**Iğdır Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 14(3), 974-987, 2024**
***Journal of the Institute of Science and Technology, 14(3), 974-987, 2024***

**ISSN: 2146-0574, eISSN: 2536-4618**

**Bilgisayar Mühendisliği / *Computer Engineering*** **DOI: 10.21597/jist.1505349**

**Araştırma Makalesi / *Research Article***

**Geliş tarihi / *Received*: 26.06.2024** **Kabul tarihi / *Accepted*: 21.07.2024**

**Atıf İçin:** Yücesoy, E. (2024). Konuşmacıları Kadın, Erkek ve Çocuk Olarak Sınıflandırmada Veri Artırmanın Performansa Etkisi. *Iğdır Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 14(3), 974-987.

***To Cite:*** Yücesoy, E. (2024). Effect of Data Augmentation on Performance in Classifying Speakers as Female, Male and Child. *Journal of the Institute of Science and Technology*, 14(3), 974-987.

# Konuşmacıları Kadın, Erkek ve Çocuk Olarak Sınıflandırmada Veri Artırmanın Performansa Etkisi

Ergün YÜCESOY[1]

**Öne Çıkanlar:**
- Derin öğrenme ile yaş ve cinsiyet tanıma
- Veri artırma yöntemlerinin karşılaştırılması
- Veri artırma ile performans artışı

**Anahtar Kelimeler:**
- Yaş ve cinsiyet Tanıma
- Evrişimli sinir ağları
- Veri artırma
- Perde kaydırma
- Zaman uzatma
- Gürültü ekleme

**ÖZET:**

Derin öğrenme alanındaki gelişmeler daha doğru sınıflandırıcıların oluşturulmasına olanak sağlamıştır. Ancak yüksek genelleme yeteneğine sahip derin öğrenme modellerinin oluşturulabilmesi için büyük miktarda etiketli veri kümelerine ihtiyaç duyulmaktadır. Veri artırma bu ihtiyacın karşılanmasında yaygın olarak kullanılan bir yöntemdir. Bu çalışmada konuşmacıların yaş ve cinsiyetlerine göre sınıflandırılmasında farklı veri artırma yöntemlerinin sınıflandırma performansı üzerindeki etkileri araştırılmıştır. Çalışmada yetişkin konuşmacılar erkek ve kadın olarak, çocuklar ise cinsiyet ayrımı yapılmadan tek bir sınıf olarak değerlendirilmiş ve toplamda üç (kadın, erkek ve çocuk) sınıflı bir sınıflandırma gerçekleştirilmiştir. Bu amaç doğrultusunda gürültü ekleme, zaman uzatma ve perde kaydırma olmak üzere üç veri artırma yöntemi farklı kombinasyonlarda kullanılarak yedi farklı model oluşturulmuş ve her birinin performans ölçümleri yapılmıştır. aGender veri kümesinden rastgele seçilen 5760 konuşma verisi ile geliştirilen bu modeller arasında en yüksek performans artışı üç veri artırma yönteminin birlikte kullanıldığı modelle sağlanmıştır. Bu model sınıflandırma doğruluğunu %84.583'den % 87.523'e çıkararak %3'e yakın performans artışı sağlarken veri artırmanın kullanıldığı diğer modellerde de %1 ile %2.3 arasında performans artışı sağlanmıştır.

# Effect of Data Augmentation on Performance in Classifying Speakers as Female, Male, and Child

**Highlights:**
- Age and gender recognition with Deep Learning
- Comparison of data augmentation methods
- Performance improvement with data augmentation

**Keywords:**
- Age and gender recognition
- Convolutional neural networks
- Data augmentation
- Pitch shift
- Time stretching
- Noise addition

**ABSTRACT:**

Developments in the field of deep learning have enabled the creation of more accurate classifiers. However, large amounts of labeled datasets are needed to create deep learning models with high generalization ability. Data augmentation is a widely used method to address the need for more data. This study investigates the effects of different data augmentation methods on the classification performance of speakers based on their age and gender. In this study, adult speakers are classified as male or female, while children are classified as a single group without gender discrimination, resulting in a total of three classes (female, male, and child). For this purpose, seven different models are created using combinations of three data augmentation methods: noise addition, time stretching, and pitch shifting. The performance of each model is then evaluated. Among these models, which were developed with 5760 speech data randomly selected from the aGender dataset, the highest performance increase is achieved with the model where three data augmentation methods are used together. This model increases the classification accuracy from 84.583% to 87.523%, providing a performance increase of nearly 3%, while other models using data augmentation provide a performance increase of 1% to 2.3%.

[1] Ergün YÜCESOY (**Orcid ID:** 0000-0003-1707-384X) Ordu Üniversitesi, Teknik Bilimler Meslek Yüksekokulu, Elektronik ve Otomasyon Bölümü, Ordu, Türkiye

***Sorumlu Yazar/*Corresponding Author*:** Ergün YÜCESOY, e-mail: yucesoye@odu.edu.tr

## GİRİŞ

İnsanlar iletişim kurduğu kişinin yaşına ve cinsiyetine bağlı olarak davranışlarını uyarlamaya alışkındır. Benzer şekilde, bir sistem, bu bağlamsal bilgileri karar verme süreçlerine dahil ederek kullanıcılarına daha uygun yanıtlar sağlayabilir (Lingenfelser ve ark., 2010). Konuşma sinyalinden konuşmacının yaş ve cinsiyetinin belirlenmesi gerçek hayatta önemli uygulamaları olan oldukça zorlu bir görevdir. Konuşmacı segmentasyonu, konuşmacı tanıma, insan-bilgisayar etkileşimi (HCI), oyun ve mobil uygulamalar, müşteri hizmetleri uygulamaları, etkileşimli akıllı sesli asistanlar, güvenlik sistemleri ve sesli yanıt sistemleri bu uygulamalardan bazılarıdır (Tursunov ve ark., 2021; Mavaddati, 2024; Sánchez-Hevia ve ark., 2022; Kwasny & Hemmerling, 2021).

Literatürde bulunan yaş ve cinsiyet sınıflandırma sistemlerinin çoğu yetişkinler için tasarlanmıştır. Ancak çocukların yer aldığı birçok senaryo için yetişkin ile çocuk konuşmaları arasında ayrımın yapılması gerekmektedir. Çocukların konuşmasının akustik ve dilsel özelliklerinin yetişkinlerin konuşmasından oldukça farklı olduğu iyi bilinmektedir. Örneğin; çocukların konuşması, yetişkinlerin konuşmasına göre daha yüksek perde ve formant frekanslarıyla karakterize edilir. Ayrıca, çocuğun büyümesi sırasında meydana gelen anatomik ve fizyolojik değişiklikler nedeniyle çocukların konuşma özellikleri, yaşın bir fonksiyonu olarak hızla değişir ve yaşla birlikte çocuklar artikülasyonda daha yetenekli hale gelir. Yetişkinler için geliştirilen bir ASR sisteminin performansının, çocukların konuşmasını tanımak için kullanıldığında büyük ölçüde azalmasının nedeni budur. Ayrıca çocukların konuşması üzerine eğitilmiş bir tanıma sistemi kullanıldığında bile sistemin çocuklar için performansı genellikle yetişkinlerinkinden daha düşüktür (Potamianos ve Narayanan, 2003; Gerosa ve ark., 2005).

Konuşmacının yaş ve cinsiyet gibi dil dışı (paralinguistik) içeriklerinin tanınmasına yönelik geçmişten günümüze pek çok çalışma yapılmıştır. Daha önceki yıllarda genel olarak konuşma sinyalinden elde edilen Mel-frekans kepstral katsayıları (MFCC) ve algısal doğrusal tahmin (PLP) gibi bazı akustik parametrelerin ortalamasının alınarak destek vektörü gibi algoritmalar ile sınıflandırıldığı çalışmalar yapılmıştır (Mahmoodi ve ark., 2011). Daha sonra değişken uzunluktaki ifadelerin sabit boyutlu bir gömme vektörüne (embedding vector) yerleştirilerek bu vektörün olasılıksal doğrusal diskriminant analizi (PLDA) gibi bir sınıflandırıcı ile sınıflandırılması fikrine dayanan bir yaklaşım ortaya çıkmıştır (Dehak ve ark., 2011; Li ve ark., 2013). Son yıllarda ise konuşmaya dayalı tanıma sistemlerinin de dahil olduğu birçok alanda derin öğrenme yaklaşımları ana akım olarak kullanılmaya başlanmıştır (Chai ve ark., 2021). Jasuja ve ark. (2020) tarafından yapılan çalışmada konuşmacının cinsiyetini tespit etmek için çok katmanlı algılayıcı (MLP) tabanlı bir derin öğrenme modeli önerilmiştir. Çalışmada, önerilen model farklı parametrelerle eğitilmiş ve test veri seti üzerinde %96 doğruluk ile sınıflandırma gerçekleştirilmiştir. Uddin ve ark. (2020) tarafından yapılan çalışmada ise ön işleme ile gürültüsüz düzgün veriler elde edilerek çok katmanlı bir mimari ile özellikler çıkarılmıştır. TIMIT, RAVDESS ve BGC olmak üzer üç farklı veri seti üzerinde yapılan deneylerde k-en yakın komşuluk sınıflandırıcısı (KNN) ile TIMIT veri kümesi üzerinde %96,8 cinsiyet tanıma doğruluğu elde edilmiştir. Diğer bir çalışmada ise bir ses veri setinden cinsiyet tahmini için Daha Derin Uzun Kısa Süreli Bellek (LSTM) Ağları yapısına dayalı bir yöntem önerilmiş ve bu yöntem kullanılarak %98,4 doğruluk ile cinsiyet tespiti yapılmıştır (Ertam, 2019).

Ancak bu çalışmaların hiçbirinde çocuk konuşmacılar kullanılmamıştır. Levitan ve ark. (2016) tarafından yapılan çalışmada temel frekans, enerji, ses kalitesi ve MFCC öznitelikleri kullanılarak lojistik regresyon, doğrusal regresyon, rastgele orman ve AdaBoost olmak üzere dört model geliştirilmiştir. Yapılan testlerde rastgele orman modeli konuşmacıları çocuk, erkek ve kadın olarak %85 doğruluk ile sınıflandırarak en başarılı model olmuştur. Üç sınıflı sınıflandırmanın

gerçekleştirildiği bir diğer çalışmada ise akustik çerçeve tabanlı özellikler ile ifadeye dayalı akustik, prosodik ve ses kalitesi özellikleri birlikte kullanılmıştır (Kockmann ve ark., 2010). İlgili çalışmada modelleme, Gauss Karışım Modellerine (GMM) ve Destek Vektör Makinelerine (SVM) ve ardından doğrusal Gauss arka uçlarına ve lojistik regresyon temelli füzyona dayalı olarak yapılmıştır. Çalışmada birkaç alt sistemin kombinasyonu ile üç sınıflı cinsiyet sınıflandırma kategorisinde %81.82 başarı sağlanmıştır. Yücesoy ve Nabiyev (2016) tarafından yapılan benzer bir çalışmada ise konuşmacıları yaş ve cinsiyetlerine göre sınıflandırmak için üç alt sistemin (GMM, SV, GMM-SV temelli SVM) birleşimine dayalı yeni bir model önerilmiştir. Kısa süreli telefon konuşmalarından elde edilen prosodik ve spektral öznitelikler kullanılarak geliştirilen bu modelin üç sınıflı cinsiyet sınıflandırma başarısı %90.39 olarak rapor edilmiştir. Vlaj ve Zgank (2022) tarafından yapılan çalışmada ise akıllı ev ortamında insan-bilgisayar etkileşimi için iki aşamalı bir yaş ve cinsiyet sınıflandırma sistemi önerilmiştir. Çalışmada yüksek karmaşıklıktaki özellik çıkarım yöntemleri kullanılmadan akustik sınıflandırma amaçlanmıştır. İlk aşamada GMM'lere dayalı bir sınıflandırma yapılarak her çerçevenin ilgili sınıflara ait olma olasılıkları hesaplanmıştır. Daha sonra bu olasılıklara göre her sınıfa ait kare sayısı sayılmış ve fark dörtten az ise ikinci aşamaya geçilmiştir. İkinci aşamada ise normalize edilmiş perde değerlerinin sayısı ve bu değerlerin toplamına göre sınıflandırma yapılmıştır. Son aşamada ise çoğunluk oylamasına göre konuşmacının sınıfı belirlenmiştir. TIDIGITS veri kümesi kullanılarak yapılan testlerde önerilen modelin sınıflandırma başarısı %92.25 olarak ölçülmüştür.

Derin öğrenme yaklaşımları, ham girdi verilerinden ayırt edici özellikleri otomatik olarak öğrenerek en zorlu sorunları yüksek doğrulukla çözebilmektedir. Bununla birlikte derin öğrenme modelleri, karmaşık kalıpları öğrenebilmek için büyük miktarda eğitim verisine ihtiyaç duyarlar ve veri kalitesi, modelin doğruluğunu doğrudan etkiler (Sarker, 2021). Ancak model eğitimi için her zaman yeterli miktarda veri bulunmayabilir. Bu durumda mevcut veriler üzerinde çeşitli veri artırma yöntemleri kullanılarak veri kümesi genişletilir ve model eğitimi bu verilerle gerçekleştirilir. Ses verilerinin artırılması için kullanılan çeşitli yöntemler vardır. Bu çalışmada gürültü ekleme, perde kaydırma ve zaman uzatma olmak üzere üç farklı veri artırma yönteminin yaş ve cinsiyet tanıma performansı üzerindeki etkileri incelenmiştir. Her bir yöntem önce ayrı ayrı daha sonra ise farklı kombinasyonlarda kullanılarak yaş ve cinsiyet sınıflandırma görevi için en iyi performansa sahip model belirlenmiştir. Literatürde yaş ve cinsiyet sınıflandırma görevine yönelik olarak veri artırma yöntemlerinin ayrı ayrı ele alındığı ve performanslarının karşılaştırıldığı benzer bir çalışmaya rastlanmamıştır. Çalışmanın bu yönüyle literatüre katkı sağlayacağı değerlendirilmiştir.

## MATERYAL VE METOT

### aGender Veri Kümesi

Çalışmada önerilen yaş ve cinsiyet sınıflandırma modelinin geliştirilmesinde aGender veri kümesi kullanılmıştır. aGender, 954 ücretli katılımcı tarafından kamuya açık telefon hatları üzerinden seslendirilen toplam 47 saatlik kayıtlı ve serbest metin kayıtlarından oluşmaktadır. Veri kümesindeki erkek ve kadın katılımcıların dört yaş grubuna göre dağılımı eşit olup, çocuklarda cinsiyet ayrımı yapılmamıştır. Metin içerikleri, otomatik ses hizmetlerine özgü olacak şekilde tasarlanmış ve çoğunlukla kısa komutlardan, tek sözcüklerden ve sayılardan oluşturulmuştur. Veri kümesindeki her bir kayıt çocuk, genç kadınlar, genç erkekler, olgun kadınlar, olgun erkekler, yaşlı kadınlar ve yaşlı erkekler olmak üzere yedi kategoriden birisiyle etiketlenmiştir. Toplam 65364 kayıttan oluşan veri kümesinde ortalama kayıt uzunluğu 2.58 saniye olup uzunluğu 1 saniyeden kısa kayıtlar da vardır. Ancak konuşmaya dayalı birçok tanıma sisteminde belirli uzunluktaki konuşma verileri kullanılmaktadır. Bu süre için bir standart olmamakla birlikte model eğitimleri için genellikle 3 - 5

saniye uzunluğundaki konuşma verileri kullanılmaktadır. Bu çalışmada kullanılan veriler için bir alt sınır belirlenmiş ve uzunluğu 3 saniyeden kısa olan kayıtlar veri kümesinden çıkarılmıştır. Veri kümesine bir diğer sınırlama da kayıtların dağılımı ile ilgili olarak getirilmiş ve her sınıftan (erkek, kadın ve çocuk) eşit sayıda kayıt olacak şekilde veri kümesi yeniden düzenlenmiştir. Oluşturulan veri kümesinin detayları Çizelge 1'de verilmiş olup önerilen sınıflandırma modelinin geliştirilmesinde bu veri kümesi kullanılmıştır.

**Çizelge 1.** Çalışmada kullanılan veri kümesi

| Sınıf Etiketleri | Yaş Aralıkları (Yıl) | Kayıt Sayısı |
|---|---|---|
| Çocuk | 7-14 | 1920 |
| Erkek | 15-80 | 1920 |
| Kadın | 15-80 | 1920 |

## Log-mel Spectrogram

Bu çalışmada konuşmacıların yaş ve cinsiyet sınıflarının belirlenmesinde konuşma sinyallerinden çıkarılan log-mel spectrogram öznitelikleri kullanılmıştır. Log-mel spectrogram, logaritmik bir ölçek kullanarak konuşma sinyalinin frekans bileşenlerinin zaman içinde nasıl değiştiğinin görsel bir temsilini sağlar (Zhang ve ark., 2019). İlk olarak konuşma sinyalinin kısa süreli bölümleri üzerinde kısa süreli Fourier dönüşümü (STFT) uygulanarak konuşmanın spektrogramı elde edilir. Daha sonra Mel ölçeğine göre tasarlanmış bir filtre bankası kullanılarak, spektrogram Mel ölçeğine taşınır. Mel ölçeği insanın işitsel sistemini taklit edebilen bir ses perdesi ölçeğidir ve matematiksel olarak aşağıdaki ifade ile temsil edilir (Qureshi ve ark., 2022).

$$m = 2595 * log_{10}(1 + \frac{f}{700}) \quad (1)$$

burada $f$, Hz cinsinden fiziksel frekansı, $m$ ise Mel ölçeğinde algılanan frekansı belirtir. Daha sonra logaritma fonksiyonu uygulanarak sinyalin log-mel spektrogramı elde edilir. Log-mel spektrogram ses tanıma ve ses sınıflandırması dahil olmak üzere çeşitli alanlarda yaygın olarak kullanılmaktadır.

## Veri artırma

Veri artırma, mevcut verilere çeşitli dönüşümler veya modifikasyonlar uygulanarak eğitim veri kümesinin boyutunu ve çeşitliliğini yapay olarak artırmak için makine öğreniminde yaygın olarak kullanılan bir yöntemdir (Nanthini ve ark., 2023). Bu yöntemin temel amacı orijinal verilerden yeni eğitim örnekleri oluşturmak ve böylece makine öğrenim modellerinin performansını ve sağlamlığını arttırmaktır. Veri artırmanın bir diğer amacı da gerçek dünya senaryolarını simüle eden dönüşümler sağlayarak modellerin farklı koşullara karşı daha dayanıklı hale getirilmesidir. Bu yöntem, modele öğrenebileceği daha fazla örnek sunarak aşırı uyum sorununu azaltılmasına da yardımcı olur. Ayrıca belirli sınıfların yeterince temsil edilmediği veri kümelerinde, azınlık sınıfların ek örneklerini oluşturarak dağılımı dengelemek ve böylece baskın sınıflara yönelik önyargıyı önlemek için de kullanılır (Wei ve ark., 2023). Bu çalışmada üç farklı veri artırma yöntemi kullanılmıştır ve bu yöntemlerin kısa bir açıklaması aşağıda verilmiştir.

## Gürültü ekleme

Bir paradoks gibi görünse de verilere gürültü eklemek düzenlileştirme işlevi görür ve genelleştirmeyi arttırır (Bishop, 1995). Gaussian gürültüsü ekleme, eğitim verilerinin miktarını ve çeşitliliğini arttırmak için yaygın olarak kullanılan bir veri artırma yöntemidir (Miliaresi ve ark., 2021; Liu ve ark., 2022). Gaussian gürültüsü $n(t)$, ortalaması 0 ve standart sapması 1 olan bir fonksiyon olup rastgele sayı üreticisiyle kolayca üretilir. Denklem (2) ile ifade edilen gürültü ekleme sürecinde

standart Gaussian'dan örneklenen gürültü örnekleri bir genlikle çarpıldıktan sonra veri noktalarına eklenerek yeni örnekler oluşturulur.

$$x(t) = x(t) + \sigma \, x \, n(t) \quad (2)$$

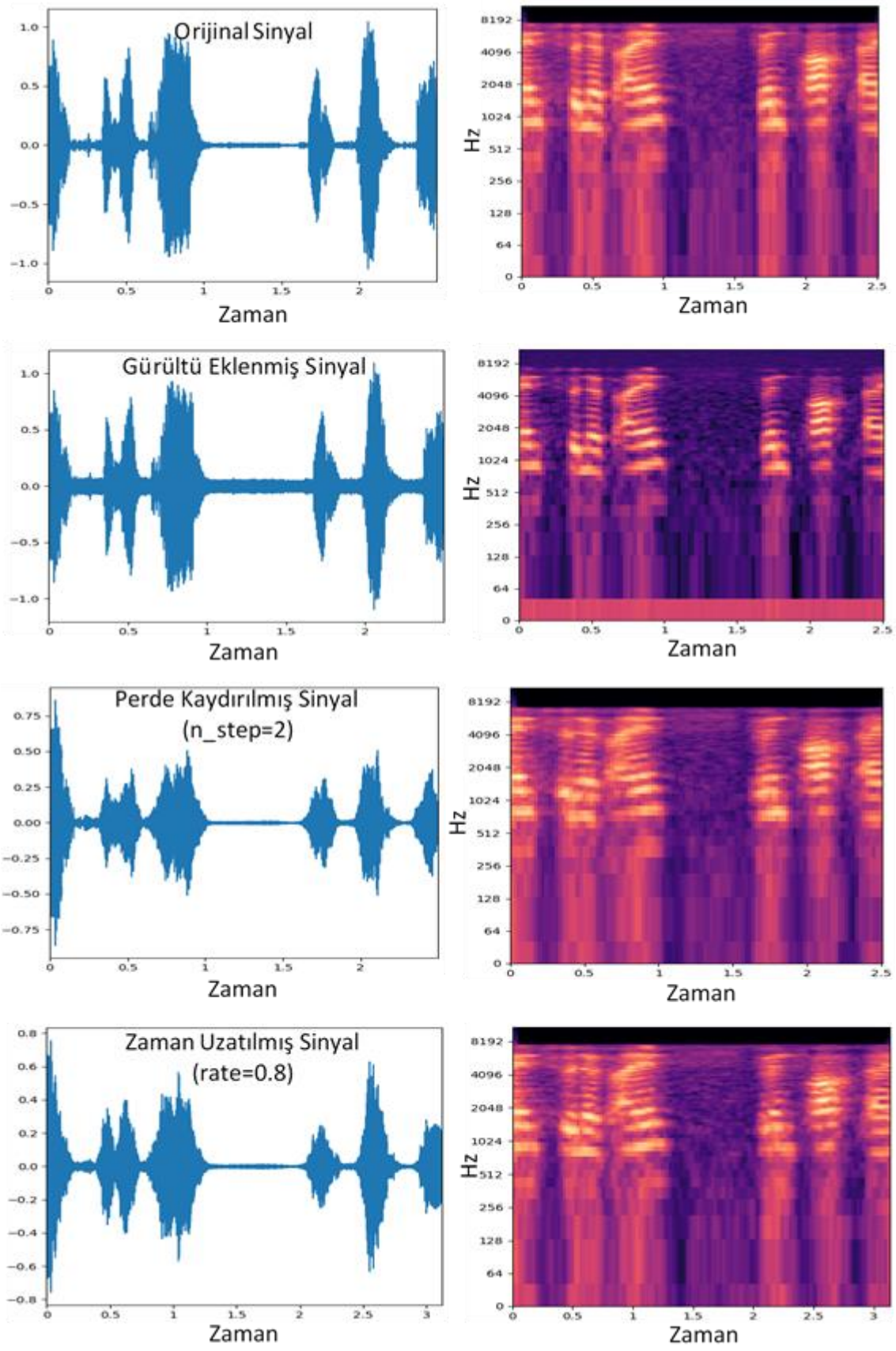
burada $n(t)$ Gaussian gürültüsü, $\sigma$ ise gürültü genliği olup belirlenmesi gereken bir hiper parametredir. Gürültü genliği çok küçük olduğunda sınıflandırıcının rahatsız edilmesi zorlaşırken, çok büyük olduğunda ise sınıflandırıcının öğrenmesi zorlaşır. Bu çalışmada gürültü genliğinin kabul edilebilir aralığı [0.005-0.015] olarak belirlenmiş ve bu aralıkta uniform dağılıma sahip rastgele değerler üretilerek $\sigma$ olarak kullanılmıştır.

**Zaman uzatma**

Zaman uzatma, ses sinyalinin perdesini etkilemeden hızını veya süresini değiştirerek sinyalin manipüle edildiği bir veri artırma yöntemidir (Nugroho ve Noersasongko, 2022). Teorik olarak bu işlem, modeli konuşma hızından bağımsız hale getirerek genellemeyi geliştirir. Konuşma hızını belirlemek için γ ile gösterilen bir ölçekleme katsayısı kullanılır. γ>1 olduğunda konuşmanın hızı artarken, γ<1 olduğunda ise konuşmanın hızı azalır. Bu çalışmada 0.9 ve 1.1 olmak üzere iki ölçekleme katsayısı kullanılarak zaman uzatma gerçekleştirilmiş ve bu yolla eğitim veri kümesindeki konuşma sayısı üç katına çıkarılmıştır. Ancak veri kümesine sonradan eklenen örneklerin uzunlukları orijinallerinden farklıdır ve bu örneklerin aynı ağa uygulanabilmeleri için eşit uzunluğa getirilmeleri gerekir. Konuşma sinyallerinin eşit uzunluğa getirilmeleri için kırpma ve sıfır dolgusu ekleme gibi yöntemler kullanılır. Uzun konuşmalar kırpılarak, kısa konuşmaların ise başına ve sonuna sıfır dolgusu eklenerek eşit uzunluğa getirilebilir. Bu çalışmada zaman uzatma işlemi için açık kaynak kodlu bir python kütüphanesi olan librosa'nın time_stretch fonksiyonu kullanılmıştır (https://librosa.org/). Eğitim veri kümesindeki en uzun konuşmanın uzunluğu belirlenmiş ve diğer konuşmaların başına ve sonuna sıfır dolgusu eklenerek her birinin uzunluğu referans uzunluğa getirilmiştir.

**Perde kaydırma**

Perde kaydırma, oynatma hızında ve genliğinde bir değişiklik yapmadan konuşmanın perde frekansının birkaç yarım ton yukarı veya aşağı kaydırıldığı bir diğer veri artırma yöntemidir (Nugroho ve Noersasongko, 2022). Perde kaydırma ile veri artırma, konuşma tanıma modellerini farklı perde aralıklarına sahip konuşmacılara karşı daha sağlam hale getirir (Arakawa ve ark., 2019). Ayrıca modelin genelleştirme yeteneğini arttırmaya yardımcı olur ve konuşmacıların perde değişimleri sergilediği senaryolarda modelin tanıma doğruluğunu arttırır. Çalışmada perde kaydırma için libroza kütüphanesinin pitch_shif fonksiyonu kullanılmıştır. {+2, -2} olmak üzere iki farklı faktör kullanılarak perde kaydırma gerçekleştirilmiş ve bu yolla model eğitiminde kullanılan veri kümesinin boyutu üç katına çıkarılmıştır. Çalışmada kullanılan üç veri artırma yönteminin bir erkek konuşmasına uygulanması sonucunda elde edilen sinyallerin zaman ve frekans uzayındaki temsilleri Şekil 1'de gösterilmiştir.




**Şekil 1.** Bir erkek konuşmasına uygulanan veri artırma yöntemlerinin zaman (sol) ve frekans (sağ) uzayındaki temsilleri

## Evrişimli sinir ağları

Verilerin hazırlanması ve özellik çıkarma işlemlerinden sonra bir sınıflandırma modeli oluşturularak konuşma sinyalleri farklı kategorilere göre sınıflandırılır. Literatürde sınıflandırma için geliştirilmiş birçok yöntem mevcuttur. Evrişimli sinir ağları (CNN) bu yöntemlerden birisidir ve bu çalışmada konuşmacıların çocuk, erkek ve kadın olarak sınıflandırılması amacıyla kullanılmıştır. CNN, hayvanların görsel korteksinin organizasyonundan ilham alınarak geliştirilmiş bir tür derin öğrenme yöntemidir (Yamashita ve ark., 2018). CNN'ler özelliklerin mekansal hiyerarşilerinin insan denetimi olmaksızın düşük ve yüksek seviye kalıplardan otomatik olarak öğrenmek için tasarlanmıştır ve bilgisayarlı görme, konuşma işleme, yüz tanıma vb. dahil birçok alanda yaygın olarak kullanılmaktadır (Bhatt ve ark., 2021; Issa ve ark., 2020; Lou ve Shi, 2020). Tipik bir CNN mimarisi birkaç evrişim katmanı ve bir havuzlama katmanından oluşan bir yığının tekrarından ve ardından bir veya daha fazla tam bağlı katmandan oluşur (Yamashita ve ark., 2018). CNN mimarisini oluşturan temel yapıların kısa bir açıklaması aşağıda verilmiştir.

## Evrişim katmanı

Evrişim katmanı CNN mimarisinin temel bileşenlerinden birisidir ve ana işlevi giriş eğitim örneklerinden özellik çıkarmaktır. Her evrişim katmanında, özellik çıkarmaya yardımcı olan, çekirdek adı verilen bir dizi filtre bulunur. Çekirdeğin boyutu giriş sinyalinin boyutuyla aynı olur ve genellikle modelin derinliği arttıkça çekirdek sayısı da artar. Böylece ilk evrişim katmanında basit özellikler yakalanırken, son evrişim katmanında daha karmaşık özellikler yakalanır. Evrişim işlemi, çekirdeğin giriş verileri üzerinde kaydırılmasıyla gerçekleştirilir. Çekirdek ile giriş sinyalinin yerel penceresinin çakışan elemanları eleman bazında çarpılır ve sonuçlar toplanarak ilgili pencerenin özellik haritalaması gerçekleştirilir (Yamashita ve ark., 2018). Bu işlem rastgele sayıda çekirdek için tekrarlanır ve böylece giriş sinyalinin farklı özelliklerini temsil eden özellik haritaları oluşturulur. 2D evrişim işleminin matematiksel temsili aşağıda verilmiştir.

$$z(i,j) = x(i,j) * w(i,j) = \sum_{s=-a}^{a} \sum_{t=-b}^{b} x(s,t)\,.\,w(i-s, j-t) \quad (3)$$

burada $x(i,j)$ giriş sinyalini, $w(i,j)$ ise evrişim çekirdeğini, "$*$" ise evrişim işlemini temsil eder. Evrişim sırasında, her çekirdeğin merkezi, giriş verilerinin en dıştaki elemanıyla örtüşmez ve bu nedenle çıkış özellik haritasının boyutu azalır. Bu sorunu çözmek için sıklıkla sıfır doldurma (zero padding) yöntemi kullanılır. Bu yöntemde girişin tüm sınırları sıfırlarla doldurulur. Böylece çekirdeğin merkez noktası ile girdi verisinin en dıştaki elemanı örtüşür ve boyut azalma önlenir. Evrişim işleminde belirlenmesi gereken bir diğer parametre de adım (stride) değeridir. Adım, ardışık iki çekirdek konumu arasındaki mesafedir ve genellikle 1 olarak seçilir. Ancak bazen özellik haritalarını alt örneklemek için 1'den büyük adım değerleri de kullanılır. Bu çalışmada sıfır dolgusu ve 1 adım değeri kullanılarak evrişim işlemleri gerçekleştirilmiştir.

## Aktivasyon fonksiyonu

Evrişim işlemi, yalnızca matris çarpımı ve toplamından oluştuğu için tamamen doğrusaldır. Ancak gerçek dünya verilerinin çoğu doğrusal değildir ve bu nedenle doğrusal olmayan özelliklerin CNN'e dahil edilmesi gerekir. Bu amaçla evrişim katmanından sonra doğrusal olmayan bir aktivasyon fonksiyonu kullanılır. Sigmoid ve hiperbolik tanjant (tanh) gibi düzgün doğrusal olmayan fonksiyonlar, biyolojik nöron davranışını matematiksel olarak temsil ettikleri için daha öncelerde yaygın olarak kullanılıyordu. Ancak günümüzde düzeltilmiş doğrusal birim (ReLU) aktivasyon fonksiyonu, gradyan tabanlı öğrenmenin yakınsamasını hızlandırdığı ve kaybolan gradyan problemini

önlediği için ana akım aktivasyon fonksiyonu olarak kullanılmaya başlandı. CNN tabanlı sınıflandırıcılarda kullanılan bir diğer aktivasyon fonksiyonu da softmax dır. Softmax, çıktısı giriş sınıflarının kategorik olasılık dağılımına eşdeğer olan bir aktivasyon fonksiyonudur ve genellikle CNN ağının son katmanında kullanılır (Gupta ve ark., 2019).

**Havuzlama katmanı**

Havuzlama katmanı (Pooling Layer), önceki katmandan gelen özellik haritalarını alt örnekleyerek yoğunlaştırılmış çözünürlüğe sahip yeni özellik haritaları üretir. Bu süreçte girdinin uzaysal boyutu büyük ölçüde azaltılırken, kritik özellikler ise korunur. Böylece hem hesaplama maliyeti azalır hem de aşırı uyum sorunuyla mücadele edilir. Evrişim işlemine benzer şekilde, havuzlama işleminde de çekirdek boyutu, adım ve dolgu hiper parametreleri belirlenmelidir. En yaygın havuzlama işlevi, her özellik haritasının yerel komşuluğundaki maksimum değeri döndüren ve diğer değerleri atan maksimum havuzlamadır. Bu çalışmada havuzlama fonksiyonu olarak maksimum havuzlama kullanılmıştır.

**Küresel ortalama havuzlama**

Küresel ortalama havuzlama (Global average pooling), yükseklik × genişlik boyutuna sahip bir özellik haritasının, her özellik haritasındaki tüm öğelerin ortalaması alınarak $1 \times 1$ dizisine alt örneklendiği, derinliğin ise korunduğu alt örneklemenin aşırı bir türüdür. Bu işlem genellikle tam bağlı katmanlardan önce yalnızca bir kez uygulanır. Küresel ortalama havuzlama öğrenilebilir parametrelerin sayısını azaltır ve CNN'nin değişken büyüklükteki girdileri kabul etmesini sağlar.

**Tam bağlı katman**

Son evrişim veya havuzlama katmanının çıktısı olan özellik haritaları tipik olarak düzleştirilir, yani bir sayı dizisine dönüştürülür ve daha sonra bir veya daha fazla tam bağlı katmana bağlanır. Tam bağlı katman, her girdinin her çıktıya öğrenilebilir bir ağırlıkla bağlandığı bir sistemdir. Bu sistemin çıktısı ağın nihai çıktısıdır ve sınıflandırma görevleri için sınıf olasılıklarıdır. Tam bağlı her katmandan sonra ReLU gibi doğrusal olmayan bir aktivasyon kullanılır. Genellikle tama bağlı son katman diğerlerinden farklıdır. Sınıf sayısına eşit sayıda çıkış düğümüne ve göreve uygun bir aktivasyon fonksiyonuna sahiptir. Çok sınıflı sınıflandırma için bu fonksiyon Softmax'tır. Softmax, çıkış değerlerini toplamı 1 olacak şekilde 0 ile 1 arasına normalleştirerek sınıf olasılıklarına dönüştürür.

**Batch normalizasyonu**

Eğitim sırasında, her katmanın girdisinin dağılımı önceki katmanın parametreleri değiştikçe değişir ve bu da eğitim sürecinin yavaşlamasına neden olur. Batch Normalizasyonu (BN) bu sorunu azaltmak için Ioffe ve Szegedy tarafından önerilmiş bir yöntemdir (Ioffe ve Szegedy, 2015). Denklem (4) ile verilen BN, önceki katmanın çıktısı $O$'dan parti ortalaması μ 'yü çıkarıp parti standart sapması $\sigma$'ya bölerek $O$ 'nun normalizasyonu gerçekleştirilir.

$$\hat{O} = \frac{o-\mu}{\sigma} \tag{4}$$

Batch normalizasyon gradyanlar patlaması sorunu olmaksızın ağın yüksek bir öğrenme oranı ile kullanılmasına imkan sunar. Ayrıca ağın genelleştirme özelliğini iyileştirerek overfitting'i azaltır ve farklı başlangıç şemalarına ve öğrenme oranlarına karşı ağın daha güçlü olmasını sağlar.

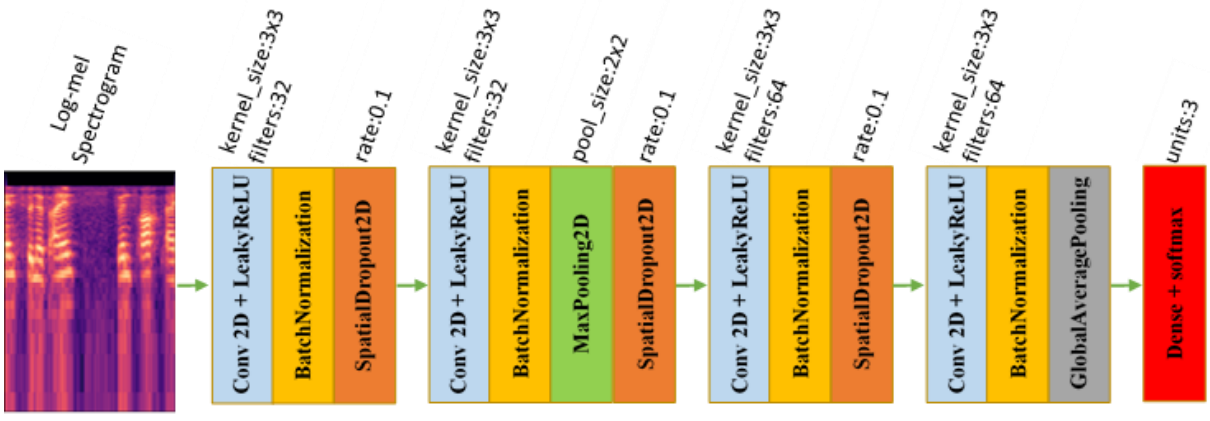
**Dropout katmanı**

Dropout, giriş verilerinin genel özelliklerinin öğrenilmesi yerine ezberlenmesi olarak ifade edilebilecek aşırı uyumun önlenmesi için geliştirilmiş bir düzenleme tekniğidir (Srivastava ve ark.,

2014). Sinir ağlarında düzenleme önemli bir görevdir ve bu amaçla geliştirilmiş çeşitli teknikler vardır (Nusrat ve Jang, 2018). Bu tekniklerin çoğu ekstra hesaplama maliyeti gerektirir ve bu da ağın daha yavaş çalışmasına neden olur. Dropout, hesaplama maliyetini arttırmayan, aksine azaltan bir tekniktir. Bu tekniğe göre, her eğitim yinelemesinde rastgele seçilen bir nöron seti bırakılarak diğer nöronlarla bağlantısı kesilir. Bağlantısı kesilen nöronlar modelin eğitim sürecini etkilemez ancak test sürecinde tam ölçekli ağ kullanılır. Dropout sayesinde özellik seçme gücü tüm nöron gruplarına eşit olarak dağıtılır ve model farklı bağımsız özellikleri öğrenmeye zorlanır.

**Önerilen CNN mimarisi**

Bu çalışmada konuşmacıların erkek, kadın ve çocuk olarak sınıflandırılması amacıyla grafiksel temsili Şekil 2'de verilen CNN modeli kullanılmıştır. Bu model dört yerel öznitelik öğrenme bloğundan (YÖÖB) oluşur ve her blok bir evrişim katmanı ile onu takip eden bir aktivasyon (LeakyReLU) ve bir BatchNormalization katmanlarını içerir. Tüm YÖÖB'lerin ilk üç katmanı aynı olup bu katmanları birinci ve üçüncü YÖÖB'de Dropout, ikinci YÖÖB'de Maxpooling ve Dropout, dördüncü YÖÖB de ise GlobalAveragePooling katmanları takip eder. İlk iki evrişim katmanında 3x3 boyutunda 32 filtre, üçüncü ve dördüncü evrişim katmanlarında ise aynı boyutlu 64 filtre kullanılmıştır. Tüm aktivasyon katmanlarının alpha parametresi 0.1, Dropout katmanlarının bırakma oranı 0.2 ve Maxpooling katmanının havuzlama boyutu 2x2 olarak belirlenmiştir. Öznitelik öğrenme bloklarında işlenen öznitelikler GlobalAveragePooling ile bir boyuta indirgendikten sonra softmax aktivasyonlu tam bağlı katmana giriş olarak uygulanmış ve ağın çıkışından giriş verilerinin belirli bir sınıfa ait olma olasılıkları alınmıştır. Süreç sonunda en yüksek olasılığa sahip sınıf, giriş verisinin tahmin edilen sınıfı olarak seçilmiş ve böylece sınıflandırma görevi tamamlanmıştır.




**Şekil 2.** Önerilen CNN modelinin grafiksel temsili

Önerilen CNN modeli Keras, scikit-learn, TensorFlow gibi derin öğrenme kütüphaneleri kullanılarak python dilinde gerçekleştirilmiştir. Model eğitimi Çizelge 2'de verilen hiper parametreler ile yapılmış ve her epoch sonunda hesaplanan doğrulama kayıp değerlerine göre en iyi model seçilmiştir.

**Çizelge 2.** Önerilen CNN modelinde kullanılan hiper parametreler

| Hiper parametreler | Değerler |
|---|---|
| Epoch | 360 |
| Batch_size | 128 |
| Loss_Function | categorical_crossentropy |
| Optimizer | Adam |
| Learning Rate | 0.0001 |

**Değerlendirme metrikleri**

Çalışmada geliştirilen CNN modelinin performansını değerlendirmek için doğruluk, F1 puanı, kesinlik ve duyarlılık olmak üzere dört farklı değerlendirme ölçütü kullanılmıştır. Bu ölçütler, konuşma ve diğer tanıma sistemlerinin performansını değerlendirmek için yaygın olarak kullanılmaktadır. Bu metriklerin hesaplanmasında modelin tahmini ile gerçek sınıflar arasındaki ilişkiyi temsil eden dört kavram kullanılır; Doğru Pozitifler (TP) , Doğru Negatifler (TN), Yanlış Pozitifler (FP) ve Yanlış Negatifler (FN). TP, pozitif olan ve doğru şekilde pozitif olarak tahmin edilen örnekleri, FN ise pozitif olan ancak yanlış bir şekilde negatif olarak tahmin edilen örnekleri temsil eder. Benzer şekilde FP, negatif olan ancak yanlış bir şekilde pozitif olarak tahmin edilen örnekleri, TN ise negatif olan ve doğru şekilde negatif olarak tahmin edilen örnekleri temsil eder. Çalışmada kullanılan değerlendirme ölçütlerinin TP, TN, FP ve FN terimlerine bağlı olarak temsilleri aşağıda verilmiştir.

$$Doğruluk(Accuracy) = \frac{TP+TN}{TP+FP+TN+FN} \quad (5)$$

$$Kesinlik(Precision) = \frac{TP}{TP+FP} \quad (6)$$

$$Duyarlılık(Recall) = \frac{TP}{TP+FN} \quad (7)$$

$$F1\ puanı = \frac{2*Duyarlılık*Kesinlik}{Duyarlılık+Kesinlik} \quad (8)$$

## BULGULAR VE TARTIŞMA

Bu bölümde çeşitli veri artırma yöntemlerinin yaş ve cinsiyet sınıflandırma performansı üzerindeki etkilerinin belirlenmesi amacıyla yapılan test sonuçları verilmiştir. Bu testlerde aGender veri kümesinden detayları bölüm 2.1 de belirtilen kriterlere göre seçilen konuşma verileri kullanılmıştır. Öncelikle veri kümesi 75/25 oranına göre iki bölüme ayrılmış ve %75 lik bölüm model eğitiminde, kalan %25 lik bölüm ise geliştirilen modellerin test aşamasında kullanılmıştır. Eğitim verilerinin bir bölümü (1/12) ise doğrulama için ayrılarak en iyi modelin belirlenmesinde bu veriler kullanılmıştır. Son aşamada ise belirlenen modelin test veri kümesi üzerindeki performans değerlendirmesi yapılarak test süreci tamamlanmıştır.

Çalışmada gürültü ekleme (GE), zaman uzatma (ZU) ve perde kaydırma (PK) olmak üzere üç farklı veri artırma yöntemi kullanılmıştır. Öncelikle, önerilen CNN modeli hiçbir veri artırma yöntemi kullanılmadan yalnızca orijinal veriler ile gerçekleştirilmiştir. Daha sonra her bir veri artırma yöntemi önce ayrı ayrı daha sonra ise birlikte uygulanarak ilgili yöntemlerin performans üzerindeki etkileri araştırılmıştır. Veri artırma yöntemleri yalnızca eğitim veri kümesine uygulanmış, test aşamasında kullanılan veriler üzerinde herhangi bir işlem yapılmamıştır. Çalışmada üç veri artırma yönteminin farklı sırada kullanımı ile yedi farklı eğitim veri kümesi oluşturulmuş, bu veri kümeleri ile de yedi farklı CNN modeli geliştirilmiştir. Bu modellerin birinde hiçbir veri artırma yöntemi kullanılmazken, üçünde birer veri artırma yöntemi, ikisinde iki veri artırma yöntemi ve birinde de üç veri artırma yöntemi birlikte kullanılmıştır. Kullanılan veri artırma yöntemine bağlı olarak model eğitiminde kullanılan veri kümesinin boyutu da değişmiştir. Örneğin gürültü ekleme yöntemi ile veri kümesinin boyutu iki katına, perde kaydırma yöntemi ile üç katına çıkmıştır. Diğer taraftan test verileri üzerinde herhangi bir işlem yapılmadığından tüm modellerin test aşamasında kullanılan veri kümesinin boyutu

değişmemiştir. Her bir modelin test veri kümesi üzerinde elde edilen sonuçları Çizelge 3'de verilmiş olup bu sonuçlar ilgili modelin beş kez çalıştırılması ile elde edilen sonuçların ortalaması alınarak hesaplanmıştır.

**Çizelge 3.** Farklı veri artırma yöntemleri ile geliştirilen CNN modellerinin test veri kümesi üzerindeki performansları

| Model No | Veri artırma | Eğitim kümesinin boyutu | Doğruluk | Kesinlik | Duyarlılık | F1 puanı |
|---|---|---|---|---|---|---|
| 1 | Yok | (4320, 40, 108) | 84.583 | 84.576 | 84.846 | 84.526 |
| 2 | GE | (8640, 40, 108) | 85.569 | 85.491 | 85.699 | 85.532 |
| 3 | PK | (12960, 40, 108) | 85.458 | 85.495 | 85.702 | 85.431 |
| 4 | ZU | (12960, 40, 120) | 86.528 | 86.466 | 86.571 | 86.494 |
| 5 | GE+PK | (17280, 40, 108) | 86.719 | 86.747 | 86.912 | 86.686 |
| 6 | GE+ZU | (17280, 40, 120) | 86.875 | 86.824 | 86.946 | 86.850 |
| 7 | GE+PK+ZU | (25920, 40, 120) | 87.523 | 87.532 | 87.712 | 87.504 |

GE: Gürültü ekleme; PK: Perde kaydırma; ZU: Zaman uzatma

Çizelge 3'de verilen sonuçlar incelendiğinde üç veri artırma yönteminin birlikte kullanımı ile geliştirilen CNN modelinin (Model No: 7) en yüksek doğruluk oranına sahip model olduğu görülmektedir. Altı kat genişletilmiş eğitim veri kümesi ile geliştirilen bu model sınıflandırma doğruluğunu %84.583'den %87.523'e çıkararak yaklaşık %3 performans artışı sağlamıştır. Yalnızca bir veri artırma yöntemi kullanılarak geliştirilen modeller arasında en yüksek doğruluk oranına zaman uzatma (ZU) yöntemi ile ulaşılmıştır. Zaman uzatma yöntemi ile doğruluk oranı %84.583'den %86.528'ye çıkarılarak, yaklaşık %2 oranında performans artışı sağlanmıştır. İki veri artırma yönteminin birlikte kullanıldığı modeller ise (GE+PK, GE+ZU) doğruluk oranını %84.583'den sırasıyla %86.719 ve %86.875 seviyesine çıkararak %2'nin üzerinde performans artışı sağlamıştır.

Hiçbir veri artırma yönteminin kullanılmadığı model (Model No: 1) ile üç veri artırma yönteminin kullanıldığı modele (Model No: 7) ait karışıklık matrisleri Çizelge 4'de verilmiştir. Karışıklık matrisi bir sınıflandırıcı tarafından sağlanan doğru ve yanlış tahminlerin sınıflara göre dağılımını gösteren bir tablodur. Bu tablonun köşegeni üzerindeki hücreler doğru tahmin edilen örneklerin sayısını, diğer hücreler ise yanlış tahmin edilen örneklerin sayısını gösterir. Karışıklık matrisleri her hücredeki örnek sayısı ilgili satırdaki toplam örnek sayısına bölünerek yüzde oranı ile de temsil edilebilir. Aşağıdaki karışık matrisinde örnek sayısı yerine doğru ve yanlış tahminlerin yüzde oranı ile temsil edildiği gösterim kullanılmıştır.

**Çizelge 4.** Veri artırma kullanılmadan geliştirilen 1 numaralı CNN modeli ile (a), veri artırma kullanılarak (b) geliştirilen 7 numaralı CNN modelinin karışıklık matrisleri

| | Çocuk | Kadın | Erkek |
|---|---|---|---|
| Çocuk | **76.31** | 18.45 | 5.24 |
| Kadın | 11.87 | **83.08** | 5.05 |
| Erkek | 1.06 | 2.34 | **96.60** |

(a)

| | Çocuk | Kadın | Erkek |
|---|---|---|---|
| Çocuk | **79.81** | 15.92 | 4.27 |
| Kadın | 7.91 | **88.13** | 3.96 |
| Erkek | 0.85 | 2.55 | **96.60** |

(b)

Çizelge 4'de verilen karışıklık matrisinden 7 numaralı modelin özellikle çocuk ve kadın konuşmacıları sınıflandırmada 1 numaralı modelden daha başarılı olduğu, erkekleri sınıflandırmada ise her iki modelin başarısının eşit olduğu görülmektedir. 1 numaralı model, test veri kümesindeki çocuk konuşmacılara ait 515 konuşmanın %76.31'ini doğru sınıflandırırken, %18.45'ini kadın ve %5.24'ünü erkek olarak yanlış sınıflandırmıştır. 7 numaralı model ise çocuk konuşmalarının %79.81'ini doğru, %15.92'sini kadın ve %4.27'sini erkek olarak yanlış sınıflandırmıştır. İki model de kadın

konuşmalarını sınıflandırmada çocuk konuşmalarına kıyasla daha yüksek başarı göstermiştir. 1 numaralı model test kümesindeki 455 kadın konuşmasının %83,03'ünü doğru, %11.87'sini çocuk ve %5.05 'ini erkek olarak yanlış sınıflandırmıştır. 7 numaralı model ise kadın konuşmalarının %88.13'ünü doğru, %7.91'ini çocuk ve %3.96'sını erkek olarak yanlış sınıflandırmıştır. Erkek konuşmaları ise her iki modelin de en yüksek doğrulukla sınıflandırdığı grup olmuştur. Her iki model de test kümesindeki 470 erkek konuşmasının %96.60'ını doğru sınıflandırmıştır. 1 numaralı model erkek konuşmalarının %1.06'sını çocuk, %2.34'ünü kadın olarak yanlış sınıflandırırken, 7 numaralı model ise %0.85'ini çocuk, %2.55'ini kadın olarak yanlış sınıflandırmıştır.

Çalışmada elde edilen sonuçlar ile benzer çalışmaların sonuçları Çizelge 5'te karşılaştırılmıştır. Çalışmaları karşılaştırırken yalnızca doğruluk üzerinden değerlendirme yapılmamalı, kullanılan veri kümesi, kayıt ortamı ve konuşmacıların dağılımı gibi farklılıklar göz önünde bulundurulmalıdır. Bu bağlamda bu çalışma ile aynı veri kümesinin kullanıldığı çalışmaların (Kockmann ve ark., 2010; Levitan ve ark., 2016; Yücesoy ve Nabiyev, 2016) sonuçlarının karşılaştırılması daha kolaydır. Bu çalışmada geliştirilen model, %87.523 doğrulukla Kockmann ve ark. (2010) ile Levitan ve ark. (2016)'nın önerdiği modelden daha üstün performans sağlamıştır. Diğer çalışmanın (Yücesoy ve Nabiyev, 2016) doğruluk oranı %90.39 olarak belirtilmesine rağmen bu sonucun sınıf dağılımı eşit olmayan bir test kümesi üzerinden elde edildiği ve dengeli bir veri kümesi kullanılması durumunda ilgili modelin doğruluğunun kaba bir hesaplama ile %83.84 olacağı modelin karışıklık matrisinden görülebilir. Bu durum göz önünde bulundurulduğunda bu çalışmada önerilen modelin performansının ilgili çalışmadaki modelin performansından daha iyi olduğu anlaşılmaktadır.

**Çizelge 5.** Geliştirilen modelin doğruluk açısından diğer ilgili çalışmalarla karşılaştırılması

| **Çalışma** | **Veri kümesi** | **Doğruluk (%)** |
|---|---|---|
| Kockmann ve ark., 2010 | aGender | 81.82 |
| Levitan ve ark., 2016 | aGender | 85 |
| Yücesoy ve Nabiyev, 2016 | aGender | 90.39 |
| Vlaj ve Zgank, 2022 | TIDIGITS | 92.25 |
| ***Bu çalışma*** | ***aGender*** | ***87.52*** |

Vlaj ve Zgank (2022) tarafından yapılan çalışmada ise geliştirilen modelin doğruluk oranının %92.25 olduğu belirtilmiştir. Bu oran bu çalışmada elde edilen doğruluk oranından daha yüksek olmakla birlikte, ilgili çalışmada gürültüsüz ortamda kaydedilen konuşmaların bu çalışmada ise telefon hattı üzerinden kaydedilen konuşmaların kullanıldığı göz önünde bulundurulduğunda iki model arasındaki farkın makul olduğu değerlendirilmektedir.

## SONUÇ

Bu çalışmada, gürültü ekleme, perde kaydırma ve zaman uzatma olmak üzere üç farklı veri artırma yönteminin, yaş ve cinsiyet sınıflandırma başarısı üzerindeki etkileri araştırılmıştır. Bu amaçla dört yerel öznitelik öğrenme bloğundan (YÖÖB) oluşan bir CNN modelinin kullanımı önerilmiştir. Konuşma sinyallerinden çıkarılan mel spectrogram öznitelikleri, bu modele giriş olarak uygulanmış ve modelin yaş ve cinsiyet tahminleri çıkış olarak alınmıştır. Çalışmada aGender veri kümesinden rastgele seçilen konuşma verileri kullanılmıştır. Öncelikle herhangi bir veri artırma yöntemi kullanılmadan yalnızca orijinal veri kümesindeki konuşmalar ile model gerçekleştirilmiş ve modelin performans değerlendirmesi yapılmıştır. Daha sonra her bir veri artırma yöntemi önce ayrı ayrı, sonra birlikte kullanılarak eğitim kümesinin boyutu arttırılmış ve bu veriler ile eğitilen modellerin performansları karşılaştırılmıştır. Veri artırma yöntemleri yalnızca eğitim veri kümesine uygulanmış, test aşamasında kullanılan veriler üzerinde herhangi bir işlem yapılmamıştır. Çalışmada üç veri artırma

yöntemi farklı kombinasyonlarda kullanılarak yedi farklı model geliştirilmiştir. Bu modeller arasında üç veri artırma yönteminin birlikte kullanıldığı model (Model No:7) en başarılı model olmuş ve sınıflandırma doğruluğunu %84.583'den %87.523'ye çıkarmıştır. Veri artırma yöntemlerinin kullanıldığı diğer modellerin performanslarında da %1 ile %2.3 arasında artış sağlanmıştır. Bu sonuçlar yaş ve cinsiyet sınıflandırma performansının iyileştirilmesinde veri artırma yöntemlerinin kullanılmasının etkinliğini göstermektedir. Gelecek çalışmalarda zaman maskeleme, frekans maskeleme ve dinamik karıştırma gibi farklı veri artırma yöntemlerinin kullanımının sınıflandırma performansı üzerindeki etkilerinin araştırılması önerilmektedir.

## Çıkar Çatışması

Makale yazarı çıkar çatışması olmadığını beyan eder.

## KAYNAKLAR

Arakawa, R., Takamichi, S., & Saruwatari, H. (2019). Implementation of DNN-based real-time voice conversion and its improvements by audio data augmentation and mask-shaped device. *In: Proc. ISCA Workshop Speech Synthesis*, (pp. 93–98). Vienna, Austria.

Bhatt, D., Patel, C., Talsania, H., Patel, J., Vaghela, R., Pandya, S., ... & Ghayvat, H. (2021). CNN variants for computer vision: History, architecture, application, challenges and future scope. *Electronics*, 10(20), 2470.

Bishop, C. M. (1995). Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, 7(1), 108–116.

Chai, J., Zeng, H., Li, A., & Ngai, E. W. (2021). Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications*, 6, 100134.

Gerosa, M., Giuliani, D., & Brugnara, F. (2005). Speaker adaptive acoustic modeling with mixture of adult and children's speech. *In Interspeech*, (pp. 2193-2196). Lisbon, Portugal.

Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-End Factor Analysis for Speaker Verification. *IEEE Trans. Audio Speech Lang*, 19, 788–798.

Ertam, F. (2019)An effective gender recognition approach using voice data via deeper LSTM networks. *Appl. Acoust*., 156, 351–358.

Gupta, A., Harrison, P. J., Wieslander, H., Pielawski, N., Kartasalo, K., Partel, G., ... & Wählby, C. (2019). Deep learning in image cytometry: a review. *Cytometry Part A*, 95(4), 366-380.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *In International Conference on Machine Learning*, (pp 448-456). Lille France.

Issa, D., Demirci, M. F., & Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*, 59, 101894.

Jasuja, L., Rasool, A., Hajela, G. (2020) Voice Gender Recognizer Recognition of Gender from Voice using Deep Neural Networks. *In Proceedings of the 2020 International Conference on Smart Electronics and Communication (ICOSEC)*, (pp. 319–324). Trichy, India.

Kwasny, D., & Hemmerling, D. (2021). Gender and age estimation methods based on speech using deep neural networks. *Sensors*, 21(14), 4785.

Kockmann, M., Burget, L., & Cernocký, J. (2010). Brno university of technology system for interspeech 2010 paralinguistic challenge. *In Interspeech*, (pp. 2822-2825). Makuhari, Chiba, Japan.

Levitan, S. I., Mishra, T., & Bangalore, S. (2016). Automatic identification of gender from speech. *In Proceeding of speech prosody*, (pp. 84-88). Boston, USA.

Li, M., Han, K. J., & Narayanan, S. (2013). Automatic speaker age and gender recognition using acoustic and prosodic level information fusion. *Computer Speech & Language*, 27(1), 151-167.

Lingenfelser, F., Wagner, J., Vogt, T., Kim, J., & André, E. (2010). Age and gender classification from speech using decision level fusion and ensemble based techniques. *In Eleventh Annual Conference of the International Speech Communication Association*, (pp. 2798-2801). Makuhari, Chiba, Japan.

Liu, X., Wang, H., Zhang, Y., Wu, F., & Hu, S. (2022). Towards efficient data-centric robust machine learning with noise-based augmentation, *arXiv preprint arXiv*:2203.03810.

Lou, G., & Shi, H. (2020). Face image recognition based on convolutional neural network. *China communications*, 17(2), 117-124.

Mahmoodi, D., Marvi, H., Taghizadeh, M., Soleimani, A., Razzazi, F. & Mahmoodi, M. (2011, July). Age Estimation Based on Speech Features and Support Vector Machine. *In Proceedings of the 2011 3rd Computer Science and Electronic Engineering Conference (CEEC)*, (pp. 60–64). Colchester, UK.

Mavaddati, S. (2024). Voice-based Age, Gender, and Language Recognition Based on ResNet Deep model and Transfer learning in Spectro-Temporal Domain. *Neurocomputing*, (580), 127429.

Miliaresi, I., Poutos, K., & Pikrakis, A. (2021). Combining acoustic features and medical data in deep learning networks for voice pathology classification. *In 2020 28th European Signal Processing Conference (EUSIPCO)*, (pp. 1190-1194). Amsterdam, Netherlands.

Nanthini, K., Sivabalaselvamani, D., Chitra, K., Gokul, P., KavinKumar, S., & Kishore, S. (2023). A Survey on Data Augmentation Techniques. *In 2023 7th International Conference on Computing Methodologies and Communication (ICCMC),* (pp. 913-920). Erode, India.

Nugroho, K., & Noersasongko, E. (2022). Enhanced Indonesian ethnic speaker recognition using data augmentation deep neural network. *Journal of King Saud University-Computer and Information Sciences*, 34(7), 4375-4384.

Nusrat, I., & Jang, S.B. (2018). A comparison of regularization techniques in deep neural networks. *Symmetry*, 10(11):648.

Potamianos, A., & Narayanan, S. (2003). Robust recognition of children's speech. *IEEE Transactions on speech and audio processing*, 11(6), 603-616.

Qureshi, M. F., Mushtaq, Z., ur Rehman, M. Z., & Kamavuako, E.N. (2022) Spectral image-based multiday surface electromyography classification of hand motions using CNN for human–computer interaction. *IEEE Sens. J.*, 22, 20676–20683.

Sánchez-Hevia, H. A., Gil-Pita, R., Utrilla-Manso, M., & Rosa-Zurera, M. (2022). Age group classification and gender recognition from speech with temporal convolutional neural networks. *Multimedia Tools and Applications*, 81(3), 3535-3552.

Sarker, I. H. (2021). Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN Computer Science*, 2(6), 420.

Srivastava, N., Hinton, G., Krizhevsky, A,, Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929-1958.

Tursunov, A., Mustaqeem, Choeh, J. Y., & Kwon, S. (2021). Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms. *Sensors*, 21(17), 5892.

Uddin, M. A., Hossain, M. S., Pathan, R. K., & Biswas, M. (2020). Gender Recognition from Human Voice using Multi-Layer Architecture. *In Proceedings of the 2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, (pp. 1–7).Novi Sad, Serbia.

Vlaj, D., & Zgank, A. (2022). Acoustic Gender and Age Classification as an Aid to Human–Computer Interaction in a Smart Home Environment. *Mathematics*, 11(1), 169.

Wei, S., Sun, Z., Wang, Z., Liao, F., Li, Z., & Mi, H. (2023). An efficient data augmentation method for automatic modulation recognition from low-data imbalanced-class regime. *Applied Sciences*, 13(5), 3177.

Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, 9, 611-629.

Yücesoy, E., & Nabiyev, V. V. (2016). A new approach with score-level fusion for the classification of a speaker age and gender. *Computers & Electrical Engineering*, 53, 29-39.

Zhang X., Chen A., Zhou G., Zhang Z., Huang X.,& Qiang X. (2019). Spectrogram-frame linear network and continuous frame sequence for bird sound classification. *Ecol. Inform.*, 54, 101009.