BLM508 Yapay Zekâ ve Bilişsel Sistemler

Proje Raporu

Adı Soyadı: Akif YAVUZSOY

Numarası: 235113001

Projenin Adı: TIMIT Veri Tabanından LSTM/RNN Kullanarak Cinsiyet Tanıma İçin Bilişsel Bir Sistem

1. Giriş:

Cinsiyet tanıma, konuşma işleme sistemlerinde önemli bir rol oynamaktadır. Bu proje, TIMIT veri tabanındaki ses verilerinden cinsiyet tanımlama için bir bilişsel sistem geliştirmeyi amaçlamaktadır. Ses sinyallerinden cinsiyet tanımlama, konuşma işleme sistemlerinde yaygın olarak kullanılan bir uygulamadır. Bu çalışmada, LSTM (Long Short-Term Memory) ve RNN (Recurrent Neural Network) gibi derin öğrenme yöntemlerini kullanarak bu görevi gerçekleştirmeyi hedeflemektedir.

İnsan seslerinin sanatsal ve akustik özellikleri, konuşma verilerinden çıkarılabilir ve konuşma işleme sistemleri için özellikler olarak kullanılabilir. Güvenlik sistemleri gibi alanlarda konuşmacının cinsiyetini belirlemenin önemli olduğu durumlarda kullanılır. Ayrıca, film analizi gibi uygulamalarda da önemlidir.

Çalışmanın başarılı olabilmesi için kullanılacak olan veri setinin özellik çıkarımı oldukça önemlidir ve bu çalışmada kullanılacak olan özellik çıkarıma yöntemlerinden daha ayrıntılı bahsedilecektir.

2. Materyal ve Metot:

Konuşma tanıma sistemleri oluştururken ses verilerine ve o ses verilerinin etiketlerine ihtiyaç duymaktayız. Öznitelik çıkarımı ile sesten ürettiğimiz matrisleri Sinir Ağları ile oluşturduğumuz modele vererek cinsiyet karşılığını öğrenen bir sistem oluşturmaktayız. Bu sistemi test verileri ile test ederek başarım oranlarını hesaplıyoruz.

2.1. Veri Toplama ve Hazırlık:

Proje için kullanılan veri seti TIMIT veri tabanından elde edilmiştir. Bu veri seti, farklı konuşmacılardan alınan ses kayıtlarını içerir ve her bir kayıt, konuşmacının cinsiyetine ilişkin etiket bilgisiyle birlikte gelir. Veri seti, önceden tanımlanmış özniteliklerle birlikte kullanılmak üzere ses sinyallerine dönüştürülür.

TIMIT (Texas Instruments/Massachusetts Institute of Technology) veri seti, konuşma işleme ve konuşma tanıma alanında sıkça kullanılan bir veri kaynağıdır. Bu veri seti, ABD'nin farklı bölgelerinden toplanmış 630 konuşmacının İngilizce ses kayıtlarını içerir. TIMIT, özellikle konuşmanın farklı aksanlarını ve diyalektlerini içeren çeşitli konuşmacı profillerini kapsayan geniş bir veri yelpazesine sahiptir. Her ses dosyası (.wav), 16kHz örnekleme oranıyla kaydedilmiştir. Şekil 1’de TIMIT veri setinin eğitimi ve test oranları yer alırken, Şekil 2’de ise cinsiyet oranları yer almaktadır.

TIMIT veri seti, iki ana bileşenden oluşur:

Fonem Etiketleri: Her ses kaydı, kayıttaki konuşmanın doğru bir transkripsiyonunu içeren fonem etiketleriyle eşleştirilmiştir. Bu fonem etiketleri, konuşmanın hangi seslerden oluştuğunu belirtir ve model eğitimi ve değerlendirmesi için önemli bir rehber sağlar.

Dalga Biçimi Dosyaları: TIMIT veri setindeki her konuşmacının, farklı cümlelerde veya ses koşullarında yaptığı konuşmayı içeren ses dalga biçimi dosyaları bulunur. Bu ses dosyaları, genellikle bir mikrofon kullanılarak kaydedilmiş ve çeşitli frekans bileşenlerini içeren gerçek dünya konuşma verilerini temsil eder.

Bölge (Aksan)	Test 30%	Train 70%	Total
New England	11	38	49
Northern	26	76	102
North Midland	26	76	102
South Midland	32	68	100
Southern	28	70	98
New York City	11	35	46
Western	23	77	100
Army Brat	11	22	33
Total	**168**	**462**	**630**

Şekil 1: *Veri seti eğitim oran tablosu*

Bölge (Aksan)	Male 70%	Female 30%	Total
New England	31	18	49
Northern	71	31	102
North Midland	79	23	102
South Midland	69	31	100
Southern	62	36	98
New York City	30	16	46
Western	74	26	100
Army Brat	22	11	33
Total	**438**	**192**	**630**

Şekil 2: *Veri seti cinsiyet oran tablosu*

2.2. Literatür Araştırmaları:

Gender and Age Estimation Methods Based on Speech Using Deep Neural Networks, Baseline X-Vector, QuartzNet Tabanlı X-Vector, D-Vector (LSTM RNN) yöntemleri kullanılmıştır. Konuşma sinyallerinden yaş tahmini ve cinsiyet sınıflandırması bağlamında farklı sinir ağı mimarileri, transfer öğrenme düzenleri ve çoklu görev öğrenme kullanımı araştırılırmıştır. Kullanılan yöntemlerden en iyi sonuç veren yöntem D-Vector (LSTM RNN) olduğu sonucuna varılmıştır.

Derin Öğrenme Algoritmalarını Kullanarak Görüntüden Cinsiyet Tahmini, AlexNet, VGG-16 yöntemleri kullanılmıştır. Bu çalışmada görüntülerden cinsiyet tahmini yapmak için derin öğrenme algoritmaları kullanılmıştır. 3170 eğitim ve 318 test verileri kullanılmıştır. AlexNet, VGG-16 ve yeni bir model karşılaştırılmış ve VGG-16 modelinin daha başarılı olduğu sonucuna varılmıştır. Veri çeşitliliği ve sayısını artırılması önerilmiştir.

Ses Olayı Tanıma ve Akustik Sahne Geri Getirimi, MLP, RNN, LSTM, CNN modelleri kullanılmıştır. Bu çalışmada çevresel seslerden oluşmuş ses klipleri içerisindeki ses olaylarının tanımlanması sağlanmıştır. Sınıflandırma eğitimi için MLP, RNN, LSTM, CNN yöntemleri kıyaslanmıştır. Leaky ReLu aktivasyon fonksiyonu tercih edilmiştir. AlexNetish ve VGGish mimarisi üzerine GRU ve LSTM algoritmalarının eklenmesiyle başarılı sonuç elde edilmiştir.

A Dual-Staged heterogeneous stacked ensemble model for gender recognition using speech signal, Gaussian Mixture Model (GMM) super-vector-based SVM, K-Nearest Neighbours (K-NN) ve Support Vector Machine (SVM), LSTM ve GRU, Random Forest Recursive Feature Elimination yöntemleri kullanılmıştır. Bu çalışma, insan sesinin cinsiyetini tanımlamak için çeşitli öznitelik çıkarım yöntemlerinin ve makine öğrenimi algoritmalarının kullanıldığı bir araştırmadır. Öncelikle, ses verilerinden özniteliklerin çıkarılması ve daha sonra bu özniteliklerin makine öğrenimi veya derin öğrenme algoritmalarıyla eğitilerek cinsiyetin sınıflandırılması işlemi ele alınmıştır. Bu çalışmanın sonuçlarına göre, önerilen DH-SEM (dual-staged heterogeneous stacked ensemble) yöntemi, cinsiyet tanımlama probleminde daha başarılı bulunmuştur.

A new pitch-range based feature set for a speaker’s age and gender classification, ön işleme, özellik çıkarma (MFCC, RASTA_PLP, F0 ve pitch-range (PR)), sınıflandırma (kNN, SVM) yöntemleri kullanılmıştır. Bu çalışma, konuşmacının sesinden yaş ve cinsiyet bilgisi çıkarma problemini ele almaktadır. Veri kümesi, konuşmacı yaşını ve cinsiyetini algılamayı desteklemek amacıyla Interspeech 2010 Paralinguistic Challenge organizasyonu tarafından sağlanmıştır. Korpus, 795 konuşmacıdan gelen 49 saatlik telefon konuşması içermektedir ve eğitim (23 saat, 471 konuşmacı), geliştirme (14 saat, 299 konuşmacı) ve test setleri (12 saat, 175 konuşmacı) olarak bölünmüştür. Yöntemin üç ana adımı vardır. Önişleme adımı, sessizlikten konuşma sesini ayırmak için VAD içerir. Bu çalışmada, VAD için enerji ve sıfır geçiş oranları kullanılmıştır. Özellik çıkarma adımı, sınıflandırmada kullanılan özellik setlerini hesaplar. MFCC, RASTA_PLP, F0 gibi bilinen özellik setlerinin yanı sıra, önerilen pitch-range (PR) özellik seti hesaplanmıştır. Sınıflandırma adımı, kNN ve SVM kullanır.

Voice gender recognition under unconstrained environments using self-attention, özellik çıkarımı (MFCC), modelleme (LSTM), sınıflandırma (kNN, SVM) yöntemleri kullanılmıştır. Bu çalışma, ses cinsiyeti tanıma (Voice Gender Recognition - VGR) üzerine odaklanmaktadır. VGR, ses kayıtlarından erkek veya kadın olarak insan sesini tanımlamayı amaçlamaktadır. İlk yaklaşım, ses verisinin sayısal özellikleri ve karakteristiklerini kullanmaktır. Örneğin, ortalama frekans, mod, standart sapma gibi özellikler üzerine yoğunlaşır. İkinci yaklaşım ise sesin spektral özelliklerini kullanır. Bu spektral özellikler arasında MFCC'ler, Log-Mel özellikleri gibi yöntemler bulunur. Relief algoritmasını özellik seçimi yöntemi olarak kullanmışlar ve ardından sınıflandırma için çift katmanlı derin Long Short-Term Memory (LSTM) kullanmışlardır. Sınıflandırma için ise k-Nearest Neighbors (kNN) ve Support Vector Machine (SVM) kullanmışlardır. MFCC'nin TIMIT veri tabanında en yüksek doğruluğu elde ettiğini belirtmişlerdir. Veri seti seçimi, daha önce bahsedilen zorlukları aşabilen bir sonuç elde etmek için önemlidir. Bu çalışmada, daha zorlu ve kısıtlı olmayan ortamlarda daha iyi çalışacak bir ses cinsiyeti tanıma sistemi için VoxCeleb veri seti kullanılmıştır. İlk model, saf self-attention katmanlarından oluşurken, ikinci model daha karmaşık bir yapıya sahiptir. İlk modelin aşırı uyuma ve yavaş performansa sahip olduğu gözlemlenmiştir, bu nedenle ikinci model önerilmiştir. Bu iki modelin de etkili bir şekilde çalıştığı ve daha zorlu ortamlarda da başarılı olduğu sonucuna varılmıştır.

An effective gender recognition approach using voice data via deeper LSTM networks, Long Short-Term Memory (LSTM) yöntemi kullanılmıştır. Bu çalışmada, bir ses veri seti üzerinde cinsiyet tanımlama amaçlı olarak derin Long Short-Term Memory (LSTM) ağları kullanılmıştır. Çalışma, ses veri setinden cinsiyeti %98,4 başarı oranıyla tahmin etmeyi başarmıştır. Önerilen yaklaşım üç ana adımdan oluşmaktadır: özelliklerin azaltılması, Derin LSTM'lerin oluşturulması ve oluşturulan Derin LSTM ağlarının test edilmesi. Çalışmanın sonuçları, önerilen yöntemin %98,4 doğruluk oranı ile cinsiyet tahmininde oldukça başarılı olduğunu göstermektedir. Ayrıca, duyarlılık ve özgünlük gibi performans metrikleri de hesaplanmıştır. Önerilen yöntemin, duyarlılık ve özgünlük değerlerinin sırasıyla %97,2 ve %99,5 olduğu belirtilmektedir. Sonuç olarak, bu çalışma ses veri setleri üzerinde cinsiyet tanımlama için Derin LSTM ağlarının etkili bir şekilde kullanılabileceğini göstermektedir.

A novel octopus based Parkinson’s disease and gender recognition method using vowels, SVD (Singular Value Decomposition - Tekil Değer Ayrışımı), NCA (Neighborhood Component Analysis - Komşuluk Bileşen Analizi) yöntemleri kullanılmıştır. Bu çalışmada, cinsiyet ve Parkinson hastalığının tanınması için bir yöntem önerilmiştir. Önerilen yöntem, ses verilerini analiz etmek için kullanılan bir dizi özellik havuzlama yöntemi ve SVD (Singular Value Decomposition - Tekil Değer Ayrışımı) gibi bir özellik çıkarma tekniği içermektedir. Ardından, öznitelik seçimi için NCA (Neighborhood Component Analysis - Komşuluk Bileşen Analizi) tabanlı bir yöntem kullanılmış ve seçilen öznitelikler geleneksel makine öğrenimi yöntemleriyle sınıflandırılmıştır. Son olarak, tahmin edilen değerlere dayalı olarak bireysel sonuçlar elde etmek için mod tabanlı bir son işleme yöntemi uygulanmıştır.

A machine learning approach for gender identification using statistical features of pitch in speeches, Evrişimli Sinir Ağları (CNN 1D), Çok Katmanlı Algılayıcılar (MLP), Support Vector Machine (SVM), Lojistik Regresyon (LR) yöntemleri kullanılmıştır. Bu çalışma, konuşma işleme sistemlerinde cinsiyet tanımlamanın önemli bir rol oynadığını ve bu alanlardaki performansları artırabileceğini vurgulamaktadır. Sıkça kullanılan özellik çıkarma yöntemleri arasında MFCC, Lineer Tahmin Katsayıları (LPC), Mel Frekans Spektral Katsayıları (MFSC), Spektrogramlar, Enerji Entropisi, Sıfır Çaprazlama Oranı, Kısa Süreli Enerji ve Spektral Akış bulunmaktadır. Cinsiyete özgü özellikler arasında ses perdesi, yoğunluk ve formant frekansları bulunmaktadır. Çoğu model, cinsiyet tanımlamak için bir özellik olarak ses perdesini kullanmış ve %60 ile %98,65 arasında doğruluk oranı sağlamıştır.
Bu çalışmada, "PFG" olarak adlandırılan Pitch Feature set for Gender identification adlı özellik seti önerilmektedir. Bu özellik seti, istatistiksel ölçümler ve makine öğrenimi sınıflandırıcıları aracılığıyla elde edilmiştir. Özellik setinin performansı, evrişimli sinir ağı (CNN 1D) ve diğer geleneksel makine öğrenimi sınıflandırıcıları - Çok Katmanlı Algılayıcı (MLP), Destek Vektör Makineleri (SVM) ve Lojistik Regresyon (LR) ile değerlendirilmiştir.

Parallel Gated Recurrent Unit Networks as an Encoder for Speech Recognition, RNN, GRU, LAS yöntemleri kullanılmıştır. Bu çalışmada, değiştirilmiş LAS ağıyla fonem tanıma performansı incelendi. Bu çalışmada TIMIT veritabanı kullanıldı. Kodlayıcı kısmın tek GRU yerine, iki ve dört GRU paralel olarak kullanıldı. 128 ve 64 gizli düğüm boyutları, temel ağın 256 düğümüne kıyasla düşünüldü. İki GRU'nun 64 gizli düğümü, temel modele kıyasla daha kötü performans gösterdi, ancak performans kaybı %1'in altında kaldı ve yaklaşık olarak 3 kat daha az öğrenilebilir parametre kullanıldı. Diğer değişiklikler ise daha iyi performanslar elde etti. En iyi performans, 128 gizli düğümlü dört GRU ile elde edildi ve %16,71 PER olarak gerçekleşti. En iyi durum için göreceli hata azalması %5.43'tür. Sonuçlar, kodlayıcıdan-dekodlayıcıya türünden uçtan uca ağların performanslarının bu hafif değişiklikle artırılabileceğini göstermektedir. Paralel GRU'ların kullanılmasıyla, öğrenilebilir parametre sayısı azaltıldı ve performansı tehlikeye atmadan. Ancak, ağ küçüldükçe performansı azalacaktır. Dolayısıyla, farklı ağları deneysel olarak karşılaştırarak bir tatlı nokta hedeflenmelidir.
Daha ileri iyileştirmeler için, farklı aktivasyon fonksiyonları, farklı dropout oranları vb. gelecekteki çalışmalarda düşünülecektir.

Speech-to-Gender Recognition Based on Machine Learning Algorithms, Lojistik Regresyon, Karar Ağaçları, Random Forests yöntemleri kullanılmıştır. Bu çalışmada ses sinyallerinden cinsiyet tahmini için MFCC ve spektrogram kullanarak özellik çıkarımı yapılmıştır. Farklı makine öğrenimi algoritmaları (Lojistik Regresyon, Karar Ağaçları, Rastgele Ormanlar, XGB vb.) test edilmiş ve hibrit model ile en iyi doğruluk oranı (%89) lojistik regresyon kullanılarak elde edilmiştir. Model, hatırlama, kesinlik ve f-skoru değerleri açısından yüksek performans göstermiş ve cinsiyet sınıflandırmasında başarılı bulunmuştur.

Gender Recognition Using Cognitive Modeling, Linear Discriminant Analysis (LDA), Support Vector Machine (SVM) yöntemleri kullanılmıştır. Araştırma, yüzlerin “cinsiyet gücünü” tahmin etmek için kognitif modelleme kullanıyor. Bu, geleneksel ikili cinsiyet etiketlemesinin ötesine geçen sürekli bir sınıf değişkenidir. Çalışma, dört farklı veri setini kullanarak sonuçların tek bir protokole karşı önyargılı olmadığını göstermeye çalışıyor. Ayrıca, eğitim örneklerini çıkarmak suretiyle bilinen sınıflandırma algoritmalarının performansının nasıl iyileştirilebileceğini değerlendiriyor.

Convolutional LSTM model for speech emotion recognition, Evrişimli Sinir Ağları (CNN), Long Short-Term Memory (LSTM), Convolutional LSTM (CoLSTM) yöntemleri kullanılmıştır. Metin, sanal kişisel asistanların duygu tanıma yeteneklerinin geliştirilmesi gerektiğini vurguluyor. Ses temelli duygu analizi, sinyallerin işlenmesi, öznitelik çıkarımı ve modellerle analiz edilmesi süreçlerini içeriyor. Önemli zorluklar arasında doğru veri setlerinin elde edilmesi, farklı veri setlerinin birleştirilmesi ve bireysel farklılıkların etkisi bulunuyor. RAVDESS, TESS ve EmoDB gibi veri setleri bu alanda yaygın olarak kullanılıyor.

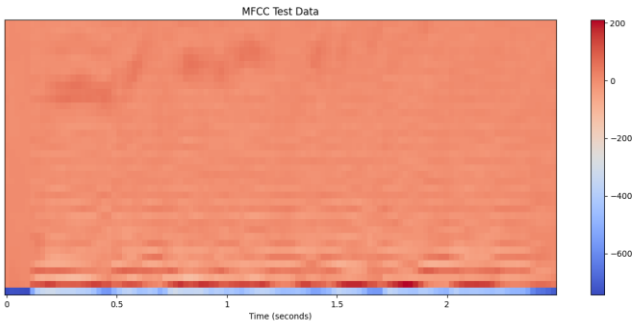
2.3. Veri Özellik Çıkarımı:

2.3.1. MFCC (Mel-Frequency Cepstral Coefficients):

MFCC, ses sinyalinin frekans ve zamandaki özelliklerini temsil etmek için kullanılan bir özellik setidir. İnsan işitmesinin frekans algısına daha yakın bir şekilde çalışmak için ses sinyalinin frekans ölçeğini mel ölçeğine dönüştürür Denklem 1'deki formül kullanılır.

$$Mel = 2595 * log_{10}\left(1 + \frac{f}{700}\right) \qquad (1)$$

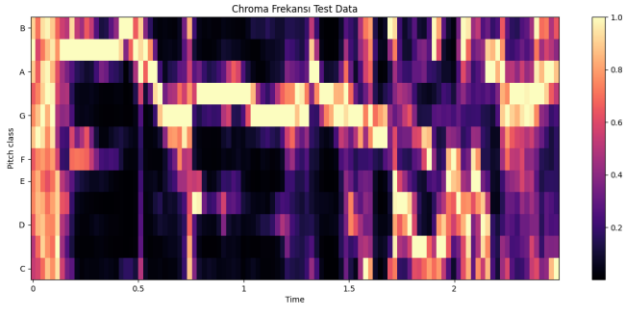
Ardından, mel ölçeğindeki özelliklerin zamanda değişimini temsil etmek için cepstral analiz uygulanır. MFCC, konuşma tanıma ve konuşmacı tanıma gibi birçok ses işleme uygulamasında yaygın olarak kullanılan etkili bir özellik setidir. Şekil 3'de TIMIT test veri setinden örnek bir ses sinyalinin MFCC grafiği yer almaktadır.



Şekil 3: *Test veri setinden örnek MFCC*

2.3.2. Chroma Frekansı:

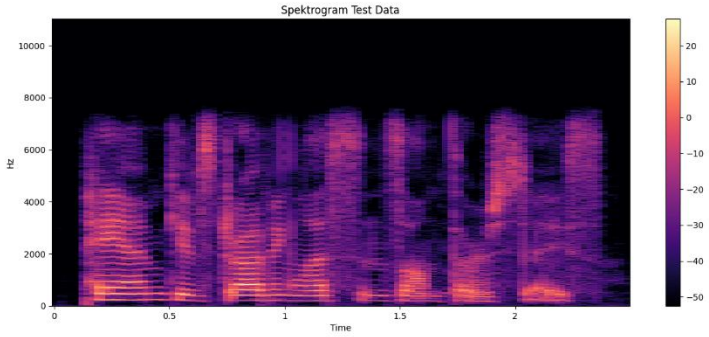
Spektrum müzikal oktavının 12 farklı yarı tonunu (chroma) temsil eden 12 parçanın belirttiği ses için güçlü bir sunumudur. Şekil 4'te TIMIT test veri setinden örnek bir ses sinyalinin Chroma Frekans grafiği yer almaktadır.



Şekil 4: *Test veri setinden örnek Chroma Frekansı*

2.3.3. Spektrogram:

Belirli bir dalga formunda bulunan çeşitli frekanslarda bir sinyalin sinyal gücünü veya yüksekliğini temsi eden görseldir. Şekil 5'te TIMIT test veri setinden örnek bir ses sinyalinin Spektrogram grafiği yer almaktadır.



Şekil 5: *Test veri setinden örnek Spektrogram*

3. Deneysel Kurulum ve Sonuçlar:

3.1. Model Geliştirme:

Model geliştirme aşamasında, LSTM veya RNN gibi derin öğrenme mimarileri kullanılacaktır. Bu mimariler, ardışık verileri işlemek için özel olarak tasarlanmıştır ve zaman serileri gibi dinamik veri yapılarını modellemek için oldukça etkilidir. Model eğitimi için TIMIT veri tabanından ses örnekleri kullanılacak ve doğruluk oranını artırmak için modelin hiper parametreleri ayarlanacaktır.

3.1.1. Recurrent Neural Network (RNN):

RNN'ler, zaman serisi verileri gibi ardışık girdileri işlemek için tasarlanmış yapay sinir ağlarıdır. Her adımda, RNN bir önceki adımdan gelen girdiye ek olarak mevcut girdiyi de alır ve bir sonraki adıma aktarır. RNN'lerin birçok uygulaması vardır, özellikle dil işleme, metin sınıflandırma, zaman serisi tahmini gibi alanlarda kullanılırlar. Ancak, geleneksel RNN'lerin sorunu, uzun vadeli bağımlılıkları başarılı bir şekilde öğrenememeleridir. Bu durum, "gradientsiz kaybolma" olarak adlandırılan bir soruna yol açar.

$$h_t = \tanh(W_h * x_t + U_h * h_{t-1} + b_h) \;\rightarrow\; SimpleRNN \qquad (2)$$

Modelde kullanılan RNN katmanının denklemi Denklem 3'te yer almaktadır. Denklem 2'deki parametrelerin açıklamaları:

W_h: Giriş verisi ağırlıkları

U_h: Gizli durum (hidden state) ağırlıkları

b_h: Bias vektörü

h_t: Her adımda hesaplanan gizli durum

$$\hat{y} = \sigma(w * ReLU(W_d(Dropout(\tanh(W_h * x_t + U_h * h_{t-1} + b_h)) + b_d) + b) \qquad (3)$$

3.1.2. Long Short-Term Memory (LSTM):

LSTM, RNN'lerin uzun vadeli bağımlılıkları öğrenme yeteneğini geliştirmek için tasarlanmış bir türüdür. LSTM'ler, hafıza hücreleri ve bir dizi kontrol kapısı kullanarak geleneksel RNN'lerden farklıdır. Hafıza hücreleri, bilginin uzun vadeli saklanmasını sağlar, kontrol kapıları ise hangi bilginin hafızada saklanacağını ve hangisinin unutulacağını kontrol eder. Bu yapı, uzun vadeli bağımlılıkları daha etkili bir şekilde öğrenmeyi mümkün kılar. Dil işleme, metin oluşturma, çeviri, duygu analizi ve zaman serisi tahmini gibi birçok alanda başarılı bir şekilde kullanılmıştır.

$$i_t = \sigma(W_i * x_t + U_i * h_{t-1} + b_i) \rightarrow Giriş\ Kapısı \quad (4)$$

$$f_t = \sigma\left(W_f * x_t + U_f * h_{t-1} + b_f\right) \rightarrow Unutma\ Kapısı \quad (5)$$

$$o_t = \sigma(W_o * x_t + U_o * h_{t-1} + b_o) \rightarrow Çıkış\ Kapısı \quad (6)$$

$$\tilde{c}_t = \sigma(W_c * x_t + U_c * h_{t-1} + b_c) \rightarrow Hücre\ Adayı \quad (7)$$

3.1.3. Gated Recurrent Unit (GRU):

GRU, LSTM'in daha basitleştirilmiş bir versiyonu olarak görülebilir. GRU, LSTM'deki gibi bilgiyi saklama ve atma mekanizmalarını kullanır fakat daha az parametre ile bu işlemi gerçekleştirir.

$$z_t = \sigma(W_z * x_t + U_z * h_{t-1} + b_z) \rightarrow Giriş\ Kapısı \quad (8)$$

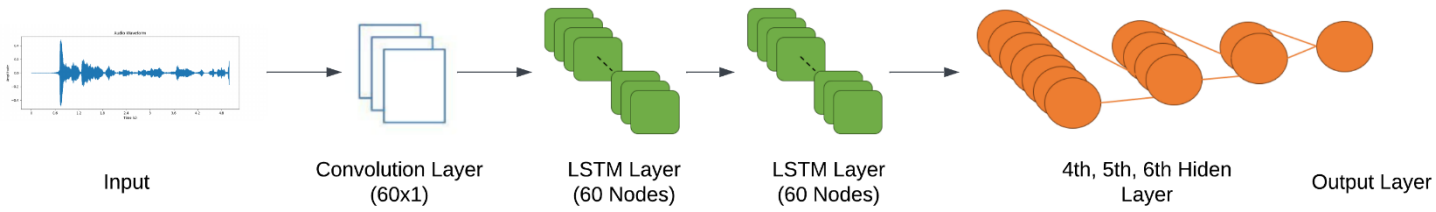
$$r_t = \sigma(W_r * x_t + U_r * h_{t-1} + b_r) \rightarrow Unutma\ Kapısı \quad (9)$$

$$\tilde{h}_t = tang(W_h * x_t + U_h * (r_t \odot h_{t-1}) + b_h) \rightarrow Aday\ Gizli\ Durum \quad (10)$$

$$h_t = r_t \odot \tilde{h}_t + (1 - z_t) \odot h_{t-1} \rightarrow Gizli\ Durum\ Güncellemesi \quad (11)$$

3.1.4. Convolutional LSTM (Co-LSTM):

Geleneksel LSTM'de girişler ve gizli durumlar tam bağlantılı katmanlar ile işlenirken, Co-LSTM'de bu işlemler evrişimsel (convolutional) filtreler ile gerçekleştirilir. Co-LSTM'de, hücre durumu (C_t) ve gizli durum (H_t) matrisleri, her zaman adımında evrişimsel filtreler ile güncellenir. Şekil 6'da modelin şeması verilmiştir.



Şekil 6: *Co-LSTM Model Akış Şeması*

3.1.5. CNN ve LSTM:

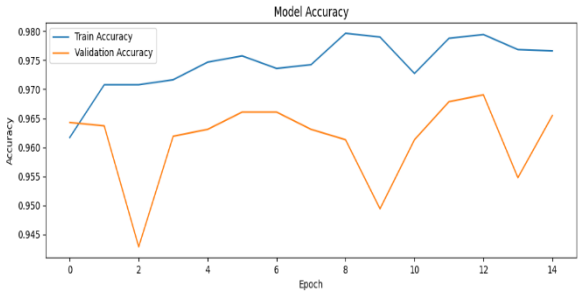
Ham ses dosyalarının (wav formatta) doğrudan CNN katmana vererek öznitelik çıkarım işlemi gerçekleştirmiş olunur. Sonrasında LSTM katmanlarından faydalanarak model tamamlanmış olur. CNN katmanları, ses sinyalinden yerel örüntüleri öğrenirken, LSTM katmanları, sesin zaman içerisindeki dinamiklerini modellemiştir. Bu yaklaşım, özellik çıkarımı adımını veri işleme sürecinin bir parçası haline getirerek geleneksel yöntemlerden bağımsız bir yapı sunmuş ve daha esnek bir model elde edilmesini sağlamıştır. Model, TIMIT veri seti üzerinde eğitilmiş ve test edilmiş olup, ham veriyle yapılan tahminlerde yüksek bir başarı oranı sergilemiştir.

Şekil 6'da modelin akış şeması yer almaktadır. Her çalışma 15 döngü (epoch) çalıştırılmış, 0.001 öğrenme adımı (learning rate) kullanılmış ve grup boyutu (batch) olarak 32 ayarlanmıştır.

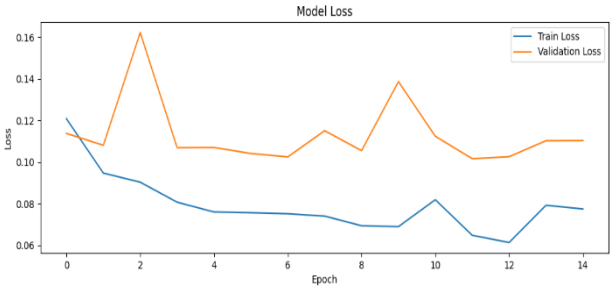
Şekil 7: *CNN + LSTM Model Akış Şeması*

3.2. Deneysel Sonuçlar ve Tartışma:

Şekil 8 ve 9'da RNN model ile 15 epochs eğitimin başarı ve kayıp grafiği yer almaktadır.

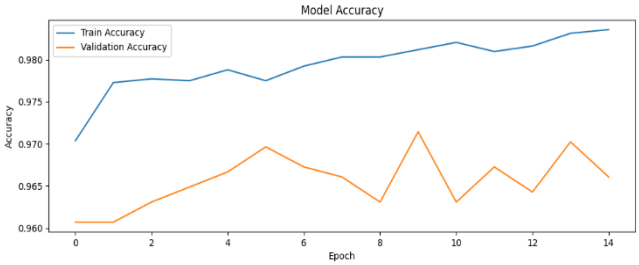


Şekil 8: *15 Epochs için RNN Model Başarısı*

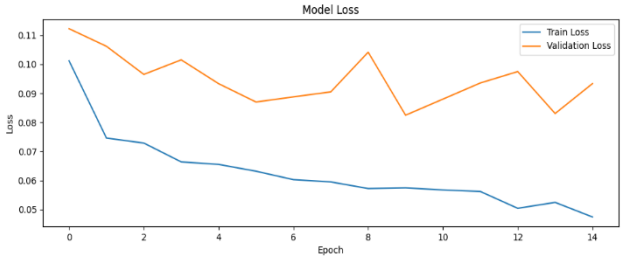


Şekil 9*: 15 Epochs için RNN Model Kaybı*

Şekil 10 ve 11'de LSTM modeli ile 15 epochs eğitimin başarı ve kayıp grafikleri yer almaktadır.

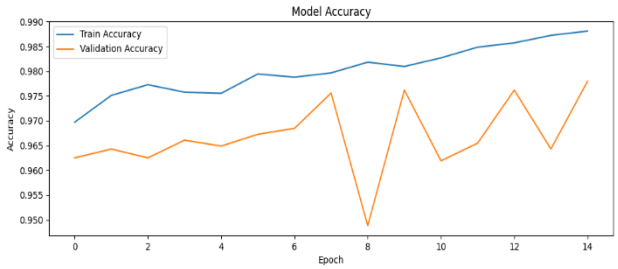


Şekil 10: *15 Epochs için LSTM Model Başarısı*

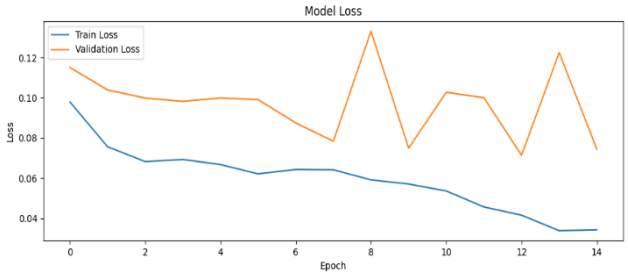


Şekil 11: *15 Epochs için LSTM Model Kaybı*

Şekil 12 ve 13'de GRU model ile 15 epochs eğitimin başarı ve kayıp grafiği yer almaktadır.

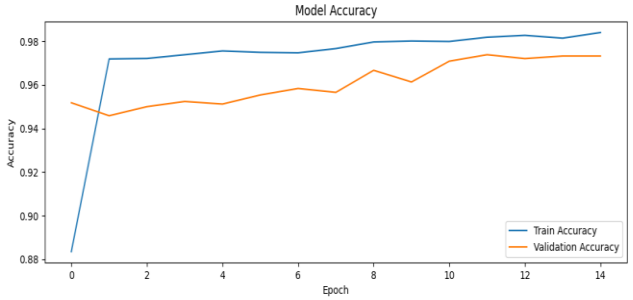


Şekil 12: *15 Epochs için GRU Model Başarısı*

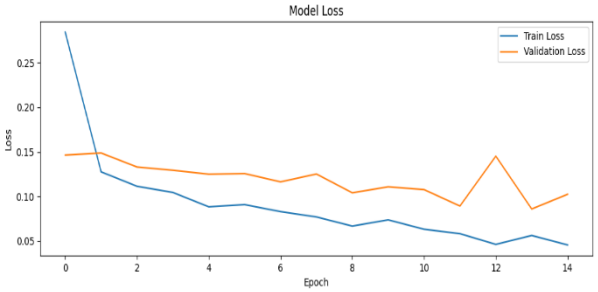


Şekil 13: *15 Epochs için GRU Model Kaybı*

Şekil 14 ve 15'te Co-LSTM model ile 15 epochs eğitimin başarı ve kayıp grafiği yer almaktadır.

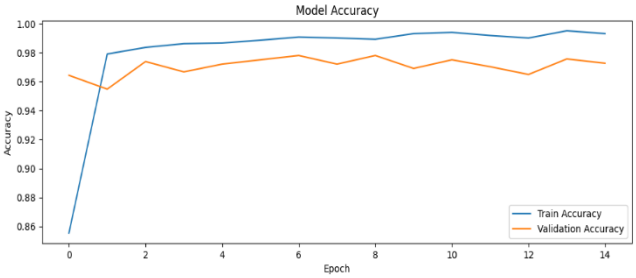


Şekil 14: *15 Epochs için CoLSTM Model Başarısı*

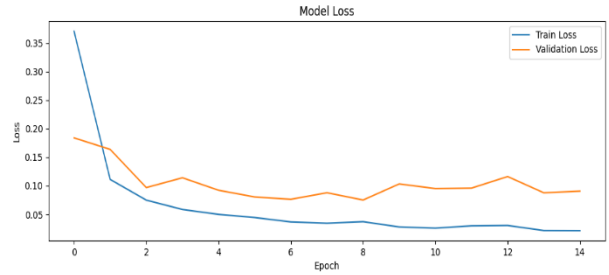


Şekil 15: *15 Epochs için CoLSTM Model Kaybı*

Şekil 16 ve 17'de CNN + LSTM modelin 15 epochs eğitiminin başarı ve kayıp grafiği yer almaktadır.



Şekil 16: *15 Epochs CNN+LSTM Model Başarısı*



Şekil 17: *15 Epochs CNN+LSTM Model Kaybı*

4. Sonuç:

Bu proje kapsamında, TIMIT veri seti kullanılarak ses verilerinden cinsiyet tahmini gerçekleştirilmiştir. Farklı yapay zekâ mimarileri (RNN, LSTM, GRU, Co-LSTM ve CNN+LSTM) karşılaştırılmış, eğitim süreci boyunca modellerin başarı ve kayıp oranları grafiklerle incelenmiştir.

LSTM modeli, cinsiyet tanımlama görevi için yüksek doğruluk oranlarına ulaşmıştır. Grafiklerden de görüldüğü gibi, eğitim ve doğrulama başarı oranları diğer modellere kıyasla daha iyi bir performans sergilemiştir. CNN+LSTM mimarisi, özellik çıkarımı ve zaman içindeki dinamiklerin modellenmesinde etkili bir yaklaşım sunmuş, ancak eğitim sürecindeki kayıp oranları nispeten yüksek kalmıştır. Co-LSTM modeli, geleneksel LSTM'den farklı olarak evrişimsel filtreleri kullanarak yerel örüntüleri daha iyi öğrenmiş ve bazı metriklerde daha dengeli sonuçlar elde etmiştir. GRU modeli, düşük parametre sayısıyla performans kaybını minimumda tutmuş ve enerji verimliliği sağlamıştır. RNN, ardışık verilerle çalışma yeteneği nedeniyle temel bir yöntem olarak ele alınmış, ancak uzun vadeli bağımlılıkların öğrenilmesinde zorluklar yaşamıştır.

Deneysel sonuçlar, LSTM ve türevlerinin (Co-LSTM ve CNN+LSTM) bu tür ardışık veri işleme görevlerinde diğer yöntemlere göre daha başarılı olduğunu göstermektedir. Bu çalışmada kullanılan modellerin başarı oranları, cinsiyet tanımlama görevinin TIMIT veri tabanı üzerinde yüksek doğruluklarla gerçekleştirilebileceğini ortaya koymuştur.

Kaynakça:

Nasef, M. M., Sauber, A. M., & Nabil, M. M. (2021). Voice gender recognition under unconstrained environments using self-attention. *Applied Acoustics*, *175*. https://doi.org/10.1016/j.apacoust.2020.107823

Kwasny, D., & Hemmerling, D. (2021). Gender and age estimation methods based on speech using deep neural networks. *Sensors*, *21*(14). https://doi.org/10.3390/s21144785

GÜNDÜZ, G., & CEDİMOĞLU, İ. H. (2019). Derin Öğrenme Algoritmalarını Kullanarak Görüntüden Cinsiyet Tahmini. *Sakarya University Journal of Computer and Information Sciences*, *2*(1). https://doi.org/10.35377/saucis.02.01.517930

Barkana, B. D., & Zhou, J. (2015). A new pitch-range based feature set for a speaker's age and gender classification. *Applied Acoustics*, *98*. https://doi.org/10.1016/j.apacoust.2015.04.013

Ertam, F. (2019). An effective gender recognition approach using voice data via deeper LSTM networks. *Applied Acoustics*, *156*. https://doi.org/10.1016/j.apacoust.2019.07.033

Tuncer, T., & Dogan, S. (2019). A novel octopus based Parkinson's disease and gender recognition method using vowels. *Applied Acoustics*, *155*. https://doi.org/10.1016/j.apacoust.2019.05.019

Shagi, G. U., & Aji, S. (2022). A machine learning approach for gender identification using statistical features of pitch in speeches. *Applied Acoustics*, *185*. https://doi.org/10.1016/j.apacoust.2021.108392

TÜFEKCİ, Z., & DİŞKEN, G. (2022). Parallel Gated Recurrent Unit Networks as an Encoder for Speech Recognition. *European Journal of Science and Technology*. https://doi.org/10.31590/ejosat.1103714

HIZLISOY, S., ÇOLAKOĞLU, E., & ARSLAN, R. S. (2022). Speech-to-Gender Recognition Based on Machine Learning Algorithms. *International Journal of Applied Mathematics Electronics and Computers*, *10*(4). https://doi.org/10.18100/ijamec.1221455

Jo, A. H., & Kwak, K. C. (2023). Speech Emotion Recognition Based on Two-Stream Deep Learning Model Using Korean Audio Information. *Applied Sciences (Switzerland)*, *13*(4). https://doi.org/10.3390/app13042167

Zhang, S., Zhao, X., & Tian, Q. (2022). Spontaneous Speech Emotion Recognition Using Multiscale Deep Convolutional LSTM. *IEEE Transactions on Affective Computing*, *13*(2). https://doi.org/10.1109/TAFFC.2019.2947464

YALMAN, H. İ., & TÜFEKCİ, Z. (2022). Yeni Bir Türkçe Sesli Kitap Veri Seti Üzerinde Convolutional RNN+CTC, LSTM+CTC ve GRU+CTC Modellerinin Karşılaştırılması. *European Journal of Science and Technology*. https://doi.org/10.31590/ejosat.1082109