Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı 32, S. 471-483, Aralık 2021

Derleme Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 32, pp. 471-483, December 2021

Review Article

Konuşmadan Duygu Tanıma Üzerine Detaylı bir İnceleme: Özellikler ve Sınıflandırma Metotları

Emel Çolakoğlu[1*], Serhat Hızlısoy[2], Recep Sinan Arslan[3]

[1*] Kayseri Üniversitesi, Lisansütü Eğitim Enstitüsü, Hesaplamalı Bilimler ve Mühendislik Anabilim Dalı, Kayseri, Türkiye, (ORCID: 0000-0003-1755-3130), emelcolakoglu@gmail.com

[2] Kayseri Üniversitesi, Mühendislik Mimarlık ve Tasarım Fakültesi, Bilgisayar Mühendisliği Bölümü, Kayseri, Türkiye (ORCID: 0000-0001-8440-5539), serhathizlisoy@kayseri.edu.tr

[3] Kayseri Üniversitesi, Mühendislik Mimarlık ve Tasarım Fakültesi, Bilgisayar Mühendisliği Bölümü, Kayseri, Türkiye (ORCID: 0000-0002-3028-0416), sinanarslanemail@gmail.com

(International Conference on Design, Research and Development (RDCONF) 2021 – 15-18 December 2021)

(**DOI:** 10.31590/ejosat.1039403)

ATIF/REFERENCE: Çolakoğlu, E., Hızlısoy, S. & Arslan, R. S. (2021). Konuşmadan Duygu Tanıma Üzerine Detaylı Bir İnceleme: Özellikler ve Sınıflandırma Metodları. *Avrupa Bilim ve Teknoloji Dergisi*, (32), 471-483.

Öz

Konuşma insanlar arasındaki hızlı ve en doğal iletişim yöntemlerindendir. Konuşmadan duygu tanıma çalışmaları, konuşma sırasında çıkan ses sinyalinden anlam bilgisini elde etmeye çalışmaktadırlar. Son yıllarda konuşma sinyalleri üzerinden duygu analizi ile ilgili olarak birçok çalışma yapılmıştır. Bu çalışmalarda duygu analizinde 3 önemli yön dikkate alınarak detaylı bir araştırma yapılmıştır. Birinci konu konuşma sinyallerinden öznitelik çıkarma, ikinci konu bu özniteliklerden sınıflandırmaya olumlu katkısı olacakların seçimi ve üçüncü konu ise sınıflandırma şemalarının tasarımı ve performans değerlendirmesidir. Özniteliklerin doğru belirlenmesi, öznitelikler üzerinde seçme işleminin başarılı bir şekilde uygulanması performansı büyük ölçüde etkilemektedir. Ancak sesten özniteliklerin çıkarılması, ve sınıflandırılmasında farklı yöntemler tercih edilse de performans veri setlerine, duygu durumlarına, dillere, eğitim setinin kullanım yöntemine göre değişebilmektedir. İncelenen makaleler kapsamında sınıflandırıcı olarak en sık SVM ve öznitelik olarak da MFCC kullanılmıştır. En yüksek tanıma oranı ise TESS veri setinde oto-kodlayıcı ve Alex-net CNN ile sağlanmış ve %98 başarı elde edilmiştir.

Anahtar Kelimeler: Konuşmadan Duygu Tanıma, Derleme, Öznitelik Çıkarım Teknikleri, Sınıflandırma.

A Detailed Survey on Speech Emotion Recognition: Features and Classification Methods

Abstract

Speech is one of the fastest and most natural communication methods between people. Emotion recognition studies without speech try to obtain semantic information from the sound signal during speech. In recent years, many studies have been carried out on emotion analysis over speech signals. In these studies, detailed research was conducted by considering 3 important aspects in sentiment analysis. The first topic is feature extraction from speech signals, the second topic is the selection of these features that will contribute positively to the classification, and the third topic is the design and performance evaluation of the classification schemes. The correct determination of the features and the successful implementation of the selection process on the features greatly affect the performance. However, although different methods are preferred in the extraction and classification of features from the voice, the performance may vary according to the data sets, moods, languages, and the method of use of the training set. Generally, among the articles examined, SVM was used as the classifier and MFCC was used as the feature. The highest recognition rate was achieved with the auto-encoder, TESS dataset and Alex-net CNN and 98% success was achieved.

Keywords: Speech Emotion Recognition, Survey, Feature Reduction Techniques, Classification.

[*] Sorumlu Yazar: emelcolakoglu@gmail.com

1. Giriş

İnsanoğlunun var oluşundan beri iletişim bilgi alışverişinin temelidir. İletişimi daha doğru, net ve anlaşılır kılmak için kelimeler ve duygular birbirlerini takip etmektedir. İnsanların duygusal durumuna bağlı olarak vücut hareketleri, kan basıncı, nabız, ses tonu gibi bazı fizyolojik değişiklikler olmaktadır. Nabız, kan basıncı gibi değişiklikler özel bir cihazla tespit edilirken, ses tonu, yüz ifadesi gibi değişiklikler ise cihaz gerektirmeden anlaşılabilmektedir. Duygu tahminleri için genellikle makineler kullanılmaktadır. (Süha Gökalp ve diğerleri, 2021).

Konuşma insanlar arasındaki hızlı ve en doğal iletişim yöntemlerindendir. Bu nedenle araştırmacılar insan ve makine etkileşimini daha hızlı ve verimli hale getirmek için konuşma sinyalini kullanmaya başlamıştır. Konuşma sinyalleri, konuşmacının yaşı, ruh hali, cinsiyeti, fizyolojisi, dili gibi birçok bilgiyi aynı anda barındırabilen karmaşık bir yapıya sahiptir. Konuşmadan duygu tanıma çalışmaları, konuşma sırasında çıkan ses sinyalinden anlam bilgisini elde etmeye çalışmaktadırlar. (Süha Gökalp ve diğerleri, 2021).

Duygu tanıma çalışmaları son zamanlarda oldukça ilgi görmeye ve ilerleme kaydetmeye başlamıştır. Makinelerden elde edilen mekanik sesler insanlarda olumsuz etkiler oluşturabilmektedir. Bu sorunu çözebilmek adına makinelere duygu içerikli konuşmalar yapabilmeleri ve insanların duygularını anlayabilmeleri için yeni özellikler kazandırılmaya çalışılmaktadır. (Cevahir Parlak ve diğerleri 2014).

2. Konuşmadan Duygu Tanıma

Konuşma sinyalleri insanlar arasındaki hızlı ve en doğal iletişim yöntemlerindendir. Bu durum araştırmacıları insan ve makine etkileşimini daha hızlı ve verimli hale getirmek için konuşma sinyalini kullanmaya yöneltti. Ancak bu durum makinelerin insan sesini tanımada kusursuz çalışması sonucunu doğurdu. (Onur Erdem Korkmaz, 2016).

Konuşmadan duygu tanıma özellikle insan-bilgisayar arasında doğal etkileşim gerektiren web filmleri ve kullanıcının algılanan duygusuna bağlı olarak tepki veren bilgisayarlı öğretici uygulamalarında kullanılmaktadır. Ayrıca sürücünün zihinsel durumuna göre güvenlik sistemlerini ayarlayan otomobil uygulamaları da bulunmaktadır. Bunlara ilaveten tedavi uzmanları için hastanın duygusunu teşhis eden araçlar tasarlanabilir. Konuşmacının duygu durumunun önemli olduğu otomatik çeviri sistemlerinde de kullanımı faydalı olabilir.

Uçak kokpitlerinde konuşmacı tanıma sistemleri bulunur. Stres halinde bu sistemlerim çalışması verimsiz hale gelmektedir. Duygu tanıma sistemleriyle bu istenmeyen durum ortadan kaldırılmaya çalışılmıştır. Konuşmadan duygu tanıma ayrıca çağrı merkezi sistemlerinde ve mobil haberleşme uygulamalarında da kullanılmaktadır. (Onur Erdem Korkmaz, 2016).

3. Konuşmadan Duygu Tanıma Mimarileri

Konuşma duygu tanıma genellikle üç bölümden oluşur: öznitelik çıkarma, öznitelik seçimi ve sınıflandırma (Shadi Langari ve diğerleri, 2020).

3.1. Öznitelik Çıkarma

Bir problemi makine öğrenmesi ile çözebilmek için uygun özniteliklere sahip olmamız gerekir. Ama her zaman elimizdeki problemde doğrudan kullanabileceğimiz nitelikler olmayabilir. Bu durumda veriden özniteliklerin çıkarılması gerekmektedir. "İşaret İşleme" zaman serileri ile uğraşan bilim dalıdır. "Görüntü İşleme" fotoğraf, video gibi görsel veriler ile uğraşan bilime verilen isimdir. "Örüntü tanıma" ise içeriğinde hem zaman serileri hem de görüntüler olabilen her türlü işaretten özellik çıkarmayı amaçlayan bilim dalıdır. Yapılan önceki çalışmaları incelediğimizde çok farklı yöntem ve araçlar ile öznitelik çıkarımı yapıldığı görülmektedir. Ancak kullanılacak özniteliklerin optimum seviyede seçilmesi önem arzetmektedir. Çünkü ne kadar çok öznitelik kullanılırsa makine öğrenme sürecinde maliyetler de o kadar artar. Ayrıca amaca yönelik doğru özniteliklerin seçilmesi de çok önemlidir. (Orhan, 2021)

Genel olarak, konuşmadan duygu tanımada (SER) iki öznitelik kategorisi kullanılır: bunlar prozodik öznitelik ve ses yolu sistemi özniteliklerdir. İlk kategori Perde (Pitch), Enerji (Energy) ve Süre (Duration) gibi prozodik verilerden elde edilir. İkinci kategori, Mel Frekansı Kepstrum Katsayısı (MFCC), Doğrusal Öngörülü Kepstrum Katsayıları (LPCC), Formantlar (Formants) ve Ayrık Fourier Dönüşümü (DFT) harmonikleri gibi Kepstral katsayılarını içeren ses yolu ile ilgilidir.

Çoğu konuşmadan duygu tanıma (SER) çalışması, Doğrusal Öngörülü Cepstral Katsayıları (LPCC), Mel Frekansı Kepstrum Katsayıları (MFCC) ve Formantlar (Formants) gibi ses yolundan çıkarılan veriler olarak spektral öznitelikleri kullanır. (Shadi Langari ve diğerleri, 2020). İncelenen makalelere bakıldığında sesten duygu tanıma da aşağıdaki özniteliklerin genellikle kullanıldığı gözlenmiştir: Enerji (Energy), Perde (pitch), Doğrusal Öngörülü Cepstral Katsayıları (LPCC), Mel Frekansı Kepstrum Katsayıları (MFCC), Mel-Enerji Spektrumu Dinamik Katsayıları (MEDC), Mel Ölçekli Spektrogram (Mel-scaled spectrogram), Kromagram (Chromagram), Spektral Kontrast Özelliği (Spectral contrast feature), Tonnetz Temsili (Tonnetz representation), Mel Frekans Büyüklük Katsayısı (Mel frequency magnitude coefficient), Log Frekansı Güç Katsayısı (LFPC), Üst Düzey İstatistiksel Fonksiyonlar (HSF), Sıfır Geçiş Oranı (ZCR), Teager Enerji Operatörü (TEO), Harmonik Gürültü Oranı (HNR), Ayrık Kesirli Fourier Dönüşümü (DFrFT), Ayrık Fourier dönüşümü (DFT), Dalgacık Paket Sayısı (WPC).

Özniteliklerin çıkarılmasında ise aşağıdaki programlar kullanılmıştır:

- OpenSMILE
- Praat
- MIRtoolbox
- JAudio

3.2. Öznitelik Seçme

Genel olarak, yüksek boyutluluk yüksek olasılıkla sınıflandırmanın doğruluğunu ve verimliliğini etkiler. Bu yüzden ideal doğruluğu sağlamak ve daha kısa bilgi işlem süresi için öznitelik boyutu küçültülmelidir. Konuşmadan duygu tanımada (SER) genelde kullanılan bazı öznitelik seçme metodları vardır. Bunlar; Temel Bileşenler Analizi (PCA), Doğrusal Ayrım Analizi (LDA), İleri Seçimli Sarmalayıcı Yaklaşımı, İleri Özellik Seçimi (FFS), Geriye Doğru Özellik Seçimi (BFS), Sıralı Kayan İleri Seçim (SFFS) metodlarıdır. (Shadi Langari ve diğerleri, 2020)

İncelenen makelelerde ise genellikle aşağıdaki özellik seçme metodları kullanılmıştır: İleri Özellik Seçimi (FFS), Temel Bileşen Analizi (PCA), Aykırı Değer Algılama (ADA), Otomatik Kodlayıcı (auto-encoder), Korelasyon Tabanlı Öznitelik Seçme Yöntemi (Correlation-based Feature Selection), Genetik Algoritmalar (Genetic Algorithm), Cuckoo Arama (Cuckoo Search), Sıralı Kayan İleri Arama (SFFS), Bilgi Kazancı (Information Gain), Ki Kare Analizi (ChiSquared).

3.3. Sınıflandırma

Sınıflandırma kısmında ise farklı sınıflandırıcılar kullanılmaktadır. İncelenen makelelerde kullanılan sınıflandırıcılar ve genel özellikleri aşağıdaki şekilde tanımlanmaktadır.

3.3.1. Naive Bayes

Naive Bayes Bayes teoremine dayanan olasılıklı bir makine öğrenme algoritmasıdır. Bayes teoremi ise koşullu olasılıkları hesaplamak için kullanılan bir matematiksel formüldür. Koşullu olasılık bir olayın meydana gelme olasılığının bir ölçüsüdür. Az bir eğitim verisi ile çok başarılı sonuçlar elde etmek mümkündür.

Naive Bayes ile basit ama hızlı, güvenilir ve doğru sonuçlar elde etmek mümkündür. Birçok çalışmada başarılı sonuçlar vermesiyle birlikte özellikle doğal dil işleme (NLP) alanındaki problemler ile çalışır. Naive Bayes 3 sınıflandırıcı türü vardır. Bunlar: Çok terimli sınıflandırıcı, Bernoulli sınıflandırıcı ve Gauss Saf Bayes sınıflandırıcıdır. (Chauhan, 2021)

Kullanım alanları çok sınıflı tahmin, gerçek zamanlı tahmin, spam filtreleme, duyarlılık analizi, metin sınıflandırması ve öneri sistemleri olarak örneklendirilebilir.

3.3.2. DVM (Destek Vektör Makineleri)

DVM ler iki sınıflı ya da çok sınıflı sınıflandırma problemlerinin çözümü için geliştirilmiş makine öğrenme algoritmasıdır. DVM'ler örüntü tanıma, aykırı değerlerin belirlenmesi, sınıflandırma, regresyon için kullanılmaktadır. Bu yöntem diğer makine öğrenmesi yöntemleri ile karşılaştırıldığında performans ve yeteneği özellikle doğrusal olmayan problemlerde daha iyidir. (Akpınar, 2021)

Destek Vektör makinesinin amacı N boyutlu (N-öznitelik sayısı) bir problem uzayında bi hiper düzlem bulmaktır. Bunu bulabilmek için birçok yöntem vardır. Ancak amaç her iki sınıfın veri noktaları arasındaki maksimum mesafeye sahip bir düzlem bulmaktır. (Gandhi, 2021)

SMO (Sıralı Minimal Optimizasyon) DVM için geliştirilmiş bir eğitim algoritmasıdır. DVM nin yüksek hesaplama ve bellek kullanımı problemine çözüm olarak geliştirilmiştir. DVM için en çok kullanılan methodlardandır. Doğrusal DVM için iyi bir performans sergiler. Özetle SMO optimize edilecek değişkenlerin ikişerli gruplar halinde alt uzaylarda çözülmesi ve farklı kombinasyonlarla oluşturulan ikililerin çözülmesi esasına dayanır. (Akpınar, 2021)

3.3.3. Karar Ağaçları

Karar Ağaçları, tahmin, regresyon ve sınıflandırma da kullanılan oldukça güçlü ve çok tercih edilen bir araçtır. Karar ağaçlarının ilk hücrelerine **kök** (root veya root node) denir. Her bir gözlem kökteki koşula göre "Evet" veya "Hayır" olarak sınıflandırılır. Kök hücrelerinin altında **düğümler** (interval nodes veya nodes) bulunur. Her bir gözlem düğümler yardımıyla sınıflandırılır. Düğüm sayısı arttıkça modelin karmaşıklığı da artar. Karar ağacının en altında **yapraklar** (leaf nodes veya leaves) bulunur. Yapraklar, bize sonucu verir. (Akca, 2021)

3.3.4. Lojistik Regresyon

Lojistik regresyon bağımlı değişken ikili olduğunda yapılacak uygun regresyon analizidir. Tahmine dayalı bir analiz gerçekleştirilir. Logistik regrsyon bir bağımlı ikili değişken ile bir veya daha fazla bağımsız değişken arasındaki ilişkiyi açıklamak için kullanılır. (Statistics Solutions Team, 2021)

3.3.5. RNN (Recurrent Neural Network)

RNN'ler genelde bir sonraki adımı tahmin etmek için kullanılan bir çeşit Derin Öğrenme yapılarıdır. Diğer derin öğrenme yapılarından en büyük farkları ise hatırlamalarıdır. Bir diğer farkları ise, diğer sinir ağlarında her girdi birbirinden bağımsız iken RNN'lerde girdiler birbiri ile ilişkilidir. RNN'ler bir sonraki adımı takip edebilmek için girdiler arasında ilişki kurarlar ve eğitilirken tüm ilişkilerini hatırlarlar. RNN'ler kurmuş oldukları ilişkilerin kalıcı olması için kendi içlerinde dönen döngü benzeri bir yapı kullanırlar. (Akca, 2021)

3.3.6. DNN (deep neural network)

Derin öğrenme, makine öğreniminin bir alt kümesidir. Bilgisayar algoritmalarını inceleyerek kendi kendine öğrenmeye ve geliştirmeye dayanan bir alandır. Derin öğrenme için ilham insan beyninin bilgiyi filtreleme şeklidir. Amacı insan beyninin yapabildiğini yapabilmek için nasıl çalıştığını taklit etmektir. Kelimenin tam anlamıyla yapay bir sinir ağıdır. Derin Sinir Ağları (DNN'ler), her katmanın görüntü, ses ve metin anlamını taşıyan temsil ve soyutlama gibi karmaşık işlemleri gerçekleştirebileceği ağ türleridir. (Protopars Team, 2021)

3.3.7. Evrişimli Sinir Ağı (CNN)

Evrişimli sinir ağı piksel verilerini işlemek için özel olarak tasarlanmış görüntü tanıma ve işlemede kullanılan bir YSA türüdür. CNN nin nöronları insanlarda ve diğer hayvanlarda görsel uyaranların işlenmesinden sorumlu alan olan ön lobunkilere benzer şekilde düzenlenmiştir. Nöron katmanları, geleneksel sinir ağlarının parça parça görüntü işleme probleminden kaçınarak tüm görsel alanı kaplayacak şekilde düzenlenmiştir. (TechTarget Team, 2021)

Bir CNN modeli normalde bir giriş katmanı, çoklu evrişim katmanları, havuz katmanları, tam bağlantılı katmanlar, normalleştirme katmanları ve bir çıkış katmanından oluşur. Günümüzde çok fazla bilinen bazı CNN mimarileri nedir dersek onlar da şu şekilde listeleyebiliriz. (Doğan, 2021)

- GoogLeNet
- ResNet
- LeNet
- MobileNet
- AlexNet
- VGGNet
- ZFNet

Evrişimli Sinir Ağları (CNN) dediğimizde genellikle görüntü sınıflandırması için kullanılan 2 boyutlu bir CNN'yi kastediyoruz. Ancak gerçek dünyada kullanılan 1 boyutlu ve 3 boyutlu CNN'ler olan iki başka Evrişim Sinir Ağı türü daha vardır.

- 3D evrişimli sinir ağları (CNN)
- 1D evrişimli sinir ağları (CNN)

3.3.8. Uzun Kısa Süreli Bellek Sinir Ağı (LSTM)

LSTM uzun kısa süreli bellek ağları anlamına gelmektedir. Uzun vadeli bağımlılıkları öğrenebilen tekrarlayan bir RNN türü sinir ağıdır. LSTM geri besleme bağlantılarına sahiptir. Bu şekilde tüm veri dizisini işleyebilirler. LSTM dil modelleme, makine çevirisi, konuşma tanıma, el yazısı tanıma, resim yazısı, soru cevaplama, videodan metne dönüştürme gibi pekçok problemde oldukça iyi performans gösteren bir RNN türüdür. (Intellipaat Team, 2021)

Tekrarlayan sinir ağları önceki bilgileri kullandıkları için "kısa süreli belleğe" sahiptir. LSTM ise uzun süreli belleği tekrarlayan sinir ağı sürecine dahil eder. Bu nedenle bilgileri uzun süre hatırlayabilmeleri öğrenmek için mücade ettikleri bir süreç olmayıp doğal davranışlarıdır. (Bilimci, 2021)

3.3.9. Gauss Karışımları Modeli (GMM)

Gauss karışım modeli bir tür makine öğrenme algoritmasıdır. Verileri olasılık dağılımına göre farklı kategorilere ayırır. Temelde bir kümeleme algoritmasıdır. Gauss karışım modelleri denetimsiz öğrenme algoritmasıdır. Bu durumda eldeki veri etiketsiz ise kullanılabilmesi mümkün olan bir algoritmadır. Bu model verileri parçalara bölmek yerine farklı olan veri gruplarını belirler. Tahmin ve kararlarda daha başarılı sonuç almak için kullanılabilecek her kategori için bir olasılık sağlar. K-ortalamalara göre başarı oranı da oldukça yüksektir. (Kumar, 2021)

3.3.10. K-En Yakın Komşu (KNN)

KNN en basit anlamı ile içerisinde tahmin edilecek değerin bağımsız değişkenlerinin oluşturduğu vektörün en yakın komşularının hangi sınıfta yoğun olduğu bilgisi üzerinden sınıfını tahmin etmeye dayanır. (Aslan, 2021)

K-En yakın komşu (KNN) algoritması iki temel değer üzerinden tahmin yapar;

- **Uzaklık (Distance):** Tahmin edilecek noktanın diğer noktalara uzaklığı hesaplanır.
- **K (komuşuluk sayısı):** En yakın kaç komşu üzerinden hesaplama yapılacağını söyler.

3.3.11. HMM (Hidden Markov Model).

HMM modeli 1970 lerin başında tanıtılmıştır. İlk olarak konuşma tanıma da kullanıldı. Biyolojik dizilerin analizinde ise süregelen bir başarısı vardır. Bayes ağlarının özel bir biçimi olarak kabul edilir. HMM optik karakter tanıma, el yazısı tanıma, hesaplamalı biyoloji, konuşma tanıma gibi birçok alanda başarılı bir şekilde uygulanmaktadır. İyi bir HMM gözlemlenen gerçek verilerin gerçek dünya kaynağını doğru bir şekilde modeller ve kaynağı simüle etme yeteneğine sahiptir. İstatistiksel temelinin sağlam olması, kavramsal olarak basitliği ve şekillendirilebilirliği tercih edilmesini sağladı. (Franzese, 2021)

3.3.12. TCN (temporal convolutional neural networks)

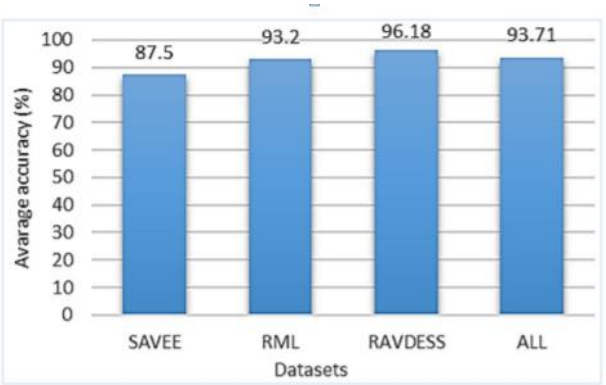
Temporal Convolutional Network'ün kısaltması olan bir TCN, aynı giriş ve çıkış uzunluklarına sahip genişletilmiş, nedensel 1D evrişim katmanlarından oluşur. (Lassig, 2021)

4. Araştırma Sonuçları ve Tartışma

4.1. Makale Değerlendirmesi

Orhan Atila ve Abdulkadir Şengür (2021), giriş konuşma sinyallerini konuşma görüntülerine dönüştürmek için spektrogram, mel-frekans kepstral katsayısı (MFCC), koleagram ve fraktal boyut yöntemleri kullanılmıştır. 3D CNN-LSTM modelinde altı adet 3D evrişim katmanı, iki toplu normalleştirme (BN) katmanı, beş Rektifiye Doğrusal Birim (ReLu) katmanı, üç adet 3D maksimum havuzlama katmanı, bir LSTM, bir düzleştirme, bir bırakma katmanı ve iki tam bağlı katman bulunmaktadır. Veri seti olarak Ryerson Duygusal Konuşma Görsel-İşitsel Veritabanı (RAVDESS), SAVEE, RML kullanılmıştır. Dikkat yönlendirmeli 3D CNN-LSTM ağına dayalı derin bir CNN mimarisi geliştirilmiştir. Geliştirilen derin CNN ağı 28 katmandan oluşmakta ve eğitimi uçtan uca gerçekleştirilmektedir. Deneysel çalışma için kodlar hem MATLAB hem de Python üzerinde çalıştırılmıştır. Sinyalden görüntüye dönüştürme Python'da (Librosa kütüphanesi) yapılırken, önerilen derin model MATLAB'da geliştirilmiş ve eğitilmiştir. Deney ve doğrulukta 10 kat çapraz doğrulama yaklaşımı göz önünde bulundurulmuştur. SAVEE veri seti için en yüksek doğruluk puanı %98,33 (Doğal) ve en düşük doğruluk puanı %75 (Kızgın) duyguları için elde edilmiştir. RML veri seti için, yüksek ve en düşük doğruluk puanları %99,17 (Kızgın) ve %87,5 (İğrenme) olarak hesaplanmıştır. RAVDESS veri seti için tüm duygular %90 doğruluk puanları üzerinden sınıflandırılmıştır. Son olarak TÜM veri setleri için %97,92 (Sakin) ve %88,44 (İğrenme) doğruluk puanlarının hesaplandığı en yüksek ve en düşük doğruluk puanları elde edilmiştir. Şekil 1 de incenelen veri setleri için elde edilen ortalama doğruluk puanları yer almaktadır.

Şekil 1. incelenen veri setleri için elde edilen ortalama doğruluk puanları



Süha GÖKALP1 ve İlhan AYDIN'ın (2021), çalışması Python 3.6 üzerinde Tensorflow = 2.4, Keras = 2.2, librosa, pandas, numpy, PIL, matplotlib kütüphaneleri kullanılarak yapılmıştır. Veri seti olarak Ryerson Duygusal Konuşma ve Şarkının Görsel-İşitsel Veritabanı (RAVDESS) ve Toronto Duygusal Konuşma Seti (TESS) kullanılmıştır. Yapay sinir ağları olarak AlexNet (CNN türü), Resnet50 (CNN türü), MobileNet (CNN türü) ve Squeezenet (DNN türü) kullanılmıştır. Tercih edilen sınıflandırıcılar karar ağaçları, destek vektör makinesi (SVM) ve Otomatik Kodlayıcıdır. Model 100 devir boyunca eğitim verileriyle eğitilmiştir. Elde edilen sonuçlar Tablo 1 de gösterilmektedir.

Tablo 1. Evrişimli Sinir ağlarında bazı modellerin karışık veri setlerindeki sonuçları

Özellikler	Evrişimli Sinir Ağı	Sınıflandırıcı	Veri Seti	Eğitim/ Test	Başarı
MFCCler,Spektogramlar	Alexnet	Otomatik Kodlayıcı	TESS	%20 Test	98%
MFCCler,Spektogramlar	Resnet50	Otomatik Kodlayıcı	TESS	%20 Test	90,60 %
MFCCler,Spektogramlar	-	Karar Ağacı	Ravdess+Tess	%33 Test	64%
MFCCler,Spektogramlar	Squeeze Net	-	Ravdess+Tess	%20 Test	81,40 %

Cevahir PARLAK ve Banu DİRİ (2014), konuşmadan duygu tanıma ile ilgili çalışmalarında genelde eğitim ve test sürecinde tek bir veritabanı üzerinde durmuşlardır. Bu çalışmalarda da doğruluk oranları oldukça yüksek değerdedir. Bu makalede ise iki veri seti kullanılmış ve biri eğitim seti diğeri test seti olarak konumlandırılmıştır. Veri seti olarak EmoSTAR ve EmoDB kullanılmıştır. Özellik çıkarımı için OpenSMILE ile beraber gelen Emobase ve Emo_large öznitelik setleri kullanılmıştır. Öznitelik seçme ve sınıflandırma ise Weka aracıyla yapılmıştır. Kullanılan sınıflandırıcılar Naive Bayes (NB) ve Sıralı Minimal Optimizasyon (SMO)' dur. Tablo 2 de EmoSTAR ve EmoDB eğitim ve test seti sonuçları yer almaktadır.

Tablo 2. EmoSTAR ve EmoDB eğitim ve test seti olarak test sonuçları

	EmoSTAR Eğitim EmoDB Test		EmoDB Eğitim EmoSTAR Test	
	NB	SMO	NB	SMO
Emobase	43,65	**52,8**	41,73	43
Emo_large	45,13	**64,3**	41,98	43,25

Özelliklerin seçilmesi sonucunda elde edilen değerler aşağıdaki tablo 3-6 arasında yer almaktadır.

Tablo 3. EmoDB'de özellik seçme (Emobase)

EmoDB	Emobase (1482)	
	NB (57,00)	SMO (87,85)
CfsSub+LFS	76,44	81,68
InfoG+Rank	69,9	88,41
Chi+Rank	69,9	88,41
PCA(145)	46,91	74,01

Tablo 4. EmoDB'de özellik seçme (Emo_large)

EmoDB	Emo_large (8190)	
	NB (70,46)	SMO (87,28)
CfsSub+LFS(102)	77,75	83,55
InfoG+Rank(6512)	70,09	87,28
Chi+Rank(6512)	69,9	87,47

Tablo 5. EmoSTAR'da özellik seçme (Emobase)

EmoSTAR	Emobase (1482)	
	NB (83,2)	SMO (95,92)
CfsSub+LFS(75)	87,27	94,65
InfoG+Rank(1236)	83,96	**96,18**
Chi+Rank(1236)	83,96	**96,18**
PCA(105)	72,26	81,67

Tablo 6. EmoSTAR'da özellik seçme (Emo_large)

EmoSTAR	Emo_large (8190)	
	NB (86,00)	SMO (96,69)
CfsSub+LFS(95)	89,05	94,14
InfoG+Rank(6755)	85,49	**97,2**
Chi+Rank(6755)	84,98	**97,2**

Sonuç olarak bakıldığında çapraz veri seti testlerinde orta seviyede bir başarı elde edilmiştir. Diğer önemli bir sonuçta öznitelik seçme algoritmalarının başarısıdır. Öznitelik seçiciler büyük öznitelik sayıları olmasına rağmen, seçim uygulanmamış özniteliklerden daha iyi sonuç elde edilmesini sağlamıştır.

Turgut ÖZSEVEN (2019), çalışmasında veri seti olarak EmoDB kullanmıştır. Veri üzerinde ön işleme yapılmış ve öznitelik kümesi akustik analiz ile elde edilmiştir. Çalışmada 149 adet öznitelik kullanılmıştır. Öznitelik çıkarımında OpenSMILE ve Praat kullanılmıştır. Öznitelik seçiminde ise temel bileşen analizi (TBA), aykırı değer algılama (ADA) ve ileri doğru seçim (IDS) kullanılmıştır. Sınıflandırıcı olarak da ÇKA, DVM ve k-NN sınıflandırıcıları kullanılmış ve sınıflandırma işlemi WEKA programı ile yapılmıştır. Tüm veri sınıflandırmasında elde edilen sonuç sonrası öznitelik kümesine z-puan normalizasyon uygulanmış ve yeni değerler hesaplanmıştır.

Normalizyon işlemi sonrası DVM sınıflandırıcı da hem gerçekleşme süresi hem de doğruluk olarak ciddi başarı artışı elde edilmiş. Ancak aynı işlem ÇKA ve k-NN sınıflandırıcı da herhangi bir değişikliğe neden olmamıştır. Öznitelik seçim yöntemleri için de karşılaştırma yapılmıştır. Bu karşılaştırma da en iyi performansı ADA elde etmiştir. (Tüm sınıflandırıcılar için) Diğer bir analizde ise filtelerin ve gürültü azaltmanın başarı üzerindeki etkisi araştırılmış. (En iyi sonuç verdiği için ADA bu analizde kullanılmıştır.) Sonucunda en yüksek başarıyı yüksek geçiren filtre elde etmiştir. Diğerlerin ise (alçak geçiren, bant geçiren ve gürültü) başarı oranı düşmüştür. Diğer bir analizde cinsiyet ve yaşın başarıdaki etkisi incelenmiş ve (ÇKA sınıfladırıcı, yüksek geçiren filtre ve ADA ile) 20-29 yaş aralığındaki bireylerin duygularını seslerine daha çok yansıtabildiği belirlenmiştir. Çalışmanın genel sonucu olarak en yüksek başarı oranı %90,3 ile z-puan normalizasyon, ÇKA sınıflandırıcı, yüksek geçiren filtre ve ADA öznitelik seçimi kombinasyonu ile olmuştur.

Serhat Hızlısoy ve Zekeriya Tüfekci (2020), çalışmalarında veri seti olarak her Türkçe duygusal müzik veri tabanını kullanılmışlardır. Öznitelik seçme yöntemi olarak korelasyon tabanlı öznitelik seçme tercih edilmiştir. Elde edilen özniteliklerin üzerinde öznitelik seçim işlemi de uygulanmıştır. Sınıflandırıcı olarak Bayes Ağları kullanılmıştır. Bu şekilde %94,35 lik bir başarı oranı elde edilmiştir. Başka bir analizde de tüm araçlardan elde edilen öznitelikler birleştirilip bunun üzerinden öznitelik seçme işlemi uygulanmış. Yine Bayes ağları ile sınıflandırma yapılmış ve sonuç olarak önceki değere göre %1,6 artış ile %95,96 başarı elde edilmiştir. Sınıflandırma başarısını artırmak için gereken öznitelikler MIRtoolbox, JAudio ve OpenSMILE araçları birarada kullanılarak elde edilmiştir. Bu çalışmanın sınıflandırma aşamasında WEKA aracı kullanılmıştır.

Yapılan müzikten duygu tanıma çalışmasında sınıflandırma çalışmasında lojistik regresyon, sıralı minimal optimizasyon, karar ağaçları ve Bayes ağları kullanılmıştır. WEKA da varsayılan parametreleri ile uygulama yapılmıştır. 10 kat çapraz doğrulama yapılmıştır. Sonuç olarak öznitelik seçim işleminin başarı performansını etkilediği ve tüm sınıflarda bu işlem ile başarının arttığı görülmüştür. En iyi sonuç ise öznitelik seçimi öncesi SMO seçim sonrası ise Bayes ağlarıdır.

Tablo 7. Araçlar birlikte kullanıldığında elde edilen sonuçlar

Araçlar	Bayes Ağları Doğruluk Sonuçları
MIRtoolbox	94,35
JAudio	91,93
OpenSMILE	94,35
MIRtoolbox+JAudio+OpenSMILE	95,96

Cevahir Parlak ve Banu Diri (2013), çalışmalarında veri seti olarak televizyon kanallarından elde edilen konuşma örnekleri ve Berlin EmoDB kullanmışlardır. Çalışma bu nedenle konuşmadan bağımsız bir modeldir. İnsan sesinden duygunun belirlenmesinde konuşma kalitesi, konuşmanın hızı, enerji, f0 gibi özellikler kullanılabilmektedir. Bu çalışma da ise f0'ın konuşma sinyali boyunca gösterdiği değişim ve sessiz duraklamalar kullanılmıştır. F0 tespiti için harmoniklerden yararlanılmıştır. Mutlu ve kızgın konuşmalarda f0 genel olarak yüksek değerdedir ve standart sapma yüksektir. Ayrıca nötür, mutlu ve kızgın konuşmalarda duraklamalar da azdır. Üzgün konuşmada ise fazladır ayrıca f0 değeri de düşüktür. Bu bilgiler ışığında sınıflandırma yapılmaya çalışılmıştır. SoundGarden programı bu çalışmada duygu çıkarımı için kullanılmıştır. Çerçeve fonksiyonu olarak dikdörtgen çerçeve kullanılmıştır.
Berlin EmoDB veri setinde kızgın %82 ve nötür %76 başarı oranları elde edilmiştir (Konuşmacı bağımsız). Önceki çalışmalardan farklı olarak bu çalışmada MFCC vektörleri kullanılmamıştır. Konuşma sinyalleri içindeki duraklamalar hesaplanmış ve bu duraklamalar nötür ve üzgün ayrımında çok belirleyici olmuştur.

Yixiong Pan ve diğerleri (2012), ise veri seti olarak Berlin duygu veritabanı (5 sınıf var 3 ü kullanılmış) ve SJTU Chinese duygu veritabanı (kendileri oluşturmuş) kullanılmışlardır. Özellik çıkarma işlemi sonrasında enerji, perde, formant, LPCC, MFCC ve MEDC özelliklerini elde etmişler ve onların farklı kombinasyonunu kullanarak tanıma oranı karşılaştırılmışlardır. Sınıflandırmada ise DVM (Support Vector Machine) metodu uygulanmıştır. Modellerin çapraz doğrulamasını yapmak ve sonuçları analiz etmek için Matlab'da libsvm aracı kullanılmıştır. Eğitim alt kümesi %90 ve test alt kümesi olarak %10'dur. Berlin duygu veritabanında 5 farklı model ve SJTU Chinese duygu veritabanında 2 farklı model test edilmiştir. Elde edilen sonuçlar Tablo 8 ve Tablo 9 da gösterilmiştir.

Tablo 8. Berlin duygu veritabanı

Eğitim Modeli	Özelliklerin Kombinasyonu	Çapraz Doğruluk Oranı	Tanıma Oranı
Model 1	Enerji+Perdeleme	66,6667%	33,3333%
Model 2	MFCC+MEDC	90,1538%	86,6667%
Model 3	MFCC+MEDC+LPCC	72,5275%	86,6667%
Model 4	MFCC+MEDC+Enerji	95,0549%	91,3043%
Model 5	MFCC+MEDC+Enerji+Perdeleme	94,5055%	90%

Tablo 9. SJTU Chinese duygu veritabanı

Eğitim Modeli	Özelliklerin Kombinasyonu	Çapraz Doğruluk Oranı	Tanıma Oranı
Model 2	MFCC+MEDC	88,6168%	80,4763%
Model 4	MFCC+MEDC+Enerji	95,1852%	95,0874%

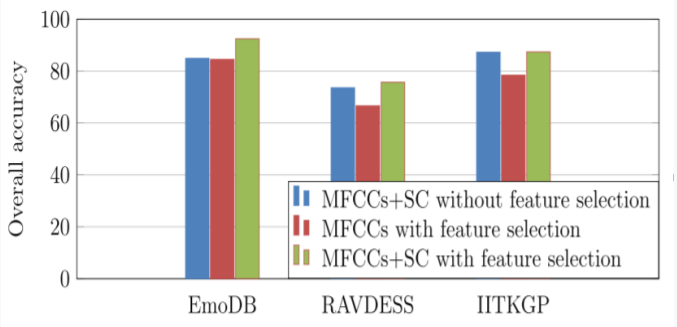
Berlin Veritabanında en iyi özellik kombinasyonu, gerçek zamanlı olmayan tanıma için çapraz doğrulama oranının %95'e kadar çıkabildiği MFCC+MEDC+Enerji ile elde edilmiştir. SJTU veritabanında da MFCC+MEDC+Enerji'nin özellik kombinasyonu burada da en iyi performans göstermiştir. Konuşmanın sadece spektrum özelliklerini kullanan sistemin duygu tanıma oranı, konuşmanın sadece prozodik özelliklerini kullanan sisteme göre biraz daha yüksektir. Hem spektral hem de prozodik özellikleri kullanan sistem, yalnızca spektrum veya prosodik özellikleri kullanan sistemden daha iyidir.

Anjali Bhavan ve diğerleri (2019), çalışmalarında veri seti olarak Berlin EmoDB, RAVDESS, IITKGP-SEHSC kullanmışlardır. Özellik çıkarımı sonrası MFCC, Delta and Delta–Delta MFCCs, Spectral Centroids ler elde edilmiştir. Boruta kullanılarak özellik seçim işlemi yapılmış. Veriler için model olarak bir torbalama topluluğu (bagging ensemble) yöntemi uygulamışlardır. Önyükleme toplamanın kısaltması olan torbalama, topluluğun çeşitli temel tahmin edicileriyle beslenen eğitim örneklerinden oluşur. Temel tahminci Gauss çekirdeğine sahip bir destek vektör makinesidir. 20 destek vektör makinesinden oluşan torbalama topluluğu kullanılmıştır. Eğitim setinden değiştirme ile 20 örnek set çekilmiş ve ardından her bir temel tahmin edici üzerinde paralel bir şekilde eğitilmiştir. Elde edilen sonuçlar daha sonra nihai tahminleri vermek için ortalama alma kullanılarak toplanmıştır. Öznitelik vektörü elde edildikten sonra veriler önce (0, 1) aralığına ölçeklendi ve ardından 90:10 oranında eğitim ve test verilerine bölünmüş. Tüm prosedür makine öğrenimi algoritmaları ve kaynakları için Scikit-learn paketi kullanılarak gerçekleştirilmiş. Veri kümeleri üzerinde eğitim ve değerlendirme, çapraz doğrulama metriği olarak seçilen doğrulukla 10 katlı çapraz doğrulama kullanılarak yapılmış.

Tablo 10. EmoDB, RAVDESS ve IITKGP-SEHSC veritabanlarında MFCC'ler ve spektral centroid özellikleri ile eğitim ve test doğruluklarını göstermektedir.

VeriSeti	Eğitim Doğruluğu	Bekleme Seti Doğruluğu
EmoDB	96,25%	92,45%
RAVDESS	79,85%	75,69%
IITKGP+SEHSC	85,72%	84,11%

Şekil 2. Tanıma Oranlarının Karşılaştırılması

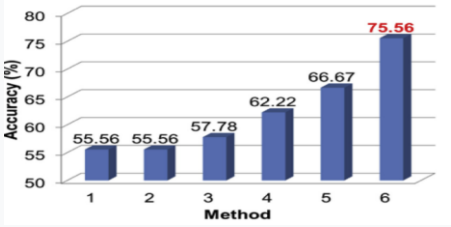


Sonuç olarak sınıflandırma yönteminin etkisini anlamak için duygu tanıma sisteminde EmoDB ve RAVDESS veritabanları için sırasıyla %86,69 ve %72,91 genel doğruluk sağlayan basit bir destek vektör makinesi sınıflandırıcısı kullanılarak değerlendirilmiştir. Sınıflandırma için destek vektör makinelerinden oluşan torbalı bir grubun kullanılmasıyla, bu doğruluk kabaca %5 oranında daha da artırılmıştır.

Gökhan POLAT ve Halis ALTUN (2008), çalışmalarında veri seti olarak Berlin Veri Setini kullanmışlardır. Öznitelik çıkarım işlemi sonrası 38 öznitelik vektörü elde edilmiştir. Sınıflandırma işleminde çok katmanlı bir YSA kullanılmıştır. Uygulama Matlab üzerindeki Neural Network Toobox kullanılarak geliştirilmiştir. Oluşturulan YSA yapısı 38 girişli, 4 çıkışlı ve farklı saklı katman sayısına sahiptir. Bu farklı gizli katman sayılarından en iyi başarı 9 gizli katmanlı YSA ile elde edilmiştir. Eğitim setinde %94,4 ve test setinde %81,65 başarı oranı elde edilmiştir. En iyi duygu üzüntüdür. (%90) (Test veri setinde) Mutluluk ise %58 ile en düşük başarı oranına sahip olmuştur.

Kun-Yi Huang ve diğerleri (2018), veri seti olarak CHI-MEI duygu durum bozukluğu veritabanı, MHMC (Multimedya İnsan-Makine İletişimi duygu veritabanı) kullanmışlardır. Özellik çıkarımında OpenSMILE kullanılmıştır. Çıkarılan özellikler sıfır geçiş oranı (zero-crossing rate), kök kare ortalama(root-mean-square), temel frekans (fundamental frequency), Harmonik-Gürültü-Oranı (Harmonic-Noise-Ratio) ve MFCC dir. Her konuşma yanıtı için EP'yi (Emotion Profile-Duygu Profili) oluşturmak üzere bir dikkat mekanizmasına sahip CNN kullanılmış. Bu çalışmada, altı video tabanlı yanıttan oluşan tüm görüşme boyunca ortaya çıkan yanıtların EP'lerinin seyrini modellemek için LSTM ağları kullanılmış. Çünkü yanıtın tamamı bazı alakasız bilgiler içerebileceğinden, tüm yanıtlar arasında video tabanlı önemli yanıtları vurgulamak için bir dikkat mekanizması kullanılmıştır.

Şekil 3. Farklı Methodların Karşılaştırılması



1) OpenSMILE + SVM
2) OpenSMILE + LSTM.
3) OpenSMILE + CNN
4) HSC + EP + SVM (termed SVM-based): Bu yöntem, 32 boyutlu ham özniteliklerden veri uyarlaması için HSC algoritmasını kullanmış. MHMC duygu veri tabanının uyarlanmış 32 boyutlu özelliği, CNN tabanlı duygu üretme modelini eğitmek için kullanılmış. EP'ler CNN kullanılarak CHI-MEI duygudurum bozukluğu veri tabanından elde edilmiş. Son olarak, duygudurum bozukluğu tespiti için bir SVM oluşturuldu ve kullanılmış)
5) HSC + EP + CNN
6) HSC + EP + LSTM

Sonuç olarak önerilen yöntem, %75,56'lık bir algılama doğruluğu sağlayarak, genel olarak kullanılan SVM tabanlı (%62,22) ve CNN tabanlı (%66,67) sınıflandırıcılardan daha iyi performans göstermiştir.

Haytham M. Fayek ve diğerleri (2017), veri seti olarak IEMOCAP kullanmışlardır. Bu yazıda ileri beslemeli mimariler için kullanılan aktivasyon fonksiyonu Rektifiye Doğrusal Birim (ReLU) kullanılmıştır. Önerilen model derin, çok katmanlı bir sinir ağıdır. Konuşma işleme ve analiz için Kaldi tools kullanılmış. Sinir ağları ve eğitim algoritmaları Matlab ve C'de uygulanmış. İleri beslemeli mimari olarak her biri BatchNorm, ReLU ve aralarına serpiştirilmiş bırakma katmanları ve bir softmax çıktı katmanı ile 1024 tam bağlantılı üniteden oluşan 5 gizli katmana sahip bir DNN kullanılmıştır. Yinelenen mimari olarak da her bir gizli katmanda 256 birim ve aralarına serpiştirilmiş bırakma ve bir softmax çıktı katmanına sahip 2 katmanlı bir LSTM-RNN kullanılmıştır.
Sonuç olarak önerilen SER sistemini, ileri beslemeli ve tekrarlayan sinir ağı mimarilerini ve bunların varyantlarını deneysel olarak araştırmak için kullanmışlar.

Tablo 11. Her mimariden en iyi modeli ve bunların ilgili doğruluğunu ve aynı veri alt kümeleri ve deney koşulları altında eğitilmiş ve değerlendirilmiş UAR'yi listeler

Model	Test Doğruluğu(%)	Test UAR(%)
FC(1024)x5-Softmax(1024)	62,55	58,78
Conv(16x10x10)-Conv(32x10x10)-FC(716)x2-Softmax(716)	**64,78**	**60,89**
LSTM-RNN(256)x2-Softmax(256)	61,71	58,05

Tablo 11, konuşmadaki statik bileşenin (ConvNet ve DNN), SER için dinamik bileşenden (LSTM-RNN) daha ayırt edici olduğunu göstermektedir. Çalışmada ConvNet en iyi doğruluğu ve UAR'yi, ardından DNN'yi ve ardından LSTM-RNN'yi vermiştir.

Dias Issa ve diğerleri. (2020), veri seti olarak RAVDESS, EMO-DB ve IEMOCAP kullanmışlardır. Özellik çıkarma işleminde Librosa ses kitaplığını kullanılmış. Beş farklı spektral temsil kullanılmıştır. Bunlar Mel-frekans Kepstrum Katsayıları (MFCCs), Mel ölçekli spektrogram (Mel-scaled spectrogram), Kromogram (Chromagram), Spektral kontrast özelliği (Spectral contrast feature) ve Tonnetz sunumu (Tonnetz representation) dır. Bir ses dosyasından çıkarılan özelliklere dayalı olarak duyguların sınıflandırılması için evrişimli sinir ağını (CNN) kullanılmıştır. Temel model, bırakma, toplu normalleştirme ve etkinleştirme katmanlarıyla birleştirilmiş 1D CNN içerir. Aktivasyon fonksiyonu olarak ReLU kullanılmıştır. RAVDESS modelde %71,61 başarı oranı elde edilmiştir. Model A'da %82,86, Model B'de %96,34, Model'C de %82,4, Model'D de %84,76 ve Model'E de C ile D den en yüksek olan değer alınmıştır. (Veri Seti EMO-DB). IEMOCAP modelde ise %64,30 başarı elde edilmiştir.

Zengwei Yao ve diğerleri (2020), veri seti olarak IEMOCAP kullanmışlardır. Sınıflandırmada da derin sinir ağları (DNN), evrişimli sinir ağları (CNN), ve tekrarlayan sinir ağlarını (RNN) birlikte kullanılmıştır. HSF'lerden üst düzey temsil elde etmek için DNN'yi, spektrogramlardan zaman ve frekans alanını modellemek için CNN'yi ve LLD'lerden uzun bağlamlı bilgileri öğrenmek için RNN'yi kullanılmıştır. Farklı girdi türlerini kullanarak üç sınıflandırıcıyı ayrı ayrı eğitilmiş ve daha sonra nihai sonuca ulaşmak için karar düzeyinde füzyon stratejisini uygulanmıştır. Özellik çıkarımında OpenSMILE ve librosa araçları kullanılmıştır. Aktivasyon fonksiyonu olarak tanh ve softmax kullanılmıştır. Burada multi-task bir öğrenme yapısı tasarlanmıştır. Birinci analizde MS-CNN and LLD-RNN

sınıflandırıcıları havuzlama modellerine göre karşılaştırılmış ve en iyi sonuç her iki modelde de ağırlıklı havuzlama da elde edilmiştir. (MS-CNN %48,3 (UA) ve LLD-RNN %54,1 (UA)) İkinci analizde çok görevli öğrenme çerçeveleri karşılaştırılmış ve tüm sınıflarda Multi-task (V&A) yapısı en iyi sonucu vermiş. (MS-CNN %50,1 (UA), HSF-DNN %55,6 (UA) ve LLD-RNN %55,6 (UA)) Üçüncü analizde ise oluşturulan fusion model, Baseline model ve diğer 3 sınıflandırma karşılaştırılmıştır. Spesifik olarak, HSF-DNN ve LLD-RNN sınıflandırıcıları kızgın, mutlu ve üzgün durumları tanımada güçlüyken, MS-CNN sınıflandırıcısı nötr durumu tanımada güçlü olarak belirlenmiş. Önerilen füzyon yöntem her bir bireysel sınıflandırıcınınkinden önemli ölçüde daha yüksek %57,1'lik WA ve %58,3'lük UA'ya ulaşmış.

Hadhami Aouani ve Yassine Ben Ayed (2020), veri seti olarak RML duygu veritabanı kullanmışlardır. Özellik çıkarımında 39 MFCC, Sıfır geçiş oranı (ZCR), Teager Enerji operatörü (TEO) and Harmonik gürültü oranı (HNR) kullanılmıştır. Özellik seçme işlemi otomatik-kodlayıcı (AE) ile yapılmıştır. Burada 2 tür otomatik-kodlayıcı (AE) kullanılmıştır (Temel AE ve yığılmış otomatik-kodlayıcı). Sınıflandırma işleminde ise destek vektör makinesi (SVM) seçilmiştir. SVM Linear, Polynomial, RBF çekirdek işleviyle kullanılmıştır. Bu üç özellik içinde en iyi sonuç Kernel RBF de elde edilmiştir. (%65,43) Otomatik-kodlayıcı ile özellik seçimi sonrası tüm özelliklerin dahil edildiği analize (%65,43) göre daha yüksek bir başarı elde edilmiştir. Ayrıca Temel otomatik-kodlayıcının performansı (%74,07) yığılmış otomatik-kodlayıcının performansından (%72,83) daha iyidir. Çalışma ayrıca otomatik kodlayıcı ile boyut küçültme uygulamasının tanımlama oranını iyileştirdiğini göstermiştir.

Shadi Langari ve diğerleri (2020), veri seti olarak EMO-DB, SAVEE, PDREC kullanmışlardır. Çalışmanın Ön işleme bölümü, pre-emphasis block, windowing ve frame blocking içerir. Bu çalışmada, klasik Ayrık Fourier dönüşümü (DFT) ile özvektörleri ve özdeğerlerinin genişletilmesiyle elde edilen Ayrık Kesirli Fourier Dönüşümü (DFrFT) temelli uyarlanabilir bir zaman-frekans öznitelik çıkarma yöntemi kullanılmıştır. Öznitelik seçimi için Genetik Algoritma ve Cuckoo Search'ten oluşan hibrit bir yöntem kullanılmış. Sınıflandırıcı olarak 2D CNN kullanılmıştır. İlk aşama, bir konuşma spektrogramı için seyrek bir otomatik kodlayıcı kullanarak yerel değişmez özellikleri öğrenmektir ve sonraki aşamada tanımayı iyileştirmek için PCA kullanılarak ayırt edici özellikler çıkarılır. Sınıflandırıcıyı eğitme ve test etmede 10 kat çapraz doğrulama kullanılmıştır. Önerilen yöntem her bir veri setinde uygulandı ve EMO-DB için (%97,57(WA), %97,21(UA)), SAVEE için (%80(WA), %80(UA)) ve PDREC için (%91,46(WA), %87,31(UA)) doğruluk oranlarına ulaştı. Bu çalışmada, SER'yi iyileştirmek için uyarlanabilir zaman-frekans katsayılarına dayalı yeni bir öznitelik çıkarma yöntemi öneriliyor. Çalışmanın ana katkısı, Kesirli Fourier Dönüşümüne dayalı Uyarlanabilir Zaman-Frekans öznitelikleri adı verilen yeni öznitelikleri çıkarmak ve bunları Kepstral öznitelikleri ile birleştirmektir.

Kunxia Wang ve diğerleri (2020), veri seti olarak EMODB ve EESDB kullanmışlardır. Bu çalışmada konuşmacıdan bağımsız duygu tanıma için konuşma sinyallerinden dalgacık paket analizini kullanarak duygu özellikleri çıkarılmıştır. Duyguyu tanımak için dalgacık paket katsayısı (WPC) özelliklerine dayanan yeni bir yöntem önerilmiştir. Özellik seçiminde Sıralı Kayan İleri Arama (SFFS) kullanılmıştır. Sınıflandırma için de SVM kullanılmıştır. Sınıflandırıcı olarak Linear SVM (LSVM) ve radyal temel işlevli (RBF) çekirdekli SVM seçilmiştir. Özellik seçiminde MFCC özelliklerini ve WPC özelliklerini çıkarılmış. Dalgacık paketi, Daubechies dalgacık filtresi ve Gabor filtresi gibi birçok algoritmayı destekler. Bu çalışmada Daubechies dalgacıklarının aileleri seçilmiştir. Seviye 4 dalgacık paket ayrıştırması 16 katsayı üretir ve seviye 5 dalgacık paket ayrıştırması 32 katsayı üretir. Seviye 5'teki dalgacık paket katsayısı, tanıma performansını ortalama %6,7 artırır. Bu da Seviye 5'teki dalgacık paket katsayısı özelliklerinin, duyguyu tanımak için daha iyi doğruluk elde ettiğini gösterir. Tekli dalgacık katsayısı ile elde edilen başarı yüzdeleri EMODB (%57,4) ve EESDB (%55,1) dir. Çoklu dalgacık katsayısı ile ise EMODB (%64,5) ve EESDB (%62,8) dir. Çoklu WPC özellikleri ile performans, tek WPC ile olandan daha iyi olsa da çok fazla özellik ile iyi sonuçlar almak zordur. Diğer bir analizde MFCC, WPC8 ve WPC+SFFS özellikleri karşılaştırılmış ve en iyi tanıma performansı iki veri setinde de WPC+SFFS metodunda ve RSVM ile sağlanmıştır. Ortalama tanınma oranları sırasıyla %79,5 (EMODB) ve %60,7 (EESDB)'dir. Önerilen WPC özellikleri MFCC özellikleri ile karşılaştırıldığında, EMODB ve EESDB veritabanlarında konuşmacıdan bağımsız duygu tanımayı yaklaşık %1,8 ve %5,5 oranında iyileştirebilir. Tanıma oranları, iki veritabanında SFFS özellik seçimi kullanılarak %16,2 ve %7,2 oranında daha da artırılabilir.

Ning Jia ve Chunjun Zheng (2021), çalışmalarında SER görevine iki seviyeli bir ayırt edici model uygulayarak sorunu ele almaktadır. Birinci düzey, bireysel bir konuşmacıyı, konuşmacının özelliklerine göre belirli bir konuşmacı grubuna yerleştirir. İkinci seviye, dalga alanı dinamiği modelini ve bir çift kanallı genel SER modelini kullanarak her bir konuşmacı grubu için kişiselleştirilmiş bir SER modeli oluşturur. Veri seti olarak SER ve IEMOCAP kullanılmıştır. Giriş özellikleri olarak Mel-frekans cepstral katsayıları (MFCC'ler) uygulanmış. Konuşma sinyalinin karmaşıklığı, bağıl spektral (RASTA) filtreleme uygulaması yoluyla arka plan gürültüsünü ortadan kaldırarak azaltılmış. Mevcut çalışma, bir konuşma dalga biçimi bir zaman serisi olduğundan ve bu nedenle bağlamlar arasındaki korelasyonları içerdiğinden bir BiMLSTM modeli uygulanmış. BiMLSTM modeli, zıt dönüşümlere sahip iki tekrarlayan katmanı benimser. Mevcut çalışmada, daha yüksek konuşmacı sınıflandırma doğruluğu için kişiselleştirilmiş bağlam bilgisinden yararlanmak için üç katmanlı bir BiMLSTM modeli oluşturmuştur. BiMLSTM modelinin çıktısı, tam bağlı bir katmana ve son olarak bir softmax katmanına geçer. Her iki veri seti için yapılan deneylerde beşli çapraz doğrulama yöntemi kullanılmış ve verilerin %80'i model eğitimi için, geri kalan veriler ise doğrulama ve test için kullanılmıştır. Birinci analizde Tablo 12 da konuşmacı sınıflandırma modeli için elde edilen sonuçlar yer almaktadır.

Tablo 12. Farklı konuşmacı sınıflandırma modelleri (1.Seviye) için sonuçlar

Model	Custom Corpus EER	Custom Corpus Min_DCF
Baseline: MFCC+ tek katmanlı LSTM	16,60	0,77
Model 1: MFCC+tek katmanlı LSTM	16,10	0,76
Model 2: RASTA-MFCC+tek katmanlı LSTM	15,60	0,75
Model 3: RASTA-MFCC+üç katmanlı LSTM	14,80	0,72
Model 4: önerilen yöntem (RASTA-MFCC+üç katmanlı BiMLSTM)	14,50	0,70

Önerilen model bu veri seti için en iyi performansı sağladı. İkinci seviyedeki SER modelin sonuçları ise Tablo 13 de gösterilmiştir.

Tablo 13. İkinci seviye SER Modeli

Model	Custom Corpus UA	Custom Corpus wA	IEMOCAP UA	IEMOCAP WA
Baseline	0,56	0,55	0,51	0,50
Model 1	0,58	0,56	0,56	0,55
Model 2	0,68	0,68	0,60	0,61
Model 3	0,74	0,73	0,64	0,65
Model 4	0,75	0,74	0,65	0,63
Model 5	**0,77**	**0,78**	**0,69**	**0,70**

Baseline: kişiselleştirilmiş model yok, genel modelin 1. kanalı
Model 1: kişiselleştirilmiş model yok, genel modelin 2. kanalı.
Model 2: Kişiselleştirilmiş model yok, iki kanallı genel model.
Model 3: kişiselleştirilmiş model + genel modelin 1. kanalı.
Model 4: kişiselleştirilmiş model + genel modelin 2. kanalı.
Model 5: önerilen yöntem (yani, kişiselleştirilmiş model + iki kanallı genel model).

Önerilen yöntemin her iki veri seti için de en iyi WA ve UA değerleri ile en iyi performansı elde ettiğini görebiliriz. Önerilen yöntemde kızgın ve mutlu sınıflar, üzgün ve tarafsız sınıflardan daha doğru bir şekilde tanınır. Kişiselleştirilmiş SER model analizinde ise Tablo 14'teki sonuçlar elde edilmiştir.

Tablo 14. Kişiselleştirilmiş SER Model Karşılaştırması

Model	UMASK UA	UMASK WA	MASK UA	MASK WA
Baseline	0,69	0,68	0,67	0,67
Model 1	0,70	0,69	0,65	0,66
Model 2	0,71	0,72	0,70	0,69
Model 3	0,71	0,72	0,68	0,67
Model 4	**0,74**	**0,73**	**0,73**	**0,72**

Baseline: kişiselleştirilmiş model yok, iki kanallı genel model
Model 1: Sadece havada sabit bir dalga hızına sahip kişiselleştirilmiş model kullanılarak.
Model 2: Yalnızca uyarlanabilir dalga hızına sahip kişiselleştirilmiş model kullanılarak.
Model 3: Havada sabit bir dalga hızı ile önerilen model.
Model 4: Uyarlanabilir dalga hızına sahip önerilen model.

Sonuçlar, konuşmacının maske takıp takıp takmadığına (yani UMASK) göre sıralanır. Temel model en düşük SER doğruluğunu elde etmiştir çünkü genel model çapraz ortam koşullarında mükemmel şekilde uygulanamıyor. Ayrıca, Model 1 ve 3'te havada sabit bir dalga hızının benimsenmesinin, maske takmayan konuşmacılar için iyi SER doğruluğu ile sonuçlandığını, SER sonuçlarının ise maske takan konuşmacılar için önemli ölçüde daha az doğru olduğunu not ediyoruz. Model 2 ve 4'te uyarlanabilir bir dalga hızının benimsenmesi hem maskeli hem de maskesiz konuşmacılar için neredeyse tek tip SER doğruluğu ile sonuçlanır. Sonuç olarak IEMOCAP veri setinde önerilen model, tanıma doğruluğunu %7 oranında iyileştirmektedir.

Yousef Pourebrahim ve diğerleri (2021), önerdikleri yöntem ile etiketsiz numunelerde bulunan bilgileri çıkararak ve etiketli numunelerdeki bilgilerle birleştirerek sınıflandırma işlemini gerçekleştirmiştir. SSAE yöntemi, bu bildiride otomatik kodlayıcı tabanlı bir yapı önerilmiştir. Önerilen yöntem, ayırt edici olan girdi özelliklerinin yeni bir temsilini çıkarmaya yardımcı olabilecek paralel paylaşılan kodlayıcılarla oluşturulmuştur. Yarı denetimli bir yöntem olarak, özellikle etiketli örneklerin yeterli olmadığı dillerde, etiketli örneklerin kullanımını azaltmak için otomatik kodlayıcılara dayalı yeni bir yapı önerilmiştir. SSPSE modelinde The supervised path, girdi örneklerinin sınıflandırılmasına yönelik bir softmax katmanından oluşur. The unsupervised path ise, girişleri yeniden yapılandırmak için bir kod çözücüden oluşur. Veri seti olarak IEMOCAP, FAUAEC, PESD, EmoDB ve RAVDESS şarkıları kullanılmıştır. Karşılaştırmalar, az sayıda etiketlenmiş eğitim örneğinden yararlanan önerilen yöntemin (SSPSE) denetimli öğrenme yöntemlerine meydan okuyabildiğini ve veri dağılımındaki değişikliklere karşı dayanıklı olduğunu göstermiştir.

Tablo 15. IEMOCAP/EmoDB/AEC veri setlerindeki etiketlenmemiş veriler ve SS-AE semi-supervised öğrenme metodu ile geliştirilen modelin karşılaştırılması

Öğrenme Metodu	Etiketsiz Veri			AEC test setindeki etiketli veri sayısı			
	AEC	EmoDB		100	200	300	400
SS-AE	+			40,1	43,1	43,1	43,6
	+	+		38,9	42,9	43,3	42,9
		+		39,7	41,6	43,3	42,9
			Ortalama:	**39,59**	**42,53**	**43,23**	**43,13**
SSPSE	+			40,7	43,6	44,5	45,3
	+	+		40,4	42,9	44,3	44,8
		+		40,3	42,7	44,0	44,6
			Ortalama:	**40,46**	**43,06**	**44,26**	**44,9**

Tablo 15'e göre SSPSE (geliştirilen model) SSAE den daha iyi performans göstermiştir.

Tablo 16. IEMOCAP/EmoDB/AEC veri setlerindeki etiketlenmemiş veriler ve SS-AE semi-supervised öğrenme metodu ile geliştirilen modelin sonuçların karşılaştırılması

Öğrenme Metodu	Etiketsiz Veri				IEMOCAP test setindeki etiketli veri sayısı			
	IEMOCAP	EmoDB	AEC		300	600	1200	2400
SSGAN	+				51,9	55,3	57,8	59,3
	+	+			51,8	55,4	57,5	59,0
	+		+		52,4	55,6	57,2	58,6
	+	+	+		52,1	55,3	57,3	58,6
		+			51,6	55,3	57,6	58,8
		+	+		52,0	55,3	57,5	58,4
			+		52,0	55,5	57,8	58,8
				Ortalama	**51,9**	**55,3**	**57,5**	**58,7**
SSPSE	+				52,7	56,2	58,6	62,2
	+	+			52,5	56,1	57,9	61,7
	+		+		52,1	55,8	58,1	61,4
	+	+	+		51,4	55,4	58,1	61,1
		+			51,9	55,5	58,3	61,5
		+	+		51,7	55,9	58,3	61,7
			+		51,3	56,2	58,5	62,1
				Ortalama	**51,9**	**55,8**	**58,2**	**61,6**

Sonuç olarak, bu makale, etiketlenmemiş ve etiketlenmiş veri bilgilerini aynı anda kullanabilmek için konuşma duygularını tanımlamak için otomatik kodlayıcılara dayalı yarı denetimli bir öğrenme modeli sağlamaya çalışmıştır.

J. Ancilin ve A. Milton. (2021) Bu yazıda, Mel frekans kepstral katsayısının çıkarılmasında, enerji spektrumu yerine büyüklük spektrumu kullanılarak ve ayrık kosinüs dönüşümü hariç tutularak ve Mel Frekans Büyüklük Katsayısı'nın çıkarılmasında iki değişiklik yapılmıştır. Mel frekans büyüklük katsayısı, Mel frekansı kepstral katsayısı, log frekansı güç katsayısı (log

frequency power coefficient) ve lineer tahmin kepstal katsayısı (linear prediction cepstral coefficient) test edilmiştir. Veri seti olarak Berlin, Ravdess, Savee, EMOVO, eNTERFACE ve Urdu kullanılmıştır. Sınıflandırma ise çok sınıflı SVM kullanılmıştır. Bağımsız bir özellik olarak Mel frekans büyüklük katsayısı, duyguyu Berlin için %81,50, Ravdess için %64,31, Savee için %75,63, EMOVO için %73,30, eNTERFACE için %56,41 ve Urdu veritabanları için %95,25 doğrulukla tanır. MFMC, iki adım dışında Mel frekansı cepstral katsayısı ile aynı şekilde çıkarılır: ilk olarak, büyüklük karesi yerine hızlı Fourier dönüşümünün büyüklüğü kullanılır. İkinci olarak, korelasyon amacıyla MFCC ekstraksiyonunda kullanılan ayrık kosinüs dönüşümü hariç tutulur. Sınıflandırma deneyleri 10 katlı çapraz doğrulama şeması ile gerçekleştirilmiştir. Sonuçlar Tablo 17 de gösterilmiştir.

Tablo 17. Farklı Katsayılarına Göre LPCC, LFPC, MFCC ve MFMC nin doğruluklarının karşılaştırılması

Özellik		LPCC	LFPC	MFCC	MFMC
Veritabanı	#Katsayı	Doğruluk Yüzdesi			
Berlin	12	69,91	72,15	73,08	80,37
	24	70,09	73,27	72,34	80,19
	30	69,72	74,21	73,64	81,50
Ravdess	12	46,53	45,35	30,83	58,89
	24	47,22	49,44	32,01	62,85
	30	47,08	50,49	31,25	64,31
Savee	12	58,54	57,29	58,33	71,67
	24	60,63	57,71	66,67	72,50
	30	59,58	60,64	69,17	75,63
EMOVO	12	53,06	61,22	53,23	64,12
	24	65,31	64,80	52,04	70,92
	30	67,86	64,12	53,40	73,30
eNTERFACE	12	41,72	44,06	50,58	53,22
	24	43,90	44,06	49,18	54,55
	30	44,76	43,82	48,33	56,41
Urdu	12	88,50	90,75	88,25	90,75
	24	92,25	93,00	87,50	93,25
	30	89,25	92,25	85,75	95,25

MFMC özelliği, altı veritabanının tümünün duygularını tanımada diğer üç özellikten daha iyi performans göstermiş.

Tablo 18. Duygu Durumuna Göre En İyi Özellik Çıkarıcının Listesi

Veritabanı	**Berlin**	**Ravdess**	**Savee**	**EMOVO**	**eNTERFACE**	**Urdu**
Sinirli	MFMC-88,98	MFMC-70,83	MFMC-80,00	MFMC,LPCC-78,57	MFMC-73,95	LPCC,LFPC, MFMC-97
Can Sıkıntısı	MFMC-74,07	-	-	-	-	-
Şaşırma	-	MFMC-60,94	MFMC-60,00	MFMC-66,67	MFMC-51,16	-
İğrenme	MFMC-84,78	MFMC-65,1	MFMC-75,00	MPMC-73,81	MFMC-42,79	-
Korkma	MFMC-78,26	MFMC-59,89	MFMC-63,33	LPCC-67,86	MFMC-47,44	-
Mutlu	MFMC,LFPC-71,83	MFMC-58,33	MFMC-61,67	MFMC-67,86	MFMC-59,91	MFMC-92
Üzgün	MFMC,MFCC-93,55	MFMC-57,81	MFMC-80,00	MFMC-84,52	MFMC-63,26	LFPC-96
Doğal	MFMC,LPCC-74,68	MFCC-80,21	MFMC-92,50	LPCC-77,38	-	MFMC-97

Deneysel sonuçlardan açıkça görülmektedir ki, MFMC sadece altı veritabanının tümünün duygularını en yüksek doğrulukla tanımakla kalmaz, aynı zamanda MFCC, LFPC ve LPCC'den daha yüksek tanıma oranlarıyla veritabanlarının bireysel duygularını da tanır.

Ziping ZHAO ve diğerleri (2021), modeli (SATN) ilk önce öz-dikkat tabanlı bir kodlayıcı-kod çözücü ile eğitir. Konuşma tanıma ağındaki gizli katmanların parametrelerini öğrendikten sonra model, parametrelerini dondurur. Bir sonraki adımda, model bir konuşma tanıma görevi aracılığıyla dikkat ağırlıklarını eğitir ve ardından bu ağırlıkları konuşma duygu tanıma sistemine besler. Aktivasyon fonksiyonu olarak tanh kullanılmıştır. Veri seti olarak IEMOCAP kullanılmıştır. Tüm modeller PyTorch1 framework kullanılarak uygulanmış. Deneyde, her deneydeki modeller 100 devir için eğitilmiş. 5 katmanlı 1D evrişimli modüllere sahip bir TCN kullanıldı ve konuşma tanıma ön eğitim görevinde kodlayıcı için 1, 2, 4, 8 ve 16 genişletme faktörleri kullanıldı. Kod çözücü için 256 tek bellekli hücre bloğu içeren uzun bir kısa süreli bellek (LSTM) kullanıldı. Karşılaştırma içinde, TCN'nin yerine BLSTM ağını kullandık. Karşılaştırma deneyinde, 128 ileri gizli nodülü ve 128 geriye gizlenmiş nodülü olan tek katmanlı bir BLSTM ağına girdi olarak 3D log-mels spektrogramını beslenmiş.

Tablo 19. Geliştirilen model ile diğer seçilen modellerin IEMOCAP veri setinde karşılaştırılması

	Geliştirme		Test	
Method	WA[%]	UA[%]	WA[%]	UA[%]
Önceki raporlanan yöntemler				
DNN+ELM(44,45)	-	-	57,9	52,1
RNN+ELM(45)	-	-	62,9	63,9
Attention+RNN(3)	-	-	63,5	58,8
GMM+HMM(46)	-	-	55	60,3
Önerilen self-attention modeller				
BLSTM+soft attention	63,8	62,9	59,6	59,7
BLSTM+self attention	63,7	64,5	60	60,5
BLSTM+soft attention w/AT	65,6	65,6	62,1	62,2
BLSTM+self attention w/AT	66,9	68,1	63,8	64,5
TCN+soft attention	65,5	66,6	61,8	62,5
TCN+self attention	67,5	67,4	63,7	64,2
TCN+soft attention w/AT	67,2	67,9	63,4	64,4
TCN+self attention w/AT	**68,6**	**69,5**	65	66,1

Test setindeki en iyi WA (%65,0) ve UA (%66,1) ve geliştirme setindeki en iyi WA (%68,6) ve UA'nın (%69,5) geliştirilen modelde (self-attention-based TCN model) elde edildiği görülebilir.

Mesut Durukal, A. Köksal Hocaoğlu (2015), çalışmalarında veri seti olarak Berlin Duygu Tabanlı Konuşma kullanmışlardır. Öznitelik olarak MFCC çıkarılmış ve sınıflandırmada ise SVM kullanılmıştır. Veri setinin %5 i test kalanı eğitime ayrılmıştır. 20 kez döngü ile tüm veri seti test eğitim olarak kullanılmıştır. En son olarak bu 20 döngüden elde edilen değerlerin ortalaması alınmıştır. Çıkarılan öznitelikler üzerinde farklı kombinasyonlar denenip başarı oranları karşılaştırılmıştır. İlk analizde MFCC sayısı 20'den az olduğunda performans düşmekte ve 40 fazla olduğunda da ekstra bir artış gözlenmemektedir. MFCC vektörleri öznitelik seçimi olmadan sınıflandırıldığında %47 başarı elde edilmiştir. Doğrudan kullanım yerine öznitelik seçim işlemi yapıldığında (ortalama, maksimum, minimum, standart sapma, genlik) başarı oranı %64 olmuştur. Sadece ortalama ve standart sapma kullanılırsa %75 başarıya ulaşılmıştır. Önceki çalışmalarda sıklıkla kullanılan parametreler ile en yüksek %75 değerine ulaşılmışken çalışma sırasında yapılan analizler ile belirlenen parametreler ile başarı oranı %88 olmuştur.

4. Sonuç

Bu makalede konuşmadan duygu tanıma üzerine daha önce yapılan çalışmalar incelenmiştir. Temelde konuşmadan duygu tanıma mimarileri üç ana kısımdan oluşmaktadır. Bunlar özniteliklerin çıkarılması, özniteliklerin seçilmesi ve sınıflandırmadır. İncelenen çalışmalar kapsamında bu üç kısım için hem metrikleri hem de yöntemleri genel hatlarıyla açıklanmıştır. Sonrasında makale bazlı uygulanan yöntemleri ve elde edilen başarı oranlarını özetlenmiştir.

Çoğunlukla çalışmaların tamamına yakını hazır veri setleri üzerinde yapılmıştır. Bir kısmı aynı veri setini hem eğitim hemde test süresince kullanırken bir kısmı da farklı veri setlerini çapraz olarak eğitim ve test süresinde kullanmıştır. Makale kapsamında incelenen çalışmalar için özet tablo aşağıdaki şekildedir.

Tablo 20. Tüm Çalışmaların Özet Sonucu

ID	Öznitelik Çıkarma	Öznitelik Seçme	Sınıflandırma	Veri Seti	En İyi Sonuç
1	MFCC, koleagram ve fraktal boyut		3D CNN-LSTM	RAVDESS 96.18 SAVEE 87.5 RML 93.2 ALL 93.71	RAVDESS 96.18
2	MFCC,Spektrogram		karar ağaçları destek vektör makinesi(SVM) oto kodlayıcı	RAVDESS TESS	Alexnet+Oto Kodlayıcı+TESS(%98)
3	CfsSub,LFS,InfoG,Rank,PCA		Naive Bayes(NB) SMO	EmoSTAR EmoDB	InfoG+Rank(6755)(%97.2) Chi+Rank(6755)(%97.20)
4	149 öznitelik	TBA ADA IDS	ÇKA DVM k-NN	EmoDB	z-puan normalizasyon+ÇKA+yüksek geçiren filtre+ADA (%90.3)
5		korelasyon tabanlı öznitelik seçme	Bayes Ağları lojistik regresyon sıralı minimal optimizasyon karar ağaçları	Türkçe duygusal müzik veri tabanı	MIRtoolbox+Jaudio+OpenSMILE+Bayes %95.96
6	F0			televizyon kanallarından elde edilen konuşma örnekleri EmoDB	EmoDB %82(kızgın %76(nötür)
7	Enerji, Perde, formant, LPCC, MFCC ve MEDC		SVM	Berlin duygu veritabanı SJTU Chinese duygu veritabanı	MFCC+MEDC+Enerj SJTU DB %95.08
8	MFCC, Delta and Delta–Delta MFCCs, Spectral Centroids	Boruta	SVM	EmoDB RAVDESS IITKGP-SEHSC	EmoDB %92.45
9	38 öznitelik		Çok Katmanlı YSA	Berlin Veri Seti	9 gizli katmanlı model %81.65
10	sıfır geçiş oranı		CNN		HSC+EP+LSTM %75.56
	kök kare ortalama,temel frekans,Harmonik-Gürültü-Oranı ve MFCC		LSTM	CHI-MEI MHMC	
11			DNN LSTM-RNN CNN	IEMOCAP	CNN+FC+Softmax %64.78
12	Mel-frekans Kepstrum Katsayıları (MFCCs) , Mel ölçekli spektrogram(Mel-scaled spectrogram), Kromogram(Chromagram), Spektral kontrast özelliği(Spectral contrast feature) ve Tonnetz sunumu(Tonnetz representation)		1D CNN	RAVDESS, EMO-DB IEMOCAP	Model B'de 96.34%
13	MS, LLD, HSF		DNN CNN RNN	IEMOCAP	Füzyon Yöntem %58.3(UA)
14	39 MFCC, Sıfır geçiş oranı (ZCR), Teager Enerji operatörü (TEO) and Harmonik gürültü oranı (HNR)	otomatik-kodlayıcı	SVM	RML	Basit Oto Kodlayıcı %74.07
15	Ayrık Kesirli Fourier Dönüşümü (DFrFT) dayalı Uyarlanabilir Zaman-Frekans öznitelikleri Cepstral öznitelikler	Genetik Algoritma ve Cuckoo Arama	2D CNN	EMO-DB SAVEE PDREC	EMO-DB %97.57(WA)
16	MFCC WPC SFFS	Sıralı Kayan İleri Arama	SVM	EMODB EESDB	WPC+SFFS+RSVM %79.5(EMODB)
17	MFCC		BiMLSTM	SER IEMOCAP	kişiselleştirilmiş model + iki kanallı genel model %78 (SER)
18			Otomatik Kodlayıcı	IEMOCAP FAUAEC PESD EmoDB RAVDESS	SSPSE %61.6
19	Mel frekans büyüklük katsayısı, Mel frekansı kepstral katsayısı, log frekansı güç katsayısı(log frequency power coefficient) ve lineer tahmin kepstal katsayısı(linear prediction cepstral coefficient)		SVM	Berlin Ravdess Savee EMOVO eNTERFACE Urdu	MFMC %95.25(Urdu)
20	3D log-mels spektrogramı		1D evrişimli modüllere sahip bir TCN LSTM BLSTM	IEMOCAP	self-attention-based TCN model %66.1(UA)
22	MFCC	ortalama, maksimum minimum, standart sapma, genlik	SVM	Berlin Duygu Tabanlı Konuşma	0,88

Kaynakça

Ancilin, J., & Milton, A. (2021). Improved speech emotion recognition with Mel frequency magnitude coefficient. Applied Acoustics, 179, 108046.

Aouani, H., & Ayed, Y. B. (2020). Speech emotion recognition with deep learning. Procedia Computer Science, 176, 251-260.

Atila, O., & Şengür, A. (2021). Attention guided 3D CNN-LSTM model for accurate speech based emotion recognition. Applied Acoustics, 182, 108260.

Bhavan, A., Chauhan, P., & Shah, R. R. (2019). Bagged support vector machines for emotion recognition from speech. Knowledge-Based Systems, 184, 104886.

Durukal, M., & Hocaoğlu, A. K. (2015, May). Performance optimization on emotion recognition from speech. In 2015 23nd Signal Processing and Communications Applications Conference (SIU) (pp. 308-311). IEEE.

Fayek, H. M., Lech, M., & Cavedon, L. (2017). Evaluating deep learning architectures for Speech Emotion Recognition. Neural Networks, 92, 60-68.

GÖKALP, S., & AYDIN, İ. (2021). Farklı Derin Sinir Ağı Modellerinin Duygu Tanımadaki Performanslarının Karşılaştırılması. Muş Alparslan Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi, 2(1), 35-43.

Hızlısoy, S. & Tüfekci, Z. (2020). Türkçe Müzikten Duygu Tanıma. Avrupa Bilim ve Teknoloji Dergisi, Ejosat Special Issue 2020 (ICCEES), 6-12. DOI: 10.31590/ejosat.802169

Huang, K. Y., Wu, C. H., & Su, M. H. (2019). Attention-based convolutional neural network and long short-term memory for short-term detection of mood disorders based on elicited speech responses. Pattern Recognition, 88, 668-678.

Issa, D., Demirci, M. F., & Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks. Biomedical Signal Processing and Control, 59, 101894.

Jia, N., & Zheng, C. (2021). Two-level discriminative speech emotion recognition model with wave field dynamics: A personalized speech emotion recognition method. Computer Communications, 180, 161-170.

Korkmaz, O. E. (2016). Ses sinyalinden duygu tanıma (Doctoral dissertation, Karadeniz Teknik Üniversitesi).

Langari, S., Marvi, H., & Zahedi, M. (2020). Efficient speech emotion recognition using modified feature extraction. Informatics in Medicine Unlocked, 20, 100424.

Wang, K., Su, G., Liu, L., & Wang, S. (2020). Wavelet packet analysis for speaker-independent emotion recognition. Neurocomputing, 398, 257-264.

Monica, F., & Antonella, I. (2019). Correlation Analysis. Encyclopedia of Bioinformatics and Computational Biology.

Özseven, T. (2019). Konuşma Tabanlı Duygu Tanımada Ön İşleme ve Öznitelik Seçim Yöntemlerinin Etkisi. Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi, 10(1), 99-112.

PARLAK, C., & Banu, D. İ. R. İ. (2014). FARKLI VERİ SETLERİ ARASINDA DUYGU TANIMA ÇALIŞMASI. Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi, 16(48), 21-29.

Parlak, C., & Diri, B. (2013). İnsan Sesinden Duygu Çıkarma. Sinyal İşleme ve Uygulamaları Kurultayı.

Pan, Y., Shen, P., & Shen, L. (2012). Speech emotion recognition using support vector machine. International Journal of Smart Home, 6(2), 101-108.

POLAT, G., & ALTUN. H. (2008). SES ÖZNİTELİK VEKTÖRLERİNİN DUYGUSAL DURUM SINIFLANDIRILMASINDA KULLANIMI.

Pourebrahim, Y., Razzazi, F., & Sameti, H. (2021). Semi-supervised parallel shared encoders for speech emotion recognition. Digital Signal Processing, 118, 103205.

Yao, Z., Wang, Z., Liu, W., Liu, Y., & Pan, J. (2020). Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN. Speech Communication, 120, 11-19.

Zhao, Z., Bao, Z., Zhang, Z., Cummins, N., Sun, S., Wang, H., & Schuller, B. W. (2021). Self-attention transfer networks for speech emotion recognition. Virtual Reality & Intelligent Hardware, 3(1), 43-54.

Umut Orhan, Makine Öğrenmesi, (21, Kasım, 2021). Erişim Adresi https://bmb.cu.edu.tr/uorhan/DersNotu/Ders11.pdf

Nagesh Singh Chauhan, Naive Bayes, 22, Kasım, 2021). Erişim Adresi (https://www.kdnuggets.com/2020/06/naive-bayes-algorithm-everything.html.

Robith Gandhi, Support Vector Machine- Introduction to Machine Learning Algorithms, (20, Kasım, 2021). Erişim Adresi https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Betül Akpınar, Adaptif Sıralı Minimal Optimizasyon ile Destek Vektör Makinesi, (20, Kasım, 2021). Erişim Adresi https://prezi.com/m7epydjvyf37/adaptif-sral-minimal-optimizasyon-ile-destek-vektor-makine/

Mehmet Fatih Akca, Karar Agaçları, (22, Kasım, 2021). Erişim Adresi https://medium.com/deep-learning-turkiye/karar-a%C4%9Fa%C3%A7lar%C4%B1-makine-%C3%B6%C4%9Frenmesi-serisi-3-a03f3ff00ba5

Statistics Solutions Team, What is Logistic Regresssion, 26, Kasım, 2021). Erişim Adresi https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-logistic-regression/

Mehmet Fatih Akca, RNN Nedir? Nasıl Çalışır? (26, Kasım, 2021). Erişim Adresi https://medium.com/deep-learning-turkiye/rnn-nedir-nas%C4%B1l-%C3%A7al%C4%B1%C5%9F%C4%B1r-9e5d572689e1

Protopars Team, Derin öğrenme (Deep learning) nedir?, (25, Kasım, 2021). Erişim Adresi https://www.protopars.com/derin-ogrenme-deep-learning-nedir/

TechTarget Team, Convolutional neural network, (21, Kasım, 2021). Erişim Adresi https://searchenterpriseai.techtarget.com/definition/convolutional-neural-network

Özgür Doğan, CNN (Convolutional Neural Networks) Nedir?, (22, Kasım, 2021). Erişim Adresi https://teknoloji.org/cnn-convolutional-neural-networks-nedir/

Intellipaat Team, What is LSTM, (25, Kasım, 2021). Erişim Adresi https://intellipaat.com/blog/what-is-lstm/

Veri Bilimci Ekibi, Uzun/Kısa Süreli Bellek, (17, Kasım, 2021). Erişim Adresi https://veribilimcisi.com/2017/09/26/uzun-kisa-sureli-bellek-long-short-term-memory/

Ajitesh Kumar, Gaussian Micture Models: What are they and when to use? (27, Kasım, 2021). Erişim Adresi: https://vitalflux.com/gaussian-mixture-models-what-are-they-when-to-use/

Evren Aslan, Makine Öğrenmesi- KNN Algoritması Nedir, (17, Kasım, 2021). Erişim Adresi https://medium.com/@arslanev/makine-%C3%B6%C4%9Frenmesi-knn-k-nearest-neighbors-algoritmas%C4%B1-bdfb688d7c5f

Francesco Lassig, Temporal Convolutional Networks and Forecasting (11, Kasım, 2021). Erişim Adresi https://unit8.com/resources/temporal-convolutional-networks-and-forecasting/