

EmotionUnet: Konuşma Duygu Tanıma için U-Net Tabanlı Özgün Derin Öğrenme Modeli

Yasin Görmez, Sivas Cumhuriyet Üniversitesi, Yönetim Bilişim Sistemleri, Doktor Öğretim Üyesi, yasingormez@cumhuriyet.edu.tr, 0000-0001-8276-2030, 0000-0001-8276-2030

ÖZ

Konuşma, insanlar arasındaki iletişimin en temel ve etkili yolu olarak değerlendirilmektedir. İnsanlar konuşma yolu ile duygu, düşünce ve bilgilerini paylaşmakta, ilişkilerini güçlendirmekte ve toplumsal bağlarını pekiştirmektedir. Konuşma sırasında karşıdaki kişinin duygu durumunun anlaşılması, empati kurarak daha etkili ve anlamlı bir iletişim sağlamak için önemlidir. Günümüzde telefon gibi araçlarla yapılan uzaktan konuşmalarda ifade edilen duygu tonlarının anlaşılması için konuşma duygu tanıma yöntemlerinden sıklıkla faydalanılmaktadır. Konuşma duygu tanıma müşteri hizmetleri, sağlık, eğitim, eğlence ve akıllı sistemler gibi birçok alanda kullanılmaktadır. Konuşma duygu tanımada sinyal işleme, istatistiksel analiz ve biyometrik teknikler gibi yöntemler kullanılırken, son zamanlarda derin öğrenme yöntemleri de yaygınlaşmıştır. Bu çalışmada konuşma duygu tanıma için evrimsel sinir ağları kullanılarak U-Net tabanlı özgün derin öğrenme modeli önerilmiştir. Önerilen modelin hiper-parametre optimizasyonları için Bayesian optimizasyon yönteminden faydalanılmıştır. Önerilen model Türkçe, İngilizce, Arapça ve Bangla dillerinden dört farklı veri ile analiz edilmiştir. Önerilen model ile farklı veri setlerinde %56,55 ile %99,71 arasında doğruluk hesaplanmıştır.

Anahtar Kelimeler

: Konuşma Duygu Tanıma, Derin Öğrenme, Evrimsel Sinir Ağları, U-Net, Makine Öğrenmesi

EmotionUnet: A Novel Deep Learning Model Based on U-Net for Speech Emotion Recognition

ABSTRACT

Speech is considered to be the most basic and effective way of communication between people. Through speaking, people share their feelings, thoughts and information, strengthen their relationships and reinforce their social bonds. It is important to understand the emotional state of the other person during the conversation in order to provide a more effective and meaningful communication by empathizing. Today, speech emotion recognition methods are frequently used to understand the emotional tones expressed in remote conversations via tools such as telephones. Speech emotion recognition is used in many fields such as customer service, healthcare, education, entertainment, and intelligent systems. While signal processing, statistical analysis and biometric techniques are used in speech emotion recognition, deep learning methods have recently become widespread. In this study, a novel U-Net based deep learning model for speech emotion recognition using convolutional neural networks is



proposed. Bayesian optimization method is used for hyper-parameter optimization of the proposed model. The proposed model is analyzed with four different datasets from Turkish, English, Arabic and Bangla languages. The accuracy of the proposed model is calculated between 56,55% and 99,71% on different datasets.

Keywords : Speech Emotion Recognition, Deep Learning, Convolutional Neural Network, U-Net, Machine Learning

EXTENDED ABSTRACT

INTRODUCTION

Effective communication relies not only on the exchange of thoughts, feelings, and information, but also on the ability to strengthen relationships and reinforce social bonds. A key component of this process is understanding the emotional state of the other person, as it facilitates empathy and leads to more meaningful interactions. In face-to-face communication, cues such as gestures and facial expressions make it easier to discern emotional states. However, in remote communication—such as phone conversations—this becomes more challenging. In such cases, emotional understanding is often derived from vocal elements such as tone, accent, speech rate, volume, pauses, and word choice. Speech Emotion Recognition (SER) methods have been developed to automatically detect emotional states in spoken language, offering a technological solution to this challenge.

In this study, a novel deep learning model for SER is developed using the U-Net architecture. The model incorporates both Convolutional Neural Network (CNN) and Fully Connected Layer (FCL) structures. To enhance model performance, hyper-parameters are optimized using the Bayesian optimization approach. The proposed model is evaluated on datasets from four languages: Turkish, English, Arabic, and Bangla. The primary objective of this study is to assess the performance of the U-Net-based deep learning model in the context of SER. A comprehensive review of the existing literature reveals that U-Net-based architectures have not previously been applied to SER tasks. Therefore, the model developed in this study is considered to be a novel contribution to the field.

MATERIALS AND METHODS

In this study, a novel U-Net-based deep learning model incorporating CNN and FCL layers was developed. The model was trained on the TurEV-DB, BanglaSER, TESS, and ANAD datasets, and its performance was evaluated on each. Key performance metrics, including accuracy, precision, recall, and area under the curve, were employed to assess the model's effectiveness. Spectrogram representations of the audio data, used as inputs, were generated using Short-Time Fourier Transform. Additionally, the model's hyperparameters were optimized through a Bayesian optimization approach.

RESULTS AND DISCUSSION

In the initial phase of the study, each dataset was divided into three subsets: training, validation, and testing. Specifically, 25% of the data was randomly selected for the testing set, 10% for the validation set, and the remaining 65% for the training set. The training set was used to fit the model, the validation set to monitor model performance during hyper-parameter optimization and to trigger callback functions during training, and the test set to evaluate the final performance of the trained models. Once the datasets were partitioned, the hyper-parameters of the proposed model were optimized individually for each dataset. This optimization was performed using the Bayesian approach, implemented through the scikit-optimize library in Python. The results of the hyper-parameter optimization are presented in Table 2.

After the hyper-parameter tuning phase, the model was trained separately on each dataset using the optimized hyper-parameters. During training, two callback functions were incorporated for learning rate reduction and early stopping. The learning rate reduction function halved the learning rate if the validation loss did not improve for two consecutive iterations, while the early stopping function halted training if the validation loss remained unchanged for six consecutive iterations, preventing overfitting. The performance metrics, calculated after training, are displayed in Table 3. These results indicate that the model achieved the highest accuracy on the ANAD dataset and the lowest accuracy on the BanglaSER dataset, with noticeably lower accuracy rates observed on the BanglaSER dataset compared to the others.

To further evaluate model performance, it is critical to analyze results on a class-by-class basis. For this purpose, precision-recall curves were plotted for each dataset, as shown in Figures 2–5. For the TurEV-DB dataset, the best performance was observed in the "sad" class, while the "calm" class yielded the weakest performance. The "angry" and "happy" classes showed similar patterns. In the ANAD dataset, all classes exhibited near-perfect curves. For the BanglaSER dataset, the "neutral" and "angry" classes outperformed the others. In the TESS dataset, nearly all classes showed near-perfect curves except for the "surprised" class, which had a slightly lower performance.

When compared to the existing literature, the proposed model demonstrates superior results on the TESS and ANAD datasets, achieving almost flawless performance. For the BanglaSER dataset, the model's results are close to the literature but still show an improvement over previously reported finding.

CONCLUSIONS

In this study, a novel U-Net-based deep learning model for Speech Emotion Recognition is proposed, utilizing CNN and FCL layers. The model was trained on datasets from four different languages, and its performance was evaluated based on key metrics.

Hyper-parameter tuning was conducted using Bayesian optimization, which offers significant advantages over traditional methods such as grid search for deep learning applications. The experimental results reveal that the proposed model achieved accuracies ranging from 56.55% to 99.71%. When compared to the existing literature, the model underperforms on the TurEV-DB dataset, while slightly outperforming the literature on the BanglaSER dataset. For the ANAD and TESS datasets, the model demonstrates superior performance compared to previously reported results.

GİRİŞ

Konuşma insanlar arasındaki iletişimin en temel ve etkili yoludur. Konuşma sayesinde insanlar; duygu, düşünce ve bilgilerini paylaşabilmekte, ilişkilerini güçlendirebilmekte ve toplumsal bağları pekiştirebilmektedir. Konuşma esnasında karşıdaki kişinin duygu durumunun bilinmesi, empati kurarak daha etkili ve anlamlı bir iletişim sağlanması için büyük önem arz etmektedir. Bu ön bilgi sayesinde insanlar rahatlıkla iş birliği yaparak, birlikte sorunlara çözümler bulabilmektedir. Bu nedenle insanlar konuşma esnasında karşıdaki insanın duygu durumunu anlamak için bir çaba sarf etmektedir. Yüz yüze konuşma esnasında karşıdaki kişinin duygu durumunu anlamak, jest ve mimik gibi unsurlar sayesinde daha rahat olmakta ancak sohbet esnasında kişilerin birbirini görmediği durumlarda duygu durumlarının anlaşılması daha zor olmaktadır. Günümüzde sıklıkla telefon vb. araçlar sayesinde yapılan uzaktan konuşmalarda duygu durumunu; ses tonu, vurgular, hız, ses seviyesi, duraklamalar ve kelime seçimi gibi unsurların analiziyle anlaşılabilir. Konuşma esnasındaki kişilerin duygu durumlarının otomatik bir şekilde elde edilmesi için ise konuşma duygu tanıma (Speech Emotion Recognition- SER) yöntemlerinden faydalanılmaktadır.

SER, konuşma sırasında ifade edilen duygusal tonların ve ruh hallerinin otomatik olarak belirlendiği süreci ifade etmektedir. SER; müşteri hizmetleri, sağlık, eğitim, eğlence ve akıllı sistemler gibi birçok alanda farklı amaçlar ile kullanılmaktadır (Madanian et al., 2023; Zhu et al., 2017). SER kullanım alanlarına acil çağrı merkezlerinde tehlikeli veya yaşamı tehdit eden durumları tanımlamak, araç içi interaktif sesli yanıt sisteminde yorgun sürücülerini tanımlamak, zihinsel sağlık teşhislerini desteklemek ve çevrimiçi eğitim sistemlerinde öğrenci tepki durumlarını anlamak örnek olarak verilebilmektedir (Ahmad et al., 2016; Kerkeni et al., 2020; D. Mishra & rawat, 2015; Zhou et al., 2016).

Geliştirilen SER uygulamalarında konuşulan dil büyük önem arz etmektedir çünkü her dilin kendine özgü ses tonları, ritimleri, vurgu düzenleri ve duygu ifadeleri bulunmaktadır. Bu unsurlar duyguların nasıl ifade edildiğini ve algılandığını büyük ölçüde etkilemekte bu nedenle farklı dillerdeki vurgu, ritim ve tonlama farklılıkları duygusal durumların

tanınmasında kritik rol oynamaktadır. Bu nedenle bir dilde başarılı sonuçlar elde eden bir modelin başka bir dilde istenilen başarıyı elde etmemesi olası bir durumdur (Ververidis & Kotropoulos, 2006). Bahsedilen sebeplerden ötürü tasarlanan bir uygulamanın farklı dillerde de analizinin yapılması, uygulamanın genelleme yapabilme kabiliyetini ölçmek için önemlidir.

SER uygulamalarında sinyal işleme, istatistiksel analiz ve biyometrik teknikler gibi yöntemler kullanılabilir. Bu yöntemlerin arasında fourier dönüşümü, wavelet dönüşümü, cepstral analiz, elektromiyografi, elektrodermal aktivite ve kural tabanlı sistemler gibi yaklaşımlar bulunmaktadır (Wagner et al., 2005). Bu yöntemlerin yanı sıra makine öğrenmesi ve derin öğrenme yöntemleri de SER için sıklıkla kullanılmaktadır. Son zamanlarda birçok problemde olduğu gibi SER uygulamalarında da derin öğrenme yaklaşımlarının model performanslarında iyileşmeler yaptığı görülmektedir (Satt et al., 2017).

Bu çalışmada SER için özgün bir derin öğrenme modeli U-Net mimarisi kullanılarak geliştirilmiştir. Geliştirilen modelde evrimsel sinir ağları (Convolutional Neural Network-CNN) ve tam bağlı katman (Fully Connected Layer- FLC) yapılarından faydalanılmıştır. Derin öğrenme modellerinde başarı oranını etkileyen en önemli unsurlardan biri ise kullanılan hiper-parametrelerdir. Çalışmada önerilen modelin hiper-parametreleri Bayesian optimizasyon yaklaşımı ile optimize edilmiştir. Önerilen model Türkçe, İngilizce, Arapça ve Bangla dili olmak üzere dört farklı dilden veri seti kullanılarak analiz edilmiştir. Bu çalışmanın amacı U-Net tabanlı geliştirilen derin öğrenme modelinin SER üzerindeki performansının ölçülmesidir. Çalışma kapsamında yapılan literatür taramaları doğrultusunda daha önce U-Net tabanlı mimarinin SER için kullanılmadığı görülmüştür. Bu nedenle çalışma kapsamında geliştirilmiş olan model özgün olarak değerlendirilmektedir. Çalışmanın devamında sırası ile literatür taramasından bahsedilmiş, yöntemler anlatılmış, analiz sonuçlarına değinilmiş ve tartışma kısmı ile çalışma sonlandırılmıştır.

1. LİTERATÜR TARAMASI

Literatürde hem geleneksel makine öğrenmesi hem de derin öğrenme yöntemleri kullanılarak SER çalışmaları yapılmıştır. Lin ve Wei yapmış oldukları çalışmada farklı öznitelik teknikleri kullanarak eğittikleri saklı Markov modelleri, destek vektör makinaları ve k en yakın komşu modellerinde sırasıyla %99,5, %88,9 ve %84,5 doğruluk elde etmişlerdir (Lin & Wei, 2005). Jha ve diğerleri RAVDESS veri setinden altı farklı yöntemle öznitelik çıkararak eğittikleri naive Bayes, rastgele orman, k en yakın komşu, destek vektör makinaları ve yapay sinir ağı modellerinden en başarılı yöntemlerin destek vektör makinaları ve yapay sinir ağları olduğu sonucuna varmışlardır (Jha et al., 2022). Krishna ve diğerleri destek vektör makinaları ve yapay sinir ağı sınıflandırma yöntemleri ile geliştirmiş oldukları modellerde %86,50 doğruluğa ulaşmışlardır (Krishna et al., 2022). Liu ve diğerleri CASIA veri seti üzerine korelasyon analizi ve Fisher yöntemi ile öznitelik seçimi uygulamış ve aşırı öğrenme

makinaları yöntemi kullanarak eğittikleri modelde %89,6 doğruluk elde etmişlerdir (Liu et al., 2018). Ghai ve diğerleri mel-frekansı kepsral katsayıları ile öznitelik çıkararak eğittikleri destek vektör makinaları, rastgele karar ormanı ve gradyan artırma yöntemleriyle %81,05 doğruluğa ulaşmışlardır (Ghai et al., 2017).

Harrar ve arkadaşları önermiş oldukları CNN tabanlı derin öğrenme modelini EMO-DB veri seti ile eğiterek %96,97 doğruluğa ulaşmışlardır (Harár et al., 2017). Issa ve arkadaşları geliştirmiş oldukları CNN tabanlı derin öğrenme modelini kullanarak RAVDESS ve EMO-DB veri setlerinde sırasıyla 71,61% ve 95,71% doğruluğa ulaşmışlardır (Issa et al., 2020). Anvarjon ve arkadaşları duygu konuşma tanıma için önerdikleri hafif etkili CNN modeliyle IEMOCAP veri setinde %77,01, EMO-DB veri setinde %92,02 doğruluk elde etmişlerdir (Anvarjon et al., 2020). Mustaqeem ve arkadaşları kısa zamanlı Fourier dönüşümünden (Short-time Fourier transform- STFT) elde edilen spektrogram değerlerini girdi olarak kullanan CNN ve uzun kısa süreli bellek (Long Short-Term Memory - LSTM) tabanlı derin öğrenme modelinde IEMOCAP, EMO-DB ve RAVDESS veri setlerinde sırasıyla 74%, 86% ve 81% F1 skoru elde etmişlerdir (Mustaqeem et al., 2020). Mustaqeem ve Kwon SER için önermiş oldukları tek boyutlu genişletilmiş CNN tabanlı çoklu öğrenme mimarisinde IEMOCAP ve EMO-DB veri setlerinde sırasıyla %73 ve %90 doğruluk elde etmişlerdir (Mustaqeem & Kwon, 2021). Nediyanath ve diğerleri çok kafalı dikkat ağları ve log mel filtre bankası enerjileri kullanarak geliştirmiş oldukları modelle IEMOCAP veri setinde %76,4 doğruluğa ulaşmışlardır (Nediyanath et al., 2020). Singh ve arkadaşları CNN ve destek vektör makinalarını birlikte kullanarak önerdikleri modelde farklı spektrogram yöntemlerini deneyerek EMO-DB veri setinde %79,86, RAVDESS veri setinde %52,54 doğruluğa ulaşmışlardır (Singh et al., 2023). Liu ve arkadaşları dikkat tabanlı LSTM ve CNN kullanarak geliştirmiş oldukları modelde IEMOCAP ve IMPROV veri setlerinde sırasıyla %70,27 ve %60,90 ortalama duyarlılık skoru elde etmişlerdir (Liu et al., 2023). Khan ve diğerleri önerdikleri çok modlu SER modeli sayesinde IEMOCAP ve MELD veri setlerinde doğruluğu %4,5 iyileştirmişlerdir (Khan et al., 2024). Mishra ve arkadaşları farklı öznitelik çıkarma tekniklerini kullanarak elde ettikleri girdiler ile derin sinir ağı eğiterek %87,48 doğruluğa ulaşmışlardır (S. P. Mishra et al., 2024). Mary Little Flower ve diğerleri veri çoğaltma teknikleri sayesinde CNN tabanlı modelin doğruluğunu AESDD, CAFE, EmoDB, IEMOCAP ve MESD veri setlerinde sırasıyla %99,2, %99,5, %97,5, %92,4, ve %96,9 olarak hesaplamışlardır (Mary Little Flower et al., 2024).

2. MATERYAL VE METOT

Çalışmada CNN ve FLC katmanları kullanılarak U-Net tabanlı özgün derin öğrenme modeli geliştirilmiştir. Yapılan literatür taraması doğrultusunda konuşma duygu tanıma alanında U-Net tabanlı bir mimari yapının kullanılmamış olması ise çalışmamızda önerilen modelin özgün olarak değerlendirilmesini sağlamaktadır. Geliştirilen model dört farklı dilden

veri setleri kullanılarak eğitilmiş ve modelin bu veri setlerindeki performans skorları hesaplanmıştır. Girdi olarak kullanılan ses verilerinin spektrogram gösterimleri STFT kullanılarak elde edilmiş ve modelin hiper-parametreleri Bayesian optimizasyon yaklaşımı ile optimize edilmiştir. Girdi olarak STFT spektrogram verileri kullanmak zamana bağlı frekans bilgisini modele sunarken, sesin içeriğini daha iyi temsil eden bir özellik uzayı oluşturmayı sağlamıştır. Hiper-parametre optimizasyonu aşamasında Bayesian yaklaşımının kullanılması ise daha az deneme sayısı ile modeller için en uygun hiper-parametrelerin tespit edilmesini sağlamıştır.

2.1. VERİ SETİ

Çalışmada Türkçe dilinden TurEV-DB (Canpolat et al., 2020), Arapça dilinden ANAD (Altamimi & Alayba, 2023), Bangla dilinden BanglaSER (Das et al., 2022) ve İngilizce dilinden TESS (Pichora-Fuller & Dupuis, 2020) veri setleri kullanılmıştır. Bu veri setleri, her biri için sunulan referanslarda yer alan bağlantılardan indirilmiştir. ANAD veri setinde 741 kızgın, 505 mutlu ve 137 şaşırılmış olmak üzere toplam 1383 örnek bulunmaktadır. BanglaSER veri setinde 306 kızgın, 306 mutlu, 306 üzgün, 306 şaşırılmış ve 243 nötr olmak üzere toplam 1467 örnek bulunmaktadır. TurEV-DB veri setinde 487 kızgın, 408 sakin, 357 mutlu ve 483 üzgün olmak üzere toplam 1735 örnek bulunmaktadır. TESS veri setinde ise korku, şaşırılmış, üzgün, sinirli, iğrenmiş, mutlu ve nötr sınıflarının her birinden 400 olmak üzere 2800 örnek bulunmaktadır. Veri setleri hakkındaki detaylı bilgilere ilgili makalelerden ulaşılabilmektedir.

2.2. KISA ZAMANLI FOURIER DÖNÜŞÜMÜ

STFT, sinyali kısa zaman aralıklarına bölerek her bir aralığın Fourier dönüşümünü hesaplayan zaman-frekans analiz yöntemidir. STFT sinyalin zaman içerisindeki frekans içeriğini incelemeye olanak tanıdığı için özellikle konuşma ve müzik sinyalleri gibi zamanla değişen sinyalleri analiz etmek için sıklıkla kullanılmaktadır. Her bir kısa zaman aralığı, genellikle Hamming veya Hanning gibi bir pencere fonksiyonu ile çarpılmakta ve bu pencere fonksiyonu sinyalin belirli bir kısmını izole etmektedir. STFT hem zamansal hem de frekans bilgilerini koruyarak sinyalin daha kapsamlı bir analizini sağlamaktadır. Zaman-frekans çözünürlüğü, pencere uzunluğuna bağlı olarak değişmektedir. Daha kısa pencereler daha iyi zaman çözünürlüğü sağlarken, daha uzun pencereler daha iyi frekans çözünürlüğü sağlamaktadır (Allen & Rabiner, 1977).

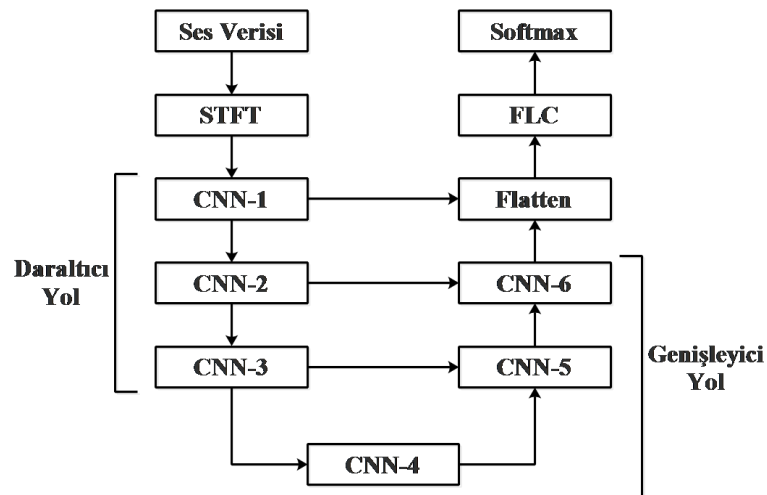
2.3. BAYESIAN OPTIMIZATION

Bayesian optimizasyon, özellikle hesaplamalı olarak pahalı ve zaman alıcı fonksiyonları optimize etmek için kullanılmaktadır. Bayesian optimizasyon yaklaşımı, bir fonksiyonun belirli bir noktadaki değerini tahmin etmek için Bayes teoremini kullanılmaktadır. Bu yöntemde genellikle fonksiyonun bilinmeyen kısımlarını modellemek

için Gaussian süreçlerinden faydalanılmaktadır. Bayesian optimizasyon, özellikle hiper-parametre optimizasyonu gibi alanlarda yaygın olarak kullanılmakta ve derin yapay sinir ağları gibi hiper-parametre sayısı çok ve hiper-parametre uzayı geniş olan modellerde tercih edilmektedir. Bayesian optimizasyon, daha önce kendisine sunulan örneklerden öğrenerek yeni denemelerin nerede yapılması gerektiğini belirleyerek optimize edilecek fonksiyonun minimum veya maksimum değerine daha hızlı ulaşmayı amaçlamaktadır. Öngörücü model ve kazanım fonksiyonu olmak üzere iki temel bileşene dayanmaktadır. Kazanım fonksiyonu, yeni denemelerin yapılacağı noktaları seçerken keşif ve sömürü arasında denge kurmaktadır. Bu denge, yeni bölgeleri keşfetmek ile mevcut bilgiye dayanarak en iyi bilinen çözümü sömürmek arasında bir seçim yapılmasını sağlamaktadır (Snoek et al., 2012).

2.4. U-NET TABANLI DERİN ÖĞRENME MODELİ

Çalışmada altı adet CNN modülü ve bir adet FLC katmanı içeren, EmotionUnet ismi verilen U-Net tabanlı derin öğrenme modeli geliştirilmiştir. ProteinUnet (Kotowski et al., 2021) modeline benzer bir şekilde, Python dilinde var olan keras (*Keras: Deep Learning for Humans*, 2024) kütüphanesi ile geliştirilmiş olan bu modelin mimari yapısı Şekil 1’de gösterilmektedir.



Şekil 1: Emotionunet Modeli Mimari Diyagramı

Şekil 1’de de görüleceği üzere model, CNN katmanlarını daraltıcı yol (contracting path) ve genişleyici yol (expansive path) olmak üzere iki ana katmana ayırmıştır. Daraltıcı yol birbirine seri şekilde bağlı üç CNN modülünden oluşurken genişleyici yol hem birbirine seri şekilde bağlı hem de daraltıcı yolda kendisine karşılık gelen CNN modülüne bağlı üç CNN modülünden oluşmaktadır. Her bir CNN modülü sırasıyla tek boyutlu evrişim ve Batch Normalizasyonu katmanlarından oluşmaktadır. Tek boyutlu evrişim katmanı için filtre sayıları her bir modül için ayrı ayrı optimize edilmiştir. Daraltıcı yoldaki CNN modülleri için çekirdek genişliği 7, genişleyici yoldaki CNN modülleri için ise çekirdek genişliği 5 olarak

ayarlanmıştır. Bu katmandaki aktivasyon fonksiyonu relu ve padding parametresi same olarak ayarlanırken diğer parametreler ise varsayılan olarak bırakılmıştır. CNN katmanlarından sonra ise modele aktivasyon fonksiyonu relu olan bir FLC katmanı eklenmiş daha sonra model softmax katmanı ile sonlandırılmıştır. FLC katmanındaki nöron sayısı da optimize edilmiş, kalan parametreler ise varsayılan ayarda bırakılmıştır. Modelde optimizasyon algoritması Adam ve kayıp fonksiyonu categorical_crossentropy olarak ayarlanmıştır. Modelin öğrenme adımı, döngü sayısı ve parça sayısı değerleri ise optimize edilmiştir. Çalışmada konuşma duygu tanıma için önerilmiş olan U-Net mimarisinin çeşitli faydalarının olduğu düşünülmektedir. Bunlardan ilki iniş-çıkış yaklaşımıyla hem geniş kapsamlı (global) hem de ince ayrıntılı (local) desenleri yakalayabilen yapısı sayesinde ses verisinde farklı zaman aralıklarında oluşan kısa ya da uzun süreli olayları tanıyabilmesidir. Özellikle görsel verilerinde başarılı sonuçlar veren U-Net mimarisi, spektrogram gibi görsel özellikleri kullanan ses verileri ile yüksek uyuma sahiptir. Daraltıcı yol bölümünde özellikler sıkıştırılırken, genişleyici yol bölümünde özellikler genişletilmektedir. Bu duruma en olarak daraltıcı yol ve genişleyici yol bölümlerinde geçiş bağlantısının olması ise veri kaybını minimize etmektedir.

3. DENEY SONUÇLARI

Çalışmanın ilk aşamasında her bir veri seti eğitim, test ve validasyon olmak üzere üçe bölünmüştür. Bu kapsamda örneklerin rastgele %25'i seçilerek test, %10'u seçilerek validasyon ve kalanlar ile ise eğitim veri seti oluşturulmuştur. Her bir veri seti için eğitim, test ve validasyon veri setindeki örnek sayıları tablo 1'de gösterilmektedir.

Tablo 1: Her Bir Veri Setinden Elde Edilen Eğitim, Test ve Validasyon Veri Setleri İçin Örnek Sayıları

Veri Seti Adı	Eğitim Örnek Sayısı	Test Örnek Sayısı	Validasyon Örnek Sayısı
TurEV-DB	1129	433	173
ANAD	900	345	138
BanglaSER	955	366	146
TESS	1431	549	219

Oluşturulan veri setlerinden, eğitim veri seti modeli eğitmek, validasyon veri seti hiper-parametre optimizasyonu aşamasında model performansı hesaplamak ve model eğitimi esnasında geri çağırma (callback) fonksiyonlarını tetiklemek, test veri seti ise eğitilmiş modellerin performansını hesaplamak için kullanılmıştır. Eğitim, test ve validasyon veri setleri oluşturulduktan sonra önerilen modelin hiper-parametreleri her bir veri seti için ayrı ayrı optimize edilmiştir. Hiper-parametre optimizasyonu için kullanılan Bayesian optimizasyon yöntemi Python dilinde var olan scikit-optimize (skopt) (*Scikit-Optimize*, 2025) kütüphanesi kullanılarak geliştirilmiştir. Yöntem için bu kütüphanede var olan gp_minimize

fonksiyonu $acq_func='EI'$, $n_calls=50$ parametre ayarları ile kullanılmıştır. Bu yöntem çizge aramadan farklı olarak, her bir hiper-parametre için bir aralık ya da kategorik bir uzay olarak, belirlenen uzaydan en uygun değerleri Gauss sürecini kullanarak tespit etmeye çalışmaktadır. Optimize edilen hiper-parametreler için hiper-parametre uzayı, hiper-parametre adı, hiper-parametre türü ve her bir veri seti için bulunan optimum değer tablo 2’de gösterilmektedir.

Tablo 2: Optimize edilen hiper-parametreler için hiper-parametre adı, uzayı ve optimum değerler

Hiper-parametre Adı	Hiper-parametre Türü	Hiper-parametre uzayı	TurEV-DB değeri	optimum	ANAD optimum değeri	BanglaSE R optimum değeri	TESS optimum değeri
öğrenme adımı	Gerçek Sayı	En düşük = 10^{-4} , En yüksek = 10^{-1}	0,000232		0,0001208	0,0001	0,0011830
döngü sayısı	Tam Sayı	En düşük = 5, En yüksek = 100	100		52	85	29
parça sayısı	Kategorik	[1, 2, 4, 8, 16, 32, 64, 128]	16		4	8	16
filtre sayısı (CNN-1)	Tam Sayı	En düşük = 16, En yüksek = 256	256		128	68	34
filtre sayısı (CNN-2)	Tam Sayı	En düşük = 16, En yüksek = 256	106		39	16	242
filtre sayısı (CNN-3)	Tam Sayı	En düşük = 16, En yüksek = 256	256		57	256	194
filtre sayısı (CNN-4)	Tam Sayı	En düşük = 16, En yüksek = 256	196		93	256	124

filtre sayısı (CNN-5)	Tam Sayı	En düşük = 16, En yüksek = 256	203	256	116
filtre sayısı (CNN-6)	Tam Sayı	En düşük = 16, En yüksek = 256	169	256	186
nöron sayısı (FLC)	Tam Sayı	En düşük = 100, En yüksek = 1500	596	794	864

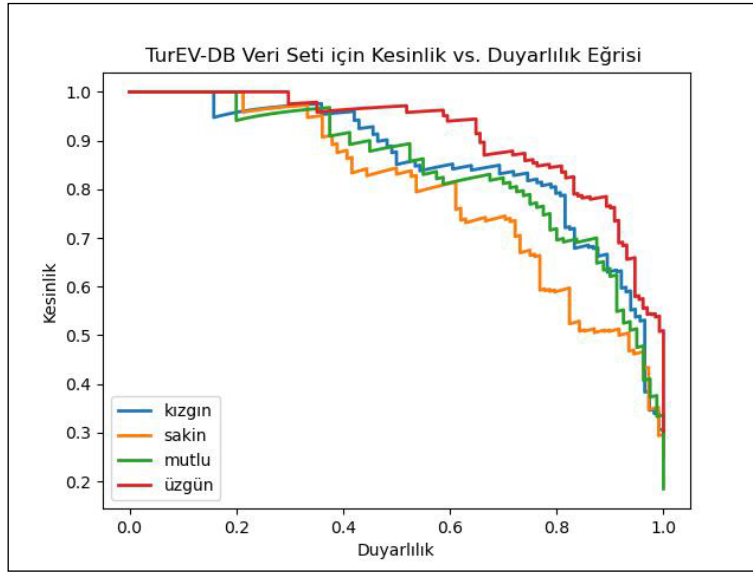
Tablo 2’de yer alan öğrenme adımı (learning rate), her adımda model ağırlarının değiştirilme katsayısını; döngü sayısı (epoch), modelinin eğitim verisinin tamamını kaç kez geçeceğini; parça sayısı (batch size), modelinin eğitim sırasında aynı anda işlediği veri örneklerinin sayısını; filtre sayısı, ilgili evrişim katmanındaki pencere sayısını; nöron sayısı ise FLC katmanındaki birim sayısını ifade etmektedir. Filtre sayısı hiper-parametresi her bir CNN modülü için ayrı ayrı optimize edildiği için tabloda altı farklı filtre sayısı verilmiştir. Filtre sayısı hiper-parametresinde hangi modül için kullanıldığı parantez içerisinde verilmiştir.

Hiper-parametre optimizasyonu aşamasından sonra, belirlenen hiper-parametreler kullanılarak önerilen model her bir veri seti için eğitim kümeleri kullanılarak ayrı ayrı eğitilmiştir. Eğitim esnasında modele öğrenme hızı azalması ve erken durdurma için iki adet geri çağırma fonksiyonu eklenmiştir. Öğrenme hızı azaltma fonksiyonu, validasyon veri kümesi üzerinde art arda iki kez kayıp değeri değişmez ise öğrenme oranının iki ile bölünmesini sağlarken erken durdurma fonksiyonu, validasyon veri kümesi üzerinde art arda altı kez kayıp değeri değişmez ise maksimum döngü sayısına ulaşmadan model eğitiminin bitirilmesini sağlamaktadır. Model eğitimi tamamlandıktan sonra modellerin performanslarını gözlemlemek için test verisi üzerinde doğruluk (accuracy), kesinlik (precision), duyarlılık (recall) ve roc eğrisi altında kalan alan (auc), değerleri hesaplanmıştır. Tablo 3’te her bir veri seti için hesaplanmış olan değerler gösterilmektedir.

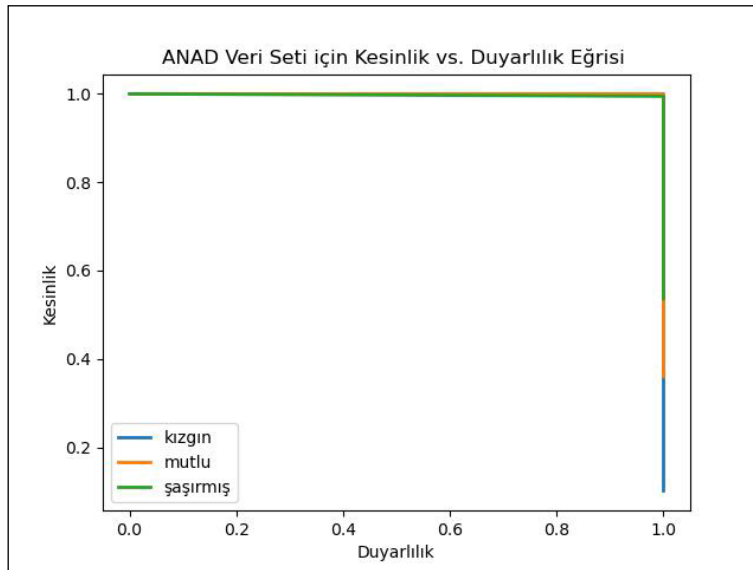
Tablo 3: Optimum Hiper-Parametreler Kullanılarak Eğitilmiş Modellerin Performans Ölçüt Skorları

Veri Seti Adı	Doğruluk	Kesinlik	Duyarlılık	AUC
TurEV-DB	%77,82	%78,01	%76,95	%84,97
ANAD	%99,71	%99,82	%99,04	%99,68
BanglaSER	%59,55	%59,08	%61,17	%74,98
TESS	%98,90	%98,80	%98,77	%99,34

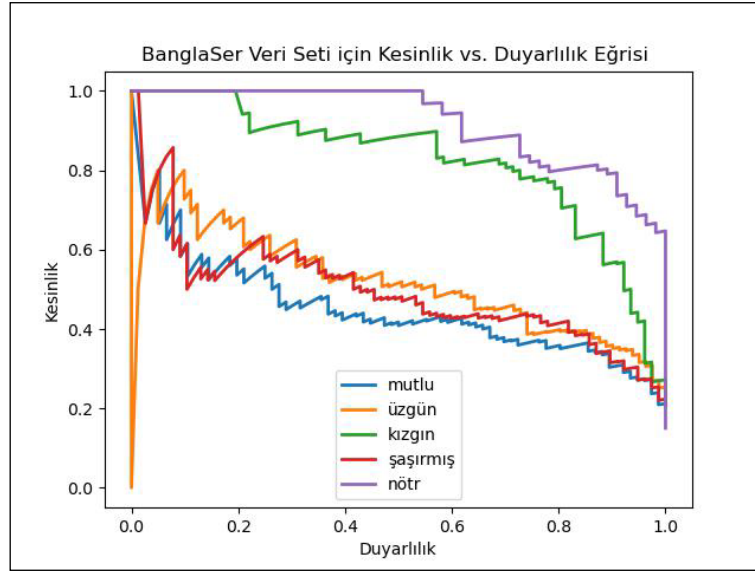
Tablo 3'te yer alan sonuçlar incelendiğinde en yüksek doğruluğun ANAD veri setinde, en düşük doğruluğun ise BanglaSER veri setinde elde edildiği görülmüştür. Özellikle BanglaSER veri setinde doğruluk oranları diğer veri setlerine göre düşük olarak gözlemlenmiştir. Model performanslarının daha iyi değerlendirilebilmesi için sınıf bazında performans skorlarına da bakılması büyük önem arz etmektedir. Bu bağlamda her bir veri seti için sınıf bazında kesinlik-duyarlılık grafikleri çizilmiştir. Şekil 2-5 arasında çizilen bu grafikler gösterilmektedir.



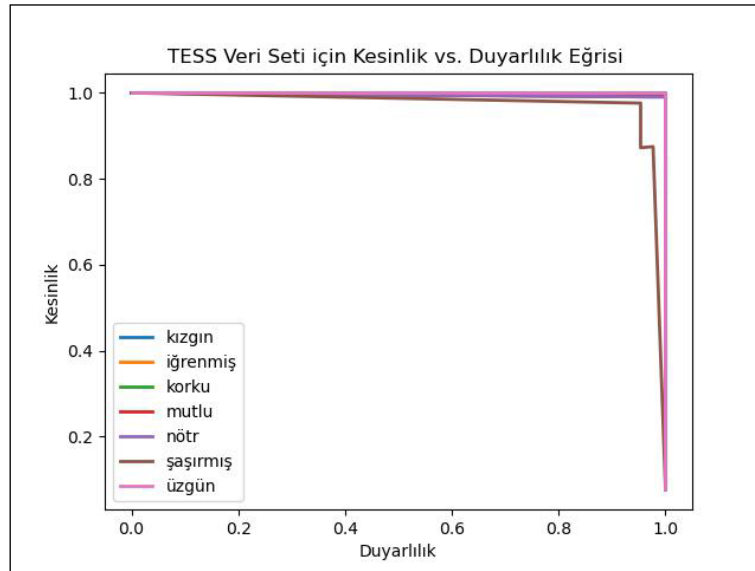
Şekil 2: TurEV-DB Veri Seti için Kesinlik vs Duyarlılık Eğrisi



Şekil 3: ANAD Veri Seti için Kesinlik vs Duyarlılık Eğrisi



Şekil 4: BanglaSER Veri Seti için Kesinlik vs Duyarlılık Eğrisi



Şekil 5: TESS Veri Seti İçin Kesinlik vs Duyarlılık Eğrisi

Şekil 2’de yer alan grafik incelendiğinde, TurEV-DB veri seti için en iyi eğrinin “üzgün”, en kötü eğrinin ise “sakin” sınıf ile elde edildiği gözlemlenmiştir. “Kızgın” ve “mutlu” sınıflarına ait eğrilerin ise birbirine yakın bir çizge takip ettiği görülmektedir. Şekil 3’te yer alan grafik incelendiğinde ANAD veri set için tüm sınıfların neredeyse mükemmel bir eğriye sahip olduğu görülmektedir. Şekil 4’te yer alan grafik incelendiğinde BanglaSER veri seti için “nötr” ve “kızgın” veri sınıflarına ait eğrilerin diğer sınıflara göre daha iyi bir performansa sahip olduğu görülmektedir. Şekil 5’te yer alan grafik incelendiğinde TESS veri seti için “şaşırmış” sınıfı hariç tüm sınıfların neredeyse mükemmel bir eğriye sahip olduğu, “şaşırmış” sınıfının ise diğerlerine göre biraz daha kötü bir eğriye sahip olduğu görülmektedir.

Elde edilen sonuçlar literatür ile karşılaştırıldığında TurEV-DB veri seti için literatürden daha kötü sonuç elde edilmiş, BanglaSER veri seti için literatür ile benzer fakat biraz daha iyi sonuçlar elde edilmiş, ANAD ve TESS veri seti için ise literatürden daha iyi sonuçlar elde edilmiştir. Bu bağlamda, çalışmamızda elde edilen sonuçlar BanglaSER veri seti için Aziz ve diğerleri tarafından veri çoğaltma yapılmadan (çalışmamızdakine benzer şekilde) elde edilen sonuçlar ile (Aziz et al., 2023), TESS veri seti için Sankara Pandiammal ve diğerleri tarafından yapılan çalışma ile (Sankara Pandiammal et al., 2024), ANAD veri seti için ise Albasan tarafından yapılan ve İsmail ve diğerleri tarafından yapılan çalışmalar (Alsabhan, 2023; Ismaiel et al., 2024) ile karşılaştırılmıştır. Model karşılaştırma sonuçları tablo 4'te gösterilmektedir.

Tablo 4: Önerilen Model ile Literatür Sonuçlarının Karşılaştırılması

Model Adı	Veri Seti Adı	Doğruluk
(Aziz et al., 2023)	BanglaSER	59,00%
(Sankara Pandiammal et al., 2024)	TESS	91,00%
(Alsabhan, 2023)	ANAD	96,72%
(Ismaiel et al., 2024)	ANAD	97,40%
Önerilen Model	BanglaSER	59,55%
Önerilen Model	TESS	98,90%
Önerilen Model	ANAD	99,71%

Tablo 4'te yer alan sonuçlar incelendiğinde önerilen modelin TESS ve ANAD veri setleri için en iyi sonuçları elde ettiği ve neredeyse hiç hata yapmadığı gözlemlenmiştir. BanglaSER veri seti için ise literatüre yakın ama literatürden daha iyi bir sonuç elde edildiği görülmektedir.

SONUÇLAR VE TARTIŞMA

Bu çalışmada, SER için U-Net tabanlı özgün bir derin öğrenme modeli önerilmiş ve modelin performansı Türkçe (TurEV-DB), İngilizce (TESS), Arapça (ANAD) ve Bangla (BanglaSER) dillerinden dört farklı veri seti kullanılarak değerlendirilmiştir. Önerilen model, CNN ve FLC yapılarını bir araya getiren U-Net mimarisini temel almakta olup, hiper-parametre optimizasyonu için Bayesian optimizasyon yöntemi kullanılmıştır. Deney sonuçları, modelin veri setlerine bağlı olarak %56,55 ile %99,71 arasında doğruluk oranları elde ettiğini göstermektedir. Deney sonuçları kapsamında elde edilen bu sonuçlar, modelin farklı dillerdeki genelleme kabiliyetini ve U-Net mimarisinin SER alanındaki potansiyelini ortaya koymaktadır.

Modelin performansı, veri setleri arasında belirgin farklılıklar göstermiştir. Özellikle ANAD veri setinde %99,71 doğruluk oranıyla neredeyse mükemmel bir performans elde edilmişken, TESS veri setinde %98,90 doğruluk oranıyla yüksek bir başarı sağlanmıştır. Bu

sonuçlar, literatürdeki diğer çalışmalara kıyasla önemli bir gelişme sunmaktadır. Önerilen modelin ANAD veri setindeki üstün performansı, U-Net mimarisinin spektrogram verilerindeki zaman-frekans özelliklerini etkin bir şekilde yakalayabildiğini ve Bayesian optimizasyonun hiper-parametre ayarlamasında sağladığı avantajları göstermektedir. Benzer şekilde.

Bununla birlikte, BanglaSER veri setinde elde edilen %59,55 doğruluk oranı, diğer veri setlerine kıyasla daha düşük bir performans sergilemiştir. BanglaSER veri setindeki düşük performansın, veri setinde var olan ses özelliklerinin benzerliği gibi faktörlerden kaynaklandığı ön görülmektedir. Özellikle Şekil 4'te gösterilen kesinlik-duyarlılık eğrileri, "nötr" ve "kızgın" sınıfların diğer sınıflara göre daha iyi performans sergilediğini ortaya koymaktadır. Bu durum, Bangla dilindeki duygu ifadelerinin tonlama ve ritim farklılıklarının model tarafından ayırt edilmesinde zorluklar yaratabileceğini düşündürmektedir. TurEV-DB veri setinde ise %77,82 doğruluk oranı elde edilmiş olup, bu sonuç literatürdeki diğer çalışmalara kıyasla daha düşük bir performans göstermiştir. Şekil 2'de sunulan kesinlik-duyarlılık eğrileri, "üzgün" sınıfının en iyi performansı sergilediğini, ancak "sakin" sınıfının daha düşük bir performans gösterdiğini ortaya koymaktadır. Bu durum, Türkçe dilindeki sakın duygu ifadelerinin spektrogram özelliklerinde diğer duygulara kıyasla daha az belirgin olabileceğini işaret etmektedir.

Önerilen U-Net tabanlı modelin başarısının ardındaki temel faktörler, mimarinin daraltıcı ve genişleyici yollarının spektrogram verilerindeki hem yerel hem de küresel desenleri etkin bir şekilde yakalayabilmesi ve Bayesian optimizasyonun hiper-parametre seçiminde sağladığı verimliliğidir. Özellikle, daraltıcı yolun özellik sıkıştırma kapasitesi ve genişleyici yolun özellik genişletme yeteneği, ses verilerindeki kısa ve uzun süreli olayların tanınmasında önemli bir avantaj sağlamaktadır. Bununla birlikte, modelin bazı veri setlerinde (özellikle BanglaSER ve TurEV-DB) daha düşük performans göstermesi, dil-spesifik özelliklerin ve veri seti boyutlarının modelin genelleme kabiliyetini etkilediğini göstermektedir.

Gelecek çalışmalar için önerilen modelin performansını artırmak amacıyla birkaç strateji önerilmektedir. İlk olarak, ses verileri gibi zaman serisi verilerinde sıklıkla başarılı sonuçlar veren uzun kısa vadeli bellek (LSTM) ağları ve Transformer tabanlı katmanlar modele entegre edilerek özellikle BanglaSER ve TurEV-DB veri setlerindeki performansın iyileştirilmesi hedeflenmektedir. İkinci olarak, veri çoğaltma teknikleri kullanılarak veri setlerinin boyutunun artırılmasının, modelin genelleme kabiliyetini güçlendirebileceği ön görülmektedir. Üçüncü olarak, farklı dillerden daha fazla veri seti ile modelin dil bağımsızlığı test edilerek, U-Net tabanlı mimarinin evrensel bir SER çözümü olarak potansiyelinin değerlendirilmesi amaçlanmaktadır.

TEŞEKKÜR

Bu çalışmada yer alan tüm nümerik hesaplamalar TÜBİTAK ULAKBİM, Yüksek Başarım ve Grid Hesaplama Merkezi'nde (TRUBA kaynaklarında) gerçekleştirilmiştir.

KAYNAKÇA

- Ahmad, J., Muhammad, K., Kwon, S., Baik, S. W., & Rho, S. (2016). Dempster-Shafer Fusion Based Gender Recognition for Speech Analysis Applications. *2016 International Conference on Platform Technology and Service (PlatCon)*, 1–4. <https://doi.org/10.1109/PlatCon.2016.7456788>
- Allen, J. B., & Rabiner, L. R. (1977). A unified approach to short-time Fourier analysis and synthesis. *Proceedings of the IEEE*, 65(11), 1558–1564. <https://doi.org/10.1109/PROC.1977.10770>
- Alsabhan, W. (2023). Human–Computer Interaction with a Real-Time Speech Emotion Recognition with Ensembling Techniques 1D Convolution Neural Network and Attention. *Sensors*, 23(3), Article 3. <https://doi.org/10.3390/s23031386>
- Altamimi, M., & Alayba, A. M. (2023). ANAD: Arabic news article dataset. *Data in Brief*, 50, 109460. <https://doi.org/10.1016/j.dib.2023.109460>
- Anvarjon, T., Mustaqeem, & Kwon, S. (2020). Deep-Net: A Lightweight CNN-Based Speech Emotion Recognition System Using Deep Frequency Features. *Sensors*, 20(18), Article 18. <https://doi.org/10.3390/s20185212>
- Aziz, S., Arif, N. H., Ahbab, S., Ahmed, S., Ahmed, T., & Kabir, Md. H. (2023). Improved Speech Emotion Recognition in Bengali Language using Deep Learning. *2023 26th International Conference on Computer and Information Technology (ICCIT)*, 1–6. <https://doi.org/10.1109/ICCIT60459.2023.10441053>
- Canpolat, S. F., Ormanoğlu, Z., & Zeyrek, D. (2020). Turkish Emotion Voice Database (TurEV-DB). In D. Beermann, L. Besacier, S. Sakti, & C. Soria (Eds.), *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)* (pp. 368–375). European Language Resources association. <https://aclanthology.org/2020.sltu-1.52>
- Das, R. K., Islam, N., Ahmed, Md. R., Islam, S., Shatabda, S., & Islam, A. K. M. M. (2022). BanglaSER: A speech emotion recognition dataset for the Bangla language. *Data in Brief*, 42, 108091. <https://doi.org/10.1016/j.dib.2022.108091>
- Ghai, M., Lal, S., Duggal, S., & Manik, S. (2017). Emotion recognition on speech signals using machine learning. *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*, 34–39. <https://doi.org/10.1109/ICBDACI.2017.8070805>

- Harár, P., Burget, R., & Dutta, M. K. (2017). Speech emotion recognition with deep learning. *2017 4th International Conference on Signal Processing and Integrated Networks (SPIN)*, 137–140. <https://doi.org/10.1109/SPIN.2017.8049931>
- Ismail, W., Alhalangy, A., Mohamed, A. O. Y., & Musa, A. I. A. (2024). Deep Learning, Ensemble and Supervised Machine Learning for Arabic Speech Emotion Recognition. *Engineering, Technology & Applied Science Research*, 14(2), Article 2. <https://doi.org/10.48084/etasr.7134>
- Issa, D., Fatih Demirci, M., & Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*, 59, 101894. <https://doi.org/10.1016/j.bspc.2020.101894>
- Jha, T., Kavya, R., Christopher, J., & Arunachalam, V. (2022). Machine learning techniques for speech emotion recognition using paralinguistic acoustic features. *International Journal of Speech Technology*, 25(3), 707–725. <https://doi.org/10.1007/s10772-022-09985-6>
- Keras: Deep Learning for humans. (2024, July 20). <https://keras.io/>
- Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., Ali Mahjoub, M., & Cleder, C. (2020). Automatic Speech Emotion Recognition Using Machine Learning. In A. Cano (Ed.), *Social Media and Machine Learning*. IntechOpen. <https://doi.org/10.5772/intechopen.84856>
- Khan, M., Gueaieb, W., El Saddik, A., & Kwon, S. (2024). MSER: Multimodal speech emotion recognition using cross-attention with deep fusion. *Expert Systems with Applications*, 245, 122946. <https://doi.org/10.1016/j.eswa.2023.122946>
- Kotowski, K., Smolarczyk, T., Roterman-Konieczna, I., & Stapor, K. (2021). ProteinUnet—An efficient alternative to SPIDER3-single for sequence-based prediction of protein secondary structures. *Journal of Computational Chemistry*, 42(1), 50–59. <https://doi.org/10.1002/jcc.26432>
- Krishna, K. V., Sainath, N., & Posonia, A. M. (2022). Speech Emotion Recognition using Machine Learning. *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*, 1014–1018. <https://doi.org/10.1109/ICCMC53470.2022.9753976>
- Lin, Y.-L., & Wei, G. (2005). Speech emotion recognition based on HMM and SVM. *2005 International Conference on Machine Learning and Cybernetics*, 8, 4898–4901 Vol. 8. <https://doi.org/10.1109/ICMLC.2005.1527805>
- Liu, Z.-T., Han, M.-T., Wu, B.-H., & Rehman, A. (2023). Speech emotion recognition based on convolutional neural network with attention-based bidirectional long short-term

- memory network and multi-task learning. *Applied Acoustics*, 202, 109178. <https://doi.org/10.1016/j.apacoust.2022.109178>
- Liu, Z.-T., Wu, M., Cao, W.-H., Mao, J.-W., Xu, J.-P., & Tan, G.-Z. (2018). Speech emotion recognition based on feature selection and extreme learning machine decision tree. *Neurocomputing*, 273, 271–280. <https://doi.org/10.1016/j.neucom.2017.07.050>
- Madanian, S., Chen, T., Adeleye, O., Templeton, J. M., Poellabauer, C., Parry, D., & Schneider, S. L. (2023). Speech emotion recognition using machine learning—A systematic review. *Intelligent Systems with Applications*, 20, 200266. <https://doi.org/10.1016/j.iswa.2023.200266>
- Mary Little Flower, T., Jaya, T., & Christopher Ezhil Singh, S. (2024). Data augmentation using a 1D-CNN model with MFCC/MFMC features for speech emotion recognition. *Automatika*, 65(4), 1325–1338. <https://doi.org/10.1080/00051144.2024.2371249>
- Mishra, D., & rawat, A. (2015). Emotion Recognition through Speech Using Neural Network. *International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE)*, 5.
- Mishra, S. P., Warule, P., & Deb, S. (2024). Speech emotion recognition using MFCC-based entropy feature. *Signal, Image and Video Processing*, 18(1), 153–161. <https://doi.org/10.1007/s11760-023-02716-7>
- Mustaqeem, & Kwon, S. (2021). MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach. *Expert Systems with Applications*, 167, 114177. <https://doi.org/10.1016/j.eswa.2020.114177>
- Mustaqeem, Sajjad, M., & Kwon, S. (2020). Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM. *IEEE Access*, 8, 79861–79875. IEEE Access. <https://doi.org/10.1109/ACCESS.2020.2990405>
- Nediyanchath, A., Paramasivam, P., & Yenigalla, P. (2020). Multi-Head Attention for Speech Emotion Recognition with Auxiliary Learning of Gender Recognition. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7179–7183. <https://doi.org/10.1109/ICASSP40776.2020.9054073>
- Pichora-Fuller, M. K., & Dupuis, K. (2020). *Toronto emotional speech set (TESS)* [Dataset]. Borealis. <https://doi.org/10.5683/SP2/E8H2MF>
- Sankara Pandiammal, K., Karishma, S., Harine Sakthe, K., Manimaran, V., Kalaiselvi, S., & Anitha, V. (2024). Emotion Recognition from Speech – an LSTM approach with the TESS Dataset. *2024 5th International Conference on Innovative Trends in Information Technology (ICITIIT)*, 1–6. <https://doi.org/10.1109/ICITIIT61487.2024.10580351>

- Satt, A., Rozenberg, S., & Hoory, R. (2017). Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms. *Interspeech 2017*, 1089–1093. <https://doi.org/10.21437/Interspeech.2017-200>
- scikit-optimize: Sequential model-based optimization toolbox*. (Version 0.10.2). (2025). [Python; MacOS, Microsoft :: Windows, POSIX, Unix]. <https://scikit-optimize.readthedocs.io/en/latest/contents.html>
- Singh, P., Sahidullah, M., & Saha, G. (2023). Modulation spectral features for speech emotion recognition using deep neural networks. *Speech Communication*, 146, 53–69. <https://doi.org/10.1016/j.specom.2022.11.005>
- Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. *Advances in Neural Information Processing Systems*, 25. <https://proceedings.neurips.cc/paper/2012/hash/05311655a15b75fab86956663e1819cd-Abstract.html>
- Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9), 1162–1181. <https://doi.org/10.1016/j.specom.2006.04.003>
- Wagner, J., Kim, J., & Andre, E. (2005). From Physiological Signals to Emotions: Implementing and Comparing Selected Methods for Feature Extraction and Classification. *2005 IEEE International Conference on Multimedia and Expo*, 940–943. <https://doi.org/10.1109/ICME.2005.1521579>
- Zhou, X., Guo, J., & Bie, R. (2016). Deep Learning Based Affective Model for Speech Emotion Recognition. *2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCoM/IoP/SmartWorld)*, 841–846. <https://doi.org/10.1109/UIC-ATC-ScalCom-CBDCoM-IoP-SmartWorld.2016.0133>
- Zhu, L., Chen, L., Zhao, D., Zhou, J., & Zhang, W. (2017). Emotion Recognition from Chinese Speech for Smart Affective Services Using a Combination of SVM and DBN. *Sensors*, 17(7), Article 7. <https://doi.org/10.3390/s17071694>