Dicle University
Journal of Engineering
https://dergipark.org.tr/tr/pub/dumf
duje.dicle.edu.tr

Araştırma Makalesi / Research Article

Konuşmalardaki duygunun evrişimsel LSTM modeli ile tespiti

Convolutional LSTM model for speech emotion recognition

Ömer Faruk Öztürk[1*], Elham Pashaei[2]

[1] İstanbul Gelişim Üniversitesi, Bilgisayar Mühendisliği Bölümü, 170403057@ogr.gelisim.edu.tr
ORCID: https://orcid.org/0000-0003-1780-3152

[2] İstanbul Gelişim Üniversitesi, Bilgisayar Mühendisliği Bölümü, epashaei@gelisim.edu.tr
ORCID: https://orcid.org/0000-0001-7401-4964

MAKALE BİLGİLERİ

Makale Geçmişi:

Geliş 5 Temmuz 2021
Revizyon 7 Eylül 2021
Kabul 18 Eylül 2021
Online 28 Eylül 2021

Anahtar Kelimeler:

Konuşmada Duygu Tanıma (SER), Uzun-Kısa Süreli Bellek (LSTM), Tekrarlayan Sinir Ağı (RNN), Evrişimli Sinir Ağı (CNN), RAVDESS veri seti MFCC öznitelikleri

ÖZ

Konuşmada duygu tanıma İngilizce adıyla Speech emotion recognition (SER), duyguların konuşma sinyalleri aracılığıyla tanınması işlemidir. İnsanlar, iletişiminin doğal bir parçası olarak bu işlemi verimli bir şekilde yerine getirebilse de programlanabilir cihazlar kullanarak duygu tanıma işlemi hali hazırda devam eden bir çalışma alanıdır. Makinelerin de duyguları algılaması, onların insan gibi görünmesini ve davranmasını sağlayacağından dolayı, konuşmada duygu tanıma, insan-bilgisayar etkileşiminin gelişmesinde önemli bir rol oynar. Geçtiğimiz on yıl içerisinde çeşitli SER teknikleri geliştirilmiştir, ancak sorun henüz tam olarak çözülmemiştir. Bu makale, Evrişimsel Sinir Ağı (Convolutional neural networks -CNN) ve Uzun-Kısa Süreli Bellek (Long Short Term Memory-LSTM) olmak üzere iki derin öğrenme mimarisinin birleşimine dayanan bir konuşmada duygu tanıma tekniği önermektedir. CNN lokal öznitelik seçiminde etkinliğini gösterirken, LSTM büyük metinlerin sıralı işlenmesinde büyük başarı göstermiştir. Önerilen Evrişimsel LSTM (Convolutional LSTM – Co-LSTM) yaklaşımı, insan-makine iletişiminde etkili bir otomatik duygu algılama yöntemi oluşturmayı amaçlamaktadır. İlk olarak, Mel Frekansı Kepstrum Katsayıları (Mel Frequency Cepstral Coefficient- MFCC) kullanılarak önerilen yöntemde konuşma sinyalinden bir görüntüsel öznitelikler matrisi çıkarılır ve ardından bu matris bir boyuta indigenir. Sonrasında modelin eğitimi için öznitelik seçme ve sınıflandırma yöntemi olarak Co-LSTM kullanılır. Deneysel analizler, konuşmanın sekiz duygusunun tamamının RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) ve TESS (Toronto Emotional Speech Set) veri tabanlarından sınıflandırılması üzerine yapılmıştır. MFCC Spektrogram öznitelikleri kullanılarak Co-LSTM ile %86,7 doğruluk oranı elde edilmiştir. Elde edilen sonuçlar, önceki çalışmalar ve diğer iyi bilinen sınıflandırıcılarla karşılaştırıldığında önerilen algoritmanın etkinliğini ikna edici bir şekilde kanıtlamaktadır.

ARTICLE INFO

Article history:

Received 5 July 2021
Received in revised form 7 September 2021
Accepted 18 September 2021
Available online 28 September 2021

Keywords:

Speech Emotion Recognition (SER), Long Short-Term Memory (LSTM), Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), RAVDESS dataset, MFCC features

Doi: 10.24012/dumf.1001914

* Sorumlu Yazar

ABSTRACT

Speech emotion recognition (SER) is the task of recognizing emotions from speech signals. While people are capable of performing this task efficiently as a natural aspect of speech communication, it is still a work in progress to automate it using programmable devices. Speech emotion recognition plays an important role in the development of human-computer interaction since adding emotions to machines makes them appear and act in a human-like manner. Various SER techniques have been developed over the last few decades, but the problem has not yet been completely solved. This paper proposes a speech emotion recognition technique based on the hybrid of two deep learning architectures namely Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM). Deep CNN has demonstrated its effectiveness in local feature selection, whereas LSTM has shown great success in the sequential processing of large texts. The proposed Convolutional LSTM (Co-LSTM) approach aims to create an efficient automatic method of emotion detection in human-machine communication. In the suggested method, Mel Frequency Cepstral Coefficient (MFCC) is used to extract a matrix of spectral features from the speech signal and afterward is converted to 1-dimensional (1D) array. Then, Co-LSTM is employed as a feature selection and classification method to learn the model for emotion recognition. The experimental analyses were carried out on the classification of all the eight emotions of the speech from RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) and TESS (Toronto Emotional Speech Set) databases. An accuracy of 86.7% was achieved with Co-LSTM using MFCC Spectrogram features. The obtained results convincingly prove the effectiveness of the proposed algorithm when compared to the previous works and other well-known classifiers.

Giriş

Birleşmiş Milletler'in bir raporuna göre [1], artan insan nüfusu ile birlikte önümüzdeki beş yıl içerisinde insanlar, çevresindeki bireylerden ziyade sesli asistanlar etkileşime girecek. Günlük etkileşimlerimizde Aple Siri, Cortana, Amazon Alexa ve Google Asistan gibi Sanal Kişisel Asistanların (Virtual Personal Assistants- VPA) yaygınlaşmasıyla, sorularımızı yanıtlama ve taleplerimizi hızlı ve doğru bir şekilde yerine getirme rolünü üstleniyorlar. Bu asistanlar komutlarımızı anlasalar da ruh halimizi tanıma ve buna göre tepki verme konusunda yeterince yetkin değiller. Bu nedenle, bu asistanların yeteneklerini artırabilecek ve tüm endüstride devrim yaratabilecek verimli bir duygu tanıma sistemi geliştirmek önemlidir [2]. Ayrıca duygu tanıma sistemlerinin eğitim [3], sağlık [4][5], güvenlik [6] gibi birçok alanda farklı amaçlarla kullanılabilmesi bu konuda yapılan çalışmaların sayısında önemli bir artışa neden olmuştur [7].

Duygu kelimesi canlıların bir olaya veya nesneye verdiği tepki olarak değerlendirilebilir. Duyguların makineler yardımı ile tespit edilmesini duygu analizi başlığı altında toplamak mümkündür. Duygu analizi metin [8], ses [9], video [10], konuşma [11], Yüz ifade [12] ve beyin sinyal (ElektroEnsefaloGrafi- EEG) [13] verileri üzerinde yapılabilmektedir. Çalışmanın da ana konusu olan ses verilerinden duygu analizi en basit haliyle, alınan sinyallerin bir ön işleme işleminden geçirilmesi, bu sinyallerin bazı öznitelik çıkarım metotları ile özniteliklerin elde edilmesi, çıkarılan özniteliklerin makine öğrenmesi [14] veyahut derin öğrenme [15][16] algoritmaları ile modellenmesi ve ardından bu model ile ses dosyalarının analizi şeklinde özetlenebilir.

Sinyalden çıkarılan öznitelikler dört farklı kategoride gruplandırılabilir ve bu kategoriler akustik, dilsel, bağlamsal bilgi ve farklı öznitelik kümelerinin birleşiminden elde edilen hibrit öznitelikler olarak gruplandırılır. Akustik öznitelikler perde, formant frekansları, enerji, entropy, sıfır geçiş oranı (zero crossing rate), Mel frekansı kepstrum katsayısı (Mel Frequency Cepstral Coefficient- MFCC) [17], doğrusal tahmin katsayıları (Linear Prediction Coefficients- LPC), güç spektral yoğunluğu (Power Spectral Density- PSD), Chroma gibi çeşitli öznitelikler içermektedir [16]. Bu öznitelikler ses verilerinden anlam çıkarmak için kullanılan özniteliklerdir. Bu çalışmada MFCC özniteliği kullanılmıştır.

Bu alanda mevcut olan zorluklardan bahsetmek gerekirse kullanılacak veri seti içerisinde aktarılmak istenen duyguların net ve doğru bir şekilde yansıtıldığı veri setlerinin elde edilmesidir. Bu zorluk göze alınarak araştırmalar için geçerli görülmüş ve günümüzde popüler olan veri setleri the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [18], Toronto Emotional Epeech Set (TESS) [19], a Database of German motional peech (EmoDB) [20] gibi hazır veri setleri veyahut bu alan üzerinde araştırmalar yapan insanların kendilerinin oluşturduğu veri setleri günümüzdeki çalışmalarda kullanılabilmektedir.

Bir başka zorluk da bu veri setlerinin birlikte kullanımını sağlamak. Her veri seti kendine has bir şekilde dosyaları etiketlemekte ve bu veri setlerinden iki veyahut daha fazlasını kullanmak istediğimizde bu veri setlerinin etiketlerini birlikte kullanıma uygun hale getirmek gerekmektedir. Aynı zamanda bir diğer zorluk da hangi duyguların hangi ses öznitelikleri üzerinde etkili olduğunun tespiti yaş ve cinsiyet dahil olmak üzere sesteki bireysel farklılıklardan dolayı zorluk oluşturmaktadır [21].

İlgili Çalışmalar

Ses verilerinden duygu tahmini literatürde önemli bir yer edinmiş bir çalışma alanıdır. Canalys'e göre, akıllı hoparlör kurulu sayısı 2020'de 225 milyona ve 2022'de 320 milyona yaklaşacak. Amazon Echo ve Google Home cihazlarının ABD'nin %50'sinden fazlasında bulunduğu düşünülüyor. Juniper Research'e göre ise 2022 yılına kadar hane halkı ve sesli asistanlara yapılan küresel reklam harcamaları aynı yıl 19 milyar dolara ulaşacak [22]. Gün geçtikçe insanların hayatında yer eden yapay zekâ sistemleri ile birlikte sesli asistanlar ve duygu analizi de bu gelişmeler ile birlikte daha da önem kazanmıştır.

Duyguların ses dosyaları ile analizi ile alakalı literatürde bu alanda çok sayıda çalışma yapılmıştır [2], [11], [23][24][25]. Bu çalışmalar neticesinde farklı metotlar ve farklı veri setleri ile birçok sonuç elde edilmiştir. Ses verilerinin de spektrogramlara dönüştüklerinde bir resim dosyasına benzer bir şekilde iki boyutlu matrise dönüşmesi bu verilerin iki boyutlu öğrenme modelleri üzerinde de çalışmalarını sağlamıştır.

Issa ve diğerleri [11], 8 farklı duyguyu içeren RAVDESS veri seti kullanılarak sesin MFCC, Chromagram, Mel, Contrast, Tonnetz gibi öznitelikleir kullanılarak 1 boyutlu Evrişimsel Sinir Ağı (Convolutional Neural Network- CNN) algoritması eğitimiyle %71,61 doğruluk oranı elde etmişler [11]. Bu özniteliklerin dışında Log-Mel, MFCC, perde ve enerji gibi öznitelikler dikkate alınarak bu özniteliklerin Uzun-Kısa Süreli Bellek (Long Short Term Memory- LSTM), CNN, Gizli Markov Modelleri (Hidden Markov Model- HMM) ve Derin Sinir Ağları (Deep Neural Network- DNN) gibi farklı öğrenme algoritmaları uygulanarak karşılaştırıldığı bir makalede de Log-Mel Spektrogramı öznitelikleri kullanılarak 4 katmanlı 2 boyutlu CNN ile %68 doğruluk elde edilmiştir [2].

CNN'nın mimarilerinden biri olan VGG-16 modeli de bu tarz veri setlerinde kullanılmıştır. Mel spektrogramı ve VGG-16 mimarisi kullanılarak %71 oranında doğruluk elde edilebildiği gözlemlenmiştir [23].

2005 yılında yayınlanan, RAVDESS veriseti gibi içerisinde etiketli ses dosyalarını barındıran başka veri seti olan EmoDB üzerinde Hibrit DNN ve HMM kullanılarak %77,92 oranında bir doğruluk oranı elde edilmiştir [24]. RAVDESS veri seti ses dosyalarının dışında video verileri de içermekte. Bu verilerin hepsinin ses dosyasına dönüştürülmesi ve bunların MFCC özniteliklerinin bir boyutlu CNN ile eğitilmesi sonucu %91 F1 skoru (F1 score) sağlamak mümkündür [25].

Bu çalışmada yapıldığı gibi CNN ve LSTM derin öğrenme algoritmalarının birlikte kullanıldığı örnekler de bulunmaktadır. 2020 yılında yayınlanan ve Berlin EmoDB üzerinde yapılan CNN+LSTM çalışması sonucunda %63 oranında bir doğruluk oranına erişilmiştir [26].

Materyal ve Metot

Veri kümeleri

Veri kümesi olarak konuşma ve şarkı içeren ses dosyaları RAVDESS veri kümesi ve bununla birlikte TESS veri kümesi kullanıldı.

RAVDESS [18] veri kümesi içerisinde 24 seslendirme sanatçısı ve her birine ait 60 konuşma kaydı, 44 şarkı kaydı (23. seslendirme sanatçısının şarkı kaydı bulunmakta) bulunup toplamda 2452 kayıt içermektedir. 12'si kadın, 12'si erkekten oluşan bu seslendirme sanatçıları, Kuzey Amerika aksanı ile İngilizce olan 2 cümleyi seslendirmektedirler. Bu veri seti sakin, mutlu, üzgün, kızgın, korkulu, şaşırmış ve tiksinmiş duygu hallerini içinde bulundurur (7 sınıf etiketi).

2452 ses dosyanın her birinin benzersiz bir dosya adı vardır. Her dosya adı kısa çizgi ile ayrılan toplamda 7 kısımdan (ör. 03-01-06-01-02-01-12.wav) oluşur. Bütün kısımların kendine ait farklı anlamları vardır. Bu kısımları soldan sağa biçimde sırasıyla tanımlamak gerekirse;

1. Dosya tipi (01 = Video ve Ses, 02 = yalnızca video, 03 = yalnızca ses)
2. Söyleme Tarzı (01 = konuşma, 02 = şarkı)
3. Duygu (01 = nötr, 02 = sakin, 03 = mutlu, 04 = üzgün, 05 = kızgın, 06 = korkulu, 07 = iğrenmiş, 08 = şaşırmış)
4. Duygusal yoğunluk (01 = normal, 02 = güçlü).
5. Cümle (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door")
6. Tekrar (01 = 1. tekrar, 02 = 2. tekrar)
7. Seslendirme Sanatçı (01'den 24'e kadar. Tek numaralılar erkek, çift numaralılar kadın seslendirme sanatçıları)

Modelde kullanılan ikinci veri seti, konuşma tabanlı duygu tanıma alanında araştırmacılar tarafından yaygın olarak kullanılan Toronto Emotional Speech Set (TESS)'dir [19]. Toplamda 2800 adet ses dosyası içeren TESS, 26 ve 64 yaşlarında iki farklı ses sanatçısının İngilizce konuşmalarından oluşmaktadır. Her iki ses sanatçısı da Toronto bölgesinde yaşayan, ana dili İngilizce olan, üniversite ve müzik eğitimi almış insanlardır. Bu 2800 verinin toplamında 1 saat 36 dakikalık bir veri seti elde ederiz. RAVDESS veri kümesine nazaran 7 adet duygu içermektedir. Bunlar mutlu, üzgün, kızgın, tiksinmiş, nötr, şaşırmış ve korkulu olma durumlarını kapsamaktadır. RAVDESS veri kümesinden farklı olarak dosya isimleri anlamlı sayılar ile kodlanmamıştır. Bunun yerine dosya isminde ses sanatçısının cinsiyeti, ses dosyasında verdiği duygular yer almaktadır.

Bu çalışmada iki veri seti birlikte kullanıldığı için ilk iş bu veri setlerini birlikte kullanıma hazır hale getirmektir. Bu işlemi yerine getirdikten sonra kullanılan RAVDESS+ TESS veri seti üzerinde birtakım işlemler uygulanarak sesin MFCC öznitelikleri çıkarılmıştır. İlaveten Tablo 1, her duygu için toplam örnek sayısını göstermektedir.

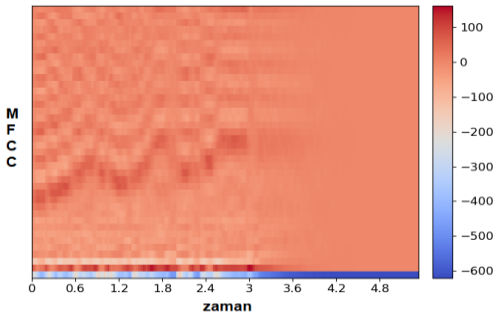
Tablo 1'den, RAVDESS+TESS veri kümesinin dengeli (balanced) olduğu sonucuna varılabilir.

Tablo 1. Toplam numune sayısı [27]

Duygu sınıfı	TESS	RAVDESS	TESS+RAVDESS
Nötr	200	90	290
Sakin	201	185	386
Mutlu	400	185	585
Üzgün	398	185	583
Kızgın	399	186	584
Korkulu	399	186	584
İğrenmiş	399	186	584
Şaşırmış	400	186	586

Öznitelik Çıkarımı

CNN tabanlı modellere öznitelik çıkarımı yapılmadan verilen sinyallerden elde edilen özniteliklerin modelin performansı üzerinde etkisinin az olduğu gözlemlenmiştir [28]. MFCC, insan işitme algılarına dayanmaktadır. İnsanlar, 1Khz üzerindeki frekansları mevcut duyuları ile lineer olarak algılayamazlar [29]. Buradan yola çıkarak bir ses özelliği olan MFCC, insanların algıladıkları seslerin bir özelliğe dökümünü ortaya koymuştur.



Şekil 1. Bir ses dosyasının MFCC ile görünümü

Bu öznitelikler ile birlikte ses dosyası aynı bir resim dosyası gibi spektrogramlar yardımı ile iki boyutlu bir matrise dönüştürülür. Bu sayede ses makine öğrenmesi ve derin öğrenme algoritmalarına hazır hale getirilir.

MFCC öznitelik çıkarımı istenildiği takdirde bir boyutlu bir veriye de indirgenebilir. Bu indirgeme işlemi hali hazırda iki boyutlu matris olan MFCC'nin dikey olarak değerlerinin ortalamasının alınmasıyla sağlanmaktadır. Bu işlem kullanılan donanım üzerindeki yükü azaltacağı gibi doğruluk oranında da etkisi olacaktır. Bahsedilen yöntem performans faktörü göz önünde bulundurularak, bu çalışmada kullanıldı.

Ayrıca insan konuşması için daha uygun olduğu gerekçesiyle yüksek ve alçak geçiren filtreler MFCC özniteliği oluşturulurken 30Khz ve 2700Khz arasındaki veriler dikkate alındı [23]. Bu çalışmanın uygulaması python programlama dili ile yazılmış ve bu işlem için Librosa [30] kütüphanesi kullanılmıştır. Ses verilerinden öznitelik çıkarımı esnasında yüksek ve alçak geçiren filtreler uygulanmıştır.

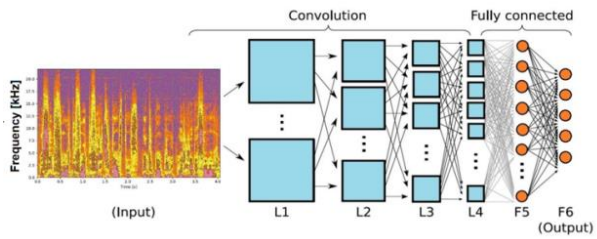
Verilerin eğitim ve test olarak ayrılması

Oluşturulan modelin öğrenmesi için veriler belirli oranlarda bölünür. Bunlardan bir kısmı öğrenmek için ve kalan kısmı da öğrendiklerini test etmek üzerine olur. Bu çalışmada da kullanılan yüzdelik ayırma (Bekletme çapraz doğrulama- Holdout cross valdation) [31]–[33] metodu bu amaca hizmet etmektedir. Veriler bu metod yardımıyla yüzde 80 oranında öğrenme verisi, yüzde 20 oranında test verisi olmak üzere X ve Y verileri kendi içerisinde 2 farklı gruba ayrıldı.

Bunu yapmaktaki amaç öğrenme veri setiyle eğittiğimiz veriyi tekrar test etmek için kullanıldığında, modelin başarısının ölçülmesi doğru sonuçlar vermeyecektir. Modelin daha önce karşılaşmadığı bir veri kümesine modele verdiğimiz zaman, modelin öğrenme başarısının ölçümü gerçeğe yakın bir sonuç verecektir.

CNN

Literatürde de görüldüğü gibi çok sayıda çok katmanlı algılayıcı (Multi-Layer Perceptron- MLP)[34], CNN [35], LSTM [36] gibi derin öğrenme algoritmaları veya rastgele orman (Random Forest- RF) [37], karar ağacı (Decision Tree- DT) [38], destek vektör makinesi (Support Vector Machine- SVM) [39][40], k-en yakın komşu (K-nearest Neighbor- KNN)[41] gibi makine öğrenme yöntemleri duygu analizi çalışmalarında kullanılmıştır. CNN en popüler yapay sinir ağ (YSA) tabanlı modellerinden biridir. YSA, bir insanın beyninde bulunan bilgiyi işleme şekli göz önüne alınarak geliştirilmiş olan bir tekniktir ve yapay zekâ problemlerini matematik ve istatistikten yardım alarak çözmektedir. Bu çalışmada oluşturulan modelin ilk kısmında indirgenmiş olan tek boyutlu MFCC öznitelik vektörü input olarak CNN modele verilmiştir.



Şekil 2. Örnek bir CNN [9]

CNN genel olarak evrişim katmanı, ortaklama katmanı ve tam bağlı katmanı olmak üzere 3 ayrı katmandan oluşmaktadır. Yıllar boyunca yapılan çeşitli araştırma ve uygulamalardan sonra bu alanla ilgilenen insanlar bu modeli farklı veri türleri, farklı veri setleri üzerinde denedikçe CNN modellerini de genel yapıyı bozmayacak şekilde farklı mimariler oluşturarak iyileştirmeye çalıştılar. Bunun üzerine LeNet, AlexNet, VGG, GoogLeNet, ResNet gibi birçok mimari literatüre kazandırılmış oldu.

LSTM

LSTM, derin öğrenme alanında kullanılan bir tekrarlayan sinir ağı (Recurrent Neural Network- RNN) mimarisidir. RNN mimarilerinde bulunan ve diğer mimarilere göre farklı olduğu en temel konu hatırlayabilmesidir. Bir sonraki adım için verilen girdiler arasında bir ilişki ararlar ve bu girdiler içerisinde buldukları ilişkileri hatırlarlar. Kendi içinde bulunan yenilemeli yapısı ve sonucu bir sonraki girdiye aktarması sayesinde tekrarlayan sinir ağı mimarilerinde hatırlama denilen olay yapay sinir ağı çerçevesinde gerçekleşir [42].

Bir RNN mimarisi olan LSTM, bu mimarilerdeki sorunlardan biri olan Kaybolan Gradyan İnişi (Vanishing Gradient Descent) problemini çözmeyi amaçlamıştır. Vanishing Gradient Descent problemi, normalde büyük farklılıklar bulunan girdilerin aktivasyon fonksiyonları sonrasında ufak değer aralıklarına indirgenmesiyle, bunların gittikçe sıfıra yaklaşması ve model tarafından algılanamaması, dolayısıyla da iyi bir öğrenilememesine sebep olur. LSTM bu problemi yapısında bulundurduğu hücre durumu ve unutma, girdi, çıktı kapılarıyla çözmektedir. Bu nedenle, LSTM modelleri iyi performans gösteriyorlar.

Önerilen Co-LSTM

Evrişimsel LSTM (Convolutional LSTM- Co-LSTM) [43] modeli iki önemli derin öğrenme mimarisi olan LSTM ve CNN birleşiminden meydana gelmiş bir mimaridir. CNN ile detaylıca çıkarılan özniteliklerin, LSTM algoritmasının veriyi ilişkilendirme ve hatırlama/unutma mekanizmasıyla birleşmesi bu mimaride önemli rol oynamaktadır. Co-LSTM [43], sosyal medyada bulunan büyük veriden duygu analizi yapmayı amaçlayan metin verilerinde, sıkça kullanılan Naive Bayes (NB) algorimasına göre çok daha iyi bir doğruluk oranı yakalamıştır. Bu çalışmada, konuşma duygu tanıma için önerilen Co-LSTM'nin performansı incelenmiştir. Önerilen yöntemin adımları ve yapısı şekil 3 ve şekil 4'te sırasıyla gösterilmiştir.

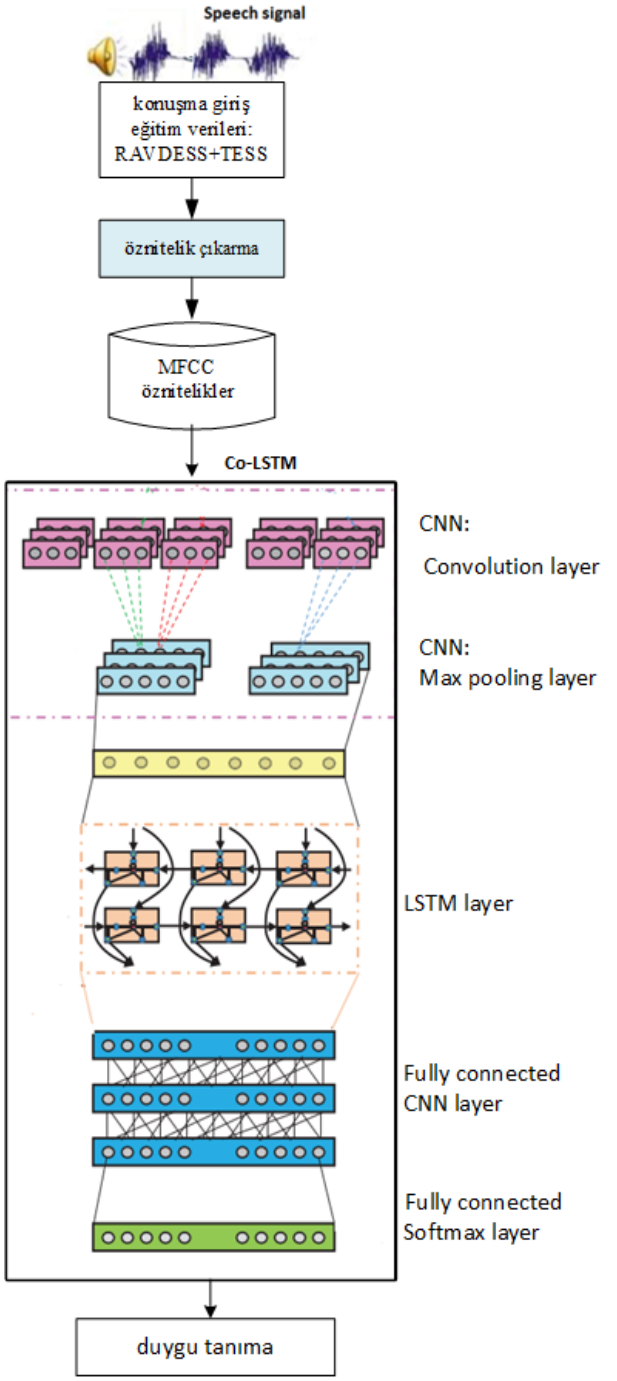
Araştırma Sonuçları ve Tartışm

Gerçekleştirim

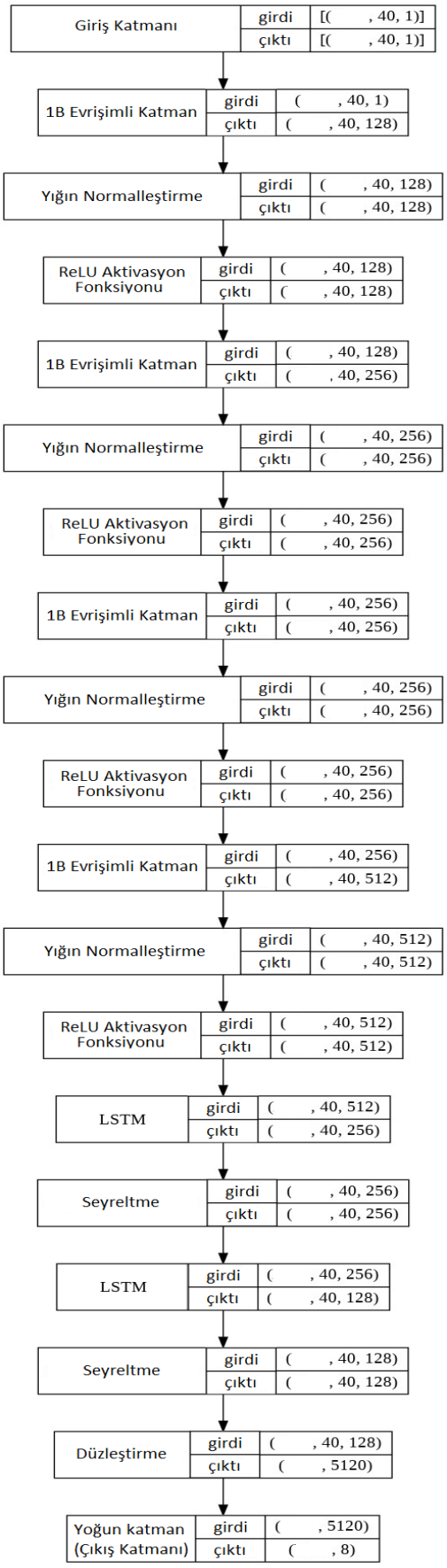
Bu çalışmada MFCC ses özniteliği kullanılmıştır. MFCC ses özniteliğinin sayısı hem kullanılan donanım nedeniyle hem de literatürde sıkça bu değerin kullanıldığından dolayı 40 olarak belirlenmiştir. Oluşturmayı planladığımız yapıda kullanılan evrişimli katmanların (Convolution layers) performans ve başarım açısından düşünülerek 1 boyutlu olması gerektiğine karar verildi. Bu yapının en başında yer alan bir boyutlu CNN (1 Dimensional CNN- 1D CNN) içerisinde her biri sırasıyla 128, 256, 256 ve 512 filtreden ayrıca çekirdek boyutu (kernel size) 3 olan dört evrişimli katman kullanılmıştır. Evrişimli katmana ek olarak Doğrultulmuş Lineer Birim (Rectified Linear Unit- ReLU)

aktivasyon fonksiyonu uygulanmış ve ardından bu katmana yığın normalleştirme işleminden geçirilmiştir.

Bu işlemler bahsi geçen dört katmana da uygulanmıştır Ortaklama katmanlarının (pooling layers), öznitelik sayımızın 40 olması ve ortaklama katmanları kullanılarak bu sayının çok küçük sayılara inmesi bunun sonucunda da başarının düşmesi nedeniyle bu modelde kullanılmamasına kanaat getirilmiştir. CNN'nı tanımladıktan sonra yapıya LSTM algoritması da dahil edildi. LSTM algoritması oluşturulan bu yapıda 2 katmandan oluşmakta.



Şekil 3. Önerilen duygu tanıma modelinde derin öğrenme adımlarının şematik diyagramı ([44]'den uyarlandı)



Şekil 4. Bu çalışmada oluşturulan Co-LSTM yapısı

Bu iki katman da içerisinde 256 birim barındırmaktadır. Ve her bir LSTM katmanının ardından 0,1 oranında bir seyreltme (dropout) katmanı eklenmiştir. Verimizi düzleştirme (flatten) işleminden sonra elde edilen vektörü sadece çıktı katmanı olarak sınıflandırma yapılması için veri setinin duygularının sayısı olan 8 birimlik bir tam bağlantılı katman (Fully Connected Layer) kullanılmıştır. Bu tam bağlantılı katman çıkış katmanı olduğu için Softmax aktivasyon fonksiyonu kullanılmıştır.

Optimizasyon algoritması olarak 2019 yılında yayınlanan ve diğer optimizasyon algoritmalarına nazaran statik öğrenme oranı (learning rate) almayan, Adam optimizasyon algoritması kadar hızlı ve Olasılıksal Dereceli Azalma (Stochastic Gradient Descent- SGD) kadar iyi bir optimizasyon algoritması olduğunu iddia eden AdaBound [45] optimasyon algoritması kullanıldı. AdaBound tek bir öğrenme oranı almak yerine iki adet öğrenme oranı alarak mevcut yapıya bu iki oran arasında bir optimizasyon uygular.

Co-LSTM'yi eğitmek için belirli bir numune sayısı (batch size) seçilmemiştir, mevcut olan varsayılan (default vlaue) ile eğitim tamamlanmıştır. Ayrıca model için 50 tur (epoch) sayısı belirlenmiştir. Bunun nedenini 50. tur sonrasında kayıp fonksiyonumuzdaki artış doğruluk oranımızdaki sabitlik olarak açıklanabilir. Ayrıca, sinir ağlarını optimize etmek için seyrek kategorik çapraz entropi (sparse categorical cross entropy) bir kayıp fonksiyon (loss function) olarak tercih edilmiştir Son olarak MFCC ile özniteliklerine ayırdığımız verileri ve bunların etiketi olan 8 ayrı duyguya göre bir model oluşturulmak istenmiştir.

Oluşturulan modelin başarısını ölçmek için kesinlik (precission), duyarlılık (recall), F1 skoru ve doğruluk oranı (accuracy score) [46][47] gibi değerlendirme parametrelerine bakılmaktadır. Co-LSTM modelin uygulamaları python derin öğrenme kütüphanesi Keras ile yapılmıştır. Ayrıca, bir python kütüphanesi olan Sklearn, makine öğrenme modelleri geliştirmek ve değerlendirme parametrelerinin hesaplanmak için kullanılmıştır.

Derin bir mimari için öğrenme oranı, dropout, katman sayısı vb. gibi hiperparametrelerin seçilmesi çok önemlidir. Hiperparametre optimizasyonunun amacı, derin ağın performansını bağımsız bir veri setinde iyileştirmektir [16]. Deneylerimizde, hiperparametreleri optimize etmek için önerilen Co-LSTM'de rastgele arama (random search) metodu uygulanmıştır [48].

Performans değerlendirme ve karşılaştırma

Bu çalışmada çeşitli öğrenme algoritmaları denenerek mevcut ses veri setinin MFCC ile öznitelikleri çıkarılıp, 8 farklı sınıflandırıcıda test edilerek çeşitli sonuçlar alınmıştır. Bu sonuçlar Tablo 2 de gösterilmektedir. Bu tabloda var olan tüm öğrenme algoritmalarına ızgara arama tabanlı parametre iyileştirme işlemleri uygulanarak (grid search based-hyberparameter tuning) alınabilecek en yüksek doğruluk oranları alınmaya çalışıldı. 40 adet öznitelik içeren çok kısıtlı bir öznitelik kümemiz olmasına rağmen tüm sınıflandırıcıların %60 değeri üzerinde değerler doğruluk oranları aldığı gözlemlenmiştir. Buradan yola çıkarak birleştirmiş olduğumuz iki veri kümesinin birbiriyle uyumlu olduğu söylenebilir. Bu tablodaki değerlendirme parametrelerine bakıldığı zaman en başarılı sonucun %86,7 ile Co-LSTM derin öğrenme algoritmasında olduğu gözükmektedir. Bu ise bize Co-LSTM öğrenme algoritmasının metin verilerinde olduğu kadar ses verilerinde de duygu tahmini için başarılı olduğu varsayımına ulaştırabilir. Co-LSTM algoritmasının ardından gelen algoritmaya baktığımızda SVM geldiği görülmektedir. SVM parametre iyileştirme işlemi sırasında parametrelere göre verdiği doğruluk oranları göz önünde bulundurulduğunda bu algoritmanın mevcut veri seti için çekirdeğin Radyal Temelli Fonksiyon (Radial basis function- RBF) olduğu, marjinin dar olduğu ve dağılım genişliğinin geniş olduğu durumlarda doğruluk oranının arttığı gözlemlenmiştir.

Literatürdeki benzer çalışmalar ile önerilen yöntemin öznitelikleri ve doğruluk oranları Tablo 3 de gösterilmiştir. Literatürdeki benzer çalışmalar göz önüne alındığında oluşturulan modelin kısıtlı bir öznitelik kümesi ile herhangi bir öznitelik seçim algoritmasına tabii olmadan bu sonucu elde etmiş olması modelin bu veri seti için uygun olduğunu göstermektedir.

Tablo 2. Modellerin, RAVDESS+TESS veri setinde MFCC özniteliği kullanılarak elde ettikleri değerlendirme parametreleri

Model	Kesinlik	Duyarlılık	F1 Skoru	Doğruluk Oranı
DT	0.6232	0.6232	0.6233	0.6233
MLP	0.7589	0.7145	0.7150	0.7117
KNN	0.8125	0.8106	0.8100	0.8106
RF	0.8473	0.8372	0.8384	0.8372
SVM	0.8566	0.8553	0.8555	0.8553
LSTM	0.7859	0.7792	0.7802	0.7792
1 D CNN	0.8200	0.8192	0.8187	0.8192
önerilen Co-LSTM	**0.8698**	**0.8677**	**0.8682**	**0.8677**

Deneysel sonuçlar, önerilen yöntemimizin altı son teknoloji yaklaşımının beşinden daha yüksek sınıflandırma doğruluğu elde ettiğini göstermektedir. Önerilen Co-LSTM yönteminin doğruluğu Ref. [19] tekniğinden biraz daha düşük olmasına rağmen, Co-LSTM daha basittir ve Ref. [19]'dan daha düşük zaman karmaşıklığına sahiptir. Bunun nedeni, Ref. [19] modelinin yapısında çeşitli özellik çıkarma yöntemlerini ve parçacık sürüsü optimizasyonu (Partcile swarm Optimization -PSO) stratejisini kullanmasıdır.

Tablo 3. Co-LSTM ve son teknoloji yöntemler (state-of-art methods) arasındaki karşılaştırmanın sonuçları

Çalışma	**Veri Seti**	**Öznitelik**	**Metod**	**Doğruluk Oranı**
Aldeneh ve Provost (2017) [49]	RAVDESS	Mel Spektrogramı	CNN (VGG-16)	0.7100
Darekar ve Dhande (2018) [50]	RAVDESS	MFCC, NMF, Pitch	ANN & PSO-FF	**0.8870**
Bhavan ve diğerleri (2019) [51]	RAVDESS	MFCC	Bagged ensemble of SVM	0.7569
Mekruksavanich ve diğerleri (2020) [52]	RAVDESS TESS	MFCC	DCNN	0.7583 0.5571
Keesing ve diğerleri (2020) [53]	TESS	Mel Spektrogramı	CNN (2D)	0.5510
Tangriberganov ve diğerleri (2020) [26]	EmoDB	-	Co-LSTM	0.6370
Mustaqeem ve Kwon (2020) [35]	RAVDESS	Üst düzey özellikler	Deep stride CNN (DSCNN)	0.81
Singh ve diğerleri (2021)[54]	RAVDESS	MFCC	CNN (1D)	0.8293
Önerilen Co-LSTM	RAVDESS+TESS	MFCC	Co-LSTM	**0.8677**

Sonuç

Duygu Analiz sistemleri canlıların seslerinden, yüz ifadelerinden veya yazdıkları üzerinden duygularının analizi ve tahminlenmesini amaçlamaktadır. Bu çalışmadaki konu olan ses verilerinden duygu analizi işlemini gerçekleştirebilmek için bu seslerden çeşitli özniteliklerin çıkarılması ve bunlardan sistem için anlamlı olanlarının sisteme dahil edilmesi gerekmektedir.

Bu çalışmada, Adabound optimizasyon fonksiyonun, CNN ile LSTM birlikte çalışmasının bir ürünü olan Co-LSTM mimarisinin ses verileri üzerinde duygu analizi için ortaya çıkardığı sonuçları gözlemlendi. Co-LSTM yöntemiyle birlikte diğer çalışmalarda LSTM, CNN ver diğer mimarilerle alınan doğruluk oranlarına kıyasla daha iyi bir sonuç verdiği gözlemlendi.

Yapılan bu çalışmanın bir sonraki aşamasında hayvanlardan alınan veriler üzerine duygularının gerçek zamanlı olarak sınıflandırılması hedeflenmektedir. Gelecek çalışmalarda aynı zamanda ses dosyalarının yanı sıra o canlının görüntüsü veya video kesitleriyle birlikte hazırlamış olduğumuz Co-LSTM modelinden gelen ses tahmininin diğer görsel etkenler ile kıyaslanması ve ortak bir tahminleme yapılması planlanmaktadır.

Kaynakça

[1] "United Nations Educational, Scientific, and Cultural Organization. (2019). I'd blush if I could: closing gender divides in digital skills through education," *2)., (Programme Document GEN/2019/EQUALS/1 REV*. [Online]. Available: http://unesdoc.unesco.org/images/0021/002170/217073e.pdf.

[2] K. Venkataramanan and H. R. Rajamohan, "Emotion Recognition from Speech," *SpringerBriefs Speech Technol.*, pp. 31–32, Dec. 2019.

[3] L. B. Krithika and G. G. Lakshmi Priya, "Student Emotion Recognition System (SERS) for e-learning Improvement Based on Learner Concentration Metric," *Procedia Comput. Sci.*, vol. 85, pp. 767–776, Jan. 2016, doi: 10.1016/J.PROCS.2016.05.264.

[4] A. E. Wells, L. M. Hunnikin, D. P. Ash, and S. H. M. van Goozen, "Improving emotion recognition is associated with subsequent mental health and well-being in children with severe behavioural problems," *Eur. Child Adolesc. Psychiatry 2020*, vol. 1, pp. 1–9, Sep. 2020, doi: 10.1007/S00787-020-01652-Y.

[5] J. R. I. Coleman, K. J. Lester, R. Keers, M. R. Munafò, G. Breen, and T. C. Eley, "Genome-wide association study of facial emotion recognition in children and association with polygenic risk for mental health disorders," *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.*, vol. 174, no. 7, pp. 701–711, Oct. 2017, doi: 10.1002/AJMG.B.32558.

[6] M. Bebawy, S. Anwar, and M. Milanova, "Active Shape Model vs. Deep Learning for Facial Emotion Recognition in Security," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10183 LNAI, pp. 1–11, 2016, doi: 10.1007/978-3-319-59259-6_1.

[7] H. Aouani and Y. Ben Ayed, "Speech Emotion Recognition with deep learning," *Procedia Comput. Sci.*, vol. 176, pp. 251–260, Jan. 2020, doi: 10.1016/J.PROCS.2020.08.027.

[8] B. Kratzwald, S. Ilić, M. Kraus, S. Feuerriegel, and H. Prendinger, "Deep learning for affective computing:

Text-based emotion recognition in decision support," *Decis. Support Syst.*, vol. 115, pp. 24–35, Nov. 2018, doi: 10.1016/J.DSS.2018.09.002.

[9] E. Frant, I. Ispas, V. Dragomir, M. Dascalu, E. Zoltan, and I. C. Stoica, "Voice Based Emotion Recognition with Convolutional Neural Networks for Companion Robots," *Rom. J. Inf. Sci. Technol.*, vol. 20, no. 3, pp. 222–240, 2017.

[10] V. Sreenivas, V. Namdeo, and E. V. Kumar, "Group based emotion recognition from video sequence with hybrid optimization based recurrent fuzzy neural network," *J. Big Data 2020 71*, vol. 7, no. 1, pp. 1–21, Aug. 2020, doi: 10.1186/S40537-020-00326-5.

[11] D. Issa, M. Fatih Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomed. Signal Process. Control*, vol. 59, p. 101894, May 2020, doi: 10.1016/j.bspc.2020.101894.

[12] M. A. Ozdemir, B. Elagoz, A. Alaybeyoglu, R. Sadighzadeh, and A. Akan, "Real time emotion recognition from facial expressions using CNN architecture," *TIPTEKNO 2019 - Tip Teknol. Kongresi*, Oct. 2019, doi: 10.1109/TIPTEKNO.2019.8895215.

[13] M. A. Ozdemir, M. Degirmenci, E. Izci, and A. Akan, "EEG-based emotion recognition with deep convolutional neural networks," *Biomed. Tech. (Berl).*, vol. 66, no. 1, pp. 43–57, Feb. 2020, doi: 10.1515/BMT-2019-0306.

[14] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M. A. Mahjoub, and C. Cleder, "Automatic Speech Emotion Recognition Using Machine Learning," *Soc. Media Mach. Learn.*, Mar. 2019, doi: 10.5772/INTECHOPEN.84856.

[15] A. Saxena, A. Khanna, and D. Gupta, "Emotion Recognition and Detection Methods: A Comprehensive Survey," *J. Artif. Intell. Syst.*, vol. 2, no. 1, pp. 53–79, Feb. 2020, doi: 10.33969/AIS.2020.21005.

[16] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomed. Signal Process. Control*, vol. 47, pp. 312–323, Jan. 2019, doi: 10.1016/J.BSPC.2018.08.035.

[17] N. A. Zaidan and M. S. Salam, "MFCC Global Features Selection in Improving Speech Emotion Recognition Rate," *Lect. Notes Electr. Eng.*, vol. 387, pp. 141–153, 2016, doi: 10.1007/978-3-319-32213-1_13.

[18] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north American english," *PLoS One*, vol. 13, no. 5, p. e0196391, May 2018, doi: 10.1371/journal.pone.0196391.

[19] M. K. Pichora-Fuller and K. Dupuis, "Toronto emotional speech set (TESS)." Scholars Portal Dataverse, 2020, doi: doi/10.5683/SP2/E8H2MF.

[20] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology*, 2005.

[21] B. Zupan, D. Neumann, D. R. Babbage, and B. Willer, "The importance of vocal affect to bimodal processing of emotion: Implications for individuals with traumatic brain injury," *Journal of Communication Disorders*, vol. 42, no. 1. pp. 1–17, Jan-2009, doi: 10.1016/j.jcomdis.2008.06.001.

[22] "Voice-enabled smart speakers to reach 55% of U.S. households by 2022, says report | TechCrunch." [Online]. Available: https://techcrunch.com/2017/11/08/voice-enabled-smart-speakers-to-reach-55-of-u-s-households-by-2022-says-report/. [Accessed: 05-Sep-2021].

[23] A. S. Popova, A. G. Rassadin, and A. A. Ponomarenko, "Emotion Recognition in Sound," in *Studies in Computational Intelligence*, 2018, vol. 736, pp. 117–124, doi: 10.1007/978-3-319-66604-4_18.

[24] L. Li *et al.*, "Hybrid Deep Neural Network - Hidden Markov Model (DNN-HMM) based speech emotion recognition," in *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, 2013, pp. 312–317, doi: 10.1109/ACII.2013.58.

[25] M. G. De Pinto, M. Polignano, P. Lops, and G. Semeraro, "Emotions Understanding Model from Spoken Language using Deep Neural Networks and Mel-Frequency Cepstral Coefficients," in *IEEE Conference on Evolving and Adaptive Intelligent Systems*, 2020, vol. 2020-May, doi: 10.1109/EAIS48028.2020.9122698.

[26] G. Tangriberganov, T. Adesuyi, and B. M. Kim, "(PDF) A Hybrid approach for speech emotion recognition using 1D-CNN LSTM," in *Korea Computer Congress (KCC 2020)*, 2020.

[27] G. Agarwal and H. Om, "Performance of deer hunting optimization based deep learning algorithm for speech emotion recognition," *Multimed. Tools Appl. 2020 807*, vol. 80, no. 7, pp. 9961–9992, Nov. 2020, doi: 10.1007/S11042-020-10118-X.

[28] R. Sarkar, S. Choudhury, S. Dutta, A. Roy, and S. K. Saha, "Recognition of emotion in music based on deep convolutional neural network," *Multimed. Tools Appl.*, vol. 79, no. 1–2, pp. 765–783, Jan. 2020, doi: 10.1007/s11042-019-08192-x.

[29] E. Yucesoy and V. V. Nabiyev, "Gender identification of a speaker using MFCC and GMM," in *ELECO 2013 - 8th International Conference on Electrical and Electronics Engineering*, 2013, pp. 626–629, doi: 10.1109/eleco.2013.6713922.

[30] B. McFee *et al.*, "librosa: Audio and Music Signal Analysis in Python," in *Proceedings of the 14th Python in Science Conference*, 2015, pp. 18–24, doi: 10.25080/majora-7b98e3ed-003.

[31] E. Pashaei, M. Ozen, and N. Aydin, "Splice sites prediction of human genome using AdaBoost," in *3rd IEEE EMBS International Conference on Biomedical and Health Informatics, BHI 2016*, 2016, doi: 10.1109/BHI.2016.7455894.

[32] E. Pashaei, M. Ozen, and N. Aydin, "Random Forest in Splice Site Prediction of Human Genome," in *XIV*

Mediterranean Conference on Medical and Biological Engineering and Computing 2016, 2016, vol. 57, pp. 518–523, doi: 10.1007/978-3-319-32703-7_99.

[33] E. Pashaei and E. Pashaei, "Gene Selection using Intelligent Dynamic Genetic Algorithm and Random Forest," in *2019 11th International Conference on Electrical and Electronics Engineering (ELECO)*, 2019, pp. 470–474, doi: 10.23919/ELECO47770.2019.8990557.

[34] H. K. Palo, M. Chandra, and M. N. Mohanty, "Emotion recognition using MLP and GMM for Oriya language," *Int. J. Comput. Vis. Robot.*, vol. 7, no. 4, pp. 426–442, 2017, doi: 10.1504/IJCVR.2017.084987.

[35] Mustaqeem and S. Kwon, "A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition," *Sensors 2020, Vol. 20, Page 183*, vol. 20, no. 1, p. 183, Dec. 2019, doi: 10.3390/S20010183.

[36] F. Tao and G. Liu, "Advanced LSTM: A Study about Better Time Dependency Modeling in Emotion Recognition," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2018-April, pp. 2906–2910, Sep. 2018, doi: 10.1109/ICASSP.2018.8461750.

[37] L. Chen, W. Su, Y. Feng, M. Wu, J. She, and K. Hirota, "Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction," *Inf. Sci. (Ny).*, vol. 509, pp. 150–163, Jan. 2020, doi: 10.1016/J.INS.2019.09.005.

[38] Z. T. Liu, M. Wu, W. H. Cao, J. W. Mao, J. P. Xu, and G. Z. Tan, "Speech emotion recognition based on feature selection and extreme learning machine decision tree," *Neurocomputing*, vol. 273, pp. 271–280, Jan. 2018, doi: 10.1016/J.NEUCOM.2017.07.050.

[39] L. Sun, B. Zou, S. Fu, J. Chen, and F. Wang, "Speech emotion recognition based on DNN-decision tree SVM model," *Speech Commun.*, vol. 115, pp. 29–37, Dec. 2019, doi: 10.1016/J.SPECOM.2019.10.004.

[40] E. Pashaei, A. Yilmaz, and N. Aydin, "A combined SVM and Markov model approach for splice site identification," *2016 6th Int. Conf. Comput. Knowl. Eng. ICCKE 2016*, no. Iccke, pp. 200–204, 2016, doi: 10.1109/ICCKE.2016.7802140.

[41] J. Umamaheswari and A. Akila, "An Enhanced Human Speech Emotion Recognition Using Hybrid of PRNN and KNN," *Proc. Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput. Trends, Prespectives Prospect. Com. 2019*, pp. 177–183, Feb. 2019, doi: 10.1109/COMITCON.2019.8862221.

[42] T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li, "Spatial-Temporal Recurrent Neural Network for Emotion Recognition," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 939–947, Mar. 2019, doi: 10.1109/TCYB.2017.2788081.

[43] R. K. Behera, M. Jena, S. K. Rath, and S. Misra, "Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data," *Inf. Process. Manag.*, vol. 58, no. 1, p. 102435, Jan. 2021, doi: 10.1016/j.ipm.2020.102435.

[44] V. Passricha and R. K. Aggarwal, "A Hybrid of Deep CNN and Bidirectional LSTM for Automatic Speech Recognition," *J. Intell. Syst.*, vol. 29, no. 1, pp. 1261–1274, Jan. 2020, doi: 10.1515/JISYS-2018-0372.

[45] L. Luo, Y. Xiong, Y. Liu, and X. Sun, "Adaptive Gradient Methods with Dynamic Bound of Learning Rate," *7th Int. Conf. Learn. Represent. ICLR 2019*, Feb. 2019.

[46] M. A. Ozdemir, G. D. Ozdemir, and O. Guren, "Classification of COVID-19 electrocardiograms by using hexaxial feature mapping and deep learning," *BMC Med. Informatics Decis. Mak. 2021 211*, vol. 21, no. 1, pp. 1–20, May 2021, doi: 10.1186/S12911-021-01521-X.

[47] M. A. Ozdemir, O. K. Cura, and A. Akan, "Epileptic EEG Classification by Using Time-Frequency Images for Deep Learning," *https://doi.org/10.1142/S012906572150026X*, May 2021, doi: 10.1142/S012906572150026X.

[48] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for Hyper-Parameter Optimization," *Adv. Neural Inf. Process. Syst.*, vol. 24, 2011.

[49] Z. Aldeneh and E. M. Provost, "Using regional saliency for speech emotion recognition," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2017, pp. 2741–2745, doi: 10.1109/ICASSP.2017.7952655.

[50] R. V. Darekar and A. P. Dhande, "Emotion recognition from Marathi speech database using adaptive artificial neural network," *Biol. Inspired Cogn. Archit.*, vol. 23, pp. 35–42, Jan. 2018, doi: 10.1016/j.bica.2018.01.002.

[51] A. Bhavan, P. Chauhan, Hitkul, and R. R. Shah, "Bagged support vector machines for emotion recognition from speech," *Knowledge-Based Syst.*, vol. 184, p. 104886, Nov. 2019, doi: 10.1016/J.KNOSYS.2019.104886.

[52] S. Mekruksavanich, A. Jitpattanakul, and N. Hnoohom, "Negative Emotion Recognition using Deep Learning for Thai Language," in *2020 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering, ECTI DAMT and NCON 2020*, 2020, pp. 71–74, doi: 10.1109/ECTIDAMTNCON48261.2020.9090768.

[53] A. Keesing, I. Watson, and M. Witbrock, "Convolutional and Recurrent Neural Networks for Spoken Emotion Recognition," in *Proceedings of the The 18th Annual Workshop of the Australasian Language Technology Association*, 2020, pp. 104–109.

[54] P. Singh, G. Saha, and M. Sahidullah, "Deep scattering network for speech emotion recognition," May 2021.