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

TÜRKİYE'DE KONUŞULAN DİLLERİN DERİN ÖĞRENME YÖNTEMLERİYLE SINIFLANDIRILMASI

HAZIRLAYAN

MÜGE ERTEKİN

YÜKSEK LİSANS TEZİ

ANKARA - 2025

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

TÜRKİYE'DE KONUŞULAN DİLLERİN DERİN ÖĞRENME YÖNTEMLERİYLE SINIFLANDIRILMASI

HAZIRLAYAN

MÜGE ERTEKİN

YÜKSEK LİSANS TEZİ

TEZ DANIŞMANI

PROF. DR. HAMİT ERDEM

ANKARA - 2025

BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Elektrik Elektronik Mühendisliği Anabilim Dalı Elektrik Elektronik Mühendisliği Tezli Yüksek Lisans Programı çerçevesinde Müge ERTEKİN tarafından hazırlanan bu çalışma, aşağıdaki jüri tarafından Yüksek Lisans Tezi olarak kabul edilmiştir.

Tez Savunma Tarihi: 24 / 06 / 2025

Tez Adı: Türkiye'de Konuşulan Dillerin Derin Öğrenme Yöntemleriyle Sınıflandırılması

Tez Jüri Üyeleri (Unvanı, Adı- Soyadı, Kurumu)		**İmza**
Prof. Dr. Hamit ERDEM	Başkent Üniversitesi	……
Doç. Dr. Selda GÜNEY	Başkent Üniversitesi	……
Dr. Öğr. Üyesi KORAY AÇICI	Ankara Üniversitesi	……

ONAY

Prof. Dr. Dilek ÇÖKELİLER SERDAROĞLU
Fen Bilimleri Enstitüsü Müdürü
Tarih: … /… / 2025

BAŞKENT ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZ ÇALIŞMASI ORİJİNALLİK RAPORU

Tarih: … / … / 2025

Öğrencinin Adı, Soyadı : Müge ERTEKİN

Öğrencinin Numarası : 22320221

Anabilim Dalı : Elektrik-Elektronik Mühendisliği Ana Bilim Dalı

Programı : Elektrik-Elektronik Mühendisliği Tezli Yüksek Lisans Programı

Danışmanın Unvanı/Adı, Soyadı : Prof. Dr. Hamit ERDEM

Tez Başlığı: Türkiye'de Konuşulan Dillerin Derin Öğrenme Yöntemleriyle Sınıflandırılması

Yukarıda başlığı belirtilen Yüksek Lisans tez çalışmamın; Giriş, Ana Bölümler ve Sonuç Bölümünden oluşan, toplam 62 sayfalık kısmına ilişkin, 24/06/2025 tarihinde tez danışmanım tarafından Turnitin adlı intihal tespit programından aşağıda belirtilen filtrelemeler uygulanarak alınmış olan orijinallik raporuna göre, tezimin benzerlik oranı %4'dir. Uygulanan filtrelemeler:

1. Kaynakça hariç
2. Alıntılar hariç
3. Beş (5) kelimeden daha az örtüşme içeren metin kısımları hariç

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

Öğrenci İmzası:…….

ONAY

Tarih: … / … / 2025

Öğrenci Danışmanı:

Prof. Dr. Hamit ERDEM

……..

TEŞEKKÜR

Yüksek lisans öğrenimim süresince bilgi, tecrübe ve rehberliğiyle her zaman yanımda olan, çalışmamın her aşamasında değerli katkılarını esirgemeyen, bilimsel yaklaşımı ve akademik desteğiyle beni yönlendiren saygıdeğer danışmanım Prof. Dr. Hamit Erdem'e en içten teşekkürlerimi sunarım.

Hayatım boyunca bana inanan, destekleyen ve her koşulda yanımda olan sevgili annem Yasemin ALTINGEYİK'e, babam Haluk ALTINGEYİK'e ve ablam Aylin ALTINGEYİK'e; eğitim hayatımda bana her zaman güç veren, sabrıyla ve sevgisiyle yanımda duran kıymetli eşim Berkay ERTEKİN'e teşekkürlerimi sunarım.

ÖZET

Müge ERTEKİN

TÜRKİYE'DE KONUŞULAN DİLLERİN DERİN ÖĞRENME YÖNTEMLERİYLE SINIFLANDIRILMASI

Başkent Üniversitesi Fen Bilimleri Enstitüsü

Elektrik Elektronik Mühendisliği Anabilim Dalı

2025

Günümüzde konuşan dil tanıma sistemleri; insan-makine etkileşimi, robotik uygulamalar, çok dilli iletişim ve güvenlik gibi çeşitli alanlarda yaygın olarak kullanılmaktadır. Bu sistemler, otonom robotik etkileşim, çok dilli müşteri hizmetleri ve otomatik çeviri gibi uygulamalarda etkin rol oynamaktadır. Ayrıca, güvenlik ve adli bilişimde anonim ses kayıtlarının dilini belirlemek; göçmen profilleme ve istihbarat analizleri açısından önem taşımaktadır. Bu tez çalışmasında, Türkiye'de yaygın olarak konuşulan diller olarak, Türkiye'ye yerleşen ve Türkiye'yi sıkça ziyaret eden yabancı uyruklu bireylerin konuştukları diller esas alınarak bir veri seti oluşturulmuştur. Bu kapsamda hazırlanan veri setinde; Türkçe, Kürtçe, Arapça, Rusça, Farsça ve Almanca dillerine ait ses kayıtları yer almaktadır. Konuşulan dili otomatik olarak tanıma problemi, bir sınıflandırma problemi olarak ele alınmıştır. Bu problemi çözmek amacıyla farklı derin öğrenme mimarileri kullanılmıştır. İlk olarak, spektral özelliklerin zamansal örüntülerini yakalamada etkili olan İki Yönlü Uzun Kısa Süreli Bellek (BiLSTM) ağları uygulanmıştır. Ayrıca, mekânsal örüntüleri çıkarmada başarılı olan Evrişimsel Sinir Ağları (CNN) ile birlikte, dil tanıma problemlerinde son yıllarda dikkat çekici başarılar gösteren Transformer tabanlı Encoder mimarisi de değerlendirilmiştir. Bunun yanı sıra, CNN ve self-attention mekanizmalarını aynı yapıda birleştirerek hem yerel hem de küresel bağlamı yakalayabilen Conformer Encoder mimarisi de çalışmada uygulanmış ve karşılaştırmalı analiz yapılmıştır. Modelin performansını değerlendirmek amacıyla doğruluk (accuracy), duyarlılık (sensitivity/recall) ve diğer ilgili performans metrikleri kullanılmıştır.

ANAHTAR KELİMELER: Konuşulan Dil Tanıma, Derin Öğrenme, Çok Dilli Sınıflandırma, BiLSTM, Conformer Encoder

ABSTRACT

Müge ERTEKİN

CLASSIFICATION OF LANGUAGES SPOKEN IN TÜRKİYE USING DEEP LEARNING METHODS

Başkent University Institute of Science

Department of Electric Electronic Engineering

2025

Today, spoken language identification systems are widely used in various domains such as human-machine interaction, robotics, multilingual communication, and security. These systems play an important role in applications like autonomous robotic interaction, multilingual customer services, and automatic translation. Moreover, identifying the language of anonymous voice recordings has become increasingly significant in fields such as security, forensic informatics, immigration profiling, and intelligence analysis. In this thesis, a dataset was created based on the languages commonly spoken in Türkiye, including those spoken by foreign nationals who have settled in or frequently visit the country. The dataset includes voice recordings in Turkish, Kurdish, Arabic, Russian, Persian, and German. The task of automatically recognizing the spoken language is addressed as a classification problem. To solve this problem, various deep learning architectures were implemented. First, Bidirectional Long Short-Term Memory networks were used due to their effectiveness in capturing the temporal patterns of spectral features. Additionally, Convolutional Neural Networks, which are successful in extracting spatial patterns, and Transformer-based encoder architectures, which have shown remarkable performance in recent language recognition tasks, were also evaluated. Furthermore, the Conformer encoder architecture, which combines CNN and self-attention mechanisms to capture both local and global contexts simultaneously, was applied and compared with other models. To evaluate model performance, accuracy, recall (sensitivity), and other relevant performance metrics were utilized.

KEYWORDS: Spoken Language Identification, Deep Learning, Multilingual Classification, BiLSTM, Conformer Encoder

İÇİNDEKİLER

TABLOLAR LİSTESİ

Sayfa

ŞEKİLLER LİSTESİ

Sayfa

SİMGELER VE KISALTMALAR LİSTESİ

1D-CNN	1 Boyutlu Evrişimli Sinir Ağı
Adam	Uyarlamalı Moment Tahmini
AI	Yapay Zeka
ASR	Otomatik konuşma tanıma
BLSTM	Çift Yönlü Uzun Kısa Süreli Bellek Ağı
CNN	Evrişimli Sinir Ağları
CRNN	Evrişimli Tekrarlayan Sinir Ağı
DenseNet	Yoğun Bağlantılı Evrişimli Sinir Ağları
DNN	Derin Sinir Ağı
FN	Yanlış Negatif
FP	Yanlış Pozitif
GMM	Gauss Karışım Modeli
HMM	Gizli Markov Modeli
LSTM	Uzun Kısa Süreli Bellek Ağı
MFCC	Mel-Frekans Kepstral Katsayıları
NLP	Doğal Dil İşleme
ODTÜ	Orta Doğu Teknik Üniversitesi
PLDE	Fonetik Öğrenilebilir Sözlük Kodlaması
ReLU	Düzeltilmiş Doğrusal Birim
RNN	Tekrarlayan (Yinelemeli) Sinir Ağı
SLID	Konuşulan Dil Tanıma
SVN	Destek Vektör Makineleri
TCN	Zamansal Evrişimli Ağ
TDNN	Zaman Gecikmeli Sinir Ağı
TN	Doğru Negatif
TP	Doğru Pozitif

1. GİRİŞ

Dil sınıflandırma, doğal dil işleme (NLP) ve yapay zekâ (AI) alanlarında temel bir araştırma konusu olarak öne çıkmaktadır. Günümüzde konuşan dil tanıma sistemleri; insan-makine etkileşimi, robotik uygulamalar, çok dilli iletişim sistemleri ve güvenlik alanları gibi birçok farklı sektörde etkin şekilde kullanılmaktadır. Ayrıca, robotik sistemlerde otonom etkileşimlerin artırılması, müşteri hizmetlerinde çok dilli destek sağlanması ve çeviri hizmetlerinin otomatikleştirilmesi gibi kullanım senaryolarında da dil sınıflandırma algoritmalarına ihtiyaç duyulmaktadır. Bununla birlikte, güvenlik ve adli bilişim alanlarında, anonim ses kayıtlarının hangi dile ait olduğunu belirlemek; suçla mücadele, göçmen profilleme ya da istihbarat amaçlı analizlerde kritik rol oynayabilmektedir.

1.1. Konunun Tanımı

Konuşma, insanlar arasında en yaygın ve doğal iletişim yöntemidir. Günümüzde teknolojinin hızla gelişmesiyle birlikte, makinelerin insan sesini işleyip anlamlandırması mümkün hale gelmiş; bu da insan-makine etkileşimini yaygınlaştırarak iletişim alanında yeni uygulamaların önünü açmıştır. Bu gelişmeler sayesinde, hem günlük hayatın dijitalleşmesi kolaylaşmakta hem de bireylerin teknolojik olanaklara erişimi artmaktadır.

Bu tez çalışmasında, çeşitli dillerdeki konuşmaların hangi dile ait olduğunun otomatik olarak belirlenmesini amaçlayan bir ses tabanlı dil tanıma sistemi geliştirilmektedir. Sistem, ham ses sinyallerinden öznitelik çıkarımı yapılmasını ve bu özniteliklerin derin öğrenme mimarileriyle işlenmesi yoluyla çalışmaktadır. Özellikle günümüzde yaygın biçimde kullanılan BiLSTM (İki Yönlü Uzun Kısa Süreli Bellek) CNN (Convolutional Neural Network), Transformer tabanlı encoder, ve Conformer encoder gibi derin öğrenme mimarileri, bu sistemin temel bileşenlerini oluşturmaktadır. Bu yapılar, ses sinyallerinden çıkarılan özniteliklerin işlenmesinde kullanılmış ve model performansları karşılaştırmalı olarak analiz edilmiştir.

Çalışma kapsamında, konuşma verilerinden elde edilen MFCC (Mel-Frequency Cepstral Coefficients) gibi öznitelikler kullanılarak, farklı dillerin akustik yapılarının makine öğrenmesi yöntemleriyle ayrıştırılması hedeflenmektedir. Geliştirilen sistem, hem çok dilli ortamlarda konuşma analizine katkı sağlamayı hem de dil tanıma alanında BiLSTM, CNN, Transformer ve Conformer gibi modern derin öğrenme mimarilerinin sunduğu avantajları ortaya koymayı amaçlamaktadır.

1.2. Konuya İlişkin Yapılan Önceki Çalışmalar

İlgili çalışmada, konuşulan dili tanıma alanında doğruluğu artırmak amacıyla, x-vector temelli bir sistemin mimarisi geliştirilmiş ve TDNN (zaman gecikmeli sinir ağları) üzerinde yapılandırılmış yeni bir model sunulmuştur. Model, Indo-Avrupa, Sami ve Doğu Asya dil ailelerinden seçilmiş İspanyolca, Fransızca, İtalyanca, Rusça, Arapça, Hausa, Farsça, Çince, Japonca ve Tayca olmak üzere on dil üzerinde eğitilmiştir. Mozilla Common Voice veri kümesi kullanılmış, veriler ön işleme, dengeleme ve veri artırma adımlarından geçirilmiştir. 1x1 TDNN ara katmanları, grid search ile hiperparametre optimizasyonu ve huni yapılı katmanlar mimariye entegre edilmiştir. Modelin başarımı %97 doğruluk oranına ulaşmış ve accuracy metriği üzerinden değerlendirilmiştir [1]

MCSLD-AIBER (Multi-Class Spoken Language Detection with Artificial Intelligence using Fractal Al-Biruni Earth Radius Optimization) algoritmasının önerildiği bu çalışmada, kısa süreli ve gürültü içerikli ses kayıtlarında konuşulan dilin yüksek doğrulukla tanımlanması amaçlanmıştır. Sınıflandırma işlemi LSTM ağı ile gerçekleştirilmiş ve deneylerde İngilizce, İspanyolca ve Almanca dillerinden oluşan bir veri seti kullanılmıştır. Modelin performansı %96.19 doğruluk, %94.30 kesinlik, %94.30 duyarlılık ve %97.15 özgüllük ortalamaları ile değerlendirilmiştir [2].

PLDE (Phonetic Learnable Dictionary Encoding) adlı yeni bir havuzlama katmanının sunulduğu bu çalışmada, fonetik özelliklere dayalı hafif ve verimli bir konuşulan dil tanıma sistemi geliştirilmiştir. OLR2020 ve OLR2021 veri setleri kullanılarak Mandarin, Kantonca, Japonca, Korece, Rusça, Endonezce, Vietnamca, Kazakça, Tibetçe ve Uygurca dillerinde değerlendirme yapılmıştır. Önerilen PLDE yapısı, geleneksel havuzlama yöntemlerine kıyasla daha düşük parametre maliyetiyle başarılı sonuçlar üretmiştir. Model, fonetik özellikleri önceden eğitilmiş bir ASR (Otomatik konuşma tanıma) modelinden transfer öğrenme ile alır ve bu bilgileri tek katmanlı PLDE üzerinden sınıflandırır. Başarım, Cavg (ortalama maliyet) ve EER (eşit hata oranı) gibi metriklerle değerlendirilmiş; PLDE özellikle düşük parametreli yapılarda gürültülü ortamlarda yüksek doğruluk sağlamıştır [3].

Singh ve arkadaşları (2021), farklı konuşmacılardan gelen ses kayıtlarını kullanarak konuşulan dili tanıyan bir derin öğrenme modeli önermiştir. Çalışmada, ses dosyaları log-Mel spektrogramlara dönüştürülmüş ve CNN, Word Embedding ile Bernoulli Naive Bayes gibi yöntemlerle sınıflandırma yapılmıştır. Model, İngilizce, Almanca, İspanyolca ve 22 farklı dili içeren veri kümelerinde test edilmiş ve CNN modeliyle %98'e varan doğruluk elde edilmiştir. Ayrıca, veri artırma ve çeşitli ön işleme teknikleri kullanılarak modelin

genellenebilirliği artırılmıştır. Bu çalışma, dil tanımlamada görüntü tabanlı yaklaşımların etkinliğini vurgulamaktadır [4].

Başka bir çalışmada, konuşulan dili tanıma amacıyla filtre bank özelliklerinden yararlanılarak CNN tabanlı bir model önerilmiştir. Sistem, İngilizce, Almanca ve İspanyolca olmak üzere üç dili sınıflandırmak üzere eğitilmiştir. Veri kümesi Kaggle üzerinden temin edilmiş, örnekler cinsiyet ve konuşmacı açısından dengelenmiş ve veri artırma (pitch, hız, gürültü) teknikleri uygulanmıştır. Model, %91.3 doğruluk ve %92.74 ortalama güven skoruyla değerlendirilmiş; sınıf bazlı performans ise precision, recall ve F1-score metrikleri ile ölçülmüştür. Sonuçlar, CNN tabanlı yapının dil ayrımı görevlerinde güçlü bir performans sergilediğini göstermiştir [5].

Hint dillerinin sınıflandırılmasında doğruluk ve esnekliği artırmak amacıyla olan bu çalışmada, ensembe derin öğrenme modellerini temel alan yeni bir konuşulan dil tanıma sistemi önermektedir. Çalışmada Bengalce, Hintçe, Marathi, Tamilce ve İngilizce dillerinden oluşan bir veri kümesi kullanılmış; CNN, CRNN (Convolutional Recurrent Neural Network), 1D-CNN (One-Dimensional Convolutional Neural Network), LSTM ve TCN(Temporal Convolutional Network) modelleri çeşitli kombinasyonlarla entegre edilerek Stacking, Fine-tuning ve Voting tabanlı ensembe yapılar tasarlanmıştır. Veri seti, VoxLingua107 ve Common Language kaynaklarından elde edilen ses dosyalarıyla oluşturulmuş ve öznitelikleri çıkarılarak işlenmiştir. Sistem başarımı, doğruluk, kesinlik, duyarlılık ve F1-skoru gibi metriklerle değerlendirilmiş; en yüksek başarı oranı %93.5 doğruluk ve 0.92 F1-skoru ile Fine-tuned EDL-2 modeli tarafından elde edilmiştir [6].

X-vector temelli sistemin kullanıldığı bir başka çalışmada, değişken uzunluktaki ses segmentlerinden sabit boyutlu öznitelikler çıkarılarak konuşulan dil sınıflandırılmıştır. Model, 2017 NIST Language Recognition Evaluation (LRE17) kapsamında Arapça, Çince, İngilizce, Rusça, İspanyolca ve Portekizce gibi birbirine yakın diller ve lehçeler üzerinde test edilmiştir. Sistem, çok katmanlı bir DNN ile dil sınıflandırması için eğitilmiş; ses verileri MFCC ve bottleneck özelliklerle temsil edilmiştir. Elde edilen x-vector öznitelikleri, discriminative Gaussian classifier ile sınıflandırılmış; başarı, NIST'in Cprimary metrik değerleriyle değerlendirilmiştir. Sonuçlar, x-vector sisteminin mevcut i-vector temelli yaklaşımlardan daha yüksek doğruluk sağladığını göstermektedir [7].

Senone temelli derin sinir ağı (DNN) yaklaşımlarını karşılaştırmalı olarak ele alan bu çalışmada, konuşulan dilin tanınmasına yönelik farklı mimariler değerlendirilmiştir. Sistemler, NIST LRE09 kapsamında yer alan 23 dil ile DARPA RATS veri setinde bulunan Farsça, Urduca, Paştuca, Levanten Arapçası ve Darice dilleri üzerinde test edilmiştir.

DNN'lerden elde edilen bottleneck ve posterior öznitelikler, GMM ya da DNN tabanlı i-vector modelleri ile sınıflandırılmıştır. Performans değerlendirmeleri, farklı test süreleri dikkate alınarak Cavg (ortalama maliyet) metriği ile yapılmış; özellikle bottleneck + i-vector + GMM (Gaussian Mixture Model) tabanlı yapıların diğer yöntemlere kıyasla daha başarılı sonuçlar verdiği gösterilmiştir [8].

Bu çalışma, konuşulan dilin tanınması görevinde kullanılan DNN tabanlı telefon ayrıştırıcılarının (tokeniser) denetimsiz çapraz dil adaptasyonu ile farklı dillere uyarlanmasını incelemektedir. Temel tokeniser, İngilizce verilerle eğitildikten sonra Arapça, Mandarin, İngilizce, Lehçe ve İspanyolca gibi sekiz dile adapte edilmiştir. Sınıflandırma işlemi, bottleneck i-vector (BNIV-LR) ve fonotaktik TFIDF-SVM sistemleriyle gerçekleştirilmiştir. Performans, minDCF (minimum detection cost function) metriği ile ölçülmüş; özellikle sistem füzyonu ile %7–18 oranında iyileşme elde edilmiştir. Adaptasyonun, farklı tokeniser'lar arasında tamamlayıcı bilgi çeşitliliği sağladığı ve sınıflandırma başarımını olumlu yönde etkilediği gösterilmiştir [9].

Oruh ve arkadaşları tarafından yürütülen çalışmada, geleneksel LSTM mimarisi geliştirilerek, içine RNN tabanlı bir "unutma kapısı" entegre edilmiş ve bu sayede kesintisiz ses akışlarının daha verimli işlenmesi hedeflenmiştir. Model, İngilizce konuşulan rakamları içeren açık erişimli bir ses veri kümesi ile eğitilmiş; her biri 0–9 arasındaki sayıları temsil eden 2400 ses dosyası kullanılmıştır. Öznitelik çıkarımı için MFCC yöntemi uygulanmış ve önerilen LSTM-RNN mimarisi, diğer CNN ve sıralı modellerle karşılaştırıldığında en yüksek doğruluk değerine (%99.36) ulaşmıştır. Modelin sınıflandırma başarımı accuracy, precision, recall ve F1-score gibi yaygın metrikler ile değerlendirilmiştir [10].

Yapılan literatür incelemesi, konuşulan dil tanıma alanında birçok çalışmanın İngilizce, Fransızca, Almanca, İspanyolca ve Asya dilleri gibi yüksek kaynaklı dillere odaklandığını; ancak Türkçe ve Türkiye'de konuşulan diğer dillerin bu bağlamda büyük ölçüde ihmal edildiğini ortaya koymaktadır. Bu tez çalışması, bu eksikliği gidermek amacıyla Türkçe'nin yanı sıra Türkiye'de yaygın şekilde karşılaşılan Kürtçe, Arapça, Farsça, Rusça ve Almanca dillerine odaklanmaktadır. Sınıflandırma işlemleri, CNN, BiLSTM, Transformer tabanlı encoder ve Conformer mimarileri kullanılarak gerçekleştirilmiştir. Model başarımı accuracy, precision, recall ve F1-score gibi metriklerle değerlendirilmiştir. Kullanılan veri setleri; dengeli, Türkiye'nin demografik dağılımına göre dengesiz ve Türkçe hariç bırakılmış üç ayrı yapıdan oluşmakta olup, her biri 2–12 saniyelik ses dosyalarını içermektedir. Bu yönüyle çalışma, hem yerel dil çeşitliliğini esas alan hem de farklı veri dağılımlarına karşı model dayanıklılığını test eden özgün bir yaklaşım sunmaktadır.

2. PROBLEMİN TANIMI

Türkiye, farklı etnik kökenlerden bireylerin bir arada yaşadığı kozmopolit bir yapıya sahip olmasının yanı sıra, aynı zamanda birçok farklı ülkeden göç alan ve yoğun şekilde turist ağırlayan bir ülkedir. Bu çok dillilik durumu, özellikle güvenlik, iletişim ve insan-makine etkileşimi gibi alanlarda konuşulan dili otomatik olarak tanıma ihtiyacını ortaya çıkarmaktadır. Ancak, mevcut literatür incelendiğinde Türkiye bağlamında konuşan dili tanımaya yönelik özel olarak oluşturulmuş ve çok dilli yapıyı temel alan kapsamlı bir çalışmanın bulunmadığı görülmektedir. Bu çalışma, söz konusu boşluğu doldurmak amacıyla tasarlanmıştır.

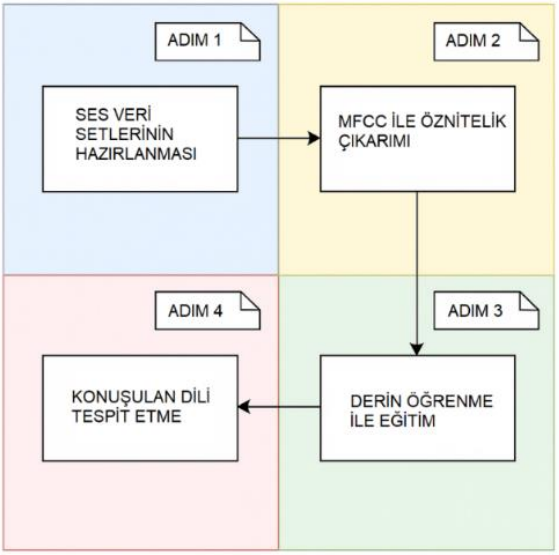
Bu kapsamda, çalışmada üç farklı veri seti oluşturulmuştur. İlk veri seti dengeli olarak tasarlanmış ve her bir dile eşit sayıda örnek dahil edilmiştir. İkinci veri seti ise dengesiz olup, gerçek dünyadaki dağılıma daha yakın olacak şekilde oluşturulmuştur. Üçüncü veri setinde ise Türkçe dil verisi çıkarılmış ve kalan dillerin sınıflandırma performansı bu durumda ayrı olarak analiz edilmiştir. Her üç veri seti üzerinde, CNN, BiLSTM, Transformer tabanlı encoder ve Conformer mimarileri kullanılarak derin öğrenme tabanlı analizler gerçekleştirilmiştir. Bu sayede, veri dağılımının model performansına etkisi sistematik olarak değerlendirilmiştir. Şekil 2.1'de problem tanımına yönelik blok diyagramı verilmiştir.

Şekil 2.1. Konuşulan dil tanıma blok diyagramı

2.1. Konuşulan Dilin Tanımlanması

Konuşulan dil tanıma (Spoken Language Identification – SLID), bir ses kaydı içerisindeki konuşmanın hangi doğal dile ait olduğunu, konuşmacının kimliği bilinmeden otomatik olarak belirlemeyi amaçlayan çok sınıflı bir sınıflandırma problemidir. Bu teknoloji, yalnızca konuşma içeriğine dayanarak dil etiketlemesi yapabilme yeteneğiyle,

konuşmacı bağımsız tanıma görevlerinde önemli bir rol üstlenmektedir. İnsanlar, biyolojik ve bilişsel mekanizmaları sayesinde dil tanıma konusunda oldukça yüksek doğruluk oranlarına sahip olsa da, bu becerinin bir makine tarafından taklit edilmesi, özellikle kısa konuşmalar, gürültülü ortamlar, aksan çeşitliliği ve düşük kaynaklı diller gibi birçok zorluk içermektedir [11]. SLID sistemleri, günümüzde çok sayıda pratik uygulama alanında aktif olarak kullanılmaktadır. Özellikle çok dilli konuşma tanıma sistemlerinin ön uç birimlerinde, konuşmanın hangi dilde olduğunu erken aşamada belirlemek, sonraki süreçlerde dil tabanlı modellerin doğru şekilde devreye girmesi açısından büyük önem taşımaktadır. Ayrıca, çağrı merkezlerinde gelen çağrının hangi dilde olduğunu tespit ederek otomatik müşteri yönlendirmesi sağlamak, kriz anında yönlendirme, çok dilli kullanıcı deneyimini geliştirmektedir. Bunun yanı sıra SLID, güvenlik ve istihbarat amaçlı konuşma gözetleme sistemlerinde, web tabanlı çok dilli içeriklerin otomatik olarak sınıflandırılmasında ve dijital medya arşivlerinin dil etiketleme süreçlerinde de yaygın olarak kullanılmaktadır [12]. Tipik bir konuşan dil tanımlama sistemi dört temel bileşenden Şekil 2.2 ’de ki gibi oluşmaktadır: veri toplama, öznitelik çıkarımı, model ile eğitim ve dil sınıflandırması. İlk aşamada, hedeflenen dilleri temsil edecek nitelikte ses verileri toplanmakta; ikinci aşamada, bu verilerden akustik özellikler çıkarılmakta; üçüncü aşama olarak çıkarılan bu özelliklerin modele öğretilmesi ve son aşamada ise çıkarılan bu öznitelikler bir sınıflandırıcıya (örneğin DNN, CNN, SVM vb.) aktarılmak suretiyle ilgili dilin tahmini yapılmaktadır. Başarılı bir SLID sisteminin oluşturulabilmesi için, uygun çeşitlilikte ve dengeli yapıya sahip bir konuşma veri tabanına erişim büyük önem arz etmektedir [13].

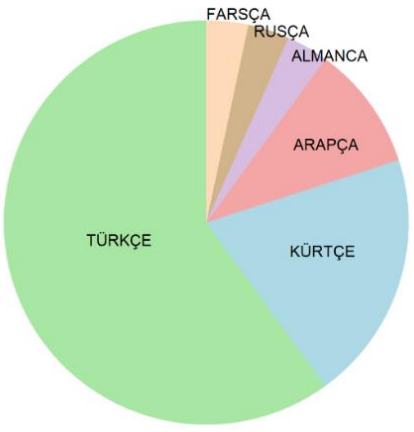


Şekil 2.2. Konuşulan dil tanıma sisteminin dört temel adımdan oluşan iş akışı

Bu çalışmada, konuşulan dil tanıma sistemi geliştirilirken yalnızca modelin teknik doğruluğu değil, aynı zamanda Türkiye'nin çok dilli sosyodemografik yapısı da göz önünde bulundurulmuştur. Türkçe'nin yanı sıra ülkede sayıca kayda değer nüfuslar tarafından konuşulan diğer dillerin sistem içinde temsil edilmesi, çalışmanın kapsamını genişletmek açısından önem arz etmektedir.

Türkiye'de konuşulan dillerin çeşitliliği, göç hareketleri ve etnik yapı nedeniyle oldukça zengindir. Özellikle Arapça, Kürtçe, Farsça, Rusça ve Almanca gibi diller, belirli göçmen toplulukları veya etnik gruplar aracılığıyla Türkiye'de aktif olarak konuşulmaktadır. Arapça, Irak, Suriye, Mısır, Filistin, Fas ve Yemen gibi ülkelerden gelen bireyler sayesinde yaygın olarak kullanılmaktadır ve toplamda yaklaşık 3.2 milyon kişilik bir Arapça konuşan nüfus mevcuttur. Kürtçe ise, Türkiye Cumhuriyeti vatandaşları arasında en yaygın konuşulan ikinci dil olup, yaklaşık 13 milyon Kürt kökenli vatandaş tarafından konuşulmaktadır. Farsça konuşan nüfus, ağırlıklı olarak İran uyruklu bireylerden oluşmakta ve yaklaşık 96 bin kişiyi kapsamaktadır. Rusça, özellikle turistik bölgelerde yoğunlaşan 85 bin civarındaki Rusya vatandaşları tarafından konuşulurken, Almanca da hem Almanya'dan gelen bireyler hem de Almanya kökenli Türkler aracılığıyla 116 bin kişi tarafından konuşulmaktadır. Bu veriler, Türkiye'deki dilsel çeşitliliğin sadece yerli nüfusla sınırlı kalmayıp göçmen ve yabancı topluluklar sayesinde daha da genişlediğini göstermektedir [14], [15].

Bu çok dilli yapı, yalnızca akademik anlamda teknik bir sınıflandırma sisteminin geliştirilmesini değil, aynı zamanda Türkiye'deki dilsel çeşitliliğin dikkate alındığı, uygulama potansiyeli yüksek bir altyapının oluşturulmasını da gerekli kılmaktadır. Şekil 2.3'teki görselde, Türkiye'de konuşulan dillerin yaklaşık nüfuslarına göre tahmini dağılımı sunulmaktadır.



Şekil 2.3. Türkiye'de konuşan dil istatistiği grafiği

2.2. Konuşma Tanımlamada Kullanılan Geleneksel Yöntemler

Derin öğrenme temelli modellerin ortaya çıkışından önce, konuşma sinyallerinin tanımlanmasında ve dolayısıyla konuşulan dilin sınıflandırılmasında en yaygın kullanılan yöntemler arasında Gaussian Karışım Modelleri (GMM) tabanlı yöntemler [16], Gizli Markov Modelleri (HMM) ve Destek Vektör Makineleri (SVM) tabanlı yaklaşımlar [17] yer almaktaydı. Bu istatistiksel yaklaşımlar, konuşma sinyallerinin zaman ve frekans boyutlarındaki özelliklerini modellemek için kullanılmıştır.

GMM, kısa süreli veya statik konuşma kesitlerinin spektral özelliklerini modellemek amacıyla kullanılır. Bu model, sesin akustik özelliklerini çok değişkenli olasılık dağılımları ile temsil eder. Statik vektör temsilleri üzerinden çalışan GMM, özellikle her bir sınıfa (örneğin dil) ait özniteliklerin Gauss dağılımlarının karışımı şeklinde modellendiği senaryolarda etkili olmuştur. Bu yönüyle GMM, dil sınıflandırma görevlerinde ilk nesil öznitelik temelli sistemlerin temelini oluşturmuştur.

Öte yandan, konuşma gibi zamana bağlı süreçleri modellemek için daha gelişmiş bir yapı olan Gizli Markov Modeli (HMM) geliştirilmiştir. HMM, zamana yayılmış rastgele değişken dizilerini modelleyerek, her bir zaman adımında sistemin belirli bir "durum"da olduğunu varsayar. Bu durumlar gözlemlenemese de, her bir durumun gözlenen çıktılarla ilişkili bir olasılık dağılımı vardır. Bu bağlamda HMM, çift aşamalı olasılıksal bir süreç olarak tanımlanır: birincisi gözlenemeyen (gizli) durum geçişleri; ikincisi bu durumlara bağlı olarak ortaya çıkan gözlemler.

HMM'nin konuşma tanımadaki başarısı, üç temel algoritmanın verimli bir şekilde uygulanmasına dayanır; Olasılık hesaplama için Forward-Backward algoritması, Parametre öğrenimi için EM algoritmasının özel bir formu olan Baum-Welch, algoritması, Durum dizisinin çözümü (decoding) için Viterbi algoritması.

GMM ve HMM tabanlı yaklaşımlar, konuşulan dil tanıma sistemlerinin ilk örneklerini oluşturmuştur. Tablo 2.1'de bu iki yöntemin çeşitli yönlerden karşılaştırılması sunulmuştur.

Tablo 2.1. HMM ve GMM karşılaştırması

Özellik	**GMM**	**HMM**
Model tipi	Karışım modeli (statik)	Durum geçişli stokastik model
Zaman bağımlılığı	Zaman bağımsız	Zaman bağımlı
Veri tipi	Statik öznitelik sektörleri	Zaman dizisi (sekans)
Parametre sayısı	Orta	Yüksek

Tablo 2.1. devam ediyor.

Özellik	**GMM**	**HMM**
Veri gereksinimi	Orta	Yüksek
Gürültüye dayanıklılık	Düşük	Orta
Yorumlanabilirlik	Yüksek	Orta
Genişletilebilirlik	Sınırlı	Yüksek

Geleneksel konuşulan dil tanıma sistemlerinde, GMM ve HMM iki temel yöntem olarak öne çıkmaktadır. GMM'ler, kısa süreli ve zamandan bağımsız öznitelikleri modellemede daha basit ve yorumlanabilir bir yapı sunarken; HMM'ler, zaman bağımlı yapıları modelleme yeteneği sayesinde dilin dinamik özelliklerini daha doğru şekilde temsil edebilmektedir. HMM'ler daha fazla parametre gerektirir ve hesaplama açısından daha karmaşık olsalar da, zaman dizisi içeren görevlerde genellikle daha yüksek başarı sağlar. Bu bağlamda, görev türüne ve veri yapısına bağlı olarak HMM'ler GMM'lere göre daha güçlü bir modelleme yeteneği sunmaktadır.

HMM, konuşma öznitelik dizilerinin üretilmesinde etkili bir model olmakla birlikte, her bir durumun zamanla değişmeyen sabit bir dağılımla temsil edilmesi, modelin gerçek dünyadaki karmaşık konuşma örüntülerini yansıtmasında sınırlayıcı olmuştur. Bu nedenle, HMM zamanla değişen, bağımlı ve daha dinamik yapıları modelleyememeye başlamıştır. Bu sınırlılıklar sonucunda, HMM tabanlı modeller yerlerini daha esnek, veri odaklı ve güçlü temsil yeteneğine sahip olan derin öğrenme tabanlı yöntemlere bırakmıştır.

2.3. Konuşma Tanımlamada Kullanılan Yeni Nesil Yöntemler

Geleneksel modelleme yaklaşımlarının (GMM, HMM vb.) zaman içindeki yetersizlikleri, konuşma tanıma sistemlerinin daha çok çoklu veriyi anlaması, karmaşık dil yapılarını tanıması ve zaman bağlı örüntüleri etkin şekilde işlemesi ihtiyacını doğurmuştur. Bu eksiklikleri gidermek üzere geliştirilen yeni nesil derin öğrenme tabanlı yöntemler, konuşma tanıma ve konuşulan dil sınıflandırması alanında son dönemde yaygın olarak kullanılmaktadır. Bu yöntemler, özellikle zaman bağlılık içeren ses sinyallerini anlamada ve anlamlı temsillere dönüştürmede geleneksel tekniklere göre daha üstün performans göstermektedir.

2.3.1. BiLSTM tabanlı yöntemler

BiLSTM mimarisi, özellikle zaman serisi verilerdeki bağlamsal örüntüleri

yakalayabilme yeteneği sayesinde, konuşma temelli dil tanıma sistemlerinde yaygın olarak kullanılmaktadır. BiLSTM, ses sinyallerinden elde edilen sıralı öznitelikleri hem ileri hem de geri yönde işleyerek geçmiş ve gelecek bağlamı aynı anda dikkate alır. Bu çift yönlü yapı, konuşma içerisindeki dil spesifik zamansal geçişlerin daha ayrıntılı biçimde öğrenilmesine olanak tanır. BiLSTM tabanlı modelin genel yapısı şu temel katmanlardan oluşmaktadır:

- **Giriş Katmanı:** MFCC tabanlı zaman serisi öznitelikleri modele giriş olarak verilir. Bu öznitelikler ses sinyalinin zaman içindeki akustik yapısını temsil eder.
- **Ön İşlem Katmanı (Conv2D + Reshape):** BiLSTM girişinden önce düşük seviyeli uzamsal ilişkileri yakalayabilmek amacıyla konvolüsyon işlemi uygulanır ve ardından veriler sıralı formatına dönüştürülür.
- **BiLSTM Katmanları:** Zaman içinde hem ileri hem de geri yönde örüntüleri analiz eden iki ardışık BiLSTM katmanı bulunur. Bu katmanlar, giriş dizilerinin bağlamsal ilişkilerini çok boyutlu biçimde işler.
- **Tam Bağlantılı Katman (Dense Layer):** BiLSTM katmanlarının çıktısı, doğrusal olmayan aktivasyonlarla çalışan dense katmana iletilerek yüksek düzeyli soyut temsiller üretilir.
- **Dropout Katmanı:** Aşırı öğrenmeyi önlemek amacıyla bazı bağlantılar eğitim sırasında rastgele devre dışı bırakılır.
- **Çıkış Katmanı:** Softmax aktivasyon fonksiyonu ile model, giriş ses örneğinin hangi dil sınıfına ait olduğunu olasılık tabanlı olarak tahmin eder.

Bu yapı sayesinde BiLSTM mimarisi, farklı dillere ait konuşma örneklerini ayırt etmek için hem kısa süreli hem de uzun menzilli ilişkileri öğrenebilmekte ve yüksek doğruluk oranlarına ulaşabilmektedir [18][19].

2.3.2. CNN tabanlı yöntemler

Son yıllarda, derin öğrenme yöntemleri, özellikle ses tanıma ve görüntü işleme gibi örüntü tanıma problemlerinde çığır açan sonuçlar üretmiştir. Bu başarıların temelinde yer alan yaklaşımlardan biri de, Evrişimli Sinir Ağı (CNN) yapılarıdır.

CNN, özellikle görüntü, ses ve zaman serisi verilerinde örüntü tanıma için geliştirilmiş derin öğrenme mimarileridir. CNN'ler, yüksek boyutlu verilerde uzamsal ve zamansal korelasyonları etkili bir biçimde öğrenme yetenekleri sayesinde ses tanıma, bilgisayarlı görü ve doğal dil işleme gibi birçok alanda yaygın olarak kullanılmaktadır [20][21]. CNN mimarisi genel olarak aşağıdaki temel katmanlardan oluşur:

- **Giriş Katmanı :** Ham veri (örneğin, bir görüntü ya da ses spektrogramı) modele giriş olarak verilir. Veri genellikle sabit boyutlu bir matris (örneğin, 224x224 piksel bir resim) şeklindedir [20].
- **Evrişim Katmanı (Convolution Layer):** Bu katmanda, küçük boyutlu filtreler (kernels) veri üzerinde kaydırılarak yerel özellikler çıkarılır. Filtreler aracılığıyla elde edilen özellik haritaları (feature maps), verinin düşük seviyeli örüntülerini temsil eder (örneğin kenarlar, tonlar) [21].
- **Aktivasyon Fonksiyonu (ReLU - Rectified Linear Unit):** Evrişim sonucunda elde edilen değerler, doğrusal olmayan bir dönüşümden geçirilir. ReLU fonksiyonu, negatif değerleri sıfıra kırparak ağın doğrusal olmayan örüntüleri öğrenmesine imkân tanır [21][22].
- **Havuzlama Katmanı (Pooling Layer):** Pooling işlemi (genellikle max pooling), özellik haritalarının boyutunu azaltarak hesaplama yükünü düşürür ve en önemli bilgilerin korunmasını sağlar. Bu katman, modelin gürültüye karşı daha dayanıklı olmasına da yardımcı olur [21].
- **Flatten Katmanı:** Havuzlama katmanından çıkan çok boyutlu veri, tam bağlantılı katmanlara (dense layers) giriş yapabilmesi için tek boyutlu bir vektöre dönüştürülür [22].
- **Tam Bağlantılı Katmanlar (Fully Connected Layers - Dense Layers):** Bu katmanlar, öğrenilen öznitelikleri bir araya getirerek nihai sınıflandırmayı gerçekleştirir. Genellikle en son katmanda, hedef sınıf sayısı kadar nöron bulunur ve softmax gibi bir aktivasyon fonksiyonu kullanılarak olasılık tahminleri yapılır [20][21].

Bu yapı sayesinde CNN'ler, veri içindeki düşük seviyeli (kenar gibi) özellikleri ilk katmanlarda, daha karmaşık yapısal ilişkileri ise derin katmanlarda öğrenir. CNN mimarisinin, öğrenilmesi gereken parametre sayısını azaltması ve bölgesel benzerlikleri kullanarak verimli öğrenim sağlaması, onu büyük veri setleri üzerinde güçlü bir seçenek haline getirmiştir [21][22].

2.3.3. Transformer encoder tabanlı yöntemler

Geleneksel RNN ve CNN mimarilerinin sınırlamaları, konuşma tanıma ve konuşulan dil tanıma alanında Transformer tabanlı yaklaşımların yükselmesine neden olmuştur. Özellikle uzun süreli bağımlılıkları modelleme, bağlamı daha geniş ölçekte öğrenebilme ve

paralel işlem kapasitesi gibi avantajları, Transformer mimarisini ses işleme uygulamaları için cazip bir alternatif haline getirmiştir [23], [24]. Transformer mimarisi, self-attention mekanizmasına dayalı olarak giriş dizisindeki tüm öğelerin birbirleriyle etkileşim kurmasını sağlar. Temel yapı taşları şunlardır:

- **Çok Başlı Dikkat (Multi-Head Attention):** Verideki farklı bilgi parçalarını paralel olarak modelleyebilmek için birden fazla dikkat mekanizması kullanılır.
- **Pozisyonel Kodlama (Positional Encoding):** Transformer, giriş verisinin sırasını doğrudan algılayamadığından, her girişe konum bilgisi eklenir.
- **Katman Normalizasyonu ve Artımsal Bağlantılar (Residual Connections):** Derin ağlarda eğitim stabilitesini artırmak ve bilgi kaybını azaltmak için kullanılır.
- **Feed-Forward Ağlar:** Her katmandan sonra, doğrusal olmayan dönüştürmelerle bilgiyi işler.

Transformer mimarilerinin konuşma tanıma sistemlerine uygulanabilirliğini artırmak amacıyla çeşitli mimari iyileştirmeler yapılmıştır. Yeh ve ark. (2019) tarafından geliştirilen Transformer-Transducer yapısı, klasik RNN-Transducer mimarisine alternatif olarak önerilmiş ve akustik modelleme birimi olarak Transformer'ın kullanılabilirliğini göstermiştir [6]. Bu yapıda, girişteki ham ses verileri önce VGGNet tabanlı evrişimsel bloklar ile ön işleme tabi tutulmuş, ardından pozisyonel kodlamalarla Transformer'a aktarılmıştır. Ayrıca, truncated self-attention kullanılarak hesaplama yükü azaltılmış ve modelin akışkan (streamable) bir biçimde gerçek zamanlı çalışması sağlanmıştır [25], [26]. LibriSpeech veri kümesi üzerinde yapılan deneylerde, %6.37 WER (test-clean) ve %15.30 WER (test-other) gibi sonuçlarla, LSTM ve BLSTM tabanlı modellere göre daha yüksek doğruluk sağlanmıştır.

Zhu ve arkadaşları (2020) tarafından yapılan başka bir çalışmada ise, dil kimliği (Language ID) bilgisi self-attention mekanizmasına dahil edilerek dil odaklı dikkat başlıkları (Language-Specific Attention Heads) ve dil bağımlı bağlam pencereleri (Language Dependent Attention Spans) geliştirilmiştir. Bu yöntemle, Danca, Norveççe, İsveççe, Fince ve Felemenkçe dillerini kapsayan testlerde, baz model ile karşılaştırıldığında %8 ila %20 arasında göreceli doğruluk artışı sağlanmıştır. Bu sonuçlar, özellikle düşük kaynaklı dillerde Transformer tabanlı mimarilerin sağladığı kazanımları açıkça ortaya koymaktadır.

Sonuç olarak, Transformer tabanlı modeller, SLID sistemlerinde uzun süreli bağımlılıkların modellenmesi, paralel hesaplama, dil kimliği gibi meta verilerin dikkat mekanizmalarına entegre edilmesi gibi avantajlarıyla yeni bir standart oluşturmaktadır. Ancak bu sistemlerin yüksek veri ve hesaplama gücü gereksinimi, mimarilerin dikkatli bir

şekilde tasarlanmasını ve optimize edilmesini gerekli kılmaktadır [23],[25].

2.3.4. Conformer encoder tabanlı yöntemler

Son yıllarda, hem yerel hem de küresel örüntülerin aynı anda öğrenilmesini sağlayan Conformer mimarisi, konuşulan dil tanıma sistemlerinde öne çıkan yaklaşımlardan biri hâline gelmiştir. Conformer, CNN lokal bağlam çıkarımı kabiliyeti ile Transformer mimarisinin kendine dikkat (self-attention) mekanizmasını birleştirerek, ses sinyallerindeki hem düşük seviyeli hem de uzun menzilli örüntüleri aynı anda modelleyebilen bir yapı sunmaktadır [27][28]. Conformer mimarisi temel olarak aşağıdaki katmanlardan oluşur.

- **Altörnekleme Katmanları (Subsampling Layers):** Giriş ses verisini zamansal olarak sıkıştırarak işlem yükünü azaltır ve önemli bilgilerin daha kısa diziler üzerinden temsil edilmesini sağlar.
- **Conformer Blokları :** Her blok; iki ileri beslemeli ağ (FFN), çok başlı kendine dikkat mekanizması (MHSA) ve evrişim modüllerini içeren bir Macaron benzeri yapıya sahiptir. Bu yapı sayesinde hem lokal hem de global temsil öğrenimi sağlanır [28].
- **Dikkat Tabanlı Havuzlama Katmanı (Attentive Statistics Pooling):** Zaman ekseni boyunca elde edilen temsil vektörlerini ağırlıklı ortalama yoluyla sabit uzunluklu bir gömme vektöre (embedding) dönüştürür. Bu yaklaşım, modelin girişin önemli bölümlerine daha fazla odaklanmasını sağlar [28].
- **Gömme Katmanları ve Sınıflandırıcı :** Havuzlanan çıktılar, doğrusal projeksiyon ve normalizasyon katmanları ile işlenerek nihai dil sınıflandırmasına gönderilir. Eğitim süreci genellikle çapraz entropi kaybı ile gerçekleştirilir [27].

Conformer mimarisi, ses tanıma ön-eğitimli modellerle birlikte kullanıldığında konuşulan dil tanıma performansını büyük ölçüde artırmaktadır. Bartley vd. (2023) tarafından yapılan çalışmada, önceden eğitilmiş Conformer katmanlarının yalnızca alt düzey bölümleri kullanılarak bile yüksek doğrulukla dil tanıma gerçekleştirilebildiği gösterilmiştir. Özellikle alt katmanlardan alınan gömme temsillerin hem bilinen hem de daha önce görülmemiş diller üzerinde yüksek genelleme yeteneği sergilediği belirtilmiştir [27].

Wang vd. (2023) ise, Conformer mimarisine segment maskeleme temelli özbilgili damıtma (self-knowledge distillation) stratejisi entegre ederek özellikle kısa süreli ses örnekleri üzerindeki başarımı artırmıştır. Bu yaklaşımda, aynı ses örneğinin tam hâli ve maskeleme uygulanmış versiyonları modelden geçirilerek çıkış dağılımları arasında

benzerlik sağlanır; böylece modelin öğrenme yetkinliği kısa süreli seslerde de korunur [28].

Sonuç olarak, Conformer tabanlı modeller hem parametre verimliliği açısından avantajlıdır hem de yüksek performansları ile dikkat çekmektedir. Bu tez kapsamında uygulanan Conformer mimarisi, CNN ve Transformer gibi klasik mimarilerle karşılaştırmalı olarak değerlendirilmiş ve yüksek başarı göstermiştir.

Yeni nesil derin öğrenme modelleri, ses sinyallerinden doğrudan öznitelik öğrenebilme, zamansal bağlamları daha isabetli şekilde çözümleme ve büyük veri kümeleri üzerinde eğitilebilme avantajlarıyla, geleneksel modellere kıyasla çok daha yüksek performans sunmaktadır. Özellikle CNN, BiLSTM, Transformer tabanlı encoder ve Conformer gibi mimariler, farklı öznitelik türleriyle entegre çalışarak konuşma temelli dil tanıma sistemlerinin evriminde belirleyici bir rol oynamaktadır.

3. TEZDE YAPILAN ÇALIŞMALAR

Bu tez çalışmasında, çok dilli konuşma verileri üzerinden otomatik dil tanıma sistemi geliştirmek amacıyla bir dizi aşamalı işlem gerçekleştirilmiştir. İlk olarak, Türkçe, Kürtçe, Arapça, Farsça, Rusça ve Almanca dillerine ait ses dosyalarından oluşan bir veri seti hazırlanmış ve sistemde kullanılmak üzere yüklenmiştir. Ardından, her ses örneğinden anlamlı öznitelikler elde edebilmek için MFCC yöntemi kullanılmış; bu işlem öncesinde frekans uzayında düşük enerjili bileşenlerin filtrelenmesiyle temel düzeyde bir gürültü azaltma uygulanmıştır. Özellik çıkarımı tamamlandıktan sonra veriler rastgele bir şekilde eğitim (%70), doğrulama (%15) ve test (%15) alt kümelerine ayrılarak modelin güvenilirliğini artırmaya yönelik bir yapı benimsenmiştir.

Bu altyapının üzerine, dört farklı derin öğrenme mimarisi geliştirilmiştir: CNN, BiLSTM, Transformer tabanlı encoder ve Conformer mimarisi. Bu modeller, eğitim verisiyle beslenerek konuşulan dili doğru bir şekilde tahmin edebilecek düzeye gelene kadar optimize edilmiş; doğrulama verisi yardımıyla ise modelin aşırı öğrenme (overfitting) eğilimi kontrol edilmiştir. Eğitilen modeller daha sonra, daha önce hiç görülmemiş test verileri üzerinde denenerek, sistemin genel başarımı değerlendirilmeye alınmıştır. Bu aşamada doğruluk oranı, sınıflandırma başarımı ve hata matrisi gibi metriklerle birlikte sistemin etkinliği ortaya konmuştur. Son olarak, elde edilen sonuçlar eğitim süreci boyunca oluşan doğruluk ve kayıp grafiklerinin yanı sıra, test performansını yansıtan görselleştirmelerle desteklenmiş ve tezde sunulmuştur. Genel süreci ve modeli ayrıntılandıran akış şeması Şekil 3.1'de gösterilmiştir.

Şekil 3.1. Önerilen modelin akış şeması

3.1. Veri Seti

Bu tez çalışmasında kullanılan veri kümesi, farklı açık kaynak platformlarından derlenmiş olup, Türkçe, Kürtçe, Arapça, Farsça, Rusça ve Almanca dillerine ait ses kayıtlarını içermektedir. Veri seti oluşturulurken; Türkçe ses kayıtları Orta Doğu Teknik Üniversitesi (ODTÜ) tarafından sağlanan kaynaklardan toplanmış, diğer diller için ise Kaggle [29], [30], [31], [32] ve Github [33] platformlardan elde edilen veriler kullanılmıştır Tablo 3.1' de veri setleri ile ilgili bilgiler bulunmaktadır.

Tablo 3.1. Kullanılan veri setlerinin özellikleri

Dil	**Kişi Sayısı**	**Ses dosyalarının Uzunlukları**
Türkçe	15	2sn- 8sn
Kürtçe	8	2sn- 10sn
Rusça	8	2sn- 7sn
Almanca	12	3sn- 12sn
Farsça	10	3sn- 11sn
Arapça	10	2sn- 14sn

Ses kayıtları, dil isimlerine göre klasörleme işlemleri yapılarak etiketlenmiştir. Kayıtlar üzerinde, spektral analiz yöntemiyle gerçekleştirilen bir gürültü temizleme işlemi uygulanmıştır. Bu işlemde, ses sinyallerinin FFT (Fast Fourier Transform) spektrumundaki düşük enerjili bileşenler filtrelenerek, yalnızca belirgin frekanslar korunmuştur. Tüm ses dosyaları 16 kHz örnekleme oranına normalize edilmiş ve gürültü azaltımı sonrası MFCC çıkarımı gerçekleştirilmiştir. Veri seti, rastgele bir şekilde %70 eğitim, %15 doğrulama ve %15 test alt kümelerine ayrılmıştır. Konuşulan dili tanıma sistemlerinin performansını değerlendirmek amacıyla, Türkiye'de konuşulan başlıca dilleri temsil eden üç farklı veri seti tasarlanmıştır. Bu veri setleri; dengeli, dengesiz ve veri setlerine Türkçe dahil edilmeden yapılan örneklerden oluşmaktadır. Veri setlerinin bu şekilde çeşitlendirilmesi, hem dil çeşitliliğini temsil etme hem de sınıflandırma modellerinin farklı veri dağılımı koşullarındaki (örneğin veri dengesi veya baskın dilin çıkarılması gibi) performanslarını karşılaştırmalı olarak değerlendirme amacı taşımaktadır.

3.1.1. Dengeli veri seti

Dengeli veri seti, her dil sınıfı için eşit sayıda örnek içerecek şekilde oluşturulmuştur.

Bu kapsamda, her bir dil için 1000 adet ses dosyası olmak üzere toplam 6000 ses örneği kullanılmıştır. Ses dosyaları 2 ila 12 saniye uzunluğundadır. Bu yapı sayesinde, modellerin farklı dillerdeki öğrenme başarımı, veri miktarından bağımsız olarak karşılaştırılabilir hale getirilmiştir. Tablo 3.2'de dengeli veri setine ait özellikler belirtilmiştir.

Tablo 3.2. Dengeli veri seti

	Dil	**Dengeli Veri Seti (Ses Dosya Sayısı)**
1	Türkçe	1000
2	Kürtçe	1000
3	Arapça	1000
4	Almanca	1000
5	Rusça	1000
6	Farsça	1000

3.1.2. Dengesiz veri seti

Dengesiz veri seti, Türkiye'deki gerçek dil dağılımlarını yansıtacak şekilde oluşturulmuştur. Bu bağlamda, Türkçe konuşan bireyler veri setinin %78'ini, Kürtçe konuşanlar %15'ini, Arapça konuşanlar %5'ini ve diğer dilleri konuşanlar ise %2'sini oluşturmaktadır. Bu dağılım, sınıflandırma modellerinin dengesiz veri koşulları altındaki performansını değerlendirebilmek açısından önemli bir temel sağlamaktadır. Tablo 3.3'de dengesiz veri setine ait özellikler belirtilmiştir.

Tablo 3.3. Dengesiz veri seti

	Dil	**Dengesiz Veri Seti (Ses Dosya Sayısı)**
1	Türkçe	4700
2	Kürtçe	900
3	Arapça	300
4	Almanca	70
5	Rusça	70
6	Farsça	70

3.1.3. Türkçe'siz veri Seti

Bu son veri setinde, Türkçe verileri dışarıda bırakılarak yalnızca azınlık ve yabancı kökenli diller üzerinde sınıflandırma yapılmıştır. Bu veri setinde Arapça, Kürtçe, Farsça,

Rusça ve Almanca dillerine ait ses dosyaları eşit sayıda olacak şekilde dağıtılmış olup toplamda 6000 adet ses dosyasından oluşmuştur. Bu yapı, Türkçe baskınlığının model performansına olan etkisini gözlemlemek amacıyla tasarlanmıştır. Tablo 3.4'te Türkçe'siz veri setine ait özellikler belirtilmiştir.

Tablo 3.4. Türkçe'siz veri seti

	Dil	**Türkçe'siz Veri Seti (Ses Dosya Sayısı)**
1	Kürtçe	1000
2	Arapça	1000
3	Almanca	1000
4	Rusça	1000
5	Farsça	1000

3.2. BiLSTM Tabanlı Dil Sınıflandırma Yaklaşımları

Konuşma sinyallerinin sınıflandırılmasında, geleneksel öznitelik çıkarım süreçlerinin ardından uygulanan derin öğrenme mimarileri arasında BiLSTM ağları, zaman serilerindeki ardışık bağıntıları modelleme konusundaki üstünlükleriyle dikkat çekmektedir. BiLSTM modelleri, ses sinyallerinin hem geçmiş hem de gelecek bağlam bilgisini eş zamanlı olarak işleyebilme yeteneği sayesinde, dil tanıma görevlerinde etkili bir çözüm sunmaktadır. Bu çift yönlü yapı, özellikle konuşmanın doğal akışında ortaya çıkan zamansal örüntülerin daha doğru şekilde temsil edilmesine olanak tanımaktadır.

Dil sınıflandırma uygulamalarında, ham ses sinyali belirli pencereleme teknikleriyle çerçevelenir ve bu çerçevelerden MFCC, log-Mel spektrogram veya benzeri zaman-frekans temsilleri elde edilir. Bu öznitelik dizileri sıralı yapılar şeklinde BiLSTM ağına giriş olarak verilerek, farklı dillerin karakteristik zamansal geçişleri, fonetik yapıları ve artikülasyonel örüntüleri öğrenilmeye çalışılır. BiLSTM mimarisi, bu ardışık özniteliklerdeki bağlamsal ilişkileri ileri ve geri yönde analiz ederek, ses sinyallerine ait derin temsil öğrenimini mümkün kılar.

Ayrıca, BiLSTM modellerinin uzun bağımlılıkları hatırlayabilme kapasitesi, konuşma boyunca zaman içinde dağılmış kritik ipuçlarının bütüncül bir şekilde işlenmesini sağlar. Bu yönüyle, sınırlı miktarda veriye sahip durumlarda dahi dil-spesifik zamansal kalıpların başarıyla öğrenilmesine olanak verir. Bu sayede, BiLSTM temelli modeller; konuşma örneklerini sadece statik özniteliklere değil, zaman içerisindeki değişimlere dayalı olarak da ayırt ederek, dil sınıflandırma alanında yüksek doğrulukta sonuçlar elde edebilmektedir.

3.3. CNN Tabanlı Dil Sınıflandırma Yaklaşımları

Konuşma sinyallerinin sınıflandırılmasında, geleneksel özellik çıkarım yöntemlerinin ardından uygulanan derin öğrenme mimarileri arasında CNN modelleri, güçlü örüntü tanıma yetenekleri nedeniyle öne çıkmaktadır. CNN'ler, özellikle görüntü verileri üzerinde başarılı sonuçlar üretmek üzere tasarlanmış olmalarına rağmen, ses sinyallerinden elde edilen spektrogramlar veya MFCC gibi zamana bağlı iki boyutlu temsiller sayesinde, konuşma verileri üzerinde de etkili şekilde uygulanabilmektedir.

Dil sınıflandırma uygulamalarında, öncelikle ham ses sinyali belirli bir pencere boyutunda çerçevelenir ve bu çerçevelerden MFCC, log-Mel spektrogram ya da Mel-spektrogram gibi ses özellikleri elde edilir. Bu iki boyutlu özellik matrisleri, CNN'e giriş olarak verilerek, farklı dillerin fonetik, artikülasyonel ve akustik özelliklerinin uzamsal örüntüleri öğrenilir. CNN mimarileri, bu giriş verilerindeki dil-specifik özellikleri filtrelerle yakalayarak derin katmanlarda genelleştirir ve sınıflandırma katmanına aktarır.

Ayrıca, CNN'lerin parametre paylaşımı ve yerel algılayıcılara (local receptive fields) dayalı yapısı, sınırlı boyuttaki dil verilerinden öğrenmeyi mümkün kılarken, uzamsal yapıyı bozmadan özellik çıkarımı yapılmasına olanak tanır. Bu sayede, CNN modelleri farklı dillere ait konuşma örneklerini yüksek doğrulukla ayırt edebilecek yapay öğrenme yetenekleri geliştirebilmektedir.

Son yıllarda SLID alanında derin öğrenme temelli modellerin yaygınlaşması, özellikle CNN kullanımıyla önemli bir ivme kazanmıştır. CNN mimarileri, ses verisinin mel spektrogram gibi görsel temsilleri üzerinden dil sınıflandırması yapmada oldukça başarılı sonuçlar vermektedir. Draghici vd. (2020) tarafından yapılan bir çalışmada, mel spektrogramlara uygulanan CNN ve CRNN mimarilerinin, dil aileleri arasında benzer fonetik yapılara sahip dillerin ayrımında etkili olduğu gösterilmiştir. Araştırmada, CNN modelinin sınıflandırma başarımının veri kümesine bağlı olarak %87'ye kadar çıktığı ve dengesiz veri kümelerinde ağırlıklandırma ve normalizasyon gibi stratejilerin önemli rol oynadığı belirtilmiştir [34]. Benzer şekilde, Das ve Roy (2021) tarafından Hindistan'da konuşulan diller üzerine yapılan çalışmada, CNN tabanlı BiLSTM mimarisi ile geliştirilen bir hibrit model, transfer öğrenme ve dikkat mekanizması (attention) sayesinde %98.1 doğruluğa ulaşmış ve geleneksel yöntemlerin üzerinde bir performans sergilemiştir [35]. Bu çalışmalar, CNN mimarisinin dil sınıflandırmada hem spektral hem de zamansal bilgiyi anlamlı şekilde modelleyebildiğini ve konuşma sinyallerinden öğrenilen derin temsiller aracılığıyla yüksek doğruluk elde ettiğini göstermektedir.

3.4. Transformer Encoder Tabanlı Dil Sınıflandırma Yaklaşımları

Son yıllarda derin öğrenme mimarilerinin evrilmesiyle birlikte, özellikle doğal dil işleme (NLP) ve dil sınıflandırma alanlarında Transformer tabanlı modeller öne çıkmıştır. İlk olarak Vaswani vd. (2017) tarafından önerilen ve "attention is all you need" ilkesi üzerine inşa edilen Transformer yapısı, dizisel verilerde uzun menzilli bağımlılıkların öğrenilmesinde önceki RNN ve LSTM tabanlı yöntemlere kıyasla çok daha etkili sonuçlar üretmektedir. Dil sınıflandırma uygulamalarında Transformer mimarisi, ses veya metin temelli girişlerdeki anlamlı örüntüleri konum bilgisi (positional encoding) ile birlikte değerlendirerek güçlü bağlamsal temsiller üretir.

Transformer tabanlı modellerde, ses sinyali ya doğrudan spektral temsiller üzerinden (örneğin MFCC, log-Mel) ya da bunlardan elde edilen öznitelik vektörleri kullanılarak işlenir. Bu vektörler modele giriş olarak sunulur ve çok katmanlı attention mekanizmaları aracılığıyla farklı dillerin fonetik ve anlamsal özellikleri öne çıkarılır. Transformer'ların en önemli avantajlarından biri, paralel hesaplama yapabilmeleri ve bu sayede çok büyük veri kümelerinde dahi verimli öğrenme sağlayabilmeleridir. Ayrıca, self-attention yapısı sayesinde hem yakın hem uzak bağlamdaki dilsel örüntüleri eşzamanlı olarak değerlendirebilirler.

Transformer tabanlı modeller, son yıllarda konuşulan dil tanıma görevlerinde önemli performans kazanımları sağlamıştır. Radfar vd. (2020) tarafından geliştirilen çalışmada, end-to-end bir Transformer mimarisi ile Fluent Speech Commands (FSC) veri kümesi üzerinde %98,1 doğruluk oranına ulaşılmıştır. Bu model, konvolüsyonel veya tekrarlayan katmanlar yerine doğrudan kendi kendine dikkat (self-attention) mekanizmasını kullanarak ses sinyalinden anlamlı temsil vektörleri öğrenmiştir. Ayrıca, bu mimari geleneksel RNN+CNN tabanlı yapılara kıyasla %25 daha az parametreye sahip olup, yüksek paralelleştirilebilirliği ile özellikle cihaz içi uygulamalar için avantaj sunmaktadır [36]. Benzer şekilde, Wang vd. (2023) tarafından önerilen LAMASSU adlı sistem, çok dilli konuşma tanıma ve çeviri görevlerini dilden bağımsız olarak gerçekleştirebilmiş, bu sırada hem parametre sayısını azaltmış hem de hedef dil tanımlama ve CTC düzenlemesi gibi stratejilerle modelin genellenebilirliğini artırmıştır. Bu sistem, hedef dili tahmin etmek için ek bir giriş vektörüyle Transformer transducer mimarisini geliştirmiş ve düşük kaynaklı dillerde bile monolingual modellerle benzer performansa ulaşmıştır [37]. Bu bulgular, Transformer tabanlı mimarilerin hem doğruluk hem de verimlilik açısından geleneksel modellere kıyasla kayda değer bir üstünlük sunduğunu açıkça göstermektedir.

3.5. Conformer Encoder Tabanlı Dil Sınıflandırma Yaklaşımları

Konuşma sinyallerinin sınıflandırılmasında, son yıllarda geliştirilen Conformer encoder mimarisi, hem zamansal hem de uzamsal bilgiyi aynı anda modelleyebilme yeteneği sayesinde dikkat çekmektedir. Conformer, geleneksel sinir ağı yapılarının ötesine geçerek konvolüsyonel katmanlar ile self-attention mekanizmasını birleştiren hibrit bir yapı sunmakta ve bu sayede konuşma verilerindeki hem yerel hem de küresel bağlamı etkin biçimde yakalayabilmektedir. Bu özellikleri sayesinde, konuşma sinyallerinde sıklıkla rastlanan kısa süreli akustik değişimleri ve uzun süreli bağımlılıkları aynı anda modelleme yeteneğine sahiptir.

Dil sınıflandırma uygulamalarında, ham ses verilerinden elde edilen Mel Frekans Kepstrum Katsayıları (MFCC), log-Mel spektrogram veya benzeri iki boyutlu temsiller, Conformer mimarisine giriş olarak verilerek işlenir. Bu temsillerin konvolüsyonel bileşenler tarafından yerel akustik desenleri tanımlaması sağlanırken, self-attention katmanları tüm zaman dizisi boyunca uzun menzilli bağlamları yakalayarak daha güçlü temsil öğrenimi gerçekleştirir. Böylece, farklı dillere özgü fonetik ve prosodik kalıpların çok katmanlı düzeyde ayrıştırılması mümkün olur.

Conformer mimarisinin çok ölçekli bağlam öğrenimi, farklı dil yapılarına sahip ses örneklerinin detaylı biçimde ayrıştırılmasına olanak tanırken; modelin derin yapısı, düşük kaliteli ya da gürültülü ortamlardaki sinyallerden dahi anlamlı dil-spesifik özelliklerin çıkarılmasını mümkün kılar. Bu yönüyle Conformer tabanlı sistemler, konuşma sinyallerini hem zamansal hem de spektral düzeyde ayrıntılı biçimde işleyerek, dil sınıflandırma alanında yüksek başarıma ulaşan güçlü bir çerçeve sunmaktadır.

3.6. Performans Değerlendirme Metrikleri

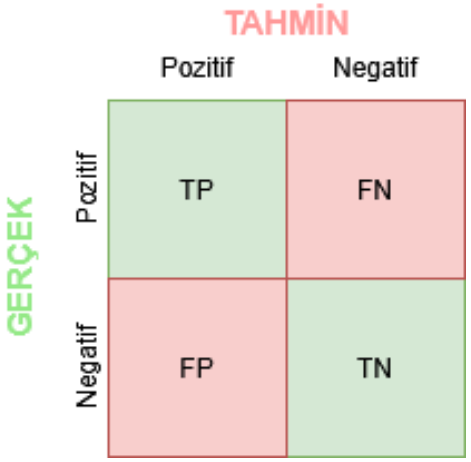
Bu tez çalışmasında kullanılan BiLSTM, CNN ve Transformer ve Conformer Encoder tabanlı modellerin başarımını değerlendirebilmek için çeşitli performans ölçüm metriklerinden yararlanılmıştır. Bu metrikler, modellerin dil sınıflandırma görevindeki doğruluğunu, genelleme yeteneğini ve sınıf bazlı başarımını nicel olarak analiz etmeye olanak tanımaktadır. Öğrenme sürecinde elde edilen doğruluk (accuracy) ve kayıp (loss) değerlerinin epoch bazlı değişimini gösteren öğrenme eğrileri, modellerin öğrenme davranışlarını izlemek açısından kritik rol oynamaktadır. Bunun yanı sıra, sınıflar arası tahmin doğruluğunun görselleştirilmesini sağlayan karmaşıklık matrisi (confusion matrix), özellikle hangi dillerin birbirine karıştırıldığını analiz etmek açısından önem taşımaktadır.

Ayrıca, sınıf bazlı detaylı performans değerlendirmesi için kesinlik (precision), Duyarlılık (recall) ve F1-Skoru (F1-score) gibi istatistiksel ölçütler de kullanılmış ve modellerin genel ve sınıf düzeyindeki başarımı kapsamlı biçimde ortaya konmuştur. Bu metrikler sayesinde, geliştirilen modellerin güçlü ve zayıf yönleri sistematik olarak değerlendirilmiş ve model karşılaştırmaları nesnel temellere oturtulmuştur.

3.6.1. Karışıklık matrisi (confusion matrix)

Karışıklık matrisi, bir sınıflandırma probleminin tahmin sonuçları ile gerçek etiketler arasındaki ilişkileri özetleyen temel araçlardan biridir. Bu matris, her bir sınıfın bağımsız olarak değerlendirilmesine imkân tanıyarak modelin genel performansını sınıf bazında görselleştirir. Matrisin mantığı belirli bir sınıfın "pozitif" olarak seçilmesi ve geri kalan tüm sınıfların "negatif" olarak kabul edilmesi esasına dayanır. Bu doğrultuda; doğru pozitif (TP), doğru negatif (TN), yanlış pozitif (FP) ve yanlış negatif (FN) olmak üzere dört temel değer elde edilir. Şekil 3.2'te örnek bir karışıklık matrisi sunulmuştur. Bu kavramlar aşağıda tanımlanmıştır:

- **Doğru Pozitif (TP)**: Belirli bir sınıfa ait verilerin, model tarafından doğru şekilde o sınıfa ait olarak tahmin edilme sayısıdır.
- **Doğru Negatif (TN)**: Belirli bir sınıfa ait olmayan verilerin, model tarafından da doğru bir şekilde o sınıfa ait olmadığı şeklinde tahmin edilme sayısıdır.
- **Yanlış Pozitif (FP)**: Farklı bir sınıfa ait verilerin, model tarafından hatalı biçimde ilgili sınıfa ait olarak tahmin edilme sayısıdır.
- **Yanlış Negatif (FN)**: İlgili sınıfa ait verilerin, model tarafından yanlış şekilde farklı bir sınıfa ait olarak tahmin edilme sayısıdır.



Şekil 3.2. İkili sınıflandırma için karışıklık matrisi

Genel anlamda karışıklık matrisinin yorumlanması bu şekildedir. Ancak sınıf sayısı arttıkça bu yapı daha karmaşık bir hâl alır. Bu nedenle sınıflandırmanın türü dikkate alınmalıdır. Sınıflandırmalar temel olarak ikiye ayrılır: ikili (binary) ve çok sınıflı (multi-class). İkili sınıflandırma yalnızca iki sınıftan oluşur ve 2x2'lik bir matrisle temsil edilirken; çok sınıflı sınıflandırma üç veya daha fazla sınıftan oluşur ve 3x3, 4x4 gibi daha büyük boyutlarda matrislerle ifade edilir.

- **Doğruluk (Accuracy)**

Doğruluk, modelin doğru sınıflandırdığı örneklerin, toplam örnek sayısına oranı olarak tanımlanır. Çok sınıflı sınıflandırma problemlerinde en sık kullanılan temel başarı ölçütlerinden biridir. Ancak sınıflar arasında belirgin dengesizlikler varsa, modelin başarısını olduğundan yüksek gösterebilir. Bu nedenle doğruluk, özellikle dengesiz veri setlerinde tek başına yeterli bir değerlendirme kriteri olmayabilir ve diğer performans metrikleriyle birlikte yorumlanmalıdır. Bu durum, Denklem (3.1)'de gösterildiği gibi ifade edilir.

$$Doğruluk = \frac{Doğru\ Tahmin\ Sayısı}{Toplam\ Tahmin\ Sayısı} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

- **Kesinlik (Precision)**

Kesinlik, modelin pozitif olarak tahmin ettiği örneklerin, gerçekten o sınıfa ait olma oranını gösterir. Başka bir ifadeyle, modelin yaptığı doğru pozitif tahminlerin, tüm pozitif tahminlere oranıdır. Özellikle yanlış pozitiflerin (False Positive) kritik olduğu durumlarda, örneğin tıbbi teşhis veya güvenlik uygulamalarında, kesinlik önemli bir değerlendirme ölçütüdür. Bu durum, Denklem (3.2)'de gösterildiği gibi ifade edilir.

$$Kesinlik = \frac{TP}{TP + FP} \tag{3.2}$$

- **Duyarlılık (Recall)**

Duyarlılık, modelin bir sınıfa ait olan gerçek örneklerin ne kadarını doğru bir şekilde tespit edebildiğini gösterir. Diğer bir ifadeyle, doğru pozitif tahminlerin, o sınıfa ait toplam gerçek örneklere oranıdır. Özellikle yanlış negatiflerin (False Negative) kritik olduğu durumlarda — örneğin hastalık taraması veya tehlike algılama sistemlerinde — duyarlılık önemli bir başarı ölçütü olarak öne çıkar. Bu durum, Denklem (3.3)'de gösterildiği gibi ifade edilir.

$$Duyarlılık = \frac{TP}{TP + FN} \quad (3.3)$$

- **F1-Skoru (F1-Score)**

F1-Skoru, modelin kesinlik ve duyarlılık değerlerinin harmonik ortalaması olarak tanımlanır. Bu metrik, iki ölçüt arasında denge kurarak her ikisinin de önemli olduğu durumlarda daha doğru bir performans değerlendirmesi sağlar. Özellikle sınıf dağılımının dengesiz olduğu veri setlerinde, F1-Skoru model başarısını daha adil bir şekilde yansıtır. Bu durum, Denklem (3.4)'de gösterildiği gibi ifade edilir.

$$F1 - Skoru = 2 * \frac{Kesinlik * Duyarlılık}{Kesinlik + Duyarlılık} \quad (3.4)$$

Bu metriklerin her biri, modelin farklı yönlerdeki başarısını ortaya koymakta ve özellikle dengesiz veri setleri ile yapılan sınıflandırmalarda daha güvenilir ve ayrıntılı değerlendirmeler sunmaktadır. Bu bağlamda, sistemin genel doğruluğu ölçülürken aynı zamanda her sınıf için ayrıntılı sınıflandırma raporu da elde edilmiştir. Böylece modelin sadece genel başarımı değil, her dil sınıfına özel tahmin yeteneği de analiz edilmiştir.

3.6.2. Öğrenme eğrileri

Öğrenme eğrileri, sınıflandırma modellerinin eğitim sürecindeki performanslarını değerlendirmek amacıyla kullanılan temel araçlardandır. Bu eğriler genellikle doğruluk (accuracy) ve kayıp (loss) eğrileri olmak üzere iki ana başlıkta incelenir. Doğruluk eğrileri, modelin eğitim ve doğrulama verileri üzerindeki başarı düzeyini zaman içinde gösterirken; kayıp eğrileri, modelin hata oranlarını aynı veri kümeleri üzerinden takip eder. Bu çalışma kapsamında, modelin öğrenme sürecini ve genelleme yeteneğini değerlendirmek amacıyla hem eğitim hem de doğrulama setlerine ait doğruluk ve kayıp eğrileri analiz edilmiştir.

- **Doğruluk Eğrisi:** Bu grafik genellikle iki bileşenden oluşur: eğitim doğruluğu ve doğrulama doğruluğu. Eğitim doğruluğu, modelin her epoch sonunda eğitim verisi üzerindeki performansını gösterirken; doğrulama doğruluğu, modelin daha önce görmediği veriler üzerindeki başarısını yansıtır. Bu eğri, modelin öğrenme süreciyle birlikte, bilinmeyen veriler üzerindeki genelleme yeteneği hakkında da bilgi verir. Genellikle x ekseninde epoch sayısı, y ekseninde ise doğruluk değeri yer alır.
- **Kayıp Eğrisi:** Kayıp eğrisi de doğruluk eğrisi gibi iki bileşenden oluşur: eğitim kaybı ve doğrulama kaybı. Eğitim kaybı, modelin eğitim verileri üzerindeki hata miktarını, doğrulama kaybı ise modelin yeni veriler üzerindeki hata oranını gösterir.

Bu grafik, modelin ne derece "öğrendiği" ya da "ezberlediği" hakkında önemli ipuçları sunar. Tıpkı doğruluk eğrisinde olduğu gibi, x ekseninde epoch sayısı, y ekseninde ise kayıp değeri yer alır.

Bu eğrilerin doğru yorumlanması, modelin aşırı öğrenme ya da yetersiz öğrenme gibi sorunlarının tespit edilmesinde kritik rol oynar. Aşağıda bu grafiklerin yorumlanmasına yönelik temel noktalar özetlenmiştir:

- Eğitim ve doğrulama eğrilerinin birbirine yakın seyretmesi idealdir.
- Her iki eğri de düşük performans sergiliyorsa, model yüksek olasılıkla az öğrenme problemine sahiptir. Bu durum, modelin fazla basit olması, yeterli özellik içermemesi veya aşırı regularizasyon uygulanmasıyla ilişkili olabilir.
- Eğitim doğruluğu hızla yükselirken doğrulama doğruluğu geride kalıyorsa, bu genellikle aşırı öğrenme işaretidir; model eğitim verisini ezberlemekte, ancak yeni veriler üzerinde yeterli genelleme yapamamaktadır. Bu durum, modelin karmaşıklığının yüksekliği ya da veri miktarının yetersizliğiyle ilişkilendirilebilir.
- Eğitim ve doğrulama kayıp eğrilerinin kesişim noktasına dikkat edilmelidir. Doğrulama kaybı azalmayı bırakıp artmaya başladığı anda eğitim süreci sonlandırılmalıdır. Bu noktadan önce eğitim durdurulursa model yetersiz öğrenmiş olur; daha geç durdurulursa modelin aşırı öğrenme eğilimi artar.

Bu bağlamda, öğrenme eğrileri yalnızca modelin başarısını izlemekle kalmaz, aynı zamanda bias/variance dengesinin değerlendirilmesine de katkı sunar. Özellikle doğrulama kaybındaki artış eğilimi, modelin genelleme kabiliyetinin sınırlandığını ve durdurulması gerektiğini gösterir. Böylece modelin performansı optimize edilirken, aşırı öğrenmenin önüne geçilmiş olur [38].

4. DENEYLER VE ANALİZLER

Derin öğrenmeye dayalı sınıflandırma sistemleri, yüksek doğruluk ve esneklik sağlayan yapıları sayesinde konuşma tanıma ve dil sınıflandırma uygulamalarında yaygın olarak kullanılmaktadır. Bu çalışmada, konuşulan dillerin otomatik olarak tanımlanması amacıyla dört farklı derin öğrenme tabanlı mimari uygulanmıştır: CNN, BiLSTM, Transformer ve Conformer. Bu yapıların tamamı, ses sinyallerinden elde edilen MFCC temsillerini giriş olarak almakta ve bu temsiller üzerinden dil-spesifik akustik örüntüleri öğrenmeyi hedeflemektedir.

Model eğitiminin ilk aşaması, verinin hazırlanması aşamasıdır. Bu kapsamda her dil için toplanan ses dosyaları, gürültü azaltması ve MFCC çıkarımı gibi önişleme adımlarından geçirilmekte olup, elde edilen veriler eğitim, doğrulama ve test setlerine ayrılmıştır. Ses temelli verilerin kare biçimli grid temsillerine dönüştürülmesi, bu verilerin CNN gibi örüntü tanıma algoritmalarında işlenebilir hale gelmesini sağlamıştır. Giriş verileri, MFCC matrisleri şeklinde normalize edilerek modelin giriş katmanına aktarılmıştır.

BiLSTM mimarisi, konuşma sinyallerinden çıkarılan özniteliklerin zamansal örüntülerini öğrenmeye yönelik olarak tasarlanmıştır. Modelde ilk olarak bir Conv2D katmanı ile düşük seviyeli uzamsal özellikler elde edilmiş, ardından bu çıktı bir yeniden şekillendirme işlemiyle zaman serisi formatına dönüştürülerek iki ardışık BiLSTM katmanına aktarılmıştır. Bu katmanlar, ses verisindeki bağlamsal ilişkileri hem ileri hem geri yönde analiz ederek güçlü zaman bağımlılıkları yakalamaktadır. Sonrasında gelen tam bağlantılı katman (Dense) ve Dropout birimi ile aşırı öğrenme engellenmiş; son çıkışta yer alan softmax aktivasyonlu katman ile model, girişin hangi dil sınıfına ait olduğunu tahmin etmektedir.

Konvolüsyonel Sinir Ağı (CNN) mimarisi, üç adet konvolüsyon ve havuzlama bloğundan oluşacak şekilde tasarlanmıştır. Her blok, uzamsal özellikleri çıkarmaya yarayan bir Conv2D katmanı, ardından boyut küçültme ve aşırı öğrenmeyi önleme amacıyla bir MaxPooling2D ve Dropout katmanı içermektedir. Bu katmanların ardından elde edilen özellik haritaları, Flatten katmanı aracılığıyla vektöre dönüştürülmüş ve iki adet tam bağlantılı (Dense) katmanla sınıflandırma yapılmıştır. Son çıkış katmanında softmax aktivasyonu ile model, girişin hangi dil sınıfına ait olduğunu tahmin etmektedir.

Transformer tabanlı model ise, sıralı verilerdeki bağlamsal ilişkileri öğrenme yeteneği ile ön plana çıkmaktadır. Bu modelde giriş olarak alınan MFCC matrislerine pozisyonel

kodlama uygulanmakta ve bu sayede zaman boyutundaki sıralı yapı korunmaktadır. Ardından gelen çok başlı dikkat (multi-head attention) ve ileri beslemeli katmanlardan oluşan üç adet encoder bloğu, dilin yapısal özelliklerini modellemek üzere ardışık biçimde çalışmaktadır. Encoder bloklarının ardından global ortalama havuzlama, tam bağlantılı katman ve softmax çıkış katmanı yer almaktadır. Böylece, model sıralı zaman adımlarındaki örüntüleri öğrenerek doğru dil tahmininde bulunmaktadır.

Conformer tabanlı model mimarisi, konuşma verilerinden çıkarılan özniteliklerin hem yerel hem de küresel bağlam ilişkilerini aynı anda öğrenebilecek şekilde yapılandırılmıştır. Modelin girişinde uygulanan konvolüsyon katmanı ile temel uzamsal özellikler elde edilmekte, ardından giriş tensörü zaman serisi formatına dönüştürülerek Conformer bloklarına aktarılmaktadır. Bu bloklar; çok başlı dikkat mekanizması, konvolüsyonel işlemler ve iki aşamalı beslemeli sinir ağı (feed-forward) yapısını bir araya getirerek dil-specifik örüntülerin çok katmanlı biçimde öğrenilmesini sağlamaktadır. Global ortalama havuzlama ve tam bağlantılı katmanların ardından, softmax aktivasyonlu çıkış birimi ile model, giriş ses örneğini beş farklı dil sınıfından birine ait olacak şekilde sınıflandırmaktadır.

Modelin eğitimi, etiketlenmiş bir veri kümesi üzerinde gerçekleştirilmiştir. Eğitim süreci; giriş verilerinin ileri besleme (feedforward) yoluyla modele aktarılması, model çıktısının gerçek etiketlerle karşılaştırılması ve oluşan hatanın geri yayılım (backpropagation) algoritması ile ağ ağırlıklarını güncellemesi adımlarını içermektedir. Optimizasyon amacıyla yaygın olarak kullanılan Adam algoritması tercih edilmiş ve öğrenme oranı sabit 0.0001 olarak belirlenmiştir. Eğitim süresince, her epoch sonunda doğrulama verisi kullanılarak modelin genel performansı izlenmiş; aşırı öğrenme (overfitting) etkilerini azaltmak amacıyla dropout gibi düzenleme (regularization) yöntemlerinden faydalanılmıştır.

Modelin nihai performansı, ayrılmış test verisi üzerinde değerlendirilmiş; doğruluk (accuracy), duyarlılık (recall), kesinlik (precision) ve F1-skoru gibi metriklerle raporlanmıştır. Böylece, konuşma sinyalleri üzerinden öğrenilen özelliklerin sınıflandırma üzerindeki etkisi her iki mimari için karşılaştırmalı olarak analiz edilebilmiştir.

4.1. Hiperparametre Ayarları

Derin öğrenme modellerinin performansını doğrudan etkileyen en kritik unsurlardan biri, doğru hiperparametrelerin belirlenmesidir. Bu çalışmada, CNN, BiLSTM, Transformer ve Conformer mimarilerine dayalı dört farklı model geliştirilmiştir. Modellerin öğrenme başarımını en üst düzeye çıkarmak amacıyla; öğrenme oranı, epoch sayısı, batch size,

dropout oranı ve katman boyutları gibi çeşitli hiperparametreler optimize edilmiştir.

Model eğitimi sırasında, Adam optimizasyon algoritması tercih edilmiş ve tüm deneylerde sabit bir öğrenme oranı (0.0001) kullanılmıştır. Epoch sayısı 25 olarak belirlenmiş, erken durdurma tekniği uygulanmamıştır. Eğitim verilerinin %70'i modelin öğrenmesi için, %15'i doğrulama, %15'i test amacıyla ayrılmıştır. Batch boyutu olarak hem bellek optimizasyonu hem de modelin kararlılığı göz önünde bulundurularak 64 seçilmiştir. Aşırı öğrenmenin önlenmesi amacıyla derin öğrenme metotlarında dropout katmanları eklenmiştir. MFCC özelliklerinin zaman-boyutlu temsili 2 boyutlu tensörlere dönüştürülerek modellerine giriş olarak verilmiştir.

4.1.1. BiLSTM tabanlı model için hipermetre ayarları

Bu çalışmada, zaman serisi özellikleri taşıyan konuşma sinyallerinin dil sınıflandırma problemine uygulanması amacıyla BiLSTM tabanlı model mimarisi tercih edilmiştir. LSTM yapıları, sıralı veri içerisindeki uzun vadeli bağımlılıkları öğrenme yetenekleri sayesinde, özellikle konuşma ve dil tanıma gibi zaman serisi problemlerinde yüksek başarı sağlamaktadır.

Modelin başarımını doğrudan etkileyen hiperparametreler; LSTM hücre boyutları, dropout oranı, fully-connected katman yapısı ve öğrenme oranı gibi temel bileşenler dikkate alınarak yapılandırılmıştır. Giriş verisi olarak kullanılan MFCC temsilleri, 40 katsayı ve 150 kareden oluşan sabit boyutta özellik matrisleri halinde modellenmiş ve böylece ses sinyalinin hem frekans hem de zamansal içerikleri korunmuştur.

Model mimarisinde ilk olarak 16 filtreli ve 3×3 çekirdek boyutuna sahip Conv2D katmanı ile düşük seviyeli uzaysal öznitelik çıkarımı gerçekleştirilmiştir. Conv2D katmanının ardından uygulanan batch normalization işlemi ile eğitim stabilitesi artırılmıştır. Konvolüsyon katmanı sonrası çıkış, zaman ve özellik boyutlarını birleştirecek şekilde yeniden şekillendirilmiş (reshape) ve sıralı özellik vektörlerine dönüştürülmüştür.

Bu sıralı özellikler, BiLSTM bloklarına giriş olarak verilmiştir. İlk BiLSTM katmanında 128, ikinci BiLSTM katmanında ise 64 hücre birimi kullanılmış ve böylece hem geçmiş hem de gelecek zaman bağımlılıkları eş zamanlı olarak öğrenilmiştir. LSTM blokları sonrasında 128 nöron içeren bir fully-connected katman uygulanmış ve ardından %30 oranında dropout uygulanarak aşırı öğrenme riski azaltılmıştır. Çıkış katmanı ise softmax aktivasyon fonksiyonuna sahip 6 sınıftan oluşmaktadır. Türkçe'siz veri seti için sınıf sayı 5 olarak düzenlenmiştir. Özet olarak Tablo 4.1'de BiLSTM tabanlı model için kullanılan hiperparametre değerleri verilmiştir.

Tablo 4.1. BiLSTM Tabanlı Modelin Hiperparametre Ayarları

Hiperparametre	**Değer**	**Açıklama**
Giriş boyutu	(150, 40, 1)	MFCC matris boyutu
CNN filtre sayısı	16	Conv2D (3×3 çekirdek boyutu, ReLU aktivasyon)
BiLSTM katmanları	128, 64	BiLSTM (İki katmanlı)
Dense katman	128 → 6 (softmax)	Tam bağlantılı sınıflandırma katmanı
Dropout oranı	0.3	Aşırı öğrenmeyi azaltmak için
Epoch sayısı	25	Model eğitimi boyunca
Batch boyutu	64	Eğitim verisi yükleme boyutu
Öğrenme oranı	0.0001	Adam optimizasyon algoritması
Optimizasyon algoritması	Adam	Adaptif moment tahmini
Kayıp fonksiyonu	Categorical Crossentropy	Çok sınıflı sınıflandırma için
Değerlendirme metriği	Accuracy	Doğruluk bazlı değerlendirme

4.1.2. CNN tabanlı model için hipermetre ayarları

CNN mimarisi, giriş olarak verilen MFCC özellik matrislerinden örüntüleri çıkarmaya yönelik olarak yapılandırılmıştır. Modelin temel yapısı üç adet evrişimli katman, bu katmanların ardından yer alan maksimum havuzlama (MaxPooling) ve dropout katmanları ile iki adet yoğun (fully connected) katmandan oluşmaktadır. Bu yapı sayesinde, farklı dillerin fonetik yapıları ve akustik örüntüleri etkili biçimde öğrenilmiştir. Modelin hiperparametreleri Tablo 4.2'de özetlenmiştir.

Tablo 4.2. CNN Tabanlı Modelin Hiperparametre Ayarları

Hiperparametre	**Değer**	**Açıklama**
Giriş boyutu	(150, 40, 1)	MFCC matris boyutu
Conv2D filtreleri	64, 128, 256	5×5 ve 3×3 çekirdek boyutlarıyla
Aktivasyon fonksiyonu	ReLU	Tüm katmanlarda kullanıldı
MaxPooling boyutu	(2, 2)	İlk iki Conv katmanından sonra

Tablo 4.2. devam ediyor.

Hiperparametre	Değer	Açıklama
Dropout oranı	0.2, 0.3, 0.4	Aşırı öğrenmeyi azaltmak için üç farklı seviye
Dense katmanlar	512→256→6 (softmax) 512→256→5 (softmax)	Son katman sınıf sayısı kadar, İlki Dengeli ve Dengesiz veri seti İkincisi Türkçe'siz veri seti için
Epoch sayısı	25	Model eğitimi boyunca
Batch boyutu	64	Eğitim verisi yükleme boyutu
Öğrenme oranı	0.0001	Adam optimizasyon algoritması ile
Optimizasyon algoritması	Adam	Adaptif moment tahmini
Kayıp fonksiyonu	Categorical Crossentropy	Çok sınıflı sınıflandırma için
Değerlendirme metriği	Accuracy	Doğruluk bazlı değerlendirme

Bu yapı sayesinde model, eğitim verisi üzerinden genelleme kabiliyetini yüksek düzeyde koruyarak öğrenme sürecini tamamlamıştır. Dropout katmanları, aşırı öğrenmenin önüne geçilmesinde etkili olmuş; derinlik seviyesi yeterli tutulduğu için farklı diller arasındaki ayrıştırıcı özellikler etkin biçimde çıkarılmıştır.

4.1.3. Transformer encoder tabanlı model için hipermetre ayarları

Transformer Encoder mimarisi, özellikle doğal dil işleme ve zaman serisi verilerinde gösterdiği yüksek başarı nedeniyle bu çalışmada konuşma temelli dil sınıflandırma amacıyla tercih edilmiştir. Modelin öğrenme başarısını doğrudan etkileyen hiperparametreler; encoder derinliği, dikkat başlık sayısı, dropout oranı ve feed-forward katman genişliği gibi temel bileşenler göz önünde bulundurularak yapılandırılmıştır.

Bu mimaride giriş verisi olan MFCC temsillerine sinüzoidal temelli sabit pozisyonel kodlama uygulanmış ve böylece zaman eksenindeki örüntülerin korunması sağlanmıştır [23]. Encoder bloklarında ise çok başlı kendine dikkat (multi-head self-attention) mekanizması kullanılarak farklı temsillerin paralel olarak öğrenilmesine olanak tanınmıştır. Bu yapı, konuşma verisi gibi zamansal örüntüler içeren sinyallerin işlenmesinde yüksek

doğruluk sağlamaktadır.

Modelde toplam 3 adet encoder bloğu, her biri içinde 4 başlıklı self-attention mekanizması ve 128 boyutlu feed-forward ağ yapılandırılmıştır. Encoder çıkışına uygulanan GlobalAveragePooling1D katmanı ile zaman boyutu boyunca ortalama alınarak temsil yoğunluğu azaltılmış, ardından dense katmanları ile sınıflandırma yapılmıştır.

Transformer Encoder yapısına dayalı dil tanıma sistemlerinde, dikkat başlık sayısının 4–8 aralığında tutulmasının, modelin genelleme yeteneğini artırdığı literatürde çeşitli çalışmalarda da rapor edilmiştir [39], [40]. Tablo 4.3'de Transformer Encoder için kullanılan hiperparametre değerleri verilmiştir.

Tablo 4.3. Transformer Encoder Tabanlı Modelin Hiperparametre Ayarları

Hiperparametre	**Değer**	**Açıklama**
Giriş boyutu	(150, 40)	MFCC matrisinden gelen giriş (zaman × katsayı)
Positional Encoding	Sinüs/kosinüs temelli	Sabit, öğrenilmeyen pozisyon kodlaması
Transformer katman sayısı	3	Encoder bloğu tekrar sayısı
Başlık sayısı (num_heads)	4	Çok başlı dikkat mekanizması
d_model (özellik boyutu)	40	Girişteki MFCC özellik boyutu
Feed-forward katman boyutu	128	Her encoder içindeki yoğun katman (FFN) genişliği
Dropout oranı	0.3	Encoder blokları içinde ve çıkış katmanlarında uygulanır
GlobalAveragePooling	Var	Encoder çıkışlarının ortalaması alınır
Dense katmanlar	64 → 6 (softmax)	Sınıflandırma için çıkış katmanları
Epoch sayısı	25	Eğitim süreci boyunca
Batch boyutu	64	Her eğitim adımında işlenen örnek sayısı
Öğrenme oranı	0.0001	Adam optimizasyon algoritması ile

Tablo 4.3. devam ediyor.

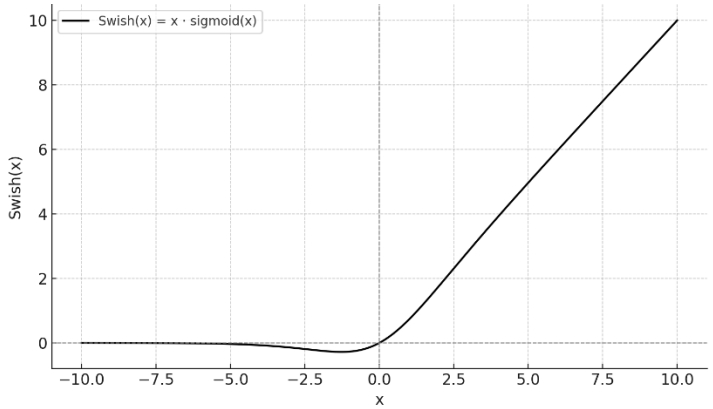
Hiperparametre	Değer	Açıklama
Optimizasyon algoritması	Adam	Adaptif öğrenme algoritması
Kayıp fonksiyonu	Categorical Crossentropy	Çok sınıflı sınıflandırma için
Değerlendirme metriği	Accuracy	Model doğruluğu üzerinden izleme yapılır

4.1.4. Conformer encoder tabanlı model için Hipermetre ayarları

Bu çalışmada, ses sinyallerinin hem yerel hem de küresel zaman bağımlılıklarını eş zamanlı olarak modelleyebilmek amacıyla Conformer tabanlı mimari tercih edilmiştir. Conformer yapıları, hem dikkat mekanizmasının uzun menzilli bağımlılıkları yakalama yeteneğini hem de konvolüsyonel katmanların yerel zaman örüntülerini öğrenme kapasitesini birleştirerek konuşma ve dil tanıma gibi görevlerde yüksek doğruluk sağlamaktadır.

Modelin öğrenme başarısını doğrudan etkileyen hiperparametreler; encoder blok sayısı, dikkat başlık sayısı, feed-forward genişliği, dropout oranları ve konvolüsyon çekirdek boyutları dikkate alınarak yapılandırılmıştır. Giriş verisi olarak kullanılan MFCC temsilleri, 40 katsayı ve 150 kareden oluşan sabit boyutta öznitelik matrisleri halinde hazırlanmıştır. Gürültü temizleme işlemi için FFT tabanlı spektral filtreleme uygulanmış, ardından MFCC çıkarımı gerçekleştirilmiştir.

Model mimarisi başlangıçta 64 filtreli Conv2D (3x3) katmanı ile düşük seviyeli öznitelik çıkartımı yapmış; sonrasında batch normalization ile stabilite sağlanmıştır. Konvolüsyon çıkışı zaman ve özellik boyutları birleştirilerek sıralı giriş formatına dönüştürülmüştür. Buradan itibaren Conformer encoder blokları devreye girmiştir. Toplamda 4 adet Conformer bloğu yapılandırılmış; her blok içerisinde 4 başlıklı multi-head self-attention, residual scaling ile beslemeli feed-forward katmanları ve depthwise konvolüsyon modülleri yer almıştır. Feed-forward katmanlarında Swish aktivasyon fonksiyonu kullanılmış ve katman genişliği 256 olarak belirlenmiştir. Swish, doğrusal olmayan ve kendiliğinden öğrenilebilir bir aktivasyon fonksiyonudur. Google tarafından önerilen bu fonksiyon, özellikle derin sinir ağlarında ReLU'ya göre daha yumuşak geçişlere sahiptir ve bazı görevlerde daha iyi performans sağlamaktadır. Negatif değerleri tamamen sıfırlamaz, bu da bilgi kaybını azaltır. Swish fonksiyonu Şekil 4.1' de gösterilmiştir.



Şekil 4.1. Swish Aktivasyon Fonksiyonu

GlobalAveragePooling katmanı sonrası özellik yoğunluğu azaltılmış ve 256 nöronlu Dense katmanı uygulanmıştır. Son katmanda softmax aktivasyon fonksiyonuyla 6 sınıflı çıktı üretilmiştir. Türkçe'siz veri seti için 5 sınıflı çıktı üretilmektedir. Tablo 4.4'te Conformer Encoder modeli için kullanılan hiperparametre değerleri verilmiştir.

Tablo 4.4. Conformer Encoder Tabanlı Modelin Hiperparametre Ayarları

Hiperparametre	**Değer**	**Açıklama**
Giriş boyutu	(150, 40, 1)	MFCC matris boyutu
Conv2D filtre sayısı	64	Başlangıç evrişim katmanı
Conformer blok sayısı	4	Encoder blok adedi
Attention başlık sayısı	4	Multi-head attention başlık sayısı
Feed-forward genişliği	256	Conformer FFN katman genişliği
Depthwise Conv çekirdek	31	Yerel zaman örüntüsü öğrenimi için
Dropout oranı	0.3	Aşırı öğrenmeyi azaltmak için
Dense katman	256 → 6 (softmax)	Tam bağlantılı sınıflandırıcı katmanları
Epoch sayısı	25	Eğitim süresi
Batch boyutu	64	Eğitim verisi yükleme boyutu

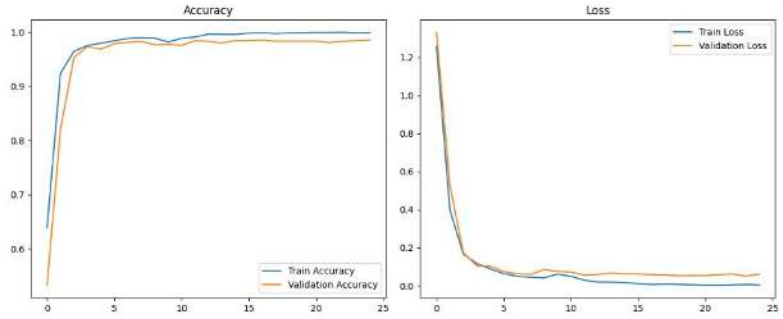
Tablo 4.4. devam ediyor.

Hiperparametre	Değer	Açıklama
Öğrenme oranı	0.0001	Adam optimizasyonu ile
Optimizasyon algoritması	Adam	Adaptif moment tahmini
Kayıp fonksiyonu	Categorical Crossentropy	Çok sınıflı sınıflandırma
Değerlendirme metriği	Accuracy	Doğruluk bazlı değerlendirme

4.2. Dengeli Veri Seti ile Eğitilen Model Sonuçları

4.2.1. BiLSTM tabanlı eğitilen model sonuçları

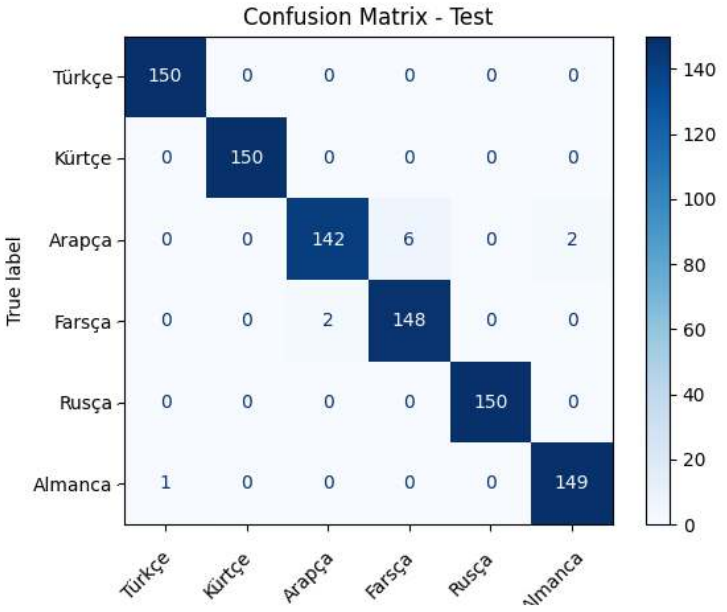
Bu çalışmada önerilen BiLSTM tabanlı model, konuşulan altı farklı dili sınıflandırmak amacıyla eğitilmiştir. Modelin eğitim sürecinde modelin doğruluk (accuracy) ve kayıp (loss) değerleri izlenerek genel başarımı değerlendirilmiştir. Şekil 4.2 (a)'da verilen doğruluk grafiğinde görüldüğü üzere, modelin hem eğitim hem de doğrulama seti üzerindeki doğruluğu yaklaşık 5. epoktan itibaren %95 düzeyine ulaşmıştır. Bu durum, BiLSTM modelinin zaman bağımlı verilerde güçlü bir öğrenme performansı gösterdiğini ve daha önce görmediği doğrulama verisi üzerinde de başarılı genelleme yapabildiğini göstermektedir. Şekil 4.2 (b)'de yer alan kayıp eğrisi incelendiğinde ise, hem eğitim hem de doğrulama setleri için kayıp değerlerinin hızlı ve istikrarlı bir biçimde azaldığı görülmektedir. Eğitim süreci boyunca kayıp değerlerinin birbirine yakın seyretmesi ve önemli bir dalgalanma göstermemesi, modelde aşırı öğrenme eğiliminin olmadığını teyit etmektedir. Bu durum, modelin hem öğrenme hem de genellemesinin dengeli olduğunu ortaya koymaktadır.



(a) (b)

Şekil 4.2. Dengeli veri seti için BiLSTM tabanlı modelin eğitim ve doğrulama grafikleri (a) Doğruluk (Accuracy) grafiği (b) Kayıp (Loss) grafiği

Şekil 4.3'te verilen karmaşıklık matrisi incelendiğinde, modelin sınıflar arası ayrım performansının oldukça yüksek olduğu görülmektedir. Tüm sınıflarda doğru sınıflandırma oranı oldukça yüksek olup, yalnızca Arapça verilerin bir kısmının Farsça ve Almanca ile, Farsça verilerin ise sınırlı ölçüde Arapça ile karıştırıldığı dikkat çekmektedir. Bu durumun, diller arası fonetik yakınlık ve bazı ortak ses örüntülerinden kaynaklandığı düşünülmektedir. Örneğin, Arapça ve Farsça'da yer alan benzer fonem grupları ya da ses aralıkları modelin ayrım yapmasını zorlaştırmış olabilir. Ancak genel tablo değerlendirildiğinde, BiLSTM modelinin dil sınıfları arasında yüksek ayrım başarısı gösterdiği ve sınıflar arası karışmaların çok sınırlı düzeyde kaldığı söylenebilir.

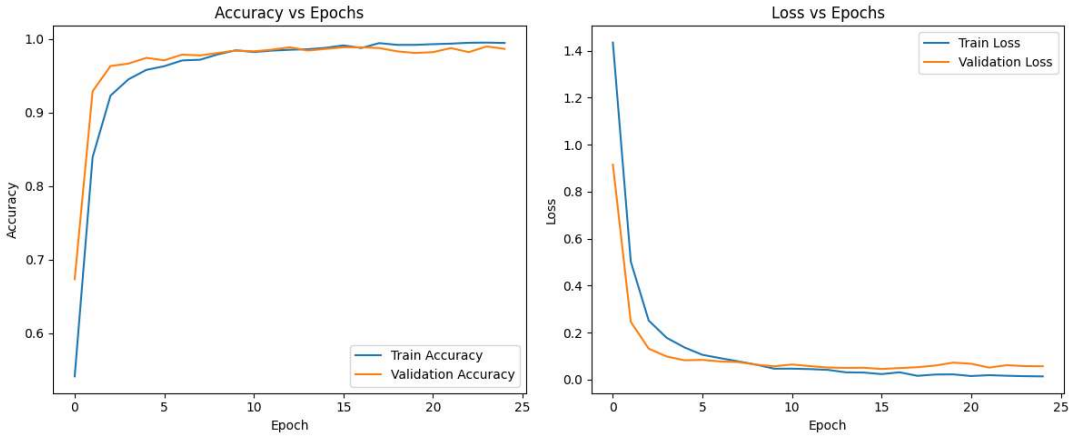


Şekil 4.3. Dengeli veri seti için BiLSTM tabanlı modelin karışıklık matrisi (confusion matrix - test seti)

Modelinin test seti üzerindeki genel doğruluğu %98.78 olarak hesaplanmıştır. Sınıflandırma raporu incelendiğinde, Türkçe sınıfı %100 doğruluk, %100 duyarlılık (recall) ve %100 F1-skor değeri ile en başarılı sonuçları vermiştir. Kürtçe ve Almanca sınıflarında da sırasıyla %99.34 ve %99.01 gibi oldukça yüksek F1-skorları elde edilmiştir. Arapça, Farsça ve Rusça sınıflarında ise birkaç örnek hatalı sınıflandırılmış olmakla birlikte F1-skorları %97 seviyesinin üzerinde kalmıştır. Genel olarak BiLSTM mimarisi, dengeli veri seti üzerinde hem baskın hem de görece azınlıkta kalan sınıflar için yüksek tutarlılıkta sınıflandırma başarısı göstermiştir.

4.2.2. CNN tabanlı eğitilen model sonuçları

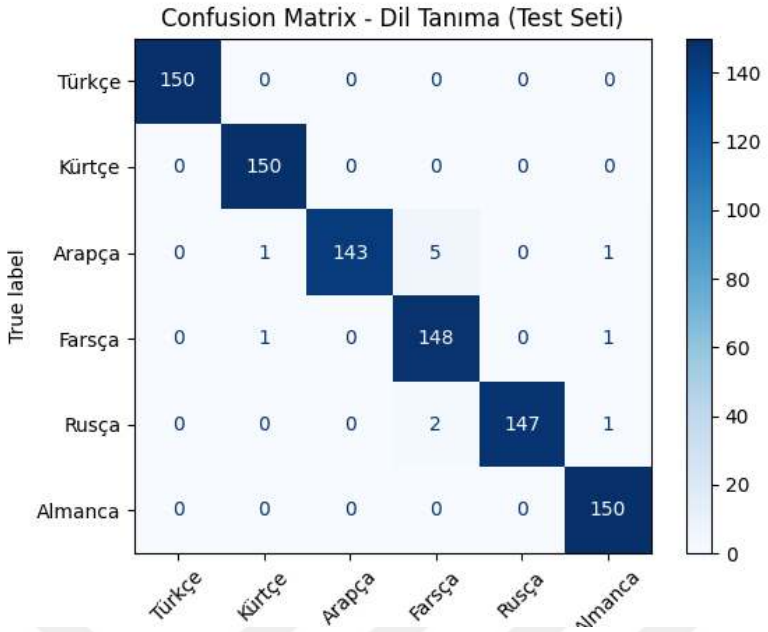
Bu çalışmada önerilen CNN tabanlı model, konuşulan altı farklı dili sınıflandırmak amacıyla eğitilmiştir. Şekil 4.4 (a)'de yer alan doğruluk grafiğinde, modelin hem eğitim hem de doğrulama seti üzerindeki başarımı yaklaşık 10. epoktan itibaren %95 düzeyine ulaşmıştır. Bu durum, modelin eğitilen veri üzerinde olduğu kadar daha önce görmediği doğrulama verisi üzerinde de iyi genelleme yapabildiğini göstermektedir. Şekil 4.4 (b)'de verilen kayıp eğrisi ise hem eğitim hem de doğrulama seti için kararlı bir şekilde azalarak aşırı öğrenme belirtileri göstermediğini doğrulamaktadır.



(a) (b)

Şekil 4.4. Dengeli veri seti için CNN tabanlı modelin eğitim ve doğrulama grafikleri (a) Doğruluk (Accuracy) grafiği (b) Kayıp (Loss) grafiği

Şekil 4.5'te verilen karmaşıklık matrisi incelendiğinde, Arapça verilerin sınırlı oranda Farsça ve Almanca ile karıştırıldığı, benzer şekilde Farsça sınıfında Kürtçe ve Almanca ile karışmaların gözlemlendiği dikkat çekmektedir. Bunun olası nedenleri arasında dillerin fonetik benzerlikleri ve bazı ses öbeklerinin örtüşmesi yer almaktadır. Ancak buna rağmen karışıklıklar oldukça sınırlı düzeyde kalmıştır.

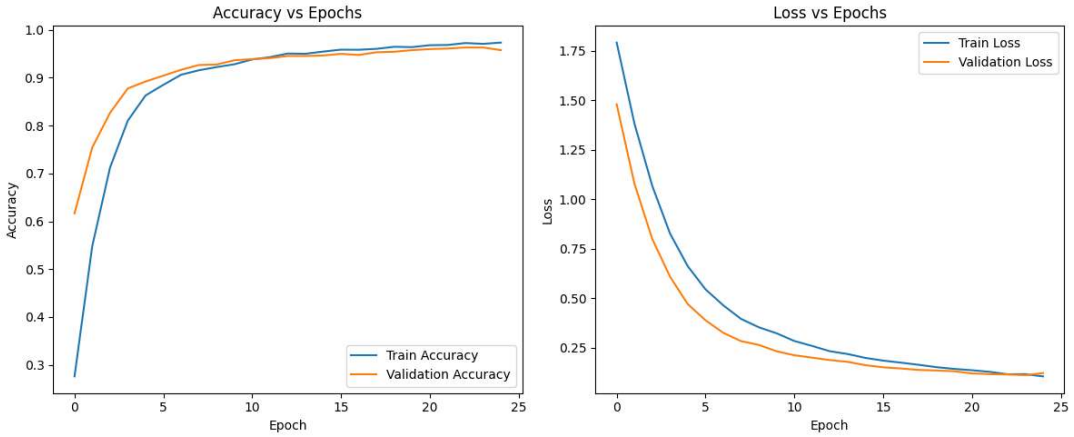


Şekil 4.5. Dengeli veri seti için CNN tabanlı modelin karışıklık matrisi (confusion matrix - test seti)

Modelin test seti üzerindeki genel doğruluğu %98.67 olarak hesaplanmıştır. Sınıflandırma raporuna göre, Türkçe dili %100 kesinlik, duyarlılık ve F1-skor değerleri ile en başarılı sınıflandırılan dil olmuştur. Bunun yanında Almanca ve Kürtçe sınıflarında da %99'un üzerinde F1-skor değerleri elde edilmiştir. Arapça, Farsça ve Rusça gibi sınıflarda ise birkaç örnek hatalı sınıflandırılmış olsa da genel olarak F1-skor değerleri %97'nin üzerindedir.

4.2.3. Transformer encoder tabanlı eğitilen model sonuçları

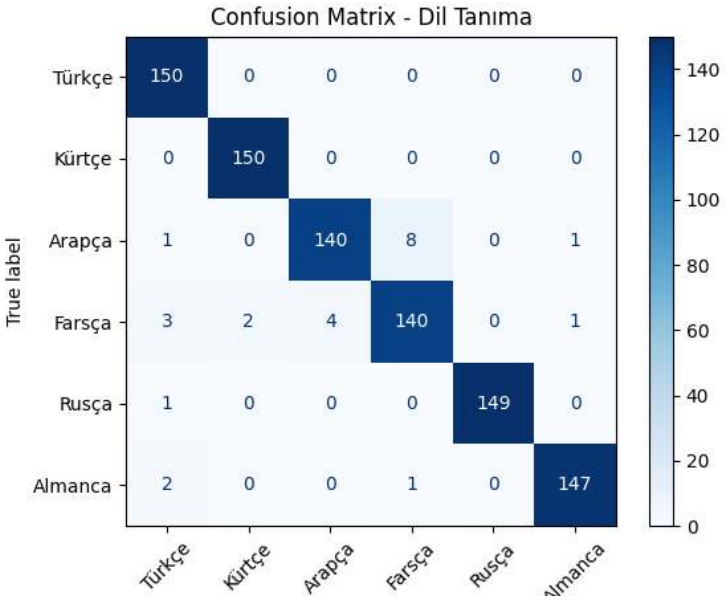
Model, eğitim süreci boyunca hızlı ve kararlı bir şekilde öğrenmiş; özellikle doğrulama kümesindeki başarım, aşırı öğrenme eğilimi göstermeden yüksek seviyede korunmuştur. Şekil 4.6 (a)'da modelin eğitim ve doğrulama doğruluk değerlerinin epoklarına göre değişimi, Şekil 4.6 (b)'de ise eğitim ve doğrulama kayıp değerlerinin eğrileri gösterilmiştir. Eğitim doğruluğu yaklaşık olarak %98'e ulaşırken, doğrulama doğruluğu %97.33 seviyelerinde sabitlenmiş ve bu durum modelin genel performansının dengeli olduğunu ortaya koymuştur.



(a) (b)

Şekil 4.6. Dengeli veri seti için transformer encoder tabanlı modelin eğitim ve doğrulama grafikleri (a) Doğruluk (Accuracy) grafiği (b) Kayıp (Loss) grafiği

Sınıflandırma başarımı, Şekil 4.7'te yer alan karışıklık matrisi ile desteklenmektedir. Tüm dillerin 150 örneklik test kümesi üzerinden değerlendirildiği bu matrise göre, özellikle Kürtçe ve Türkçe dillerinde hatasız sınıflandırma gerçekleştirilmiştir. Arapça, Farsça ve Almanca sınıflarında ise bazı sınıflar arasında karışıklık gözlemlenmiş ancak bu durum genel başarıyı olumsuz etkilememiştir.



Şekil 4.7. Dengeli veri seti için transformer encoder tabanlı modelin karışıklık matrisi (confusion matrix - test seti)

Sınıflandırma raporuna göre modelin genel doğruluğu %97.33 olup, tüm diller için kesinlik, duarlılık ve F1-skor değerleri oldukça yüksek çıkmıştır. Özellikle Kürtçe için

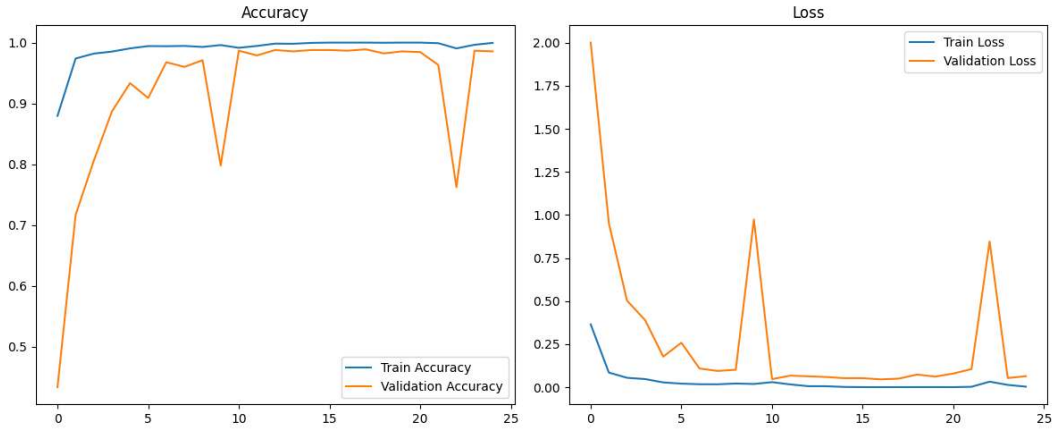
%99.34 F1-skora ulaşılmış; Türkçe için %97.72, Arapça için %95.24, Farsça için %93.65, Rusça için %99.67 ve Almanca için %98.33 F1-score elde edilmiştir. Bu durum, modelin çoğu sınıf için dengeli performans sunduğunu göstermektedir.

Sonuç olarak, Transformer Encoder mimarisiyle eğitilen model; öğrenme hızı, genelleme yetisi ve doğrulama başarımı açısından oldukça iyi performans sergilemiş, özellikle yüksek destekli sınıflarda başarılı tahminler yapmıştır. Ancak Farsça ve Arapça sınıflarındaki karışıklıklar, fonetik yakınlık ve sesletim benzerliğinden kaynaklanabilecek hatalara işaret etmektedir.

4.2.4. Conformer encoder tabanlı eğitilen model sonuçları

Bu çalışmada uygulanan Conformer encoder tabanlı model, konuşma sinyallerindeki hem kısa süreli frekans bileşenlerini hem de uzun süreli zaman bağımlılıklarını eşzamanlı olarak işleyebilme yeteneğine sahiptir. Model, evrişimsel katmanların yerel özellikleri çıkarma başarısını, Transformer mekanizmasının küresel bağlam yakalama becerisiyle birleştirmektedir. Şekil 4.8 (a)'da gösterilen doğruluk eğrisi incelendiğinde, modelin eğitim seti üzerinde erken epoklardan itibaren yüksek doğruluk değerlerine ulaştığı, ancak doğrulama seti doğruluğunun bazı epoklarda ani düşüşler yaşadığı gözlemlenmektedir. Bu dalgalanmalar, doğrulama verisi üzerindeki genelleme başarısında zaman zaman kararsızlıklar oluştuğunu göstermektedir. Şekil 4.8 (b)'de yer alan kayıp eğrisi, benzer şekilde eğitim kaybının istikrarlı biçimde azaldığını; ancak doğrulama kaybının bazı epoklarda keskin sıçramalar gösterdiğini ortaya koymaktadır. Özellikle 10. ve 22. epok civarında meydana gelen ani artışlar, modelin bazı dönemlerde aşırı öğrenme eğilimi gösterdiğini ve doğrulama verisine karşı duyarlılığının arttığını göstermektedir.

Genel olarak değerlendirildiğinde, Conformer encoder modeli yüksek öğrenme kapasitesine sahip olmasına rağmen, doğrulama verisi üzerindeki dalgalı performansı dikkat çekmektedir. Bu durum, modelin hiperparametre ayarları, veri dengesizliği veya belirli sınıflarda yetersiz genelleme gibi faktörlerden etkilenmiş olabileceğini düşündürmektedir.

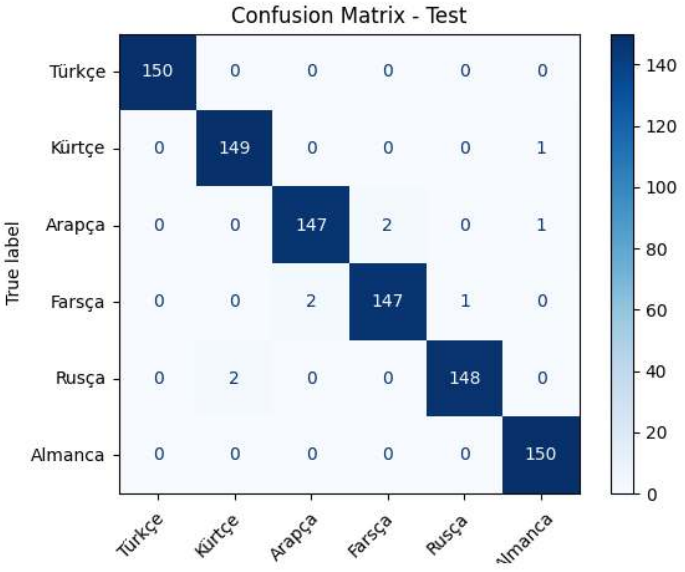


(a) (b)

Şekil 4.8. Dengeli veri seti için conformer encoder tabanlı modelin eğitim ve doğrulama grafikleri (a) Doğruluk (Accuracy) grafiği (b) Kayıp (Loss) grafiği

Şekil 4.9'de sunulan karmaşıklık matrisi, Conformer encoder modelinin sınıflar arası ayırt etme performansının oldukça yüksek olduğunu göstermektedir. Tüm sınıflarda doğru sınıflandırma oranı yüksektir; ancak az sayıda örnekte Arapça–Farsça, Farsça–Arapça ve Rusça, ayrıca Kürtçe–Almanca arasında sınırlı karışmalar gözlemlenmiştir.

Bu karışıklıkların, bazı dil çiftleri arasında bulunan fonetik benzerliklerden veya sınırlı örnek sayısından kaynaklanabileceği düşünülmektedir. Bununla birlikte, genel doğruluk oranı oldukça yüksek olup, modelin farklı dil sınıflarını başarıyla ayırt edebildiği anlaşılmaktadır.



Şekil 4.9. Dengeli veri seti için conformer encoder tabanlı modelin karışıklık matrisi (confusion matrix - test seti)

4.2.5. Modellerin karşılaştırılması

Dengeli veri seti üzerinde gerçekleştirilen analizlerde, dört farklı derin öğrenme mimarisi (CNN, BiLSTM, Transformer ve Conformer) arasında karşılaştırma yapılmıştır. Conformer modeli, %99.00 doğruluk oranıyla tüm modeller arasında en yüksek genel başarıyı sergilemiştir. BiLSTM modeli ise %98.78 doğrulukla ikinci sırada yer almış ve özellikle f1-score değerlerinde 0.98–1.00 aralığında istikrarlı sonuçlar üretmiştir. CNN modeli %98.67 doğrulukla yüksek başarı göstermiş; Farsça ve Arapça sınıflarında bazı karışıklıklar gözlenmiş olsa da genel olarak dengeli sınıflandırma performansı sunmuştur. Transformer modeli ise %97.33 doğrulukla diğer modellerin gerisinde kalmıştır. Özellikle Arapça ve Farsça sınıflarında belirgin sınıf karışıklıkları tespit edilmiştir. Eğitim ve doğrulama eğrileri incelendiğinde, CNN, BiLSTM ve Conformer modelleri kararlı ve hızlı öğrenme göstermiştir. Elde edilen bulgular, Conformer ve BiLSTM yapılarını dengeli veri setleri üzerinde öne çıkan mimariler olarak konumlandırmaktadır.

4.3. Dengesiz Veri Seti ile Eğitilen Model Sonuçları

4.3.1. BiLSTM tabanlı eğitilen model sonuçları

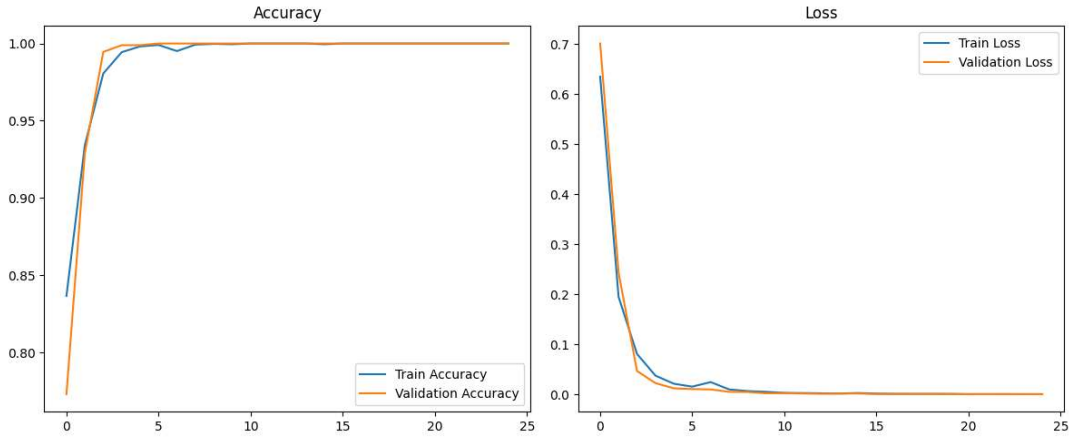
Bu çalışmada kullanılan dengesiz veri setinde, Türkçe yaklaşık 4700 ses dosyası ile veri kümesinin baskın dilini oluştururken, diğer dillerin örnek sayıları belirgin şekilde azdır: Kürtçe 900, Arapça 300, Farsça, Rusça ve Almanca dillerine ait yalnızca 70'er örnek bulunmaktadır. Bu dengesiz yapı, çok dilli gerçek dünya uygulamalarında sıkça karşılaşılan dil dağılımlarını yansıttığından, modelin sınıf bazlı genelleme yeteneğini test etme açısından önemli bir zorluk teşkil etmektedir.

Ancak bu dengesizliğe rağmen, BiLSTM tabanlı modelin eğitim performansı son derece başarılıdır. Şekil 4.10 (a)'da yer alan doğruluk eğrisine göre, hem eğitim hem de doğrulama doğrulukları ilk birkaç epok içinde hızla artarak %99 seviyesine ulaşmış ve bu yüksek doğruluk oranı eğitim süreci boyunca istikrarlı biçimde korunmuştur. Bu durum, modelin baskın sınıfa aşırı uyum göstermeden azınlık sınıfları da öğrenebildiğine işaret etmektedir.

Benzer biçimde, Şekil 4.10 (b)'de gösterilen kayıp eğrilerinde hem eğitim hem de doğrulama verileri için kayıp değerlerinin erken evrelerde keskin şekilde azaldığı ve yaklaşık 5. epoktan itibaren oldukça düşük seviyelerde sabitlendiği görülmektedir. Kayıp değerlerinin birbirine yakın seyretmesi, modelin aşırı öğrenme eğilimi göstermediğini ve sağlam bir

öğrenme süreci gerçekleştirdiğini ortaya koymaktadır.

Sonuç olarak, BiLSTM modelinin çift yönlü yapısı ve zaman serisi özniteliklerini değerlendirme kabiliyeti, onu sınıf dengesizliğine karşı dirençli kılmakta ve dengesiz veri setleri üzerinde dahi yüksek doğruluk ve kararlılıkla çalışmasını sağlamaktadır.

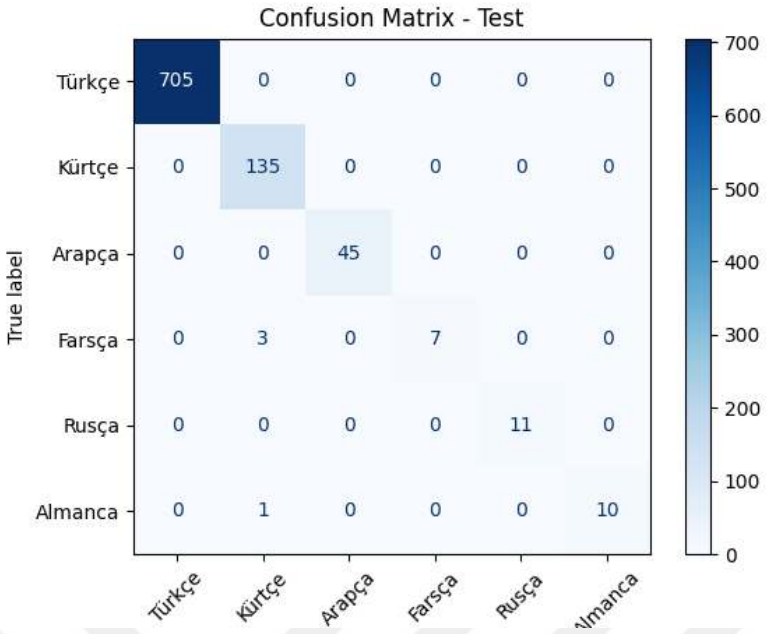


(a) (b)

Şekil 4.10. Dengesiz veri seti için BiLSTM tabanlı modelin eğitim ve doğrulama grafikleri (a) Doğruluk (Accuracy) grafiği (b) Kayıp (Loss) grafiği

Şekil 4.11'da sunulan karmaşıklık matrisi, BiLSTM modelinin dengesiz veri seti üzerindeki sınıflandırma başarımını ortaya koymaktadır. En yüksek doğruluk Türkçe sınıfında elde edilmiş olup, model bu sınıfa ait örneklerin neredeyse tamamını doğru şekilde sınıflandırmıştır. Bu durum, Türkçenin veri setinde baskın biçimde temsil edilmesiyle doğrudan ilişkilidir.

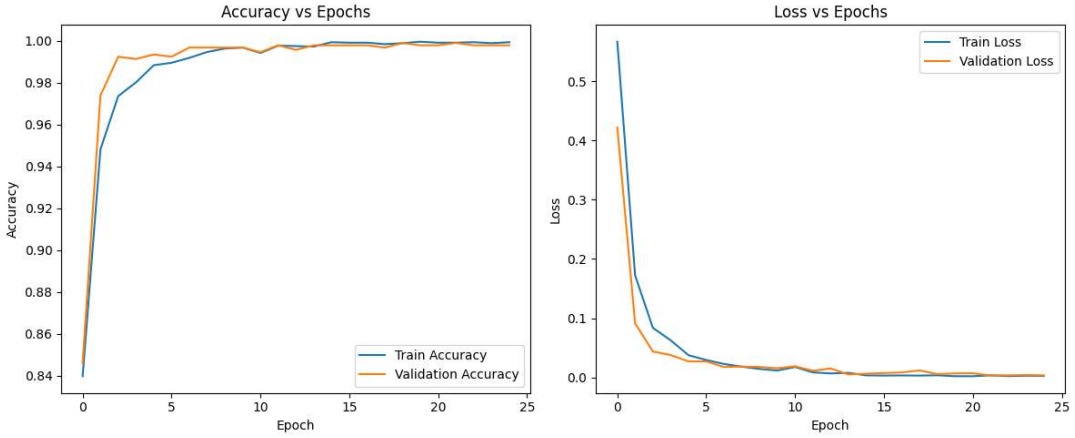
Buna karşın, veri setinde az sayıda yer alan Farsça, Rusça ve Almanca sınıflarındaki örneklerin çoğunluğu da doğru sınıflandırılmıştır. Bu sonuç, modelin sadece baskın sınıfı değil, aynı zamanda daha az temsil edilen sınıfları da etkili bir şekilde öğrenebildiğini ve sınıflar arası ayrım yapma kapasitesinin yüksek olduğunu göstermektedir.



Şekil 4.11. Dengesiz veri seti için BiLSTM tabanlı modelin karışıklık matrisi (confusion matrix - test seti)

4.3.2. CNN tabanlı eğitilen model sonuçları

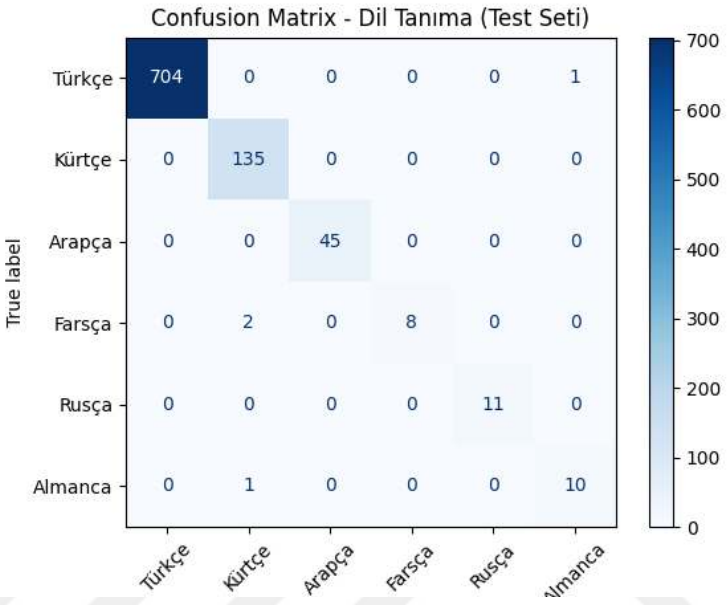
Bu çalışmada modelin çıktıları incelendiğinde, eğitimin genel olarak başarılı olduğu ve yüksek doğruluk elde edildiği görülmektedir. Şekil 4.12'da görüldüğü üzere, eğitim ve doğrulama doğrulukları hızlı bir şekilde artmakta ve yaklaşık 5. epoktan itibaren %99 seviyelerine ulaşmaktadır. Benzer şekilde, kayıp değerleri de belirgin bir şekilde azalarak erken dönemde stabil hale gelmiştir. Bu grafikler, modelin eğitim süreci boyunca istikrarlı bir öğrenme gerçekleştirdiğini ve aşırı öğrenme eğilimi göstermediğini ortaya koymaktadır.



(a) (b)

Şekil 4.12. Dengesiz veri seti için CNN tabanlı modelin eğitim ve doğrulama grafikleri (a) Doğruluk (Accuracy) grafiği (b) Kayıp (Loss) grafiği

Şekil 4.13'te yer alan karışıklık matrisi, modelin en yüksek başarıyı Türkçe sınıfında elde ettiğini göstermektedir (705 örnekten 704'ü doğru). Bu beklenen bir sonuçtur çünkü veri setindeki en fazla temsil edilen dil Türkçedir. Öte yandan, düşük örnek sayısına sahip diller (örneğin Farsça, Rusça ve Almanca), daha düşük sayılarda örneklenmesine rağmen yüksek sınıflandırma doğruluğu göstermiştir. Bu durum, modelin genel olarak sınıflar arası ayrımı başarıyla öğrenebildiğini göstermektedir.



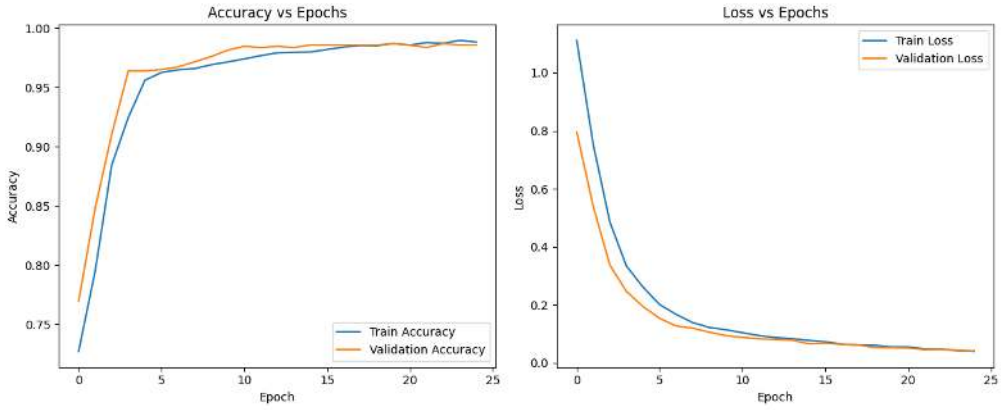
Şekil 4.13. Dengesiz veri seti için CNN tabanlı modelin karışıklık matrisi (confusion matrix - test seti)

Modelin dengesiz veri seti üzerinde elde ettiği sınıflandırma sonuçları oldukça başarılıdır. En yoğun veri grubunu oluşturan Türkçe sınıfı %99.93 F1-skora ulaşırken, Kürtçe ve Arapça gibi diğer yüksek örnek sayısına sahip sınıflarda da benzer yüksek başarılar gözlenmiştir. Veri sayısı düşük olan Farsça, Rusça ve Almanca sınıflarında kısmi performans düşüşleri olsa da genel doğruluk oranı %99.56 olarak hesaplanmıştır. Genel olarak Farsça ve Kürtçe dilleri karıştırılmıştır. Bu sonuçlar, modelin Türkçe sınıfına ağırlık vererek eğitilmesine rağmen diğer sınıflarda da genelleme yapabildiğini ve dil tanıma görevinde etkin çalıştığını göstermektedir.

4.3.3. Transformer encoder tabanlı eğitilen model sonuçları

Transformer Encoder tabanlı modelin dengesiz veri seti üzerindeki performansı değerlendirildiğinde, genel doğruluk oranı %98.80 gibi oldukça yüksek bir değere ulaşmıştır. Eğitim ve doğrulama süreçleri boyunca elde edilen doğruluk ve kayıp grafikleri Şekil 4.14'de görüldüğü gibi, modelin istikrarlı bir şekilde öğrenme gerçekleştirdiğini ve

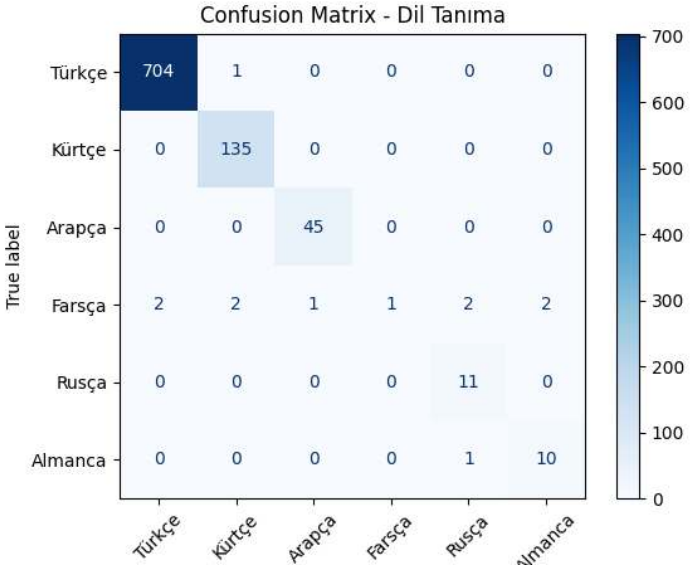
aşırı öğrenme problemi yaşamadığını göstermektedir.



(a) (b)

Şekil 4.14. Dengesiz veri seti için transformer encoder tabanlı modelin eğitim ve doğrulama grafikleri (a) Doğruluk (Accuracy) grafiği (b) Kayıp (Loss) grafiği

Şekil 4.15'da karışıklık matrisi incelendiğinde, modelin Türkçe örnekleri neredeyse kusursuz şekilde tanıdığı; ancak az sayıda örneğe sahip dillerde karışıklıkların daha sık yaşandığı gözlenmektedir. Örneğin, Farsça örneklerin bazıları Kürtçe, Arapça ve Almanca olarak sınıflandırılmıştır. Bu durum, dengesiz veri dağılımının model genellemesine doğrudan etkisi olduğunu göstermektedir.



Şekil 4.15. Dengesiz veri seti için transformer encoder tabanlı modelin karışıklık matrisi (confusion matrix - test seti)

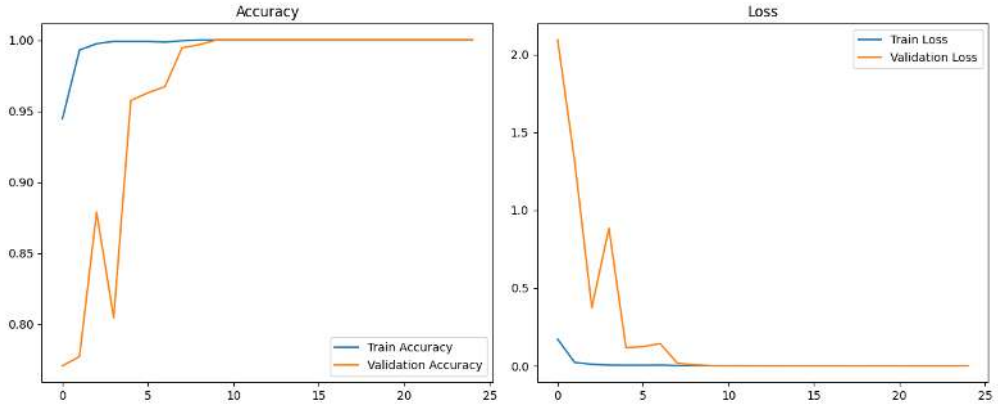
Sınıflandırma raporuna göre model, Türkçe (%99.7 F1-skor), Kürtçe (%98.9) ve Arapça (%98.9) sınıflarında oldukça başarılı sonuçlar üretmiştir. Ancak, veri setinde çok daha az temsil edilen Farsça, Rusça ve Almanca gibi sınıflarda başarı oranı düşmüştür.

Özellikle Farsça sınıfı için f1-skor değeri %18.18 ile oldukça zayıftır. Bu durum, sınıf dengesizliğinin Transformer yapısında da ciddi bir etki oluşturduğunu göstermektedir. Özellikle Farsça sınıfındaki hatalar modelin duyarlılık oranını %10'a kadar düşürmüştür, bu da bu sınıftaki örneklerin çoğunun doğru tahmin edilemediği anlamına gelir.

4.3.4. Conformer encoder tabanlı eğitilen model sonuçları

Bu bölümde, Conformer encoder tabanlı modelin dengesiz veri seti üzerindeki başarımı değerlendirilmiştir. Şekil 4.16 (a)'da sunulan doğruluk grafiğine göre, model eğitim sürecinin erken evrelerinde hızlı bir doğruluk artışı göstermiş ve kısa sürede hem eğitim hem de doğrulama verileri üzerinde yüksek başarı düzeyine ulaşmıştır. Eğrilerin genellikle birbirine yakın seyretmesi ve son epoklara doğru doğrulama başarımının sabitlenmesi, modelin genelleme yeteneğini koruyabildiğini göstermektedir. Şekil 4.16 (b)'deki kayıp grafiği incelendiğinde, doğrulama kaybında ilk epoklarda gözlenen dalgalanmaların birkaç epok sonrasında ortadan kalktığı ve hem eğitim hem de doğrulama kaybının minimum seviyeye yaklaştığı görülmektedir. Bu durum, modelin başlangıçtaki sınıf dengesizliğine bağlı kararsızlıkları zamanla dengeleyebildiğini ve aşırı öğrenme (overfitting) eğilimi göstermeden istikrarlı bir öğrenme süreci geçirdiğini ortaya koymaktadır.

Genel olarak değerlendirildiğinde, Conformer encoder modeli dengesiz sınıf dağılımına sahip veri seti üzerinde etkili bir öğrenme gerçekleştirmiş ve yüksek doğrulukla birlikte tutarlı bir kayıp eğrisi sunmuştur.



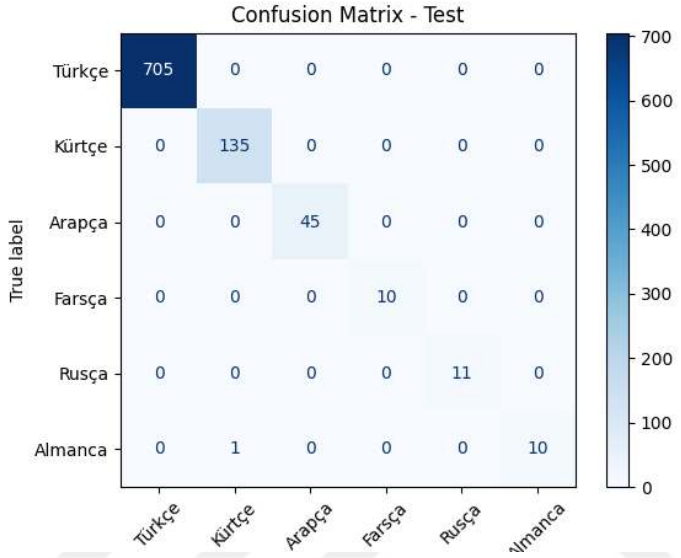
(a) (b)

Şekil 4.16. Dengesiz veri seti için conformer encoder tabanlı modelin eğitim ve doğrulama grafikleri (a) Doğruluk (Accuracy) grafiği (b) Kayıp (Loss) grafiği

Şekil 4.17'da sunulan karmaşıklık matrisi, Conformer encoder modelinin dengesiz veri seti üzerindeki sınıflandırma başarımını göstermektedir. Model, en yüksek başarıyı Türkçe

sınıfında elde etmiş; bu sınıfa ait tüm örnekler doğru sınıflandırılmıştır. Bu durum, veri setinde en fazla temsil edilen sınıfın Türkçe olması nedeniyle beklenen bir sonuçtur.

Daha az sayıda örneğe sahip olan Farsça, Rusça ve Almanca sınıflarında da genel doğruluk yüksek olmakla birlikte, bu sınıflarda sınırlı sayıda karışıklık gözlemlenmiştir. Bu sonuçlar, Conformer modelinin baskın sınıf lehine aşırı bir eğilim göstermeksizin, azınlık sınıfları da öğrenebildiğini ve genel olarak güçlü bir ayrım yeteneği sunduğunu göstermektedir.



Şekil 4.17. Dengesiz veri seti için conformer encoder tabanlı modelin karışıklık matrisi (confusion matrix - test seti)

4.3.5. Modellerin karşılaştırılması

Dengesiz veri seti ile yapılan deneylerde, dört derin öğrenme mimarisi (CNN, BiLSTM, Transformer ve Conformer) farklı performans eğilimleri göstermiştir. BiLSTM modeli, test doğruluğu %99.56'ya ulaşarak genel olarak en yüksek başarımı göstermiştir. Özellikle azınlık sınıflarda (örneğin Farsça ve Almanca) dahi oldukça yüksek F1-skor değerleri elde edilmiş, karışıklık matrisi de sınıflar arası ayrımın güçlü olduğunu ortaya koymuştur. CNN modeli ise dengeli doğruluk-eğri yapısı ve düşük kayıp değerleri ile dikkat çekmiş; özellikle eğitim ve doğrulama eğrilerinin uyumu, modelin aşırı öğrenmeden uzak, kararlı bir şekilde eğitildiğini göstermiştir. Conformer mimarisi, genel test doğruluğu açısından (%99.89) son derece başarılı bir sonuç üretmiş olsa da doğrulama eğrilerinde gözlemlenen dalgalanmalar, modelin bazı küçük sınıflarda tutarsızlıklar yaşadığını göstermektedir. Nitekim F1-skor değerlerinde de Almanca gibi az temsil edilen sınıflarda göreli bir düşüş dikkat çekmektedir. Transformer modeli ise daha karmaşık yapısı nedeniyle

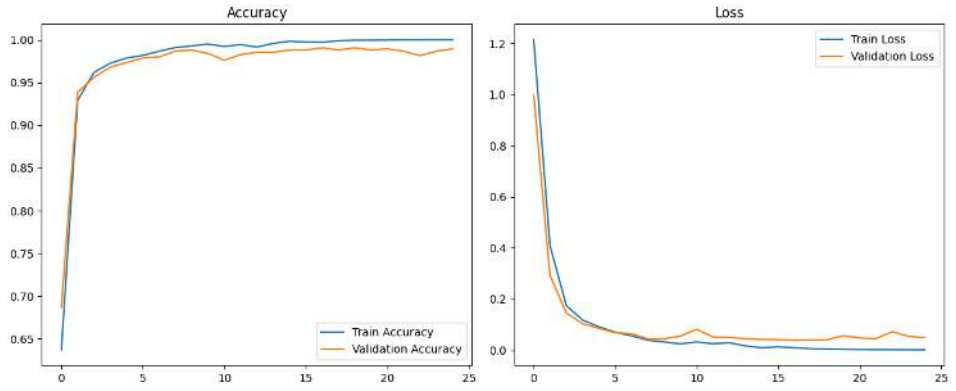
dengesiz veri dağılımına karşı en kırılgan performansı sergilemiştir. Özellikle Farsça sınıfında F1-skor değerinin 0.18 gibi oldukça düşük bir seviyeye inmesi, dikkat mekanizmasının yeterli veriye duyduğu ihtiyacı ortaya koymuştur.

4.4. Türkçe'siz Veri Seti ile Eğitilen Model Sonuçları

4.4.1. BiLSTM tabanlı eğitilen model sonuçları

Bu bölümde, veri setinde baskın sınıf etkisi ortadan kaldırılarak modelin yalnızca sınıflar arası dilsel farklara dayalı olarak öğrenim gerçekleştirmesi hedeflenmiştir. Bu yapı, sınıflar arası ayırt edici gücün daha sağlıklı biçimde test edilmesine olanak tanımaktadır.

Şekil 4.18 (a)'da sunulan doğruluk eğrisi incelendiğinde, modelin hem eğitim hem de doğrulama doğruluklarında erken epoklardan itibaren hızlı bir artış gözlemlenmektedir. Eğrilerin birbirine paralel şekilde ve oldukça yakın seyretmesi, modelin öğrenme sürecini dengeli bir biçimde sürdürdüğünü ve tüm sınıflar arasında genelleme yeteneğini koruduğunu göstermektedir. Şekil 4.18 (b)'de gösterilen kayıp grafiği ise, her iki veri kümesinde de kayıp değerlerinin erken safhada belirgin şekilde azaldığını ve daha sonraki epoklarda düşük seviyelerde sabitlendiğini ortaya koymaktadır. Bu durum, modelin eğitim verisine hızlı şekilde adapte olduğunu ve öğrenme sürecinde aşırı öğrenme eğilimi göstermeden istikrarlı bir yapı sergilediğini teyit etmektedir. Sonuç olarak, BiLSTM modelinin zaman bağımlı verilerde gösterdiği temsil gücü ve dikkatli öğrenme kapasitesi sayesinde, dengeli veri seti üzerinde yüksek doğruluk ve düşük kayıpla kararlı bir performans sunduğu söylenebilir.



(a) (b)

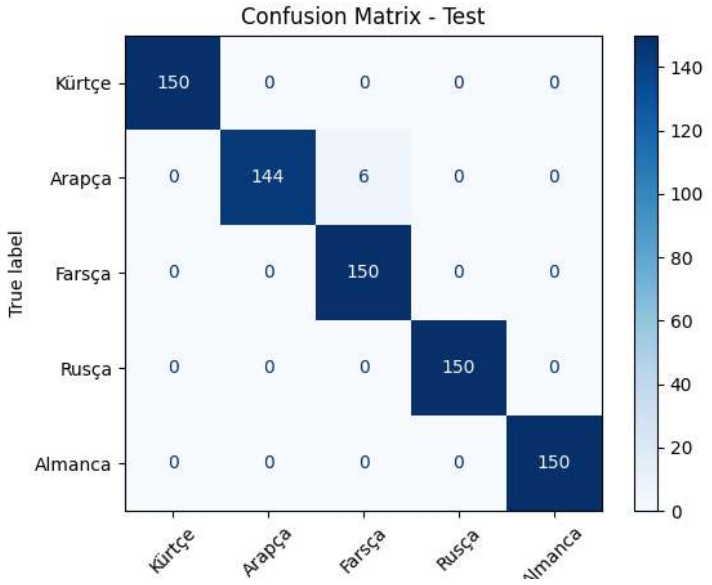
Şekil 4.18. Türkçe'siz veri seti için BiLSTM tabanlı modelin eğitim ve doğrulama grafikleri (a) Doğruluk (Accuracy) grafiği (b) Kayıp (Loss) grafiği

Şekil 4.19'de sunulan karışıklık matrisi, BiLSTM modelinin Türkçe sınıfı hariç tutularak dengelenmiş veri seti üzerinde gösterdiği yüksek sınıflandırma başarısını ortaya

koymaktadır. Tüm sınıflarda diyagonal değerlerin belirgin biçimde yüksek olması, modelin doğru sınıflandırma oranının oldukça iyi olduğunu göstermektedir.

Kürtçe, Farsça, Rusça ve Almanca sınıflarında hiçbir yanlış sınıflandırma yapılmamıştır. Arapça sınıfında yalnızca birkaç örneğin Farsça ile karıştığı görülmektedir. Bu karışıklık, muhtemelen bu iki dil arasındaki fonetik benzerliklerden kaynaklanmakta olup, genel doğruluk üzerinde önemli bir etki yaratmamaktadır.

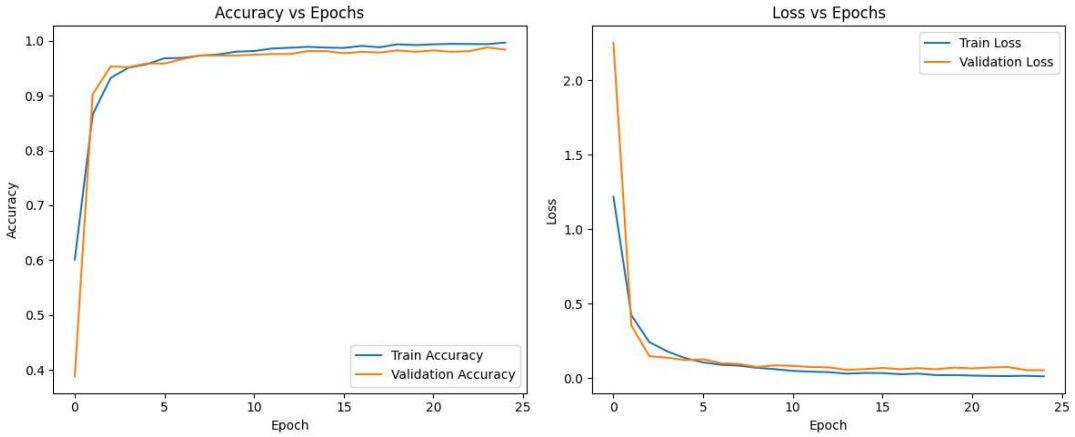
Sonuç olarak, BiLSTM modeli, Türkçe verisi olmaksızın da çok dilli sınıflandırma görevini yüksek doğruluk ve sınıflar arası ayırt edicilikle başarıyla yerine getirmiştir.



Şekil 4.19. Türkçe'siz veri seti için BiLSTM tabanlı modelin karışıklık matrisi (confusion matrix - test seti)

4.4.2. CNN tabanlı eğitilen model sonuçları

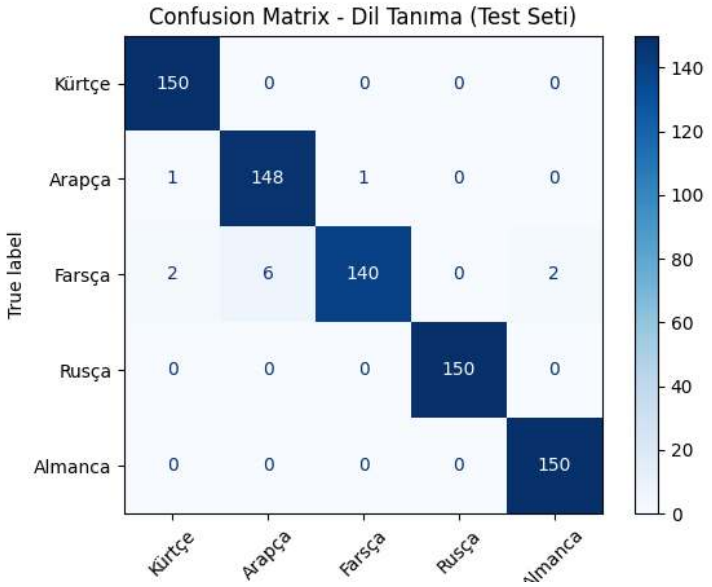
Bu çalışmada, modelin eğitim süreci boyunca hem doğruluk hem de kayıp değerleri istikrarlı bir seyir izlemiş; erken epoklardan itibaren yüksek doğruluk seviyelerine ulaşılmıştır. Şekil 4.20 (a) görüldüğü üzere, doğrulama ve eğitim doğrulukları birbirine oldukça yakın seyretmiş, bu da öğrenme sürecinin dengeli olduğunu ve modelin aşırı öğrenmeye eğilim göstermediğini ortaya koymuştur. Benzer şekilde, Şekil 4.20 (b)'de yer alan kayıp grafiğinde hem eğitim hem de doğrulama kaybı erken safhada hızlı bir şekilde azalmış ve daha sonraki epoklarda sabitlenmiştir. Bu durum, modelin eğitim verisine hızlı şekilde adapte olduğunu ve öğrenme sürecinde kararlılık sağladığını göstermektedir.



(a) (b)

Şekil 4.20. Türkçe'siz veri seti için CNN tabanlı modelin eğitim ve doğrulama grafikleri (a) Doğruluk (Accuracy) grafiği (b) Kayıp (Loss) grafiği

Şekil 4.21'de yer alan karışıklık matrisi, modelin her bir dil sınıfını oldukça başarılı bir şekilde ayırt ettiğini göstermektedir. Tüm sınıflarda diyagonal değerlerin belirgin biçimde yüksek olması, modelin doğru sınıflandırma oranının çok yüksek olduğunu ve sınıflar arası karışıklığın son derece sınırlı kaldığını ortaya koymaktadır. Özellikle Kürtçe, Rusça ve Almanca sınıflarında hiçbir yanlış sınıflandırma yapılmamış, bu da modelin bu dillere ait akustik örüntüleri net biçimde öğrenebildiğini göstermektedir. Arapça ve Farsça sınıflarında birkaç örneğin karıştığı görülmüştür ama genel başarıyı etkilemeyecek düzeydedir. Bu sonuçlar, modelin Türkçe sınıfı olmadan da çok dilli sınıflandırma görevinde yüksek ayırt ediciliğe ve genelleme kabiliyetine sahip olduğunu göstermektedir.



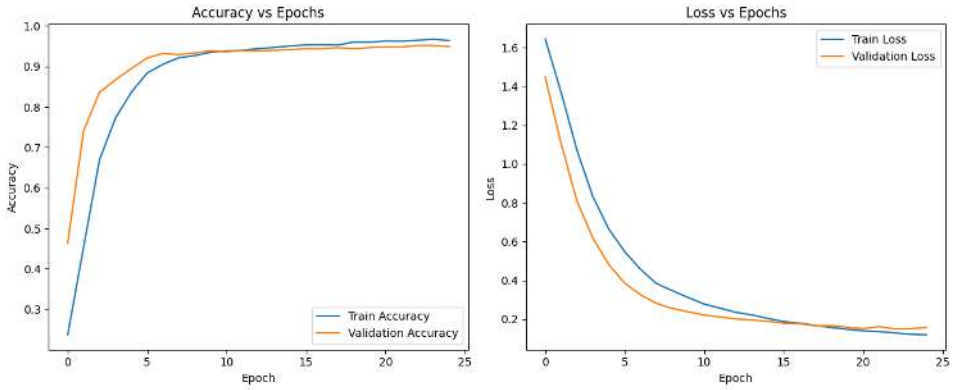
Şekil 4.21. Türkçe'siz veri seti için CNN tabanlı modelin karışıklık matrisi (confusion matrix - test seti)

Sınıflandırma raporuna göre CNN modeli, Türkçesiz veri setiyle eğitildiğinde de oldukça yüksek performans sergilemiştir. Özellikle Rusça ve Kürtçe sınıflarında sırasıyla %100 ve %99.01 F1-skoru ile model, bu dillerde neredeyse hatasız sınıflandırma yapmıştır. Arapça sınıfı %97.37 F1-skoru ile biraz daha düşük bir başarı göstermiş olsa da, genel doğruluk oranı %98.40 gibi oldukça yüksek bir seviyeye ulaşmıştır. Farsça sınıfında %96.22 F1-skoruna rağmen, duyarlılık değerinin %93.33 olması, bazı örneklerin yanlış sınıflandırıldığını göstermektedir. Bu sonuçlar, CNN mimarisinin sınıf dengesi sağlanmış senaryolarda dil tanıma görevini yüksek doğruluk ve genelleme başarısıyla gerçekleştirebildiğini göstermektedir.

4.4.3. Transformer encoder tabanlı eğitilen model sonuçları

Model, girişte pozisyonel kodlama uygulanan MFCC temsilleri ile beslenmiş ve ardından gelen çok başlı dikkat mekanizmaları içeren encoder blokları ve yoğun bağlantılı (dense) katmanlar yoluyla sınıflandırma gerçekleştirilmiştir. Eğitim süreci boyunca modelin doğruluk (accuracy) ve kayıp (loss) değerleri izlenerek genel performans değerlendirilmiştir.

Şekil 4.22 (a)'da gösterilen doğruluk grafiğinde, modelin hem eğitim hem de doğrulama veri setleri üzerinde yaklaşık 10. epoktan itibaren %95'in üzerinde doğruluk elde ettiği görülmektedir. Bu, modelin yalnızca eğitim verisini öğrenmekle kalmayıp, daha önce görmediği doğrulama verisi üzerinde de etkili genelleme yapabildiğini göstermektedir. Şekil 4.22 (b)'de yer alan kayıp eğrisi ise hem eğitim hem de doğrulama için istikrarlı bir azalma göstermekte olup, modelin aşırı öğrenme eğilimi göstermediğini doğrulamaktadır.

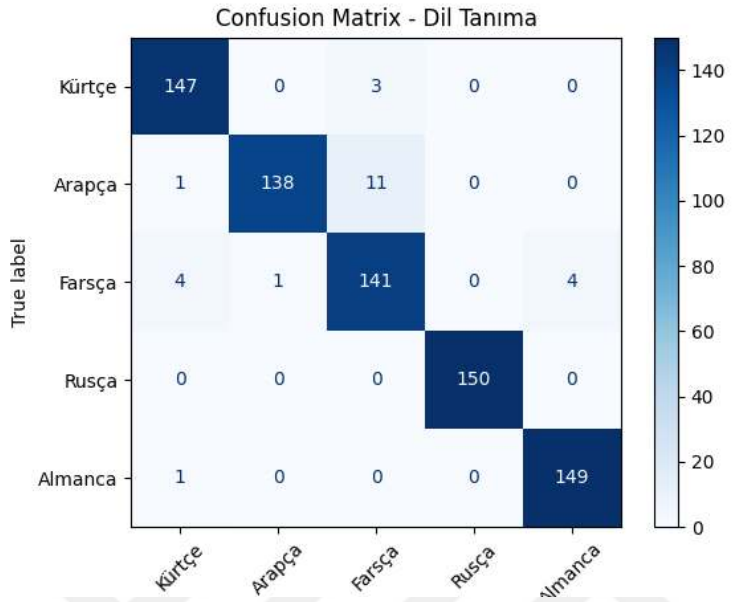


(a) (b)

Şekil 4.22. Türkçe'siz veri seti için transformer encoder tabanlı modelin eğitim ve doğrulama grafikleri (a) Doğruluk (Accuracy) grafiği (b) Kayıp (Loss) grafiği

Şekil 4.23'te yer alan karışıklık matrisi, modelin özellikle Rusça ve Almanca sınıflarında yüksek doğrulukla tahmin yaptığını göstermektedir. Arapça ve Farsça

sınıflarında ise sınırlı sayıda karışıklık gözlenmiş, bazı Farsça örnekler diğer dillere yanlış atanmıştır. Bu durum, sınıflar arası ayrımın bazı dil çiftlerinde daha zor olduğunu ortaya koymaktadır.



Şekil 4.23. Türkçe'siz veri seti için transformer encoder tabanlı modelin karışıklık matrisi (confusion matrix - test seti)

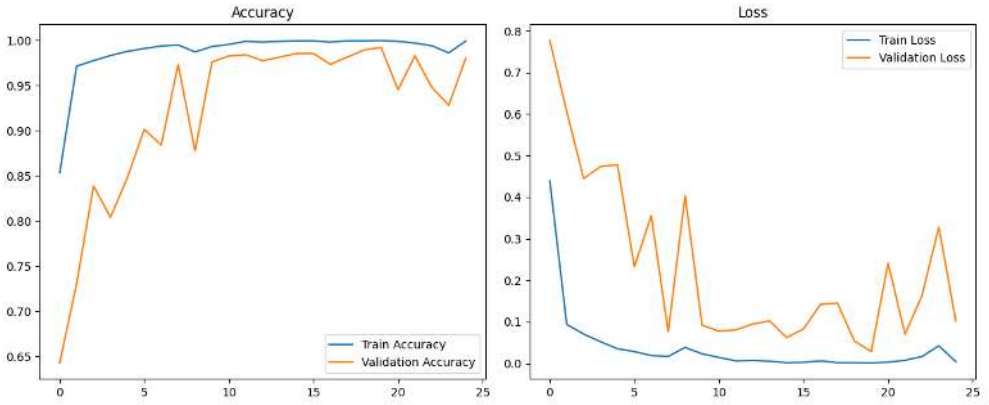
Sınıflandırma raporuna göre Transformer Encoder tabanlı model, Türkçesiz veri setiyle eğitildiğinde de yüksek düzeyde başarı göstermiştir. Özellikle Rusça ve Almanca sınıflarında sırasıyla %100 ve %98.35 F1-skoru ile model, bu dillerde oldukça isabetli tahminler yapmıştır. Kürtçe sınıfı da %97.03 F1-skoru ile güçlü bir performans sergilemiştir. Arapça sınıfında %95.50, Farsça sınıfında ise %92.46 F1-skoru elde edilmiş olup, bu iki dilde diğerlerine göre daha fazla hata gözlenmiştir. Buna rağmen genel doğruluk oranı %96.67 seviyesine ulaşmış, bu da Transformer mimarisinin dengeli veri setiyle başarılı bir şekilde çalıştığını göstermiştir.

4.4.4. Conformer encoder tabanlı eğitilen model sonuçları

Şekil 4.24 (a)'da görüldüğü üzere, eğitim doğruluğu erken epoklarda hızla artmış ve üst seviyelere ulaşmıştır. Doğrulama doğruluğu ise genel olarak yüksek seyretmekle birlikte zaman zaman dalgalanma göstermiştir. Bu durum, modelin öğrenme kapasitesinin güçlü olduğunu ancak doğrulama verisi üzerinde dönemsel kararsızlıklar yaşadığını göstermektedir. Şekil 4.24 (b)'deki kayıp grafiğinde, eğitim kaybı istikrarlı şekilde azalırken doğrulama kaybında dalgalanmalar gözlenmiştir. Bu dalgalı yapı, modelin bazı epoklarda

aşırı öğrenme eğilimi gösterebildiğine işaret etmektedir.

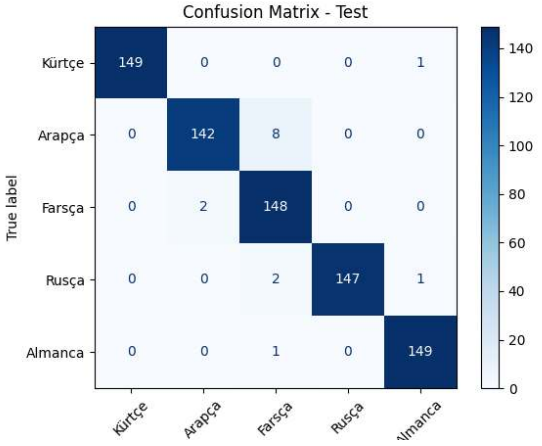
Genel olarak model başarılı sonuçlar üretmiş olsa da, doğrulama tarafındaki düzensizlikler genelleme kapasitesinin sınırlı olabileceğini düşündürmektedir.



(a) (b)

Şekil 4.24. Türkçe'siz veri seti için conformer encoder tabanlı modelin eğitim ve doğrulama grafikleri (a) Doğruluk (Accuracy) grafiği (b) Kayıp (Loss) grafiği

Şekil 4.25'de sunulan karmaşıklık matrisi, Conformer encoder modelinin Türkçe sınıfı hariç tutularak oluşturulan dengeli veri seti üzerindeki sınıflandırma başarımını yansıtmaktadır. Model, genel olarak yüksek doğrulukla sınıflandırma yapmış; tüm dillerde diyagonal değerler güçlü bir şekilde öne çıkmıştır. Ancak Arapça sınıfında Farsça ile, Farsça ve Rusça sınıflarında ise birkaç örnekte karşılıklı karışıklık gözlemlenmiştir. Bunun dışında Kürtçe ve Almanca sınıflarında yalnızca birer örnekte hata yapılmıştır. Bu karışıklıklar çok sınırlı düzeyde olup, modelin genel başarımını olumsuz etkilememektedir. Sonuç olarak, Conformer modeli, Türkçe verisi olmadan da çok dilli sınıflar arasında etkili ayrım yapabilmiş ve sınıflar arası öğrenme başarısını büyük ölçüde korumuştur.



Şekil 4.25. Türkçe'siz veri seti için conformer encoder tabanlı modelin karışıklık matrisi (confusion matrix - test seti)

4.4.5. Modellerin karşılaştırılması

Türkçesiz ve dengeli veri seti üzerinde gerçekleştirilen sınıflandırma deneylerinde, dört farklı modelin (BiLSTM, CNN, Transformer Encoder ve Conformer Encoder) performansları karşılaştırılmıştır. Doğruluk bazında en başarılı sonuç BiLSTM modeliyle elde edilmiştir; bu model %99.2 test doğruluğu sağlayarak 750 örnekten 744'ünü doğru sınıflandırmıştır. CNN modeli ise %98.4 doğruluk ile ikinci sırada yer almış ve özellikle Arapça, Kürtçe ve Farsça sınıflarında güçlü bir sınıflandırma performansı göstermiştir. Transformer Encoder modeli %96.67 doğrulukla daha düşük bir başarı sağlamış; Arapça, Farsça ve Kürtçe arasında sınırlı fakat belirgin karışıklıklar göstermiştir. Conformer modeli ise %98 doğruluk değeriyle genel başarı açısından Transformer'ın önüne geçmiştir, ancak eğitim sürecinde doğrulama eğrilerinde dalgalanmalar dikkat çekmiştir. Karışıklık matrisleri ve F1-skor değerleri incelendiğinde, BiLSTM modeli yalnızca Arapça - Farsça sınıfında hata yapmış, diğer sınıflarda neredeyse mükemmel sonuçlar üretmiştir. Eğitim ve doğrulama eğrileri incelendiğinde, BiLSTM ve CNN modellerinin hem doğruluk hem kayıp açısından stabil ve hızlı bir öğrenme eğrisi sunduğu; Conformer modelinin ise yüksek doğruluk üretmesine rağmen doğrulama kayıplarında dalgalanmalar gösterdiği gözlenmiştir. Transformer modelinde ise eğitim süreci daha dengeli ilerlemiş olsa da, düşük örnekli sınıflarda başarı oranı görece düşük kalmıştır.

Sonuç olarak, BiLSTM modeli bu veri seti özelinde en yüksek doğruluk ve sınıf bazlı kararlılığı sağlayarak öne çıkmıştır. CNN modeli ise sınıf dengesine duyarlı yapısıyla genel başarıda BiLSTM'in hemen ardından gelmiştir. Transformer ve Conformer modelleri daha karmaşık mimarilere sahip olmalarına rağmen, düşük örnekli sınıflarda daha fazla hata üretmiş; bu da daha sade mimarilerin az veriyle daha etkili çalışabildiğini göstermiştir.

5. SONUÇLARIN İNCELENMESİ

Bu çalışmada, SLID (Spoken Language Identification – Konuşulan Dili Tanıma) problemini çözmek amacıyla BiLSTM, CNN, Transformer Encoder ve Conformer Encoder tabanlı dört farklı derin öğrenme mimarisi önerilmiş ve karşılaştırmalı olarak değerlendirilmiştir. Modelin temel amacı, Türkiye'de yaygın olarak konuşulan Türkçe, Kürtçe, Arapça, Farsça, Rusça ve Almanca dillerini otomatik olarak sınıflandırmaktır. Literatürde, çok dilli sesli veri sınıflandırmasına yönelik güçlü ve genellenebilir modellerin eksikliği göz önüne alındığında, bu çalışma hem dengeli hem de gerçek dünya senaryolarını yansıtan dengesiz veri kümeleri üzerinden kapsamlı bir analiz sunmaktadır.

Çalışmanın ilk aşamasında, çeşitli açık kaynaklardan derlenen ses verileri işlenerek üç farklı veri seti oluşturulmuştur: Türkçe dahil dengeli veri seti, Türkçenin baskın olduğu dengesiz veri seti ve Türkçe verisi hariç tutularak oluşturulan dengeli veri seti. Türkçenin hariç bırakıldığı bu son veri seti, özellikle Göç İdaresi Genel Müdürlüğü, turizm danışma merkezleri ve sınır kapıları gibi Türkçe dışındaki dillerin daha baskın biçimde kullanıldığı kurumsal veya geçici etkileşim ortamlarını simüle etmeyi amaçlamaktadır.

Tüm ses verileri MFCC öznitelik çıkarımı ve gürültü azaltma süreçlerinden geçirilmiş, eğitim doğrulama test kümelerine ayrılmıştır. Böylece, her bir modelin hem teknik başarımı hem de farklı veri dağılımı senaryolarına verdiği tepkiler değerlendirilmiştir.

Dengeli veri setiyle yapılan analizlerde, en yüksek doğruluk oranı %99.00 ile Conformer modeline ait olup, modelin genel doğruluk, macro ve weighted ortalamalarda tüm metriklerde 0.99 seviyesinde performans sergilediği gözlemlenmiştir. BiLSTM modeli %98.78 doğruluk ile benzer şekilde yüksek performans göstermiş; özellikle Türkçe, Kürtçe ve Rusça sınıflarında f1-score değerleri 1.00 seviyesine ulaşmıştır. CNN modeli %98.67 doğruluk sağlamış ve tüm sınıflarda dengeli bir başarı sunmuştur; Arapça dışındaki tüm dillerde f1-score değerleri 0.97 ve üzerindedir. Buna karşın, Transformer modeli %97.33 doğruluk oranıyla diğer modellerin gerisinde kalmıştır. Özellikle Arapça ve Farsça sınıflarında f1-score değerleri 0.95'in altına düşmüş, bu da modelin bu dillerde ayrıştırma performansında zayıfladığını göstermektedir. Genel olarak, Conformer mimarisi hem doğruluk hem de sınıflar arası tutarlılık açısından en yüksek başarıyı sağlamıştır.

Dengesiz veri seti ile yapılan sınıflandırma analizlerinde, en yüksek doğruluk oranı %99.89 ile Conformer modeline aittir. Conformer, özellikle baskın sınıf olan Türkçe'de tüm metriklerde mükemmele yakın sonuçlar üretmiş; daha az örneğe sahip olan Almanca

sınıfında ise yalnızca bir hata yapmıştır. BiLSTM ve CNN modelleri, eşit şekilde %99.56 doğruluk sağlamış ve küçük sınıflarda (Farsça, Almanca) makul düzeyde başarı elde etmiştir. Buna karşın, Transformer modeli %98.80 doğruluk ile diğer modellerin gerisinde kalmış; özellikle Farsça sınıfında f1-score değeri yalnızca 0.18 olup ciddi bir sınıflandırma zayıflığı ortaya koymuştur. Modelin bu sınıfta recall değeri %10 seviyesinde kalmış ve yalnızca 1 örneği doğru tahmin edebilmiştir. Sonuçlar, özellikle dengesiz veri ile çalışılan senaryolarda, Conformer, BiLSTM ve CNN gibi modellerin küçük sınıflarda dahi yüksek tutarlılıkla çalışabildiğini, Transformer yapısının ise belirli sınıflarda kararsızlık yaşayabildiğini göstermektedir.

Türkçe sınıfının çıkarıldığı dengeli veri seti analizlerinde, BiLSTM modeli %99.2 doğruluk oranıyla en yüksek performansı göstermiş ve tüm dillerde f1-score değerlerini 0.98–1.00 aralığında koruyarak güçlü bir genelleme yeteneği sergilemiştir. CNN modeli %98.40 doğrulukla istikrarlı bir yapı sunarken, Arapça ve Farsça sınıflarında küçük karışıklıklar gözlemlenmiştir. Conformer mimarisi %98.00 doğruluk üretmiş; özellikle Arapça ve Farsça arasında sınırlı çapraz hata kaydedilmiştir. Transformer modeli ise %96.67 doğruluk ile diğer modellere kıyasla daha düşük bir başarı göstermiş ve Farsça sınıfında f1-score değerlerinde düşüş yaşanmıştır. Bu senaryo, modellerin azınlık diller üzerinde ayırt edici performansını test etmesi açısından kritik önem taşımaktadır.

Performans metrikleri genel olarak tüm modellerin SLID görevinde yüksek doğruluk ve sınıflandırma başarısı sergilediğini göstermektedir. Özellikle CNN ve BiLSTM modelleri; hem dengeli hem dengesiz senaryolarda, hem de Türkçe sınıfı hariç tutulduğunda dahi yüksek f1-score**,** hızlı öğrenme**,** düşük kayıp ve kararlı doğruluk gibi açılardan öne çıkmıştır. Conformer Encoder modeli ise dengeli ve dengesiz veri setlerinde en yüksek doğruluk oranlarına ulaşarak yüksek genelleme başarısı göstermiştir. Bununla birlikte, doğrulama sürecinde bazı sınıflarda dönemsel dalgalanmalar gözlemlenmiş, ancak bu durum genel test doğruluğuna olumsuz yansımamıştır. Transformer modeli ise özellikle dengesiz veri setinde küçük sınıflarda belirgin performans düşüşü göstermiştir.

Bu çalışmanın en önemli katkılarından biri, sadece genel doğruluk değil, aynı zamanda sınıf bazlı başarı**,** karışıklık matrisi ve f1-score gibi detaylı metrikler üzerinden model başarımının analiz edilmiş olmasıdır. Ayrıca, sınıflar arası karışıklığın hangi dil çiftlerinde yoğunlaştığı görselleştirilmiş ve akustik benzerlik kaynaklı sınırlı hataların nedenleri tartışılmıştır.

Sonuç olarak, bu çalışma MFCC özniteliklerine dayalı CNN, BiLSTM, Transformer ve Conformer mimarilerini çok dilli ses sınıflandırmasında karşılaştırmalı biçimde ele alarak

alana katkı sağlamaktadır. Farklı veri dağılım senaryolarında test edilen bu yaklaşımlar, gerçek dünya uygulamalarına doğrudan uyarlanabilir nitelikte yüksek doğrulukta sonuçlar üretmiştir. Elde edilen bulgular, çağrı merkezi otomasyonu, sınır güvenliği, dil temelli eğitim sistemleri ve çok dilli insan–makine etkileşimi gibi pek çok alanda konuşulan dili tanıma sistemlerinin geliştirilmesine katkı sunabilecek niteliktedir. Ayrıca, sınıf dengesizliği veya baskın sınıf etkisinin öğrenme sürecine olan etkileri, veri mühendisliği süreçlerinin önemini bir kez daha ortaya koymuştur.

KAYNAKLAR

[1] O. H. Anidjar and R. Yozevitch, “Enhancing Neural Spoken Language Recognition: An Exploration with Multilingual Datasets,” Jan. 2025, [Online]. Available: arxiv.org/abs/2501.11065

[2] N. I. Al-Shathry *et al.*, “MULTI-CLASS SPOKEN LANGUAGE DETECTION USING ARTIFICIAL INTELLIGENCE with FRACTAL AL-BIRUNI EARTH RADIUS OPTIMIZATION ALGORITHM,” *Fractals*, 2024, doi: 10.1142/S0218348X25400547.

[3] Z. Li, Y. Xu, D. Ke, and K. Su, “PLDE: A lightweight pooling layer for spoken language recognition,” *Speech Commun*, vol. 158, Mar. 2024, doi: 10.1016/j.specom.2024.103055.

[4] G. Singh, S. Sharma, V. Kumar, M. Kaur, M. Baz, and M. Masud, “Spoken Language Identification Using Deep Learning,” *Comput Intell Neurosci*, vol. 2021, 2021, doi: 10.1155/2021/5123671.

[5] S. Mukherjee, N. Shivam, A. Gangwal, L. Khaitan, and A. J. Das, “Spoken language recognition using CNN,” in *Proceedings - 2019 International Conference on Information Technology, ICIT 2019*, Institute of Electrical and Electronics Engineers Inc., Dec. 2019, pp. 37–41. doi: 10.1109/ICIT48102.2019.00013.

[6] S. Garai and S. Samui, “Optimizing Performance of Spoken Language Identification Systems for Indian Languages Using Ensemble Deep Learning Models,” in *2024 IEEE Calcutta Conference, CALCON 2024 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/CALCON63337.2024.10914168.

[7] D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey, and S. Khudanpur, “Spoken Language Recognition using X-vectors,” in *Speaker and Language Recognition Workshop, ODYSSEY 2018*, The International Society for Computers and Their Applications (ISCA), 2018, pp. 105–111. doi: 10.21437/Odyssey.2018-15.

[8] L. Ferrer, Y. Lei, M. McLaren, and N. Scheffer, “Study of senone-based deep neural network approaches for spoken language recognition,” *IEEE/ACM Trans Audio*

Speech Lang Process, vol. 24, no. 1, pp. 105–116, Jan. 2016, doi: 10.1109/TASLP.2015.2496226.

[9] R. W. M. Ng, M. Nicolao, and T. Hain, "Unsupervised crosslingual adaptation of tokenisers for spoken language recognition," *Comput Speech Lang*, vol. 46, pp. 327–342, Nov. 2017, doi: 10.1016/j.csl.2017.05.002.

[10] J. Oruh, S. Viriri, and A. Adegun, "Long Short-Term Memory Recurrent Neural Network for Automatic Speech Recognition," *IEEE Access*, vol. 10, pp. 30069–30079, 2022, doi: 10.1109/ACCESS.2022.3159339.

[11] Y. K. Muthusamy, E. Barnard, and R. A. Cole, "Reviewing Automatic language Identification."

[12] G. Montavon, "Deep learning for spoken language identification." unpublished, Berlin Institute of Technology, 2010.

[13] P. Kumar, A. Biswas, A. N. Mishra, and M. Chandra, "Spoken Language Identification Using Hybrid Feature Extraction Methods," 2010.

[14] T.C. İçişleri Bakanlığı Göç İdaresi Başkanlığı. Accessed: May 18, 2025. [Online]. Available: goc.gov.tr

[15] TÜİK - Veri Portalı. Accessed: May 18, 2025. [Online]. Available: data.tuik.gov.tr/Kategori/GetKategori?p=Nufus-ve-Demografi-109

[16] L. Burget, P. Matějka, and J. Černocký, "Discriminative training techniques for acoustic language identification," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Toulouse, France, May 2006, pp. I-209–I-212.

[17] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," in *Computer Speech and Language*, Apr. 2006, pp. 210–229. doi: 10.1016/j.csl.2005.06.003.

[18] *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) : ASRU 2019 : proceedings : December 15-18, 2019, Guadeloupe, West Indies*. IEEE,

2019.

[19] H. S. Das and P. Roy, “A CNN-BiLSTM based hybrid model for Indian language identification,” *Applied Acoustics*, vol. 182, Nov. 2021, doi: 10.1016/j.apacoust.2021.108274.

[20] S. Indolia, A. K. Goswami, S. P. Mishra, and P. Asopa, “Conceptual Understanding of Convolutional Neural Network- A Deep Learning Approach,” in *Procedia Computer Science*, Elsevier B.V., 2018, pp. 679–688. doi: 10.1016/j.procs.2018.05.069.

[21] “Convolutional Neural Networks (CNN) — Architecture Explained | by Dharmaraj | Medium.” Accessed: May 14, 2025. [Online]. Available: medium.com/@draj0718 /convolutional-neural-networks-cnn-architectures-explained-716fb197b243

[22] “Introduction to convolution neural network | GeeksforGeeks.” Accessed: May 14, 2025. [Online]. Available: geeksforgeeks.org/introduction-convolution-neural-network

[23] A. Vaswani *et al.*, “Attention Is All You Need.” in Proc. 31st Conf. Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, Dec. 2017, pp. 5998–6008.

[24] J. Chorowski and D. Bahdanau, “Attention-Based Models for Speech Recognition.” arXiv preprint arXiv:1506.07503, Jun. 2015.

[25] C.-F. Yeh *et al.*, “TRANSFORMER-TRANSDUCER: END-TO-END SPEECH RECOGNITION WITH SELF-ATTENTION”.

[26] A. Mohamed, D. Okhonko, and L. Zettlemoyer, “Transformers with convolutional context for ASR,” Apr. 2019, [Online]. Available: arxiv.org/abs/1904.11660

[27] T. M. Bartley, F. Jia, K. C. Puvvada, S. Kriman, and B. Ginsburg, “Accidental Learners: Spoken Language Identification in Multilingual Self-Supervised Models,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICASSP49357.2023.10096407.

[28] F. Wang, L. Huang, T. Li, Q. Hong, and L. Li, “Conformer-based Language Embedding with Self-Knowledge Distillation for Spoken Language Identification,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, International Speech Communication Association, 2023, pp. 5286–5290. doi: 10.21437/Interspeech.2023-1557.

[29] “Kurdish_splited_audios.” Accessed: May 19, 2025. [Online]. Available: kaggle.com/datasets/saadali5/kurdish-splited-audios?resource=download

[30] “Arabic Speech Corpus.” Accessed: May 19, 2025. [Online]. Available: kaggle.com/datasets/haithemhermessi/arabic-speech-corpus

[31] “Russian Speech Disorder Audio (RSDA).” Accessed: May 19, 2025. [Online]. Available: kaggle.com/datasets/mhantor/russian-voice-datasetselect=Normal+Voice

[32] “German Single Speaker Speech Dataset.” Accessed: May 19, 2025. [Online]. Available: kaggle.com/datasets/bryanpark/german-single-speaker-speech-dataset

[33] “persian-ctc-asr/dataset at main · meytiii/persian-ctc-asr · GitHub.” Accessed: May 19, 2025. [Online]. Available: github.com/meytiii/persian-ctc-asr/tree/main/dataset

[34] A. Draghici, J. Abeßer, and H. Lukashevich, “A study on spoken language identification using deep neural networks,” in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Sep. 2020, pp. 253–256. doi: 10.1145/3411109.3411123.

[35] H. S. Das and P. Roy, “A CNN-BiLSTM based hybrid model for Indian language identification,” *Applied Acoustics*, vol. 182, Nov. 2021, doi: 10.1016/j.apacoust.2021.108274.

[36] M. Radfar, A. Mouchtaris, and S. Kunzmann, “End-to-end neural transformer based spoken language understanding,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, International Speech Communication Association, 2020, pp. 866–870. doi: 10.21437/Interspeech.2020-1963.

[37] P. Wang et al., “LAMASSU: Streaming Language-Agnostic Multilingual Speech Recognition and Translation Using Neural Transducers,” Nov. 2022, [Online]. Available: arxiv.org/abs/2211.02809

[38] “What Is a Learning Curve in Machine Learning? | Baeldung on Computer Science.” Accessed: May 19, 2025. [Online]. Available: baeldung.com/cs/learning-curve-ml

[39] *ICASSP : 2018 IEEE International Conference on Acoustics, Speech and Signal Processing : proceedings : April 15-20, 2018, Calgary Telus Convention Center, Calgary, Alberta, Canada*. Institute of Electrical and Electronics Engineers, 2018.

[40] Z . Xu, B. Zhang, Y. Bai, D. Li, and G. Fan, “Learning to Coordinate via Multiple Graph Neural Networks,” Apr. 2021, [Online]. Available: arxiv.org/abs/2104.03503