiyonları bulunduran dcase_util ve hesaplama aracı olan *sed_eval* kütüphaneleri bu tez çalışmasında kullanılmıştır. Bu araçlar sayesinde çeşitli çevresel ses içeren veri kümeleri üzerinde çalışma imkânı ve çeşitli algoritmaların kolaylıkla kullanılması sağlanmıştır. Ayrıca, derin öğrenme sırasında önerdiğimiz alternatif havuzlama modeli olan ve He v.d.[52], tarafından geliştirilen Spatial Pyramid Pooling (SPP) algoritması için SPP kütüphanesi [65] kullanılmıştır.

Bu araştırmalarda yer alan tüm veya kısmi nümerik hesaplamalar TÜBİTAK ULAKBİM, Yüksek Başarım ve Grid Hesaplama Merkezi'nde (TRUBA kaynaklarında) gerçekleştirilmiştir [66].

4.1.1 Kullanılan Veri Kümeleri

İlk çalışma için; deneylerde kullanılmak üzere, ses olayı tanıma konusunda yaygın olarak kullanılan DCASE performans veri kümesi [10] seçilmiştir. Bu veri kümesi toplam 70 dakika uzunluğunda çevresel ses senaryosu için sokak ortamında kayıt

altına alınmış ve etiketlenme işlemleri altı temel sınıf (araba, çocuklar, büyük araç, insan konuşmaları, insan yürümesi, fren sesi) kapsamında tasarlanmış ve geliştirilmiştir. 24 adet 3-5 dakika uzunluklara sahip bu ses verileri *Soundman OKM II Classic/studio A3 elektret* kulak içi mikrofon ve *Roland Edirol R-09* ses kayıt aracı ile kayıt altına alınmıştır. Veri kümesi eğitim, doğrulama ve test olmak üzere üç parçada sunulmaktadır. Test sonuçları, benzer çalışmalarla kıyaslanabilmesi için dört katlı çapraz doğrulama yöntemi ile elde edilmiştir.

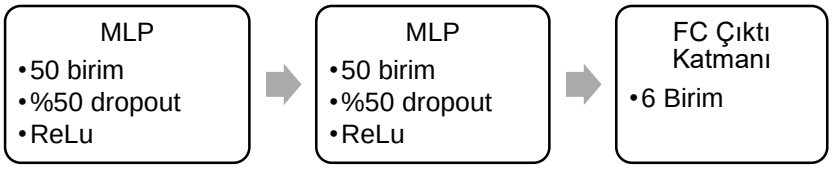
Diğer çalışmalarda DCASE tarafından akustik sahnelerin sınıflandırma problemleri için oluşturulan 2018 veriseti [11] kullanılmıştır. Burada ses kayıtları *Soundman OKM II Klassik/studio A3, electret* kulak içi mikrofonu ve *Zoom F8 audio recorder* ile 48 kHz örnekleme oranı ve 24 bit çözünürlük ile kayıt işlemi yapılmıştır. Bu veri kümesi için birçok farklı lokasyonda 5-6 dakika ses kayıtları ile toplam 28 saatlik kayıt toplanmıştır. Veri kümesinin geliştirme aşaması için hazırlanan *development dataset* kısmı kullanılmıştır. Bu kısım *airport, Indoor shopping mall, metro station, pedestrian street, public square, traffic, travelling by a tram, travelling by a bus, travelling by an underground metro* ve *park* olmak üzere 10 adet sınıf bulundurmaktadır. Veri setinin %70'inin eğitim için ayrılması ve geri kalan %30'unun test aşaması için kullanılması uygun görüldü.

4.1.2. MLP ve LSTM Mimarisi ile Ses Olay Tanıma Problemi

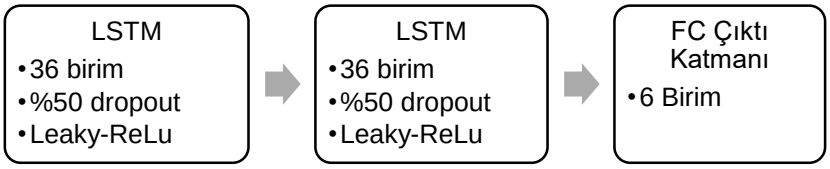
Tez kapsamında yapılan çalışmalardan ilk olarak; MLP ve LSTM sınıflandırma modelleri üzerinde performans incelemesi yapılmıştır. Eğitim modelleri için MFCC ve mel enerji öznitelikleri kullanılmıştır. Öznitelik çıkarım aşamasında farklı analiz pencere sürelerinin başarıma etkisini incelemek için 0,04 pencere boyutu ile 0,02 atlama boyutu değerlerinin yanı sıra 0,02 pencere boyutu ve 0,01 atlama boyutu parametreleri girilerek farklı öznitelikler elde edilmiştir. İki katmanlı 50 birimli MLP algoritması ve FC katmanı mimarisi [57] ile iki katmanlı 36 birimli LSTM mimarisi üzerine eğitim için elde edilen öznitelikler gönderilip sonuçlar karşılaştırılarak irdelenmiştir (Çizelge 4.1). Burada 6 birimli sınıflandırma için yoğun katman ve %20 *dropout* parametresi kullanılmıştır. Geliştirilen modellerin tasarımı Şekil 4.2 ve Şekil 4.3'de gösterilmektedir.

Çizelge 4.1. Tasarlanan sinir ağı mimarileri.

Ara not türü	**Mimari Özellikleri**
MLP	2 x (50 birim) MLP %20 *dropout* ve ReLU aktivasyon 6 birimli yoğun katman çıktı
LSTM + LreLU	2 x (36 birim) LSTM %50 *dropout* ve Leaky ReLU aktivasyon 6 birimli yoğun katman çıktı



Şekil 4.2. MLP + Yoğun Katman modeli.



Şekil 4.3. Önerilen LSTM + Yoğun Katman modeli.

Değerlendirme kapsamında literatür çalışmaları ile karşılaştırılabilir olması için ölçüt olarak F1 değerlendirme ve hata oranı değerleri kullanılmıştır. Bu ölçümler için DCASE tarafından sağlanan TUT Sound Events 2017 sistem çıktısı çok sesli ses olay sezimi ölçütleri referans alınmıştır [67]. Parça bazlı F değerlendirme (F1) ve hata oranı (ER) denklemleri aşağıdaki gibi hesaplanır:

$$F1 = \frac{2.\sum_{k=1}^{K} TP(k)}{2.\sum_{k=1}^{K} TP(k) + \sum_{k=1}^{K} FP(k) + \sum_{k=1}^{K} FN(k)} \qquad (4.1)$$

$$ER = \frac{\sum_{k=1}^{K} S(k) + \sum_{k=1}^{K} D(k) + \sum_{k=1}^{K} I(k)}{\sum_{k=1}^{K} N(k)} \qquad (4.2)$$

Parça bazlı *(k)* değerlendirmede *TP(k)* tahmin ve gerçekte "doğru" olarak etiketlenen ses olaylarını temsil etmektedir. *FP(k)* ise gerçekte "yanlış" ama tahmin olarak "doğru" etiketlenmiş ses olaylarını temsil etmektedir. *FN(k)* değeri ise ses

olaylarının gerçek ortamda hatalı ve tahmin olarak “doğru” etiketlenmiş olayları temsil etmektedir. Hata oranında ise *N(k)* değeri toplam *k* parça başı ses olaylarını, *I(k)* değeri değişimlerden sonra kalan yanlış pozitifleri, *D(k)* değişimlerden sonra kalan yanlış negatifleri, *S(k)* değeri ise sistemin doğru olarak tespit edemediği yanlış negatif ve yanlış pozitifleri temsil etmektedir.

Önerilen sinir ağı modeline *MFCC* ve *Mel* enerji öznitelikleri çıkarımı kullanılmıştır. *MFCC* ve *Mel* öznitelik çıkarımı için ayrı ayrı için pencere ve atlama uzunluğu parametrelerinde değişiklik yapılmıştır. Pencere süresi için 0,04 *ms* (MFCC_A ve Mel_A) ve 0,02 *ms* (MFCC_B ve Mel_B) değerleri kullanılmış, kaydırma boyutu için %50 atlama oranı kullanılması tercih edilmiştir. Bu seçim bellek ve hesaplama karmaşıklığını artırmakla birlikte, başarım ve hata oranına etkilerinin incelenmesi amaçlanmıştır. Deney sonucunda, MLP algoritması testlerinde birbirine yakın sonuçlar elde edilmiştir. LSTM algoritması kullanılarak yapılan testlerde öznitelik çıkarımı değişimi öncesi yapılan testlerde hata oranı MEL_MLP_A modeline göre belirli derecede arttığı gözlemlenmiştir. Bununla birlikte, çıkarılan öznitelik sayısının artırılması, F1 değerinde de küçük de olsa bir artış sağlamaktadır.

Çizelge 4.2. Analiz pencere sürelerinin başarıma etkisi.

A: PENCERE SÜRESİ: 0,04 ms, KAYDIRMA BOYUTU: 0,02 ms.

B: PENCERE SÜRESİ: 0,02 ms, KAYDIRMA BOYUTU: 0,01 ms.

Model İsmi	Model Özellikleri	F1	ER
MEL_MLP_A	40 Mel-bandı enerji –> MLP	**55,87**	**0,69**
MFCC_MLP_A	60 MFCC –> MLP	55,06	0,70
MEL_MLP_B	40 Mel-bandı enerji –> MLP	54,49	0,72
MFCC_MLP_B	60 MFCC –> MLP	**55,60**	0,71
MEL_LSTM_A	40 Mel-bandı enerji –> LSTM	53,42	0,77
MFCC_LSTM_A	60 MFCC –> LSTM	53,87	0,80
MEL_LSTM_B	40 Mel-bandı enerji –> LSTM	53,51	0,77
MFCC_LSTM_B	60 MFCC –> LSTM	53,55	0,80

Daha fazla öznitelik çıkarımı sağlayan küçük analiz çerçeve boyutları ile öznitelik çıkarımı yapıldığı bu çalışmada LSTM ve MLP mimari ile eğitilmesi sonucu başarım

değerleri incelenmiştir. MEL_MLP_A ve MFCC_MLP_B modellerimizin F1 başarım değerlerinin yakın çıktığı görülmektedir. Pencere analiz boyutlarındaki değişimin incelendiği bu sonuçlar DCASE TUT 2017 veri kümesi üzerindeki literatür çalışmaları ile karşılaştırılmıştır. Deneyler sonucunda küçük analiz çerçeve boyutları kullanımı sonrası F1 ve ER oranlarında kayda değer bir farklılık oluşturmadığı görülmüştür. Bu nedenle, bellek maliyeti açısından standart çıkarım parametreleri kullanılarak deneylere devam edilmesi tercih edilmiştir. Sinir ağı aktivasyon işlevleri açısından, Leaky ReLU işlevi LSTM+MFCC özniteliğinde F1 oranını artırmaktadır.

4.1.3 CNN+SPP Mimarisi ile Akustik Sahne Sınıflandırma Problemi

Tez kapsamında bir başka deney çalışmasında; makine öğrenme algoritmaları ile otomatik akustik sahne sınıflandırılması problemi üzerinde araştırma yapılmıştır. Bu problem kapsamında CNN mimarisi ile SPP havuzlama katmanının kullanımının performans analizi yapılmıştır. Çalışmada gelişigüzel en-boy oranlı boyutlara sahip verileri girdi olarak kullanarak tanımlama oranının düşmesinin önlenmesi amacı ile belirsiz gelişigüzel en-boy oranlı boyutlarda imge verileri üzerinde başarılı çalışan SPP yönteminin kullanımı amaçlanmıştır. Nesne deformasyonuna karşı dayanıklı olması nedeniyle bu yöntemin irdelenmesi tercih edilmiştir. Yayınladığımız Basbug vd. [68] çalışmasında, bilindiği kadarıyla SPP algoritması ilk defa akustik sahne sınıflandırma problemlerinde uygulanmıştır.

Bu çalışma kapsamında her biri 7x7 çekirdek boyutu olacak şekilde 3 adet konvolüsyonel katman yapısı kullanılmıştır. Bu katmanlarda 32, 64, 128 adet süzgeç yapısı ile 0,3 bırakma oranı parametreleri mimariye entegre edilmiştir. Son konvolüsyonel katman hariç diğer konvolüsyonel katmanlarda 2x2 Maksimum havuzlama (MP) katmanına yer verilmiştir (Çizelge 4.3). Bunun yanında aktivasyon katmanında bazı gradyan değerlerinin kırılgan olabileceği ve ayrıca eğitim sırasında ReLU aktivasyon yapısında kaybolacağı için Leaky ReLU uygulanması tercih edilmiştir. Eğitim sırasında ağırlıkların güncellenmesi sırasında ölü nöronlar oluşması, Leaky ReLU aktivasyon yapısıyla önleyecektir. Bu yapıyla birimler aktif olmadığı anda sıfıra yakın küçük pozitif değerli gradyanlar üzerinde işlemler yapılabilmesini sağlanmaktadır. Son konvolüsyonel katmanında kullanılan piramit havuzlama için 3 katmanlı SPP (1x1, 2x2, 4x4) havuzlama yöntemi kullanılmıştır. Çıktı katmanında ise 0,3 bırakma oranlı 100 birimli yoğun katman ve 10 katmanlı

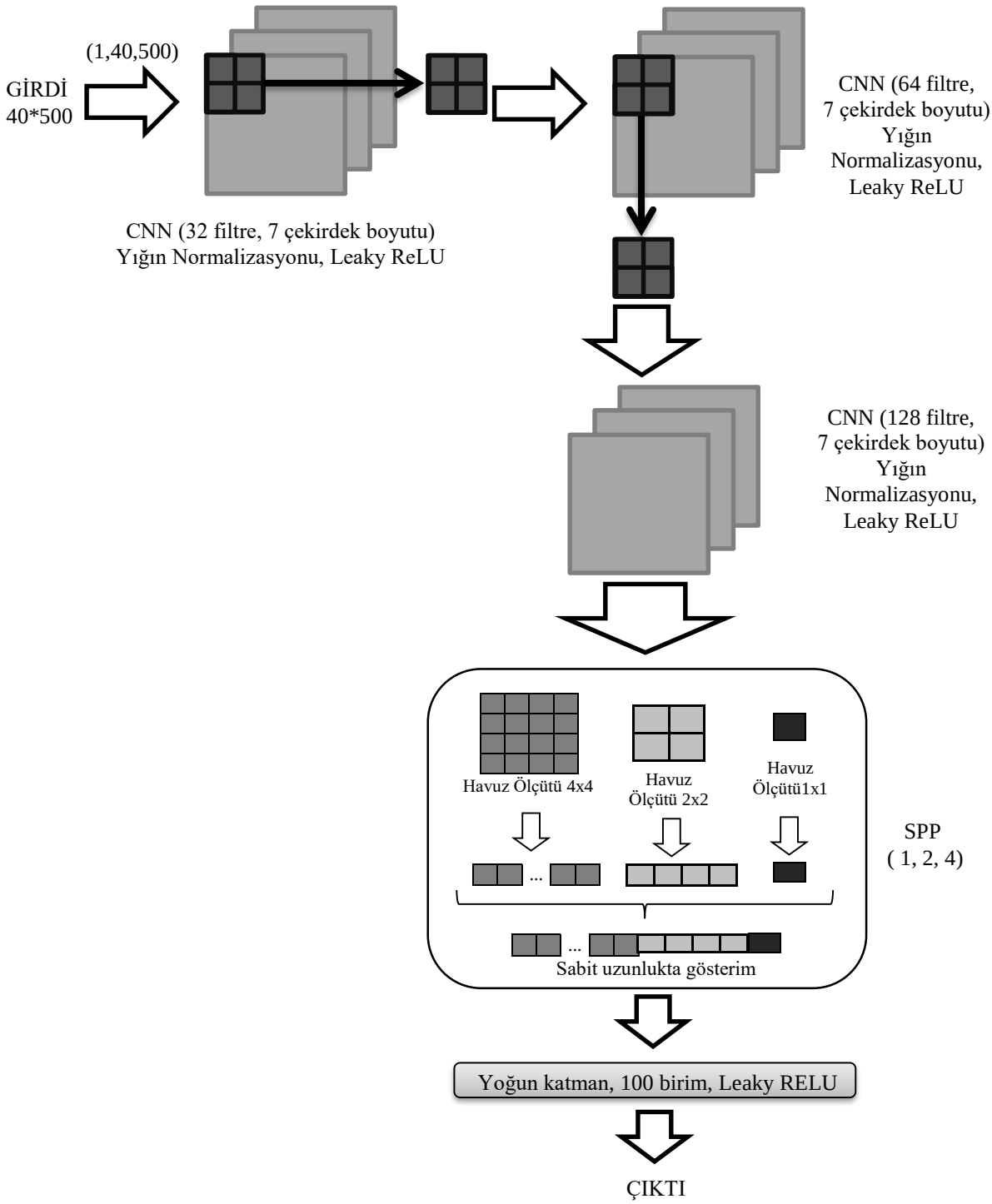
SoftMax sınıflandırma çıktı katmanı kullanılması tasarlanmıştır (Şekil 4.4). Şekil 4.4'te geliştirilen mimarinin aşamaları gösterilmiştir.

Çizelge 4.3. Geliştirilen CNN-SPP mimarisi.

Girdi katmanı
Conv2d(32 süzgeç) – Leaky ReLU + Yığın Normalizasyonu 2x2 MP + Dropout(0,3)
Conv2d(64 süzgeç) – Leaky ReLU + Yığın Normalizasyonu 2x2 MP + Dropout(0,3)
Conv2d(128 süzgeç) – Leaky ReLU + Yığın Normalizasyonu SPP (1, 2, 4) + Dropout (0,3)
100 birimli Yoğun Katman – Leaky ReLU Dropout (0,3)
10 birimli Yoğun Katman – SoftMax
Çıktı (*10*)

Öznitelik çıkarımı aşamasında *mel* enerjileri, *MFCC* ve Spektrogram öznitelik çıkarımları kullanılmış olup her geliştirilen model ile ayrı ayrı test edilmiştir. Öncelikle pencere boyutu 0,04 ve atlama boyutu 0,02 parametre değerlerinin girilmesi ile 40 *mel* enerjileri oluşturulmuş; bu oluşturulan özniteliklerin geliştirilen mimari üzerinde eğitimi sonucu %59,5 doğruluk oranı sonucuna varılmıştır. Ardından pencere boyutu ve buna paralel olarak atlama boyutu %50 düşürülerek, yeni değerlerin 0,02 ile 0,01 olarak belirlenen 40 *mel* enerjileri ve 60 vektörlü *MFCC* öznitelik çıkarımı ile geliştirilen mimari ayrı ayrı test edilerek önceki sonuç ile karşılaştırılmıştır. *Mel* ve *MFCC* öznitelik çıkarımı yöntemleri sonucu önceki yöntemden daha düşük bir performans elde edildiği gözlemlenmiştir. Son olarak Spektrogram öznitelik çıkarımı ile geliştirilen mimarinin kullanımı test edilmiştir. Spektrogram ses sinyalinin frekans spektrumunun zamansal değişkenliğinin görsel bir gösterimidir [43][44]. SPP yönteminin görsel sınıflandırma problemlerindeki başarısı sebebi ile geliştirilen CNN-SPP yönteminin Spektrogram öznitelikleri karşısında başarısı, performansın geliştirdiğini tespit ettik. Elde ettiğimiz sonuçlara göre spektrogram ile geliştirilen mimari kullanımı sonucu %60,1 doğruluk oranı elde edilmiştir. Ayrıca bir konvolüsyonel katmanı daha eklenmesi sonucu %61,6 doğruluk

Şekil 4.4. Geliştirilen CNN-SPP mimarisinin görünümü.



Çizelge 4.4. Geliştirilen mimari ile uygulanama sonuçları.

Yöntem	Öznitelik	airport	bus	metro	metro station	park	public square	shopping mall	street, pedestrian	street, traffic	tram	Genel Toplam
Mesaros vd. [70]	Mel_A	0,729	0,629	0,512	0,554	0,791	0,404	0,496	0,5	0,805	0,551	0,597
2CNN+SPP	Mel_B	0,823	0,649	0,506	0,51	0,893	0,412	0,315	0,433	0,74	0,46	0,574
2CNN+SPP	MFCC_B	0,713	0,674	0,536	0,506	0,855	0,396	0,326	0,462	0,78	0,433	0,568
2CNN+SPP	Mel_A	0,645	0,607	0,406	0,571	0,847	0,38	0,62	0,486	0,841	0,544	0,595
2CNN+SPP	Spectrograms	0,66	0,607	0,521	0,514	0,789	0,384	0,602	0,603	0,817	0,51	0,601
3CNN+SPP	Mel_A	0,834	0,591	0,506	0,405	0,781	0,407	0,591	0,478	0,846	0,529	0,605
3CNN+SPP	**Spectrograms**	0,72	0,619	0,551	0,531	0,891	0,371	0,61	0,621	0,83	0,531	**0,616**
4CNN+SPP	Mel_A	0,675	0,616	0,525	0,471	0,744	0,417	0,448	0,579	0,874	0,527	0,592
4CNN+SPP	Spectrograms	0.71	0,61	0,54	0,501	0,79	0,32	0,59	0,61	0,801	0,49	0,599

A: Pencere boyutu: 0,04 ms, atlama zamanı: 0,02 ms.
B: Pencere boyutu: 0,02 ms, atlama zamanı: 0,01 ms

oranı elde edildiği gözlemlenmiştir. Elde edilen sınıf bazlı sonuçlar Çizelge 4.4'de gösterilmektedir.

Eğitim aşamasında için 200 eğitim tur sayısı (epoch) ve bu turlar arası veri karıştırma ile birlikte 16 *batch size* kullanılmış, deneylerimize dayanarak öğrenim oranı (learning rate) 0,001 ve Adam [69] optimizasyon parametreleri ile öğrenim modeli geliştirilmiştir. Her epoch sonrası ayarlanan doğrulama ile değerlendirilen en iyi performans gösteren model seçilmiştir. Çizelge 4.4'de belirtildiği gibi DCASE 2018 Challenge development veri setini kullanarak elde ettiğimiz sonuçları temel sistem olarak geliştirilmiş Mesaros vd. [70] DCASE task1 baseline system yöntemi ile karşılaştırdığımızda %25 daha kısa eğitim süresi geçirdiğini ve ayrıca bu akustik sahne sınıflandırma görevi için *metro_station*, *park*, *shopping_mall*, *street_traffic* ve *street_pedestrian* gibi bazı sınıfların sınıflandırma başarısında olumlu etki ettiği gözlemlenmiştir. Spektrogram öznitelik çıkarımı ile bu problem karşısında sınıflandırma doğruluğunun başarılı şekilde yükseldiği; önerilen ağın performans, ağ parametreleri, normalizasyon vb. ayarlamalar sonrası daha da geliştirilebileceği düşünmekteyiz.

4.1.4. LSTM ve GRU sınıflandırma mimarileri

Tez kapsamında son olarak AlexNetish ve VGGnish modellerinden temel alınarak dört ve sekiz katmanlı CNN yöntemlerine [28] ek olarak sınıflandırma için LSTM ve GRU yöntemleri mimarinin son konvolüsyonel katmanında bulunan GM havuzlama katmanının ardına eklenmiştir. Girdi olarak ses sinyallerinin öznitelik çıkarımı için 40 *ms* pencere boyutu ve %50 kaydırma boyutu kullanılarak 64 *mel* öznitelik çıkarımından yararlanılmıştır. Çalışma; 2019 yılı Sinyal İşleme ve İletişim Uygulamaları Kurultayında yayınlanmıştır [71]. Çalışmada dört konvolüsyonel katmanlı mimaride (Çizelge 4.5 a); 64, 128, 256, 512 süzgece sahip 5x5 çekirdek boyutlu CNN mimarisi kullanılmıştır. Her bir konvolüsyonel katman ardında 2x2'lik MP katmanları, Relu aktivasyonu ve yığın normalizasyonu kullanılmaktadır. Sekiz katmanlı CNN modelinde ise (Çizelge 4.5 b); her katman ikişer konvolüsyona sahip, dört adet 64, 128, 256 ve 512 süzgeçli 3x3 çekirdek boyutlu CNN katmanları kullanılmaktadır. Her katman 2x2'lik MP, Relu aktivasyonu ve yığın normalizasyonu kullanılmaktadır. Her iki modelde de son konvolüsyonel katmanının ardından GM

havuzlama kullanılıp 10 birimli softmax aktivasyonu içeren çıktı katmanıyla model tamamlanmaktadır. Çizelge 4.5'de dört ve sekiz katmanlı CNN modelleri gösterilmektedir.

Çizelge 4.5. Dört ve sekiz katmanlı CNN modelleri.

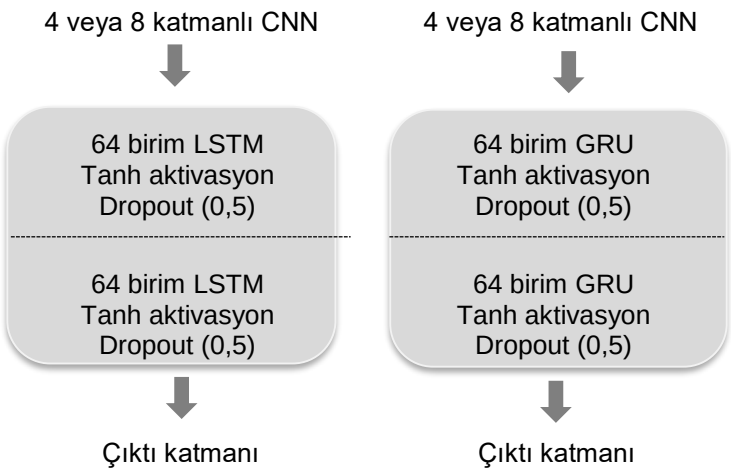
a) Dört Katmanlı CNN Modeli

Girdi (64x320)
5x5, 64 Conv2d-ReLU Yığın Normalizasyonu 2x2 MP - Dropout (0,3)
5x5, 128 Conv2d-ReLU Yığın Normalizasyonu 2x2 MP - Dropout (0,3)
5x5, 256 Conv2d-ReLU Yığın Normalizasyonu 2x2 MP - Dropout (0,3)
5x5, 512 Conv2d-ReLU Yığın Normalizasyonu 2x2 MP - Dropout (0,3)
Global Maksimum Havuzlama

b) Sekiz Katmanlı CNN Modeli

Girdi (64x320)
3x3, 64 Conv2d-ReLU Yığın Normalizasyonu 3x3, 64 Conv2d-ReLU Yığın Normalizasyonu 2x2 MP Dropout (0,3)
3x3, 128 Conv2d-ReLU Yığın Normalizasyonu 3x3, 128 Conv2d-ReLU Yığın Normalizasyonu 2x2 MP Dropout (0,3)
3x3, 256 Conv2d-ReLU Yığın Normalizasyonu 3x3, 256 Conv2d-ReLU Yığın Normalizasyonu 2x2 MP Dropout (0,3)
3x3, 512 Conv2d-ReLU Yığın Normalizasyonu 3x3, 512 Conv2d-ReLU Yığın Normalizasyonu 2x2 MP Dropout (0,3)
Global Maksimum Havuzlama

Çalışma kapsamında [71] dört ve sekiz katmanlı CNN modellerine sınıflandırma için modellerdeki GM havuzlama katmanının ardından yinelemeli sinir ağları algoritmalarından 64 birimli iki katmanlı LSTM ve 64 birimli iki katmanlı GRU tasarlanmış, *Tanh* aktivasyonu ve 0,5 bırakma oranı katmanlara eklenmiştir. Aşağıdaki şekilde dört ve sekiz katmanlı CNN modellerine eklenen LSTM ve GRU modelleri gösterilmektedir.



Şekil 4.5. CNN katmanlarının ardından eklenen LSTM ve GRU modelleri.

Geliştirilen LSTM sınıflandırma yöntemleri dört ve sekiz katmanlı CNN modellerine eklenmesi sonucu sırası ile %68,5 ve %68,2 doğruluk oranı elde edilmiştir (Çizelge 4.6). Ayrıca geliştirilen 64 birimli GRU sınıflandırma yöntemlerinin dört ve sekiz katmanlı CNN modellerine eklenmesi sonucu %70,1 ve %69,9 doğruluk oranı elde edildiği görülmüştür (Çizelge 4.6). Doğruluk oranı dört katmanlı CNN mimarisinde biraz yüksek olması ve öğrenme süresinin sekiz katmanlı CNN mimarisinden kısa olması ile birlikte daha performanslı bir mimari olduğu söyleyebiliriz.

Çizelge 4.6. Elde edilen test sonuçları.

Yöntem	**Model**	**Öznitelik**	**Doğruluk (%)**
Mesaros [70]	2 CNN + FC	mel	59,7
Kong [28]	CNN4	mel	67,6
	CNN8	mel	68
Önerilen Sistem 1	CNN4+LSTM	mel	68,5
	CNN4+GRU	mel	**70,1**
Önerilen Sistem 2	CNN8+LSTM	mel	68,2
	CNN8+GRU	mel	69,9

Son olarak bahsettiğimiz yöntemler ve geliştirilen CNN4-GRU modelimiz temel sistemden [70] %10,4 daha yüksek ve Kong vd. [28] çalışmasında kullandığı CNN4 yöntemine göre %2,5 daha yüksek başarı elde etmektedir. CNN8-GRU modeli ise CNN4-GRU modeli sonucundan %0,2 az da olsa düşük sonuç verdiği gözlemlenmiştir. Öğrenme süreci uzunluğuna bakılacak olunursa, CNN4-GRU

modeli CNN8-GRU modelinden daha kısa sürede öğrenme süreci tamamlaması (Çizelge 4.7) ve diğerine göre küçük bir doğruluk oranı farkına sahip olması nedeniyle daha performanslı bir öğrenme süreci geçirdiğini söyleyebiliriz (Çizelge 4.6). Gelecek çalışmalar arasında farklı öznitelik çıkarım modellerinin probleme uygulanması yer almaktadır. Deneysel çalışma sonucunda aşağıdaki bulgular kaydedilmiştir.

Çizelge 4.7. Önerilen sistemlerin öğrenme süreçleri.

Yöntem	**Model**	**Süreç**
Önerilen Sistem 1	CNN4+LSTM	**64 saat**
	CNN4+GRU	
Önerilen Sistem 2	CNN8+LSTM	147 saat
	CNN8+GRU	

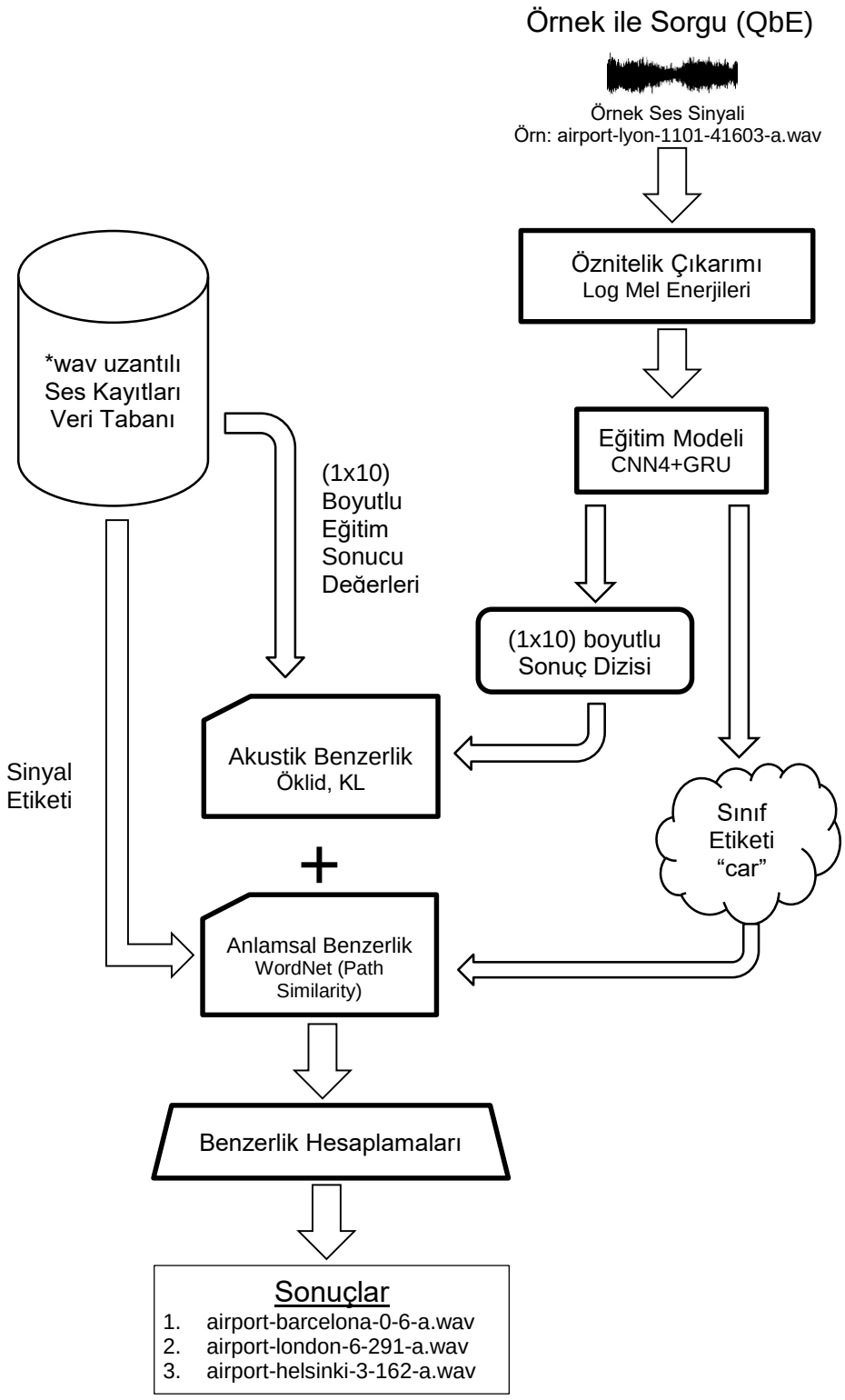
- GRU ve LSTM yapılarının eklenmesi akustik sahne sınıflandırma başarımını artırmaktadır.
- GRU yapısının sınıflandırma başarımı LSTM yapısından %1,6 daha yüksektir.
- GRU algoritmasının CNN4 ve CNN8 modellerine eklenmesi sonucunda CNN4+GRU mimari modelinin CNN8 modeline göre daha yüksek doğruluk yüzdesi vermektedir.

5 SES SAHNE GERİ GETİRİMİ

Teknolojinin gelişmesi ile birlikte internet ortamında verilerin hızlıca paylaşılıp erişilmeye çalışılması; çokluortam verilerinin çoğalıp depolama alanlarında kapladıkları yerlerin artmasına neden olmuştur. Bu çokluortam verilerinin artışı ile birlikte milyonlarca veri arasında istenilen verinin arama motoru sistemlerinde geri getirim problemi üzerinde çalışmalar yoğunlaşmıştır. Kullanıcıların aradıkları ses verisinin veri tabanlarında bulunan diğer ses verilerinin arasından kolayca aranılıp, hemen erişilebilme imkânlarının geliştirilmesine büyük önem verilmektedir. Bu işlemler için büyük ses dosyaları içeren çokluortam veri tabanlarında indekslenme performansının arttırılması ve ilgili arama motorlarının veriye hemen erişebilecek şekilde geliştirilmesi araştırmacılar tarafından son dönemlerdeki güncel çalışmalar arasında yer almaktadır. Terabaytlara varan ses dosyalarının indekslenmesi ve ileri düzeyde işlenmesi bu problemler için önemlidir.

Tez çalışmasının bu bölümünde bilgisayar ortamlarında ses sinyallerinin otomatik sınıflandırma işlemlerini gerçekleştirmek üzere geliştirdiğimiz modeller ile hedeflenen sinyal verilerinin geri getirimi işleminin sağlanabilmesi ve sistemin performans artışı sağlaması üzerine araştırmalar yapılması hedeflenmiştir. Bu amaç doğrultusunda ses sinyalleri içerisindeki akustik sahnelerinin öznitelikleri çıkarılmış, bu öznitelikler Ses Olay ve Akustik Sahne Sınıflandırıcı bölümünde bahsedilen sınıflandırıcılar yardımıyla tanımlama yapılabilmesi üzerine çalışılmıştır. Yine bu bölümde geliştirilen model ile eğitilen ses verileri üzerinde, içerik bazlı arama yapabilmek için örnek ile sorgulama (QbE) modeli kapsamında yaklaşık 5000 adet test için ayrılmış işitsel ses dosyaları sorgu girdisi olarak gönderilmek istenmiştir. Bu sorgu verilerine göre çokluortam veri tabanımız içinde bulunan benzer ses verilerinin getirilmesi ve sistemin performans artışı üzerine araştırma çalışmaları yapılmıştır.

Bu çalışma kapsamında içerik tabanlı işitsel benzerlik çalışmaları ile birlikte etiketlerinin anlamsal benzerliğinin değerlendirilmesi sonucu ses sinyalleri üzerinde veri geri getirim işlemlerinin geliştirilmesi üzerinde çalışmalar yapılmıştır. Anlamsal benzerlik hesaplamaları için büyük bir sözcüksel veri tabanı olan WordNet [36]



Şekil 5.1. Önerilen geri getirim sisteminin genel görünümü.

kullanılması düşünülmüştür. WordNet üzerinde fiiller, zarflar, sıfatlar ve isimler gibi etiketler bulunmakta ve bu etiketler arasındaki ilişkilerin her biri ayrı bir kavram ifade eden bilişsel eşanlamlı (*synsets*) kümeleri halinde gruplandırılmaktadır. Böylece etiketler bulunmakta ve bu etiketler arasındaki ilişkilerin her biri ayrı bir kavram ifade eden bilişsel eşanlamlı (*synsets*) kümeleri halinde gruplandırılmaktadır. Böylece anlamsal ilişkili bir hiyerarşi olarak temsil edilebilmektedir. Bu aynı kavramı ifade eden eşanlamlı ve birçok bağlamda birbirinin yerine geçebilen ifadeler kavramsal, anlamsal ve sözcüksel ilişkiler aracılığıyla birbirine bağlanmaktadır. Bir bilişsel eş anlamlı veri öğesi içinde çok farklı manaya sahip anlamsal olarak eş değer kabul edilen veri elemanları grubu içerebilmektedir. WordNet kavramında birçok bağlamda birbirinin yerine geçebilen kelimeler üzerinde benzerlik yöntemleri mevcuttur.

Önerilen yöntem ile ortalama hassasiyet öçütü *mAP* puanları ve P@k değerleri üzerindeki sonuçlar gözlemlenmiştir. Örnek tabanlı sorgulama için önerdiğimiz sınıflandırma mimarisinin genel görünümü Şekil 5.1'de gösterilmektedir. Bu tez kapsamının bu bölümündeki hesaplama çalışmaları Tensorflow [62] kütüphanesinin GPU kipinde kullanımı ile birlikte iki adet NVidia M2090 GPU grafik kartı kullanılmıştır.

5.1 Akustik Sahnelerde İşitsel Benzerlik

İşitsel benzerlik, ses sinyallerinin çerçeve tabanlı gösterimleri arasındaki mesafe ile ya da bu sinyallerin istatistiksel modelleri arasındaki mesafe ölçümlerine dayanan objektif ölçümler kullanılarak ölçülebilir [72]. Tez çalışmasının bu bölümünde; çevresel ses kayıtları içerisindeki akustik sahnelerin işitsel benzerliği ele alınmış; ses verisi geri getirimi konusu üzerinde çalışılmıştır. Çalışmanın ilk ayağında örnek ses sinyallerinin öznitelik çıkarımı yapıldıktan sonra çokluortam veri tabanı üzerinde bulunan ses verilerinin öznitelikleri ile benzerlik performansı araştırılmıştır. En temel eski yöntemlerden olan öznitelik vektörleri ile benzerlik karşılaştırması deneylerinden sonra çalışmanın bir sonraki ayağında bu tez çalışmasının önceki bölümde anlatılan *CNN4-GRU* sınıflandırma eğitim modelinin eklenmesi ile geri getirim (retrieval) sistemi geliştirilmeye çalışılmıştır. Geliştirme sonrası deneylerdeki performans kazanımları gözlemlenerek tarafımızca çalışma kapsamında raporlanmıştır.

Her iki deneyde de ses sinyal verilerinden temel frekans veya sinyal yüksekliği gibi anlamsal verilerin çıkarılması için öznitelik çıkarım adımları uygulanmıştır. Bu işlem kapsamında her ses sinyali için frekans aralığı 0-22050 *Hz*. olan 40 *ms*. pencere boyutu ve %50 atlama oranı ile 64 adet *Mel* enerjileri özniteliği çıkarımı kullanılmıştır. İşitsel benzerlik için popüler olarak kullanılan *KL-divergence* [72] ve Öklid uzaklığı(*Euclidean distance*) [73] hesaplamaları kullanılması düşünülmüştür. Ses sinyallerinde genellikle Öklid uzaklığı; iki öznitelik vektörü arasındaki uzaklığının hesaplanarak birbirine yakın değerlerin bulunması amacıyla birçok işitsel benzerlik problemlerinde kullanılmıştır. Öklid uzaklık denklemi aşağıdaki denklemde herhangi bir düzlemde bulunan iki noktanın (*q* ve *p)* birbirlerine Öklid uzaklığı (*d*) hesabı ölçümü gösterilmektedir.

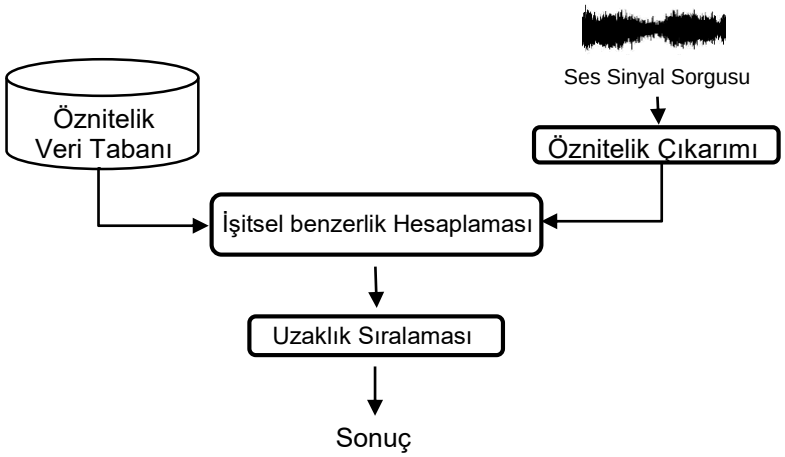
$$d(p,q)^2 = (q_1 - p_1)^2 + (q_2 - p_2)^2 \qquad (5.1)$$

KL-divergence ise; teoriksel olarak iki olasılık dağılımı arasındaki dağılımın ölçülmesinden ilham alınmıştır. Gözlenen özellik dağılımı ve sınıfın özellik dağılımı arasında en düşük *KL-divergence* ölçülerek gözlemin hangi sınıfa dahil olabileceği tahmin edilmesi hesaplamalarında kullanılmaktadır. Simetrik bir uzaklık sonucu oluşmayacağı için simetrize edilmektedir. İki öznitelik vektörü işleme alındığında, Öklid uzaklığında olduğu gibi, uzaklık farkı sonucu ne kadar düşük olursa birbirine de o kadar benzerdir. Bu yüzden aşağıdaki denklemde sonucunun minimum değerleri işitsel benzerlik probleminde dikkate alınır. Aşağıdaki denklemde *KL* uzaklığı ölçümünde olasılık dağılımı *P(x)* ile yaklaşık dağılım *Q(x)* logaritmik farkları ele alınarak yapılan ölçüm gösterilmektedir:

$$D_{KL}(P \,||\, Q) = \sum_{x \in X} P(x) \log\left(\frac{P(x)}{Q(x)}\right) \qquad (5.2)$$

Tez çalışması kapsamında ilk olarak; örnek ses sorgusu üzerinde öznitelikleri ve çokluortam veri tabanı üzerinde saklanan ses sinyallerinin öznitelikleri çıkarılmıştır. Ardından elde edilen öznitelik verileri arasında işitsel benzerlik kapsamında bir uzaklık hesabı yapılmıştır. Bu kapsamda işitsel benzerlik uygulamalarında popüler olarak kullanılan Öklid uzaklığı ve *KL-divergence* hesaplamaları ile öznitelik vektörleri arasındaki uzaklık hesaplanmıştır. Bu hesaplama ile en ilişkili yani yakın seslerin en yakın uzaklıkta olup, ilişkisi az olan veya herhangi bir benzerlik

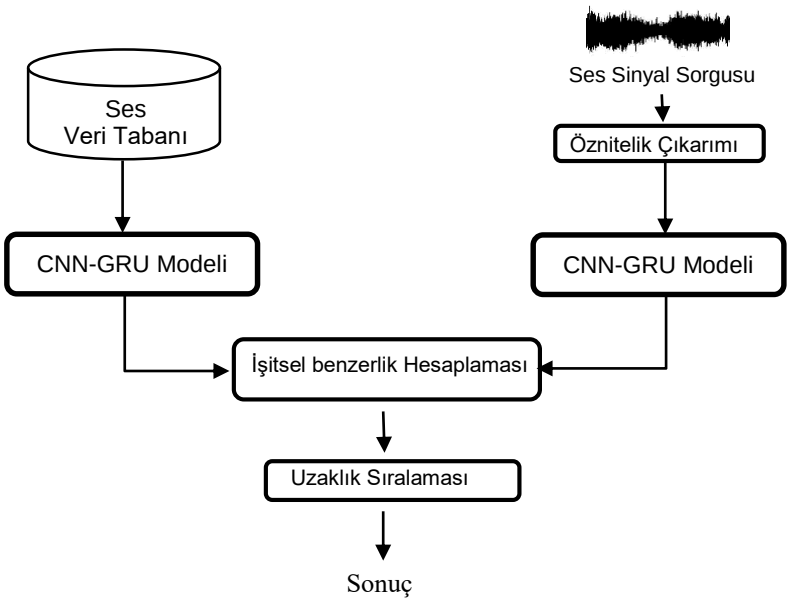
hesaplanamayan ses sinyallerinin ise en uzakta olacak şekilde sıralanabilmektedir. Bu sıralamaya göre özniteliklerin çıkarımı ile ses sinyalleri benzerliği işlemi yapılmış ve performans izleminleri alınmıştır.



Şekil 5.2. Öznitelikleri çıkarılmış ses sinyali sorgusu ile öznitelik veri tabanı arası işitsel benzerlik uygulaması genel bakışı.

Çalışma kapsamında içerik tabanlı erişim yöntemi için Bölüm 4'de yer alan geliştirdiğimiz sınıflandırma modelinin kullanımı ile çalışmanın QbE kısmının geliştirilmesi planlanmıştır. Eğitimde geliştirdiğimiz model olan *CNN4-GRU* yöntemi kullanılarak işitsel benzerlik probleminin performans gelişimi gözlemlenmiştir. Bu çalışma kapsamında kullanılan eğitim modeli 3x3 çekirdek boyutlu 32, 64, 128, 256 süzgeç değerlerine sahip dört konvolüsyonel katmanlarından oluşmaktadır. CNN katmanının ardından iki adet 64 birimli GRU katmanları kullanılarak *softmax* sonuç katmanıyla model oluşturulmuştur. Sorgu ile gönderilen örnek ses sinyalinde ilk olarak 64 adet *mel* enerjileri özniteliği çıkarımı yapılmıştır. Bu öznitelik verileri daha sonra eğitim modeline yönlendirilmiştir. Eğitim modeline yönlenmesi ile model tahmin sonuçları (probability) değerlerini kullanarak; eğitim sürecinde veri tabanında bulunan eğitilmiş ses sinyallerinin model tahmin sonuçları (probability) ile işitsel benzerlik değerleri hesaplanmıştır. Ayrıca model kullanımı ile yapılan bir başka deneyde; sorgu ses sinyalinin modelde tahmin öncesi son katmanda oluşan özellik değerleri ile veri tabanında bulunan eğitilmiş ses verilerinin model tahmin öncesi son katman değerleri arasında işitsel benzerlik değerleri hesaplanmıştır. Deneyler

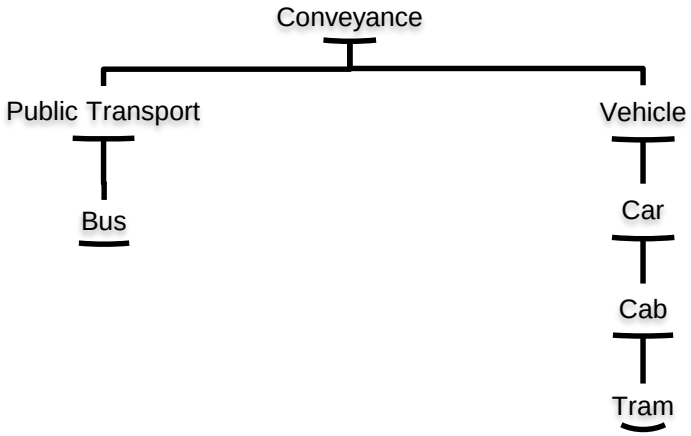
sonucunda model tahmin sonuçları üzerinde benzerlik işlemi ile “CNN-GRU (probability)” ve tahmin katmanı öncesi elde edilen değerlerin benzerlik işlemi “CNN-GRU” olmak üzere iki ayrı sonuç elde edilmiştir. İşitsel benzerlik hesaplamaları Öklid ve KL-divergence hesaplamaları ile ayrı ayrı ölçüm deneyleri yapılmıştır.



Şekil 5.3. Model ile QbE sisteminin genel bakışı.

5.2 Anlamsal Benzerlik

Anlamsal benzerlik kapsamında WordNet üzerinde tanımlanan birçok benzerlik yönteminden; hiyerarşi üzerinde karşılaştırılan iki kelime arasındaki en kısa yolun tersi olarak hesaplanan yol benzerliği (*path similarity*) yönteminin kullanılması tercih edilmiştir. Bu yönteme örnek verilirse; iki ingilizce kelimeden oluşan etiketteki (*bus* ve *tram*) *path similarity* değeri aralarındaki düğümleri içeren yolun tersi nedeniyle 0,25'tir. Şekil 5.4’de bu örnek olarak açıklanabilir. Bu sistemde, yol benzerliğinin sınır değerleri 0 (benzer değil) ile 1 (eş) arasında olabilir.



Şekil 5.4. WordNet üzerinde *bus* ve *tram* sınıflarının arasındaki yol benzerliği.

Benzerlik hesaplanmasında ses verilerinin akustik (*A*) ve semantik (*S*) benzerlik hesaplamalarının birlikte ele alınabilmesi için P@k hesaplamaları sırasında her bir örnekte ortak bir benzerlik denklemi geliştirildi. Test olarak sorguda gönderilen ses sinyali verisinin içerik bazlı arama kapsamında etiket ile anlamsal benzerlik araması yapılmıştır. Çokluortam veri tabanında bulunan ses sinyallerinin etiketleri ile anlamsal benzerlik kurularak geri getirim sisteminin performansı gözlemlendi. Anlamsal benzerlik ile akustik benzerlik belirlenen bir *w* semantik ağırlık değerleri ile çarpılarak toplamı ortak bir benzerlik sonucu *C* elde edilmek üzere aşağıdaki denkleme göre hesaplanmıştır. Denklemde *w* ağırlık değeri 0 ile 1 arasında verilecektir. 0 ile akustik ağırlık değerleri ele alınmış, 1 ile de semantik benzerlik değerleri hesaplamaya alınarak benzer sinyallerin veri tabanındaki ses verilerinin geri getirimi hesaplandı. Bu denkleme göre anlamsal benzerlik (Query-by-Keyword) ile akustik benzerliğin QbE çatısı altında hesaplanmasında destek olması düşünülmüştür.

$$C = (1 - \mathrm{w}) \cdot \mathrm{A} + \mathrm{w} \cdot \mathrm{S} \qquad (5.3)$$

5.3 Deneysel Çalışmalar

Açık kaynak kodlu yazılım dili olan Python ve bu programlama diline bağlı kütüphanelerin kullanımı ile çalışmanın bu kısmı gerçekleştirilmiştir. Çalışma kapsamında kullanılan kütüphanelerden bazıları; derin öğrenme algoritmaları ve öznitelik çıkarımı için kullanılan Tensorflow [62], Keras [63] ve Theano [64]

kütüphaneleridir. Semantik benzerlik için kullanılan WordNet için nltk [74] kütüphanesi kullanılmıştır.

Bu araştırmada yer alan tüm veya kısmi nümerik hesaplamalar TÜBİTAK ULAKBİM, Yüksek Başarım ve Grid Hesaplama Merkezi'nde (TRUBA kaynaklarında) gerçekleştirilmiştir [66].

5.3.1 Kullanılan Veri Kümeleri

Çalışmanın bu bölümünde veri kümesi gerçek hayattan alınmış çevresel ses kayıtlarında bulunan akustik sahnelerden oluşturulmuştur. Karşılaştırmanın veri tabanı kısmı; eğitim için kullandığımız DCASE topluluğu tarafından geliştirilen TUT Urban Acoustic Scenes 2018 [11] veri kümesi tercih edilmiştir. Bu veri kümesi üzerinde, benzer ses verilerini arama ve benzer sonuçların geri getirim işlemleri için test sorguları olarak yine DCASE topluluğu tarafından geliştirilen TAU Urban Acoustic Scenes 2019 [12] veri kümesi kullanılması planlanmıştır. Bu veri kümesinden önceki veri kümelerinde aynı ses dosyalarının ayrıştırılması ile birlikte yaklaşık 5670 adet test verisi elde edilmiştir. Bu test verileri ile örnek bazlı sorgulama işlemlerinde bu çalışma kapsamında kullanılacaktır. Sorgu olarak kullanılacak test verileri; her bir dosyanın süresi 10 saniye olmak üzere toplamda 12 saatlik ses dosyaları içermektedir. Her iki veri setinde kayıt edilmiş sesler; *airport, indoor shopping mall, metro station, pedestrian street, public square, traffic, travelling by a tram, travelling by a bus, travelling by an underground metro* ve *park* olmak üzere on adet sınıf bulunmaktadır.

5.3.2 Deneyler ve Sonuçları

Çalışmada ilk olarak sadece öznitelik matrislerinin karşılaştırılması ile geliştirilen yöntemde; ses sinyalleri veri tabanı ile ses sorgusu üzerinde öznitelik çıkarımı yapıldıktan sonra işitsel benzerlik hesaplamaları uygulanmıştır. Bu çalışmanın ikinci deneyinde ise; öznitelik çıkarım adımından sonra eğitilmiş CNN-GRU modelinin eklenmesi ile veri geri getirimi ve çıkarımı problemi performansı gözlemlenmiştir. CNN-GRU yönteminde tahmin olarak gönderilen *Sigmoid* çıktı katmanında oluşan değerlerin veri tabanında bulunan eğitilmiş ses sinyallerinin de *Sigmoid* çıktı katmanında oluşan değerleri arasında bir akustik benzerlik sonucu hesaplanmıştır.

Bu hesaplama sonuçları Çizelge 5.1'de *"CNN+GRU (probability)"* olarak gösterilecektir. Ardından CNN-GRU yönteminde çıktı katmanından önce oluşan değerler (Tahmin Katmanından önce *Sigmoid* çıktı katmanına gönderilen değerler) ile çokluortam veri tabanımızda tutulan ses sinyallerinin eğitim aşamasında çıktı katmanından önce üretilmiş değerler arasında benzerlik hesaplamaları yapılarak işitsel benzerlik hesaplanmıştır. Böylece içerik tabanlı geri getirim sistemi için geliştireceğimiz üç ayrı QbE yönteminin işitsel benzerlik hesaplamaları öncesi akustik benzerlik sistemine gönderilecek girdiler hazırlanmıştır.

İşitsel benzerlik ölçümünde Öklid ve *KL-divergence* hesaplamaları ayrı ayrı kullanılmıştır. Fakat iki yöntemin kullanımlarında performans sonuçlarına dair önemli farklılıklar görülmemiştir. Hesaplamalar sonucu; sorgu ile veri tabanı arasındaki en düşük uzaklık en yakın kabul edilerek benzer veri olduğu belirtilecektir. Hesaplama aşamasından sonraki en düşük uzaklıktan en uzağa olan sıralama aşamasında P@k ortalama hassasiyet (AP) [75][76] ölçümleri her k değerinin 3, 5, 10 ve 20 değerleri için yapılmıştır. Ortalama hassasiyet hesaplanması ise şu şekildedir:

$$AP = \frac{\left(\sum_{n=1}^{R} \frac{n}{rank_n}\right)}{R} \qquad (5.5)$$

En temel eski yöntem olarak belirttiğimiz ilk yöntemimizde %15 civarında bir P@k tahmin değer sonuçları başarımı gözlemlenmiştir. Bu yöntemde etiketlerin kullanımı ile anlamsal benzerlik yöntemi eklenmemiştir. Ardından ikinci deneyde modelimizi kullanarak işitsel benzerlik deneylerinde elde edilen P@k tahmin değer sonuçları %60 civarında olduğu gözlemlenmiştir. Sadece işitsel benzerlikte akustik içerik temel alınan erişim yönteminde elde edilen bu sonuç ardından anlamsal benzerlik yöntemi eklenmesi sonucu ayrı bir deney sonucu elde edilmiştir. CNN+GRU modelinin tahmin çıktı katmanı değerlerinin ele alındığı bu anlamsal-akustik benzerlik yöntemi sonucunda %66 civarında ortalama tahmin sonuçları elde edilmiştir. Değerler Çizelge 5.1'de CNN+GRU (probability) geri getirim modeli altında gösterilmektedir.

Çizelge 5.1. Yöntemlerde elde edilen P@k sonuçları.

Geri Getirim Modeli	***W*=0**	**P@k=3**	**P@k=5**	**P@k=10**	**P@k=20**
İlk Yöntem		0,1518	0,1528	0,1493	0,1423
CNN+GRU (probability)	0	0,5994	0,5995	0,5998	0,6003
CNN+GRU (probability) + Anlamsal	0,3	0,6103	0,6112	0,6116	0,6122
CNN+GRU (probability) + Anlamsal	0,9	0,6390	0,6392	0,6411	0,6422
CNN+GRU (probability) + Anlamsal	**1**	**0,65**	**0,654**	**0,658**	**0,661**
CNN+GRU	0	0,5568	0,5571	0,5572	0,5573
CNN+GRU + Anlamsal	0,3	0,5656	0,5652	0,5670	0,5686
CNN+GRU + Anlamsal	0,5	0,5765	0,5767	0,5770	0,5771
CNN+GRU + Anlamsal	0,9	0,5922	0,5924	0,5926	0,593
CNN+GRU + Anlamsal	1	0,597	0,5972	0,5984	0,5994

Deneylerde *CNN4+GRU* modelimizin *Sigmoid* çıktı katmanı öncesinde elde edilen değerler ile benzerlik yöntemi üzerinde çalışılmıştır. Sorgulanan ses sinyali ile veri tabanında bulunan ses sinyali arasında yapılan akustik benzerlik karşılaştırmalarına etiketlerin semantik benzerlik sonuçları eklenmesi ile Çizelge 5.1'de CNN+GRU geri getirim modeli başlığı altında sonuçlar elde edilmiştir. Ayrıca semantik ağırlık değeri (*w*) kullanım sonuçları yine Çizelge 5.1'de gösterilmektedir. Çizelge 5.1; benzerlik sıralaması sonucunda ilk 3, 5, 10, 20 benzerlik sırası içerisinde doğru tahmin edilme sonucunu ölçen, her k değeri için P@k tahmin sonuçları gösterilmiştir.

Gözlemlenen deneyler sonucunda en iyi başarıma ulaşmış model olan CNN+GRU (probability) anlamsal benzerlik yönteminde elde edildiği görülmüştür. Sınıf bazında *AP* sonuçları Çizelge 5.2 ile verilmektedir. Bu tabloda elde edilen sonuçları gözlemlediğimizde hassaslık değerlerine göre yaklaşık %80 P@k başarım sonucunu geçebilen *metro*, *park*, *shopping_mall* ve *street-traffic* kategorisindeki sınıflar olduğu görülmektedir. Bu sınıflar ile elde edilen başarım gelişimi yüksek olduğu söylenebilir. Bazı sınıflardaki başarım artışına rağmen önerilen yöntemde yaklaşık %20 ve %29 P@k değerlerinde gözlemlenmiştir. Bu sınıflar *public_square* ve *street_pedestrian* kategorileri olduğu görülmektedir. Elde edilen bu sonuçların sınıf bazında diğer sınıflara göre düşük olduğu söylenebilir. Diğer sınıflarda P@k değerlerinin yaklaşık %50 civarında olduğu gözlemlenmiştir.

Çizelge 5.2. Geliştirilen yöntem ile sınıf bazlı sonuçlar.

Sınıflar	**P@k=3**	**P@k=5**	**P@k=10**	**P@k=20**
Airport	0,549	0,549	0,548	0,549
Bus	0,463	0,464	0,453	0,466
Metro	0,876	0,877	0,877	0,878
metro_station	0,517	0,516	0,518	0,519
Park	0,919	0,919	0,919	0,92
public_square	0,286	0,287	0,291	0,293
shopping_mall	0,819	0,818	0,817	0,817
street_pedestrian	0,189	0,191	0,192	0,190
street_traffic	0,865	0,864	0,864	0,864
Tram	0,509	0,509	0,508	0,507

İçerik bazlı erişim için geliştirilen yöntemlerin bu çalışmalarda elde edilen P@k sonuçlarına göre akustik benzerlik ile başarım artışının gerçekleştiği görülmektedir. Bu artışın anlamsal benzerliğin de yönteme eklenmesi ile arttığı gözlemlenebilmektedir. Mesaros vd. [35] akustik benzerlik çalışmalarında GMM ve *MFCC* öznitelik çıkarımı yöntemlerinin kullanımı sonucu elde edilen sonuçları; yöntemimiz ile karşılaştırdık. Bu karşılaştırma için çalışmalarında 0,9 anlamsal ağırlık kullanıldığında k=20 değerinde elde edilen P@k değerin için yaklaşık %58 civarında bir hassaslık sonucu elde edildiği görülmektedir. Ardından anlamsal ağırlık yönteminin çalışma kapsamında 1 verilmesi sonucunda k=20 değeri için P@k sonucu %85 hassasiyet sonucu elde edildiği görülmüştür. Çalışmamızda geliştirdiğimiz mimaride *mel* öznitelik çıkarımları CNN-GRU yönteminin tahmin çıktıları kullanılması ile elde edilen sonuçlarda; semantik ağırlığı 0,9 verildiği zaman 0,6422 P@k değeri elde edilmesinin yanı sıra 0,281 *mAP* sonucu elde edildiği görülmektedir. Ayrıca yine aynı modelde anlamsal ağırlık değeri 1 verildiğinde P@K değeri yaklaşık 0,661 ve *mAP* sonucu 0,292 olduğu gözlemlenmiştir. Anlamsal ağırlık 0 verildiği deneylerde ise bu değerlerden biraz düşük sonuçlar elde edildiği gözlemlenmektedir. Bu sonuç elde edilirken k değerinin 20 olduğu ve anlamsal benzerlik yönteminin çalışmanın bu sonucuna eklenmediğini belirtmemizde fayda görmekteyiz. Çalışmamızın bir başka yönteminde geliştirilen CNN-GRU yöntemimize anlamsal benzerlik yönteminin de eklenmesi sonucu 0,661 ortalama hassasiyet değerinin geliştirdiği görülmektedir. Çizelge 5.3'de önerilen yöntemin P@k değeri ile birlikte *mAP* değerlerinin sonuçları ve önceki Mesaros vd. [35] yöntemi ile karşılaştırılması gösterilmiştir.

Çizelge 5.3. Önerilen geri getirim modelinin P@k ve mAP sonuçları.

Geri Getirim Modeli	**Öznitelik**	**Anlamsal Ağırlık**	**Öğrenim Modeli**	**mAP**	**P@k=20**
Önerilen model	Mel	0	CNN+GRU (probability)	0,261	0,6003
Önerilen model	Mel	0,9	CNN+GRU (probability)	**0,281**	**0,6422**
Önerilen model	Mel	1	CNN+GRU (probability)	**0,292**	0,661
Mesaros [35]	MFCC	0,9	GMM	0,09	0,58
Mesaros [35]	MFCC	1	GMM	0,16	0,85

Elde edilen sonuçları irdeleyecek olursak çevresel seslerde akustik sahnelerin işitsel benzerlik yönteminin anlamsal benzerlik ile hesapladığımız bu çalışmamızda, geliştirilen mimarimizde performans kazanımları elde edildiği söyleyebiliriz. *mAP* skoruna bakılacak olunursa geliştirilen modelin ses veri geri getirim probleminde akustik içerik ve anlamsal benzerliklerin kullanımı ile birlikte performans kazanımı elde edildiği söylenebilir.

Çalışma kapsamında ayrı ayrı kullanılan *KL-divergence* ve Öklid uzaklığı hesaplamalarının kullanıldığı, fakat herhangi bir gözle görülecek şekilde farklı sonuçlar elde edilmediği görülmektedir. Model kullanımı ile öznitelik karşılaştırmada kullanılan özellik matris dizi boyutundan (örn: 64x320 matris dizi boyutundan 1x10 matris dizi boyutuna) daha düşük boyutta matrisler elde edimi ile benzerlik ölçümlerinde sistem için avantaj sağlanmaktadır. Ayrıca anlamsal ağırlık değeri 0-1 aralığı verildiğinde elde edilen sonuçlara bakılacak olunursa, modelimizin akustik benzerlik hesaplamalarının etiket bazlı benzerlik işlemi hesaplamalarına göre sonuçların çok da büyük bir fark olmayacağı görülmektedir. Akustik geri getirim yönteminin etiket bazlı geri getirim yöntemine yakın başarıda sonuçlar verdiği görülmektedir.

Anlamsal benzerlik yönteminin geliştirilerek çalışmaya eklenmesi ve üzerinde geliştirilmelere devam edilmesi ile gelecek çalışmalarda sistemin daha da geliştirilmesi için bir yol açacağı ön görülmektedir. Sınıf bazında sonuçları irdelediğimizde yüksek başarım elde edilen sınıfların olması ile birlikte geliştirilen modelin etkili olduğu söylenebilir.

6 SONUÇLAR VE DEĞERLENDİRME

Tez çalışması kapsamında, çevresel seslerden oluşmuş ses klipleri içerisindeki ses olayları tanımlanması ve akustik sahnelerin sınıflandırılması problemleri üzerine çalışmalar yapılmıştır. Ayrıca akusik sahneler üzerinde geri getirim sistemi geliştirilmiştir. Her çalışmada yapılan deneylerin başarım ve performans sonuçları gözlemlenmiştir. Çalışmada *MFCC,* mel, *spektrogram* gibi farklı öznitelik çıkarımları teknikleri kullanılmış; *MFCC* ve *mel* özniteliklerinin çıkarım aşamalarında parametre değerlerinde değişiklik yapılarak sonuçlara etkisi incelenmiştir. Sınıflandırma eğitimi için MLP, RNN, LSTM, CNN gibi çeşitli sinir ağları algoritmaları kullanımı ile geliştirilen modellerin ses olayı tanıma ve akustik sahne sınıflandırma problemleri karşısında performansı incelenmiştir. Ayrıca geliştirilen sınıflandırıcı modeli kullanımı ile örnek tabanlı sorgulama yapılarak QbE geri getirimi sistemi üzerinde çalışılmış ve performans karşılaştırılması yapılmıştır.

Çalışmanın ses olayı tanıma kısmı için yapılan deneyde elde edilen sonuçlara göre daha fazla öznitelik çıkarımı sağlayacak küçük analiz çerçeve boyutları ile elde edilen özniteliklerin başarıma olumlu bir katkı vermediği görülmesi çalışmamız açısından bir dezavantaj olmuştur. Yine aynı pencere boyutu değişikliğinin yapıldığı ikinci deneyimiz akustik sahne sınıflandırma deneyimizde ise başarıma olumsuz katkı verdiği görülmüştür. Ayrıca deneyimlerimize dayanarak bellek açısından daha maliyetli bir çalışma olduğunu söyleyebiliriz. Bu yüzden pencere boyutunun standart çıkarım parametrelerinin kullanılması sonraki deneylerimizde tercihimiz olmuştur. Ses olay tanıma problemi için ayrıca daha farklı sınıfları içeren daha fazla ses kayıt dosyası bulunabilecek bir veri kümesi üzerinde kullanımının tercih edilmesi, ileriki çalışmalar için düşünülmektedir. Ses olay tanıma probleminde avantaj olarak aktivasyon parametrelerinin denenmesi sonucu çalışmada olumlu sonuç verdiği görülmüştür. Bunun üzerine akustik sahne sınıflandırma probleminde de Leaky ReLU işlevinin kullanımına tercih edilmiştir.

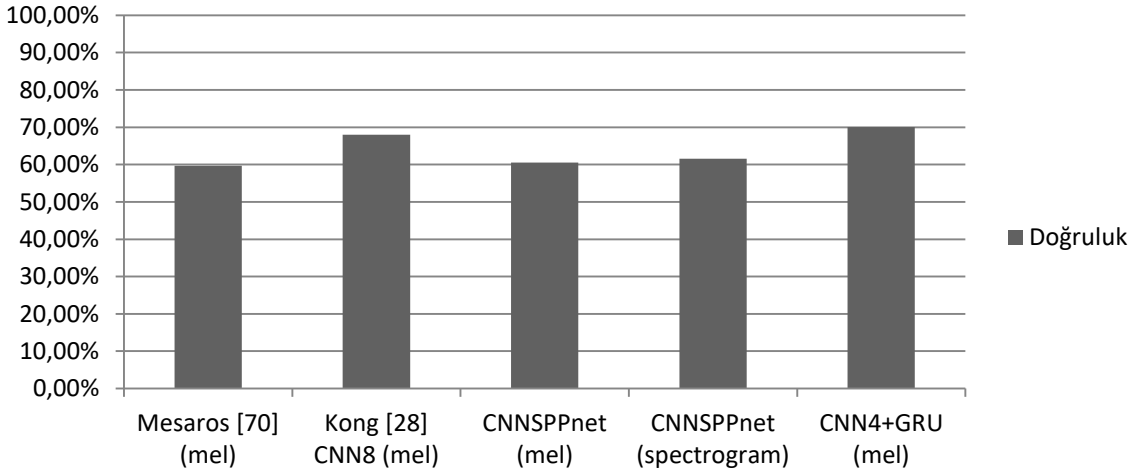
Ses olaylarının akustik sahnelerinin sınıflandırılması problemi üzerinde yaptığımız çalışmada imgesel sınıflandırma problemlerinde başarılı sonuçlar veren SPP yönteminin ilk kez bu problem kapsamında kullanılması, çalışmanın en önemli avantajı olarak görülmektedir. Bu problem için zamanla değişen ses sinyalinin

frekans ve genlik bilgisinin görsel temsili olan spektrogramlar öznitelik olarak kullanılması düşünülmüştür. CNN-SPPnet mimarimizin spektrogram öznitelikleri ile kayda değer başarım performansı elde edilmiştir. Sonuçlara baktığımızda *metro*, *park*, *street pedestrian* ve *street traffic* gibi sınıflar üzerinde başarılı sonuçlar verildiği görülmektedir. Her bir ses kayıt dosyasının sabit uzunlukta olması, ve dosya uzunluklarının değiştirilememesi çalışmamızı kısıtlaması üzerine dezavantaj oluşturmaktadır. Farklı sürelere sahip olan ses sinyal dosyaları içeren veri kümesi kullanımı tercih edilebilir. Sonuç olarak daha başarılı doğruluk oranı elde edildiği ve eğitim süresinde daha da kısaldığı gözlemlenmiştir. Eğitim süresi bakımından %25 oranında eğitim zaman kısalması tespit edilmiştir. İlerleyen çalışmalar için farklı veri kümesi üzerinde geliştirdiğimiz mimari ile yapılan çalışmalar devam etmektedir. Ayrıca spektrogram özniteliklerinin daha gelişmiş derin sinir ağları ile oluşturulmuş mimarilerin kullanımı düşünülmektedir.

Tez çalışması kapsamında çevresel sesler içeren ses klipleri içinde meydana gelen çeşitli ses olaylarının zamansal bilgiler içermesi ve bu bilgilerin işlenerek sınıflandırabilmesi için geliştirilen yinelemeli sinir ağları yöntemlerinin CNN mimarileri üzerine denemeler yapılmıştır. Bu kapsamda AlexNetish ve VGGish mimarileri üzerine GRU ve LSTM algoritmalarının eklenmesi sonucu başarılı sonuçlar elde edilmiştir. Kong vd. [28], çalışmasında elde edilen sonuçlar ile kıyaslandığında dört katmanlı CNN yönteminin GRU algoritması eklenmesi sonucu başarım oranının yüksek çıkarması ve eğitim süresinin daha da kısaltması çalışmanın en büyük avantajı olduğu söylenebilir. Eğitim zamanının düşürerek öğrenim maliyetinin düşürülebildiği gözlemlenmesi çalışmamızda bir başka avantaj olarak görülmektedir. Çalışmanın dezavantajından bahsedilecek olunursa; bazı sınıflar üzerinde (örneğin; *metro*, *metro station*, *tram*) sınıflandırma sırasında birbirine yakın sesler olduğundan ötürü birbiri ile karışması sonucu hatalı sınıflandırma yapılabildiği görülmektedir. Bu sorun ASC problemi için geliştirilen CNN-SPPnet mimarimizde de karşılaştığımızı söyleyebiliriz. İlerleyen çalışmalar için derin sinir ağları kullanımı ile çıkarılmış derin ses özniteliklerin öğrenim mimarisi üzerinde kullanılmasının başarım performansına etkisinin incelenmesi düşünülmektedir. Ayrıca gelecek çalışma planı olarak, CNN'nin veri birleştirme katmanında kullandığımız SPP yönteminde farklı piramit seviyeleri kullanarak analiz edilmesi düşünülmektedir.

Son geliştirdiğimiz CNN4-GRU modelimiz temel sistemden [70] %10 civarı daha yüksek, Kong vd. [28] çalışmasında kullandığı CNN8 yöntemine göre ortalama %2 daha yüksek başarı elde etmektedir. Ayrıca önceki çalışmamız CNNSPPnet mimarimizden %9 civarında yüksek başarı elde edildiği görülmektedir (Çizelge 6.1). Bu sonuçlar üzerine CNN4-GRU mimarimizi ses veri geri getirimi uygulamasının işitsel benzerlik yöntemi için kullanılmasına karar verilmiştir. Anlamsal benzerlik yönteminin geliştirilme yapılması gelecek çalışmalarda sistemin daha da geliştirilmesi için bir yol açacağı ön görülmektedir.

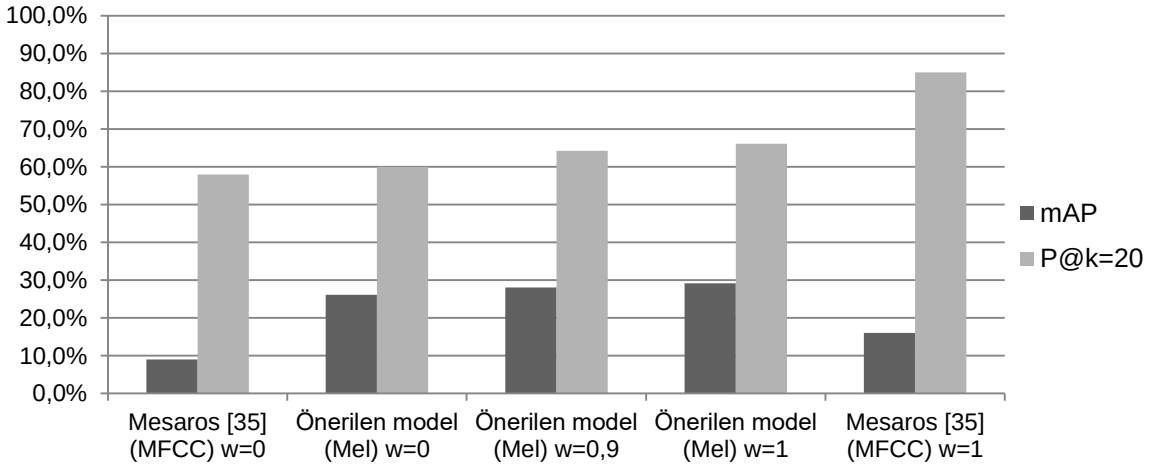
Çizelge 6.1. Önerilen yöntemlerin ve karşılaştırılan çalışmaların doğruluk sonuçları grafiği.



Ses olayında akustik sahneleri geri getirimi problemi kapsamında öznitelikler üzerinden benzerlik hesaplama, CNN4-GRU mimarimizin çıktı katmanı sonucu tahmin değerleri ile benzerlik hesaplama ve ayrıca CNN4-GRU mimarimizin çıktı katmanı öncesi elde edilen son GRU katmanı çıktısı değerleri ile benzerlik hesaplama deneyleri yapılmıştır. Mimarimizin çıktı katman sonucu tahmin değerleri ile yapılan deney sonuçlarında etiket bazlı aramaya yakın performans göstermesi avantaj olarak gösterilmektedir. Sınıflandırma eğitimi sırasında yaşanan benzer sınıfları karıştırabilme sorunu burada da devam etmiştir. Çalışmamızda ek olarak etiket benzerliğinden anlamsal benzerlik hesaplamalarını anlamsal ağırlık değeri ile eklenerek çalışma genişletilmiştir. Burada sonuçları irdelediğimizde, sadece akustik benzerlik sonucu ile sadece anlamsal benzerlik sonucu arasında %6 civarı bir fark olduğu gözlemlenmesi, akustik benzerlik sonucunun; sadece etiketin anlamsal benzerlik sonucu ile benzerlik kurulmasına yakın bir sistem olduğunu

göstermektedir (Çizelge 5.1). Ayrıca Mesaros vd. [35] geliştirdiği yöntem ile kıyasladığımızda mAP skorunda iyileştirme yapılması çalışmanın avantajı olarak görülebilir. Çalışmada anlamsal ağırlık değerinin 0,9 değerine kadar Mesaros vd. [35] geliştirdiği yöntem sonucunu geçebildiği; anlamsal ağırlık değeri 1 verildiğinde ise geçemediği görülmektedir (Çizelge 6.2). Anlamsal ağırlık değerinin 1 olarak verilmesi sonucu Mesaros vd. [35] geliştirdiği yöntem %85 civarında P@k sonucu elde etmesi ve bizim geliştirdiğimiz yöntemin bunun üzerinde bir iyileştirme sağlayamaması dezavantaj olarak görülmektedir. Anlamsal benzerlik yönteminin geliştirilme yapılması gelecek çalışmalarda sistemin daha da geliştirilmesi için bir yol açacağı ön görülmektedir.

Çizelge 6.2. Önerilen geri getirim modeli ve Mesaros [35] çalışmasının P@k=20 ve mAP yüzdelik sonuç grafiği.



KAYNAKLAR LİSTESİ

[1] BUGALHO, M., Portelo, J., Trancoso, I., Pellegrini, T., Abad, A., “Detection Audio Events For Semantic Video Search”, in Interspeech, pp. 1151-1154, 2009

[2] ERONEN, A. J., Peltonen, V. T., Tuomi, J. T., Klapuri, A. P., Fagerlund, S., Sorsa, T., Lorho, G., Huopaniemi, J. Audio-based Context Recognition, IEEE Transactions on Audio Speech and Language Processing, vol. 14, no. 1, pp. 321-329, Ocak, 2006.

[3] ALIAS, F., Socoro, J. C., and Sevillano, X., A Review of Physical and Perceptual Feature Extraction Techniques for Speech, Music and Enviromental Sounds, in Applied Science, vol. 6, no. 5, pp. 143, 2016.

[4] HEITTOLA, T., Research Sound Event Detection, http://www.cs.tut.fi/~heittolt/research-sound-event-detection. [Erişim: 11/08/2019].

[5] LECUN, Y., Bengio, Y., Hinton, G., "Deep learning", Nature, vol. 521, no. 7553, pp. 436-444, 2016.

[6] SEN, D., Sert, M., “Continuous valence prediction using recurrent neural networks with facial expressions and EEG signals”, 2018 26th Signal Processing and Communications Applications Conference (SIU), 2018.

[7] KRİZHEVSKY, A., Sutskever, I. and Hinton, G.E., Imagenet classification with deep convolutional neural networks, Advances in Neural Information Processing Systems, s.1097–1105, 2012.

[8] SİMONYAN, K. and Zisserman, A., Very Deep Convolutional Networks for Large-Scale Image Recognition, Computing Research Repository (CoRR), arXiv 1409.1556, 2014.

[9] HE, K., Zhang, X., Ren, S., and Sun, J., Deep residual learning for image recognition, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), s.770-778, 2016.

[10] HEITTOLA, T., Mesaros, A., Virtanen, T., TUT Sound Event 2017 dataset, https://zenodo.org/record/400516#.XQF-qhYzbcs. [Erişim: 11/08/2019].

[11] HEITTOLA, T., Mesaros, A., Virtanen, T., TUT Urban Acoustic Scenes 2018 Development dataset, Available: https://zenodo.org/record/1228142. [Erişim: 11/08/2019].

[12] HEITTOLA, T., Mesaros, A., Virtanen, T., TAU Urban Acoustic Scenes 2019 Development dataset, https://zenodo.org/record/2589280. [Erişim: 11/08/2019].

[13] PICZAK, K.J., "Enviromental sound classification with convo-lutional neural networks", IEEE International Workshop on Machine Learning for Signal Processing, 2015.

[14] GORIN A., Makhazhanov N., Slunyrev N., "DCASE 2016 Sound Event Detection System Based On Convolutional Neural Network". Detection and Classification of Acoustic Scenes and Events (DCASE), 2016.

[15] SCHRÖDER J., Anemüller J., Goetze S., "Performance Comparison of GMM, HMM and DNN Based Approaches For Acustic Event Detection Within Task 3 Of The DCASE 2016 Challenge". Detection and Classification of Acoustic Scenes and Events (DCASE), 2016.

[16] ADAVANNE S., Parascandolo G., Pertila P., Heittola T., Virtanen T., "Sound Event Detection in Multichannel Audio Using Spatial and Harmonic Features". Detection and Classification of Acoustic Scenes and Events (DCASE), 2016.

[17] Lİ, Y., Li, X., Zhang, Y., Wang, W., Liu, M., Feng, X., "Acoustic scene classification using deep audio feature and BLSTM network", 2018 International Conference on Audio, Language and Image Processing (ICALIP), Shanghai, China, 2018.

[18] ZHOU, J., "Sound Event Detection in Multichannel Audio LSTM Network". Detection and Classification of Acoustic Scenes and Events (DCASE), 2017.

[19] ADAVANNE, S,. Drossos K., Çakır E., Virtanen T., "Stacked Convolutional and Recurrent Neural Networks For Bird Audio Detection". European Signal Processing Conference, 2017.

[20] ÇAKIR, E., Parascandolo G., Heittola T., Huttunen H., Virtanen T., "Convolutional Recurrent Neural Netwrks for Polyphonic Sound Event Detection". IEEE Transactions on Audio, Speech and Language Processing, Special Issue on Sound Scene and Event Analysis, 2017.

[21] HAN, Y., Park, J., Lee, K., “Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification”, Detection and Classification of Acoustic Scenes and Events (DCASE), 2017.

[22] ADAVANNE, S, Virtanen T., “A Report on Sound Event Detection with Different Binaural Features”. Detection and Classification of Acoustic Scenes and Events (DCASE), 2017.

[23] BAE, S. H., Choi, I., Kim, N. S., “Acoustic scene classification using parallel combination of LSTM and CNN”, IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE), 2016.

[24] VALENTİ, M., Diment, A., Parascandolo, G., “DCASE 2016 acoustic scene classification using convolutional neural networks”, Workshop on Detection and Classification of Acous-tic Scenes and Events (DCASE), Budapest, Hungary, 2016.

[25] WEI, D., Li, J., Pham, P., “Acoustic scene recognition with deep neural networks”, Detection and Classification of Acoustic Scenes and Events (DCASE), Budapest, Hungary, 2016.

[26] KUKANOV, I., Hautamaki, V., Lee, K. A., "Recurrent neural network and maximal figure of merit for acoustic event detection", Detection and Classification of Acoustic Scenes and Events (DCASE), Munich, 2017.

[27] JALLET, H., Çakır, E., Virtanen, T., "Acoustic scene classification using convolutional recurrent neural networks", Detection and Classification of Acoustic Scenes and Events (DCASE), Munich, 2017

[28] KONG, Q., Turab, I., Yong, X., Wang, W., Plumbley M. D., "DCASE 2018 Challenge Surrey Cross-Task Convolutional Neural Network Baseline", Detection and Classification of Acoustic Scenes and Events (DCASE), 2018.

[29] JACZYŃSKA, M., Bobiński, P., Pietrzak, A., “Music Recognition Algorithms Using Queries by Example”, 2018 Joint Conference – Acoustics, pp. 1-4, Ustka, Poland, 2018.

[30] HOU, J., Xie, L., Fu, Z., “Investigating neural network based query-by-example keyword spotting approach for personalized wake-up word detection in Mandarin Chinese”, 10th In-ternational Symposium on Chinese Spoken Language Processing (ISCSLP), pp. 1-5, Tian-jin, China, 2016.

[31] FEKİ, I., Ammar, A. B., Alimi, A. M., Automatic environmental sound concepts discovery for video retrieval, International Journal of Multimedia Information Retrieval, vol 5, pp. 105-115, 2016.

[32] CARMEL, D., Yeshurun, A., Moshe, Y., "Detection of alarm sounds in noisy environments", 2017 25th European Signal Processing Conference (EUSIPCO), pp. 1839-1843, 2017.

[33] DE OLIVIERA BARRA G., Lux M., Giro-i-Nieto, X., "Large scale content-based video retrieval with LIvRE", 2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI), pp. 1-4, 2016.

[34] MESAROS, A., Heittola, T., Palomäki, K., “Analysis Acoustic-Semantic Relationship for Diversely Annotated Real-World Audio Data”, 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 26-31 Mayıs, Vancouver, BC, Canada, 2013.

[35] MESAROS, A., Heittola, T., Palomäki, K., “Query-by-example Retrieval of Sound Events Using an Integrated Similarity Measure of Content and Label”, 14th International Work-shop on Image Analysis for Multimedia Interactive Services (WIAMIS), pp. 1-4, Paris, France, 2013.

[36] Princeton University, WordNet, http://wordnet.princeton.edu. [Erişim: 11/08/2019].

[37] WANG, C., Santoso, A., Mathulaprangsan, S., Chiang, C., Wu, C., Wang, J., “Recognition and retrieval of sound events using sparse coding convolutional neural network”, 2017 IEEE International Conference on Multimedia and Expo (ICME), pp. 589-594, Hong Kong, China, 2017.

[38] HALLIDAY, D., Resnick, R., Walker, J., Fundamentals of Physics, 10th Edition, Wiley.

[39] KÜÇÜKBAY, S. E., İşitsel sahnelerin tanınması için çevresel ses analizi, M.Sc. thesis, Baskent University, Ankara, Turkey, 2015.

[40] Mel Frequency Cepstral Coefficients & Implementation, http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/. [Erişim: 11/08/2019].

[41] KAYA, T., İnce, M. C., Pencere Fonksiyonu Aileleri ve Uygulama Alanları, Erciyes Üniversitesi Fen Bilimleri Enstitüsü Dergisi, vol. 26, no. 3, pp. 291-306, Eylül, 2010.

[42] FOOTE, J., An Overview of Audio Information Retrieval, *in* Multimedia Systems, vol. 7, no. 1, pp. 2-10, ACM Press/Springer-Verlag, January, 1997.

[43] NEW, T. L., Dat, T. H., Ma, B., “Convolutional Neural Network with Multi-Task Learning Scheme for Acoustic Scene Classification”, Proceedings of APSIPA Annual Summit and Conference, Malaysia, 2017.

[44] FELIPE, G. Z., da Costa, Y. M. e G., Helal, L. G., “Acoustic Scene Classification Using Spectrograms”, 2017 36th International Conference of the Chilean Computer Science Society (SCCC), Arica, Chile, 2017.

[45] OLAH, C., “Understanding LSTM Networks”, http://colah.github.io/posts/2015-08-Understanding-LSTMs, 2019. [Erişim: 11/08/2019].

[46] HOCHREITER, S., Schmidhuber J., Long Short-Term Memory, Neural Computation, vol. 9, no. 8, 735-1780, 1997.

[47] CHO, K., Merriënboer, B., Gulcehre, C., Bougares, F., “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation”, DOI: 10.3115/v1/D14-1179, 2014.

[48] LECUN, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., Jackel, L. D., “Handwritten digit recognition with a backpropagation network”, Advances in Neural Information Processing Systems 2, s.396–404, Morgan-Kaufmann, 1990.

[49] VALENTI, M., Squartini, S., Diment, A., Parascandolo, G., Virtanen, T., “A convolutional neural network approach for acoustic scene classification”, International Joint Conference on Neural Networks (IJCNN), 2017.

[50] EGHBAL-ZADEH, H., Lehner, B., Dorfer, M., Widmer, G., “A hybrid approach using binaural I-vectors and deep convolutional neural networks”, Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE), Budapest, Hungary, 2016.

[51] LAZEBNIK, S., Schmid, C., Ponce, J., “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories”, In: CVPR, 2006.

[52] HE, K., Zhang, X., Ren, S., Sun, J., “Spatial pyramid pooling in deep convolutional networks for visual recognition”, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2015.

[53] YANG, W., Chen, Y., Huang, C., ve Gao, M., Video-based human action recognition using spatial pyramid pooling and 3D densely convolutional networks, Future Internet 2018, vol. 10, no. 12, pp. 115, Aralık, 2018.

[54] BUGALHO, M., Portelo, J., Trancoso, I., Pellegrini, T., Abad, A., "Detecting Audio Events For Semantic Video Search." in Interspeech, 2009, pp. 1151–1154.

[55] ERONEN, A. J., Peltonen, V. T., Tuomi, J. T., Klapuri, A. P., Fagerlund, S., Sorsa, T., Lorho, G.Huopaniemi, J., Audio-based Context Recognition, IEEE Trans. on Audio, Speech, and Language Processing, vol. 14, no. 1, pp. 321–329, Ocak, 2006.

[56] JEONG, H.Y., Lee, S., Han, Y., Lee, K., "Audio Event Detection Using Multiple-Input Convolutional Neural Network", Detection and Classification of Acoustic Scenes and Events (DCASE), 2017.

[57] MESAROS, A., Heittola, T., Diment, A., Elizalde, B., Shah, A., Vicent, E., Raj, B., Virtanen, T. "DCASE 2017 challenge setup: tasks, datasets and baseline system", In Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE), 2017.

[58] WEIPING, Z., Jiantao, Y., Xiaotao, X., Xiangtao, L., Shaohu, P., "Acoustic scene classification using deep convolutional neural network and multiple spectrograms fusion" , Detection and Classification of Acoustic Scenes and Events (DCASE), 2017.

[59] ZHANG, X., Zou, Y., Shi, W., "Dilated convolution neural network with LeakyReLU for environmental sound classification", 22nd International Conference on Digital Signal Processing (DSP), 2017.

[60] MASS, A.L., Hannun, A. Y., Ng, A. Y., "Rectifier nonlinearities improve neural network acoustic models", In ICML, 2010.

[61] XU, B., Wang, N., Chen, T., Li, M., "Empirical evaluation of rectified activations in convolution network", ICML Deep Learning Workshop, Lille, France, 2015.

[62] Tensorflow library, https://www.tensorflow.org. [Erişim: 11/08/2019].

[63] Keras library, https://keras.io. [Erişim: 11/08/2019].

[64] Theano library, https://www.deeplearning.net/software/theano. [Erişim: 11/08/2019].

[65] https://github.com/yhenon/keras-spp. [Erişim: 11/08/2019].

[66] TRUBA, https://www.truba.gov.tr. [Erişim: 11/08/2019].

[67] MESAROS, A., Heittola, T., Virtanen, T., “Metrics for polyphonic sound event detection.” in Applied Sciences, 2016.

[68] BASBUG, Ahmet-M., Sert, M., "Acoustic Scene Classification Using Spatial Pyramid Pooling With Convolutional Neural Networks," The 13th IEEE International Conference on Semantic Computing (ICSC2019), Newport Beach, California, USA, Ocak 30 - Şubat 1, 2019.

[69] KIGMA, D.P., and BA, J. “Adam: A method for stochastic optimization”, In Proceedings of the 3rd International Conference on Learning Representations (ICLR), 2014.

[70] MESAROS, A., Heittola, T., Virtanen, T. “A multi-device dataset for urban acoustic scene classification”, In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018), 9–13 Kasım, 2018.

[71] BASBUG, Ahmet-M., Sert, M., "Analysis of Deep Neural Network Models for Acoustic Scene Classification" IEEE 27th Signal Processing and Communications Applications Conference (SIU 2019), 24-26 Nisan, Sivas, Turkey, 2019.

[72] VIRTANEN, T., Helen, M., "Probabilistic model based similarity measures for audio query-by-example," in Proceedings of WASPAA, 2007.

[73] HELEN, M., Virtanen, T., ”Query by Example of Audio Signals using Euclidean Distance between Gaussian Mixture Models”, 2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '07), 15-20 Nisan, Honolulu, USA, 2007.

[74] Wordnet Interface, http://www.nltk.org/howto/wordnet.html. [Erişim: 11/08/2019].

[75] HERLOCKER, J. L., Konstan, J., Terveen, L. G., Riedl, J. T., Evaluating Collaborative Filtering Recommender Systems, ACM Transactions on Information Systems, vol. 22, no. 1, pp. 5–53, Ocak, 2004.

[76] POTHULA, S., Dhavachelvan, P., Precision at K in Multilingual Information Retrieval, in International Journal of Computer Applications, vol. 24, no. 9, pp. 40-43, Haziran, 2011.