Muş Alparslan Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi
2(1),35-43,2021

Farklı Derin Sinir Ağı Modellerinin Duygu Tanımadaki Performanslarının Karşılaştırılması

Süha GÖKALP[1], İlhan AYDIN[2]
[1]Atatürk Üniversitesi Pasinler MYO Bilgisayar Teknoloji Bölümü, Erzurum, Türkiye
[2]Fırat Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği Bölümü, Elazığ, Türkiye
Sorumlu yazar: suha.gokalp@atauni.edu.tr

Geliş tarihi:26.05.2021; **Kabul tarihi:08.06.2021**

Özet

Teknolojinin geliştirilmesi ile insan ve makine etkileşimi her geçen gün artmaktadır. Bilim insanları bu etkileşim nedeniyle oluşan iletişimin dolayısıyla bilgi alışverişinin güçlendirilmesini amaçlamaktadırlar. Son yıllarda güçlendirme için insan sesinin ve yüz ifadelerinin analiz edilerek insan duygularının otomatik olarak tanınmasını sağlayan çalışmaların sayısında artış yaşanmaktadır. Ses sinyalinde duygu tanıma özelikle, görsel bilginin kısıtlı ya da hiç olmadığı durumlarda oldukça önemlidir. Bu çalışmada da insan sesinin analiz edilerek duyguların otomatik olarak tanımlanması üzerine kayda alınmış RAVDESS (The Ryerson Audio-Visual Database of Emotional Speech and Song) ve TESS (Toronto Emotional Speech Set) ses kayıtları veri seti olarak kullanılmış, makine öğrenmesi sınıflandırıcıları ve derin öğrenme algoritmaları kullanılarak modellerin iyi tahminler üretip üretmediğine bakılmış, algoritmalar ve yöntemler kıyaslanmıştır. Bunların yanı sıra Alexnet, Resnet50 ve SqueezeNet ağları da kıyaslamaya dahil edilmiştir. RAVDESS ve TESS veri setleriyle Alexnet ağında Karar Ağacı %44, SVM %29 isabetli sonuç elde edilirken, RAVDESS veri setine TESS eklendiğinde sonuçlar %64 ve %55 isabet oranına yükselmiştir. Ağlar arasında en iyi sonuç Squeezenet'le 100 adımdan henüz 70 adım gerçekleştiğinde tam başarım elde edilirken en kötü sonuç MobileNet'te %15 isabette kalmıştır. Evrişimsel sinir ağı derin öğrenme algoritmalarının bütün ağlarda %15-17 civarı isabetli sonuçlar verdiği gözlemlenmiştir.

Anahtar Kelimeler: Derin Öğrenme, Duygu Tanıma, MobileNet, SqueezeNet, Evrişimsel Sinir Ağı

Comparison of Different Deep Neural Network Models in Emotion Recognition

Abstract

With the development of technology, human and machine interaction is increasing day by day. Scientists aim to strengthen the communication that occurs due to this interaction, and thus the exchange of information. In recent years, there has been an increase in the number of studies that enable automatic recognition of human emotions by analyzing human voices and facial expressions for reinforcement. Emotion recognition in the audio signal is particularly important in situations where visual information is limited or absent. In this study, RAVDESS (The Ryerson Audio-Visual Database of Emotional Speech and Song) and TESS (Toronto Emotional Speech Set), which were recorded on the automatic identification of emotions by analyzing the human voice, were used as a data set, machine learning classifiers and deep learning algorithms. It was checked whether the models produced good predictions by using them, and algorithms and methods were compared. In addition, Alexnet, Resnet50, SqueezeNet and MobileNet networks are included in the benchmark. With the RAVDESS dataset, 44% of the Decision Tree and 29% of SVM were obtained in the Alexnet network, while when TESS was added to the RAVDESS data set, the results increased to 64% and 55% hit rate. Among the networks, the best result was achieved with Squeezenet when only 70 steps from 100 steps were achieved, while the worst result remained with a 15% hit on MobileNet. It has been observed that convolutional neural network deep learning algorithms give accurate results around 15-17% in all networks.

Keywords: Deep Learning, Emotion Recognition, MobileNet, SqueezeNet, Convolutional Neural Network

1. GİRİŞ

İnsanoğlunun var oluşundan beri iletişim bilgi alışverişinin temelidir. İletişimi daha doğru, net ve anlaşılır kılmak için kelimeler ve duygular birbirlerini takip etmektedir[1]. 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[2]. Duygu tahminleri için genellikle makineler kullanılmaktadır[1].

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 [3]. Konuşma sinyali, konuşanın fizyolojisi, ruh hali, yaşı, cinsiyeti, lehçesi gibi birçok bilgiyi aynı anda barındırabilen karmaşık bir işarettir. Konuşmadan duygu tanıma çalışmaları, konuşma sırasında çıkan ses sinyalinden anlam bilgisini elde etmeye çalışmaktadırlar [4].

Teknolojinin gelişmesi ile birlikte insan ve makine arasındaki etkileşim hızla artmaktadır. Son yıllarda insan yüz ifadelerinin ve sesinin analiz edilerek duyguların otomatik olarak tanımlanması üzerine çok fazla çalışma yapılmıştır. Ses sinyalleri ile duygu tanımayı analiz eden birçok çalışma bulunmaktadır. De Pinto ve ark. (2020) ses kayıtlarını kullanarak duyguların sınıflandırılması için derin sinir ağlarına dayalı bir sistem geliştirmişlerdir. Model, sekiz farklı duyguyu (tarafsız, sakin, mutlu, üzgün, kızgın, korkulu, tiksinti, şaşırmış) sınıflandırmak için eğitilmiştir[5]. Tarantino ve ark.(2019) veri tabanındaki son teknoloji sonuçları önemli ölçüde geliştirilmesini başarmışlardır[6]. Triantafyllopoulos ve ark. (2019) konuşma duygu tanıma görevi için konuşma geliştirme algoritmalarını uygulamışlardır ve performansta önemli gelişme göstermişlerdir[7]. Zhao ve ark. (2019) log-mel spektrogramlarını 2-D evrişimsel sinir ağı (CNN) LSTM ağlarına giriş verileri olarak kullanmışlardır. Çalışmalarının sonuçları, veri setinden alınan örneklerin konuşmacıya bağlı sınıflandırma için % 95.33 ve konuşmacıdan bağımsız sınıflandırması için % 95.89 ile mümkün olan en iyi doğruluğu göstermiştir[8]. Chatziagapi ve ark. (2019) ise konuşma duygusu tanıma modellerinin performansını artırmak için Generative Adversarial Network (GAN) tabanlı veri artırma yaklaşımını önermişlerdir. Yazarlar, sentetik olanlar üretmek için GAN'ı spektrogramlara uygulamışlardır. Bu teknik, % 10 bağıl performans kazancı elde etmek için veritabanı örneklerine uygulanmıştır[9]. Hossain ve Muhammad (2019) büyük veri yapıları (konuşma ve video verileri) ile derin öğrenme tabanlı bir duygu tanıma sistemi sunmuşlardır. Bu sistemde öncelikle ses sinyali frekans düzlemine geçirilerek Mel Spektrumu görüntüleri elde edilmiş bu görüntülerden özellikler çıkartılmıştır [10]. Iqbal ve ark. (2019) belirli göreve bağlı olarak cinsiyete dayalı farklılıkları yaklaşık % 40 ile % 80 arasında genel doğrulukla tanımlamak için bu çalışmada kullanılan veri kümesinde granüler bir sınıflandırma üzerinde çalışmak için Gradient Boosting, KNN ve SVM kullanılmıştır. Özellikle, önerilen sınıflandırıcılar farklı veri kümelerinde farklı performans göstermiştir [11].Salur ve Aydın (2020) en popüler sosyal medya uygulamalarından olan Twitter ortamında paylaşılan mesajlar üzerinde ikili duygu sınıflandırma (olumlu-olumsuz) işlemi gerçekleştirmişlerdir[12]. Yazarların başka bir çalışmalarında ise farklı kelime yerleştirmeyi (Word2Vec, FastText, karakter seviyesi yerleştirme) farklı derin öğrenme yöntemleriyle (LSTM, GRU, BiLSTM, CNN) stratejik olarak birleştiren yeni bir hibrit derin öğrenme modeli önermişlerdir[13].

Bu çalışmanın amacı iki farklı veri setinden, beş farklı ağ modeli ve iki farklı sınıflandırma algoritmaları ile derin öğrenme kullanarak duygu tanıma sistemi gerçekleştirmektir.

2. MATERYAL VE METOD

Bu bölümde kullanılan veri setleri, yapay sinir ağları ve sınıflandırıcılar anlatılmaktadır.

Çalışma Python 3.6 üzerinde Tensorflow = 2.4, Keras = 2.2, librosa, pandas, numpy, PIL, matplotlib kütüphaneleri kullanılarak, CPU~Quad core Intel Core i7-4790 (-MT-MCP-) speed/max~1216/4000 MHz Kernel~4.15.0-76-generic x86_64 Mem~8861.6/15948.6MB özelliklerine sahip bir bilgisayar ile yapılmıştır.

2.1. Veri Seti

Çalışmada aşağıdaki 2 farklı veri seti kullanılmıştır.

Ryerson Duygusal Konuşma ve Şarkının Görsel-İşitsel Veritabanı (RAVDESS) [14]: Veri setinin 1440 konuşma dosyası ve 1012 şarkı dosyasının yer aldığı sadece sesten oluşan kısmı kullanılmıştır. Bu veri seti, iki sözcüksel olarak eşleştirilmiş ifadeyi tarafsız bir Kuzey Amerika aksanıyla seslendiren 24 profesyonel oyuncunun (12 kadın, 12 erkek) kayıtlarından oluşmaktadır. Toplam 8 farklı duygu, seslendiriciler tarafından güçlü ve normal duygu yoğunluğu olarak iki defa seslendirilmiştir. Konuşma veri seti sakin, mutlu, üzgün, kızgın, korkulu, şaşkın, iğrenç duygularını; şarkı veri seti ise sakin, mutlu, üzgün, kızgın ve korkulu duygularını içermektedir.

Toronto Duygusal Konuşma Seti (TESS) [15]: TESS veri seti 2800 uyarandan ve toplam 5600 kayıttan oluşmaktadır. Bu veri seti Toronto bölgesinden 26-64 yaşları arasında, İngilizceyi ilk dili gibi kullanan, üniversite mezunu ve müzik eğitimi almış aktrislerin konuşmalarını içermektedir. Oyunculara yapılan odyometrik test, her iki oyuncunun da normal aralıkta eşiklere sahip olduğunu göstermiştir. Veri setinde yalnızca kadınların çok yüksek kalitede seslendirdiği konuşmalar duygu tanımada kullanılan 4 anahtar setten biridir. Diğer veri kümelerinin çoğu erkek konuşmacılara doğru çarpıktır ve bu nedenle biraz dengesizlik temsiline neden olmaktadır. Bu veri kümesi, genelleme açısından duygu sınıflandırıcı için çok iyi bir eğitim veri kümesine hizmet etmektedir. Ayrıca aşırı uyum gösterme görülmemektedir. Bu veri setinde iki aktris tarafından taşıyıcı cümlesinde 200 hedef kelimeden oluşan bir set konuşulmuştur ve sette yedi duygunun (öfke, tiksinti, korku, mutluluk, hoş sürpriz, üzüntü ve tarafsız) her birini tasvir eden kayıtlar yapılmıştır.

2.2. Kullanılan Yapay Sinir Ağları ve Sınıflandırıcılar

AlexNet, Resnet50, MobileNet ve Squeezenet kullanılmıştır.

Karar ağaçları ve destek vektör makinesi sınıflandırıcıları kullanılmıştır. Her evrişimli sinir ağı ayrıca modelleştirilip işlenmiştir.

2.2.1. AlexNet

Krizhevsky, Sutskever ve Hinton tarafından geliştirilmiş derin sinir ağ modeli, dünya çapında derin öğrenmenin duyulmasını sağlayan 2012 yılı ImageNet yarışmasını kazanmıştır. Birbirini takip eden evrişim katmanları ve ortaklama katmanları mevcuttur. Bu model bilgisayarlı nesne tanımlama doğruluk oranını %10.8 iyileştirerek %83.6 olmasını sağlamıştır.

60 milyon parametre ve 650.000 nöron içeren sinir ağı; 5 evrişim katmanı bunların birçoğunu takip eden maksimum havuzlama katmanı ve 3 tam bağlantılı katmandan oluşmaktadır [16]. ImageNet veri seti 1000 farklı örnek resim sınıfı içerdiğinden, çıktı katmanı 1000 birimden oluşmaktadır. Elde edilen model, LeNet'e çok benzer bir mimariye sahip gibi görünmesine rağmen daha derin ve daha büyük bir modeldir. Diğer modellerden farklı olarak bu modelde iki farklı grafik işlemci ünitesi (GPU) arasında bir işleyiş görülmektedir. Modelin üst kısmını farklı bir GPU, alt kısmını farklı bir GPU eğitmektedir. Bu iki GPU sadece belirli katmanlarda iletişim kurmaktadır [16]. Bu durum modelin daha hızlı eğitilmesi için avantaj sağlamaktadır. Aktivasyon fonksiyonu olarak ReLU, aşırı öğrenmeyi engellemek için seyreltme yöntemi uygulanmıştır.

2.2.2 ResNet50

ResNet50 artık değerli nöral ağların bir kısaltmasıdır. Mimarisi evrişimli sinir ağlarının geliştirilmiş bir versiyonudur. Geleneksel ardışık ağ mimarisinden (AlexNet, VggNet gibi) farklı bir yapıya sahiptir. Bu mimaride başarım oranının daha yüksek seviyelere çıkması sağlamak için bazı katmanlar arasındaki değişim dikkate alınmayarak bir sonraki alt katmana geçiş işlemi yapılmıştır. Bu model, derin ağlar yakınsamaya başladığında evrişimli sinir ağlarının performans düşümü problemini çözmeyi amaçlamaktadır[17].

ResNet50 mimarisinde 177 katmandan oluşan bir ağ ile katmanlar arası bağlantıların nasıl olacağı hakkında bilgi bulunmaktadır[17].Giriş katmanı 224x224x3 boyutundadır.

2.2.3 MobileNet

Bu model, mobil ve gömülü görme uygulamaları için tanıtılmış olup derinlik açısından ayrılabilir evrişimlere dayanmaktadır [18]. MobileNet, diğer birçok popüler modelden çok daha küçük boyutta ve performans açısından daha hızlı olan, basit bir derin evrişimsel sinir ağı modelidir. Derinlik açısından ayrılabilir evrişimlerde her bir girdi kanalına tek bir filtre uygulanır. Ardından nokta bazlı evrişimde 1 × 1 evrişimler kullanılarak derinlik bazlı katmanların çıktısının doğrusal kombinasyonu oluşturulur[18]. Derinlik açısından ayrılabilir katmanlar, tipik evrişim katmanlarının işlemini neredeyse taklit eder. Ancak çok daha hızlı ve küçük bir farkla (tipik evrişim, çıktı özelliklerine hem filtre uygular hem de birleştirir, ancak bu işlem derinlik açısından ayrılabilir evrişimde iki katmana ayrılır, filtreleme için ayrı bir katman ve birleştirme için ayrı bir katman) işlem gerçekleştirilir. Bu yaklaşım modelin boyutunu en aza indirir ve hesaplama güç taleplerini azaltır. Çıktıyı sınıflandırmak için softmax katmanını besleyen tam bağlantılı katman dışında, tüm katmanlardan sonra bir batchnorm ve ReLU doğrusal olmayan aktivasyon katmanı gelir. MobileNet, derinlikli ve noktasal evrişimler haricinde 28 katmana sahiptir [18]. MobileNet mimarisi derinlemesine ayrılabilir evrişim işleminden faydalanarak parametre sayısını, işlem hacmini ve model karmaşıklığını düşürmüş, bu sayede donanımsal olarak kısıtları olan mobil ve gömülü cihazlarda kullanılabilir hale getirilmiştir[18].

2.2.4 SqueezeNet

CNN modelinde kullanılan bir diğer popüler mimari ise SqueezeNet mimarisidir [19]. SqueezeNet mimarisi Iandola ve ark. tarafından 2016 yılında sunulmuştur [20]. Bu mimarinin amacı daha az parametreye sahip bir sinir ağı oluşturmak ve mimari 50 kat daha az parametre ile AlexNet düzeyinde doğruluk sağlamaktadır [21]. SqueezeNet mimarisinin avantajı ise daha verimli dağıtılmış katmanlar sayesinde sinir ağındaki iş yükü azalmakta ve bu sayede daha hızlı çalışmaktadır [21, 22].

Bu mimari, 2016 yılında DeepScale, California Üniversitesi, Berkeley ve Stanford Üniversitesi'ndeki araştırmacılar tarafından önerilmiştir. Avantajları;

- Dağıtılmış eğitim sırasında sunucular arasında daha az iletişim gerektirir.
- Buluttan yeni bir model dışa aktarmak için daha az bant genişliği gerektirir.

•Sınırlı belleğe sahip özelleştirilmiş donanıma yerleştirmek daha uygundur. Büyük aktivasyon haritaları, aynı sayıda parametre verildiğinde daha yüksek bir sınıflandırma doğruluğu sağlar [23].

Yazarlar, doğruluğu en üst düzeye çıkarırken parametre boyutunu azaltmak için 3 ana stratejiyi özetlemektedir:

Strateji 1: 3x3 filtreleri 1x1 filtrelerle değiştirerek ağı daha küçük hale getirmek.1x1 filtre, 3x3 filtreden 9 kat daha az parametreye sahiptir.

Strateji 2: Kalan 3x3 filtreler için giriş sayısını azaltmak. Dönüşümlü katmanlara daha az giriş daha az parametreye neden olur.

Strateji 3: Katlama katmanlarının büyük aktivasyon haritalarına sahip olması için ağın sonlarına doğru örnek almak. Daha az sayıda parametreden en iyi şekilde yararlanmak ve doğruluğu en üst düzeye çıkarmaktır. Alt örneklemeyi ağın sonlarında geciktirmek, daha büyük etkinleştirme /özellik haritaları oluşturur[23].

2.2.5. Derin oto kodlayıcılar

Oto kodlayıcılar girdi verilerinin sıkıştırılmış hallerinden en iyi özelliklerin öğrenilmesini amaçlayan bir denetimsiz öğrenme için kullanılan ileri beslemeli yapay sinir ağıdır[24]. Çalışma prensibi, girişteki veri kodlama ve kod çözme süreçlerinden sonra aynı girişi çıktı olarak görünceye kadar ağırlıklar güncellenir ve amaca varıldığında gizli katmanda yer alan düğüm miktarı ile giriş verisi temsil edilmiş olur. Derin öğrenme de kullanılan derin oto kodlayıcılar ise oto kodlayıcıların çok katmanlı yapılarından meydana gelen ve her bir katmandaki çıkışların ardışık katmanın girişlerine bağlandığı bir yapay sinir ağı modelidir [25]. Literatürde, otomatik kodlayıcıların çeşitli versiyonları önerilmiştir. Bunlar arasında, Değişken Otomatik Kodlayıcı, Gürültü Arındıran Otomatik Kodlayıcı, Seyrek Otomatik Kodlayıcı, Ters Otomatik Kodlayıcı, duygu tanımada çok popüler ve kullanışlıdır [26]. Aykırı değer tespiti ve anormallik gibi durumlarda daha iyi sonuçlar verir [27,28].

Derin oto kodlayıcı, girdi dizisi ilk gizli katmanda şifrelenmiş dizi ile oluşur. Sonraki gizli katmana elde edilen özellik I dizisi girdi olarak verilmiş ve şifreleme sonucunda özellik II dizisi elde edilmiştir. Son olarak özellik II dizisi sınıflandırıcıya giriş yapılarak işlem sonlanmış olur. Şifrelemenin tersi olarak ta düşünülen şifrenin çözülme aşamasında gizli katmanlardan geçen şifrenin çözülerek çıktı olarak verilmesi sonucunda sona erir [29].

2.2.6. Karar ağaçları

Karar ağaçları hemen hemen her tür veriye adapte edilebilir olduğundan makine öğrenme algoritmaları içerisinde en popüler yöntemlerden biridir. Amaç, birkaç girdi değişkenine dayalı olarak bir hedef değişkenin değerini tahmin eden bir model oluşturmaktır. Veriler soyağacı gibi mantıksal yapı biçiminde herhangi bir istatiksel bilgi olmadan kolayca anlaşılabilir halde gösterilir. Bu metotta diğer sınıflandırma yöntemlerindeki gibi öğrenme ve sınıflandırma işlemlerinden oluşur. İlk aşama olan öğrenmede, sınıflandırma algoritması ile önceden etiketlenmiş eğitim verisi model oluşturmak amacıyla analiz edilir. Analiz sonucu öğrenilen bilgi bir ağaç şeklinde karar ağacı olarak gösterilir [30].

2.2.7. Destek vektör makinesi

Destek vektör makinesi (SVM) en eski sınıflandırıcılardan biridir. Bu sınıflandırıcı eğitildikten sonra sağladığı tahminlerin doğruluğu açısından belirgin bir performans gösterir. Destek vektör makinesi genellikle sınıflandırma sorunlarında kullanılan gözetimli öğrenme yöntemlerindendir. Bu sınıflandırma, iki sınıfının noktaları için de maksimum uzaklıkta olmasını için bir düzlem üzerine yerleştirilmiş noktaları ayırmak için bir doğru çizer. Küçük ve orta ölçekteki karmaşık veri setleri için uygundur. [31].

2.3. Gerçekleştirilen İşlemler ve Metodoloji

Öncelikle her bir dosyanın, dosya adında alttaki sıraya göre kaydolmuş bulunan ve "-" ile ayrılmış bulunan bilgilerini üst veri çalışması olarak dosya kısayolu ve kaynağı bilgisi de eklenerek değişkenlere atılmıştır. Sözlük (dict) değişkeni kolonları sırasıyla "path, source, actor, gender, intensity, statement, repetition, emotion" şeklindedir.Kolon izahları da aşağıdaki gibidir:

Modality (01 = full-AV, 02 = video-only, 03 = audio-only).

Vocal channel (01 = speech, 02 = song).

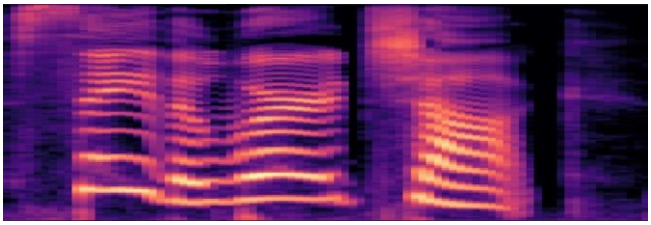
Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 =disgust, 08 = surprised).

Duygusal Yoğunluk (01 = normal, 02 = strong). "neutral" dediğimiz duygu tipinde "strong" şeklinde bir yoğunluk bilgisi bulunmamaktadır.

Kullanılan Cümle (statement) (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door").

Tekrar Miktarı (01 = 1.ci tekrar, 02 = 2.ci tekrar).

RAVDESS Dataseti için Aktör (Actor) (01'den 24'e kadar. Tekil numaralandırılmış aktörler erkek, çift numaralandırılmış aktörler kadın). TESS Dataseti için ise Aktör 25 ve Aktör 26 eklenmiştir. YAF aktör 25 olarak, OAF aktör 26 olarak isimlendirilmiştir. TESS datasetinde RAVDESS'ten 1 duygu eksiktir (calm), bu da uygun bir şekilde projeye dahil edilmiştir. Modality ve Vocal Channel bilgisi, bu projede sadece sesten duygu tanıma kullanılacağı için, ayırt edici bir nitelikte değildir, o yüzden sabit "video-only" ve "speech" alınmıştır. Alınan bu üst verilerle birlikte belirli bir örnekleme değeri (sample rate) belirlenerek spectrogramlar çizilerek her birinin resmi bir dosyaya kaydedilmiştir.



Şekil 1. TESS data setindeki "happy" duygu durumunun spektrogramı

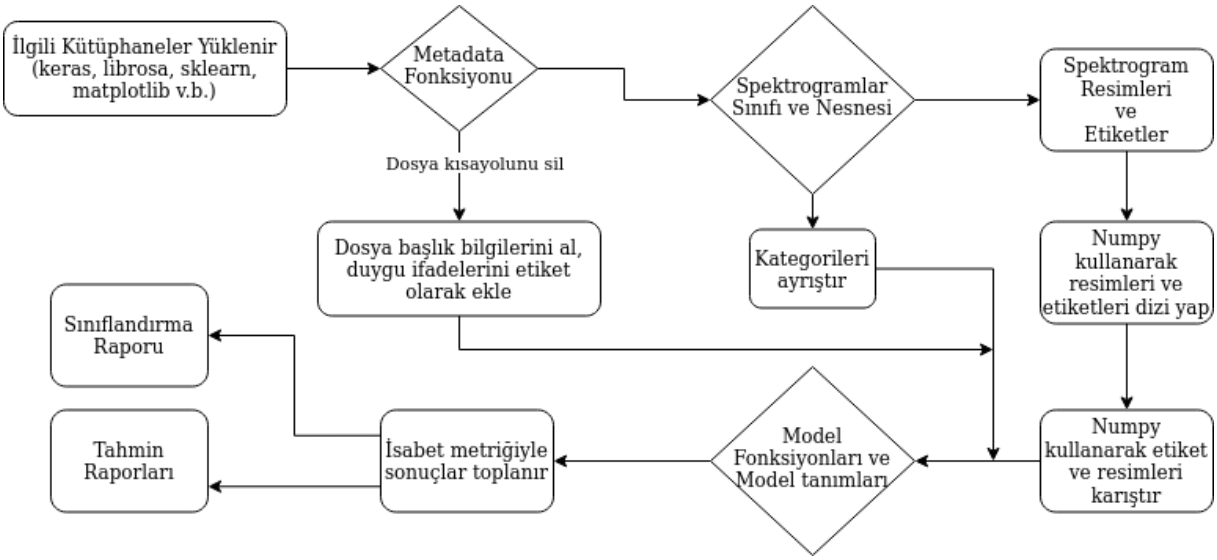
Kategorilerine göre klasörlere kaydedilmiş bulunan resimler tekrar images ve labels adında 2 liste değişkenine kaydedilmiştir. Bu images ve labels dizileri numpy dizilerine çevrilir, kalıplarının doğru alınıp alınmadığı çıktı alınarak kontrol edilmiştir. Resimler düz sırada alındığı için rastgele karıştırma işlemi yapılmıştır. Karıştırılan değişkenlerin kalıpları tekrar çıktı alınmıştır. Resimler normalizasyona tabi tutulmıştur. Daha sonra eğitim verileri ve test verileri rastgele olacak şekilde ayrıştırılarak, %20'si test verisi olacak şekilde ayrılmıştır. Sequential model üzerine hangi katmanlar uygulanacaksa tek tek eklenmiştir. Modelin özetine göz gezdirilerek model derlenmiştir. Model 100 devir boyunca eğitim verileriyle eğitilmiştir. Test verileriyle modelden tahmin istenerek sınıflandırma raporu alınmıştır. Karıştırma matrisi çıktısı alınmıştır. Modelle ilgili eğitim verisinin ve test verisinin resmi alınmıştır.

RAVDESS[14] dosyaları tek tek Python librosa "kaiser_fast" yeniden örnekleme (re-sample) dizisine yüklenerek MFCC matrisleri elde edilmiş ve MFCC bilgileri numpy kütüphanesi aracılığıyla yeni bir diziye aktarılmıştır. Bu işlemle ilgili 800 saniye - 2200 saniye gibi çeşitli çalışma süreleri gözlemlenmiştir. Bu ana dizi, numpy ile x ve y alt dizilerine ayrıştırılmıştır. Eğitim verileri ile test verileri % 33'ü test verisi olacak şekilde ayrıştırılmıştır.

Oto kodlayıcı için öncelikle bir model daha oluşturulmuştur. Bu model karıştırılmış x eğitim verisiyle 100 döngü boyunca eğitilmiş ve modelin sıkıştırılmış ses dosyaları programın dizinine kaydedilmiştir. Model verileri aşağıdaki gibidir.

Optimal makine öğrenmesi ve/veya derin öğrenme ağını bulabilmek amacıyla SqueezeNet ve Mobilenet ağları da denenmiştir. AlexNet ağı için oluşturulan spectrogramlardan burada da faydalanılmıştır. SqueezeNet ve MobileNet, AlexNet'ten farklı olarak Resnet50'nın resim kalıbını kullandığı için bu iş için oluşturulan "Resimler ve Etiketler" fonksiyonundan Resnet50 bölümünü kullanılmıştır. Öncelikle RAVDESS data seti kullanılmıştır. Diğerlerinde olduğu gibi resimler ve etiketler değişkenlere alınmıştır (bu yöntemin çok sistem kaynağı, mesela RAM tükettiği not edilmiştir). Bunlar sayı işlemlerinde ön plana çıkan NumPy kütüphanesi dizilerine çevrilmiştir. 2459 adet resim ve etiket yüklenmiştir. Resim kalıbı 224, 224, 3'tür. (Alexnet'te 227, 227, 3 idi). Resimler ve etiketler random karıştırma (shuffle) metodlarına tabi tutulmuştur. 1967'si eğitim, 492'si test verisi olarak ayrılmıştır. SqueezeNet model yapısı oluşturulmuş, model değişkenine alınmıştır. MobileNet için de daha sonra aynı işlem yapılmıştır.

Çalışmamızdaki kullandığımız metodolojiyi aşağıda Şekil.2 'deki gibi özetlenmiştir.



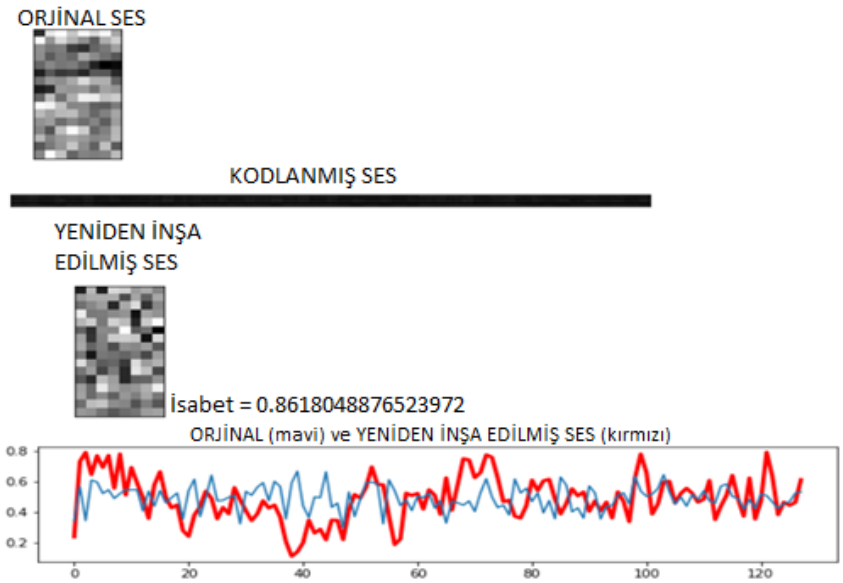
Şekil 2. Derin öğrenme duygu tanıma işlem adımları

3. MODELLERİN UYGULANMASI

RAVDESS data seti için Karar Ağacı sınıflandırıcısında; x ve y eğitim verileri sınıflandırma işlemine tabi tutulmuştur. X test verisinden tahminler üretip bunları y test verisiyle karşılaştırması istenmiştir. Karar Ağacı tahminlerinden 0.44 oranında başarım elde edilmiştir.

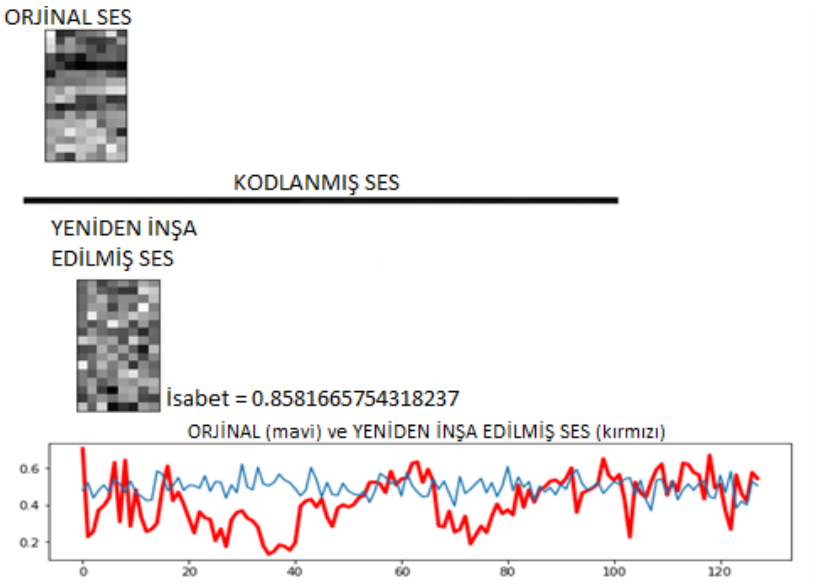
SVM Sınıflandırıcısında ; x ve y eğitim verileri işleme alınmıştır. X test verisinden tahminler üretip bunları y test verisiyle karşılaştırması istenmiştir. Sonuç 0.29679802955665024 isabet olarak elde edilmiştir. Sonuçlar yeterli görülmeyerek SVM Sınıflandırıcısı RAVDESS data setine ek olarak TESS data seti de eklenmiş, veri setinin çeşitlendirilmesi ve genişletilmesinin nasıl bir sonuç vereceği gözlemlenmeye çalışılmıştır. Sonuçlarda ciddi manada gibi bir iyileşme gözlemlenmiştir. Daha önce ~0.2968 civarında gözlemlenen sonuç isabet oranının 0.5357554786620531'e çıktığı not edilmiştir.

Alexnet-CNN Modelinde; x için eğitim verileri 1648 ve test verileri 812 olarak, y için ise eğitim verileri 1648 ve test verileri 812 olarak alınmıştır. Tensorflow'dan keras yüklenmiştir. Keras'tan da Sequential modeli yüklenmiştir. Sequential modeline ağlar eklenmiştir ve modelin özeti elde edilmiştir. 500 döngü boyunca x ve y verileri bu modele uygulanarak sonuçlar alınmıştır. Ortalama isabet 0.16379310190677643 olarak alınmıştır. Sonuç yeterli görülmeyerek RAVDESS + TESS dataseti birlikte eğitim ve test verileri olarak kullanılmış, isabet oranı 0.2116493582725525 olarak not edilmiştir. Elde edilen bu nihai veri seti ve modele Oto Kodlayıcı modeli uygulanarak yeniden kodlanmış ses dosyalarıyla eskilerin uyumu 0.8618048876523972 olarak elde edilmiştir.



Şekil 3. Bu modelde eğitilmiş rastgele seçilmiş tekil RAVDESS ses dosyası (Orjinali ve yeniden oluşturulmuş hali birlikte)

Rastgele seçilmiş üstteki tekil RAVDESS ses dosyası çıktısının karşılığı olarak RAVDESS+TESS data setinde eğitilmiş bir tekil veri de altta gösterilmiştir.

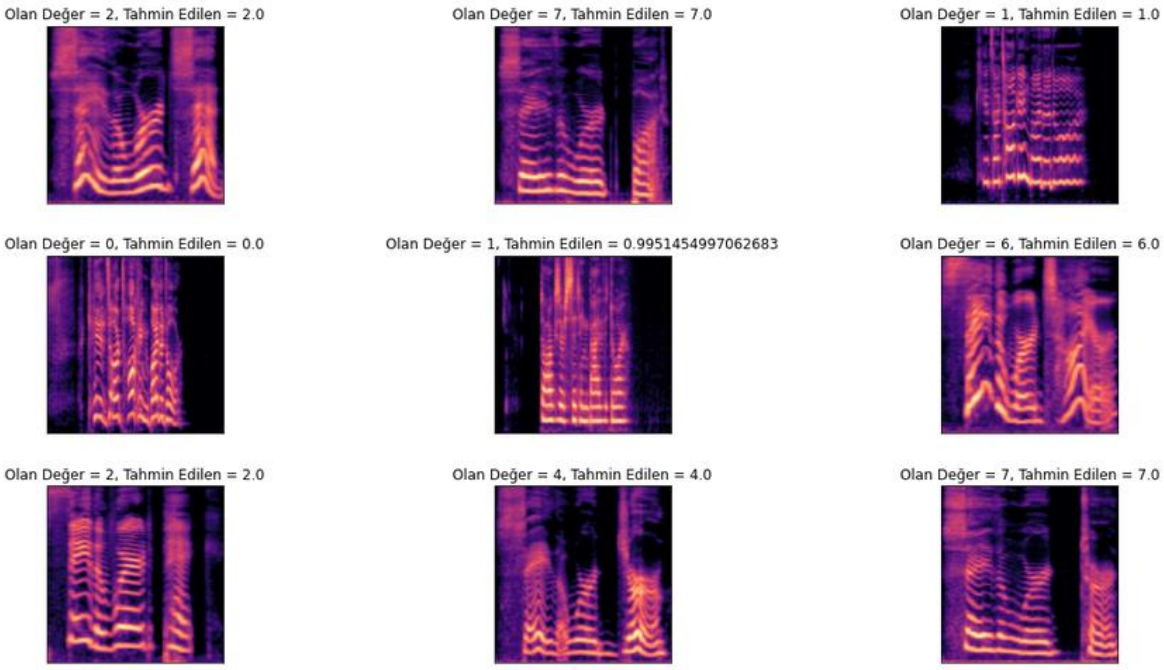


Şekil 4. Bu modelde eğitilmiş rastgele seçilmiş tekil RAVDESS + TESS ses dosyası (Orjinali ve yeniden oluşturulmuş hali birlikte)

İki veri arasındaki isabet oranındaki artış not edilmiştir. Yeniden kodlanmış verilerden tekrar eğitim ve test verileri % 33 esasına göre ayrıştırılmıştır. Bu yeniden kodlanmış verilerden elde edilen eğitim ve test verileri SVM modeline uygulanmıştır. X eğitim verilerinden tahmin alınıp y eğitim verileriyle karşılaştırılıp 0.3657635467980296 doğruluk oranı elde edilmiştir. Yeniden elde edilmiş ses dosyaları Karar Ağacı modeline uygulanarak %25'lik başarım elde edilebilmiştir. Yeniden elde edilmiş bu veriler Ardışık (Sequential) modeline de tekrar uygulanmıştır. Model 1000 döngü boyunca uygulanmış, isabet oranı 0.1391625553369522 olarak elde edilmiştir. Aynı model RAVDESS + TESS datasetine uygulanarak 0.1412918120622635 isabet oranı elde edildiği not edilmiştir.

SqueezeNet ağında veriler eğitildikçe (toplam 100 eğitim döngüsü kullanılmıştır) isabet oranlarının 20. döngüden itibaren hızla arttığı görülmüştür. Henüz 90. döngüye gelmeden tam isabet oranına ulaşılmıştır. MobileNet ağında ise isabet oranı %15 ilâ %20 arasında kalmıştır. Veri setinin çeşitlendirilmesi ve genişletilmesi kapsamında; RAVDESS datasetine (SVM, Decision Tree, CNN örneklerinde olduğu gibi) TESS dataseti eklenmiş, datasetinin boyutu 5252 resime ulaşmıştır. Veri seti karıştırılmış ve 4201'i eğitim, 1051'i test verisi olarak ayrılmıştır. Neticede SqueezeNet ağındaki düzelme henüz 13. döngüde başlamış, 63. döngüde tam isabet oranına ulaşılmıştır.

Aşağıda birkaç ses dosyasına dair analizlere yer verilmiştir.



Şekil 5. Mobilenet rastgele ses dosyaları spektrogramları, orijinal değeri ve modelin tahminleri birlikte.

MobileNet ağında ise veri setinin çeşitlendirilmesi ve genişletilmesi beklenen iyileşmeyi vermemiş, %1'lik bir iyileşme ile isabet oranı 0.15344657003879547 seviyesinde kalmıştır.

Evrişimli Sinir Ağlarında ve Karar Ağacı sınıflandırıcısında elde edilen sonuçlar Tablo 1'de toplu olarak verilmiştir.

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

Özellikler	Evrişimli Sinir Ağı	Sınıflandırıcı	Veri Seti	Eğitim / Test	Başarım
MFCCler, Spektrogramlar	Alexnet	Autoencoder	TESS	%20 Test	%98
MFCCler, Spektrogramlar	Resnet50	Autoencoder	TESS	%20 Test	%90,6
MFCCler, Spektrogramlar	-	Decision Tree	Ravdess+Tess	%33 Test	%64
MFCCler, Spektrogramlar	SqueezeNet	-	Ravdess+Tess	%20 Test	%81,4

4. SONUÇ

Bazı Evrişimli Sinir Ağları'nda sonuçların %15 isabet seviyesinde kalması, bazılarında %45-%60 civarında isabet elde edilmesi, bazılarında da henüz istenen döngü tamamlanmadan tam isabet oranına ulaşılması duygu tanıma işi için uygun ağın seçilmesi gerektiğini net bir şekilde göstermiştir. Özellikle çeşitli Sinir Ağları'nda Autoencoder'ların başarım oranını %90'ların üzerine çıkardığı tespit edilmiştir.

Ses işleme ve çoklu sınıflandırma için veri setinin doğru bir şekilde etiketlendirilmiş olması, doğru kurgulanmış ön işleme safhası, spektrumlar çıkartılırken hedef sinir ağının veya makine öğrenmesi sınıflandırıcısının girdi kalıplarının dikkate alınması, hangi sinir hücresine ne verinin gelmesi gerektiği ve ne şekilde öğreneceğinin kurgulanarak verinin buna uydurulması kritik önemdedir.

Sonuçları alınmamış ve sonuç tablosunda gösterilmemiş olan modellerdeki eğitim ve test süreçlerinde tekrar çalışmanın gerekli olmadığı, ancak yüksek başarım gösteren SqueezeNet sinir ağının ve yüksek başarım gösterme ihtimali bulunan Karar Ağacı makine öğrenmesi sınıflandırıcısının daha yüksek performanslı bilgisayar sistemlerinde tekrar, daha uzun eğitim süreçleriyle çalıştırılması gerektiği görülmüştür.

KAYNAKÇA

1. Aziz A. İletişime Giriş, Hiperlink Yayınları, pp.4-256, 2016.
2. Akleylek S., Kılıç E., Söylemez B., Ergun A. R. U. K., Aksaç C. Nesnelerin interneti tabanlı sağlık izleme sistemleri üzerine bir çalışma, Mühendislik Bilimleri ve Tasarım Dergisi, *8*(5), 80-89, 2020.
3. El Ayadi M., Kamel M. S., Karray F. Survey on speech emotion recognition: Features, classification schemes, and databases, Pattern Recognition, 44 (3), 572-587, 2011.
4. Hızlısoy S., Tüfekci Z. Türkçe müzikten duygu tanıma, Avrupa Bilim ve Teknoloji Dergisi, 6-12, 2020.
5. De Pinto M. G., Polignano M., Lops P., Semeraro G. Emotions understanding model from spoken language using deep neural networks and mel-frequency cepstral coefficients, In 2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS),pp. 1-5, IEEE, 2020.
6. Tarantino L., Garner P. N., Lazaridis, A. Self-attention for speech emotion recognition, In Interspeech, 2578-2582, 2019.
7. Triantafyllopoulos A., Keren G., Wagner J., Steiner I., Schuller B. W. Towards robust speech emotion recognition using deep residual networks for speech enhancement, In Interspeech, 1691-1695, 2019.
8. Zhao J., Mao X., Chen L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks, Biomedical Signal Processing and Control, 47, 312-323, 2019.
9. Chatziagapi A., Paraskevopoulos G., Sgouropoulos D., Pantazopoulos G., Nikandrou M., Giannakopoulos T., Katsamanis A., Potamianos A., Narayanan S. Data augmentation using gans for speech emotion recognition, In Interspeech 171-175, 2019.
10. Hossain M. S., Muhammad G. Emotion recognition using deep learning approach from audio–visual emotional big data, Information Fusion, 49, 69-78, 2019.
11. Iqbal A., Barua K. A real-time emotion recognition from speech using gradient boosting, In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), IEEE, pp. 1–5, 2019.
12. Salur M. U., Aydın İ. Sentiment classification based on deep learning, In 2018 26th Signal Processing and Communications Applications Conference (SIU), pp. 1-4, IEEE, 2018.
13. Salur M. U., Aydın İ. A novel hybrid deep learning model for sentiment classificatio,. IEEE Access, 8, 58080-58093, 2020.
14. https://zenodo.org/record/1188976, Erişim Tarihi: 25 Aralık 2020.
15. https://tspace.library.utoronto.ca/handle/1807/24487 , Erişim Tarihi: 25 Aralık 2020.
16. Krizhevsky A., Sutskever I., Hinton G. E. Imagenet classification with deep convolutional neural networks, In Advances In Neural Information Processing Systems, 2012.
17. He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition, In Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778, 2016.
18. Howard A. G., Zhu M., Chen B., Kalenichenko D., Wang W., Weyand T., Andreetto M., Adam H. MobileNets: Efficient convolutional neural networks for mobile vision applications, 2017.
19. Zavan F.H.D.B., Bellon O.R.P., Silva L., Medioni G.G. Benchmarking parts based face processing in-the-wild for gender recognition and head pose estimation, Pattern Recognition Letters,123,104-110, 2019.
20. Iandola F.N., Han S., Moskewicz M.W., Ashraf K., Dally W.J., Keutzer K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size, 3th International Conference on Learning Representations. Toulon: ICLR;2016. pp.1-13, 2016.
21. Özyurt F., Sert E., Avcı D. An expert system for brain tumor detection: Fuzzy C-means with super resolution and convolutional neural network with extreme learning machine, Medical Hypotheses,134,1-8, 2020.
22. Pathak D., El-Sharkawy M. ReducedSqNet: A shallow architecture for CIFAR-10, In 2018 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas:IEEE,pp. 380-385, 2018.
23. Mateen M., Wen J., Song S.N., Huang Z. Fundus image classification using VGG-19 architecture with PCA and SVD, Symmetry, 2019.
24. Krizhevsky A., Hinton G. E. Using very deep autoencoders for content-based image retrieval, In ESANN, 1, pp. 2, 2011.
25. Şeker A., Yüksek A.G. Stacked autoencoder method for fabric defect detection, Cumhuriyet Üniversitesi Fen-Edebiyat Fakültesi Fen Bilimleri Dergisi, 38(2), 342-354, 2017.
26. https://www.mdpi.com/1424-8220/21/4/1249 , Erişim Tarihi: 10 Şubat 2021.
27. Lyudchik O. Outlier detection using autoencoders, 2016.
28. Yadav S. Subramanian S. Detection of application layer DDoS attack by feature learning using stacked autoencoder, In 2016 International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT), pp. 361-366, IEEE, 2016.
29. Canchumuni S. W., Emerick A. A., Pacheco M. A. C. Towards a robust parameterization for conditioning facies models using deep variational autoencoders and ensemble smoother, Computers & Geosciences, 128, 87-102, 2019.
30. https://medium.com/@k.ulgen90/makine-%C3%B6%C4%9Frenimi-b%C3%B6l%C3%BCm-5-karar-a%C4%9Fa%C3%A7lar%C4%B1-c90bd7593010, Erişim Tarihi: 15 Şubat 2021.
31. https://medium.com/deep-learning-turkiye/nedir-bu-destek-vekt%C3%B6r-makineleri-makine-%C3%B6%C4%9Frenmesi-serisi-2-94e576e4223e, Erişim Tarihi: 15 Şubat 2021.