Avrupa Bilim ve Teknoloji Dergisi
Sayı 54, S. 46-54, Aralık 2024

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
No. 54, pp. 46-54, December 2024

Research Article

Ses Kaydı Analizine Dayalı Müşteri Duygu Sınıflandırması İçin CNN Hiper Parametre Optimizasyonu

Taner Hacıoğlu[1] , Sina Apak[2]

[1] Bilgisayar Mühendisliği, Lisansüstü Eğitim Enstitüsü, İstanbul Aydın Üniversitesi, Türkiye (ORCID: 0000-0001-7190-2675), thacioglu@stu.aydin.edu.tr

[2] Yönetim Bilişim Sistemleri, Uygulamalı Bilimler Fakültesi, İstanbul Aydın Üniversitesi, Türkiye (ORCID: 0000-0002-7923-5253), sinaapak@aydin.edu.tr

(İlk Geliş Tarihi 28 Kasım 2023 ve Kabul Tarihi 25 Mart 2024)

(**DOI:** 10.5281/zenodo.14176095)

ATIF/REFERENCE: Hacıoğlu, T. & Apak, S. (2024). Ses Kaydı Analizine Dayalı Müşteri Duygu Sınıflandırması İçin CNN Hiper Parametre Optimizasyonu. *Avrupa Bilim ve Teknoloji Dergisi*, (54), 46-54.

Öz

Çağrı merkezleri, müşteri hizmetleri sunan şirketler için iletişim kanalları olarak kritik bir rol oynamakta olup, gelen çağrıların etkili bir şekilde yönetilmesi müşteri memnuniyeti ve şirket performansı açısından önemlidir. Çağrı merkezi performansının değerlendirilmesi ve iyileştirilmesi için veri analitiği ve yapay zeka tekniklerinin kullanımı giderek yaygınlaşmaktadır [6]. Bu makale, müşteri hizmetlerine odaklanan tekstil ürünleri alanında faaliyet gösteren Gurmen Textiles şirketine ait bir çağrı merkezi veri setinin detaylı bir analizini sunmaktadır. Şirket, çağrı merkezini müşteri sorularına cevap vermek ve müşteri memnuniyetini maksimize etmek amacıyla kullanmaktadır. Bu çalışmada, Gurmen Textiles'in çağrı merkezi veri seti kullanılarak, öfkeli ve normal çağrıları ayırt etmek için bir derin öğrenme modeli geliştirilmiştir. Öfkeli çağrılar, müşteri memnuniyetsizliği veya sorunları işaret edebilirken, normal çağrılar genellikle rutin müşteri taleplerini içermektedir. Öfkeli ve normal çağrıların doğru bir şekilde sınıflandırılması, şirketin müşteri ilişkilerini yönetmesine ve hizmet kalitesini arttırmasına yardımcı olabilir. Veri analitiği ve derin öğrenme tekniklerini kullanarak, çağrı merkezi verilerinin analizi daha verimli bir şekilde gerçekleştirilebilir, bu da müşteri memnuniyeti üzerinde olumlu etkiler yaratabilir. Bu analizlerde yaygın olarak kullanılan konuşma analizi, çağrı merkezi verilerindeki ses kayıtlarını değerlendirerek müşteri duygusal durumları veya endişeleri hakkında değerli içgörüler sağlayabilir. Önerilen yöntem, SAVEE genel erişim veri setinde %98 ve Gurmen Tekstil veri setinde %97 sınıflandırma doğruluğu ile uyumlu sonuçlar elde etmektedir.

Anahtar Kelimeler: Evrişimli Sinir Ağı, Ses Sınıflandırma, Spektogram Dönüşümü, MFCC

CNN Hyperparameter Optimization for Customer Emotion Classification Based on Voice Recording Analysis

Abstract

Call centers play a crucial role as communication channels for companies providing customer services, and efficient management of incoming calls is essential for customer satisfaction and company performance. The use of data analytics and artificial intelligence techniques for evaluating and improving call center performance is increasingly prevalent. This article presents a detailed analysis of a call center dataset belonging to Gurmen Textiles, a company specializing in textile products with a strong focus on customer service. The company utilizes its call center to address customer inquiries and aims to maximize customer satisfaction. In this study, a deep learning model was developed using the call center dataset of Gurmen Textiles to differentiate between irate and normal calls. Irate calls may indicate customer dissatisfaction or issues, while normal calls typically encompass routine customer requests. Accurate classification of irate and normal calls can assist the company in managing customer relationships and enhancing service quality. By leveraging data analytics and deep learning techniques, analyses of call center data can be conducted more efficiently, leading to positive impacts on customer satisfaction. Speech analysis, a commonly employed technique in such analyses, can provide valuable insights into customer emotional states or concerns by evaluating audio recordings in call center data. The proposed approach achieves compatible results in SAVEE public access dataset with %98 and significant result in Gurmen Textiles dataset with %97 accuracy of classification.

Keywords: Convolutional Neural Network, Sound Classification, Spektogram Transformation, MFCC

1. Giriş

Çağrı merkezleri, müşteri hizmetleri sunan şirketlerin en önemli iletişim kanallarından biridir ve müşteri memnuniyeti ve şirket performansı açısından kritik bir rol oynamaktadır. Gelen çağrıların yönetimi, etkili bir şekilde gerçekleştirilmeli ve müşterilerin ihtiyaçlarına hızlı ve etkili bir şekilde yanıt verilmelidir. Bu nedenle, çağrı merkezlerinin performansını değerlendirmek ve iyileştirmek için veri analitiği ve yapay zeka teknikleri kullanımı giderek artmaktadır [6].

Bu makalede, Gürmen Tekstile ait bir çağrı merkezi veri seti üzerinde gerçekleştirilen analizden bahsedilecektir. Gürmen Tekstil, tekstil ürünleri üreten ve müşteri hizmetlerine büyük önem veren bir şirkettir. Şirket, çağrı merkezi aracılığıyla müşteri taleplerini karşılamakta ve müşteri memnuniyetini en üst düzeye çıkarmayı hedeflemektedir.

Bu çalışmada, Gürmen Tekstile ait çağrı merkezi veri seti kullanılarak sinirli ve normal çağrıları ayırmak için bir yapay zeka modeli geliştirilmiştir. Sinirli çağrılar, müşteri memnuniyetsizliği veya sorunları gösterebilirken, normal çağrılar genellikle sıradan müşteri taleplerini içermektedir. Bu nedenle, sinirli ve normal çağrıları doğru bir şekilde sınıflandırmak, şirketin müşteri ilişkileri yönetiminde ve hizmet kalitesinde iyileştirmeler yapmasına yardımcı olabilir.

Veri analitiği ve yapay zekâ tekniklerinin kullanımıyla, çağrı merkezi verileri üzerindeki analizler daha verimli bir şekilde gerçekleştirilebilir ve müşteri memnuniyeti üzerinde olumlu etkiler sağlanabilir. Ses analizi, bu tür analizlerde yaygın olarak kullanılan bir tekniktir ve çağrı merkezi verilerindeki ses kayıtlarının değerlendirilmesi, müşteri duygusal durumunu veya endişelerini anlamak için önemli bir bilgi kaynağı olabilir.

Bu makalede, Gürmen Tekstile ait çağrı merkezi veri seti üzerinde gerçekleştirilen analizde, sinirli ve normal çağrıları ayırmak için ses dosyaları kullanılmıştır. Ses dosyalarının analizi için ses işleme ve derin öğrenme teknikleri kullanılarak bir yapay zeka modeli geliştirilmiştir. Oluşturulan model, ses dosyalarını sınıflandırmak ve sinirli ve normal çağrıları ayırt etmek için eğitilmiştir.

Bu çalışmanın amacı, Gürmen Tekstile ait çağrı merkezi verilerini analiz etmek ve sinirli ve normal sesleri ayırmak için bir yapay zeka modeli geliştirmektir. Elde edilen sonuçlar, şirketin çağrı merkezi performansını değerlendirmek ve müşteri memnuniyetini artırmak amacıyla kullanılabilir.

Makalenin devamında, veri ön işleme adımları, kullanılan yöntemler, elde edilen sonuçlar ve değerlendirme bulguları ayrıntılı bir şekilde açıklanacaktır. Ayrıca, yapay zekâ modelinin etkinliği ve uygulanabilirliği üzerine tartışmalar sunulacak ve ilgili sektörlerde benzer analizlerin yapılması için önerilerde bulunulacaktır.

1.1. Literatür Araştırması

Bu çalışmada [1], Fransız acil servisini arayan hastaların Öfke, Korku, Olumlu ve Normal (Nötr) durumları incelenmiştir. Çalışmada 485 arayan ve 10 farklı çağrı merkezi temsilcisiyle yapılan görüşmelere odaklanılmıştır. Sınıf sayısı dört, üç ve ikiye indirgenmiştir. Sınıflandırma sayısı azaldıkça başarımın arttığını görmüşlerdir. Sonuçlar Dört sınıf için %45,6 üç sınıf için %54,4(Olumsuz, Olumlu, normal), iki sınıf için (olumsuz ve normal) %77,5 olarak ortaya çıkmıştır. Umut vadeden bu çalışmada gerçek hayatta konuşmacıların duygu çeşitliliklerin fazla olması nedeniyle net sahnelenen duygulara göre tespitinin daha karmaşık olduğu belirtilmiştir. Gelecekte Evrişimli sinir ağları yanında dil çeviri özellikleri kullanılarak çok modlu bir mimari yapısını kendilerine hedef seçmişlerdir.

Benzer şekilde bir başka çalışmada[2], veri artırma (data augmentation) konuşma duygu tanıma için 2D CNN önermektedir. EMODB veritabanı, önerilen modelin artırılmış veri ile değerlendirilmesi için kullanılmıştır. Önerilen model, hassasiyet, geri çağırma ve F1-skoru ile değerlendirilmiştir. Önerilen model, konuşma duygu tanıma alanındaki diğer mevcut şemalarla karşılaştırılmıştır. Deneysel sonuçlar, 2D CNN modeli ve veri artırmanın birleşiminin konuşma duygu tanıma doğruluğunu artırabileceğini göstermektedir. Az miktarda veri ile derin öğrenme yöntemi için veri artırmanın önemli olduğu özetlenebilir. Önerilen model, %88 doğruluk değeri elde etmiştir. Gelecekteki çalışmalarda, konuşma duygu özelliklerini çıkarmak için 3D-CNN'nin uygulanması düşünülebilir.

Başka bir sağlam duygusal tanıma çalışmasında[3], çağrı merkezi (CC) telefon hattındaki konuşmacıların duygularını tanıma için yeni bir yöntem önermektedir. Yöntem, duygusal durumların hem metin (sohbet türü) hem de ses kanallarında tanınmasını varsayar. Önerilen çözüm, çağrı merkezini arayan müşterilerin duygularını tanıma yeteneği sağladığı gibi, bu çağrıları işleyen temsilcilerin duygularını da tanıma yeteneği sunar, bu da pratik uygulamalar için önemlidir. Hazırlanan sınıflandırma modelleri için yapılan doğrulama deneylerinde, metin kanalındaki duygu tanıma sonuçları, ajan ifadeleri için %70'in üzerinde ve müşteri ifadeleri için %60'a kadar değerlere ulaşmaktadır. Ses kanalında, konuşma transkripsiyonu ve sistemde entegre edilen sözlük göz önüne alındığında, CNN sınıflandırıcısı için her iki kanalda da %68'in üzerinde değerlere ulaşmaktadır.

Derin öğrenmeye dayalı başka bir benzer çalışma[4],çağrı merkezi çalışanları ile müşteriler arasındaki telefon konuşmalarının otomatik olarak olumlu veya olumsuz şeklinde değerlendirilmesi üzerine odaklanılmıştır. Çalışmada kullanılan veri seti firma bünyesinde gerçekleştirilen telefon görüşmelerinden oluşmaktadır. Veri seti üçer saniyelik 10411 adet ses kaydını içermekte olup bu kayıtların 5408 tanesi olumlu kayıtlardan 5003 tanesi münakaşa, öfke ve hakaret içeren olumsuz kayıtlardan oluşmaktadır. Çağrı merkezi kayıtlarından duygu tanıma için anlamlı öznitelikler elde etmek amacıyla her bir ses kaydından MFCC öznitelikleri çıkarılmıştır. Çağrı merkezi kayıtlarını olumlu olumsuz olarak sınıflandırmak için önerilen CNN mimarisi MFCC öznitelikleriyle eğitilmiştir. Önerilen CNN modeli %86,1 eğitim başarısı, %77,3 doğrulama başarısı göstermiş olup test verileri üzerinde %69,4 sınıflandırma başarısı elde edilmiştir.

Benzer bir başka tezde[5], duygu tanımanın önemli bir konu olduğunu ve ses verilerinin duygu tanıma işleminde kullanılabileceğini vurgulamaktadır. Berlin duygu veri tabanı (EmoDB) üzerinde gerçekleştirilen çalışmada, cinsiyet ve kişi bağımsızlığı gözetilerek üç farklı uygulama yapılmıştır. Çalışmada, özellik çıkarımı için spektral, prozodik ve format özellikleri kullanılmıştır. Sınıflandırma için Yapay Sinir Ağları, Destek Vektör Makineleri, k En Yakın Komşuluk Algoritması ve Sade Bayes algoritmaları kullanılmıştır. Ayrıca, özellik seçimi için Etmen Tabanlı Otomatik Özellik Seçimi yaklaşımı, Bulanık C-Ortalama Algoritması (BCO) ve Derin Öğrenme Algoritmaları (AlexNet) ile sınıflandırma yapılmıştır. Elde edilen sonuçlarda, MFCC katsayılarından oluşturulan veri kümesi ile %92.98'lik en yüksek sınıflandırma doğruluğuna BCO yöntemiyle ulaşıldığı belirtilmiştir. Bu çalışma, duygu tanıma probleminde etkin özelliklerin belirlenmesi ve sınıflandırma konularında önemli katkı sunmaktadır.

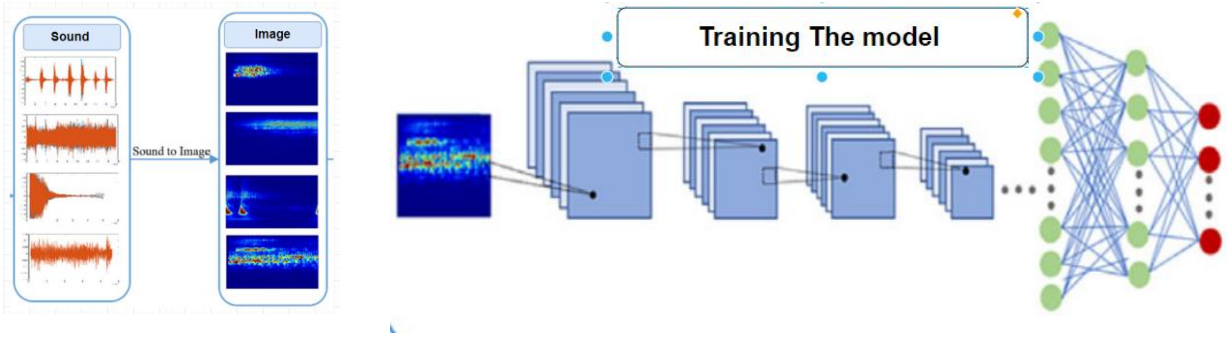
Güncel bir üniversite çalışmasında ise[16], Konuşma Duygu Tanıma (SER), alanındaki büyük gelişmeleri inceleyerek derin öğrenmenin bu alandaki önemli etkisini vurgular. SER, insan konuşmalarından duygusal durumları tanımlama amacını taşıyan bir İnsan-Bilgisayar Etkileşimi (HCI) dalıdır. Ancak, araştırmanın ilerlemesine rağmen, SER hala gerçek dünya uygulamalarında kullanılabilirliği destekleyecek nitelikli veri setlerinin ve en iyi uygulama pratiklerinin eksikliği nedeniyle çeşitli zorluklarla karşı karşıyadır. Çalışma, birleştirilmiş bir CNN modeli ve birikimli dikkat bloğu içeren bir model kullanarak SER uygulamalarında yaygın olarak kullanılan dört İngiliz veri seti üzerinde bir dizi deneme gerçekleştirir: RAVDESS, TESS, CREMA-D ve IEMOCAP. Her bir veri seti üzerinde yapılan testler, önerilen modelin sırasıyla %83, %100, %68 ve %63 ortalama doğruluk elde ettiğini göstermektedir.

Güncel başka bir pilot çalışmada[17], Konuşma duygu tanıma(SER) konusunda bir konvolüsyonel sinir ağı algoritması kullanarak gerçek zamanlı bir web tabanlı uygulamayı tanıtmaktadır. Çalışmanın temel amacı, kullanıcı dostu bir tasarıma sahip güvenilir bir araç geliştirerek konuşmadaki duyguları tahmin etmektir, bu sayede tanıma sonuçlarına kolay erişim ve görüntüleme imkânı sağlamaktadır. Platform, öfke, iğrenme, korku, mutluluk, tarafsız, üzüntü ve şaşkınlık olmak üzere yedi farklı duyguyu tanıyabilmektedir ve statik ile gerçek zamanlı konuşma sinyalleri analizi olmak üzere iki temel işlevselliğe sahiptir. Statik analiz, kullanıcıların önceden kaydedilmiş ses dosyalarını yüklemelerine olanak tanırken, gerçek zamanlı analiz, ses kaydedilirken sürekli ses işleme sağlamaktadır. Çalışma ayrıca, minimal özelliklere sahip olmasına rağmen duyguları doğru bir şekilde tanıyabilen güvenilir bir model geliştirmeye odaklanmaktadır. Modelin algoritmik performansı, genel olarak kullanılabilir veri setleri (RAVDESS, TESS ve SAVEE) kullanılarak değerlendirilmiştir ve seçilen spektral özellik olan MFCC kullanılarak statik analizde %86,46'lık bir doğruluk oranına ulaşılmıştır. Gerçek zamanlı analizin performansı, 20 katılımcı içeren bir kullanıcı çalışması aracılığıyla doğrulandı ve muhtemel bilinen faktörlere bağlı olarak gerçek zamanlı duygu tanıma başarısı %65 olarak belirlendi.

Yukarıda belirtilen araştırmalara dayanarak, bu projede ses sinyallerini görüntülere dönüştürmek için MFCC kullandık. Daha sonra, eğitim verisi sayısını artırmak amacıyla KERAS kütüphanesinde veri büyütme fonksiyonunu uyguladık. Bu yöntem, aşırı uyum ve eksik uyum sorunlarının önüne geçmeye yardımcı oldu. Ardından, etiketli görüntülerle birlikte CNN'i sıfırdan eğitimden geçirerek çalışmayı sonuçlandırmış olduk.

2. Materyal ve Metot

Bu projede çağrı merkezi verileri yerel veri tabanına wav formatında kaydedilmektedir. Konuşma içeren veriler için pydub kütüphanesinden ses segmentasyonu kullandık. Belirli sabit 10 saniyelik bölümlenmiş sinyaller MFCC yöntemi yardımıyla RGB görüntülere dönüştürülmüştür [7]. Uzman çağrı merkezi çalışanı, çalışanlarla müşteriler arasındaki konuşmayı normal ve kızgın olmak üzere iki kategoriye göre etiketliyor. Etiketler ve görüntüler, modelleri eğitmek için evrişimli sinir ağını besler[8]. Önerilen sistemler aşağıdaki şekilde sunulmuştur.



Şekil 1. ESA ile seslerin sınıflandırılması (**Figure 1.** Classification of sounds with CNN)

Firma ve public veri setlerinde MFCC öznitelikleri, ses verisinden derin öznitelikleri ortaya çıkarmak için kullanılmıştır ve bu öznitelikler, 128x128x3 boyutunda bir veri matrisi olarak modele girdi hizmeti sunmaktadır. Modellerdeki veri eşitsizliğini engellemek için sinirli olarak etiketlenmiş 60 adet veriye data augmentation uygulanmıştır. Normal veri sayısı 420 olduğundan 6 kat veri artırımı ile 360 adet sinirli veri ilave edilmiştir. Son durumda model 420 normal ve 420 sinirli ses olacak şekilde toplam 840 veri üzerinde koşmuştur. En yüksek doğruluğu elde etmek için farklı model yapıları denenmiştir. Bu modeller yalnızca evrişim katmanlarının sayısını değiştirerek oluşturulmuş ve diğer katmanlar ve parametreler aynı bırakılmıştır. Bu yüzden, Model-1 tek bir evrişim katmanı içerirken, Model-2 iki evrişim katmanı içerir Model-3 üç evrişim katmanına sahiptir, bu katmanlar ardışık olarak havuzlama katmanları tarafından takip edilir. Bu deneysel çalışmaların yapıldığı modeller, sadece evrişim katmanlarının sayısının değiştiği aynı temel model mimarisi kullanır. Model yüzde 80'e eğitim(training) yüzde 20 doğrulama(test) verisi olarak ayrılmıştır. Böylece 672 rastgele eğitim, 168 rastgele doğrulama verisi kullanılmıştır. Eğitimdeki tur (epoch) sayısı her iki dataset için 60 olarak

belirlenmiştir. Öğrenme oranı (learning rate) RMSprop algoritması için 0,0001'dir [9]. Demet boyutu (batch size) 60'tır. Az sayıdaki örneğin aşırı öğrenme davranışı göstermesinden kaçınmak amacıyla yüzde 40 özniteliği rastgele bırakma (dropout) uygulanmıştır [10]. Doğruluk, hassasiyet, duyarlılık ve F1 puanı hesaplanmıştır. Madde başında F1 puanını elde etmeye yarayan formüllere değinilmiştir.

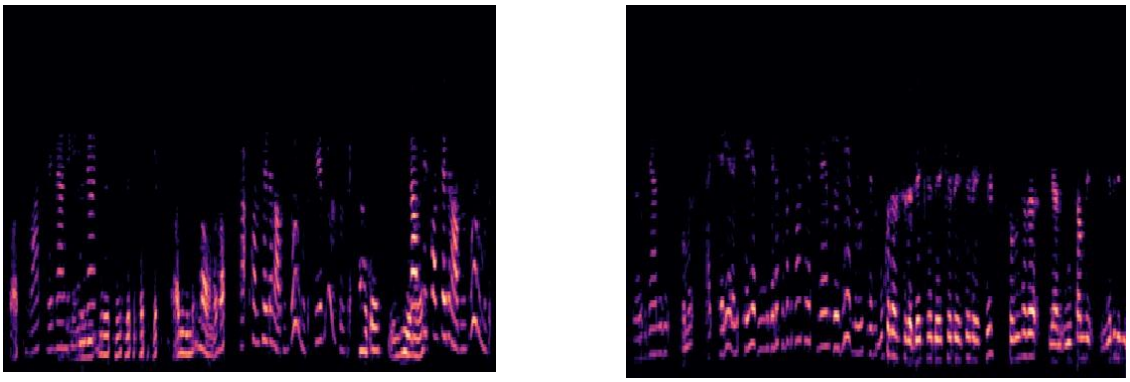
Firma veri seti laboratuvar ortamındaki seslere göre ortalama %1-1.5 daha kötü sonuçlar vermesine rağmen her iki veri seti %96 ile %99 civarı sonuçlar üretmiştir. Tablo 1 ve Tablo 2 deki sonuçlara göre Model başarısının rahatlıkla %90'ı aştığı söylenebilir. Şekil 7. Karmaşıklık matrislerinden anlaşılacağı üzere test ve doğrulama başarısı bu başarıyı destekleyecek nitelikte yüksektir. Bu başarıda firma çağrı merkezi ses kayıtlarının doğru sınıflandırılmasının yanında yüksek başarı elde etmesi için laboratuvar ortamındaki 4 saniyelik sesler gibi 10 saniyelik konuşma içeren önemli dilimlerin dikkate yakalanması önem arz etmektedir. Gerçek ortama ait çağrı merkezi ses kayıtları specktograma dönüştürülmeden önce en yüksek enerjili 5 saniye ve en düşük enerjili konuşma içeren 5 saniye birleştirilerek 10 saniyelik nitelikli veriler elde edilmiştir. Sonuçlar Türkçe ve İngilizce gibi dillerde tatmin edici olduğundan ayrıca metin madenciliği ve benzeri yöntemlere ihtiyaç duyulmamıştır. İtalyanca, İspanyolca, Çince gibi tınılı ve vurgusu yüksek diller için ek çalışmalar yapılması düşünülebilir.

İlgili model Python programlama diliyle conda kütüphaneleri üstünde derlenmiştir. Evrişimli Sinir Ağı(CNN) katmanları için Keras kütüphanesi tercih edilmiştir. Yine dengesiz veri kümelerini dengelemek amacıyla Keras ImageDataGenerator kullanılmıştır. Daha önceden seslerin düzenlenmesi, kesilmesi ve MFCC spectogramlara dönüştürülmesi için Librosa ve genel işlem ve hesaplamalar için Matloblib ile Numpy kütüphaneleri sıklıkla kullanılmıştır [11].

2.1 MFCC Dönüşümü

SAVEE ve Gürmen Tekstile ait ses verilerinin derin öğrenme algoritmaları ve yapay sinir ağlarıyla sınıflandırılabilmesi için öncelikle öznitelik(feature) çıkartılması gerekmektedir. Mel-Frequency Cepstral Coefficients yüksek performanslı sinyal işleme için çokça tercih edilen popüler bir tekniktir. MFCC İnsan kulağının algılayabildiği frekans ve bant genişliğini baz alır. Dijital ses işleme dalında önem arz eden bir ayırt edici nitelik çıkarma yöntemidir[12]. Özellikle konuşma tanıma ve ses işleme uygulamalarında tercih edilmektedir. MFCC öznitelikleri çıkartırken ses sinyallerinin kesilmesi, hamming penceresi uygulanması, spectrum hesaplama, filtre bankası uygulanması, logaritma alma ve DCT (Ayrık Kosinüs Dönüşümü) adım ve yöntemlerini uygulayarak bir öznitelik vektörü oluşturur[13].

Uzun ses sinyalleri daha küçük ve işlenebilir 20-30 milisaniyelik çerçevelere kesilir. Her çerçeve Hamming Penceresi ile çarpılarak kaynak hataları yumuşatılır. Spektrum hesaplarken(FFT) Hızlı Fourier Dönüşümü uygulanarak frekansa ait bileşenler elde edilir. Filtre bankası çerçeveye ait spektrumu farklı frekans bantlarına böler. Bu frekans bantları üst üste binen üçgen filtre bankasını oluşturur. Filtre bankasındaki bu değerler bir takım enerji hesaplamaları sonucunda MFCC hesaplamasında kullanılmak üzere ayrılır. Logaritma alma adımında bu ayrılmış enerjiler insan kulağına yakın bir taklit ile logaritmik ses yoğunluğuna uyarlanır. Son olarak Ayrık Kosinüs Dönüşümü(DCT) ile enerji dağılımındaki bu frekans farklılıkları katsayılar vasıtasıyla MFCC öznitelik vektörlerinin oluşturulması için kullanılır.



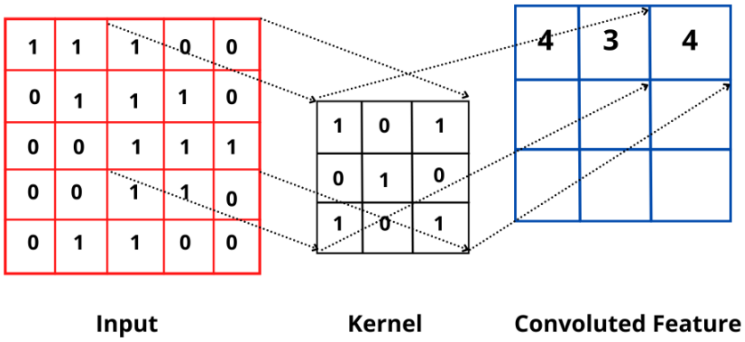
Şekil 2. Rastgele seçilmiş normal (sodaki) ve sinirli (sağdaki) konuşma sinyallerinin MFCC vektörel spectogramları (**Figure 2.** MFCC vectors of randomly selected normal (left image) and angry (right image) speech signals)

Sonuç olarak, MFCC öznitelikleri, ses sinyalinin temel spektral özelliklerini yakalar ve insan işitme sisteminin algılayabileceği biçimde düzenler. Bu nedenle, konuşma tanıma, konuşmacı tanıma, ses duygu analizi ve diğer ses işleme görevlerinde yaygın olarak kullanılır. Bu öznitelikler, ses verilerinin boyutunu azaltır ve anlamlı bilgileri vurgular, bu da daha iyi model eğitimi ve daha iyi sonuçlar elde etmek için önemlidir.

2.2 Derin Öğrenme Yöntemi

Evrişimli Sinir Ağı (Convolutional Neural Network - CNN) ileri beslemeli yapay sinir ağıdır [12]. Özellikle görüntü ve ses tanıma ve işleme alanında büyük başarı elde etmiş bir yapay sinir ağı türüdür. CNN'ler, veri içinde özellikleri belirlemek ve bu özellikleri kullanarak karmaşık desenleri öğrenmek için özel olarak tasarlanmıştır [12].

CNN'lerin temel özelliği, evrişim katmanları adı verilen özel katmanlar içermesidir. Evrişim katmanları, girdi verisi üzerinde öznitelikleri belirlemek üzere öğrenilebilen filtre veya çekirdekleri(kernel) kullanır. Bu filtreler, girdi verisi üzerinde gezinirken belirli desenleri veya özellikleri vurgular. Filtre boyutu, adım (stride) ve dolgu (padding) değerleri değiştirilebilir. Ardından havuzlama (pooling) katmanları, öznitelik haritasındaki boyutu küçültürken ve önemli bilgileri koruyarak hesaplama karmaşıklığını azaltır.



Şekil 3. ESA modeli giriş, çekirdek ve evrişim özellikleri (**Figure 3.** CNN model input, Kernel and convolution features)

CNN'ler genellikle tam bağlantılı (fully connected) katmanlarla sonlanır. Bu katmanlar, önceki evrişim ve havuzlama katmanlarından gelen özellikleri kullanarak sınıflandırma veya regresyon gibi görevleri gerçekleştirmek için eğitilir. CNN'lerin bu yapıları, özellikle görüntü verilerinde hiyerarşik özellikleri anlamak ve öğrenmek için oldukça etkilidir. CNN'ler, önceden elle belirlenmiş özellikler yerine, veriden özellikleri otomatik olarak çıkarabilme yetenekleriyle bilinir, bu nedenle geniş bir uygulama yelpazesi bulunmaktadır.

Çalışmada kullanılan CNN mimarisi şu şekildedir. Veri yükleme adımında sinirli ve normal klasöründeki sesler png formatında spectrogram resimlere dönüştürülmüştür. Bu spectrogramlarda (MFCC) 128x128 piksel RGB renk formatı kullanılmış bu görüntüler sinirli 1 normal 0 olacak şekilde etiketlenmiştir. Sinir ağındaki veriler Numpy dizisine dönüştürülerek %20 test verisi olacak şekilde eğitim/test olarak sınıflandırılmıştır. Böylece model verilerin yüzde 80'ini eğitim için kullanacak yüzde 20'siyle hiç görmediği veriler üstündeki başarısını ölçecektir. İleri besleme yapısına uygun olarak(feedforward) her katmanın bir önceki katmandan gelen çıkışları alarak işlem yapması amacıyla Keras kütüphanesindeki sequential model kullanılmıştır.

Evrişim katmanı 2 Boyutlu(Conv2D) ve havuzlama katmanı yine 2D olacak şekildedir. Havuzlama katmanı önemli bilgileri koruyup gereksiz detayları azaltmasıyla özellik haritasının boyutunu küçülterek ağın hesaplama karmaşıklığını azaltır. Veri setlerinde az veri bulunduğundan gerçek dünyadaki verimsizliği azaltmak amacıyla veri artırma(Data Augmentation) ve aşırı uydurmayı(Overfitting) azaltmak amacıyla dropout katmanı kullanılmıştır. Ağırlıklara göre kayıpları hesaplayıp kayıpların azaltılması için optimizasyon tekniği olarak ReLu (Rectified Linear Unit) tercih edilmiştir. ReLU, özellikle evrişimli sinir ağları (CNN) ve derin sinir ağları (DNN) gibi modellerde tercih edilen bir aktivasyon fonksiyonudur.

ReLU fonksiyonu, matematiksel olarak şu şekilde ifade edilir:

$$\boldsymbol{f(x) = max(0, x)} \textbf{ (1)}$$

Bu fonksiyon, girdi olan x değerini alır ve eğer x pozitifse, kendi değerini direkt olarak çıktı olarak verir. Eğer x değeri negatifse, çıktıyı sıfır yapar. Yani, ReLU fonksiyonu girişin negatif kısmını sıfır yaparak aktivasyonu geçirir. ReLU hızlı hesaplama ve Vanishing Gradient (Kaybolan Gradyan) sorunu için iyi bir aktivasyon fonksiyonu çözümü sağlamaktadır. Çalışmaya ait model derlenirken onlarca seçenekten ikisi olan Adam ve RMSprop optimizasyon algoritmalarından başarı oranı daha yüksek bir başarı oranına sahip olduğu gözlemlenen RMSprop tercih edilmiştir. Adam algoritması nerdeyse RMSprop kadar başarı sağlamaktadır [13].

3. Araştırma Sonuçları ve Tartışma

Kod çalışmaları 8 Çekirdekli, 3 GHz işlemcili, 64 bit Windows 11 işletim sistemli ve 64GB hafızaya sahip dizüstü bir bilgisayarda gerçekleştirilmiştir.

3.1. Veri Setleri

Çalışmada iki farklı veri seti tercih edilmiştir. Gerçek dünya uygulaması için Gürmen Tekstil çağrı merkezinden uzunlukları 1 ile 15 dakika arasında değişen 480 adet ses kaydı alınmıştır. Bu konuşma ses kayıtlarının son 1 dakikalık bölümleri içinden sadece konuşma içeren bölümler alınarak sessiz kısımlar göz ardı edilmiştir. Gürmen Tekstil çağrı merkezine ait konuşma içeren kısımlardan son olarak enerjisi en yüksek 5 saniye ile en düşük 5 saniye birleştirilerek 10 saniyelik kısımlar analize hazır hale getirilmiştir. Gerçek dünyadaki Gürmen Tekstil veri setinin akademik bir veri seti ile kıyaslanması için SAVEE (Surrey Audio-Visual Expressed Emotion) veri seti tercih edilmiştir. SAVEE ses veri setinde 4 farklı erkek tarafından seslendirilmiş 480 adet ingilizce sözcük ve 7 duygu bulunmaktadır[15]. Mutluluk, Üzüntü, İğrenme, Korku, Şaşkınlık, Nötr duygular normal olarak etiketlenmiş Öfke duygusu sinirli olarak etiketlenmiştir. SAVEE veri setindeki her kayıt yüksek kaliteli laboratuvarlarda kayıt edilmiş 4 saniyelik ses kayıtlarından oluşmaktadır. Bu kayıtlara ait duygular 10 kişilik juri tarafından değerlendirilerek 7 farklı duygu sınıflandırılmıştır[14]. Her iki veri setinde 420 adet normal, 60 adet sinirli olarak etiketlenmiş 480er ses kaydı yer almaktadır.

3.2. Metrikler

Çalışmada CNN modeli ile konuşma kayıtlarının normal ve sinirli olarak sınıflandırılması önerilmekte olup modellerin bu sınıflandırma problemindeki performansları aşağıdaki formüllerle hesaplanmaktadır.

Doğruluk (accuracy) değerini hesaplamak için kullanılan formül

$$\boldsymbol{doğruluk} = \frac{(\boldsymbol{DP + DN})}{(\boldsymbol{DP + DN + YP + YN})} \textbf{(2)}$$

Kesinlik (precision) değerini hesaplamak için kullanılan formül

$$\boldsymbol{kesinlik} = \frac{\boldsymbol{DP}}{(\boldsymbol{DP + YP})} \textbf{(3)}$$

Duyarlılık (recall) değerini hesaplamak için kullanılan formül

$$\boldsymbol{duyarlılık} = \frac{\boldsymbol{DP}}{(\boldsymbol{DP + YN})} \textbf{(4)}$$

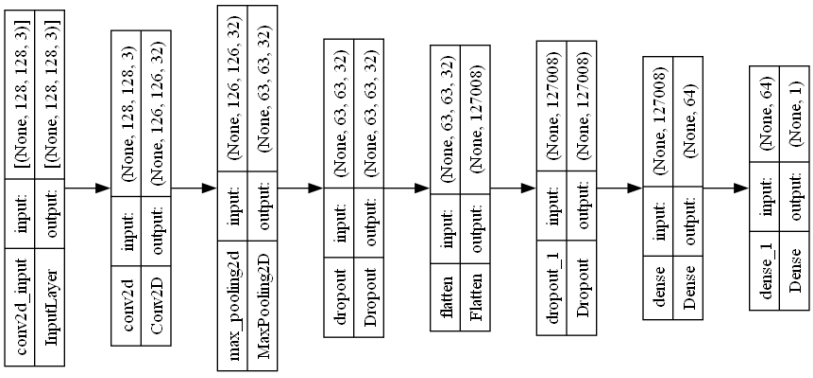
F1 skoru ($f1\ score$) değerini hesaplamak için kullanılan formül

$$\boldsymbol{f1\_skoru} = \frac{\boldsymbol{2 \times (kesinlik \times duyarlılık)}}{(\boldsymbol{kesinlik + duyarlılık})} \textbf{(5)}$$

DP-true positives, DN-true negatives, YP-false positives, YN-false negatives.

3.3. Derin Model Tasarımı ve Mimarisi

Bu çalışmada kullanılan derin model Model1, Model2 ve Model3 olarak adlandırılmıştır. Bu modellerde MAX poling katmanı, drop out, düzleştirme katmanı ve 64 boyutlu Dense katmanını kullandık. Bu modeller arasındaki temel fark, evrişim katmanlarının sayısına dayanmaktadır. İlk modelde, şekil 4'te gösterilen 3x3 pencere boyutunda 32 düğüm kullandık. Model 2'de 32 düğümlü ve 3'e 3 pencere boyutunda 2 evrişimli katman kullandık. Ayrıca Modelde aynı konfigürasyonda 3 adet evrişimsel katman uyguluyoruz.



Şekil 4. Önerilen Evrişimli Sinir Ağı Modeli (**Figure 4.** Proposed Convolutional Neural Network Model)

3.4. Sonuçlar ve Tartışma

Bu metodolojiler, sonuçları aşağıdaki alt bölümlerde sunulan iki veri kümesine dayalı olarak denenmiştir.

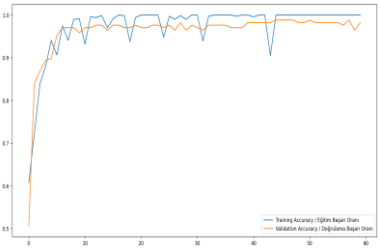
3.4.1. Savee Veriseti Sonuçları

Savee Akademik veri seti en yüksek başarı oranı Model 1'de gözlemlenmiştir. 3 model, eğitimde toplu iş boyutu (Batch size) 16 olan 60 dönem boyunca(Epochs) test edilmiştir. Analizler tablo 1 de gördüğümüz sonuçlara göre Model 1 en başarılı sonuçlara göre eğitim başarı oranı %100, doğrulama başarı oranı %98.21 F1 skorumuz %98.28 olarak gözlemlenmiştir. Model 2 genel başarı oranı diğer modellere göre daha düşüktür. Model 2 deki eğitim başarı ıranı %94.94, doğrulama başarısı %89.28 ve F1 Skoru % 89.02 olarak gerçekleşmiştir. Bu çalışmada Evrişimsel katman sayısı ve karmaşıklık arttıkça başarı oranının düştüğü gözlemlenmiştir. Bu yüzden tek katmanlı evrişimsel sinir ağları metodolojimiz için uygun bir katsayıdır.

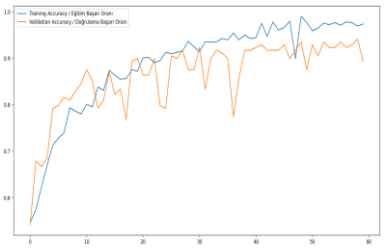
Model	Eğitim Başarısı (%)	Doğrulama Başarısı (%)	F1 Skoru (%)
Model 1	100	98.21	98.28
Model 2	94.94	89.28	89.02
Model 3	97.32	89.88	90.90

Tablo 1. Savee veri seti eğitim ve doğrulama(test) en yüksek başarı oranları (**Table 1.** Savee dataset training and validation (testing) highest success rates)

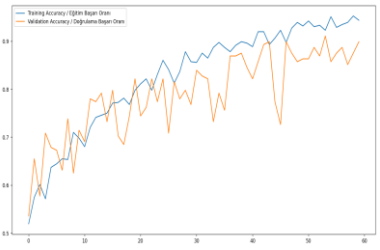
SAVEE datasetinde Model 2 ve Model 3 oskilatör figürlerinde training accuracy ve validation accuracy değerleri arasında büyük bir fark gözlemlenmiştir. Bu durumda b ve c grafiklerine göre ilgili modeller için under fitting olayı görünmektedir. Bu sinir ağı için datasetin yetersiz sayıda olduğunu ortaya koymaktadır.



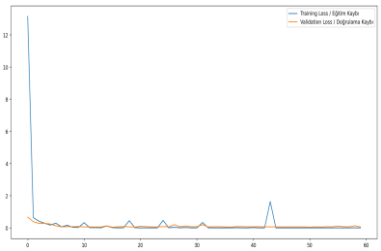
a) Savee Model-1 Eğitim Başarı Grafiği



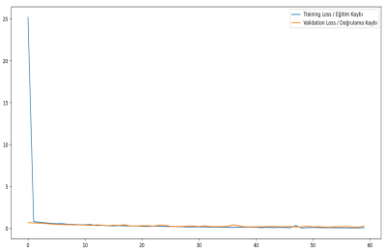
b) Savee Model-2 Eğitim Başarı Grafiği



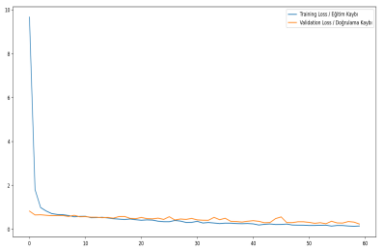
c) Savee Model-3 Eğitim Başarı Grafiği



a) Savee Model-1 Eğitim Kayıp Grafiği



b) Savee Model-2 Eğitim Kaybı Grafiği



c) Savee Model-3 Eğitim Kaybı Grafiği

Şekil 5. Savee public veri seti eğitim ve doğrulama model grafikleri (**Figure 5.** Savee public dataset training and validation model graphics)

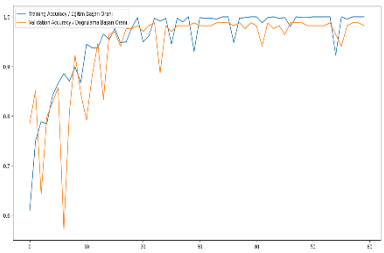
3.4.2. Gürmen RAMSEY|KIP Veriseti Sonuçları

Gürmen veri seti analizlerinde Tablo 2'de gördüğümüz sonuçlara göre Model 2 en başarısız model olarak görülmektedir. Model 2 deki sonuçlara göre eğitim başarı oranı %93.89, doğrulama başarı oranı 89.88 F1 skorumuz 86.82 olarak gözlemlenmiştir. Model 1 genel başarı oranı diğer iki modele göre daha yüksektir. Model 1 deki eğitim başarı ıranı %99.86, doğrulama başarısı %97.19 ve F1 Skoru % 97.26 olarak gerçekleşmiştir.

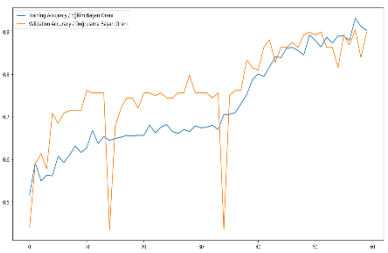
Bu çalışmada Evrişimsel katman sayısı ve karmaşıklık azaldıkça başarı oranı yükselmektedir. Bu yüzden çok katmanlı evrişimsel sinir ağları yerine tek katmanlı evrişimsel katman yapısı metodolojimiz için daha yüksek başarı sağlamıştır.

Model	Eğitim Başarısı (%)	Doğrulama Başarısı (%)	F1 Skoru (%)
Model-1	99.86	97.19	97.26
Model-2	93.89	89.88	86.82
Model-3	98.80	96.42	96.15

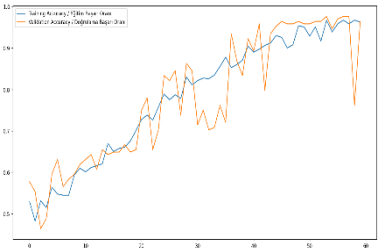
Tablo 2. Gürmen veri seti eğitim ve doğrulama(test) yüksek başarı oranları (**Table 2.** Gürmen dataset training and validation (testing) high success rates)



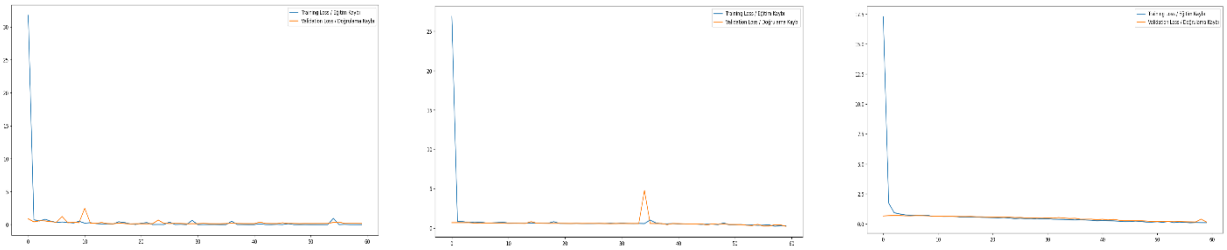
d) Gürmen Model-1 Eğitim Başarı Grafiği



e) Gürmen Model-2 Eğitim Başarı Grafiği



f) Gürmen Model-3 Eğitim Başarı Grafiği



d) Gürmen Model-1 Eğitim Kayıp Grafiği

e) Gürmen Model-2 Eğitim Kaybı Grafiği

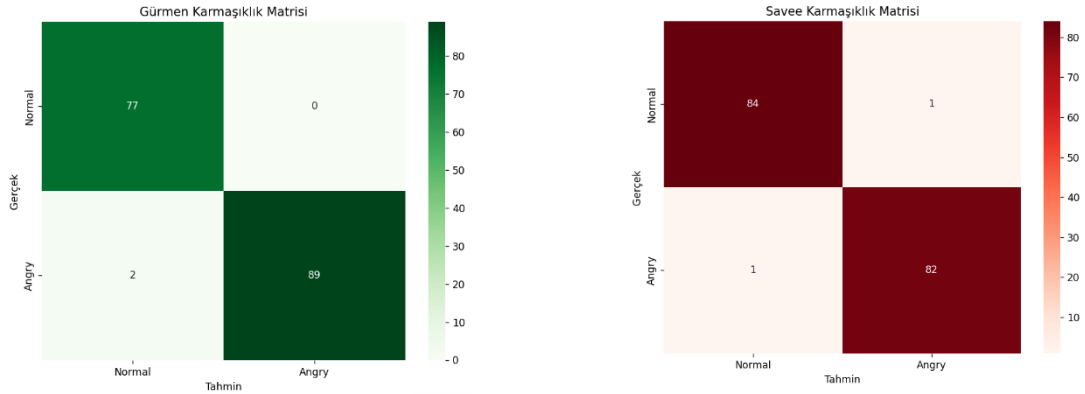
f) Gürmen Model-3 Eğitim Kaybı Grafiği

.

Şekil 6. Gürmen Tekstil firma veri seti eğitim ve doğrulama model grafikleri (**Figure 6.** Gurmen Textile company dataset training and validation model graphics)

3.5. Mukayese

Biz yukarıdaki çalışmalara göre F1 skorundaki başarısı doğrultusunda Model 1'i ana modelimiz olarak seçtik. Bu ana modele göre Gürmen ve Savee veri setleri için sonuçların kayda değer şekilde başarılı olduklarını gözlemledik. Bu tatmin edici sonuçları daha iyi incelemek için karmaşıklık matrislerini elde ettik. Aşağıda Şekil 7'de görüldüğü gibi Savee veriseti sinirli seslerden sadece 1 tanesini normal olarak teşhis etmiş ve normal olan seslerden sadece birini sinirli olarak tespit etmiştir. Geri kalan 84 normal olanı ve 82 sinirliyi doğru tespit etmiştir. Aynı şekilde Gürmen veri seti tüm sinirli sesleri doğru tahmin etmiş sadece 2 tane normal sesi sinirli olarak yorumlamıştır. Karmaşıklık matrisinde görüldüğü gibi sonuçlar çok tatmin edicidir.



Şekil 7. Firma(Gürmen) ve Public(Savee) veri seti karmaşıklık matrisleri (**Figure 7.** Company(Gurmen) and Public(Savee) dataset complexity matrices*)*

4. Sonuç

Çalışma, literatürdeki [1,2,3,4] numaralı kaynaklarla karşılaştırıldığında daha yüksek başarı elde etmiş, ancak [5,16,17] numaralı kaynaklarda elde edilen sonuçların nispeten daha yakın olduğu gözlemlenmiştir. Daha az sayıda duygu tipinin analiz edilmesi ve çağrı merkezine ait kayıtların nitelikli, anlam içeren 10 saniyelik dilimlerinin modele verilmesi çalışmanın başarısını diğer çalışmalara oranla artırmıştır. Kullanılan az sayıda veriye rağmen %99 eğitim, %97 doğrulama başarısı, sistemin gerçek hayatta kullanılabilir olması açısından tatmin edicidir. Çalışmada Türkçe ve İngilizce gibi naif diller kullanılmış olup elde edilen başarı neticesinden metin madenciliği vb. hibrit modellere yönelmeye gerek olmadığı görülmüştür. Farklı diller ve konuşma anında doğrudan duygu analizi için gelecekte model başarısının tekrardan ele alınması düşünülebilir. Modelde kullanılan Türkçe veri seti çalışmaya özgünlük katmıştır. Veri seti büyüklüğü arttıkça modeller genellikle daha dengeli sonuçlar verme eğilimine gireceğinden ilerleyen çalışmalarda çağrı merkezi kayıtlarının artırılması ve sonrasında görüşmelerin konuşma sırasındaki anlık analizi düşünülmektedir.

5. Teşekkür

Bu çalışmanın gerçekleşmesine katkıda bulunan Gürmen Tekstil'e ve özellikle Ramsey ve Kip markalarının çağrı merkezi kayıtlarını paylaşmalarına izin veren tüm paydaşlara teşekkür ederiz.

Kaynakça

[1] Berger, T., Lamel, L., & Devillers, L. (2021). End-to-End Speech Emotion Recognition: Challenges of Real-Life Emergency Call Centers Data Recordings. In 2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII) (pp. 1-6). IEEE. doi:10.1109/ACII52823.2021.9597419

[2] Mujaddidurrahman, A., Ernawan, F., Wibowo, A., Sarwoko, E. A., Sugiharto, A., & Wahyudi, R. (2021). Speech Emotion Recognition Using 2D-CNN with Data Augmentation. In Pekan, Malaysia Conference (pp. 1-5). IEEE. doi:10.1109/ICSECS52883.2021.00130

[3] Płaza, M., Kazała, R., Koruba, Z., Kozłowski, M., Lucińska, M., Sitek, K., & Spyrka, J. (2022). Emotion Recognition Method for Call/Contact Centre Systems. Kielce University, Poland. Retrieved from https://www.mdpi.com/2076-3417/12/21/10951

[4] Karataş, A. F., Mercan, Ö. B., Özdil, U., & Ozan, Ş. (2022). Çağrı Merkezlerinde Olumsuzluk İçeren Çağrıların Evrişimsel Sinir Ağları ile Tespiti. İzmir Demokrasi Üniversitesi. Retrieved from https://dergipark.org.tr/tr/pub/gazibtd/issue/75695/1156330

[5] Demircan, S. (2020). Ses sinyallerinden duygu tanıma için farklı yaklaşımlar [Different Approaches for Emotion Recognition from Speech Signals]. (Doktora Tezi, Konya Teknik Üniversitesi). Ulusal Tez Merkezi Tez No: 637322.

[6] Ahmed, A., Sivarajah, U., Irani, Z., Mahroof, K., & Charles, V. (2022). Data-driven subjective performance evaluation: An attentive deep neural networks model based on a call centre case. Annals of Operations Research. doi:10.1007/s10479-022-04874-2

[7] Van den Oord, A., et al. (2016). WaveNet: A Generative Model for Raw Audio. arXiv preprint arXiv:1609.03499.

[8] Chaudhary, N., et al. (2021). Efficient and Generic 1D Dilated Convolution Layer for Deep Learning. ACM Transactions on Computing. doi:10.1145/2104.08002.

[9] Einy, S., Oz, C., & Navaei, Y. D. (2021). IoT Cloud-Based Framework for Face Spoofing Detection with Deep Multicolor Feature Learning Model. Journal of Sensors, 2021, Article ID 5047808. doi:10.1155/2021/5047808.

[10] Al-Bander, B., Al-Nuaimy, W., Williams, B. M., & Zheng, Y. (2018). Multiscale sequential convolutional neural networks for simultaneous detection of fovea and optic disc. Biomedical Signal Processing and Control, 40, 91-101. doi:10.1016/j.bspc.2017.09.008.

[11] Liang, S., & Gu, Y. (2021). Computer-aided diagnosis of Alzheimer's disease through weak supervision deep learning framework with attention mechanism. Sensors, 21(1), 220. doi:10.3390/s21010220.

[12] Rehman, Y. A. U., Po, L. M., & Liu, M. (2020). SLNet: Stereo face liveness detection via dynamic disparity-maps and convolutional neural network. Expert Systems with Applications, 142, 113002. doi:10.1016/j.eswa.2019.113002.

[13] Shao, J., & Qian, Y. (2019). Three convolutional neural network models for facial expression recognition in the wild. Neurocomputing, 355, 82-92. doi:10.1016/j.neucom.2019.05.005.

[14] Surrey Audio-Visual Expressed Emotion (SAVEE) Database. (n.d.). Retrieved from https://kahlan.eps.surrey.ac.uk/savee/

[15] Dikbıyık, E., Demir, Ö., & Doğan, B. (2022). Derin Öğrenme Yöntemleri İle Konuşmadan Duygu Tanıma Üzerine Bir Literatür Araştırması [A Literature Review on Speech Emotion Recognition Using Deep Learning Methods]. Gazi Üniversitesi Fen Bilimleri Dergisi GU J Sci, Part C, 10(4), 765-791.

[16] Dal Ri, F. A., Ciardi, F. C., & Conci, N. (2023). Speech Emotion Recognition and Deep Learning: An Extensive Validation Using Convolutional Neural Networks. University of Trento Department of Information Engineering and Computer Science (DISI). doi:10.1109/ACCESS.2023.3326071.

[17] Kabdualiyev, D., Madiyev, A., Rakhaliyev, A., Dikhan, B., Gizhduan, K., & Ali, H. (2023). A Web-Based Platform for Real-Time Speech Emotion Recognition using CNN. Department of Computer Science, Nazarbayev University, Astana, Kazakhstan.