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

Araştırma Makalesi / Research Article

Duygu tanımada akustik verilerle derin öğrenme modellerinin karşılaştırılması: LSTM ve DenseNet üzerine bir inceleme

Comparison of deep learning models with acoustic data in emotion recognition: a study on LSTM and DenseNet

Buket İşler[1*], Fahreddin Raşit Kılıç[2]

[1] İstanbul Topkapı Üniversitesi, Yazılım Mühendisliği Bölümü, buketisler@topkapi.edu.tr
ORCID: https://orcid.org/ 0000-0002-9393-9564

[2] Konya Teknik Üniversitesi, Bilgisayar Mühendisliği Bölümü, frkilic@ktun.edu.tr
ORCID: https://orcid.org/ 0009-0001-2099-3279

MAKALE BİLGİLERİ

Makale Geçmişi:

Geliş 31 Ekim 2024
Revizyon 10 Ocak 2025
Kabul 2 Şubat 2025
Online 26 Mart 2025

Anahtar Kelimeler:

Ses tanıma, Duygu tanıma, LSTM, DenseNet

ÖZ

Duygu tanıma sistemleri, insan-makine etkileşiminde sezgisel ve etkili ara yüzlerin geliştirilmesine olanak tanıyan önemli bir teknolojidir. Son yıllarda derin öğrenme yaklaşımlarının benimsenmesiyle bu sistemlerin etkinliği artmış ve farklı model türlerinin performanslarının incelenmesi önem kazanılmıştır. Bu çalışma, ses tabanlı duygu sınıflandırma alanında uzun-kısa süreli bellek (Long-Short Time Memory-LSTM) ve DenseNet modellerinin performanslarını karşılaştırmayı amaçlamaktadır. Emotion Speech Dataset (ESD) kullanılarak ses verileri üzerinde Mel Frekans Kepstral Katsayıları (MFCC) yöntemi ile özellik çıkarımı yapılmış, veriler dengelenmiş ve modeller eğitilmiştir. Performans değerlendirmesi, doğruluk, kesinlik, duyarlılık ve F1-skoru metrikleri ile gerçekleştirilmiştir. Sonuçlar, LSTM modelinin tüm metriklerde %92 oranında başarı sergilediğini, DenseNet modelinin ise %88 oranında kaldığını göstermiştir. Üzgün duygusunda LSTM modeli %98 kesinlik ve %92 duyarlılık sunarken, DenseNet modeli %94 kesinlik ve %89 duyarlılık sağlamıştır. Şaşkın duygusunda LSTM %86 kesinlik ve %97 duyarlılık elde ederken, DenseNet %82 kesinlik ve %93 duyarlılık sunmuştur. Mutlu duygusunda her iki model daha düşük performans sergilemiş olup, LSTM %85 kesinlik ve %81 duyarlılık oranlarına, DenseNet ise %80 kesinlik ve %76 duyarlılık oranlarına ulaşmıştır. Sonuç olarak, LSTM modelinin, bu çalışma kapsamında kullanılan veri seti ve deneysel koşullarda zamansal veri işleme kapasitesi sayesinde daha yüksek performans sergilediği belirlenmiştir.

ARTICLE INFO

Article history:

Received 31 October 2024
Received in revised form 10 January 2025
Accepted 2 February 2025
Available online 26 March 2025

Keywords:

Speech Recognition, Emotion Recognition, LSTM, DenseNet

Doi: 10.24012/dumf.1576811

ABSTRACT

Emotion recognition systems are a significant technology enabling the development of more intuitive and effective human-machine interaction interfaces. In recent years, the adoption of deep learning approaches has enhanced the efficiency of these systems, and examining the performance differences among various model types has become a critical research area. This study aims to compare the performance of Long Short-Term Memory (LSTM) and DenseNet models in the field of speech-based emotion classification. The Emotion Speech Dataset (ESD) was utilised, and feature extraction was performed on the speech data using the Mel Frequency Cepstral Coefficients (MFCC) method. The data were balanced, and the models were trained. Performance evaluation was conducted using accuracy, precision, recall, and F1-score metrics. The results showed that the LSTM model achieved 92% success across all metrics, while the DenseNet model remained at 88%. For the “sad” emotion, the LSTM model provided 98% precision and 92% recall, whereas the DenseNet model yielded 94% precision and 89% recall. Regarding the “surprised” emotion, the LSTM model achieved 86% precision and 97% recall, while the DenseNet model recorded 82% precision and 93% recall. In the “happy” emotion, both models exhibited lower performance, with the LSTM model achieving 85% precision and 81% recall, and the DenseNet model attaining 80% precision and 76% recall. In conclusion, it was determined that the LSTM model demonstrated higher performance due to its temporal data processing capability, under the dataset and experimental conditions used in this study.

* Sorumlu Yazar

Giriş

Duygu, deneyimler, düşünceler veya insanlarla etkileşimler gibi girdilere tepki olarak gelişen dinamik bir bilişsel ve fizyolojik durumdur. Duygular, öznel deneyim, bilişsel süreçler, davranışsal tepkiler, fizyolojik değişimler ve iletişim süreçlerini kapsar. Bu kapsamda, duygu tanıma teknolojileri, insan duygularını çeşitli işaretlerden, örneğin ses tonu, yüz ifadeleri ve vücut dili gibi verilerden tespit etmeye yönelik olarak geliştirilmiştir. Otomatik duygu algılama sistemleri, makinelerin insan duygularını anlayarak daha duyarlı ve kişisel hizmetler sunmasını amaçlamaktadır. Duygusal durumların tespiti, insan-makine etkileşimlerini geliştirmek ve iletişim süreçlerini daha verimli hale getirmek açısından önemli bir role sahiptir. Pazarlama, sağlık hizmetleri, insan-robot etkileşimi ve güvenlik gibi alanlarda bu teknolojilere olan ilgi giderek artmaktadır. Sağlık sektöründe, duyguların izlenmesi nörolojik rahatsızlıkların yönetimi için kritik öneme sahiptir. Uyku bozuklukları, şizofreni, Parkinson hastalığı ve uyku kalitesinin değerlendirilmesi gibi sorunlarda duygu tanıma sistemlerinden yararlanılmaktadır [1],[2],[3]. Ayrıca, otizm spektrum bozukluğu, dikkat eksikliği ve hiperaktivite bozukluğu ile panik bozukluğu gibi psikolojik durumların değerlendirilmesi ve yorgunluk, uykusuzluk, depresyon ile ağrı gibi fizyolojik durumların takibinde bu sistemlerin kullanım alanı genişlemektedir. Bu sistemler, zaman içinde duygusal değişimlerin izlenmesi ile daha kapsamlı veri analizi yapılmasına olanak tanımaktadır. Yapay zekâ ve derin öğrenme yöntemlerindeki gelişmeler, duygu tanıma teknolojilerinin doğruluğunu artırarak daha güvenilir sonuçların elde edilmesini sağlamaktadır [4],[5],[6]. Literatürde, yüz ifadeleri ve ses tabanlı duygu tanıma sistemlerinde genellikle derin öğrenme modelleri tercih edilmektedir [7],[8]. Bu modeller, karmaşık veri yapılarında ve büyük veri setleri üzerinde etkili sonuçlar üreterek yüksek doğruluk oranlarına ulaşabilmektedir. Görsel ve işitsel verilerin işlenmesinde öne çıkan bu yaklaşımlar, insan duygularının doğru şekilde tanınmasına katkı sağlamaktadır. Özellikle zamansal değişimlerin takibi gereken analizlerde, derin öğrenme tabanlı yöntemler dinamik veri analizine olanak tanımaktadır. Bu durum, duygu tanıma sistemlerinin insan davranışlarını daha kapsamlı ve güvenilir bir şekilde değerlendirmesine imkân vermektedir [9],[10].

Bu çalışma, sesli duygu tanıma sistemlerinin geliştirilmesinde derin öğrenme tekniklerinin uygulanabilirliğini incelemeyi ve bu alandaki iki farklı model olan uzun-kısa süreli bellek (Long-Short Time Memory-LSTM) ve DenseNet modellerini karşılaştırarak ses verileri üzerinde duygu tanıma performansını değerlendirmeyi amaçlamaktadır. LSTM, zamansal veri işleme yetenekleri ile tanınırken, DenseNet modeli hızlı eğitim süreci ve doğrusal olmayan ilişkileri öğrenme kapasitesi ile ön plana çıkmaktadır. Emotional Speech Database (ESD) kullanılarak gerçekleştirilen araştırma, Mel Frekans Kepstral Katsayıları (MFCC) yöntemiyle özellik çıkarımı yaparak farklı derin öğrenme yöntemlerinin doğruluk oranlarını analiz etmektedir. Çalışmanın literatüre özgün katkısı, aynı veri seti ve aynı özellik çıkarımı yöntemi kullanılarak iki farklı derin öğrenme modelinin (LSTM ve DenseNet) performansının kapsamlı bir şekilde karşılaştırılmasıdır. Bu karşılaştırma, doğruluk, kesinlik, duyarlılık ve F1-skoru gibi nicel bulgularla desteklenerek net bir şekilde ortaya konulmuştur. Özellikle, kullanılan yöntemlerin birbiriyle tutarlı bir şekilde uygulanması, iki modelin güçlü ve zayıf yönlerinin belirlenmesini sağlamış ve hangi durumlarda hangi modelin tercih edilebileceğine dair araştırmacılara rehberlik sunmuştur. Son olarak, elde edilen bulgular doğrultusunda gelecekteki duygu tanıma araştırmalarına sağlam bir temel oluşturmayı amaçlamaktadır.

Çalışmanın genel akışı dikkate alındığında, “İlgili Çalışmalar” başlığı altında ses tabanlı duygu tanıma alanında literatürde yer alan çalışmalar detaylandırılmaktadır. “Materyaller ve Yöntemler” bölümünde, veri seti, kullanılan derin öğrenme tabanlı modeller, özellik çıkarım teknikleri ve deneysel prosedür gibi çalışmada kullanılan tüm materyal ve metotlar açıklanmaktadır. “Bulgular” başlığı altında, gerçekleştirilen analizler tablo ve şekillerle sunulmakta ve yorumlanmaktadır. Çalışmanın sonuçlarının değerlendirilmesi ve elde edilen bulguların literatür ile karşılaştırılarak tartışılması ise “Tartışma ve Sonuç” başlığı altında yapılmaktadır.

İlgili Çalışmalar

Literatürde yaygın olarak kullanılan bu yöntemler, duygu tanımanın önemini ve uygulama alanlarını genişleterek teknolojilerin daha sezgisel ve etkili hale gelmesine olanak tanımaktadır. Farklı yıllarda yapılan çalışmalar, derin öğrenme tekniklerinin sesli duygu tanıma alanındaki başarısını göstermektedir:

2018 yılında yapılan bir çalışmada, derin sinir ağlarının konuşmadan duygu tanıma alanında büyük potansiyel taşıdığı belirtilmiştir. Araştırmada, MSP-Podcast veri seti kullanılarak farklı DenseNet mimarileri incelenmiş; ağın genişliği, derinliği, aktivasyon fonksiyonları ve veri artırma yöntemlerinin etkileri analiz edilmiştir. Sonuçlar, daha büyük eğitim setlerinin performansı artırdığını, daha derin ağlar için batch normalizasyonun kritik olduğunu ve residual ağlarla DenseNet ağlar arasında performans farkının minimal olduğunu ortaya koymuştur [11]. Elektroansefalografi (EEG) tabanlı duygusal tanıma üzerine yapılan bir çalışmada, "kanal-füzyonlu DenseNet konvolüsyon ağı" adı verilen yeni bir derin öğrenme modeli önerilmiştir. Model, EEG sinyallerindeki zamansal ve uzamsal özellikleri öğrenmek için 1-D konvolüsyon katmanları ve DenseNet kullanmıştır. SEED ve DEAP veri setleri üzerinde yapılan deneylerde sırasıyla %90.6 ve %92.5 doğruluk oranlarına ulaşılmıştır ve bu sonuçlar, modelin diğer çalışmalara kıyasla daha yüksek performans gösterdiğini ortaya koymuştur [12]. 2021 yılında gerçekleştirilen bir çalışmada, insan-makine etkileşiminde konuşma duygularını otomatik olarak tanımak amacıyla hibrit bir model geliştirilmiştir. Bu modelde Evrişimsel Sinir Ağları (CNN) ve LSTM mimarileri birleştirilerek Convolutional LSTM (Co-LSTM) yöntemi oluşturulmuştur. MFCC kullanılarak konuşma sinyallerindeki duygular tanınmış ve RAVDESS ile TESS veri setlerinde %86.7 doğruluk oranı elde edilmiştir. Bu sonuçlar, yöntemin diğer sınıflandırıcılardan daha etkili olduğunu göstermiştir [13]. Ses temelli duygu tanıma ile ilgili bir çalışmada, ses verilerini analiz etmek için Bag-of-Audio-Words (BoAW) özellikleri ve tekrarlayan sinir ağları (RNN) tabanlı bir duygu tespit modeli önerilmiştir. IEMOCAP veri seti üzerinde yapılan

deneylerde yüksek doğruluk oranları elde edilmiştir [14]. Aynı yıl yapılan bir başka çalışmada, LSTM ağlarının çeşitli mimariler ve optimizasyon teknikleri ile otomatik konuşma tanıma (ASR) sistemlerinin doğruluğunu artırdığı gösterilmiştir. LSTM tabanlı sistemlerin, özellikle gürültülü ortamlarda geleneksel yöntemlerden daha iyi performans sergilediği belirtilmiştir [15]. 2023 yılında gerçekleştirilen bir çalışmada, sentetik duygusal konuşma verilerinin DenseNet sinir ağları ile birleştirilmesi sonucu model performansının kayda değer biçimde iyileştiği gözlemlenmiştir [16]. 2024 yılında yapılan bir çalışmada, geleneksel LSTM ağlarının yüksek güç tüketimi sorunlarını çözmek amacıyla FPGA tabanlı bir LSTM hızlandırıcısı geliştirilmiştir. Bu hızlandırıcı, konuşma tanıma görevlerinde enerji verimliliğini artırırken güç tüketimini önemli ölçüde azaltmıştır [17]. 2024'te gerçekleştirilen başka bir çalışmada, konuşma duygularını tanımada kısa ve ritmik özelliklerin birleştirildiği yeni bir çoklu özellik yöntemi önerilmiştir. LSTM ağı ile yapılan deneylerde, Emo-DB veri setinde %100, CASIA'da %98.4 ve EMOVO'da %98.8 doğruluk oranları elde edilmiştir. Bu sonuçlar, yöntemin farklı diller ve duygu sınıflarında başarılı olduğunu göstermektedir [18].

Materyaller ve Yöntemler

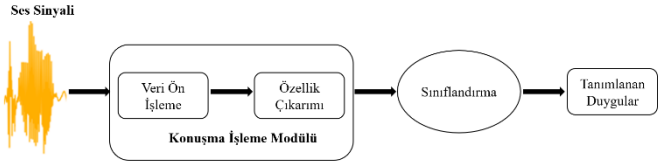
Veri seti

Bu çalışmada, duygusal konuşma tanıma (SER) alanında Singapur Ulusal Üniversitesi (NUS) ve Singapur Teknoloji ve Tasarım Üniversitesi (SUTD) tarafından sunulan ESD kullanılmıştır. ESD, nötr, mutluluk, öfke, üzüntü ve şaşkınlık duygularını içeren, her biri 10 İngilizce ve 10 Mandarin konuşmacı tarafından seslendirilen 350 paralel ifadeden oluşmakta ve yaklaşık 29 saatlik ses verisi içermektedir. Kayıtlar kontrollü bir akustik ortamda yapılmış olup, her bir .wav dosyası, konuşmacı ve duygu sınıflarına göre organize edilmiş ve benzersiz dosya adlarıyla etiketlenmiştir. ESD, mono-dil ve çapraz-dil duygusal konuşma dönüştürme, metinden duygusal konuşmaya ve duygu yüklü metinden konuşma sistemleri için uygundur [19].

Metodoloji

Bu çalışmada, duygusal ses tanıma modeli için iki derin öğrenme yöntemi, LSTM ve DenseNet, karşılaştırmalı olarak değerlendirilmiştir. Şekil 1'de sunulan geleneksel konuşma duygu tanıma sistemine benzer bir yaklaşımla, ses sinyallerinden özellik çıkarımı, özellik seçimi ve sınıflandırma aşamaları gerçekleştirilmiştir.

İlk olarak, ses dosyaları Librosa kütüphanesi kullanılarak işlenmiş ve her bir ses sinyalinden karakteristik özellikler çıkarılmıştır. Özellik çıkarımı aşamasında, yaygın olarak kullanılan Mel Frekans Kepstral Katsayıları (MFCC) yöntemi tercih edilmiştir. Elde edilen özellikler, sınıflandırma modellerine girdi olarak verilmek üzere standart hale getirilmiştir. Ön işleme ve özellik çıkarımı adımlarının tamamlanmasının ardından, bu özellikler LSTM ve DenseNet modellerine ayrı ayrı verilmiş ve her bir model, ses verilerindeki duygusal içerikleri tanıyacak şekilde eğitilmiştir. Eğitim sürecinde, veri seti için farklı eğitim-test oranları (%70-30, %75-25, %85-15) denenmiş ve elde edilen performans sonuçları değerlendirilmiştir. Yapılan bu denemeler sonucunda, %80 eğitim- %20 test oranının en iyi sonucu verdiği belirlenmiş ve bu oran esas alınarak her iki modelin eğitimi gerçekleştirilmiştir. Model hiperparametrelerinin belirlenmesinde literatürde yaygın olarak kullanılan değerler dikkate alınmış ve herhangi bir hiperparametre optimizasyon yöntemi uygulanmamıştır. Son aşamada, model çıktıları doğruluk, kesinlik, duyarlılık ve F1-skoru gibi performans metrikleri kullanılarak değerlendirilmiştir. Her iki model için elde edilen sınıflandırma raporları karşılaştırılmış ve sonuçlar doğrultusunda en iyi performansı sergileyen model belirlenmiştir. Bu çalışmanın temel amacı, ses tabanlı duygu tanıma süreçlerinde farklı derin öğrenme modellerinin performansını karşılaştırarak bu alandaki en etkili yöntemi belirlemektir.



Şekil 1. Önerilen Yöntemin Mimarisi

Kullanılan Modeller

Bu çalışmada, ESD kullanılarak, duygusal durumları tanıma amacıyla iki derin öğrenme modeli, LSTM ve DenseNet karşılaştırılmıştır.

LSTM

LSTM, RNN arasında, sıralı verilerle çalışırken uzun vadeli bağımlılıkları öğrenme yeteneği sunar. Geleneksel RNN'ler kısa vadeli bağımlılıkları yakalayabilse de zamanla yayılmış girdilerle başa çıkmakta zorlanır ve "kaybolan gradyan" sorunu ortaya çıkar. Bu sorun, küçük ağırlık değerlerinin öğrenme sürecini yavaşlatması veya durdurması ile ilgilidir [20],[21]. LSTM, bu problemi çözmek için hücre belleği ve unutma, girdi ve çıktı kapıları gibi mekanizmalar kullanır [22]. Unutma kapısı, önceki zamandan gelen hangi bilgilerin eleneceğine karar verirken; girdi kapısı, yeni bilginin ne kadarının hafızaya kaydedileceğini kontrol eder. Çıktı kapısı ise bir sonraki katmana ne kadar bilgi aktarılacağını düzenler. Bu mekanizmalar sayesinde LSTM, sıralı verilerdeki uzun süreli ilişkileri etkili bir şekilde yakalayarak RNN'lerin temel sorunlarını aşar ve konuşma, dil işleme gibi alanlarda önemli bir performans avantajı sağlar. LSTM'ler, derin öğrenme çalışmalarında CNN ve klasik RNN'lerden daha başarılı performans göstermektedir [23].

DenseNet

DenseNet sinir ağları, her katmandaki nöronların bir önceki katmandaki tüm nöronlarla bağlantılı olduğu yapay sinir ağlarıdır ve karmaşık ilişkilerin öğrenilmesine olanak tanır [24]. Bir DenseNet ağı, giriş katmanı, bir veya daha fazla tam bağlantılı gizli katman ve çıkış katmanından oluşur. Eğitim sırasında ağırlıklar ve önyargılar optimize edilerek modelin doğruluğu artırılır [25]. ReLU, sigmoid ve tanh gibi aktivasyon fonksiyonları, her katmandaki girdileri işleyerek ağın doğrusal olmayan ilişkileri modellemesini sağlar. DenseNet sinir ağları, özellikle konuşma tanıma ve duygu analizi gibi görevlerde üstün performans gösterir [26]. Bu süreçte ses sinyalleri önce MFCC gibi özellik çıkarım teknikleriyle işlenir, ardından sınıflandırma için DenseNet

sinir ağına aktarılır. Ağın tam bağlantılı yapısı, ses sinyallerindeki karmaşık kalıpları öğrenmesini sağlayarak daha doğru sınıflandırmalar yapmasına olanak tanır [27].

Sonuçlar

LSTM Modeli ve Sonuçları

Analiz süreci, MFCC özniteliklerinin çıkarılmasıyla başlamış ve bu işlemde n_mfcc=40 parametresi kullanılmıştır. Bu parametre, ses sinyallerinin spektral özelliklerini yeterli çözünürlükte temsil edebilmekte ve literatürde yaygın olarak tercih edilmektedir. Ses dosyalarından 3 saniyelik segmentler alınmış ve 0.5 saniyelik offset uygulanmıştır. Bu parametrelerin, duygusal ifadelerin karakteristik özelliklerini yakalamak için uygun zamansal çözünürlük sağladığı değerlendirilmiştir. Elde edilen MFCC matrisleri, LSTM modeline uygun formata dönüştürülmüş ve pad_sequences_2d fonksiyonu kullanılarak sabit uzunluğa getirilmiştir. Padding stratejisi olarak "post" seçilmiş ve böylece öznitelik dizilerinin başındaki kritik bilgilerin korunması sağlanmıştır. LSTM modelli için MFCC özellik çıkarım süresi 66.1 saniye olarak ölçülmüştür. Veri setinin eğitim ve test kümelerine ayrılmasında literatürde yaygın olarak kullanılan %80 eğitim- %20 test oranı esas alınmış, ancak bu oranla yapılan deneyler yalnızca tek bir seferde değil, farklı rastgele bölünmelerle birçok kez tekrarlanmıştır. Her çalıştırmada model performansı değerlendirilmiş ve sonuçların tutarlı olduğu gözlemlenmiştir. Deneysel çalışmalar sırasında farklı veri bölme oranları da (%70-30, %75-25, %85-15) denenmiş, ancak %70-30 oranında modelin yeterince öğrenemediği, %85-15 oranında ise test kümesinin küçük kalması nedeniyle değerlendirme güvenilirliğinin azaldığı belirlenmiştir. Bu doğrultuda %80 eğitim- %20 test oranının modelin öğrenme kapasitesi ve genelleme başarısı açısından optimal sonuç verdiği tespit edilmiştir. LSTM modeli için eğitim süresi 25 dakika 12 saniye, çıkarım süresi ise 101.98 ms ölçülmüştür. Çalışmada, duygusal konuşma tanıma sistemi için 5-katlı çapraz doğrulama yöntemi uygulanmıştır. Veri setinin katmanlara ayrılmasında StratifiedKFold sınıfı kullanılmış ve her katmanda sınıf dağılımlarının dengeli olması sağlanmıştır. Etiketler LabelEncoder ile sayısal değerlere dönüştürülmüş ve eğitim sürecinde kullanılmak üzere one-hot encoding yöntemi uygulanmıştır. Her bir katmanda, veri stratified splitting yöntemi ile %80 eğitim- %20 doğrulama oranında bölünmüş ve sınıf dengesi korunmuştur. Oluşturulan LSTM modeli, iki katmandan oluşmaktadır. İlk katmanda 128, ikinci katmanda 64 nöron bulunmaktadır. Literatürde benzer veri setleri ile yapılan çalışmalar dikkate alınarak nöron sayıları belirlenmiş ve modelin aşırı karmaşıklıktan kaçınarak yeterli öğrenme kapasitesine ulaşması hedeflenmiştir. İki ardışık LSTM katmanının kullanılması, veri içindeki zamansal örüntülerin daha etkin öğrenilmesini sağlamıştır. İlk LSTM katmanında return_sequences=True parametresi kullanılarak zamansal bilginin korunması amaçlanmıştır. Her iki katman arasında ReLU aktivasyon fonksiyonu tercih edilmiş ve aşırı öğrenmeyi önlemek için %20 dropout uygulanmıştır. LSTM katmanlarından sonra 32 nöronlu bir Dense katman eklenerek yüksek seviyeli özniteliklerin daha yoğun ve anlamlı bir temsili sağlanmıştır. Dense katmanında ReLU, çıkış katmanında ise çok sınıflı sınıflandırma amacıyla softmax aktivasyon fonksiyonu kullanılmıştır. Eğitim sürecinde hiperparametre olarak Adam optimizasyon algoritması tercih edilmiştir. Adam, adaptif öğrenme oranı stratejisi sayesinde hızlı yakınsama ve gradyan problemlerini etkili bir şekilde yönetme kapasitesine sahiptir. Batch size 32 olarak belirlenmiş ve bu değerin, modern GPU'ların bellek kapasitesini etkin kullanırken yeterli genelleme performansı sunduğu gözlemlenmiştir. Eğitim sürecinde epoch sayısı 50 olarak belirlenmiştir. Farklı epoch değerleri (20, 30, 40, 60 ve 100) denenmiş; 50 epoch değerinin modelin yakınsama performansı ve genel doğruluk açısından en iyi sonucu verdiği görülmüştür. Daha düşük epoch sayılarında modelin yeterince öğrenemediği, daha yüksek epoch sayılarında ise aşırı öğrenme belirtileri gözlenmiştir. Early Stopping yöntemi tercih edilmemiş; bu karar, ESD veri tabanının homojen yapısı göz önünde bulundurularak, modelin belirli bir epoch sayısına kadar kesintisiz eğitilmesinin genelleme performansını iyileştireceği değerlendirmesine dayandırılmıştır. Model hiperparametreleri belirlenirken literatürde yaygın olarak kullanılan değerler esas alınmış ve belirli bir hiperparametre optimizasyon yöntemi uygulanmamıştır. Bu seçimler, LSTM modelinin zamansal veri işleme yeteneğini ve ses verilerindeki duygusal özellikleri etkili şekilde öğrenme kapasitesini optimize etmeyi hedeflemiştir.

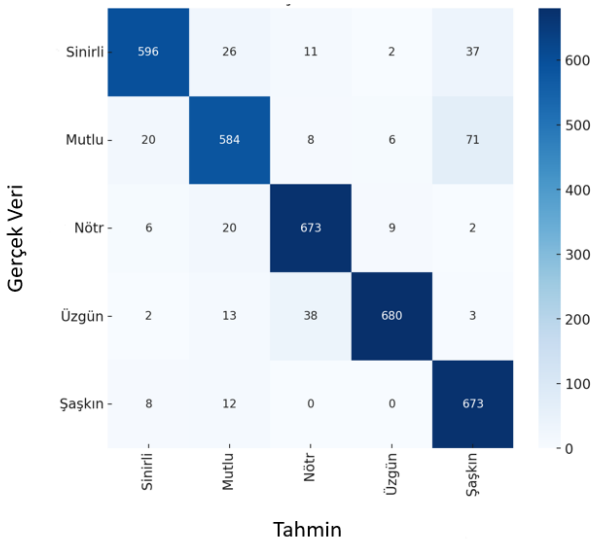
Sonuç olarak, belirlenen hiperparametreler ile eğitilen model, test verisi üzerinde %92 doğruluk oranına ulaşmıştır. Bu sonuçlar, seçilen hiperparametrelerin problem uzayı için uygun olduğunu ve modelin genel olarak etkin bir performans sergilediğini göstermektedir.

Model sonuçları, Tablo 1'de sunulan sınıflandırma metriklerine göre modelin ses verileri üzerinden duyguları tanıma performansını detaylı bir şekilde ortaya koymaktadır. Sonuçlar incelendiğinde, özellikle Üzgün duygusunun %98 kesinlik ve %95 F1-skoru ile en yüksek başarıyı sergilediği görülmektedir. Bu durum, modelin bu duyguyu tanıma konusundaki yetkinliğini açıkça göstermektedir. Benzer şekilde, Nötr sınıfı %95 duyarlılık ve %93 F1-skoru ile başarılı bir şekilde tanınmış, modelin bu sınıfta da istikrarlı bir performans sergilediği belirlenmiştir. Öte yandan, Mutlu sınıfı %85 duyarlılık ve %87 F1-skoru ile diğer duygulara kıyasla daha düşük bir performans göstermiştir. Bu sonuç, modelin bu sınıfta sınırlı bir ayırt edici yeteneğe sahip olduğunu ve geliştirmeye ihtiyaç duyduğunu işaret etmektedir. Şaşkın duygusu ise %97 duyarlılık ile yüksek bir başarı sergilemiş olmasına rağmen, %86 kesinlik oranı bu sınıfta yanlış pozitif sonuçların nispeten daha fazla olduğunu göstermektedir. Bu durum, şaşkınlık ile diğer duygular arasında benzerlik olabileceği ve modelin bu duyguyu diğer sınıflarla karıştırma eğiliminde olduğunu düşündürmektedir. Modelin genel doğruluk oranı %92 olarak belirlenmiştir. Ayrıca makro ve ağırlıklı ortalama değerlerinin tüm metriklerde %92 olması, sınıflar arasında genel olarak dengeli bir performans elde edildiğini göstermektedir. Bununla birlikte, bazı sınıflarda daha yüksek bir ayırt edicilik sağlanması gerektiği anlaşılmaktadır. Bu bulgular, modelin genel olarak ses verileri üzerinden duygu tanıma görevinde başarılı bir performans sergilediğini, ancak Mutlu ve Şaşkın sınıflarında daha iyi sonuçlar elde edebilmek için iyileştirmeye açık olduğunu göstermektedir.

Tablo 1. LSTM Duygu Tanıma Sınıflandırma Performansı

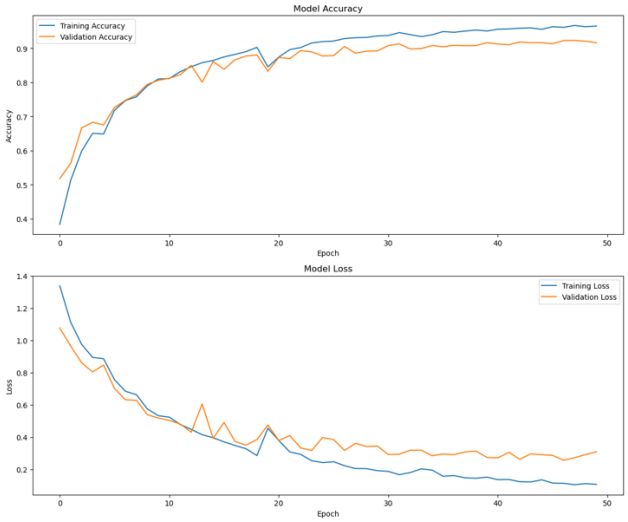
Duygu	Kesinlik	Duyarlılık	F1-Skoru	Destek
Sinirli	0.94	0.89	0.91	672
Mutlu	0.89	0.85	0.87	689
Nötr	0.92	0.95	0.93	710
Üzgün	0.98	0.92	0.95	736
Şaşkın	0.86	0.97	0.91	693
Doğruluk			0.92	3500
Makro Ort.	0.92	0.92	0.92	3500
Ağırlıklı Ort.	0.92	0.92	0.92	3500

Şekil 2'de verilen LSTM karmaşıklık matrisi, modelin ses verileri üzerinden duyguları sınıflandırma performansını ayrıntılı olarak ortaya koymaktadır. Matris incelendiğinde, Nötr ve Üzgün duygularının yüksek doğruluk oranlarıyla doğru sınıflandırıldığı, bu duyguların diğer sınıflarla karışmasının oldukça az olduğu gözlemlenmiştir. Özellikle Üzgün duygusunun yalnızca küçük bir kısmının Nötr ve Mutlu olarak yanlış sınıflandırılması, modelin bu duyguyu yüksek doğrulukla ayırt edebildiğini göstermektedir. Buna karşılık, Mutlu duygusunda diğer duygularla karışıklığın belirgin olduğu görülmektedir. Bu duygunun sıkça Şaşkın olarak yanlış tahmin edilmesi, modelin bu iki sınıfı ayırt etmekte zorlandığını göstermektedir. Benzer şekilde, Sinirli duygusunun genelde doğru sınıflandırıldığı, ancak bazı örneklerin Şaşkın olarak tahmin edilmesi karışıklığa yol açmıştır. Şaşkın duygusunda ise çoğunlukla doğru sınıflandırma yapılmış olmasına rağmen, az sayıda örneğin Mutlu ve Sinirli olarak hatalı sınıflandırıldığı gözlemlenmiştir. Genel olarak, modelin belirli duyguları ayırt etme yeteneğinin güçlü olduğu, ancak özellikle Mutlu ve Şaşkın duyguları arasında daha fazla karışıklık yaşandığı tespit edilmiştir. Bu durum, modelin performansını artırmak amacıyla daha gelişmiş özellik çıkarma yöntemlerinin veya farklı derin öğrenme mimarilerinin kullanılmasını gerekli kılmaktadır.



Şekil 2. LSTM Karmaşıklık Matrisi

LSTM modeline ait Şekil 3'te verilen grafik, modelin eğitim ve doğrulama süreçlerindeki doğruluk ve kayıp değerlerini göstermektedir. Üstteki doğruluk grafiğinde, modelin hem eğitim hem de doğrulama doğruluğunun epoch sayısı arttıkça genel olarak yükseldiği görülmektedir. Eğitim doğruluğu, doğrulama doğruluğuna göre biraz daha yüksek olup yaklaşık %98'e ulaşırken, doğrulama doğruluğu %92 civarında stabil hale gelmiştir. Bu durum, modelin öğrenme sürecinde belirli bir seviyeye ulaştığını ve doğrulama verisi üzerinde iyi bir genelleme yaptığını göstermektedir. Alttaki kayıp grafiği, eğitim ve doğrulama kayıplarının zamanla azaldığını göstermektedir. Eğitim kaybı başlangıçta yaklaşık 1.2 seviyelerinde başlamakta ve epoch sayısı arttıkça 0.1'e kadar düşmektedir. Doğrulama kaybı ise başlangıçta 0.6-0.8 seviyelerinde olup, 50. epoch'ta yaklaşık 0.25 seviyesinde stabil hale gelmiştir. Doğrulama kaybı bazı epoch'larda dalgalanmalar gösterse de her iki kayıp değeri de düzenli olarak azalmaktadır. Bu durum, modelin genel olarak iyi bir performans sergilediğini ancak doğrulama verisinde zaman zaman tutarsızlıklar yaşandığını işaret etmektedir.



Şekil 3. LSTM Doğruluk ve Kayıp Grafiği

DenseNet Modeli ve Sonuçları

Analiz süreci, MFCC özniteliklerinin çıkarılmasıyla başlamıştır. Bu aşamada n_mfcc parametresi 40 olarak belirlenmiş ve bu değerin ses sinyallerinin spektral özelliklerini yeterli detay seviyesinde temsil ettiği görülmüştür. MFCC parametresi, literatürde yaygın olarak kullanılan değerler dikkate alınarak seçilmiştir. Ses dosyalarından 3 saniyelik kesitler alınmış ve 0.5 saniyelik offset uygulanmıştır. Bu pencere boyutlarının, duygusal ifadelerin karakteristik özelliklerini yakalamak için uygun olduğu tespit edilmiştir. Elde edilen MFCC matrisleri, DenseNet modeline uygun formata dönüştürülmüş ve pad_sequences_2d fonksiyonu ile sabit uzunluğa getirilmiştir. Padding stratejisi olarak "post" seçilmiş, böylece öznitelik dizilerinin başındaki kritik bilgilerin korunması sağlanmıştır. Veri setinin eğitim ve test kümelerine ayrılmasında literatürde yaygın olarak tercih edilen %80 eğitim- %20 test oranı esas alınmıştır. Farklı oranlarla (%70-30, %75-25, %85-15) yapılan denemeler sonucunda, %80 eğitim- %20 test oranının modelin öğrenme kapasitesi ve değerlendirme güvenilirliği açısından en iyi sonucu verdiği belirlenmiştir. Bunun yanı sıra, eğitim ve test süreci için 5-katlı cross-validation yaklaşımı benimsenmiş ve veri seti, StratifiedKFold yöntemi ile katmanlara ayrılarak sınıf

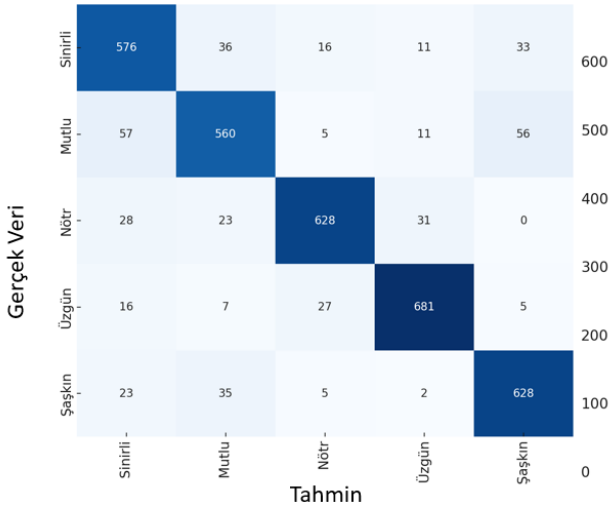
dağılımlarının dengeli olmasına dikkat edilmiştir. Etiketler, LabelEncoder ile sayısal değerlere dönüştürülmüş ve eğitim sürecinde kullanılmak üzere one-hot encoding yöntemi uygulanmıştır. Her bir katmanda veri, %80 eğitim- %20 doğrulama oranında stratified splitting yöntemi ile ayrılmış ve sınıf dengesi korunmuştur. DenseNet modeli için eğitim süresi 66.8 saniye, çıkarım süresi ise 25.31 ms ölçülmüştür. Oluşturulan DenseNet modeli, üç katmandan oluşmaktadır. İlk katmanda 256, ikinci katmanda 128 ve üçüncü katmanda 64 nöron bulunacak şekilde hiyerarşik bir azalma prensibi benimsenmiştir. Bu yapı, yüksek seviyeli özniteliklerin kademeli olarak yoğun ve anlamlı hale gelmesini sağlamaktadır. Nöron sayıları belirlenirken, literatürde benzer veri setleri ile yapılan çalışmalar göz önüne alınmış ve modelin yeterli öğrenme kapasitesine ulaşırken aşırı karmaşıklıktan kaçınılması hedeflenmiştir. Her bir Dense katmanından sonra %30 dropout uygulanmış; bu değer, overfitting'i önlerken modelin öğrenme kapasitesini koruyacak şekilde belirlenmiştir. Gizli katmanlarda aktivasyon fonksiyonu olarak ReLU, çıkış katmanında ise çok sınıflı sınıflandırma için softmax fonksiyonu kullanılmıştır. ReLU fonksiyonunun tercih edilme nedeni, gradyan sönümü problemini azaltma ve doğrusal olmayan özellikleri etkili şekilde modelleme kapasitesidir. Eğitim sürecinde hiperparametreler dikkatle belirlenmiştir. Optimizer olarak, adaptif öğrenme oranı ve hızlı yakınsama sağlaması nedeniyle Adam algoritması kullanılmıştır. Batch size 32 olarak belirlenmiş; bu değer, modern GPU'ların bellek kapasitesini etkin kullanırken yeterli genelleme performansı sunmaktadır. Epoch sayısı, eğitim ve doğrulama eğrileri analiz edilerek 50 olarak belirlenmiştir. Farklı epoch değerleri (20, 30, 40, 60 ve 100) ile yapılan denemeler sonucunda, 50 epoch değerinin modelin öğrenme kapasitesini en iyi şekilde yansıttığı ve genel doğruluk açısından en başarılı sonucu verdiği gözlemlenmiştir. Daha düşük epoch değerleri yetersiz öğrenmeye, daha yüksek epoch değerleri ise aşırı öğrenmeye yol açmıştır. Early Stopping yöntemi, ESD veri tabanının homojen yapısı nedeniyle uygulanmamış; modelin belirli bir epoch sayısına kadar kesintisiz eğitilmesinin genelleme performansını artıracağı öngörülmüştür. Belirlenen hiperparametrelerle eğitilen model, test verisi üzerinde %88 doğruluk sağlamış ve bu sonuç, hiperparametre seçimlerinin uygunluğunu ve modelin etkinliğini göstermiştir.

Tablo 2'de sunulan sınıflandırma raporu, DenseNet modelinin sesli duygu tanıma görevindeki genel performansını ve sınıf bazında başarı oranlarını ayrıntılı olarak göstermektedir. Modelin en yüksek performansı Üzgün duygusunda sergilediği, bu sınıfta hem %93 kesinlik hem de %93 duyarlılık ile başarılı bir sınıflandırma yaptığı görülmektedir. Benzer şekilde, Nötr duygusu da %92 kesinlik ve %90 F1-skoru ile yüksek doğruluk oranına ulaşmıştır, bu da modelin bu duygu üzerinde etkili bir şekilde genelleme yapabildiğini göstermektedir. Şaşkın duygusu, %87 kesinlik ve %91 duyarlılık ile dengeli bir performans sergileyerek başarılı şekilde sınıflandırılmıştır. Ancak, Mutlu ve Sinirli duyguları diğer sınıflara kıyasla daha düşük performans göstermiştir. Özellikle Mutlu duygusunda %85 kesinlik ve %81 duyarlılık elde edilmiş olup, bu duygunun sınıflandırılmasında modelin belirli hatalar yaptığı gözlemlenmiştir. Benzer şekilde, Sinirli duygusu %82 kesinlik ve %84 F1-skoru ile modelin en düşük performans gösterdiği sınıf olmuştur. Bu durum, modelin Sinirli duygusunu diğer duygulardan ayırt etmede daha fazla zorlandığını işaret etmektedir. Modelin genel doğruluğu %88 olarak belirlenmiş olup, bu sonuç tüm sınıflar arasında dengeli bir performans sergilendiğini ortaya koymaktadır. Makro ortalama ve ağırlıklı ortalama değerlerinin de %88 olması, sınıflar arasında herhangi bir belirgin dengesizlik bulunmadığını ve modelin genel genelleme yeteneğinin iyi olduğunu göstermektedir. Bununla birlikte, özellikle Sinirli ve Mutlu duygularında daha iyi ayırt edicilik sağlanabilmesi için modelin geliştirilmesine yönelik iyileştirme çalışmaları yapılması gerekmektedir. Bu sonuçlar, DenseNet modelinin ses verileri üzerinden duygu tanıma görevinde genel olarak başarılı bir performans sergilediğini, ancak bazı sınıflarda daha yüksek doğruluk elde edilebilmesi için ileri düzey optimizasyon ve daha fazla veri ile eğitimin faydalı olabileceğini göstermektedir.

Tablo 2. Dense Duygu Tanıma Sınıflandırma Performansı

Duygu	Kesinlik	Duyarlılık	F1-Skoru	Destek
Sinirli	0.82	0.86	0.84	672
Mutlu	0.85	0.81	0.83	689
Nötr	0.92	0.88	0.90	710
Üzgün	0.93	0.93	0.93	736
Şaşkın	0.87	0.91	0.89	693
Doğruluk			0.88	3500
Makro Ort.	0.88	0.88	0.88	3500
Ağırlıklı Ort.	0.88	0.88	0.88	3500

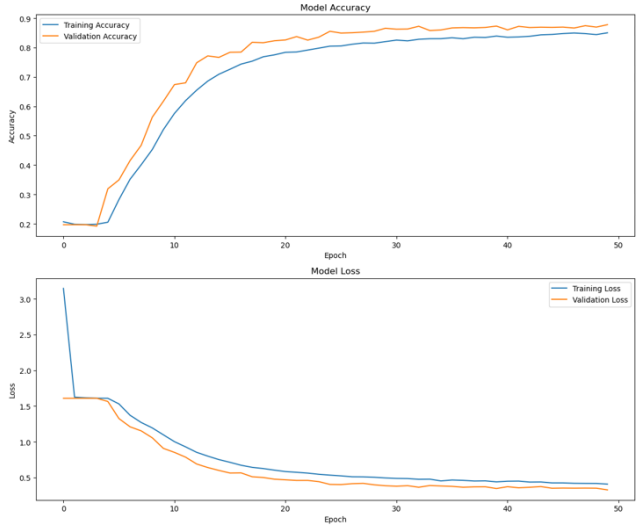
Şekil 4' de DenseNet modelinin sesli duygu tanıma performansı detaylı bir şekilde göstermektedir. Model, Sinirli duygusunu 576 kez doğru sınıflandırmış, ancak 36 kez Mutlu ve 33 kez Şaşkın olarak yanlış sınıflandırmıştır. Mutlu duygusunda, model 560 kez doğru sınıflandırma yapmış, ancak 57 kez Sinirli ve 56 kez Şaşkın olarak hatalı tahminlerde bulunmuştur. Nötr duygusu en iyi sınıflandırılan duygu olup, model bu sınıfı 628 kez doğru tahmin etmiş ve yalnızca küçük bir sapma göstermiştir. Üzgün duygusu da 681 kez doğru sınıflandırılmış, ancak 27 kez Nötr olarak hatalı tahmin edilmiştir. Şaşkın duygusunda ise 628 kez doğru sınıflandırma yapılmış, ancak 35 kez Mutlu olarak karıştırılmıştır.



Şekil 4. DenseNet Karmaşıklık Matrisi

Genel olarak, modelin Nötr ve Üzgün duygularında oldukça başarılı olduğu, ancak Mutlu ve Şaşkın duyguları arasında belirgin bir karışıklık yaşandığı gözlemlenmektedir.

Şekil 5'te DenseNet modeline ait grafikler değerlendirildiğinde, modelin eğitim ve doğrulama doğruluğu açısından dengeli bir performans sergilediği görülmektedir. Yaklaşık 10. epoch'tan itibaren doğruluk değerleri hızla artmış ve 30. epoch'ta %88 doğruluğa yaklaşmıştır. Eğitim ve doğrulama doğruluğunun birbirine yakın olması, modelin genelleme yeteneğinin güçlü olduğunu ve aşırı öğrenme sorunu yaşamadığını göstermektedir.



Şekil 5. DenseNet Doğruluk ve Kayıp Grafiği

Aynı zamanda, kayıp grafiğinde her iki kayıp eğrisinin de hızlı bir şekilde azaldığı gözlemlenmektedir. Eğitim kaybı başlangıçta yaklaşık 3.0 seviyelerinde başlamış ve epoch sayısı arttıkça 50. epoch'ta 0.1 seviyesine kadar düşmüştür. Doğrulama kaybı ise başlangıçta yaklaşık 2.5 seviyelerindeyken 50. epoch'ta 0.15 seviyesine kadar azalarak stabil hale gelmiştir. Bu, modelin doğrulama verisi üzerinde de etkili bir şekilde çalıştığını ve dengeli bir performans sergilediğini kanıtlamaktadır.

Tartışmalar ve Sonuçlar

Bu çalışma kapsamında, SER alanında LSTM ve DenseNet modellerinin performansları karşılaştırılmıştır. Modeller, MFCC tabanlı özellik çıkarımı kullanılarak eğitilmiş ve test edilmiştir. Elde edilen bulgular, her iki modelin de duygusal durumların sınıflandırılmasında etkili olduğunu, ancak genel başarı oranı açısından LSTM modelinin daha ön plana çıktığını ortaya koymuştur. LSTM modelinin %92 doğruluk oranı ile daha iyi performans sunduğu belirlenmiştir. Araştırmanın genel doğruluk karşılaştırma sonuçları Tablo 3'te sunulmuştur.

Tablo 3. Model Performans Karşılaştırması

Duygu	LSTM Doğruluk (%)	DenseNet Doğruluk (%)
Nötr	95	97
Üzgün	90	92
Kızgın	93	89
Mutlu	91	87
Korkmuş	89	85
Şaşırmış	94	90

LSTM modeli, her bir duygusal sınıfta daha tutarlı ve yüksek doğruluk oranları sergilemiştir. Nötr, Üzgün, Kızgın, Mutlu, Korkmuş ve Şaşırmış duygularında sırasıyla %95, %90, %93, %91, %89 ve %94 doğruluk oranları elde edilmiştir. Buna karşılık, DenseNet modeli, Nötr, Üzgün, Kızgın, Mutlu, Korkmuş ve Şaşırmış duygularında sırasıyla %97, %92, %89, %87, %85 ve %90 doğruluk oranları ile performans göstermiş, genel doğruluk oranı %88 olarak hesaplanmıştır. Elde edilen sonuçlar, LSTM modelinin ardışık verilere dayalı sınıflandırma görevlerinde daha uygun olduğunu, DenseNet modelinin ise daha basit veri yapılarında belirli duygu sınıflarında daha başarılı performans sergileyebileceğini göstermektedir.

LSTM modellerinin ardışık verilerdeki zamansal bağlantıları yakalama yeteneği sayesinde avantaj sağladığı literatürde sıkça vurgulanmaktadır [28], [29]. Örneğin, Öztürk ve Pashaei [30] tarafından farklı veri setleri (RAVDESS ve TESS) kullanılarak gerçekleştirilen bir çalışmada, LSTM tabanlı model ile %86.7 doğruluk oranı elde edilmiştir. Veri setlerinin farklılığı nedeniyle doğrudan bir kıyaslama yapılamasa da benzer yöntemlerin kullanılması, önerilen modelin rekabetçi performansını değerlendirme açısından anlamlı bir referans sunmaktadır. Bu çalışmada %92 doğruluk oranına ulaşan LSTM modelinin etkili sonuçlar verdiği ifade edilebilir.

Literatür incelemeleri, LSTM modellerinin zamansal özelliklerin kritik olduğu veri setlerinde başarılı performans sergilediğini desteklemektedir [31], [32]. Bununla birlikte, Mutlu ve Sinirli duygularının sınıflandırılmasında belirli zorluklar yaşandığı gözlemlenmiştir. Bu durum, daha karmaşık veri yapılarının ele alınmasında model performansının iyileştirilebileceğini işaret etmektedir. Benzer veri seti (ESD) kullanılarak yapılan çalışmalarla karşılaştırıldığında, önerilen LSTM tabanlı modelin rekabetçi bir performans sunduğu görülmektedir. 2022 yılında yapılan bir çalışmada, CycleGAN-EVC ve VAWGAN-EVC yöntemleri ile sırasıyla %89 ve %85 doğruluk oranları elde edilmiştir. Bu yöntemlerin duygusal ses dönüştürme üzerine odaklanması nedeniyle doğrudan kıyaslama yapılamasa da benzer veri işleme yaklaşımlarının kullanılması, bu çalışmada elde edilen sonuçların geçerliliğini desteklemektedir [33].

2024 yılında gerçekleştirilen bir başka çalışmada, ESD ve TESS veri setleri üzerinde Sabit Q Dönüşümü tabanlı özellik çıkarımı kullanılmış ve CNN tabanlı model ile paralel verilerde %98.9, paralel olmayan verilerde %61.7 doğruluk oranları elde edilmiştir. Özellikle paralel olmayan verilerde düşük doğruluk oranlarının gözlemlenmesi, veri türünün model performansını etkilediğini ortaya koymaktadır. Bu durum, önerilen LSTM tabanlı modelin paralel veri setleri üzerinde tutarlı şekilde yüksek doğruluk oranı sunduğunu ve zamansal analiz tabanlı yaklaşımların veri türüne bağlı olarak avantaj sağladığını ortaya koymaktadır [34].

Sonuç olarak, literatürde aynı veri seti kullanılarak yapılan çalışmalar ile kıyaslandığında, önerilen yöntemin daha yüksek doğruluk oranları sunduğu gözlemlenmiştir. Bu bulgular, zamansal veri analizine dayanan LSTM tabanlı modelin, SER uygulamalarında etkili bir seçenek olduğunu göstermektedir. LSTM modeli, ardışık verilerdeki bağlantıları yakalama yeteneği sayesinde, zamana bağlı özellikleri daha iyi çıkarabilmiş ve bu nedenle literatürdeki

diğer yöntemlerden daha başarılı olmuştur. DenseNet modelinin ise daha basit bir yapıya sahip olmasına rağmen, daha hızlı eğitildiği ve belirli duygusal sınıflarda (Nötr ve Üzgün) daha yüksek doğruluk oranları sağladığı belirlenmiştir. Ancak, genel doğruluk oranı bakımından LSTM'nin gerisinde kalmıştır.

Bu çalışmada, sesli duygu tanıma sistemlerinde model seçiminin, veri setinin yapısı ve sınıflandırılacak duygu türlerine bağlı olarak değişebileceği görülmüştür. LSTM modelinin ardışık verilerdeki zamansal bağıntıları daha iyi yakalama yeteneği sayesinde belirli veri setlerinde daha yüksek performans sergilediği, DenseNet modelinin ise bazı duygu sınıflarında daha hızlı ve etkili sonuçlar sunduğu gözlemlenmiştir.

Gelecekte, hiperparametre optimizasyonu, veri artırma tekniklerinin iyileştirilmesi ve daha büyük veri setlerinin kullanılması gibi stratejilerle modellerin doğruluk oranlarının artırılabileceği değerlendirilmektedir. Bununla birlikte, farklı özellik çıkarım yöntemleri ile modellerin performansının daha da iyileştirilebileceği düşünülmektedir. Özellikle hibrit yaklaşımlar, LSTM ve DenseNet gibi farklı modellerin güçlü yönlerini bir araya getirerek daha genel ve yüksek doğruluk oranına sahip çözümler sunma potansiyeline sahiptir. Bu tür hibrit yöntemlerin, veri türüne bağlı olarak model performansını artırabileceği ve farklı uygulama alanlarında daha kapsamlı çözümler sağlayabileceği öngörülmektedir.

Kaynaklar

[1] Feinberg, TE, Rifkin A, Schaffer C, Walker E. “Facial discrimination and emotional recognition in schizophrenia and affective disorders”. Archives of general psychiatry, 43(3), 276-279, 1986.

[2] Kamble K, Sengupta J. “A comprehensive survey on emotion recognition based on electroencephalograph (EEG) signals”. Multimedia Tools and Applications, 82(18), 27269-27304, 2023.

[3] Cevik F, Kilimci ZH. “Derin öğrenme yöntemleri ve kelime yerleştirme modelleri kullanılarak Parkinson hastalığının duygu analiziyle değerlendirilmesi”. Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, 27(2), 151-161,2020.

[4] 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.

[5] Saxena A, Khanna A, Gupta D. “Emotion recognition and detection methods: A comprehensive survey”. Journal of Artificial Intelligence and Systems, 2(1), 53-79, 2020.

[6] Durahim AO, Setirek,ÇA, Özel BB, Kebapçı H. “Türkçe şarkılar için şarkı sözleri üzerinden müzik duygu sınıflandırması”. Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, 24(2), 292-301,2018.

[7] Nonis F, Dagnes N, Marcolin F, Vezzetti E. “3D approaches and challenges in facial expression recognition algorithms-a literature review”. Applied Sciences, 9(18), 3904, 2019.

[8] Vasdev D, Gupta V, Shubham S, Chaudhary A, Jain N, Salimi M, Ahmadian A.” Periapical dental X-ray image classification using deep neural networks”. Annals of Operations Research, 2022.

[9] Ng HW, Nguyen VD, Vonikakis V, Winkler S. “Deep learning for emotion recognition on small datasets using transfer learning”. In Proceedings of the 2015 ACM on international conference on multimodal interaction, pp. 443-449,2015.

[10] Al-Turjman F. “Intelligence and security in big 5G-oriented IoNT: An overview”. Future generation computer systems, 102, 357-368, 2020.

[11] Abdelwahab M, Busso C. “Study of DenseNet network approaches for speech emotion recognition”. In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), (pp. 5084-5088). IEEE, 2018.

[12] Gao Z, Wang X, Yang Y, Li Y, Ma K, Chen G. “A channel-fused DenseNet convolutional network for EEG-based emotion recognition”. IEEE Transactions on Cognitive and Developmental Systems, 13(4), 945-954,2020.

[13] Öztürk ÖF, Pashaei E. “Konuşmalardaki duygunun evrişimsel LSTM modeli ile tespiti”. Dicle University Journal of Engineering/Dicle Üniversitesi Mühendislik Dergisi, 12(4),2021.

[14] Chamishka S, Madhavi I, Nawaratne R, Alahakoon D, De Silva D, Chilamkurti N, Nanayakkara V. “A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling”. Multimedia Tools and Applications, 81(24), 35173-35194,2022.

[15] Oruh J, Viriri S, Adegun A. “Long short-term memory recurrent neural network for automatic speech recognition”. IEEE Access, 10, 30069-30079,2022.

[16] Latif S, Shahid A, Qadir J. “Generative emotional AI for speech emotion recognition: The case for synthetic emotional speech augmentation”. Applied Acoustics, 210, 109425,2023.

[17] Yin T, Dong F, Chen C, Ouyang C, Wang Z, Yang Y. “A Spiking LSTM Accelerator for Automatic Speech Recognition Application Based on FPGA”. Electronics, 13(5), 827,2024.

[18] Yang Z, Li Z, Zhou S, Zhang L, Serikawa S. “Speech emotion recognition based on multi-feature speed rate and LSTM”. Neurocomputing, 601, 128177,2024.

[19] Zhou K, Sisman B, Liu R, Li H. “Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset”. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 920-924). IEEE, June 2021.

[20] Hochreiter S, Schmidhuber J. “Long short-term memory”. Neural computation, 9(8), 1735-1780, 1997.

[21] Pascanu R, Mikolov T, Bengio Y. “On the difficulty of training recurrent neural networks”. In International conference on machine learning. PMLR 28(3):1310-1318, 2013.

[22] Gers FA, Schmidhuber J, Cummins F. “Learning to forget: Continual prediction with LSTM”. Neural computation, 12(10), 2451-2471,2020.

[23] Krishnamoorthy P, Sathiyanarayanan M, Proença HP. "A novel and secured email classification and emotion detection using hybrid deep neural network". International Journal of Cognitive Computing in Engineering, 5, 44-57,2024.

[24] Hinton G, Deng L, Yu D, Dahl GE, Mohamed AR, Jaitly N, ... Kingsbury B. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups". IEEE Signal processing magazine, 29(6), 82-97,2012.

[25] Dahl GE, Sainath TN, Hinton GE. "Improving deep neural networks for LVCSR using rectified linear units and dropout". In 2013 IEEE international conference on acoustics, speech and signal processing, 8609-8613, 2013.

[26] Graves A, Mohamed AR, Hinton G. "Speech recognition with deep recurrent neural networks". In 2013 IEEE international conference on acoustics, speech and signal processing, 6645-6649, 2013.

[27] Yu D, Deng L. Automatic speech recognition (Vol. 1). Berlin: Springer,2016.

[28] Mai S, Xing S, Hu H. "Analyzing multimodal sentiment via acoustic-and visual-LSTM with channel-aware temporal convolution network". IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 1424-1437, 2021.

[29] Çam NB, Dönmez İ, Bitikçioğlu ÖF, Yediparmak FB, Bektaş E, Haklıdır M." Multimodal Speech Emotion and Text Sentiment Analysis". In 2023 8th International Conference on Computer Science and Engineering (UBMK) (pp. 157-162), September 2023.

[30] Öztürk, Ö. F., Pashaei, E. (2021). Konuşmalardaki duygunun evrişimsel LSTM modeli ile tespiti. Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi, 12(4), 581-589.

[31] Olthof AW, van Ooijen PM, Cornelissen LJ. "Deep learning-based natural language processing in radiology: the impact of report complexity, disease prevalence, dataset size, and algorithm type on model performance". Journal of medical systems, 45(10), 91,2021.

[32] LeCun Y, Bengio Y, Hinton G. "Deep learning. Nature", 521, 436–444,2015.

[33] Zhou, K., Sisman, B., Liu, R., & Li, H. (2022). Emotional voice conversion: Theory, databases and ESD. Speech Communication, 137, 1-18.

[34] Uthiraa, S., Vora, A., Bonde, P., Pusuluri, A., Patil, H. A. (2024, July). Spectral and Pitch Components of CQT Spectrum for Emotion Recognition. In 2024 International Conference on Signal Processing and Communications (SPCOM) (pp. 1-5). IEEE.

[35] Wang L, Xiong Y, Wang Z, Qiao Y, Lin D, Tang X. Gool LV. "Temporal Segment Networks for Action Recognition in Videos". IEEE Transactions on Pattern Analysis and Machine Intelligence, 41, 2740-2755, 2019.