**T.C.**

**BATMAN ÜNİVERSİTESİ**

**LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ**
**BİLGİ TEKNOLOJİLERİ ANABİLİM DALI**

# DERİN ÖĞRENME İLE TÜRKÇE SES İŞARETLERİNDEN RAKAM TANIMA

**YÜKSEK LİSANS**

**Abdullah EROĞLU**

**Danışman**
**Doç. Dr. Yılmaz KAYA**

**Ocak-2024**
**BATMAN**

## TEZ KABUL VE ONAYI

**ABDULLAH EROĞLU** tarafından hazırlanan "**DERİN ÖĞRENME İLE TÜRKÇE SES İŞARETLERİNDEN RAKAM TANIMA"** adlı tez çalışması 17/01/2024 tarihinde aşağıdaki jüri tarafından oy birliği / ~~oy çokluğu~~ ile Batman Üniversitesi Lisansüstü Eğitim Enstitüsü Bilgi Teknolojileri Anabilim Dalı'nda **YÜKSEK LİSANS** olarak kabul edilmiştir.

| **Jüri Üyeleri** | **İmza** |
|---|---|
| **Üye**<br>Prof. Dr. Necmettin SEZGİN | …….. |
| **Danışman**<br>Doç. Dr. Yılmaz KAYA | …….. |
| **Başkan**<br>Doç. Dr. Mehmet Recep MİNAZ | …….. |

Yukarıdaki sonucu onaylarım.

Dr. Öğr. Üyesi Ömer Murat ÖTER
Lisansüstü Eğitim Enstitüsü Müdür V.

## TEZ BİLDİRİMİ

Bu tez içinde sunulan tüm bilgilerin etik ilkeler ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, ayrıca tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade veya bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

## DECLARATION PAGE

I hereby declare that all information presented in this thesis has been obtained and presented within the framework of ethical principles and academic rules, and that the source of any statement or information that does not belong to me has been fully cited in this study prepared in accordance with the thesis writing rules.

İmza

Abdullah EROĞLU

Tarih: 17/01/2024

# ÖZET

## YÜKSEK LİSANS

## DERİN ÖĞRENME İLE TÜRKÇE SES İŞARETLERİNDEN RAKAM TANIMA

**Abdullah EROĞLU**

**Batman Üniversitesi Lisansüstü Eğitim Enstitüsü**
**Bilgi Teknolojileri Anabilim Dalı**

**Danışman: Doç. Dr. Yılmaz KAYA**

**2024, 73 Sayfa**

**Jüri**
**Prof. Dr. Necmettin SEZGİN**
**Doç. Dr. Yılmaz KAYA**
**Doç. Dr. Mehmet Recep MİNAZ**

Günümüzde, teknolojinin hızla ilerlemesiyle birlikte, ses tabanlı tanıma sistemleri birçok alanda önemli bir rol oynamaktadır. Ses, insanlar arasındaki temel iletişim araçlarından biri olmanın ötesinde, otomasyon, güvenlik ve kullanıcı deneyimi gibi birçok uygulama alanında da kritik bir faktördür. Sesin dijital ortamlarda etkili bir şekilde kullanılabilmesi, özellikle konuşma tanıma teknolojilerinin geliştirilmesi ile mümkün olmaktadır. Bu teknolojiler, ses sinyallerini analiz ederek konuşma dilini anlama ve çeşitli görevleri yerine getirme yeteneğine sahiptir. Özellikle dijital rakam sınıflandırma, bu konuşma tanıma teknolojilerinin önemli bir uygulama alanını oluşturmaktadır. Dijital rakam sınıflandırma, ses sinyallerinden elde edilen dijital rakamları doğru bir şekilde tanıyabilen ve ayırt edebilen sistemlerin geliştirilmesini içerir. Bu, telekomünikasyon sistemlerinden sesli komut sistemlerine, konuşma tabanlı güvenlik uygulamalarından çeşitli endüstriyel ve ticari uygulamalara kadar geniş bir yelpazede kullanım potansiyeli sunmaktadır. Bu bağlamda, ses tabanlı dijital rakam sınıflandırmanın önemi, günlük yaşamdan endüstriyel uygulamalara kadar geniş bir etki alanını kapsamaktadır. Bu çalışma, ses sinyalleri üzerinden gerçekleştirilen dijital rakam sınıflandırma işleminde kullanılan farklı makine öğrenimi modellerini değerlendirerek, bu alandaki teknolojik gelişmelere katkıda bulunmayı amaçlamaktadır. Ses sinyalleri üzerinden dijital rakam sınıflandırma için SVM, LSTM ve CNN modelleri değerlendirilmiştir. %80-20 eğitim-test oranında en yüksek başarı oranını %81,94 ile CNN modeli elde etmiştir. CNN modeli, özellikle "Altı (6)" dijiti için %98,2'lik bir başarı oranına ulaşarak dijit bazında yüksek performans sergilemiştir. Diğer dijitler arasında farklı başarı oranları gözlemlenmiş, örneğin "Bir (1)" ve "Dokuz (9)" için yüksek performans sergilenirken, "Üç (3)", "Dört (4)" ve "Sekiz (8)" için daha düşük başarı oranları kaydedilmiştir. Çalışma kapsamında, farklı eğitim-test oranları altında yapılan değerlendirmelerde LSTM modeli %50-50 eğitim-test oranında en yüksek başarıyı göstermiştir. SVM'nin en yüksek başarı oranını ise %80-20 eğitim-test oranında elde ettiği görülmüştür. Ancak genel olarak, derin öğrenme modelleri olan LSTM ve CNN, SVM'ye kıyasla daha yüksek başarı oranları göstermiştir. Bu sayısal sonuçlar, derin öğrenme modellerinin özellikle ses tabanlı dijital tanıma uygulamalarında etkili bir şekilde kullanılabileceğini göstermiştir.

**Anahtar Kelimeler:** CNN, CWT, LSTM, Ses dijit sınıflandırma

# ABSTRACT

## MS THESIS

## DİGİT RECOGNİTİON FROM TURKİSH SOUND SİGNALS WİTH DEEP LEARNİNG

**Abdullah EROĞLU**

**Batman University Graduate Education Institute**
**Department of Information Technologies**

**Advisor: Assoc. Prof. Yılmaz KAYA**

**2024, 73 Pages**

**Jury**
**Advisor Prof. Dr. Necmettin SEZGİN**
**Assoc. Prof. Yılmaz KAYA**
**Assoc. Prof. Mehmet Recep MİNAZ**

In today's rapidly advancing technological landscape, speech-based recognition systems play a crucial role in various fields. Sound, beyond being a fundamental means of communication among individuals, serves as a critical factor in applications such as automation, security, and user experience. The effective utilization of sound in digital environments is made possible, particularly through the development of speech recognition technologies. These technologies have the capability to analyze sound signals, comprehend spoken language, and perform various tasks. Digital digit classification, particularly, constitutes a significant application area for these speech recognition technologies. Digital digit classification involves the development of systems that can accurately recognize and distinguish digital digits obtained from sound signals. This has a wide range of potential applications, from telecommunication systems to voice command systems and from speech-based security applications to various industrial and commercial applications. In this context, the importance of speech-based digital digit classification spans from everyday life to industrial applications. This study aims to contribute to technological advancements in this field by evaluating different machine learning models used in the process of digital digit classification from sound signals. SVM, LSTM, and CNN models were assessed for digital digit classification from sound signals, with the CNN model achieving the highest success rate at 81.94% in the 80-20 training-test ratio. The CNN model demonstrated high performance, particularly achieving a 98.2% success rate for the digit "Six (6)." Different success rates were observed among other digits, with high performance for "One (1)" and "Nine (9)" but lower success rates for "Three (3)," "Four (4)," and "Eight (8)." In the scope of the study, evaluations conducted under different training-test ratios revealed that the LSTM model exhibited the highest success at a 50-50 training-test ratio. SVM achieved its highest success rate at an 80-20 training-test ratio. However, overall, deep learning models, specifically LSTM and CNN, outperformed SVM, indicating that these numerical results highlight the effective use of deep learning models, especially in sound-based digital recognition applications.

**Key Words:** CNN, CWT, LSTM, Sound digit classification

# ÖNSÖZ

Yüksek Lisans tezimin hazırlanması ve oluşturulmasında, değerli bilgilerini benimle paylaşan, bana her zaman destek olan, tecrübe ve tavsiyeleriyle bana yol gösteren, kendisine ne zaman danışsam bana kıymetli zamanını ayırıp sabırla ve ilgiyle yardımcı olan, her sorun yaşadığımda yanına çekinmeden gidebildiğim, güler yüzü ve samimiyetiyle beni destekleyen, gelecekteki mesleki hayatımda da verdiği bilgilerden faydalanacağımı düşündüğüm kıymetli danışmanım Doç. Dr. Yılmaz KAYA' ya en içten saygılarımı ve teşekkürlerimi sunuyorum.

Ayrıca çalışmamda desteklerini hiçbir zaman eksik etmeyen, bana olan güvenlerini hiç kaybetmeyen canım eşim Gamze EROĞLU ve sevgili aileme teşekkür ediyor ve onlara minnettar olduğumu dile getirmek istiyorum.

Abdullah EROĞLU
BATMAN-2024

# İÇİNDEKİLER

## SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış bazı simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

| **Simgeler** | **Açıklama** |
|---|---|
| ADR | Automatic Digit Recognition, Otomatik Rakam Tanıma |
| AI | Artificial Intelligence, Yapay Zeka |
| AM | Attention Mechanism, Dikkat Mekanizması |
| ANN | Artificial Neural Network, Yapay Sinir Ağları |
| ASC | Automatic Sound Classification, Otomatik Ses Sınıflandırma |
| ASR | Automatic Speech Recognition, Otomatik Konuşma Tanıma |
| BRNN | Bidirectional Recurrent Neural Network, Çift Yönlü Tekrarlayan Sinir Ağı |
| BSL | Bhutanese Sign Language, Bhutan İşaret Dili |
| CALL | Computer Assisted Language Learning, Bilgisayar Destekli Dil Öğrenimi |
| CAPT | Computer Assisted Pronunciation Training, Bilgisayar Destekli Telaffuz Eğitimi |
| CNN | Convolutional Neural Networks, Evrişimli Sinir Ağları |
| CRR | Correct Recognition Ratio, Doğru Tanıma Oranı |
| CWT | Continues Wavelet Transform, Sürekli Dalgacık Dönüşümü |
| DBN | Deep Belief Networks, Derin İnanç Ağları |
| DL | Deep Learning, Derin Öğrenme |
| DNN | Deep Neural Networks, Derin Sinir Ağları |
| ELM | Extreme Learning Machines, Aşırı Öğrenme Makineleri |
| FC | Fully Connected Layer, Tam Bağlantılı Katman |
| FNR | False Negative Ratio, Yanlış Negatif oranı |
| GMM | Gaussian Mixture Model, Gauss Karışım Modeli |

| | |
|---|---|
| GRU | Gated Recurrent Units, Geçitli Tekrarlayan Birimler |
| GTCC | Gammatone Cepstral Coefficients, Gammaton Cepstral Katsayıları |
| HCI | Human Computer Interaction, İnsan Bilgisayar Etkileşimi |
| HMI | Human Machine Interface, İnsan – Makine Ara Yüzü |
| HMM | Hidden Markov Model, Gizli Markov Modeli |
| IOT | Internet of Things, Nesnelerin İnterneti |
| KNN | K-Nearest Neighbors, K En Yakın Komşu |
| LSTM | Long Short Term Memory, Uzun Kısa Süreli Bellek |
| MFCC | Mel Frequency Cepstral Coefficients, Mel Frekansı Cepstral Katsayıları |
| ML | Machine Learning, Makine Öğrenimi |
| MLP | Multilayer Perceptron, Çok Katmanlı Algılayıcılar |
| MSE | Mean Squared Error, Ortalama Kare Hatası |
| NLP | Natural Language Processing, Doğal Dil İşleme |
| RC | Reservoir Computing, Rezervuar Hesaplama |
| ResNet | Residual Neural Networks, Artık Sinir Ağları |
| RF | Random Forest, Rastgele Orman |
| RNN | Recurrent Neural Networks, Tekrarlayan Sinir Ağları |
| SL | Sign Language, İşaret Dili |
| SNR | Signal to Noise Ratio, Sinyal Gürültü Oranı |
| SOM | Self Organizing Map, Kendini Organize Eden Harita |
| STFT | Short-Time Fourier Transform, Kısa Zamanlı Fourier Dönüşümü |
| SVM | Support Vector Machine, Destek Vektör Makinesi |
| VQ | Vector Quantization, Vektör Niceleme |
| VUI | Voice User Interface, Kullanıcı Ara Yüzü |
| WTS | Wavelet Time Scattering, Dalgacık Zaman Saçılma |

# TABLO LİSTESİ

# ŞEKİL DİZİNİ

# 1. GİRİŞ

İnsanlar, diğer tüm canlılardan farklı olarak konuşabilme yeteneğine sahiptir. Konuşma, insanlar arasındaki iletişimin en temel yoludur. İnsanlar, düşüncelerini, duygularını, isteklerini ve birçok diğer durumu konuşarak başkalarına iletebilirler. Bununla birlikte, insanlar zekasını kullanarak yenilikler yapabilme gücüne sahiptirler. İnsanlar bunu fark ettikten sonra hem ilgi duydukları ve merak ettikleri alanlarda çalışmaya başlamışlar hem de yaşamlarını daha kolay, daha iyi ve daha rahatlatıcı hale getirecek çözümler üretmişlerdir. Zamanla, insanların teknolojiye olan bağımlılıkları artmış ve her şeyin en kolay çözümünü istemeleriyle birlikte birçok insan konuşma tanıma üzerinde çalışmaya başlamıştır (Chowdhury vd. 2021). Son dönemde bilgi toplumundaki teknolojik ilerlemelerle birlikte, yüksek performanslı kişisel bilgisayarların yaygınlaşması hız kazanmıştır. Bu durum, bilgisayarlarla insanlar arasındaki etkileşimlerin etkin bir şekilde çift yönlü oluşmasına yön vermekle birlikte insanlar ve makineler arasındaki etkileşim sistemlerinin geliştirilmesine de etki etmiştir (Byun ve Lee, 2021). İnsan – makine etkileşimi çalışmaları, teknolojinin ilerlemesiyle birlikte hızla ilerlemektedir. Konuşma, insan hayatında büyük bir öneme sahip olduğundan, kişisel dijital asistanlar, metne dönüştürme modelleri, çağrı merkezi sistemleri ve sensör uygulamaları gibi birçok çalışma, konuşmanın makineler tarafından anlaşılması üzerine gerçekleştirilmiştir (Ozer, 2021). Konuşma işleme teknolojisi, konuşma bozukluklarının teşhis edilmesinde ve ciddiyetinin belirlenmesinde yardımcı olabilir. Ayrıca, konuşma terapisi gibi konuşma bozukluklarına destek sağlamayı hedefleyen uygulamalara da yardımcı olabilir. Sağlık bakımı ve danışmanlık gibi alanlarda, bu tür otomatik sistemlerden en iyi şekilde yararlanmak mümkündür. Konuşma tanıma sistemleri, web filmleri, bilgisayar destekli öğrenme, çağrı merkezi iletişimi, mobil iletişim ve diğer insan – makine etkileşiminin gerektiği alanlarda özellikle faydalıdır. Çünkü bu sistemler kullanıcının duyarlılığına dayanmaktadır (Patel, Patel ve Mankad, 2022).

Konuşma tanıma, bir sistem veya bir bilgisayar uygulamasının konuşma dilindeki tümceleri ve sözcükleri tanıyabilme ve bunları makine tarafından okunabilir bir biçime dönüştürebilmesidir. İnsanların konuşma dili bilgisayar diline benzemez. Bu yüzden insan dilini bilgisayarların anlayabileceği şekilde dönüştürmek gereklidir. Doğal dil işleme süreci bu durumda bize yardımcı olmaktadır. Konuşma tanıma süreci, farklı algoritmalar ve insan işitme sistemi için geliştirilmiş çeşitli dil modelleri kullanılarak gerçekleştirilir

(Sharmin, Rahut ve Huq, 2020). Ses veya konuşmanın otomatik olarak tanınması, bilgisayarın bu alandaki en karmaşık görevlerinden biridir. Konuşma tanıma, sesli komutları kullanarak cihazlarla etkileşime geçmek ve onlara erişim sağlamak, hatta o cihazları kontrol etmek için temel bir amaç taşır. Bu hedef, konuşma tanıma araştırmacılarını farklı çözüm yollarını keşfetmeye yöneltmektedir. Mevcut teknoloji ile yapılan çalışmalar piyasadaki bazı ticari ürünler için yeterli olsa da araştırmacılar sistem davranışlarının insanlar ile karşılaştırılabilir olmasını istemektedir. Sesin tanıması, rakamların tanınması, sesli aramaların yapılması, seslerin kayıt altına alınması ve konuşmanın yazılı bir metne dönüştürülmesi gibi işlemler, konuşma tanıma teknikleri kullanılarak kolayca gerçekleştirilebilmektedir (Eroğlu ve Kaya, 2023).

Konuşma tanıma, akustik sinyalin bir mikrofon aracılığıyla kaydedilerek bir dizi kelimeye dönüştürülmesi işlemidir. Elde edilen kelimeler, veri girişi, doküman hazırlama, komut ve kontrol gibi çeşitli uygulamalarda kullanılabilir. Bu sistem, insanların klavye, akıllı telefon veya tablet yerine tercih edebilecekleri bir iletişim aracıdır. Ayrıca, bedensel engellilerin bu sistem sayesinde kolaylıkla iletişim kurabilmeleri sağlanır. Ses verilerine dayalı yeni nesil otomasyon teknolojisi, bir odadaki ışıkları veya evin sıcaklığını kontrol etmek gibi işlevleri yerine getirebilir. Söz konusu evin sadece mevcut sesi tanıyıp yanıt verebilen basit bir sesli komut güvenliği sağlayan ve kolay erişilebilir elektronik bileşenler ve donanımlar üzerinde konuşma tanıma sistemi geliştirmeyi amaçlamaktadır. Bu sistem, tanımlanan belirli sözlü komutlara yanıt verecek ve ev cihazlarını, lambaları veya fanları kontrol edecektir (Rudresh, vd. 2017). Son zamanlarda elektronik sistemlerin insanların davranışlarını taklit etme işlevlerinde konuşma tanıma önemli bir yer tutmaktadır. Teknolojinin gelişmesi ile birlikte günlük hayatta kullandığımız çoğu araç gereç te gelişmektedir. Özellikle Nesnelerin İnterneti (IoT – Internet of Things) kavramı ile birlikte akıllı cihazlar yaygınlaşmakta ve o cihazları kontrol etme kolaylığı da gelişmektedir. Emlak sektöründeki ev alım satım işlemlerinde akıllı evler kavramı da sektörü çok daha ileri bir boyuta taşımış durumdadır. Akıllı evlerdeki cihazları bilgisayar, tablet, telefon, akıllı saat gibi gelişmiş güncel cihazlar ile kontrol etmenin rahatlığı o evleri tercih etmede önemli bir faktör haline gelmiştir. (Ks, Rudresh ve Shashibhushsan, 2021). Konuşma veya ses tanıma, insanlar ve akıllı cihazlar arasındaki iletişimde en doğal ve sezgisel kullanıcı ara yüzü olarak öne çıkar. Kişiselleştirilmiş ses kontrollü asistanlar, akıllı ev cihazları, yapay zekaya dayalı biyometrik kimlik doğrulama ve nesnelerin interneti altyapısı gibi kullanıcı ara yüzü (VUI – Voice User Interface) olarak dikkat çekmektedir. Ancak,

konuşma tanıma hala sınırlı ses bilgisi ve doğru işleme için uygun algoritmaların düşük tanıma oranından etkilenmektedir. Ayrıca, konuşma tanıma karmaşık süreçleri içerir, bu süreçler büyük verinin derin öğrenmeyle analizi, yoğun eğitim, konuşmacı mesafesi ve arka plan gürültüsü gibi zorlukları içerir. Bu sorunların üstesinden gelmek için karmaşık algoritmalar kullanılsa da konuşma tanıma oranını artırmak için temel araştırmalar bölgesel ses bilgilerine dayanmalıdır (Han, vd. 2018). Son yıllarda, yapay zekanın gelişimiyle birlikte bilgisayarların konuşma tanıma, doğal dil anlama, görüntü işleme gibi alanlardaki araştırmalarına büyük bir ilgi gösterilmektedir. Bu ilgi, teknolojik merakın yanı sıra, görevleri otomatikleştirme ve daha doğal makine ara yüzleri sunma isteği gibi birçok nedene dayanmaktadır. Güvenlik uygulamalarında biyometrik tanımlama sistemleri öne çıkarken, geleneksel kimlik doğrulama mekanizmaları ise geri planda kalmaktadır. Ses, güvenlik uygulamalarında önemli bir biyometrik tanımlama yöntemi olarak kabul edilmektedir ve bu nedenle çok modlu ses tanıma sistemleri geliştirilmektedir. Bu sistemler, ses tanıma sağlamlığını artırmak amacıyla ortaya çıkmaktadır (Padmanabhan ve Premkumar, 2015).

İnsanların doğal bir şekilde iletişim kurduğu yöntem olan konuşma, sinyal işleme alanında en heyecan verici araştırma konularından biridir. Konuşma işlemesi, konuşma sinyallerinin ve bu sinyallerin nasıl işlendiğinin incelenmesini içerir. Sinyaller genellikle dijital bir biçimde temsil edilerek işlenir, bu yüzden konuşma işlemesi, dijital sinyal işlemenin özel bir durumunu ifade eder. Bu disiplin, geniş bir yelpazeyi ve çeşitli teknolojileri ve uygulamaları içeren benzersiz bir alandır. Konuşma Tanıma, konuşma işleme alanındaki en önemli araştırma konularından biridir ve aynı zamanda Otomatik Konuşma Tanıma (ASR – Automatic Speech Recognition) olarak da adlandırılır. Bu süreç, bir bilgisayar programı ve uygulanan bir algoritma aracılığıyla, bir konuşma sinyalini bir kelime dizisine dönüştürme işlemidir (Karpagavalli ve Chandra, 2016). Otomatik konuşma tanıma, bilgisayarlar aracılığıyla çeşitli algoritmalar kullanarak insan müdahalesine ihtiyaç duymadan insan konuşmalarını tanıyan sistemlerdir. Otomatik konuşma tanıma (ASR), modern çağda bilgisayarlarla etkileşim kurmanın en kazançlı yollarından biridir. Ayrıca okuma yazma bilmeyenler için de avantaj sağlar. Farklı alanlarda sinyal işleme teknolojisi önemli ölçüde ilerleme kaydetmiş olmasına rağmen, bilgisayarlar ve elektronik cihazlar, kullanıcılarla etkileşim için fiziksel bir arayüze (HMI- Human Machine Interface) ihtiyaç duyar. Görme, işitme veya fiziksel engeli olan bireylerin bu arayüzden faydalanması oldukça zor olabilir. Engelli bireylerin kendi anadilleriyle konuşarak bu cihazlarla

kolayca iletişim kurmaları, bu soruna bir çözüm olabilir. Konuşma tanıma sistemleri, orta düzeyde karmaşıklıkta işlemler yapabilme konusunda önemli ilerlemeler kaydetmiş gibi görünse de doğal koşullarda yapılan konuşmaları doğru bir şekilde yorumlayabilen sistemlerin tasarlanması ve oluşturulması hala büyük bir zorluk teşkil etmektedir (Jalalvand, vd. 2015). Dijital konuşma, genellikle İnsan bilgisayar etkileşimi (HCI – Human Computer Interaction) alanında en yaygın kabul gören etkileşim yöntemidir. ASR, farklı konuşmacı türleri, konuşma tarzları, çevre, gürültü ve benzeri nedenlerden dolayı belirgin bir şekilde zorlaşır. Bu zorluklara rağmen, konuşma tanıma, televizyonda canlı altyazı, robot komut kontrolü, konuşmadan metne dönüştürme ve fiziksel engelli bireyler için klavye ve fareye alternatif olarak kullanma gibi birçok uygulamada ASR önemli bir süreçtir (Lokesh, vd. 2019). İngilizce dili için ASR üzerinde çalışmalar yaklaşık yarım yüzyıl önce başlamış olmasına rağmen, son yıllarda yerel diller için ASR geliştirme çalışmaları, insanların bilgisayarlarla daha fazla etkileşimde bulunmasının artmasıyla ortaya çıkmıştır. Bu şekilde, ana dillerini konuşan insanlar, kendi bölgesel dillerinde bilgisayarla iletişim kurabilirler. Yerel diller için ASR geliştirme süreci, yeterli kelime dağarcığı, diyalektik çeşitlilik gibi kaynakların eksikliği nedeniyle oldukça zorlu bir görevdir (Zada ve Ullah, 2020). Çeşitli insan – makine ara yüzü (HMI) sistemlerinde otomatik konuşma tanıma yaygın olarak kullanılmaktadır. ASR teknolojisinin gelişimiyle birlikte Siri ve Cortana gibi sesli yardımcıların kalitesini ve iş verimliliğini artırmasını sağlamıştır (Zhou, vd. 2020). Ancak bazı durumlarda ASR'nin çalışması zordur. Örneğin gürültü oranının yüksek olduğu kalabalık ortamlarda, uçak, tren, otobüs gibi seyahat araçlarında ses sinyalleri bu dış gürültüden ciddi oranda etkilenmektedir. Ayrıca konuşma bozuklukları gibi özel durumlarda da ses sinyalleri algılanamayacak kadar zayıflayabilmektedir (Wang, vd. 2020). Bu nedenle, uygulama senaryolarını genişletmek için, sadece sese değil, bu tarz olumsuzlukların da üstesinden gelebilecek bir konuşma tanıma yöntemi geliştirmek gerekmektedir (Tian, vd. 2022).

Sözlü rakam tanıma, ses içeriği analizi, kredi kartı numarası girişi, sesli arama ve veri girişi gibi çeşitli uygulamalar için önemli bir rol oynamaktadır. Ancak, dikkate değer bir şekilde, konuşulan rakam tanıma konusu daha az ilgi görmüştür (Sharan, 2020). Otomatik konuşma rakam tanıma sistemi, mikrofon aracılığıyla kaydedilen akustik verilerin insan müdahalesi olmadan bilgisayar tarafından değerlendirilebilen algoritmaları kullanarak insan konuşmalarından oluşan kelimeler ve cümleleri tanımayı hedefler. Bu sistem,

bir konuşma rakamı örneği alındıktan sonra, veri dizisinin önceden otomatik olarak oluşturulan ve depolanan modellere ait olup olmadığına göre sınıflandırma yapabilen bir tanıma görevinin tasarlanması ve geliştirilmesiyle ilgilenir. Bu sistem, konuşma tarzı, kelime bilgisi, konuşma modeli gibi parametrelerle belirlenir (Ks, Rudresh ve Shashibhushsan, 2021).

Sinyal İşleme ve Yapay Zekâ (AI – Artificial Intelligence) disiplinler arası araştırma alanında, modern konuşma teknolojisi sistemi için çeşitli yöntemler ve algoritmalar ortaya çıkmıştır. Makine Öğrenimi (ML – Machine Learning), doğru görevleri yerine getirmek için uygun özellikleri kullanarak doğru modeller oluşturma konusunda faydalı olabilir. Konuşma teknolojilerinde, Makine Öğrenimi paradigmasıyla ortaya çıkan yeni teknikler büyük bir ilerleme kaydetmiştir. Makine Öğreniminin temel kavramı, veri kümesinden verilen görevi analiz etmek, tespit etmek veya sonuçlandırmak için öğrenmedir (Kanchana, 2022). Makine öğrenimi tabanlı algoritmalar, etkili olabilmelerine rağmen, zaman ve emek gerektiren manuel özellik çıkarma gibi bir dezavantaja sahiptir. Ancak Derin Öğrenme (DL – Deep Learning) tabanlı algoritmaların ortaya çıkmasıyla birlikte, birçok uygulayıcı geleneksel makine öğrenimi yöntemlerinden vazgeçerek, bu daha güçlü tekniklerin ham verilerden otomatik olarak öğrenebilen özelliklerini tercih etmeye başlamıştır. Bu şekilde, Derin Öğrenme tabanlı algoritmalar zaman emek tasarrufu sağladığından daha çok tercih edilmeye başlanmıştır (Chauhan vd. 2020).

Bilgi teknolojisinin gelişmesi ve artan bilgi işleme kapasiteleriyle birlikte, derin öğrenme giderek daha yaygın olarak tanıma görevlerinde kullanılmaktadır. Derin öğrenme, el yazısı karakterlerinin ve konuşmanın tanınması gibi öğrenme ve sınıflandırma süreçlerinde etkili bir şekilde kullanılmaktadır (Fu, vd. 2020). Derin öğrenme, görüntü sınıflandırması, metin sınıflandırması, konuşma tanıma ve birçok başka alanda büyük başarılar elde etmiştir (Khan, vd. 2021). Konuşma tanıma ve çoklu dillere yönelik konuşma taklitleri üzerine birçok çalışma yapılmıştır. Ancak, tüm diller için veri bulmanın zorlu bir süreç olduğu ve veri toplamanın zaman alıcı ve sıkıcı bir iş olduğu bilinmektedir. Veri olmadan, hiçbir dil için araştırma ve geliştirme mümkün değildir. Konuşma tanımanın sağlam ve güvenilir bir şekilde ilerlemesini sürdürebilmesi, araştırmacıların konuşma odaklı sistemleri geliştirmeye devam edebilmeleri için kamuya açık verilere kolayca erişim sağlayabilmeleri gerekmektedir (Chandio vd., 2021).

Derin öğrenme, yapay zekâ için çok katmanlı yapay sinir ağıyla (ANN – Artificial Neural Network) desteklenen, veriye dayalı bir sınıflandırma matematiksel tekniğidir. Dil öğretimi ve eğitimi alanında bilgisayar destekli dil öğrenimi (CALL – Computer Assisted Language Learning) ve bilgisayar destekli telaffuz eğitimi (CAPT – Computer Assisted Pronunciation Training) gibi uygulamalar büyük ilgi görmüştür (Rogerson-Revell, 2021). Dil öğrenme ve öğretme yöntemlerini geliştirmek amacıyla sıklıkla kullanılan CALL ve CAPT sistemleri, derin öğrenme yapılarını kullanarak otomatik konuşma tanıma (ASR) teknolojisiyle konuşmayı rahatlıkla tanıyabilir seviyeye gelebilmektedir (Rukwong ve Pongpinigpinyo, 2022). Otomatik rakam tanıma (ADR – Automatic Digit Recognition), ASR'nin en zorlu alanlarından biri olarak kabul edilir. Otomatik Rakam Tanımanın giderek daha fazla önem kazanmasının temel nedeni, insan – makine etkileşimini ele alan uygulamalara olan talepten kaynaklanmaktadır. Bu uygulamalar genel olarak telaffuz edilen rakamların tanınmasıyla çalışmaktadır. Bu tür sistemlerin uygulanması, konuşma ifadelerini doğru bir şekilde tanımlayabilen güvenilir özellikler sağlamak için konuşma sinyalinin belirli bir işlemden geçirilmesini gerektirir. Literatürde konuşma sinyalini parametrik olarak temsil etmek için çeşitli teknikler önerilmiştir. En yaygın kullanılanı, insan konuşmasının bazı yönlerini taklit etmeye çalışan popüler bir teknik olan Mel Frekansı Cepstral Katsayılarıdır (MFCC – Mel Frequency Cepstral Coefficients) (Zerari, vd. 2018). Yapay Zekâ modellerini verimli bir şekilde eğitmek ve modelin sınıflandırma başarısını artırmak için, ön işleme adımlarından biri olarak MFCC kullanmak, özellik çıkarma, veri sıkıştırma ve alt örnekleme gibi süreçlere katkıda bulunacaktır. Bu yöntem, yapay zekâ modellerinin eğitimini daha etkili hale getirip sınıflandırma performansını artıracaktır (Yadav, vd. 2023).

Son yıllarda, çeşitli denetimli, denetimsiz ve takviyeli öğrenme problemlerini çözmek için derin öğrenmeye dayalı teknikler büyük bir ilgi görmektedir. Evrişimli Sinir Ağları (CNN – Convolutional Neural Networks), bu teknikler arasında en tanınmış ve yaygın olarak kullanılan bir yöntemdir. CNN'ler, girdi verilerinden ilgili özellikleri otomatik olarak çıkarabilen bir yapay sinir ağı türüdür. Özellikle bilgisayarla görü alanında problemleri çözmek için oldukça ilginç bir alternatif sunmaktadır. Aslında, bilgisayarla görü, CNN'lerdeki yeni teknikleri ve yenilikleri doğrulamak için bir test ortamı haline gelmiştir (Baldominos, Saez ve Isasi, 2019). Görüntü sınıflandırması için etkili bir derin öğrenme tekniği olan CNN, görüntü sınıflandırma görevlerinde umut verici sonuçlar ortaya çıkarmıştır. CNN'in gücü, geleneksel özellik çıkarma ve sınıflandırma tekniklerine

kıyasla ses sinyali sınıflandırması gibi görüntü dışı sınıflandırma görevlerinde daha üstün bir performans sergilemesidir. CNN'in ses sinyallerini görüntü olarak en iyi şekilde nasıl temsil edeceği önemli bir konudur. Bu sorunu çözmek için genellikle iki temel yaklaşım benimsenmektedir. İlk yaklaşımda, ses sinyalleri Spektrogram görüntülerine dönüştürülür. İkinci yaklaşımda ise Mel – Filtre bankaları kullanılarak Mel – Spektrogram görüntü temsilleri elde edilir. Spektrogram gösterimleri, eşit bant genişliği ve eşit aralıklı frekans bileşenleri sunar. Ancak bu, spektral enerjinin çoğunun düşük frekans aralığında olduğu görevlerde ideal değildir (Sharan ve Moir, 2019). Mel Filtresinin kullanılması bir dereceye kadar fayda sağlar. Bununla birlikte, çeşitli çalışmaların, bu yöntemin performansının biyolojik olarak esinlenen modellerden daha düşük olduğunu göstermektedir (Sharan, Berkovsky ve Liu, 2020).

Geleneksel otomatik ses sınıflandırma sistemleri, genellikle otomatik konuşma tanıma sistemlerinden ilham alan akustik sinyal ön işleme, özellik çıkarma ve sınıflandırma adımlarını içerir. Yıllar boyunca, akustik sinyallerin frekans içeriğini ve zamansal yapısını yakalamak için birkaç öznitelik çıkarma tekniği önerilmiştir. Mel Frekans Cepstral Katsayıları (MFCC) ve Gammaton Cepstral Katsayıları (GTCC – Gammatone Cepstral Coefficients), en yaygın olarak kullanılan öznitelik çıkarma teknikleri arasında yer almaktadır. İnsan işitsel sistemini taklit etmeleri nedeniyle, özellikle düşük frekanslı bileşenlerdeki değişikliklere daha duyarlıdırlar. Bu çerçeve tabanlı özellikler daha sonra Gauss Karışım Modeli (GMM – Gaussian Mixture Model), Gizli Markov Modeli (HMM – Hidden Markov Model), Destek Vektör Makinesi (SVM – Support Vector Machine) veya derin öğrenme modellerini eğitmek için kullanılırlar. Son yıllarda, derin öğrenme modellerinin ve ulaşılması kolay eğitim verilerinin yönlendirdiği önemli performans artışına rağmen, konuşma tabanlı otomatik ses sınıflandırma (ASC – Automatic Sound Classification) sistemlerinin mobil ve giyilebilir cihazlarda geniş çapta benimsenmesini engelleyen iki büyük zorluk bulunmaktadır. Genellikle bu tür cihazlarda yüksek güç tüketimi gerektiren yüksek performanslı bilgi işlemin bu cihazlarda sınırlı olması ve arka plan gürültüsüdür. Artan arka plan gürültüsüyle birlikte MFCC veya GTCC gibi özelliklere sahip son teknoloji GMM – HMM ve derin öğrenme modellerinin performansı, önemli ölçüde düşmektedir (Wu, vd. 2018). Yıllar süren araştırmalar boyunca Çok Katmanlı Algılayıcılar (MLP – Multilayer Perceptron), Aşırı Öğrenme Makineleri (ELM – Extreme Learning Machines), Evrişimli Sinir Ağları (CNN), Artık Sinir Ağları (ResNet – Residual Neural Networks) ve Tekrarlayan Sinir Ağları (RNN – Recurrent Neural Networks) ile

birlikte çeşitli yapay sinir ağı mimarileri de kullanılmıştır (Koolagudi, Murthy ve Bhaskar, 2018). Özellikle, Uzun Kısa Süreli Bellek (LSTM – Long Short Term Memory) ve Geçitli Tekrarlayan Birimler (GRU – Gated Recurrent Units), konuşma sinyali modellemesinde yaygın olarak kullanılmaktadır. Buna ek olarak hem özelliklerin çıkarılmasını hem de sınıflandırmayı birlikte öğrenmek için çeşitli uçtan uca mimariler de kullanılmaktadır. LSTM ve GRU ağlarına ek olarak, derin öğrenmede Dikkat Mekanizmasının (AM – Attention Mechanism) tanıtılması, sıralı veri işleme alanında bir dönüm noktası olarak kabul edilebilir. Dikkat Mekanizmasının amacı, insan görsel dikkati gibi, ilgili bilgileri seçmek ve ilgisiz olanları filtrelemektir. İlk olarak makine çevirisi görevi için tanıtılan Dikkat Mekanizması, sinir ağlarının temel bir bileşeni haline gelmiştir. Dikkat mekanizmasının kodlayıcı – kod çözücü tabanlı sinir ağlarına entegre edilmesi, makine çevirisinin performansını uzun diziler için bile önemli ölçüde artırmıştır (Luong, Pham ve Manning, 2015). Dikkat mekanizmasının makine çevirisi üzerindeki başarısı, birçok araştırmacının onu Doğal Dil İşleme (NLP – Natural Language Processing) ve konuşma işleme gibi çeşitli uygulamalarda sinir ağlarının temel bir bileşeni olarak görmesine yol açmıştır.

Derin Öğrenme, yapay sinir ağlarını kullanarak yüksek boyutlu verilere hizmet ettiği için, konuşma tanıma, temel kelime tanıma gibi uygulamalardan otomatik konuşma tanıma kullanılarak Derin Sinir Ağları (DNN – Deep Neural Networks) çağında önemli bir ilerleme kaydetmektedir. Derin Öğrenme, iletişim modeli ve bilgi işleme gibi görevleri etkileyici bir şekilde yerine getirebilir (Halageri, vd. 2015). Derin Sinir Ağı (DNN), farklı veri kümesi türlerine bağlı olarak sonucun doğruluğunu ve hızını etkileyen birçok varyasyona sahiptir. Tekrarlayan Sinir Ağı (RNN) ve Evrişimsel Sinir Ağı (CNN) gibi belirli uygulamaları gerçekleştirmek için farklı mimariler geliştirilmiştir. RNN, bir katmanın çıktısını girdiye veya bir önceki katmana geri beslemek için kaydetme ilkesine dayalı olarak çalışır. Bu sayede, ağın ilk katmanında ağırlık toplamı ve özelliklerin çarpımı ile ileri beslemeli ağlar şeklinde oluşturulan katman sonucunun tahmin edilmesine yardımcı olunur. Uygulanan RNN işlemi, her nöronun bir önceki zaman adımında sahip olduğu bilgiyi hatırlayacak şekilde uygulanır. Bu özelliği sayesinde RNN, hesaplamaları gerçekleştirirken her nöronun bir hafıza hücresi gibi işlev gördüğünü ortaya koyar (Wazir ve Chuah, 2019). Modern toplumun, çağımızın gerektirdiği modern çözümlere ihtiyacı vardır. Rastgele tahminlerle sınırlı olmadığımız, modern bir çağda yaşamaktayız. Artık tahminlerin geçerlilik, doğruluk ve sıralaması önemlidir. Uzun Kısa Süreli Bellek (LSTM) gibi tekrarlayan sinir ağı mimarileri, geri besleme bağlantılarıyla veri dizileri

üzerinde çalışabildikleri için tahminlerde önemli bir rol oynamaktadır. LSTM, zaman serisi tahmini, konuşma tanıma, sonraki kelime veya cümle tahmini, metin tahmini, el yazısı tahmini ve daha birçok alanda kullanılmaktadır (Chowdhury vd. 2021). Konuşma sinyali işleme, derin öğrenme alanında devrim niteliğinde bir etki yaratmıştır. Son zamanlarda, birçok araştırmacı, Derin İnanç Ağları (DBN – Deep Belief Networks), Evrişimli Sinir Ağları (CNN) ve Uzun Kısa Süreli Bellek (LSTM) gibi yöntemleri kullanarak belirli uygulamalarda olağanüstü sonuçlar elde etmektedir. Her geçen gün daha fazla araştırmacı, bu tekniklerin gücünü keşfetmektedir (Zhao, Mao ve Chen, 2019). Derin sinir ağları (DNN), genellikle "Kara Kutu" yaklaşımları olarak kabul edilir, çünkü nasıl çalıştıklarını anlamak son derece zor olabilir. Bu nedenle, bu sorunu incelemek için iki model veya yöntem önerilmiştir. Genellikle istatistikçiler tarafından kullanılan "Veri Modeli" yaklaşımı ile "Algoritmik Model" yaklaşımı karşılaştırıldığında Algoritmik Modelin tahmin yapmak için bir algoritma bulmaya odaklandığı görülmektedir. (Lipton, 2018). Derin Sinir Ağları (DNN) tarafından öğrenilen yüksek düzeyde soyut özelliklerin yorumlana bilirliği zayıftır. Bununla birlikte, Derin Sinir Ağları, bazı deneylerde geleneksel yaklaşımlardan önemli ölçüde daha iyi performans gösterdiği gözlemlenmedir (Zheng, Yu ve Zou, 2015).

### 1.1.Amaç

Günümüzde, teknolojinin hızla ilerlemesiyle birlikte, ses tabanlı tanıma sistemleri birçok alanda önemli bir rol oynamaktadır. Ses, insanlar arasındaki temel iletişim araçlarından biri olmanın ötesinde, otomasyon, güvenlik ve kullanıcı deneyimi gibi birçok uygulama alanında da kritik bir faktördür. Dijital rakam sınıflandırma, ses sinyallerinden elde edilen dijital rakamları doğru bir şekilde tanıyabilen ve ayırt edebilen sistemlerin geliştirilmesini içerir. Bu, telekomünikasyon sistemlerinden sesli komut sistemlerine, konuşma tabanlı güvenlik uygulamalarından çeşitli endüstriyel ve ticari uygulamalara kadar geniş bir yelpazede kullanım potansiyeli sunmaktadır. Bu bağlamda, ses tabanlı dijital rakam sınıflandırmanın önemi, günlük yaşamdan endüstriyel uygulamalara kadar geniş bir etki alanını kapsamaktadır. Bu çalışma, ses sinyalleri üzerinden gerçekleştirilen dijital rakam sınıflandırma işleminde kullanılan farklı makine öğrenimi modellerini değerlendirerek, bu alandaki teknolojik gelişmelere katkıda bulunmayı amaçlamaktadır. Ses sinyalleri üzerinden dijital rakam sınıflandırma için SVM, LSTM ve CNN modelleri değerlendirilmiştir.

## 1.2.Kapsam

Otomatik rakam tanıma giderek daha fazla önem kazanmaktadır. Bunun temel nedeni, insan – makine etkileşimini ele alan uygulamalara olan talepten kaynaklanmaktadır. Bu uygulamalar genel olarak telaffuz edilen rakamların tanınmasıyla çalışmaktadır. Bu tür sistemlerin uygulanması, konuşma ifadelerini doğru bir şekilde tanımlayabilen güvenilir özellikler sağlamak için konuşma sinyalinin belirli bir işlemden geçirilmesini gerektirir. Ancak, konuşma tanıma hala sınırlı ses bilgisi ve doğru işleme için uygun algoritmaların düşük tanıma oranından etkilenmektedir. Ayrıca, konuşma tanıma karmaşık süreçleri içerir, bu süreçler büyük verinin derin öğrenmeyle analizi, yoğun eğitim, konuşmacı mesafesi ve arka plan gürültüsü gibi zorlukları içerir. Sesin dijital ortamlarda etkili bir şekilde kullanılabilmesi, özellikle konuşma tanıma teknolojilerinin geliştirilmesi ile mümkün olmaktadır. Bu teknolojiler, ses sinyallerini analiz ederek konuşma dilini anlama ve çeşitli görevleri yerine getirme yeteneğine sahiptir. Özellikle dijital rakam sınıflandırma, bu konuşma tanıma teknolojilerinin önemli bir uygulama alanını oluşturmaktadır. Günümüz toplumu, çağın dinamiklerine uygun modern çözümlere ihtiyaç duymaktadır. Rastgele tahminlerin ötesine geçtiğimiz bu çağda, geçerlilik, doğruluk, sıralama ve sınıflandırmanın önemi göze çarpmaktadır. CNN, görüntü sınıflandırma görevlerinde umut verici sonuçlar ortaya çıkarmıştır. Ayrıca CNN geleneksel özellik çıkarma ve sınıflandırma tekniklerine kıyasla ses sinyali sınıflandırması gibi görüntü dışı sınıflandırma görevlerinde de üstün bir performans göstermektedir LSTM gibi tekrarlayan sinir ağı mimarileri de geri besleme bağlantılarıyla veri dizileri üzerinde çalışabildikleri için tahminlerde önemli bir rol oynamaktadır. Bu çalışma kapsamında, dijital rakam sınıflandırma için farklı derin öğrenme modelleri incelenmiş, özellikle CNN ve LSTM'nin incelenen diğer modellere kıyasla daha yüksek başarı oranları gösterdiği gözlemlenmiştir. Benzer çalışmaların detaylı bir şekilde gözden geçirilerek hazırlanan bu çalışmanın, bu alandaki literatüre katkı sağlayacağı düşünülmektedir.

## 2. KAYNAK ARAŞTIRMASI

Bu bölümde, dijital rakam sınıflandırması için yapılan çeşitli çalışmaların özetleri sunulmuştur. Bu çalışmalar genellikle iki temel adımdan oluşan bir yaklaşımı benimsemektedir. İlk adımda, rakam sözcüklerinden öznitelik çıkarımı gerçekleştirilirken, ikinci adımda ise çıkarılan öznitelikler kullanılarak makine öğrenimi ve derin öğrenme yöntemleriyle sınıflandırma işlemi gerçekleştirilmektedir.

Konuşma Tanıma, Doğal Dil İşleme alanında gelişen ve insan – makine etkileşimini kolaylaştırmaya yönelik bir teknolojidir. Bu dinamik alan, veri kümesi boyutu, gürültülü ortamlar, yaş, cinsiyet, lehçe, konuşmacının zihinsel ve psikolojik durumu gibi birçok zorluk içerir. Rakamların konuşma tanımasında kullanımı, cep telefonu numaraları, puanlar, hesap numaraları, kayıt kodları, sosyal güvenlik kodları gibi sayısal bilgilerin iletişiminde büyük kolaylık sağlar. Bu yöntem, insanların sayılarla ilişkili işlemleri konuşarak pratik bir şekilde gerçekleştirmesine olanak tanır. Gujarati dili, Hindistan'ın Gucerat bölgesinde yoğun olarak konuşulan bir Hint – Aryan dilidir. Hindistan'da resmi olarak tanınan 22 dilden biri olan Gujarati, en yaygın 6. dil olarak kabul edilir ve yaklaşık 55,5 milyon insan tarafından konuşulmaktadır. Tailor, vd. (2022) yaptıkları bu çalışmada, derin öğrenme yöntemini kullanarak 0'dan 9'a kadar olan 10 Gujarati rakamının tanınmasını amaçlamaktadır. Veri setlerini oluşturmak için yaşları 20 ila 40 arasında değişen 4 erkek ve 4 kadın anadil konuşmacısından çeşitli ses kayıtları alarak toplamda 2400 rakam kelimesi elde etmişlerdir. Oluşturulan bu veri seti ile Gujarati dilinde rakam tanıma işlemini gerçekleştirmek için Mel Frekans Cepstrum Katsayıları (MFCC) tekniği ile özellikler çıkarıldıktan sonra, derin öğrenme yaklaşımlarından Evrişimsel Sinir Ağını (CNN) kullanmışlardır. Önerilen yaklaşım için üç farklı veri kümesi boyutunda (1200, 1800 ve 2400) üç ayrı deney yapılmıştır. İlk deneyde 1200 veri kümesi ile Doğruluk (Accuracy) %58, Kesinlik (Precision) %58 ve Duyarlılık (Recall) %58,6 sonuçları elde edilmiştir. İkinci deneyde, 1800 veri kümesi ile Doğruluk %78, Kesinlik %78 ve Duyarlılık %78 sonuçları elde edilmiştir. Üçüncü ve en son deneyde ise 2400 veri kümesi kullanıldığında ise Doğruluk %98,7, Kesinlik %98 ve Duyarlılık %98 şeklinde oluşmuştur. 0'dan 9'a kadar tüm rakamlar için farklı veri kümesi boyutlarıyla yapılan üç deneyden en yüksek tanıma doğruluk oranı 2400 veri kümesi kullanılarak %98,7 olarak ölçülmüştür. Bu sonuçlara dayanarak, konuşma tanıma modellerinin doğruluğunun veri kümesinin boyutunu artırarak geliştirilebileceği gözlemlenmiştir. Büyük bir veri kümesiyle eğitilen CNN tabanlı derin öğrenme yaklaşımının, daha iyi performans sergilediği fark edilmektedir. Yazarlar,

gelecekte önerilen modelin, konuşma tanıma sonuçlarını etkileyen ana parametreler olan konuşmacının yaşı, cinsiyeti ve lehçesi gibi faktörleri işin içine daha fazla katarak önerdikleri bu modelin geliştirilebileceğini belirtmişlerdir (Tailor, vd. 2022).

Dünya üzerinde konuşulan farklı diller için veri setlerinin gelişimi ve yapılan çalışmalar konuşma tanıma sistemleri tasarlama ve geliştirmeye yönelik araştırma faaliyetlerini daha cazip hale getirmiştir. Hindistan ve Afganistan gibi birçok Güney Asya ülkesinde yaygın olarak konuşulan Urduca, Pakistan'ın ulusal dilidir. Chandio, vd. (2021) Urduca rakamlardan oluşan çok kapsamlı bir veri seti hazırlayıp farklı sınıflandırma algoritmaları kullanarak bir çalışma gerçekleştirmişleridir. Hazırlamış oldukları veri seti çeşitli özellik ve farklılıkları barındıran 740 katılımcı ile 25.518 ses örneğinden oluşan en büyük halka açık Audio Urdu Digits veri seti özelliğini taşımaktadır. Hazırladıkları veri setini test etmek için, Destek Vektör Makineleri (SVM), Çok Katmanlı Algılayıcılar (MLP) ve EfficientNet gibi farklı temel sınıflandırma algoritmalarını uygulamışlardır. Yazarlar çalışmalarında bu temel algoritmalar dışında ses basamak sınıflandırması için ayrıca Evrişimsel Sinir Ağını da (CNN) kullanmışlardır. Yapmış oldukları çalışma neticesinde CNN'nin daha verimli olduğunu, sınıflandırma doğruluğu bakımından CNN'nin temel algoritmalardan daha iyi performans gösterdiklerini belirtmişlerdir. Yazarlar dalga formu (Spektogram) biçiminde kapsamlı veri analizi ve Mel Spektogramı hesaplamaları sonucunda parametre sayısını azaltmak ve aşırı öğrenmenin (Overfitting) önüne geçmek için Max Pooling kullanıp filtre sayısını artırarak CNN'nin sonuçları daha da iyileştirilebileceğini belirtmişlerdir. Ayrıca, önerdikleri CNN'nin etkinliğini değerlendirmek için İngilizce rakamlardan ve Hindistan dillerinden olan Guceratça (Gujarati) rakamlarından oluşan iki farklı veri seti ile test edip ve umut verici sonuçlar elde ettiklerini belirtmişlerdir (Chandio vd. 2021).

Teknolojinin çok hızlı ilerlemesi ve teknolojik araçların gün geçtikçe hayatımızda daha çok yer edinmesinden dolayı konuşma tanıma giderek daha kullanışlı hale gelmektedir. Bengalce Bangladeş'in resmi dilidir ve Hindistan'ın ikinci en çok konuşulan dilidir. Bengalce konuşma tanıma birçok sektörde kullanımı yaygınlaşmaktadır. Sharmin, Rahut ve Huq, (2020) Bengalce konuşulan rakamların sınıflandırılması ile ilgili bir çalışma yapmışlardır. Bengalce rakam tanıma üzerine daha önce çok sayıda çalışma yapılmış ancak bu çalışmalar lehçe, cinsiyet, yaş gibi birçok parametreyi göz ardı etmişlerdir. İnsanların

sesi cinsiyete, yaşa, lehçelere göre farklılık göstermektedir. Yazarlar çalışmalarında Evrişimsel Sinir Ağı (CNN) ile sınıflandırmaya yönelik yeni bir yaklaşım üzerine çalışmışlardır. Önerdikleri modeli cinsiyet, yaş ve lehçeleri farklı on kişinin Bengalce konuştukları rakamlar ile test etmişlerdir. Çalışmalarında %98,37'lik bir doğrulukla sınıflandırma gerçekleştirip önerdikleri yaklaşımlarının geçerliliğini ispat etmişlerdir (Sharmin, Rahut ve Huq, 2020).

Ks, Rudresh ve Shashibhushsan, (2021) uygulamaları kontrol etmede sözlü olarak söylenen rakamların otomatik konuşma tanıma ile doğrulamanın önemini deneysel olarak göstermeyi amaçlamışlardır. Çalışmalarında örnek bir konuşma rakamı ifade sinyalini alıp bunu bir veri dizisi haline getirdikten sonra daha önce otomatik olarak oluşturulan modellere ait olup olmadığına göre sınıflandırmaya çalışmışlardır. Çalışmalarında Vektör Niceleme (VQ – Vector Quantization) sınıflandırma tekniğini kullanarak Cepstral öznitelik çıkarma ve Mel Frekansı Cepstral Katsayısı (MFCC) öznitelik çıkarma tekniklerini karşılaştırmalı olarak analiz etmişlerdir. 15 kişiden üç farklı oturumda sıfırdan dokuza kadar (0-9) rakamlar sözlü bir şekilde kayıt altına alıp ve bu rakam sözcüklerine hem Cepstral hem de MFCC öznitelik çıkarma tekniklerini uygulamışlardır. Daha sonra iki farklı oturumda hem Cepstral hem de MFCC özniteliklerini Vektör Niceleme sınıflandırma tekniği kullanılarak sonuçlar ayrı ayrı kaydedilmiştir. Cepstral – VQ sınıflandırma performansı ile MFCC – VQ sınıflandırma performansı karşılaştırılmıştır. MFCC – VQ sınıflandırma %93.333 ortalama performans sergilerken Cepstral – VQ sınıflandırma 86.667 ortalama performans sergilediği görülmüştür. Yazarlar çalışmalarında MFCC ile öznitelik çıkarma işleminin Vektör Niceleme sınıflandırmasında daha iyi performans gösterdiği sonucuna varmışlardır (Ks, Rudresh ve Shashibhushsan, 2021).

Peştuca ve Darice Afganistan'ın iki resmi dilidir. Bu dillerden Peştuce, Afganistan'da ve Pakistan'ın batı kesiminde yaşayan Peştunların konuştuğu dildir. Zada ve Ullah, (2020) yapmış oldukları çalışmalarında Evrişimsel Sinir Ağları (CNN) kullanarak Peştuca dili otomatik konuşma tanıma sisteminin performansını geliştirmek için izole edilmiş bir rakam tanıma sistemini geliştirmeye çalışmışlardır. 0'dan 9'a kadar olan rakamların her biri için 50 tane sözcük içeren bir veri seti kullanılmıştır. Mel Frekans Kepstral Katsayıları (MFCC) öznitelik çıkarıma algoritması tarafından çıkarılan 20 katsayı ile CNN eğitilip test edilmiştir. Çalışmada önerilen CNN sisteminde 4 evrişim katmanı (Convolutional Layer), bu katmanlar arasına Maximum Havuzlama (Max – Pooling) katmanı ve

ReLU (Rectifier) fonksiyonundan oluşmaktadır. Önerilen CNN sistemi veri setinin %50'si üzerinden eğitilip veri setinin geri kalan kısmı test için kullanılmıştır. Önerilen CNN sistemi daha önce yapılmış mevcut benzer çalışmalara kıyasla %7,32 daha iyi performans göstermiş ve ortalama %84,17 doğruluk elde edilmiştir (Zada ve Ullah, 2020).

Son yıllarda, konuşma tanıma alanında önemli bir ilerleme kaydedilmiştir. Derin Sinir Ağı (DNN), optimizasyon problemleri için hata oranını minimize etme yeteneği sayesinde, konuşma analizi için en popüler yöntemlerden biri haline gelmiştir. Wazir ve Chuah, (2019) çalışmalarında Arapça rakam konuşma tanıma modeli için Tekrarlayan Sinir Ağı (RNN) kullanan bir model önermektedirler. Yazarlar geliştirdikleri rakam tanıma modelinde, gürültü azaltma ve basamak ayırma işlemlerinden sonra, en iyi konuşma sinyali temsilini elde etmek için Mel Frekans Cepstrum Katsayıları (MFCC) tabanlı özellik çıkarımı yöntemini kullanmışlardır. Basamak konuşmasından çıkarılan bu özellikler, Uzun Kısa Süreli Bellek (LSTM) hücreleri ile donatılmış bir ağa beslenir. LSTM hücreleri, uzun süreli öğrenmeyi gerektiren zamansal bağımlılıkları ele alabilme ve RNN ile ilişkili kaybolan gradyan sorunlarını çözebilme yeteneğine sahip olduğundan, önerdikleri modelin etkili ve başarılı bir Arapça rakam konuşma tanıma yöntemi olarak öne çıktığını ifade etmişlerdir. Çalışmalarında, farklı ülkelerden ve lehçelerden toplanan 1040 Arapça rakam örneğinden oluşan bir veri seti kullanılmışlardır. Bu veri setinden 840 örnek ağı eğitmek için, diğer 200 örnek ise test amacıyla kullanılmıştır. Model eğitimi, Grafik İşleme Birimi (GPU) özellikli bir bilgisayar sistemi kullanılarak gerçekleştirilmiştir. LSTM modeli için öğrenme parametreleri optimizasyon amacıyla ayarlanmış ve model eğitimi sürecinde minimum hata ile %94'lük yüksek bir doğruluk seviyesine ulaşılmıştır. Her ses örneğinin test sonucu, ses dosyalarının kalitesine bağlı olarak değişiklik göstermektedir. Özellikle sinyalin aşırı gürültülü olduğu durumlarda, model çıkarılan ses özelliklerinden bazılarını doğru bir şekilde tanınmadığı yazarlar tarafından belirtilmiştir. Test amacıyla toplam 200 örneğin her biri ayrı bir basamak için 10 örnekten oluşan bir kümeye sahip 20 konuşmacıdan oluşmaktadır. Bu durumda, her bir rakam 20 farklı örnek üzerinde test edilmiştir. En iyi doğruluk 16 doğru ve 4 yanlış ile '0' rakamı için elde edilmiş olup %80'lik bir doğruluk oranına sahiptir. En düşük doğruluk oranı ise '6' rakamı için %60 olarak kaydedilmiştir. Diğer rakamlar yaklaşık %65 doğruluk oranına sahipken, sistem 200 veri noktası için ortalama %69 doğruluk değeri elde etmiştir. Derin öğrenme uygulamalarının çoğu genellikle iyi sonuçlar elde etmek için çok daha fazla veri noktasına ihtiyaç duymaktadır. Bu nispeten düşük genel doğruluk, eğitim veri kümesinde yer alan

840 veri noktasının sayısının sınırlı olmasından kaynaklandığı belirtilmiştir. Ayrıca, her rakam için MFCC tarafından çıkarılan özelliklerin bazı benzerlikleri, ağın test edildiğinde karışıklığa neden olmuştur. Yanlış sınıflandırmalar, bazı rakamların özelliklerinin başka rakamlara benzerliği nedeniyle meydana gelmektedir. Bu sonuçlar neticesinde yazarlar, geliştirdikleri sistemin, diğer açıklanan sistemlere göre daha basit bir yapıya sahip olduğunu, kullanılan parametrelerin, yüksek doğruluk elde etmek için optimal seviyede olduğunu, daha büyük bir eğitim veri seti veya daha fazla sayıda MFCC katsayısının kullanılması, daha iyi sonuçlar elde edilmesine katkı sağlayabileceğini belirtmişlerdir (Wazir ve Chuah, 2019).

Sözlü rakam tanıma, günümüz dijital teknolojilerinde oldukça geniş bir uygulama alanına sahiptir. Bu alanda, çeşitli öznitelik mühendisliği ve sınıflandırma stratejileri üzerinde çalışmalar yapılmıştır. Sharan, (2020) konuşulan rakamların tanınması için Evrişimli Sinir Ağlarının (CNN) potansiyelini araştıran bir çalışma yürütmüştür. CNN aslen bir görüntü sınıflandırıcı olmasına rağmen, bu araştırmada konuşulan rakamların zamansal ve frekansal gösterimini elde etmek amacıyla kullanılmıştır. Özellikle konuşma gibi düşük frekanslı sinyaller için daha iyi frekans lokalizasyonu sağladığı için zaman – frekans temsilinin oluşturulmasında dalgacık dönüşümü kullanılmaktadır. Zaman – frekans temsili, elde edilen görüntü benzeri skalogram adı verilen bir temsil biçimini alır ve bu skalogram, CNN kullanılarak konuşulan rakamların tanınmasında değerlendirilir. Sonuç olarak, bu çalışma sözlü rakam tanıma alanında önemli bir adım olarak karşımıza çıkmaktadır. Yazar, önerdiği yaklaşımı 0'dan 9'a kadar olan on sözlü basamağa ait 56.290 parça içeren bir veri kümesi üzerinde değerlendirmiştir. Bu veri kümesi, çeşitli diğer sözlü kelimelerden oluşan basamaksız ve arka plan gürültüsüyle birlikte sunulmuştur. Basamaklı, basamaksız ve arka plan sinyallerini tanımak için farklı zaman – frekans gösterimleri arasında yapılan karşılaştırmada, dalgacık skalogram gösterimleri en küçük hatayı üretmiştir. Önerilen bu yöntemle elde edilen sonuçlar oldukça etkileyicidir. %2,85 ve %2,84'lük genel doğrulama ve test hatası sonuçları ile çeşitli geleneksel yöntemlerden daha iyi performans göstermiştir. Sharan, veri setinin kapsamlılığı ve Skalogram – CNN yaklaşımının sınıflandırma performansına vurgu yaparak, CNN'lerin ayırt edici özellikleri doğrudan zaman – frekans temsillerinden öğrenme yeteneği sayesinde, önemli öznitelik çıkarma mühendisliği gerektirmeden bu yaklaşımın kullanılabileceğini vurgulamıştır. Ayrıca bu çalışmanın gelecekte sözlü rakam tanıma ve diğer benzeri uygulamalarda büyük bir avantaj olabileceğini dile getirmiştir (Sharan, 2020).

Teknolojinin gelişmesiyle birlikte, konuşma tanıma teknolojilerinin çeşitli alanlarda yaygın olarak kullanımı artmaktadır. Otomatik konuşma tanıma (ASR), insan konuşmasını kolayca anlaşılabilir metin veya kelimelere dönüştüren bir yöntemdir. Bangla sayı tanıma sistemleri üzerine yapılan önceki çalışmalarda sadece 0 ile 9 arasındaki rakamlar kullanılmış ve iki veya üç heceli sayılar genellikle göz ardı edilmiştir. Sen ve Roy, (2022) konuşmadan Bangla rakamlarını tanıyabilen bir Evrişimsel Sinir Ağı (CNN) modeli geliştirmek için bu çalışmayı yürütmüşlerdir. Bu çalışma kapsamında, Bangladeş vatandaşlarından çeşitli cinsiyet, yaş ve dillerden katılımcılar kullanılarak 0 ile 99 arasındaki Bangla sayılarının konuşma örneklerinden oluşan bir veri kümesi oluşturulmuştur. Elde edilen konuşma verileri üzerinde zaman kaydırma, hız ayarı, arka plan gürültüsü karıştırma ve ses ayarı gibi ses artırma teknikleri uygulanmıştır. Sonrasında, verilerden anlamlı özelliklerin çıkarılması için Mel Frekans Cepstrum Katsayıları (MFCC) kullanılmıştır. Bu araştırmada, Evrişimsel Sinir Ağlarına (CNN) dayalı bir yöntem kullanılarak Bangla sayı tanıma alanında önemli bir gelişme sağlanmıştır. Oluşturulan veri kümesi üzerinde yapılan testlerde, Bangla dilinde konuşulan sayıları %89,61 doğrulukla tanıma başarısı elde edilmiştir. Modelin performansı, 10 kat çapraz doğrulama yöntemiyle onaylanarak genel doğruluk oranının %89,74 olduğu tespit edilmiştir. Yazarlar, bu geliştirilen yöntemi rakam tanıma alanındaki diğer çalışmalarla karşılaştırmış ve 0 ile 9 arasındaki rakamların sınıflandırılmasında da üstünlüğünü koruduğunu vurgulamışlardır. Bangladeşli katılımcılardan kaydedilen 0 ile 99 arasındaki sayıların telaffuzunda birçok varyasyon olduğuna dikkat çekilmiş ve bu nedenle bu alanda daha fazla çalışmanın yapılmasının gerekliliği yazarlar tarafından ayrıca vurgulanmıştır (Sen ve Roy, 2022).

Otomatik konuşulan rakam tanıma, konuşma tanımadaki önemli alanlardan biridir. Teknolojik ilerlemenin bir sonraki aşaması, yerel dilde konuşulan rakamların tanınmasını içerir. Peştuca konuşulan rakam tanıma konusunda, Nisar, vd. (2017) Peştuca rakam tanıma üzerine az sayıda çalışma yapıldığını dile getirip yaptıkları bu çalışmanın oldukça önemli olduğunu ifade etmişlerdir. Ana dilleri Peştuca olan 75 erkek ve 75 kadın katılımcıdan rakamların seslendirildiği ses kayıtları alarak çok çeşitli ve kapsamlı bir veri seti oluşturmuşlardır. Bu veri setinde çeşitlilik oluşturmak için yaşları 16 ile 70 arasında değişen kişilerin tercih edildiği yazarlar tarafından dile getirilmiştir. Örüntü tanıma sürecinde, en önemli adımlardan biri özellik çıkarmadır. İlgili özellikler, bir sistemin sağlamlığına yönelik kararlar üzerinde etkilidir. Bu bağlamda, ilgili özellikler, nesne hakkında

diğer sınıflardan daha fazla bilgi sunan nitelikleri ifade eder. Mel Frekansı Kepstral Katsayılarını (MFCC) kullanılarak çıkarılan özellikler sınıflandırma amacıyla eğitim ve test olmak üzere iki kümeye ayrılmıştır. Destek Vektör Makinesi (SVM) ve K – En Yakın Komşu (KNN) sınıflandırıcıları eğitilmiş ve test edilmiştir. Yazarlar, destek vektör makinesi (SVM) sınıflandırıcısının Peştuca rakam sınıflandırmasında ilk defa kullanıldığını ve K-en yakın komşu (KNN) sınıflandırıcısıyla karşılaştırdığını belirtmişlerdir. İlk aşamada, 50 konuşmacıdan oluşan eğitim ve test veri seti kullanılarak KNN sınıflandırıcısıyla Peştuca rakam tanıma doğruluğu %76,8 olarak elde edilmiştir. Daha sonra, veri kümesi 150 konuşmacıya genişletilerek %74 eğitim ve %26 test verisi ile Peştuca rakam tanıma yapılmış ve KNN sınıflandırıcısıyla doğruluk oranı %87,75'e yükseltilmiştir. Sınıflandırıcı olarak SVM kullanıldığında ise sonuçlar oldukça tatmin edici olmuş ve tüm doğruluk %91,5'e ulaşmıştır. Yazarlar, verimlilik düzeyini koruyarak doğruluğu artırmak için bazı benzersiz özelliklerin çıkarılmasını planladıklarını ifade etmişlerdir (Nisar, vd. 2017).

Konuşma tanıma, konuşma dalgasındaki karakteristik bilgileri kullanarak konuşmacıyı tanımlama işlemidir. Güncel gelişmeler, ses tanıma teknolojisinin güvenlik sistemlerinde kullanılmasını mümkün kılmaktadır. Rudresh, vd. (2017) konuşma rakamı tanıma sistemi üzerine bir çalışma yürütmüşlerdir. Yürüttükleri bu çalışmanın, özellikle kişisel ses tanımlama, telefon bankacılığı, sesli arama ve veri tabanı erişim hizmetleri gibi kontrol ve erişim süreçlerinde kullanılabileceğini vurgulamışlardır. Yaşları ve cinsiyetleri farklı 15 kişiden birer hafta arayla 3 oturumda 0'dan 9'a kadar olan rakamların seslendirildiği ses kayıtlarını alarak bir veri seti oluşturmuşlardır. İlk iki oturumun kayıtları eğitim amacıyla kullanılmışken, üçüncü oturumun tüm verileri ise test amacıyla ayrılmıştır. Konuşma rakamı tanıma, iki temel bölümden oluşur. Bunlar özellik çıkarma ve özellik eşleştirme. Yazarlar, konuşulan rakamların tanınması sürecinde Cepstrum tekniğini özellik çıkarımı için ve Vektör Niceleme (VQ) yöntemini özellik eşleştirmesi için kullanmışlardır. Ayrıca, geliştirdikleri sistemin performansını test etmek için Öklid Uzaklığı formülünü uygulamışlardır. Çalışmalarının sonucunda en düşük başarı oranının %93.3, genel performansın ise %98.5 olduğunu belirtmişlerdir. Ayrıca, tüm rakam ifadelerinin en yüksek performansa sahip olduğunu gözlemlemişlerdir. Bu sonuçlara dayanarak yazarlar, çalışmalarının ışık, fan veya sesli arama gibi komut kontrol uygulamaları için kullanılabileceğini ifade etmişlerdir (Rudresh, vd. 2017).

Ramo ve Younis, (2021) Arapça dilinde konuşulan sayıları tanımak amacıyla akıllı konuşma sistemi çalışması yapmışlardır. Araştırmacılar çalışmalarında ön işleme aşamaları olarak ses dosyası üzerinde bir dizi işlem gerçekleştirmişlerdir. Ses dosyalarının boyutunu verimli bir şekilde azaltma ve temel özelliklerini çıkarma süreçlerinde, Max Mean Log (MMLog) adı verilen yeni bir yaklaşımı kullanmayı tercih etmişlerdir. Sonraki adımda, dosyaları tanıma işlemine hazırlamak amacıyla çeşitli ilk işlem aşamaları uygulamışlardır. Ardından, Arapça konuşulan sayının türünü tespit etmek ve söz konusu sayıyı görsel bir metne dönüştürmek için Karınca Aslanı (AntLion) adı verilen gelişmiş yapay zekâ algoritması kullanılarak rakam tanıma süreci başlatılmışlardır. Çalışmada kullanılan tüm sayılar, yaşları 15 ile 50 arasında değişen farklı sayıda kişi tarafından 176 kez tekrarlanmıştır. Bu şekilde, 1800 farklı ses dosyasından oluşan kapsamlı bir veri seti elde edilmiştir. Önerilen yöntemin etkinliği, konuşulan sayıların tanımlanması ve tanınması için sistemin doğruluk derecesini belirlemek amacıyla çeşitli ölçütler kullanılarak titizlikle değerlendirilmiştir. Yazarlar tarafından önerilen yöntem son derece etkili ve hızlı bir şekilde çalışmaktadır. Bu algoritma, söylenen sayıları tanıma konusunda %99'luk Doğruluk (Accuracy) oranına sahiptir. Veri setindeki 1.800 farklı ses dosyası için Yanlış Negatif Oranı (FNR – False Negative Ratio) sadece %1'dir. Bununla birlikte, kullanılan veri setinden farklı olan 40 ses dosyası sisteme dahil edilmiştir. Kullanılan veri setinden farklı olan, tanıdık olmayan bu 40 farklı ses üzerine ek incelemeler yapıldıktan sonra Doğru Tanıma Oranı (CRR – Correct Recognition Ratio) %72,5 ve Doğruluk (Accuracy) %93 ölçülmüştür. Bu sonuçlar, MMlog algoritmasının ses dosyalarından başarılı bir şekilde özellikler çıkarma yeteneğini ortaya koymuştur. Bu özelliklerin çıkarılması, ayırt etme sürecinde oldukça etkili nitelikler elde edilmesini sağlamıştır. Aynı zamanda, bu yöntem sayesinde ses dosyasının boyutu önemli ölçüde azalmış ve geliştirilen AntLion (Karınca Aslanı) algoritmasının işleyeceği veri miktarı büyük ölçüde azaltılmıştır. Elde edilen sonuçlar, algoritmanın ses dosyalarıyla başarılı bir şekilde çalıştığını ve Arapça konuşulan sayıları mükemmel bir şekilde tanımlayabildiğini göstermiştir. MMLog özellik çıkarma yöntemi ile yüksek düzeyde ayırt edici sonuçlar elde edilmiştir. Yazarlar, Karınca Aslanı algoritmasının (AntLion) hedefe ulaşma yeteneğinin hızlı ve doğru olduğunu vurgulamışlardır. Özellikle iterasyonun ilk aşamalarında algoritmanın sömürüye çok yakın olduğunu ve zaman zaman rastgele yürüyüşün keşif sürecinde algoritmaya yardımcı olduğunu belirtmişlerdir. Hızlı bir şekilde çözüm bulmak için keşif ve sömürü arasında gerekli dengeyi koruduklarını ifade etmişlerdir (Ramo ve Younis, 2021).

Otomatik Konuşma Tanıma (ASR), sözlü ifadelerin metne dönüştürülmesi ve belirli bir sistemi izlemek veya komuta etmek için kullanılabilir. Zerari, vd. (2018) konuşma ifadelerinin farklı uzunluklarıyla başa çıkmak için Uzun Kısa Süreli Belleği (LSTM) kullanan sıralı öğrenme tabanlı bir uçtan uca yaklaşım önermektedirler. Önerdikleri yaklaşımla, doğal insan – makine etkileşimini desteklemek amacıyla, bir sınıflandırma yöntemi kullanarak izole edilmiş bir Arapça kelimenin Arapça konuşulan rakam yazımını tanıyabileceğini belirtmektedirler. Önerdikleri sistem, ilk olarak giriş konuşma sinyalinden ilgili özellikleri çıkarmak için Mel Frekans Cepstral Katsayılarını (MFCC) kullanır ve daha sonra bu özellikleri, farklı uzunluktaki dizilere uyum sağlayabilen derin bir sinir ağı ile işler. MFCC özelliklerinin dizilerini kodlamak için Tekrarlayan Sinir Ağları (RNN) mimarilerinden olan Uzun Kısa Süreli Bellek (LSTM) ve Geçitli Tekrarlayan Birimler (GRU) kullanılarak, Çok Katmanlı Algılayıcı (MLP) ağı beslenir ve tüm sinir ağı sınıflandırıcısı uçtan uca eğitilir. Çalışmalarında, Cezayir'deki Annaba Badji – Mokhtar Üniversitesi Otomatik Sinyaller Laboratuvarı tarafından derlenen bir veri setini kullanmışlardır. Bu veri seti, Arapça rakamlara karşılık gelen MFCC'lerin zaman serilerini içerir. Veri seti, anadili Arapça olan 44 erkek ve 44 kadından alınan ses kayıtları ile toplam 8800 rakam sözcüğünü içerir. Bu rakam sözcükleri 0'dan 9'a kadar gruplandırılmış ve her rakam 10 kez kaydedilmiştir. Yazarlar önerdikleri sistemin %98,77 bir sonuç ile aynı veri seti kullanılarak daha önce yayınlanan sonuçlardan büyük bir farkla mükemmele yakın iyi performans gösterdiğini belirtmişlerdir. Arapça rakamlar üzerinde elde edilen bu sonuçlar, önerilen modelin etkinliğini doğrulamakla birlikte, umut verici sonuçlar ortaya koyduğunu ve bu sonuçlara dayanarak, MFCC özniteliklerinin bir sınıflandırma sistemine tanıtılan bu tür görevler için konuşma sinyalini karakterize edecek kadar verimli olduğunu belirtmişlerdir. Ayrıca gelecekteki çalışmalarda bu tür sistemleri, ASR için büyük zorluk olarak görülen gürültülü ortamlarda ve daha gerçekçi konuşma sinyalleriyle değerlendirmek gerektiğini de dile getirmişlerdir (Zerari, vd. 2018).

Dijital çağın etkileriyle birlikte, miyopi ve kırma kusuru gibi göz rahatsızlıkları her geçen yıl hızla artmaktadır. Bu rahatsızlıklar, yakındaki nesneleri net, uzaktaki nesneleri bulanık görmeye neden olmaktadır. Artan vakalar nedeniyle görme keskinliği testi yaptırmak isteyenlerin sayısı da artış göstermektedir. Ancak, geleneksel testler işinin ehli uzmanlar tarafından yapılmalıdır ve bu nedenle bazı kişiler bu testlere ulaşmakta zorluklar çekmektedir. Bu nedenle insanlar kendi başlarına görme keskinliği testini yapabilecekleri bir yönteme ihtiyaç duymaktadır. Evde, herhangi bir zamanda, uzman yardımı

olmaksızın bu testi kendileri gerçekleştirebilmek büyük bir kolaylık sağlayacaktır. Bu ihtiyacı karşılamak amacıyla, Taufik ve Hanafiah, (2021) çalışmalarında, mikrofonu giriş aygıtı ve monitör olarak kullanarak genel bir bilgisayar üzerinde çalışabilen otomatik bir görme keskinliği testi olan AutoVAT (Automated Visual Acuity Test) geliştirmeye çalışmışlardır. Bu otomatik test, Snellen tablosu kullanılarak rakam optotipi ile görme keskinliğini ölçmektedir. Kullanıcı, sözlü rakam testini tamamlar ve sonuçları değerlendirilir. Sözlü rakam tanımada çeşitli yöntemler kullanılmaktadır. Konuşulan rakam tanıma sistemleri genellikle üç temel bileşeni içerir. Bu bileşenler ses veri seti, ön işleme ve sınıflandırma yöntemlerinden oluşmaktadır. Herhangi bir bileşenin değiştirilmesi, sistem tarafından elde edilen sonuçlarda farklılıklara sebep olabilmektedir. Yazarlar, çalışmalarında konuşulan rakamların özniteliklerini çıkarmak için Mel Frekans Cepstral Katsayılarını (MFCC), sınıflandırmada ise Evrişimli Sinir Ağlarını (CNN) tercih etmişlerdir. Önerdikleri CNN modeli, sözlü rakam tanıma için %91.4 doğrulukla olumlu bir performans sergilemiştir. Bu sonuçlar neticesinde önerdikleri Evrişimli Sinir Ağı (CNN) modelinin konuşma rakamlarını sınıflandırmada başarılı bir performansa sahip olduğunu ve Mel Frekans Cepstral Katsayılarının (MFCC) bir konuşmanın ayırt edilebilir özelliklerini çıkarmak için güvenilir bir yöntem olduğunu belirtmişlerdir. Yazarlar, AutoVAT çalışmasının, geleneksel Snellen grafik testine kıyasla sadece 0.19 sıra farkla umut verici sonuçlar sunduğunu ifade etmişlerdir. Bu geliştirilen yöntem, insanların görme keskinliğini evlerinde daha sık ve kolay bir şekilde takip etmelerine olanak sağlayarak, göz sağlığının korunmasına katkıda bulunabilecektir. Bu sayede, dijital çağın getirdiği göz sağlığı zorluklarına karşı daha bilinçli ve tedbirli bir yaklaşım benimsememize yardımcı olabilecektir. Geliştirilen AutoVAT, sözlü rakam kullanarak Snellen grafik testini başarıyla gerçekleştirmiştir. Ancak, sözlü harf tanımayı kullanarak yaygın olarak kullanılan Snellen şeması testini uygun bir şekilde simüle edememiştir. Bu nedenle, tamamen sözlü harf tanımayı kapsayacak ve Snellen grafik testini daha etkili bir şekilde uygulayacak yeni bir yöntem için daha fazla araştırma ve çalışma yapılması gerektiğini de belirtmişlerdir (Taufik ve Hanafiah, 2021).

Konuşmadan metne dönüştürme, insan duygu ve düşüncelerini ifade etmede önemli bir araç olan Yapay Zekâ alanında oldukça önemli bir konudur. Ancak bu alanda metinden konuşmaya dönüştürme çalışmaları ile kıyaslandığında konuşmadan metne dönüştürme konusunda daha az çalışma yapılmıştır. Chowdhury, vd. (2021) sinir ağları modellerini kullanarak konuşmadan metne dönüştürmede daha doğru sonuçlar elde etmeyi

amaçlamıştır. Bu doğrultuda Bengalce kelimelerden oluşan yeni bir veri seti hazırlayarak çalışmalara başlamışlardır. Araştırmalarında hem Uzun Kısa Süreli Bellek (LSTM) hem de Geçitli Tekrarlayan Birimler (GRU) kullanarak sonuçları karşılaştırmışlardır. Yapılan karşılaştırmada, LSTM ile elde edilen doğruluk oranının %90 olduğu görülmüştür. Bu sonuç, GRU kullanımından vazgeçmelerine sebep olmuştur. Ayrıca GRU'nun kararsızlığı ve dalgalanmaya başlaması, LSTM'nin daha kararlı olması ve kayıp fonksiyonu durumunda hataları en aza indirmesi, doğruluk açısından LSTM'ye kıyasla daha az başarı elde edilmesi GRU'dan vazgeçme nedenleri olarak belirtilmiştir. Yazarlar, elde ettikleri sonuçları doğrulamak için veri testlerini manuel olarak ta yapmışlardır. Ayrıca modelin daha önce hiç beslemediği bir veri kümesinde de yaklaşık %90 doğruluk oranı elde edildiğini belirtmişlerdir. Bu çalışma doğrultusunda gelecekte, otomatik cümle tanıma, ifadenin yanıt temelini hazırlama süreci ve buna bağlı olarak duyarlılığı değiştirme ile ilgili çalışmalarda yapılabileceğini belirtmişlerdir (Chowdhury vd. 2021).

İşitme ve konuşma engelli bireyler için, geniş bir kitle iletişim aracı olan işaret dili, ses veya ses sinyallerini kullanamadıkları için bilgi iletişiminde önemli bir rol oynar. Onlar, günlük yaşamlarında duygularını ifade etmek için görsel sinyal alışverişine yönelirler çünkü konuşma yeteneğinden yoksundurlar ve duyamazlar. İşitme ve konuşma engelli bireyler işaret dilini, özellikle el hareketleri ve beden dili gibi çeşitli yöntemlerle duygu ve düşüncelerini ifade etmek için kullanılır. İşaret dilinin temelleri büyük ölçüde rakamlar ve karakterler şeklinde iki ana grupta incelenebilir. Das, vd. (2023) yapmış oldukları çalışmalarında, Bangla İşaret Dilinin (rakamlar ve alfabedeki karakterler) otomatik olarak tanınması için Rastgele Orman (RF – Random Forest) sınıflandırıcısı ve derin transfer öğrenme tabanlı bir Evrişimsel Sinir Ağı (CNN) birleşimi olan hibrit bir model önermişlerdir. Geliştirdikleri sistemin genel performansı, 'Ishara – Bochon' ve 'Ishara – Lipi' veri kümeleri üzerinde yapılan testlerle başarılı bir şekilde doğrulanmıştır. 'Ishara – Bochon' ve 'Ishara – Lipi', Bangla İşaret Dili (BSL) için ilk eksiksiz, çok amaçlı ve halka açık veri kümesidir. Bu veri kümesi izole edilmiş rakamlar ve alfabedeki karakterlerden oluşmaktadır. Yazarlar, çalışmalarının ön işleme aşamasında, modelin ilgilenilen bölgeye odaklanmasını ve bu sayede çıkarım hızını artırmayı hedefleyerek, görüntülerden gereksiz arka plan bilgilerini kaldırmak için bir arka plan eleme algoritması kullanmışlardır. Önerdikleri arka plan eleme yöntemi sayesinde, karakter tanıma için %91,67, %93,64, %91,67 ve %91,47; rakam tanıma için ise %97,33, %97,89, %97,33 ve %97,37 düzeyinde Doğruluk (Accuracy), Kesinlik (Precision), Duyarlılık (Recall) ve F1 – Skor (F1 – Score)

değerlerine ulaşmışlardır. Yazarlar, yaptıkları çalışmanın sonuçlarına göre önerilen modelin önceden eğitilmiş modelleri kullandığına dikkat çekmiştir. Bu model, daha küçük veri kümeleri için güvenilir bir mimari sunmakla birlikte, düşük öğrenme oranına sahip ince ayar süreci ile aşırı öğrenmeyi önleyerek modelin performansını artırmaktadır. Rastgele Orman (RF) sınıflandırıcısının, ince ayarlanmış özelliklerin rastgele alt kümelerini kullanarak, birkaç karar ağacının ortalamasını alarak nihai modelin varyansını azalttığı belirtilmiştir. Ayrıca, arka plan eleme gibi veri ön işleme yönteminin, işlenmemiş verilere göre daha iyi sonuçlar ürettiği ifade edilmiştir. Tüm bu bulgular, önerilen sistemin, doğruluk açısından son teknoloji tanıma sistemlerinden daha üstün performans sergileyen sistemlerin geliştirilmesine katkıda bulunabileceğini vurgulamıştır (Das vd. 2023).

Son yıllarda, çevrimiçi öğrenme gitgide daha yaygın hale gelmektedir. Örneğin, öğrencilerin değerlendirilmesine yönelik akıllı bir sürece sahip, sanal öğretmenler ve gerçek sınıf ortamında kullanılan standartlaştırılmış eğitimlere benzeyen bir çevrimiçi telaffuz uygulama sistemi geliştirilmiştir. Bu sistem günümüzde Bilgisayar Destekli Telaffuz Eğitimi (CAPT) olarak adlandırılmaktadır. Tayca ya da Merkez Tay dili, tarihi olarak Siyamca olarak bilinen Tayland devletinin tek resmi dilidir. Merkez Tay halkıyla Çin kökenli Tayların büyük çoğunluğu tarafından konuşulan bir dildir. Rukwong ve Pongpinigpinyo, (2022) yapmış oldukları çalışmalarında, Tayca sesli harfleri tanımak için Derin Öğrenme temelli bir çevrimiçi otomatik Bilgisayar Destekli Telaffuz Eğitimi (CAPT) sunmaktadır. Evrişimli Sinir Ağı (CNN) olarak adlandırılan bir derin öğrenme modeli, Tayca ünlülerin sınıflandırılmasında kullanılmış ve geliştirilmiştir. Dilbilimciler tarafından tasarlanan yeni bir veri seti, Tayca ünlüler için toplanmış ve incelenmiştir. Bu veri seti, yoldan geçen araçlar, konuşan insanlar, parkta rüzgâr ve hayvan sesleri gibi 30-40 desibel (dB) Sinyal Gürültü Oranı (SNR – Signal to Noise Ratio) gürültü içermektedir. Sesler, Tayca ana diline sahip 50 kişi tarafından cep telefonlarıyla kaydedilmiştir. Veri seti, 18 sınıfa ayrılmış olan sesli harflerden toplamda 1800 ses dosyasını içermektedir. Yazarlar, çalışmalarında Mel Spektrogramı (MS) kullanarak optimize edilmiş bir CNN modelinin %98,61 doğruluk oranına ulaştığını belirtmişlerdir. Ayrıca, Uzun Kısa Süreli Bellek (LSTM) modeli ile Mel Frekans Cepstral Katsayıları (MFCC) kullanıldığında %94,44 başarı elde edildiği, Uzun Kısa Süreli Bellek (LSTM) modeli ile Mel Spektrogramı (MS) kullanıldığında ise %90,00 başarı sağlandığını ifade etmişlerdir. Bu bulgular, çeşitli veri boyutlarının, veri tasarlama, toplama ve doğrulama süreçlerinin, konuşma ta-

nıma modelleri için yüksek kaliteli girdi verileri oluşturmak için son derece faydalı olduğunu göstermektedir. Yazarlar, çalışmalarının gerçek zamanlı çalıştığını belirterek, çalışmalarına çeşitli stratejiler uygulayarak, çalışmalarındaki sinir ağlarını optimize ederek ve gelecekte öğrenme destek sistemlerinde daha iyi performans göstermeleri için çalışmalarının hiper parametrelerini ayarlayarak benzer bir alanda CNN modellerinin geliştirilmesi için kullanılabileceğini belirtilmişlerdir (Rukwong ve Pongpinigpinyo, 2022).

Konuşma tanıma, bilgisayar bilimleri dünyasında en etkileyici konulardan biridir. Konuşma tanıma çerçevesinin doğruluğu, konuşma sinyalindeki gürültüyle yakından ilişkilidir ve bu durum doğruluğu olumsuz yönde etkileyebilir. Bu nedenle, otomatik konuşma tanıma (ASR) sistemleri için gelişim açısından gürültü giderme işlemi hayati bir öneme sahiptir. ASR, her dilin kendine özgü özelliklerinin olduğu gerçeği göz önünde bulundurularak çeşitli diller için araştırılmıştır. Tamil dili, Hindistan ve Sri Lanka'da yaşayan Tamiller tarafından ağırlıklı olarak konuşulan bir dildir ve yaklaşık 75 milyon insan tarafından kullanılmaktadır. Ayrıca, Sri Lanka ve Singapur'da resmi dil statüsündedir. Son birkaç yılda, Tamil dilindeki ASR çerçevesi ihtiyacı önemli ölçüde artmıştır. Lokesh, vd. (2019) Tamil dili konuşma tanıma için kendi Kendini Organize Eden Harita Tabanlı (SOM – Self Organizing Map) sınıflandırma yöntemiyle birlikte Çift Yönlü Tekrarlayan Sinir Ağını (BRNN – Bidirectional Recurrent Neural Network) kullanmışlardır. Özellik vektörü, kullanılan SOM özellik vektörünün doğru uzunluğunu seçmek için ölçü olarak kaydırılır. Sonunda, Tamil rakamları ve kelimeler, BRNN sınıflandırıcısı kullanılarak sıralanır ve SOM tarafından yerleşik uzunluğu belirlenen özellik vektörü, BRNN-SOM olarak adlandırılan girişe dönüştürülür. Yazarlar, Tekrarlayan Sinir Ağları (RNN) ve Derin Sinir Ağı – Gizli Markov Modeli (DNN-HMM) gibi yaklaşımları kullanarak çıkan sonuçları kendi önerdikleri modelin sonuçları ile karşılaştırmışlardır. Önerdikleri yaklaşım, Sinyal Gürültü Oranı (SNR), Sınıflandırma Doğruluğu ve Ortalama Kare Hatası (MSE – Mean Squared Error) ile ilgili sonuçlara bakıldığında %93,6 performans sergilemiş ve karşılaştırdıkları RNN ve DNN-HMM sistemlerinden daha iyi sonuçlar elde etmişlerdir. Lokesh, vd. bu sonuçlara bakıldığında, önerdikleri modelin Tamil dili için geniş kapsamlı bir ASR çerçevesi oluşturmak için gelecekte sinir ağı tabanlı planlar veya diğer çapraz hibrit stratejilerle birlikte kullanılabileceğini belirtmişlerdir (Lokesh, vd. 2019).

Bhutan Krallığı, Güney Asya'da denize kıyısı bulunmayan bir ülkedir. Ülke, Doğu Himalayalar'da olup, batısında ve güneyinde Hindistan, kuzeyinde ise Çin'in Tibet bölgesi yer alır. Bhutan hükümeti, işitme engelli bireyler ile halk arasındaki iletişim açığını kapatmak için İşitme Engelliler Okulu aracılığıyla insanları Bhutan İşaret Dili'ni (BSL – Bhutanese Sign Language) öğrenmeye teşvik etmektedir. Ancak, İşaret Dili (SL – Sign Language) öğrenmek zorlu bir süreçtir. Wangchuk, Riyamongkol ve Waranusast, (2021) Bhutan İşaret Dili Rakam Tanıma sistemi geliştirmek için çalışmalarında Evrişimsel Sinir Ağı (CNN) kullanmışlardır. BSL rakam tanıma sistemi için farklı gönüllülerden toplanan 10 statik rakamın 20.000 işaret görüntüsüne sahip ilk BSL veri kümesini oluşturarak çalışmalarına başlamışlardır. BSL rakam tanıma sistemi, Evrişimsel Sinir Ağı (CNN), Destek Vektör Makinesi (SVM), K – En Yakın Komşu (KNN), Logistic Regression ve LeNet-5 gibi beş farklı model ile değerlendirilmiştir. Bu değerlendirme süreci sonucunda, modellerin Doğruluk (Accuracy), Kesinlik (Precision), Duyarlılık (Recall) ve F-1 Score performans ölçümleri elde edilmiştir. En yüksek doğruluk oranı %97,62 ile VGGNet'e benzer bir yapıya sahip olan CNN ağında minimum eğitim süresiyle kaydedilmiştir. Yazarlar önerdikleri modelin eğitim ve test doğrulukları sırasıyla %99.94 ve %97.62 olduğunu, çalışmalarında kullandıkları diğer modellerden daha iyi performans gösterdiğini belirtmişlerdir. Ayrıca, veri kümesindeki görüntü sayısını artırarak ve VGG16, ResNet ve MobileNet gibi Transfer Öğrenimi yöntemlerini kullanarak yanlış sınıflandırma oranını düşürebileceklerini ve test doğruluğunu artırabileceklerini ifade etmişlerdir (Wangchuk, Riyamongkol ve Waranusast, 2021).

Gauss (normal) dağılımı tepe noktası tek olan (unimodal) bir dağılımdır. Birden fazla tepe noktası olan bir veriyi modellememiz gerekirse Gauss modeli işimizi görmeyecektir. Birden fazla Gauss modelini karıştırmak (mixing) bu soruna çözüm olacaktır. Karıştırmak demek karışım içindeki her Gauss'tan gelen sonuçları toplamaktır. Kısaca her veri noktasını teker teker karışımdaki tüm dağılımlara aktarıp sonuçları bir ağırlık üzerinden toplamaktır. Yüksek Sinyal Gürültü (SNR) olduğu durumlarda, Gauss Karışım Modelleri (GMM) ile Saklı Markov Modellerinin (HMM) fonksiyonları rakam tanımada yüksek performans sağladığı kabul edilmektedir. Gürültü, bozulma, eğitim ile test arasındaki uyumsuzluk gibi rakam tanıma doğruluğu üzerindeki olumsuz etkileri azaltmak için uzun yıllardır araştırmalar yapılmakta ve performansı yüksek yeni teknikler ortaya çıkmaktadır. Jalalvand, vd. (2015) temeli makine öğrenmesine dayanan Rezervuar Hesapla-

maya (RC – Reservoir Computing) dayalı akustik modelleme üzerine bir çalışma gerçekleştirmişlerdir. Bu model sabit, doğrusal olmayan bir sistemin giriş sinyallerini daha yüksek boyutlu hesaplama alanlarına eşleyen, tekrarlayan sinir ağı teorisinden türetilen bir hesaplama yapmaktadır. Doğrusal olmayan dinamik sistemler olarak rezervuarların yeni bir analizini sunarak, yeni anlayışlar elde edilir ve akustik özelliklerin dinamikleri ve modellenen akustik birimler açısından son derece basit ve oldukça anlaşılır olan yeni bir rezervuar tasarım tarifine dönüştürülür. Rezervuarı bu dinamiklere göre ayarlayarak, sadece gürültü ve diğer olumsuzluklardan yalıtılmış temiz koşullarda değil, aynı zamanda gürültülü koşullarda da iyi performans gösterebilen Rezervuar hesaplama tabanlı sistemler oluşturulabilir. Bu çalışmanın sonucunda bu sistemin gürültü varlığına karşı, aynı akustik parametrelerle çalışan geleneksel bir GMM – HMM sisteminden önemli ölçüde daha sağlam olduğu yazarlar tarafından tespit edilmiştir (Jalalvand, vd. 2015).

Kanara dili olarak ta bilinen Kannada dili Hindistan'ın güneybatısında konuşulan bir dildir. Yaklaşık 50 milyon kişinin konuştuğu bu dil dünyada en çok konuşulan 30 dilden bir tanesidir. Muralikrishna ve Ananthakrishna, (2013) özellik vektörü olarak Mel Frekans Kepstral Katsayılarını (MFCC), örüntü tanıyıcı (depolanan bilgileri gelen verilerle eşleştirme) olarak ta Saklı Markov Modeli (HMM) kullanarak Kannada izole rakam tanıma sistemi geliştirmeye çalışmışlardır. Sistem, Kannada rakamlarının izole edilmiş ifadelerini tanımak için tasarlanmıştır. Gözlem dizisini elde etmek için büyük ölçekli verileri hızlı ve etkin bir şekilde kümeleyen, uygulanması kolay ve en sık kullanılan kümeleme yöntemlerinden K-Ortalama Kümele (K-Means) yöntemi kullanılmıştır. Çalışmanın performansı, birinci ve ikinci dereceden türevleri ile birlikte MFCC'ye göre değerlendirilmiş ve karşılaştırılmıştır. Muralikrishna ve Ananthakrishna, en iyi sonuçları, MFCC özelliklerini birinci ve ikinci dereceden türevleri ile birleştirdiklerinde elde etmişlerdir. Ayrıca tanıma doğruluğunu daha da artırmak için sistemin daha büyük veri tabanı ile eğitilmesi gerektiğini dile getirmişlerdir (Muralikrishna ve Ananthakrishna, 2013).

Malalayam dili Hindistan'da konuşulan 22 tanınmış dilinden bir tanesi olup genellikle Kerala eyaletinde konuşulan ve Hindistan'da yaklaşık 35 milyon kişi tarafından konuşulan bir dildir. Renjith, Joseph ve Anish Babu, (2013) çalışmalarında Malayalam dili için konuşmacıdan bağımsız izole rakam tanıma çalışması gerçekleştirip rakam tanımanın bazı uygulama alanlarını ortaya koymaya çalışmışlardır. Çalışmalarında özellik vektörü olarak Mel Frekans Kepstral Katsayılarını (MFCC), sinyal işleme algoritması

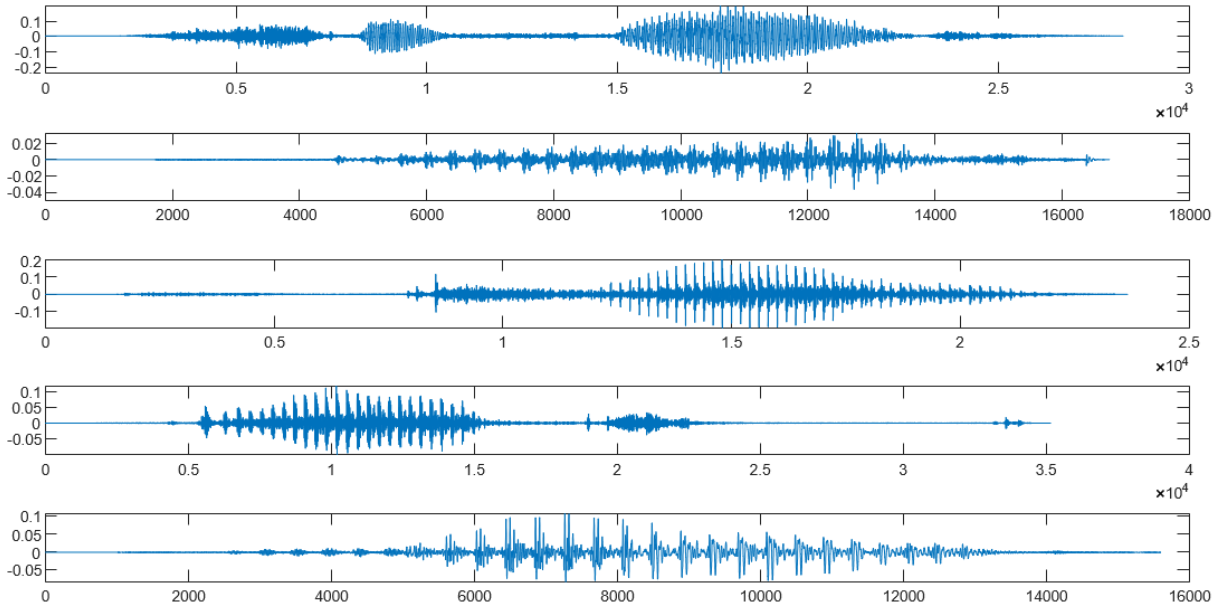
olarak Saklı Markov Modelini (HMM) kullanmışlardır. Ayrıca çalışmalarında rakam tanıma sisteminin omurga yapısı olarak kullanılan ve kullanıcı dostu iyi bir ara yüze sahip olan bir bilgisayar uygulaması geliştirmişlerdir. Kullanıcılar bu uygulama sayesinde menüdeki öğeleri sesli girişler ile kontrol edip seçebilmektedir. Renjith, Joseph ve Anish Babu, rakam tanıma sisteminin gerçek dünyadaki uygulamalarına dikkat çekmeye çalışmışlardır. Çalışmalarının sonucunda izole rakam tanıma ve buna bağlı olarak ta geliştirdikleri uygulama programı tatmin edici bir şekilde çalıştığını ifade etmişlerdir (Renjith, Joseph ve Anish Babu, 2013).

Sami dillerinden olan Arapça dili, İngilizce gibi Avrupa dillerine göre birçok farklılığı olan bir dildir. Bu farklılıklardan biri, sıfırdan dokuza kadar olan rakamların nasıl telaffuz edildiğidir. Sıfır hariç tüm Arapça rakamları çok heceli kelimelerdir. Alotaibi'nin (2005) çalışması Arapça rakam tanıma sisteminin hatalarını açıklamak ve bu hataları önlemek için bir yöntem bulmak amacıyla tüm Arap rakamlarının spektrogramlarının, hece tiplerinin ve fonem dizilerinin analizini ve karşılaştırmasını içermektedir. Rakam tanıma çoklu konuşmacı mod ve konuşmacıdan bağımsız izole mod olmak üzere iki farklı modda çalıştırılmıştır. Yapay Sinir Ağı (YSA) tabanlı bir konuşma tanıma sistemi tasarlanıp Arapça otomatik rakam tanıma ile test edilmiştir. Bant geçiren filtreler aracılığıyla, rakam konuşmalarındaki gürültü giderildikten sonra Ön Vurgu (Pre Emphasis) tekniği ile yüksek frekanslar bir miktar bastırılıp Hamming Penceresi (Hamming Window) tarafından pencerelenip bloke edilmiştir. Konuşma farklılıkları, ses uzunluklarındaki farklılıklar ve sözcük dizilimleri arasındaki yanlış hizalamaları telafi etmek için zaman hizalama algoritması kullanılmıştır. Anadili Arapça olan 17 kişiden 1700 adet rakam sözcüğü içeren bir veri seti oluşturulmuştur. Mel Frekans Kepstral Katsayıları (MFCC) öznitelik çıkarıma algoritması kullanılarak çerçeve özellikler çıkarıldıktan sonra Çok Katmanlı Algılayıcı (MLP) ağı ile rakamlar sınıflandırılmıştır. Yazarın rakam tanıma sistemi, çoklu konuşmacı modda %99,5 ve konuşmacıdan bağımsız izole modda %99,5 oranında doğru rakam tanıma performansı göstermiştir (Alotaibi, 2005).

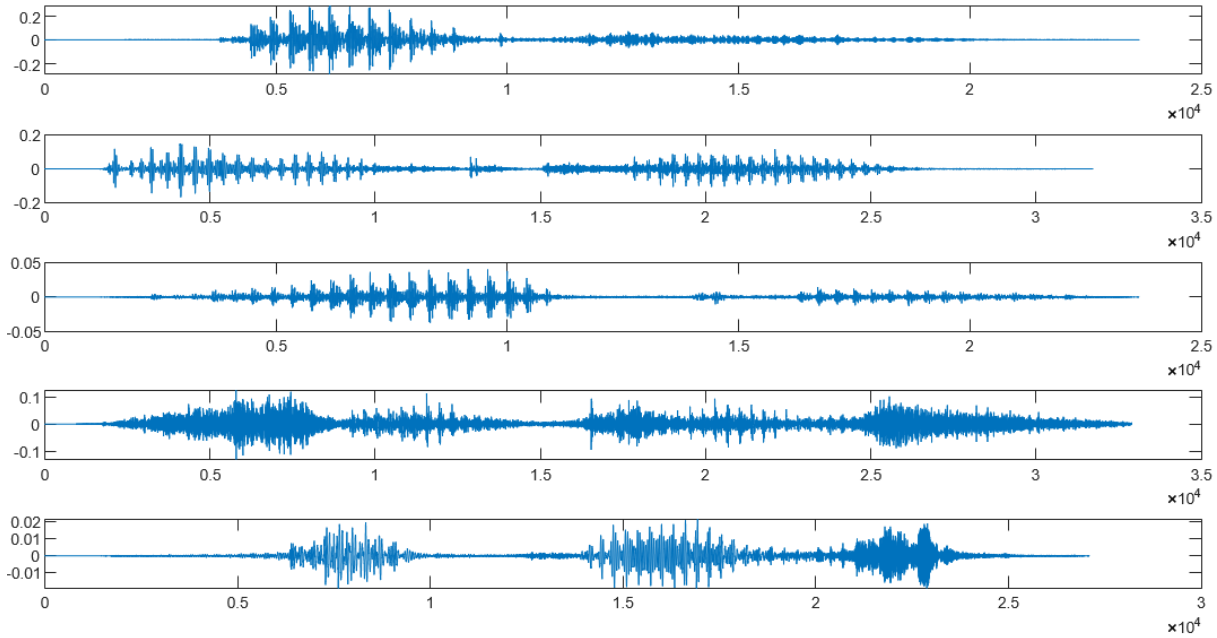
# 3. MATERYAL VE METOT

## 3.1. Materyal

Bu çalışmanın veri setini oluşturmak için, 14 ile 17 yaş aralığında olan ve sesleri farklılık gösteren 25 erkek öğrenci seçildi. Ses kayıtlarını oluşturmak için hafta içi günlerinden çarşamba ve hafta sonu günlerinden cumartesi günleri seçildi. Bu günlerin seçiminde, hafta içi öğrencilerin sabah erken saatlerde kalkmaları ve okulda derslere girip gün boyunca yoğun hareket etmeleri, hafta sonu ise daha geç uyanmaları ve daha az hareketli olmaları göz önünde bulunduruldu. Yoğun bir şekilde hareket edip yorulan bir öğrenciden alınacak ses kaydı ile daha az hareketli olup dinlenmiş bir durumda olan aynı öğrenciden alınacak ses kaydının farklı olabileceği düşünüldü. Belirlenen günlerde, sabah saatlerinde ve öğleden sonra olmak üzere iki farklı zaman diliminde ses kayıtları alındı. Bu yöntemle, her öğrenciden hafta içi ve hafta sonu olmak üzere hem sabah hem de öğleden sonra farklı tonlarda ses kaydı elde edilerek daha fazla çeşitlilik sağlandı. Her kayıt için aynı mikrofon, aynı cep telefonu ve aynı kayıt yazılımı kullanılarak her bir öğrenciden 10 tane ses kaydı alındı. Daha sonra bilgisayarda veri seti için bir ana klasör oluşturulup bu ana klasörün içinde 0'dan 9'a kadar olan rakamlar ile isimlendirilmiş 10 tane alt klasör oluşturuldu. Cep telefonu kullanılarak kaydedilen bu ses dosyaları bir bilgisayara aktarıldı. Bu ses kayıtları ilk olarak bir bilgisayar programı kullanılarak gürültü temizleme (Noise Remove) işlemine tabi tutuldu. Bu adımla, kayıt sırasında istem dışı oluşan arka plan gürültüleri temizlenmeye çalışıldı. Daha sonra, farklı bir bilgisayar programı kullanılarak bu ses kayıtlarının içinde bulunan 0'dan 9'a kadar olan rakam sözcüklerinin başlangıç ve bitiş noktaları dikkatlice işaretlenip belirlendikten sonra teker teker kesildi. Kesilen bu rakam sözcükleri ilgili oldukları alt klasörlere aktarıldı. Bu süreçte, tüm kayıtların aynı bit hızına (128 kbps) sahip olmasına özen gösterildi. Elde edilen bu rakam sözcükleri tek tek incelendi ve her bir rakam sözcüğüne ait alt klasörde 250 tane olmak üzere toplamda 2500 tane işlenmiş temiz rakam sözcüğü elde edildi.



**Şekil 3.1:** 0'dan 4'e kadar olan dijital rakam sözcüklerinin ses bilgileri (Eroğlu ve Kaya, 2023)

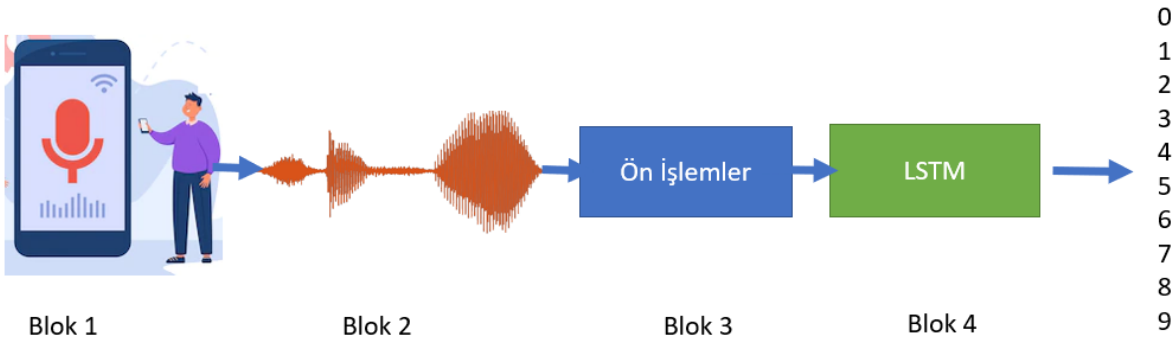


**Şekil 3.2:** 5'ten 9'a kadar olan dijital rakam sözcüklerinin ses bilgileri (Eroğlu ve Kaya, 2023)

## 3.2. Metot

### 3.2.1. Dijit tanıma için blok diyagram

Bu çalışmada önerilen dijit tanıma sistemine ait blok diyagram Şekil 3.3'te verilmiştir. Sistem 4 aşamadan oluşmaktadır. Her aşamada gerçekleştirilen işlemler aşağıda kısaca özetlenmiştir.




**Şekil 3.3:** Dijit Tanıma Sistemi (Eroğlu ve Kaya, 2023)

**Blok 1:** Bu aşamada öğrencilerden ses bilgileri toplatılmıştır. Yaşları 14-16 arasında değişen erkek öğrencilerden her dijit sınıfı için 250 toplamda 2500 ses kaydı alınmıştır.

**Blok 2:** Ses dosyaları kayıt altına alındıktan sonra ilgili klasörlere dağıtılmıştır.

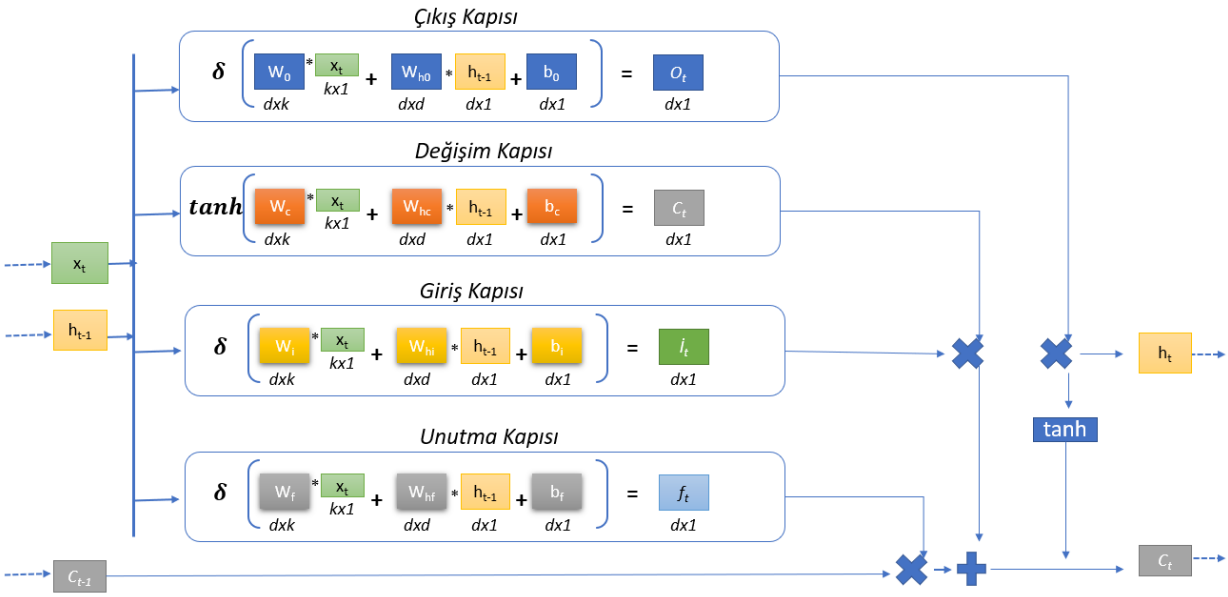
**Blok 3:** Bu aşamada ses dosyaları üzerinde ön işlemler yapılmıştır. Her dosya üzerinde sesten önce ve sonra oluşan boşluklar atılmıştır. Gürültülü bilgiler temizlenmiştir.

**Blok 4:** Son aşamada önişlemden geçirilmiş sesler derin öğrenme metotları ile sınıflandırılmıştır. Sınıflandırma işlemi LSTM modeli ile gerçekleştirilmiştir. Veri setinin %70-30 eğitim-test oranlarına dağılma oranlarına göre sınıflandırma gerçekleştirilmiştir.

### 3.2.2. Long short term memory (LSTM)

Sıralı verilerle ilişkili çeşitli problemlerin çözümünde etkili bir model olarak yaygın bir kullanıma sahip olan LSTM, uzun ve kısa süreli belleği bünyesinde barındıran bir tür tekrarlayan yapay sinir ağıdır (Yin, vd. 2021). LSTM, özel bir RNN türüdür ve basit nöronlar yerine eğitilebilir bellek hücreleri olan LSTM hücrelerini kullanır. (Londhe ve Atulkar, 2021). LSTM'nin en önemli avantajı, içinde bulunan kapılar olarak adlandırılan dahili mekanizmalardır. Bu kapılar, sıralı verilerdeki ilgili bilgileri öğrenme yeteneğine sahiptir ve bu da LSTM'nin tahmin yapma kabiliyetini güçlendirir (Hu, vd. 2020).

Bir LSTM hücresinin yapısı Şekil 3.4''te gösterilmektedir. Burada Long Term (Ct) hücre durumu ve Short Term (ht) hücre durumu bir LSTM hücresinde t zamanındaki ara çıkışı belirten iki durumu belirtmektedir. Bununla birlikte bir LSTM hücresinin iç yapısında unutma, giriş ve çıkış işlemlerinin yapıldığı eğitilebilir 3 kapı da bulunmaktadır (Chen vd. 2020). Bu kapılar, LSTM hücresinin içine alınan ve LSTM hücresinin dışına aktarılan bilgi akışını düzenleme ile birlikte, bilgi ekleme veya bilgi çıkarma kabiliyetine sahiptir. Kapıların bu özelliğinden dolayı LSTM hücresi, belirli zaman diliminde değerleri hatırlar (Hu, vd. 2020). Kapılar, hangi bilgilerin geçişine hücre durumunda izin verileceğine karar veren çeşitli sinir ağlarıdır, ayrıca bu kapılar eğitim sırasında hangi bilgilerin unutulması veya hangi bilgilerin saklanması gerektiğini de öğrenirler (Dasan ve Panneerselvam, 2021).




**Şekil 3.4:** LSTM hücresi (Akcan ve Kaya, 2023)

Hücre durumu bir tür taşıma bandı gibi düşünülebilir. Hücre içindeki temel işlevi, göreceli bilgileri taşımaktır. Bu bilgiler, kapılar tarafından titizlikle düzenlenir, ardından bu bilgiler hücrenin sonuna ardından da diğer hücrelere doğru ilerler.

LSTM hücresinde ki unutma kapısı, tek katmanlı basit bir sinir ağı tarafından kontrol edilir. Bu kapının aktivasyonu aşağıdaki eşitlik ile hesaplanır.

$$f_t = \sigma(W[x_t, h_{t-1}, C_{t-1}] + b_f) \tag{3.1}$$

Bu formüldeki $x_t$ sıralı giriş verisini, $h_{t-1}$ önceki LSTM hücresinin çıktısını, $C_{t-1}$ önceki LSTM hücresinin hafızasını göstermektedir. $b_f$ bias vektörünü, W her girdi için ağırlık vektörlerini ve $\sigma$ aktivasyon fonksiyonu belirtmektedir.

En yaygın kullanılan aktivasyon fonksiyonları tanh ve sigmoid fonksiyonudur. Sigmoid aktivasyon fonksiyonuna ait formül aşağıda verilmiştir.

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{3.2}$$

Güncel bilgiler ve bir önceki katmandan gelen bilgiler, Sigmoid fonksiyonundan geçirilir. Fonksiyonun çıktısı 0'a ne kadar yakınsa o kadar unutulmakta; 1'e ne kadar yakınsa, o kadar hatırlanmaktadır. Unutma kapısında bilgilerin unutulması için $f_t$ fonksiyonundaki formülde W her girdi için ağırlık vektörü değeri 0 sıfır yapılır.

Giriş kapısı, LSTM hücresine yeni bilgiler sağlar ve bu yeni bilgilerin hücre durumunda saklanıp saklanmayacağına karar verilir. İlk olarak, unutma kapısında olduğu gibi Sigmoid fonksiyonu uygulanır ve hangi bilginin saklanacağına karar verilir. Daha sonra, ağı düzenlemek için tanh fonksiyonu kullanılarak değerler -1 ile 1 arasına indirgenir ve bu iki sonuç çarpılır. Giriş kapısı bir sinir ağı tarafından $i_t$ yeni hafızanın tanh aktivasyon fonksiyonu ile çarpımı ve önceki hafıza bloğunun $(f_t * C_{t-1})$ etkisi ile oluşturulan bölümdür. Bu işlemler için aşağıdaki eşitlikler kullanılır.

$$i_t = \sigma(W[x_t, h_{t-1}, C_{t-1}] + b_i) \tag{3.3}$$
$$C_t = f_t * C_{t-1} + i_t * \tanh([x_t, h_{t-1}, C_{t-1}]) + b_c \tag{3.4}$$

Çıkış kapısı mevcut LSTM hücresinin çıktısının üretildiği bölüm olup bir sonraki katmana gönderilecek değere karar verir ve bu değer tahmin için kullanılır. Bu işlem için aşağıdaki eşitlikler kullanılır.

$$\sigma_t = \sigma(W[x_t, h_{t-1}, C_t] + b_0) \tag{3.5}$$
$$h_t = o_t * \tanh(C_t) \tag{3.6}$$

LSTM birimleri arasındaki bağlantılar, verinin sıralı olarak işlenmesine olanak sağlar. Bu durum, LSTM'nin zaman kavramını kavramasına ve sunulan verilerin zamansal dinamiklerini anlamasına olanak tanıyan bir iç geri bildirim durumu yaratır. Bellek hücrelerindeki sürekli hata geri yayılımı, LSTM'nin zamanla ilgili problemlerde uzun süreli gecikmeleri anlamasına olanak sağlar. Bu nedenle, LSTM ağı, zaman serilerindeki önemli olaylar arasındaki bilinmeyen gecikmeleri dikkate alarak veri sınıflandırması, işlenmesi ve zaman serilerinin tahmini için son derece uygun bir modeldir. BiLSTM, iki ayrı gizli katmanda verileri ileri ve geri yönde işleyen LSTM hücrelerini kullanan çift yönlü bir RNN'dir. Hem önceki hem de gelecekteki verileri kullanarak ileri ve geri gizli dizilerin hesaplanmasına yardımcı olur (Londhe ve Atulkar, 2021).

### 3.2.3. Spektogram analizi

Mel – Frequency Spectrogram, ses işleme ve konuşma işleme alanında kullanılan bir özelliktir. Ses dalgalarının frekans içeriğini insan işitme sisteminin frekans tepkisine daha yakın bir şekilde temsil etmek için tasarlanmıştır. Mel – Frequency Spectrogram, sinyalin frekans içeriğini zamanla birleştirilmiş bir görüntü olarak sunar. İşte Mel – Frequency Spectrogram hakkında teorik bilgiler ve formüller:

**Mel Ölçeği**: Mel – Frequency Spectrogram, öncelikle "mel" ölçeği üzerinde çalışır. Mel ölçeği, insan işitme sisteminin frekans algısını modeller. Bu ölçek, frekansları daha insan dostu bir biçimde temsil eder. Frekans f (Hz) ve mel ölçeği m arasındaki dönüşüm şu şekildedir:

$$m = 2595.log_{10}(1 + \frac{f}{700}) \quad (3.7)$$

Bu dönüşüm, düşük frekanslarda daha yoğun bir ölçekleme sağlar.

**Filtre Bankları:** Mel – Frequency Spectrogram oluştururken, sinyalin frekans bileşenleri, mel ölçeğindeki belirli noktalarda bulunan filtre bankaları kullanılarak ölçülür. Her bir filtre bankası, belirli bir frekans aralığına karşılık gelir ve genellikle üst üste binen üçgen filtreden oluşur.

**Güç Spektrumu ve Logaritmik Dönüşüm:** Sinyalin güç spektrumu, filtrelere uygulanır ve ardından logaritmik bir dönüşüm uygulanarak Mel – Frequency Cepstral Katsayıları (MFCC) elde edilir. Bu işlem, insan işitme sisteminin daha logaritmik bir şekilde işlem yapmasını taklit eder.

**FFT (Hızlı Fourier Dönüşümü):** Sinyalin frekans içeriğini analiz etmek için sıklıkla FFT kullanılır. FFT, sinyalin zaman-frekans dönüşümünü sağlar ve bu, Mel – Frequency Spectrogram oluşturulurken kullanılan filtre bankalarının genliklerinin belirlenmesine yardımcı olur.
Formül: genellikle şu formülle ifade edilir:

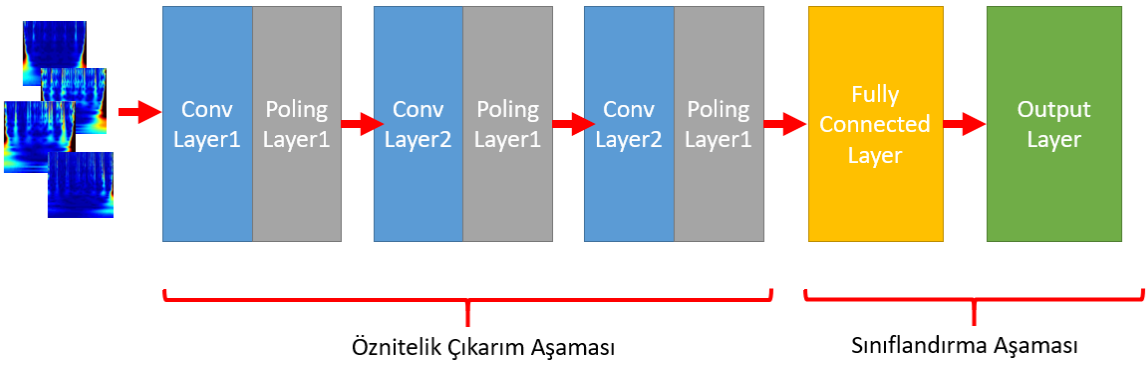
$$S_m(f,t) = \log(\sum_k P(f_k)H_k(f)\cos\left[\frac{2\pi f_k t}{T}\right]) \quad (3.8)$$

Bu formülde:
$S_m(f,t)$: Mel-Frequency Spectrogram matrisi
$P(f_k)$: FFT tarafından hesaplanan sinyalin güç spektrumu
$H_k(f)$ : k-inci filtre bankasının frekans tepkisi
$f_k$ : k-inci filtre bankasının merkez frekansı
$T$: Sinyalin toplam süresi
$t$: Zaman

### 3.2.4. Evrişimsel sinir ağları (CNN)

Evrişimli Sinir Ağı (CNN), genellikle derin öğrenme sinir ağları arasında yaygın olarak kullanılan bir mimari olup özellikle görüntü analizi alanında etkilidir. CNN, geleneksel yapay sinir ağlarına benzer bir temel yapıya sahiptir, ancak özel olarak geliştirilmiş evrişim işlemlerini kullanarak girdi olarak görüntüleri işler. Bu evrişim işlemleri, belirli bir görüntü ile önceden tanımlanmış filtreler arasındaki element – wise çarpımı içerir ve bu şekilde farklı öznitelikleri vurgular. Bu süreç, görüntülerden otomatik olarak öznitelikler çıkarmak için uygulanır. Geleneksel makine öğrenme yöntemlerinde öznitelikler genellikle manuel olarak belirlenirken, CNN ve benzeri derin öğrenme yöntemleri, bu öznitelikleri otomatik olarak evrişim gibi işlemlerle elde eder (Alom vd. 2019). CNN mimarileri genellikle iki temel aşamadan oluşur, ki bu aşamalar Şekil 3.5'te gösterilmiştir. İlk aşama, görüntülerden özniteliklerin çıkarıldığı evrişim aşamasını içerir. Bu aşamada,

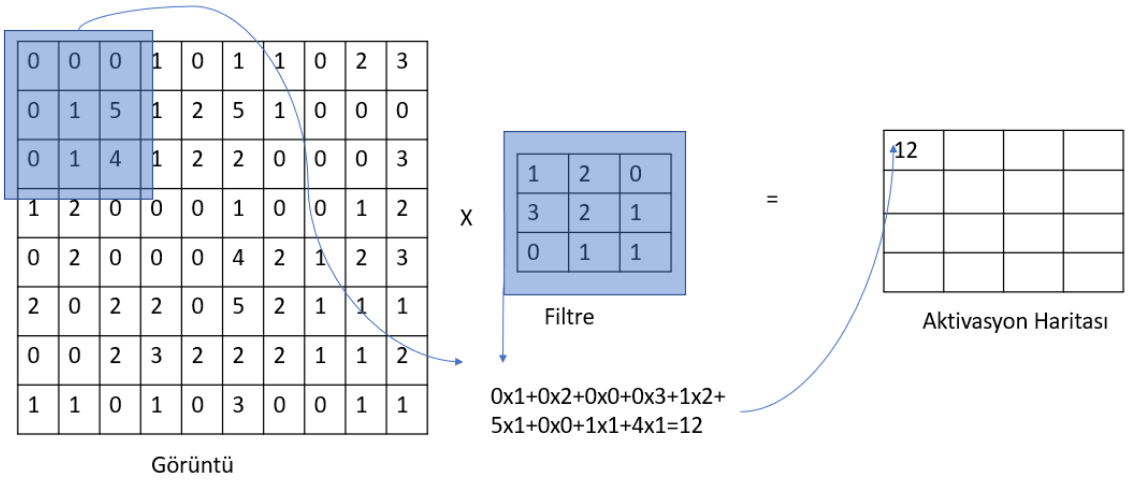
özel işlemlere tabi tutularak çeşitli öznitelikler elde edilir. Sınıflandırma aşaması ise öznitelikleri kullanarak görüntünün sınıfını belirler ve genellikle tam bağlantılı katmanları içerir. Bir CNN genellikle evrişim katmanları, havuzlama katmanları ve tam bağlantılı (FC) katmanları gibi katmanlardan oluşur ve bu katmanlar birleştirilerek CNN ağı oluşturulur. CNN ağlarının önemli bileşenleri arasında dropOut (bırakma) katmanları ve aktivasyon fonksiyonları da yer alır. Şekil 3.5'te bir CNN ağının genel yapısını göstermektedir (Alom vd. 2019). Bu yapı, özellikle görüntü işleme ve sınıflandırma görevleri için tasarlanmış güçlü bir derin öğrenme aracını temsil eder.




**Şekil 3.5:** CNN ağ yapısı

Bir CNN ağda bulunan katmanlar aşağıda kısaca açıklanmıştır.

**Convolutional Layer (Evrişimsel Katman):** Evrişimsel katman, bir CNN'nin en temel ve kritik bileşenlerinden biridir. Bu katmanda, girdi görüntülerinden özniteliklerin çıkarılma işlemi gerçekleşir. İki matris arasında element – wise çarpım operasyonuyla gerçekleştirilen bu işlemde, birinci matris giriş görüntüsünü temsil ederken, ikinci matris öznitelik çıkarma işlemi için kullanılan filtreleri içerir. Bu filtreler, giriş görüntüsünün boyutundan daha küçük olabilir, ancak derinlik (kanal sayısı) olarak aynı boyutta olur. Örneğin, bir RGB görüntü 3 kanaldan oluşurken, evrişim işlemi için kullanılan filtrelerin boyutu daha küçük olabilir, ancak her biri 3 kanalı kapsar. Evrişim katmanı genellikle görüntülerden köşeler, kenarlar ve diğer önemli özniteliklerin çıkarılmasında kullanılır. Filtreler, tüm giriş görüntüsü üzerinde kaydırılarak bu özniteliklerin çıkarılması sağlanır. Evrişim işlemi sonucunda, öznitelik haritaları elde edilir ve bu, evrişimsel katmanın çıkışı olarak adlandırılır. Aşağıda, bir görüntü ile 3x3 bir filtrenin evrişim işleminin nasıl gerçekleştiğini gösteren bir örnek verilmiştir (Shen, vd. 2019). Evrişimsel katman, bir CNN ağı içinde genellikle ilk katman olarak bulunur ve temel işlevi, giriş görüntülerinden farklı özniteliklerin çıkarılması ve temsil edilmesidir.



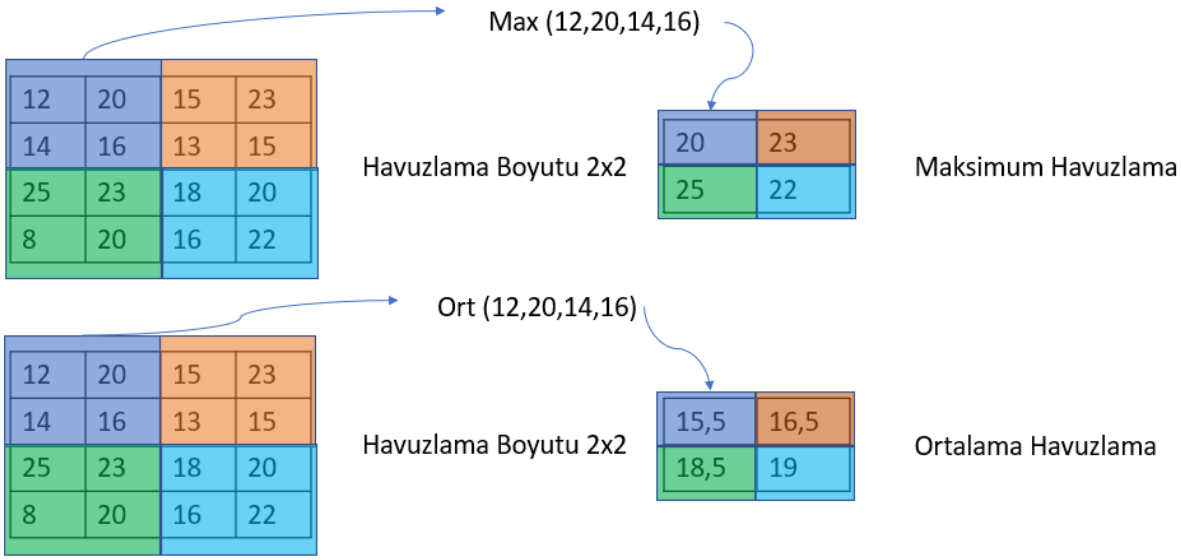

**Şekil 3.6:** Evrişim işlemi

**Havuzlama Katmanı**, genellikle Evrişimli Sinir Ağı'ndaki (CNN) evrişim katmanlarını takip eden ve genel olarak hesaplama maliyetini azaltmayı hedefleyen önemli bir bileşendir. Bu katman, öznitelik haritasının boyutunu belirli işlemler aracılığıyla küçülterek büyük boyutlu öznitelik haritalarının işlenmesinden kaynaklanan hesaplama yükünü hafifletir. Ayrıca, öznitelik haritasındaki gereksiz veya fazla detayları elemek suretiyle daha önemli özniteliklere odaklanmayı sağlar. Havuzlama katmanı, katmanlar arasındaki bağlantıları azaltarak bu işlemi gerçekleştirir.

Bu katmanın temel amaçları şunlardır:

**Hesaplama Maliyetini Azaltma:** Büyük boyutlu öznitelik haritalarının işlenmesi, hesaplama gücü açısından maliyetli olabilir. Havuzlama katmanı, öznitelik haritasının boyutunu küçülterek bu maliyeti azaltır.

**Önemli Özniteliklere Odaklanma:** Gereksiz ayrıntıları ve gürültüyü azaltarak, daha önemli özniteliklere odaklanmayı sağlar. Havuzlama katmanı, çeşitli havuzlama yöntemlerini kullanabilir, örneğin Max Havuzlama, Ortalama Havuzlama ve Toplam Havuzlama gibi. Şekil 3.7, en büyük ve ortalama havuzlama işlemlerine örnekler sunar ve bu işlemlerin CNN'nin daha yüksek seviyeli öznitelikler oluşturmasına ve hesaplama maliyetini düşürmesine nasıl yardımcı olduğunu görselleştirir (Zhou, Greenspan ve Shen, 2017).




**Şekil 3.7:** Maksimum ve Ortalama Havuzlama İşlemleri

**Tam Bağlantılı Katman (FC – Fully Connected):** Tam Bağlantılı Katman, Evrişim Katmanlarından elde edilen öznitelik haritalarının işlendiği ve sonuçların sınıflandırma veya çıkış katmanına iletilmesi amacıyla kullanılan önemli bir katmandır. Bu katman, geleneksel yapay sinir ağı (ANN) yapısına benzer ve genellikle iki aşamalı bir işlem içerir. Bunlar ağırlıklı toplam hesaplama ve aktivasyon fonksiyonlarının uygulanmasıdır.

FC katmanının temel özellikleri şu şekildedir:

**Ağırlıklı Toplam Hesaplama:** FC katmanındaki nöronlar, gelen özniteliklere belirli ağırlıklar ve eşik (bias) değerleri uygular. Bu ağırlıklar ve eşik değerleri, öğrenme süreci sırasında verilerden öğrenilir. Her nöron, girdi özniteliklerine ağırlıklı bir toplam işlemi uygular.

**Düzleştirme İşlemi:** Öznitelik haritaları matris şeklinde olduğundan, FC katmanına verilmeden önce vektör haline getirilir. Bu işlem, özniteliklerin boyutunu uygun hale getirir ve FC katmanının geleneksel yapısına uygun hale getirir.

**Öğrenme İşlemi:** FC katmanı, ağırlıkların ve eşik değerlerinin belirlenmesi için öğrenme sürecini içerir. Bu süreç, ağırlıkların girdi verilerinden en iyi şekilde öğrenilmesini sağlar ve ağın verilere uyum sağlamasını sağlar. FC katmanı, bir CNN ağının son katmanlarından birini oluşturur ve genellikle çıkış katmanından önce gelir. Sınıflandırma veya çıkış işlemi için kullanılır. Örneğin, bir giriş görüntüsünün hangi sınıfa ait olduğunu belirlemek için kullanılır. Bu katman, ağın öğrendiği öznitelikleri kullanarak sınıflandırma kararlarını alır. Sonuç olarak, FC katmanı, özniteliklerin son işlendiği ve sınıflandırma veya çıkış

kararlarının alındığı kritik bir katmandır. Bu katmanın ağırlıkları ve eşik değerleri, ağın öğrenme yeteneği sayesinde verilere uygun şekilde adapte edilir, böylece doğru sınıflandırma sonuçları elde edilir (Zhou, Greenspan ve Shen, 2017).

**Dropout (Bırakma):** Aşırı Öğrenme (Overfitting), yapay sinir ağlarının eğitimi sırasında karşılaşılan önemli bir sorundur. Bu durumda, ağ eğitim verilerini aşırı derecede ezberler ve bu verilere mükemmel şekilde uyar, ancak yeni veya test verileri üzerinde başarısız olur. Aşırı öğrenme sorunu, modelin genelleme yeteneğini kaybettiği ve eğitim verileri dışındaki verileri yanlış sınıflandırma eğiliminde olduğu bir durumu temsil eder. CNN ağlarında aşırı öğrenmeyi önlemek veya azaltmak için birkaç teknik kullanılır ve bunlardan biri Dropout (Bırakma) katmanıdır. Bu katman, FC (Tam Bağlantılı) katmanlarından sonra gelir ve ağın boyutunu küçültmek için kullanılır. Özellikle, bu katman, rastgele seçilen bir kısmını oluşturan nöronların çıkartılması işlemiyle çalışır. Örneğin, %30'luk bir dropout oranı kullanıldığında, her eğitim adımında FC katmanının %30'u rastgele seçilen nöronlarla kapatılır veya devre dışı bırakılır. Bu nöronların çıkartılması, ağın farklı öznitelikler ve ilişkiler öğrenmesini sağlar. Dropout katmanı, ağın ezberlemesi veya aşırı uyarlaması riskini azaltır çünkü her nöron eşit bir şekilde katmanın işlevini yerine getirmez ve bu nedenle ağın daha genel bir temsil öğrenmesine yardımcı olur. Dropout, bir tür düzenleme (regularization) tekniği olarak düşünülebilir, çünkü ağın aşırı öğrenmeye karşı daha dayanıklı hale gelmesini sağlar. Sonuç olarak, Dropout katmanı, aşırı öğrenmeyi azaltmak ve ağın daha iyi genelleme yapmasını sağlamak için kullanılan önemli bir düzenleme yöntemidir. Bu katman, ağın eğitimi sırasında rastgele olarak belirli nöronları devre dışı bırakarak ağın çeşitli öznitelikler ve ilişkiler öğrenmesine yardımcı olur (Shen, vd. 2019).

**Aktivasyon Fonksiyonları:** Derin öğrenme modellerinde, özellikle Evrişimsel Sinir Ağlarında (CNN), aktivasyon fonksiyonları önemli bir rol oynar. Bu fonksiyonlar, ağın giriş ve çıkış arasındaki karmaşık, doğrusal olmayan ilişkileri öğrenmesine ve ifade etmesine yardımcı olur. Aktivasyon fonksiyonları, ağın her katmanında kullanılır ve bu katmanlar arasındaki doğrusal olmayanları eklerler. İşte CNN ağlarında sıkça kullanılan bazı aktivasyon fonksiyonları:

**Sigmoid Fonksiyonu (Logistic Sigmoid):** Genellikle ikili sınıflandırma problemlerinde tercih edilir. Sigmoid fonksiyonu, girişi 0 ile 1 arasında bir değere dönüştürür, bu da çıkışın olasılık değeri olarak yorumlanabilir.

**Softmax Fonksiyonu:** Çok sınıflı sınıflandırma problemlerinde kullanılır. Softmax, bir sınıfın çıkışını diğer sınıfların çıkışlarına göre normalize eder, böylece her sınıfın olasılık dağılımını elde edebiliriz.

**ReLU (Rectified Linear Activation):** ReLU, en yaygın kullanılan aktivasyon fonksiyonlarından biridir. Girişi negatifse 0'a eşitler, pozitifse girişi doğrudan geçirir. Bu fonksiyon, ağın öğrenme hızını artırır ve aşırı öğrenmeyi azaltır.

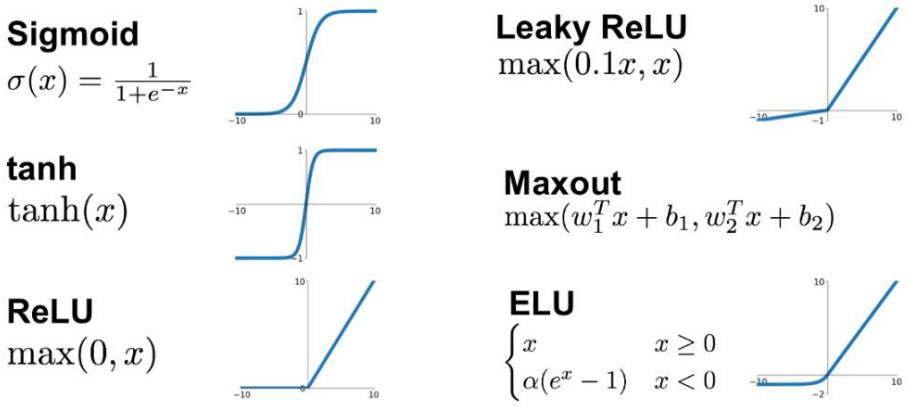
**tanh (Hyperbolic Tangent):** Tanh, girişi -1 ile 1 arasında bir değere dönüştüren bir aktivasyon fonksiyonudur. Sigmoid ile benzerdir, ancak çıkış aralığı daha geniştir.

**ELU (Exponential Linear Unit):** ELU, ReLU'nun sıfır olmayan negatif değerler için bir varyasyonudur. Bu fonksiyon, ağın aşırı öğrenmeyi azaltmasına ve daha iyi genelleme yapmasına yardımcı olabilir.

**Leaky ReLU:** Leaky ReLU, ReLU'nun negatif bölgesinde küçük bir eğim ekler. Bu, ReLU'nun bazen yaşadığı "ölü nöron" sorununu azaltmaya yardımcı olabilir.

**MaxOut:** MaxOut, girişi bir grup nöron arasında maksimum değere atan bir aktivasyon fonksiyonudur. Bu, ağın daha karmaşık ilişkileri öğrenmesine olanak tanır.

Bu aktivasyon fonksiyonları, ağın problem türüne, veriye ve eğitim sürecine bağlı olarak seçilir. Doğru aktivasyon fonksiyonunun kullanılması, ağın performansını büyük ölçüde etkileyebilir ve eğitim sürecini başarılı kılar. Bu nedenle, aktivasyon fonksiyonlarının seçimi önemlidir ve problem bağlamına uygun olarak yapılmalıdır (Zhou, Greenspan ve Shen, 2017). Aktivasyon fonksiyonlarına ait grafikler Şekil 3.8'de verilmiştir.




**Şekil 3.8:** Aktivasyon fonksiyonları (Kalyoncu, 2020)

### 3.2.5. Continues wavelet transform (CWT)

Ses gibi durağan olmayan, non-lineer karakteristiğe sahip sinyallerin zaman – frekans diyagramları genellikle Sürekli Dalgacık Dönüşümü (CWT – Continues Wavelet Transform) ve Kısa Zamanlı Fourier Dönüşümü (STFT – Short-Time Fourier Transform) yöntemleri kullanılarak elde edilmektedir. CWT yöntemi STFT yöntemine göre zaman ve frekans alanlarında üstün çözünürlüklere sahip olduğu için uygulamalarda daha çok tercih edilmektedir (Mateo ve Talavera, 2020). CWT yönteminde sinyaller sabit boyutlu pencere fonksiyonları yerine enerji odaklı temel fonksiyonlarla temsil edilir (Lee, Ratnam ve Ahmad, 2017). Bu üstün özelliklerinden dolayı CWT özellikle medikal ve ses gibi durağan olmayan sinyallerin zaman – frekans diyagramlarını oluşturmada çok etkili bir yöntemdir (Mohonta, Motin ve Kumar, 2022). CWT yönteminde sinyaller Daubechies, Morlet, Symlets ve Morse gibi dalgacık fonksiyonları kullanılarak dönüştürülür. Bu çalışmada, ses sinyali dönüşümlerinde diğer dalgacık fonksiyonlarına kıyasla daha yaygın olarak kullanılan Morse dalgacığı kullanılmıştır. Morse dalgacığı, ses sinyalleriyle olan morfolojik benzerliği ve daha iyi zaman – frekans lokalizasyonuna sahip olma özellikleri nedeniyle bu çalışmada tercih edilmiştir. CWT dönüşümü ile sinyallerdeki katsayılar aşağıda verilen formül ile hesaplanmaktadır:

$$F_{cwt}(\tau, b) = \frac{1}{\sqrt{|b|}} \int f(t) \psi^* \left(\frac{t-\tau}{b}\right) \tag{3.9}$$

Formülde, f(t) dönüştürülecek olan bir boyutlu sinyali göstermektedir. Kaydırma parametresi "$\tau$", ölçek parametresi “b” ve dalgacık fonksiyonu “ψ” sembolleri ile gösterilmiştir. Ölçek parametresi büyük olduğunda, sinyal genişler ve düşük frekanslarda kullanılır; ölçek parametresi küçük değerde olursa sinyal sıkıştırılır ve yüksek frekanslarda kullanılır. CWT dönüşümleri sonucu oluşan skalogram, sinyalin sürekli dalga dönüşümünün (CWT) mutlak değeridir ve zaman ve frekansın bir fonksiyonu olarak çizilir (Narin, 2022). Aslında sinyalin zaman kaydırma parametresi (τ) ve ölçek parametresindeki (b) enerji dağılımının bir ölçüsüdür. Enerji dağılımı aşağıda verilen formül ile bulunur.

$$E_{cwt}(\tau, b) = \sum_{\tau} \sum_{b} \left(F_{cwt}(\tau, b)\right)^2 \tag{3.10}$$

### 3.2.6. Performans ölçütleri

Dijital rakam tespiti için önerilen yaklaşımın performansını değerlendirmek için karışıklık matrisinden faydalanılır. Bu matris, veri setindeki gerçek çıkış etiketleri ile modellerin tahmin etiketlerinin yanlış ve doğru sayılarını bir tabloda gösterir.

**Tablo 3.1:** Karışıklık matrisi gösterimi

| | **Gerçek** | | |
|---|---|---|---|
| **Tahmin** | | **Gerçek** | **Yanlış** |
| | **Gerçek** | **True Pozitif (TP)** | **True Negatif (TN)** |
| | **Yanlış** | **False Negatif (FN)** | **False Pozitif (FP)** |

Burada;

**TP (True Pozitif):** Herhangi bir dijital rakam sınıfına ait etiketlenmiş bir ses dosyasının model tarafından aynı sınıfta doğru bir şekilde tahmin edilmesidir. Bu durum, doğru sınıflandırılan ses dosyalarının sayısını ifade eder.

**TN (True Negatif):** Herhangi bir dijital rakam sınıfına ait etiketlenmiş bir ses dosyasının model tarafından farklı bir sınıfta tahmin edilmesidir. Bu durum, yanlış sınıflandırılan ses dosyalarının sayısını ifade eder.

**FP (False Pozitif):** Doğru sınıflandırma olarak kabul edilmiş bir dijital rakam dışındaki bir sınıf ile etiketlenmiş ses dosyalarının, model tarafından kendi etiketlenmiş bir dijital rakam sınıfı olarak tahmin edilen ses dosyalarının sayısını belirtir.

**FN (False Negatif):** Doğru sınıflandırma olarak kabul edilmiş bir dijital rakam dışındaki bir sınıf ile etiketlenmiş ses dosyalarının, model tarafından farklı etiketlenmiş bir dijital rakam sınıfı olarak tahmin edilen ses dosyalarının sayısını belirtir.

Dijital rakam tespiti için önerilen yaklaşımın performansını değerlendirmek için aşağıdaki ölçütler kullanılmıştır.

$$Başarı = \frac{TP+TN}{TP+TN+FN+FP} \tag{3.11}$$

$$Kesinlik = \frac{TP}{TP+FP} \tag{3.12}$$

$$Hatırlatma = \frac{TP}{TP+FN} \tag{3.13}$$

$$F - \text{Ölçütü} = 2 * \frac{Kesinlik*Hatırlatma}{Kesinlik+Hatırlatma} \tag{3.14}$$

# 4. SONUÇLAR

Bu çalışma, ses sinyalleri üzerinden gerçekleştirilen bir rakam sınıflandırma prosedürünü ele almaktadır, bu doğrultuda üç farklı yöntem benimsenmiştir. İlk yaklaşım, MEL-SPEC katsayılarının çıkarılmasıyla elde edilen öznitelik matrisinin destek vektör makineleri (SVM) sınıflandırma algoritmasına giriş olarak sunulmasıdır. İkinci strateji, bir tekrarlayan yapay sinir ağı olan uzun ve kısa süreli belleğe sahip LSTM modelini içermektedir. Bu modelde, ses sinyallerine önceden kesikli dalgacık dönüşümü uygulanmakta ve ardından elde edilmiş olan zaman saçılma ortalaması öznitelik vektörleri kullanılmaktadır. Dalgacık dönüşümü, LSTM modeli için uygun girişleri düzenleyecek şekilde sıralı bir yapıya sahiptir. LSTM modeli, tek boyutlu sıralı veriler ile ilgili problemlerin çözümünde etkili bir araç olarak kabul edilmektedir. Bu model, klasik yapılarda kullanılan eğitilebilir sinir ağlarından farklı olarak kendi sinir ağlarını kullanan özel bir tekrarlayan sinir ağıdır ve bilgi akışını düzenleyebilen dahili kapılar adı verilen mekanizmalara sahiptir. Üçüncü yaklaşım ise, ses sinyallerinin dalgacık dönüşümü ile elde edilen Skolegram görüntülerine dönüştürülerek tasarlanan bir Evrişimli Sinir Ağı (CNN) ile sınıflandırılmasını içermektedir. Bu strateji, özellikle görsel verilerin analizi konusunda başarılı olan CNN modellerinin ses sinyalleri üzerinde de etkili olabileceğini öne sürmektedir. Çalışmanın veri seti, yaşları 14 – 17 arasında değişen 25 öğrenciden toplanmış olup toplamda 2500 kayıttan oluşmaktadır. Veri seti, farklı eğitim ve test oranlarıyla çeşitli deneyler gerçekleştirmek üzere kullanılmıştır. Bu kapsamlı analiz ve çeşitli metodolojilerin uygulanması, ses sinyalleri temelli rakam sınıflandırma alanında yeni perspektifler ve yöntemlerin geliştirilmesine katkı sağlamayı amaçlamaktadır.

### 4.1. LSTM Modeli ile Gözlenen Başarılar

İlk aşamada LSTM ile sınıflandırma işlemi gerçekleştirilmiştir. Kullanılan LSTM modelin yapısı ve içerdiği özellikleri aşağıdaki verilmiştir.

**Tablo 4.1:** LSTM model yapısı ve hiper parametre değerleri

| **LSTM Model** | |
|---|---|
| **Parametre** | **Değer** |
| Katmanlar | SeguenceInput<br>lstm (128)<br>bi-lstm(128)<br>fullyConnected<br>SoftMax<br>Classification |
| Giriş Sayısı | 1/sinyal uzunluğu |
| Gizli Katman Boyutu | 128 |
| Çıkış Katman Boyutu | 3 |
| Batch Boyutu | 27 |
| Optimizer | Adam |
| Öğrenme Oranı | 0.001 |
| Öğrenme Oranı gecikme Faktörü | 0.5 |

Bu araştırmada, çeşitli mimarilere sahip Long Short-Term Memory (LSTM) modelleri tasarlanabilmiştir. Kullanılan LSTM modeli, giriş, lstm, bi-lstm, FC, Softmax ve Classification katmanlarından oluşmaktadır. Batch size parametresi (27), her iterasyonda kullanılan eğitim verilerinden seçilen alt kümenin boyutunu belirtir. Maksimum Epoch sayısı, eğitim sürecinde kullanılan epoch sayısını belirler. Burada, bir epoch, tüm veri kümesinin eğitim algoritmasına gösterilme sürecini temsil eder. Epoch sayısı seçimi, öğrenme sürecini etkileyebileceği için önemlidir; örneğin, epoch sayısını düşürmek, örneklerden öğrenmeme durumuna yol açabilirken, yükseltmek ise çok fazla öğrenmeye (ezberlemeye) sebep olabilir. Bu parametreler, en yüksek doğruluğu sağlamak amacıyla optimize edilmiştir. Çalışmada, en iyi sınıflandırma sonuçlarını elde etmek için en optimum değerler olarak batch boyutu 27 ve öğrenme oranı 0.0001 seçilmiştir. Maksimum epoch sayısı ise eğitim – test setlerine bağlı olarak değişiklik göstermektedir. Dijital rakam sınıflandırma farklı eğitim – test oranlarına göre gerçekleştirilmiş olup çıkan sonuçlar Tablo 4.2'de ayrıntılı olarak sunulmuştur. Bu çeşitli eğitim – test oranları altında elde

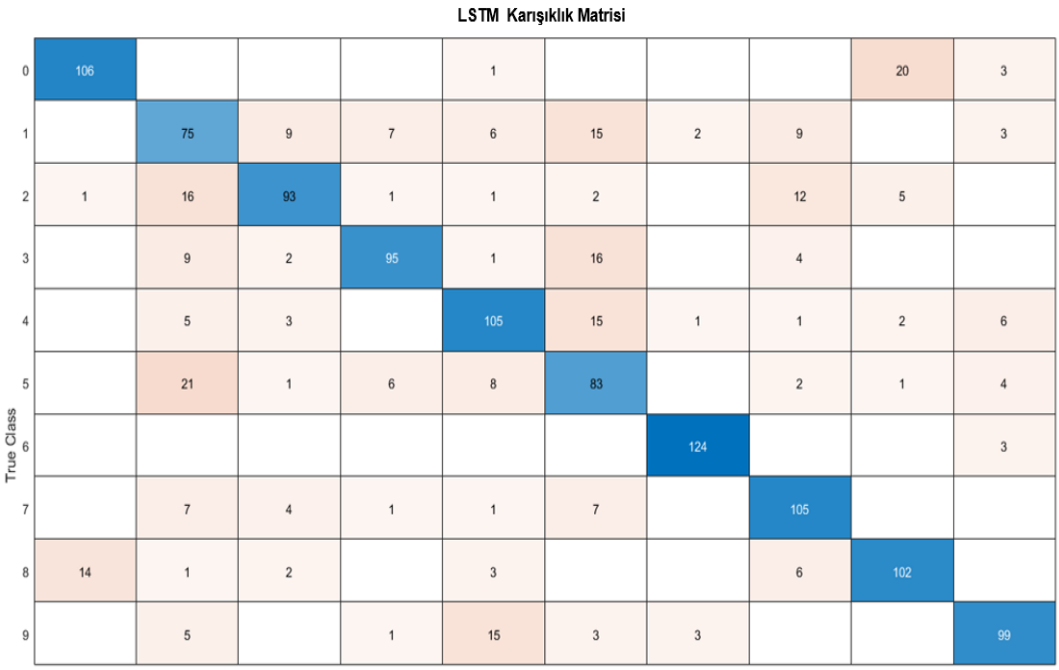
edilen sonuçlar, önerilen LSTM model mimarisi ve parametrelerinin sınıflandırma performansını optimize etme yeteneğini göstermektedir.

**Tablo 4.2:** LSTM ile Gözlenen Başarı Oranları

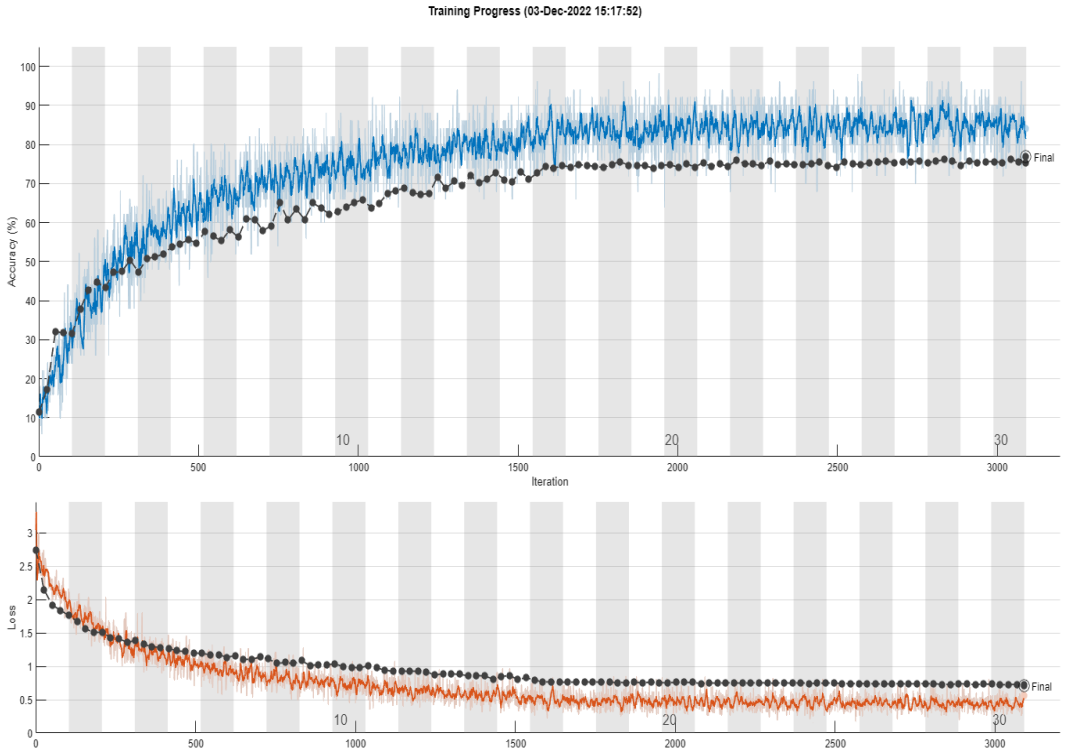
| **Veri Seti Eğitim-Test** | **Başarı** | **Kesinlik** | **Hatırlatma** | **F-Ölçütü** |
|---|---|---|---|---|
| %50-50 Oranı | 0.7741 | 0.7802 | 0.7737 | 0.7758 |
| %60-40 Oranı | 0.7500 | 0.7627 | 0.7499 | 0.7530 |
| %70-30 Oranı | 0.7280 | 0.7347 | 0.7282 | 0.7295 |
| %80-20 Oranı | 0.6757 | 0.6870 | 0.6758 | 0.6787 |

Tablo 4.2'ye yönelik analizde, en üstün başarı oranının %77.41 ile %50-50 eğitim-test oranına ait olduğu gözlemlenmektedir. Bu durum, test örnek sayısının azaldıkça başarının düşüş gösterdiği bir eğilim sergilemektedir.

Eğitim – test veri seti %50 – 50 ile elde edilen karışıklık matrisi Şekil 4.1'de detaylı bir şekilde sunulmuştur. Karışıklık matrisi (confusion matrix), bir sınıflandırma modelinin performansını değerlendirmek için kullanılan bir araçtır. Genellikle çok sınıflı (multiclass) sınıflandırma problemlerinde kullanılır ve gerçekleştirilen sınıflandırma işleminin sonuçlarını özetler. Matris, tahmin edilen sınıfların gerçek sınıflarla karşılaştırılmasını gösterir. Ayrıca, LSTM ağının performans grafiği de Şekil 4.2'de yer almaktadır. Kayıp ve performans grafikleri, bir makine öğrenimi modelinin eğitim sürecindeki başarısını görsel olarak ifade eden önemli araçlardır. Kayıp grafiği, modelin eğitim ve doğrulama veri setleri üzerindeki kayıp değerlerinin iterasyonlar boyunca nasıl değiştiğini gösterir; bu, modelin ne zaman konverge olduğunu veya aşırı öğrenmeye başladığını anlamak için kullanılır. Performans grafiği ise modelin doğruluk oranını eğitim ve doğrulama veri setleri üzerindeki değişimlerle gösterir, böylece modelin genel sınıflandırma başarısını değerlendirmek mümkün olur. Bu grafikler, modelin eğitim sürecindeki performans eğilimlerini ve olası sorunları hızlı bir şekilde görselleştirerek, modelin iyileştirilmesi için rehberlik sağlar. Şekil 4.2, LSTM ağının her iterasyondaki başarı ve kayıp değerleri grafiksel olarak gösterilmiştir. Bu grafik, modelin öğrenme sürecindeki performansını izlemeye ve değerlendirmeye olanak tanır. İterasyonlar arasındaki başarı ve kayıp değişimleri, modelin eğitim verileri üzerindeki uyumunu ve genel performansını daha ayrıntılı bir şekilde anlamamıza yardımcı olabilir.




**Şekil 4.1:** LSTM için Karışıklık Matrisi




**Şekil 4.2:** LSTM ağın performans ve kayıp grafikleri

### 4.2. SVM ile Gözlenen Başarı Oranları

Bu bölümde ses sinyalleri üzerinde Dalgacık Saçılma (Wavelet Scattering) ile elde edilen öznitelikler kullanılarak SVM ile sınıflandırma işlemi gerçekleştirilmiştir. Dalgacık Zaman Saçılma (WTS – Wavelet Time Scattering), ses sinyallerinden öznitelik çıkarma amacı güden bir işlemdir. Bu yöntem, ses sinyallerinin frekans ve zaman uzayında geniş bir kapsamda temsil edilmesini sağlayarak, makine öğrenmesi yöntemleri ile ses verilerinden bilgi çıkarmayı hedefler. WTS, ses sinyallerinin dalgacık dönüşümü (Wavelet Transform) ve zaman saçılma (Time Scattering) işlemlerini birleştirir. Dalgacık dönüşümü, ses sinyalinin farklı frekanstaki bileşenlerini ortaya çıkaran bir işlemdir. Bu, ses sinyalinin frekans uzayında temsil edilmesini sağlar. Zaman saçılma ise, sinyalin zaman içindeki değişimlerini yakalar. Bu iki işlem bir araya getirildiğinde, dalgacık zaman saçılma (WTS) katsayıları elde edilir. Bu katsayılar, ses sinyalinin frekans ve zaman içindeki özelliklerini temsil eden önemli özniteliklerdir. Makine öğrenmesi yöntemleri, bu öznitelikleri kullanarak ses sinyallerini sınıflandırabilir ve tanıyabilir. Örneğin, bu katsayılar üzerinden elde edilen öznitelik matrisi, bir destek vektör makinesi (SVM) veya derin öğrenme modeli gibi algoritmaların girişi olarak kullanılabilir. Bu şekilde, makine öğrenmesi yöntemleri, ses verilerinden çıkarılan öznitelikleri kullanarak basit rakamların (digits) tanınması gibi ses tabanlı tanıma görevlerini gerçekleştirebilir. WTS, ses sinyallerinin karmaşıklığını ve içerdikleri bilgiyi daha etkili bir şekilde temsil etme yeteneği nedeniyle bu tür uygulamalarda tercih edilen bir öznitelik çıkarma yöntemidir. Özellikle ses sinyallerinin sınıflandırılması için kullanılan dalga deseni dağılımı, frekans özelliklerini vurgulamak ve sınıflandırma performansını artırmak amacıyla kullanılmıştır. SVM modeline ait parametreler Tablo 4.3'te belirtilmiştir.

**Tablo 4.3:** SVM parametreler

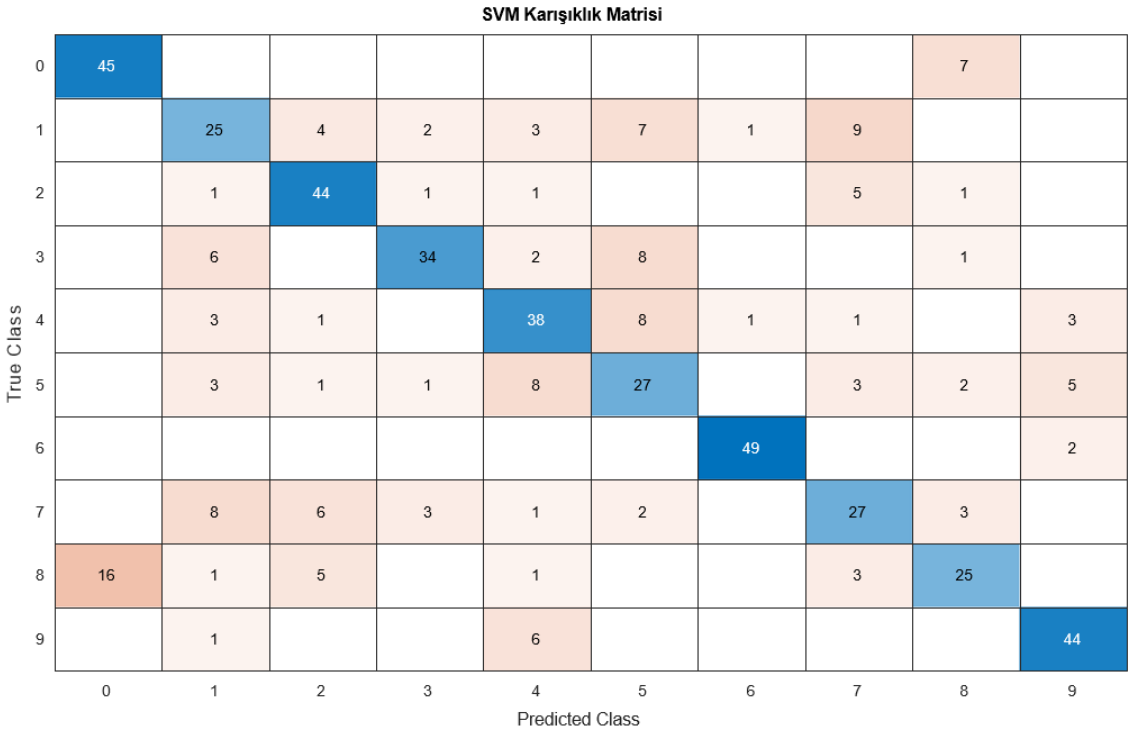
| **Parametre** | **Değer** |
|---|---|
| KernelFunction | Polynomial |
| PolynomialOrder | 2 |
| KernelScale | Otomatik |
| BoxConstraint | 1 |
| Standardize | true |

SVM modeli, ses sinyallerini sınıflandırmak amacıyla bir dalga deseni dağılımı tabanlı özellik çıkarımı ve çokgenel (polynomial) çekirdekli bir SVM sınıflandırıcı kullanmaktadır. Model, ses sinyallerinin uzunluklarını analiz ederek veri setini oluşturur, dalga deseni dağılımı uygular ve elde edilen özellik matrisleri üzerinde SVM sınıflandırıcıyı eğitip değerlendirir. SVM'nin parametreleri, polynomial çekirdek kullanımını, çekirdek ölçeğini otomatik belirlemeyi, sınırlama (BoxConstraint) uygulamayı ve giriş verilerini standartlaştırmayı içerir. Bu parametrelerin doğru seçimi, modelin performansını etkiler ve genel sınıflandırma doğruluğunu belirler. SVM modeli 5 katlı çapraz geçerlilik testine göre uygulanmış denemeler farklı eğitim test oranları ile gerçekleştirilmiş olup başarı oranları Tablo 4.4'te verilmiştir.

**Tablo 4.4:** SVM ile Gözlenen Başarı Oranları

| **Veri Seti Eğitim-Test** | **Başarı** | **Kesinlik** | **Hatır-latma** | **F-Ölçütü** |
|---|---|---|---|---|
| %50-50 Oranı | 67.4455 | 0.67357 | 0.67357 | 0.65704 |
| %60-40 Oranı | 67.8016 | 0.67686 | 0.67686 | 0.66187 |
| %70-30 Oranı | 68.5233 | 0.68474 | 0.68474 | 0.6684 |
| %80-20 Oranı | 69.5146 | 0.69371 | 0.69371 | 0.67385 |

Tablo 4.4'e bakıldığında en yüksek başarı %80-20 eğitim-test veri seti için gözlenmiştir. Başarı %69.5146 olarak elde edilmiştir. Eğitim seti oranı arttıkça başarı oranının arttığı görülmektedir. Tablo 4.4'te diğer kesinlik, hatırlatma ve F1-Ölçütü gibi performans değerleri de verilmiştir. SVM ile gözlenen en yüksek başarı oranı için karışıklık matrisi Şekil 4.3'te verilmiştir.



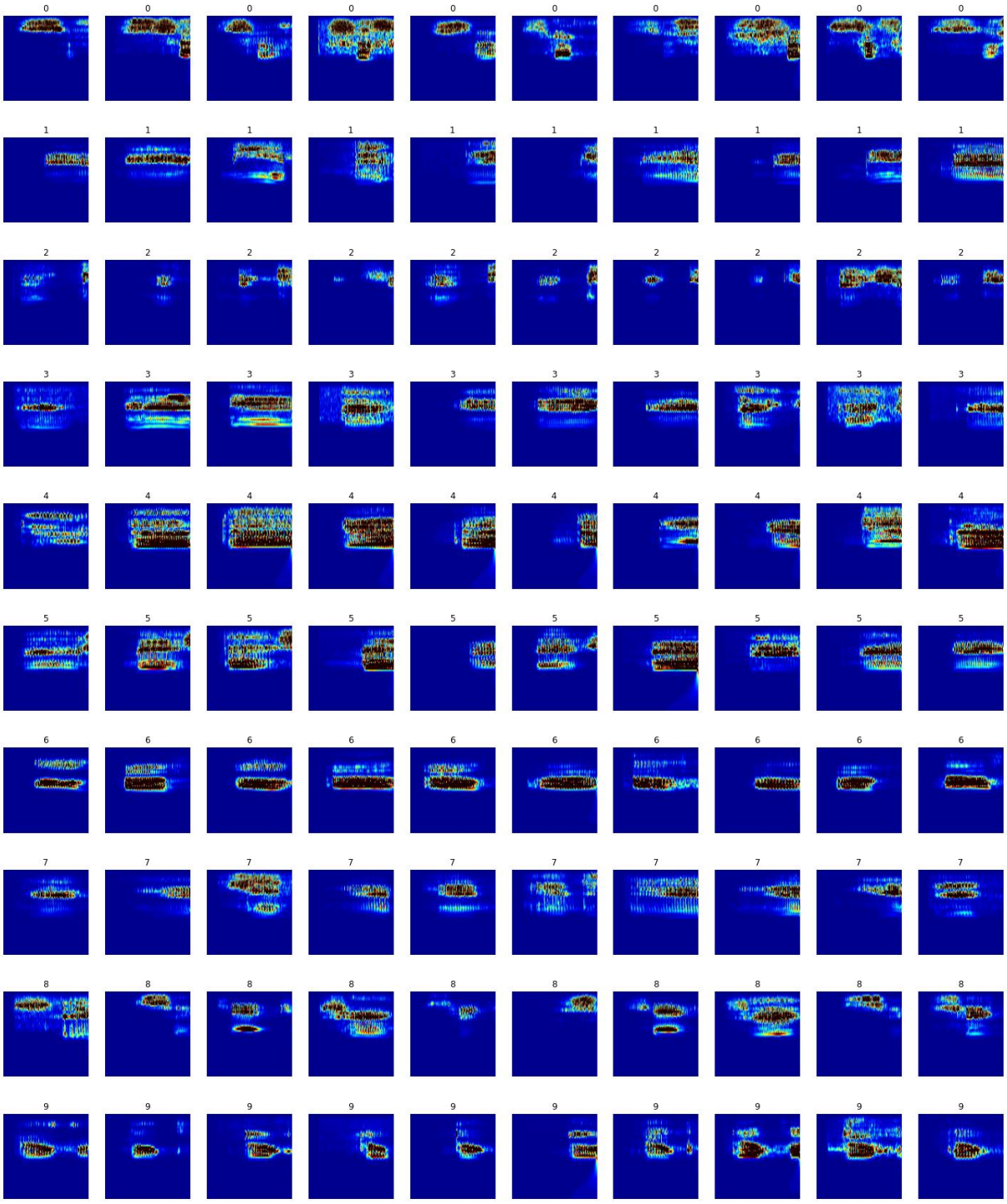

**Şekil 4.3:** SVM için Karışıklık Matrisi

### 4.3. CNN ile Gözlenen Başarı Sonuçları

Son sınıflandırma yaklaşımında, ses sinyallerinin karakteristik özelliklerini daha etkili bir şekilde çıkarmak amacıyla Continuous Wavelet Transform (CWT) uygulanarak elde edilen skolograma görüntüleri kullanılmıştır. Bu skolograma görüntüleri, önceden belirlenen frekans ve zaman çözünürlüklerine sahip olup, ses sinyallerinin zaman – frekans uzayında daha ayrıntılı bir temsilini sağlar. Elde edilen skolograma görüntüleri daha sonra özel olarak tasarlanan bir evrişimsel sinir ağı (CNN) derin öğrenme modeli kullanılarak sınıflandırılmıştır. CNN ağı, bir dizi evrişim, aktivasyon ve pooling katmanları içeren bir mimariye sahiptir. Denemeler ve doğrulama sürecinin ardından belirlenen bu CNN ağı mimarisi, Ses Tabanlı Dijital Tanıma uygulamalarındaki sınıflandırma görevlerinde optimal performans sergileme kabiliyeti açısından seçilmiştir. Mimarinin detayları, bu çalışmanın Şekil 4.4'ünde ayrıntılı bir şekilde sunulmaktadır. Ses sinyallerine CWT uygulandıktan sonra elde edilen skolograma örnek görüntüler Şekil 4.5'te verilmiştir.

| |
|---|
| imageInputLayer |
| convolution2dLayer |
| batchNormalizationLayer |
| reluLayer |
| maxPooling2dLayer |
| convolution2dLayer |
| batchNormalizationLayer |
| reluLayer |
| maxPooling2dLayer |
| convolution2dLayer |
| batchNormalizationLayer |
| reluLayer |
| maxPooling2dLayer |
| convolution2dLayer |
| batchNormalizationLayer |
| reluLayer |
| maxPooling2dLayer |
| dropoutLayer |
| fullyConnectedLayer |
| softmaxLayer |
| classificationLayer |

**Şekil 4.4:** Skogram görüntülerini sınıflandırmak için kullanılan CNN ağı



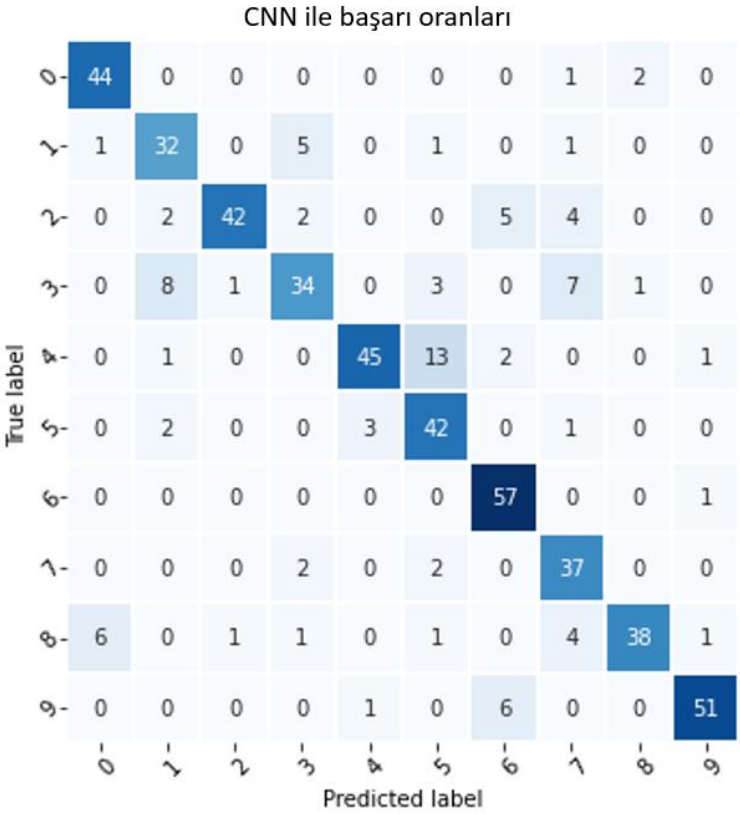
**Şekil 4.5:** Her sınıf için örnek skologram görüntüler

CWT+CNN kullanarak farklı eğitim – test oranlarına sahip veri setleri için elde edilen sonuçlar Tablo 4.5'te verilmiştir.

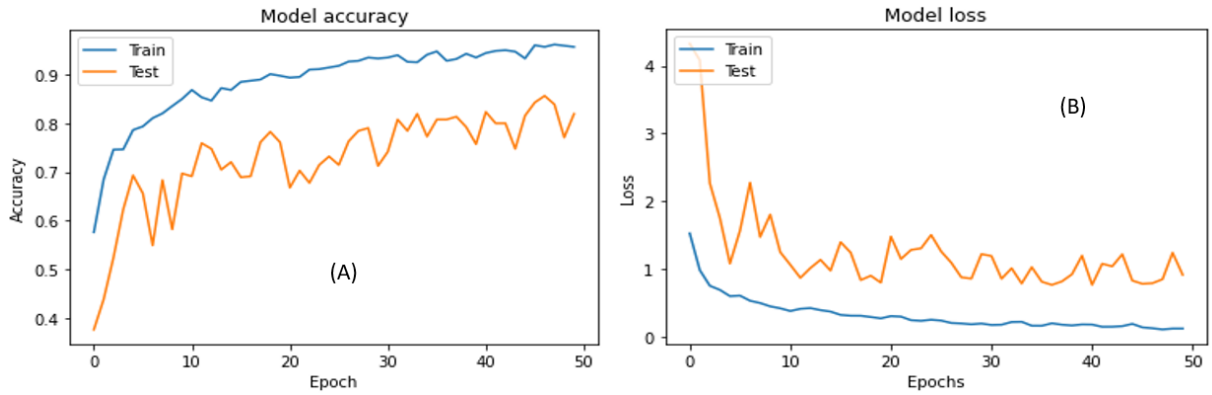
**Tablo 4.5:** CNN ile Gözlenen Başarı Oranları

| **Veri Seti Eğitim-Test** | **Başarı** | **Kesinlik** | **Hatırlatma** | **F-Ölçütü** |
|---|---|---|---|---|
| %50-50 Oranı | 0.7001 | 0.76 | 0.70 | 0.68 |
| %60-40 Oranı | 0.7534 | 0.81 | 0.75 | 0.75 |
| %70-30 Oranı | 0.750 | 0.78 | 0.75 | 0.74 |
| %80-20 Oranı | 0.8194 | 0.84 | 0.82 | 0.82 |

Tablo 4.5'e bakıldığında CNN ile diğer iki modele göre daha yüksek başarı oranlarının tespit edildiği görülmektedir. CNN ile en yüksek başarı %80-20 eğitim-test oranı için %81.94 olarak gözlenmiştir. Eğitim seti oranı arttıkça başarı oranının yükseldiği görülmektedir. CNN modeline ait karışıklık matrisi Şekil 4.6'da, eğitim ve test için başarı performans ile hata grafikleri Şekil 4.7'de verilmiştir. Başarı ve kayıp grafikleri, bir makine öğrenimi modelinin eğitim sürecindeki performansını görselleştiren kritik araçlardır. Başarı grafiği, modelin eğitim ve doğrulama veri setleri üzerindeki doğruluk oranını izlerken, kayıp grafiği ise modelin bu veri setleri üzerindeki tahmin hatalarını ölçer. İdeal durumda, başarı oranı yükselirken kayıp düşer, ancak bu her zaman böyle olmayabilir. Grafikler, modelin aşırı öğrenme, underfitting veya diğer performans sorunlarını belirlemede yardımcı olur. Eğitim sürecindeki eğilimler ve eğri arasındaki farklar, modelin genel performansını değerlendirerek, gerekirse düzeltici önlemler alınmasına olanak tanır. Başarı ve kayıp grafikleri, modelin veri setine uyumunu izleyerek, eğitim sürecinin etkiliğini anlamak ve optimize etmek adına kritik öneme sahiptir.




**Şekil 4.6:** CNN için karışıklık matrisi




**Şekil 4.7:** CNN başarı ve hata grafikleri

### 4.4. Yöntemlerin Karşılaştırılması

Bu çalışmada 3 farklı yaklaşım ile ses sinyallerinden dijit tanıma işlemi gerçekleştirilmiştir. Bu yaklaşımların her dijit için performansları Tablo 4.6'da verilmiştir. Tablo 4.6'ya bakıldığında en yüksek başarı oranı CNN modeli ile gözlenmiştir. Daha sonra LSTM modeli daha başarılı bulunmuştur. Genel olarak derin öğrenme metotlarının daha iyi olduğu görülmektedir. CNN modeli için Dijit bazlı bakıldığında en yüksek başarılar "Altı(6)" ses sinyali için gözlenmektedir. Daha sonra "Bir(1)" ve "Dokuz(9)" için yüksek başarılar gözlenmiştir. En düşük başarı oranı "Üç(3)" , "Dört(4)" ve "Sekiz(8)" için elde edilmiştir.

**Tablo 4.6:** Her sınıf için başarı oranları

| **Dijit** | **SVM** | **LSTM** | **CNN** |
|---|---|---|---|
| 0 (Sıfır) | 0,865 | 0,815 | 0,936 |
| 1 (bir) | 0.500 | 0,595 | 0,800 |
| 2 (iki) | 0,830 | 0,710 | 0,763 |
| 3 (Üç) | 0,666 | 0,748 | 0,629 |
| 4 (Dört) | 0,690 | 0,761 | 0,725 |
| 5 (Beş) | 0,540 | 0,659 | 0,875 |
| 6 (Altı) | 0,960 | 0,976 | 0,982 |
| 7 (Yedi) | 0,540 | 0,840 | 0,902 |
| 8 (Sekiz) | 0,500 | 0,797 | 0,730 |
| 9 (Dokuz) | 0,863 | 0,786 | 0,879 |
| Ortalama | **0,717** | **0,769** | **0,822** |

# 5. TARTIŞMA

İnsanlar, diğer tüm canlılardan farklı olarak konuşabilme yeteneğine sahiptir. Konuşma, insanlar arasındaki iletişimin en temel yoludur. İnsanlar, düşüncelerini, duygularını, isteklerini ve birçok diğer durumu konuşarak başkalarına iletebilirler. Bununla birlikte, insanlar zekasını kullanarak yenilikler yapabilme gücüne sahiptirler. Bunu fark ettikten sonra insanlar hem ilgi duydukları ve merak ettikleri alanlarda çalışmaya başlamışlar hem de insanların yaşamını daha kolay, daha iyi ve daha rahatlatıcı hale getirecek çözümler üretmişlerdir. Zamanla, insanların teknolojiye olan bağımlılıkları artmış ve her şeyin en kolay çözümünü istemeleriyle birlikte birçok insan konuşma tanıma üzerinde çalışmaya başlamıştır (Chowdhury vd. 2021). Konuşma tanıma teknolojileri, günümüzde hızla gelişen ve birçok alanda etkisini gösteren önemli bir araştırma ve uygulama alanı olmuştur. Bu teknolojiler, ses sinyallerini analiz ederek konuşma dilini anlama ve çeşitli görevleri yerine getirme yeteneğine sahiptir. Yapılan çalışmalarda, özellikle SVM (Destek Vektör Makineleri), LSTM (Uzun Kısa Süreli Hafıza), ve CNN (Evrişimli Sinir Ağları) gibi makine öğrenimi modellerinin konuşma tanıma süreçlerindeki performanslarına odaklanılmıştır (Wazir ve Chuah, 2019). SVM'nin kullanımı, ses sinyallerinin karmaşıklığını ele alma ve sınıflandırma görevlerini gerçekleştirme konusunda etkili olduğunu göstermektedir. Bu model, belirli bir eğitim veri setiyle çalışarak genel bir performans sergiler. Ancak, SVM'nin çoklu özelliklere adapte olma konusundaki sınırlamaları, özellikle karmaşık konuşma modelleriyle başa çıkmak için yeterli olmayabilir. LSTM, uzun vadeli bağımlılıkları modelleme yeteneğiyle öne çıkan bir reküran sinir ağı türüdür. Ses sinyallerindeki zaman bağımlılıklarını ele alarak, özellikle dilin zaman içindeki yapısını anlama konusunda etkilidir. Ancak, eğitim sürecinde karşılaşılan zorluklar ve kaynak yoğunluğu, LSTM'nin geniş ölçekli konuşma tanıma uygulamalarında kullanımını sınırlayabilir. CNN, özellikle görüntü tanıma alanında büyük başarı elde etmiş bir modeldir. Ses sinyallerini görsel özelliklere dönüştürme kabiliyeti, konuşma tanıma görevlerinde etkili olmasını sağlar. Evrişimli sinir ağları, özellikle frekans ve zaman özelliklerini başarılı bir şekilde çıkarabilir, ancak büyük veri setleri gerektirebilir. Genel olarak, yapılan karşılaştırmalar, CNN'nin diğer modellere kıyasla daha yüksek başarı oranlarına sahip olduğunu göstermektedir. CNN'nin ses sinyallerindeki karmaşıklıkları daha iyi anlama yeteneği, geniş ölçekli konuşma tanıma sistemlerinin geliştirilmesinde büyük bir avantaj sağlar. Ancak, SVM ve LSTM gibi modeller de belirli koşullarda başarılı sonuçlar verebilmektedir (Chowdhury vd. 2021). Konuşma tanıma teknolojilerindeki bu makine öğrenimi modellerinin karşılaştırılması, uygulama bağlamında belirli avantajları ve sınırlamaları

ortaya koymaktadır. Bu modellerin seçimi, uygulanacak konuşma tanıma görevine, kullanılabilir veri setine ve sistem gereksinimlerine bağlı olarak yapılmalıdır. Gelecekteki araştırmalar, bu modellerin daha karmaşık ve gerçek dünya senaryolarında nasıl performans gösterdiklerini daha ayrıntılı bir şekilde incelemelidir. Bu sayede, konuşma tanıma teknolojilerinin daha etkili ve geniş kapsamlı kullanımı mümkün olabilir.

Bu çalışmada, ses sinyalleri üzerinden gerçekleştirilen dijital rakam sınıflandırma işlemi için üç farklı yöntem olan SVM, LSTM ve CNN modellerinin performansı değerlendirilmiştir. İlk yöntem olarak SVM, MEL-SPEC katsayılarından elde edilen öznitelik matrisini kullanarak sınıflandırma yapmıştır. İkinci yöntemde, LSTM modeli, önceden kesikli dalgacık dönüşümü uygulanmış ses sinyallerini kullanarak başarılı sonuçlar elde etmiştir. Üçüncü yöntemde ise, ses sinyallerinin Continuous Wavelet Transform (CWT) ile elde edilen skolograma görüntülerine dayalı CNN modeli kullanılmıştır. Her bir yöntemin performansı, farklı eğitim-test oranları altında incelenmiştir. LSTM modeli, %50-50 eğitim – test oranında en yüksek başarıyı göstermiştir. SVM'nin en yüksek başarı oranını ise %80-20 eğitim – test oranında elde ettiği görülmüştür. Ancak, en üstün başarı oranları CNN modeli tarafından kaydedilmiştir, özellikle %80-20 eğitim – test oranında %81.94'lük bir başarı elde edilmiştir (Tablo 4.2, Tablo 4.4 ve Tablo 4.5). Dijit bazlı olarak incelendiğinde, CNN modelinin genel başarı oranının en yüksek olduğu görülmüştür. Özellikle "Altı (6)" dijiti için %98.2'lik bir başarı oranı elde edilmiştir. Diğer dijitler arasında farklı başarı oranları gözlemlenmiş, örneğin "Bir (1)" ve "Dokuz (9)" için yüksek performans sergilenirken, "Üç (3)", "Dört (4)" ve "Sekiz (8)" için daha düşük başarı oranları elde edilmiştir (Tablo 4.6). Genel olarak, derin öğrenme modelleri olan LSTM ve CNN, SVM'ye kıyasla daha yüksek başarı oranları göstermiştir. Bu sonuçlar, özellikle ses tabanlı dijital tanıma uygulamalarında derin öğrenme modellerinin etkili bir şekilde kullanılabileceğini göstermektedir. Ancak, her bir modelin avantajları ve sınırlamaları göz önüne alınarak, kullanılacak modelin uygulama bağlamına, veri setine ve sistem gereksinimlerine bağlı olarak seçilmesi gerekmektedir.

## 5.1. Kısıtlamalar ve Hedefler

Bu çalışma, ses sinyalleri üzerinden gerçekleştirilen dijital rakam sınıflandırma yöntemlerini değerlendirmekle birlikte, bazı kısıtlamalara da tabidir. İşte bu kısıtlamalardan bazıları:

**Veri Seti Kısıtlamaları:** Çalışmada kullanılan veri seti, 14 -17 yaşları arasındaki öğrencilerden toplanmış ses sinyallerinden oluşmaktadır. Bu nedenle, genel nüfusun geniş bir

yelpazesini temsil etmeyebilir ve elde edilen sonuçların genelleme yeteneğini sınırlayabilir.

**Sınıf Dengesizliği:** Veri setindeki dijitler arasında sınıf dengesizliği olabilir. Özellikle, bazı dijitlerin diğerlerine göre daha fazla temsil edilmesi, model performansını etkileyebilir ve sonuçları yanıltıcı kılabilir.

**Eğitim – Test Oranı:** Çalışmada farklı eğitim – test oranları kullanılarak modellerin performansı değerlendirilmiştir. Ancak, bu oranların seçimi, genel performans üzerinde etkili olabilir ve optimal oranın belirlenmesi için daha fazla analize ihtiyaç duyulabilir.

**Hiperparametre Seçimi:** Her bir modelin performansı, belirli hiperparametre değerleri kullanılarak değerlendirilmiştir. Bu hiperparametrelerin seçimi, genel başarı üzerinde etkili olabilir ve daha geniş bir hiperparametre araştırması yapılması gerekebilir.

**Ses Sinyali Çeşitliliği:** Çalışmada kullanılan ses sinyalleri, belirli bir yaş grubundan toplandığı için genel ses sinyali çeşitliliği sınırlı olabilir. Farklı yaş gruplarından veya çeşitli koşullardan elde edilen ses sinyalleriyle yapılan çalışmalar, sonuçların genellenmesini artırabilir.

**Ses Ortamı Koşulları:** Ses sinyallerinin toplandığı ortamın gürültü seviyeleri, akustik özellikleri ve diğer çevresel faktörler, model performansını etkileyebilir. Çalışmada bu faktörlerin standartlaştırılması veya kontrol edilmesine dair detaylı bilgi verilmemişse, bu durum sonuçların güvenilirliğini etkileyebilir.

**Model Karşılaştırma Yöntemleri:** Çalışmada kullanılan model karşılaştırma yöntemleri belirli metriklere dayalıdır. Farklı metrik seçimleri, sonuçların yorumlanmasını etkileyebilir, bu nedenle farklı metriklerin kullanılmasının sonuçların daha kapsamlı bir değerlendirmesine katkı sağlayabileceği unutulmamalıdır.

Hedeflere bakıldığında; ses sinyalleri üzerinden gerçekleştirilen dijital rakam sınıflandırma konusunda belirlenebilecek çeşitli hedefler bulunabilir. Bu hedefler, çalışmanın kapsamına, amaçlarına ve uygulama bağlamına bağlı olarak şekillenebilir. İşte bu tür bir çalışmada belirlenebilecek potansiyel hedeflerden bazıları:

**Sınıflandırma Doğruluğunun Artırılması:** Temel bir hedef, ses sinyallerini doğru bir şekilde sınıflandırmak için kullanılan modelin performansını artırmaktır. Bu, farklı özellik çıkarma yöntemlerinin, model mimarilerinin veya hiperparametrelerin test edilerek en etkili kombinasyonunun belirlenmesini içerebilir.

**Genelleme Yeteneğinin İyileştirilmesi:** Modelin, farklı yaş gruplarından veya farklı çevresel koşullardan gelen ses sinyallerini daha iyi tanıyabilme yeteneğinin artırılması hedefi. Bu, modelin gerçek dünya koşullarında daha etkili bir şekilde kullanılabilmesini sağlamayı amaçlar.

**Daha Çeşitli Veri Setlerinin İncelenmesi:** Veri setinin çeşitliliğini artırmak, modelin daha genel ve çeşitli ses sinyalleri üzerinde başarılı olma olasılığını artırabilir. Bu, farklı konuşmacıların, farklı dil ve aksanların veya farklı kayıt cihazlarının ses sinyallerini içeren veri setleri üzerinde çalışmayı içerebilir.

**Hassasiyetin ve Özgüllüğün Ayarlanması:** Modelin belirli dijitleri ayırt etme yeteneğini artırmak veya azaltmak için hassasiyet ve özgüllük ayarlarının optimize edilmesi. Bu, modelin belirli dijitleri daha belirgin bir şekilde tanıması veya yanlış pozitif veya yanlış negatif sonuçları azaltması için kullanılabilir.

**Gerçek Zamanlı İşleme Yeteneğinin İyileştirilmesi:** Modelin gerçek zamanlı ses sinyallerini sınıflandırma yeteneğini artırmak amacıyla hesaplama hızını ve işleme süresini optimize etme hedefi.

**Çeşitli Sınıflandırma Modellerinin Karşılaştırılması:** SVM, LSTM ve CNN gibi farklı sınıflandırma modellerinin karşılaştırılması. Hangi modelin belirli bir uygulama bağlamında daha iyi performans gösterdiğini anlamak için bu modellerin karşılaştırılması ve analizi.

**Dijit Bazında Performansın İncelenmesi:** Her bir dijitin sınıflandırma performansının ayrı ayrı değerlendirilmesi. Hangi dijitlerin daha kolay veya daha zor tanınabilir olduğunu anlamak, modelin güçlü ve zayıf yönlerini belirlemeye yardımcı olabilir.

**Kullanılabilirlik ve Uygulanabilirlik Testleri:** Modelin pratikte ne kadar kullanılabilir olduğunu ve gerçek dünya uygulamalarında nasıl performans gösterdiğini değerlendirmek amacıyla gerçek kullanıcılar veya belirli bir uygulama senaryosu içeren kullanılabilirlik testleri gerçekleştirme hedefi.

## 6. KAYNAKLAR


Akcan, E., & Kaya, Y. (2023). A new approach for remaining useful life prediction of bearings using 1D-ternary patterns with LSTM. *Journal of the Brazilian Society of Mechanical Sciences and Engineering,* 45(7), 378.

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., Hasan, M., Van Essen, B. C., Awwal, A. A., & Asari, V. K. (2019). A state-of-the-art survey on deep learning theory and architectures. *electronics*, *8*(3), 292.

Alotaibi, Y. A. (2005). Investigating spoken Arabic digits in speech recognition setting. *Information Sciences*, *173*(1-3), 115-139.

Baldominos, A., Saez, Y., & Isasi, P. (2019). A survey of handwritten character recognition with mnist and emnist. *Applied Sciences*, *9*(15), 3169.

Byun, S.-W., & Lee, S.-P. (2021). A study on a speech emotion recognition system with effective acoustic features using deep learning algorithms. *Applied Sciences*, *11*(4), 1890.

Chandio, A., Shen, Y., Bendechache, M., Inayat, I., & Kumar, T. (2021). AUDD: audio Urdu digits dataset for automatic audio Urdu digit recognition. *Applied Sciences*, *11*(19), 8842.

Chauhan, K., Jani, S., Thakkar, D., Dave, R., Bhatia, J., Tanwar, S., & Obaidat, M. S. (2020). Automated machine learning: The new wave of machine learning. *2nd International Conference on Innovative Mechanisms for Industry Applications*-ICIMIA 2020, Bangalore, India, 205-212.

Chen, C., Hua, Z., Zhang, R., Liu, G., & Wen, W. (2020). Automated arrhythmia classification based on a combination network of CNN and LSTM. *Biomedical Signal Processing and Control*, *57*, 101819.

Chowdhury, M. F., Sultana, Z., Jahan, N., & Alavi, S. H. (2021). Conversion of Bengali speech to text using long short-term memory (LSTM), Doctoral dissertation, *Brac University Computer Science and Engineering*, Dhaka, Bangladesh, 21-27.

Das, S., Imtiaz, M. S., Neom, N. H., Siddique, N., & Wang, H. (2023). A hybrid approach for Bangla sign language recognition using deep transfer learning model with random forest classifier. *Expert Systems with Applications*, *213*, 118914.

Dasan, E., & Panneerselvam, I. (2021). A novel dimensionality reduction approach for ECG signal via convolutional denoising autoencoder with LSTM. *Biomedical Signal Processing and Control*, *63*, 102225.

Eroğlu, A. ve Kaya, Y., 2023, Derin öğrenme ile Türkçe ses işaretlerinden rakam tanıma, *Uluslararası Bilişim Kongresi- IIC2023*, Batman, 80-90.

Fu, J., Chiba, Y., Nose, T., & Ito, A. (2020). Automatic assessment of English proficiency for Japanese learners without reference sentences based on deep neural network acoustic models. *Speech Communication*, *116*, 86-97.

Halageri, A., Bidappa, A., Arjun, C., Sarathy, M. M., & Sultana, S. (2015). Speech recognition using deep learning. *Int. J. Comput. Sci. Inf. Technol*, *6*(3), 3206-3209.

Han, J. H., Bae, K. M., *Hong*, S. K., Park, H., Kwak, J.-H., Wang, H. S., Joe, D. J., Park, J. H., Jung, Y. H., & Hur, S. (2018). Machine learning-based self-powered acoustic sensor for speaker recognition. *Nano Energy*, *53*, 658-665.

Hu, X., Yuan, S., Xu, F., Leng, Y., Yuan, K., & Yuan, Q. (2020). Scalp EEG classification using deep Bi-LSTM network for seizure detection. *Computers in Biology and Medicine*, *124*, 103919.

Jalalvand, A., Triefenbach, F., Demuynck, K., & Martens, J.-P. (2015). Robust continuous digit recognition using reservoir computing. *Computer Speech & Language*, *30*(1), 135-158.

Kalyoncu, S. (2020). Deep learning networks for stock market analysis, Yüksek lisans tezi, *İstanbul Sabahattin Zaim Üniversitesi Bilgisayar Bilimleri ve Mühendisliği*, İstanbul, 21-24.

Kanchana, S. (2022). Empirical Investigation for Predicting Depression from Different Machine Learning Based Voice Recognition Techniques. *Evidence-Based Complementary and Alternative Medicine* Volume 2022, Article ID, 9, 1-9.

Karpagavalli, S., & Chandra, E. (2016). A review on automatic speech recognition architecture and approaches. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 9(4), 393-404.

Khan, L. U., Saad, W., Han, Z., Hossain, E., & Hong, C. S. (2021). Federated learning for internet of things: Recent advances, taxonomy, and open challenges. *IEEE Communications Surveys & Tutorials*, *23*(3), 1759-1799.

Koolagudi, S. G., Murthy, Y. S., & Bhaskar, S. P. (2018). Choice of a classifier, based on properties of a dataset: case study-speech emotion recognition. *International Journal of Speech Technology*, *21*(1), 167-183.

Ks, D. R., Rudresh, M., & Shashibhushsan, G. (2021). Comparative performance analysis for speech digit recognition based on MFCC and vector quantization. *Global Transitions Proceedings*, *2*(2), 513-519.

Lee, W., Ratnam, M., & Ahmad, Z. (2017). Detection of chipping in ceramic cutting inserts from workpiece profile during turning using fast Fourier transform (FFT) and continuous wavelet transform (CWT). *Precision Engineering*, *47*, 406-423.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31-57.

Lokesh, S., Malarvizhi Kumar, P., Ramya Devi, M., Parthasarathy, P., & Gokulnath, C. (2019). An automatic tamil speech recognition system by using bidirectional recurrent neural network with self-organizing map. *Neural Computing and Applications*, *31*, 1521-1531.

Londhe, A. N., & Atulkar, M. (2021). Semantic segmentation of ECG waves using hybrid channel-mix convolutional and bidirectional LSTM. *Biomedical Signal Processing and Control*, *63*, 102162.

Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Mateo, C., & Talavera, J. A. (2020). Bridging the gap between the short-time Fourier transform (STFT), wavelets, the constant-Q transform and multi-resolution STFT. *Signal, Image and Video Processing*, *14*(8), 1535-1543.

Mohonta, S. C., Motin, M. A., & Kumar, D. K. (2022). Electrocardiogram based arrhythmia classification using wavelet transform with deep learning model. *Sensing and Bio-Sensing Research*, *37*, 100502.

Muralikrishna, H., & Ananthakrishna, T. (2013). HMM based isolated Kannada digit recognition system using MFCC. *international conference on advances in computing, communications and informatics*- ICACCI 2013, Mysore, India, 730-733.

Narin, A. (2022). Detection of focal and non-focal epileptic seizure using continuous wavelet transform-based scalogram images and pre-trained deep neural networks. *Irbm*, *43*(1), 22-31.

Nisar, S., Shahzad, I., Khan, M. A., & Tariq, M. (2017). Pashto spoken digits recognition using spectral and prosodic based feature extraction. *Ninth International Conference on Advanced Computational Intelligence*- ICACI 2017, Doha, Qatar, 74-78.

Ozer, I. (2021). Pseudo-colored rate map representation for speech emotion recognition. *Biomedical Signal Processing and Control*, *66*, 102502.

Padmanabhan, J., & Johnson Premkumar, M. J. (2015). Machine learning in automatic speech recognition: A survey. *IETE Technical Review*, *32*(4), 240-251.

Patel, N., Patel, S., & Mankad, S. H. (2022). Impact of autoencoder based compact representation on emotion detection from audio. *Journal of Ambient Intelligence and Humanized Computing*, 1-19.

Ramo, F. M., & Younis, A. N. (2021). A Novel Approach to Spoken Arabic Number Recognition Based on Developed Ant Lion Algorithm. *International Journal of Computing*, 20(2), 270-275.

Renjith, S., Joseph, A., & Anish Babu K.K. (2013). Isolated digit recognition for Malayalam-An application perspective. *International Conference on Control Communication and Computing*- ICCC 2013, Thiruvananthapuram, India, 190-193.

Rogerson-Revell, P. M. (2021). Computer-assisted pronunciation training (CAPT): Current issues and future directions. *Relc Journal*, *52*(1), 189-205.

Rudresh, M., Latha, A., Suganya, J., & Nayana, C. (2017). Performance analysis of speech digit recognition using cepstrum and vector quantization. *International conference on electrical, electronics, communication, computer, and optimization techniques*- 2017 ICEECCOT, Mysuru, India 1-6.

Rukwong, N., & Pongpinigpinyo, S. (2022). An Acoustic Feature-Based Deep Learning Model for Automatic Thai Vowel Pronunciation Recognition. *Applied Sciences*, *12*(13), 6595.

Sen, O., & Roy, P. (2022). A Novel Bangla Spoken Numerals Recognition System Using Convolutional Neural Network. *International Conference on Machine Intelligence and Emerging Technologies*- 2022, Switzerland, 344-357.

Sharan, R. V. (2020). Spoken digit recognition using wavelet scalogram and convolutional neural networks. *IEEE Recent Advances in Intelligent Computational Systems*-RAICS 2020, Thiruvananthapuram, India, 101-105.

Sharan, R. V., Berkovsky, S., & Liu, S. (2020). Voice command recognition using biologically inspired time-frequency representation and convolutional neural networks. *42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society*- EMBC 2020, Montreal, QC, Canada, 998-1001.

Sharan, R. V., & Moir, T. J. (2019). Acoustic event recognition using cochleagram image and convolutional neural networks. *Applied Acoustics*, *148*, 62-66.

Sharmin, R., Rahut, S. K., & Huq, M. R. (2020). Bengali spoken digit classification: A deep learning approach using convolutional neural network. *Procedia Computer Science*, *171*, 1381-1388.

Shen, S., Gu, K., Chen, X.-R., Yang, M., & Wang, R.-C. (2019). Movements classification of multi-channel sEMG based on CNN and stacking ensemble learning. *Ieee Access*, *7*, 137489-137500.

Tailor, J. H., Rakholia, R., Saini, J. R., & Kotecha, K. (2022). Deep learning approach for spoken digit recognition in Gujarati language. *International Journal of Advanced Computer Science and Applications*, *13*(4). 1-6.

Taufik, D., & Hanafiah, N. (2021). Autovat: An automated visual acuity test using spoken digit recognition with MEL frequency cepstral coefficients and convolutional neural network. *Procedia Computer Science*, *179*, 458-467.

Tian, H., Li, X., Wei, Y., Ji, S., Yang, Q., Gou, G.-Y., Wang, X., Wu, F., Jian, J., & Guo, H. (2022). Bioinspired dual-channel speech recognition using graphene-based electromyographic and mechanical sensors. *Cell Reports Physical Science*, *3*(10). 1-15.

Wang, M., Yan, Z., Wang, T., Cai, P., Gao, S., Zeng, Y., Wan, C., Wang, H., Pan, L., & Yu, J. (2020). Gesture recognition using a bioinspired learning architecture that integrates visual data with somatosensory data from stretchable sensors. *Nature Electronics*, *3*(9), 563-570.

Wangchuk, K., Riyamongkol, P., & Waranusast, R. (2021). Real-time Bhutanese sign language digits recognition system using convolutional neural network. *Ict Express*, *7*(2), 215-220.

Wazir, A. S. M. B., & Chuah, J. H. (2019). Spoken Arabic digits recognition using deep learning. *IEEE International Conference on Automatic Control and Intelligent Systems*- I2CACIS 2019, Selangor, Malaysia, 339-344.

Wu, J., Chua, Y., Zhang, M., Li, H., & Tan, K. C. (2018). A spiking neural network framework for robust sound classification. *Frontiers in neuroscience*, *12*, 2-5.

Yadav, H., Shah, P., Gandhi, N., Vyas, T., Nair, A., Desai, S., Gohil, L., Tanwar, S., Sharma, R., & Marina, V. (2023). CNN and Bidirectional GRU-Based Heartbeat Sound Classification Architecture for Elderly People. *Mathematics*, *11*(6), 1-3.

Yin, Y., Zheng, X., Hu, B., Zhang, Y., & Cui, X. (2021). EEG emotion recognition using fusion model of graph convolutional neural networks and LSTM. *Applied Soft Computing*, *100*, 7-9.

Zada, B., & Ullah, R. (2020). Pashto isolated digits recognition using deep convolutional neural network. *Heliyon*, *6*(2), 1-6.

Zerari, N., Abdelhamid, S., Bouzgou, H., & Raymond, C. (2018). Bi-directional recurrent end-to-end neural network classifier for spoken Arab digit recognition. *2nd International Conference on Natural Language and Speech Processing*- ICNLSP 2018, Algiers, Algeria, 1-6.

Zhao, J., Mao, X., & Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control*, *47*, 312-323.

Zheng, W., Yu, J., & Zou, Y. (2015). An experimental study of speech emotion recognition based on deep convolutional neural networks. *international conference on affective computing and intelligent interaction*- ACII 2015, Xi'an, China, 827-831.

Zhou, K., Greenspan, H., & Shen, D. (Eds.)., 2017. Deep learning for medical image analysis. *Academic Press*. United Kingdom, 8-444.

Zhou, Z., Chen, K., Li, X., Zhang, S., Wu, Y., Zhou, Y., Meng, K., Sun, C., He, Q., & Fan, W. (2020). Sign-to-speech translation using machine-learning-assisted stretchable sensor arrays. *Nature Electronics*, *3*(9), 571-578.