T.C.
SÜLEYMAN DEMİREL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

KONUŞMA SİNYALİ VE SES TELLERİ GÖRÜNTÜLERİNDEN DERİN ÖĞRENME TABANLI GLOTAL ALAN KESTİRİMİ

Yaşar Said DERDİMAN

Danışman
Dr. Öğr. Üyesi Turgay KOÇ

YÜKSEK LİSANS TEZİ
ELEKTRİK – ELEKTRONİK MÜHENDİSLİĞİ ANABİLİM DALI
ISPARTA - 2023

İÇİNDEKİLER

Sayfa

ÖZET

Yüksek Lisans Tezi

KONUŞMA SİNYALİ VE SES TELLERİ GÖRÜNTÜLERİNDEN DERİN ÖĞRENME TABANLI GLOTAL ALAN KESTİRİMİ

Yaşar Said DERDİMAN

Süleyman Demirel Üniversitesi
Fen Bilimleri Enstitüsü
Elektrik – Elektronik Mühendisliği Anabilim Dalı

Danışman: Dr. Öğr. Üyesi Turgay KOÇ

Bu tez çalışmasında glottis tespiti yapılması için U-Net tabanlı bir model önerilmiştir. Önerilen model klasik modeller ile karşılaştırılarak modelin performansının geleneksel yöntemlerin performansı ile karşılaştırması yapılmıştır. Modellerin karşılaştırmasında glottal alan büyüklüğünün performansa etkisine de bakılmıştır. U-Net, histogram, bölge büyütme ve aktif kontur olarak üç farklı klasik model ile test verisi üzerinde test edilmiştir. Glottal alanın Çok küçük açıklıklarının da bulunduğu glottal alanın sıfırdan büyük olduğu durumda hassasiyet ölçütü yönünden en yüksek başarım 0.867 ile U-Net modelinde elde edilirken, Aktif Kontur modeli kötü bir sonuç elde ederek 0.389'da kalmıştır. Geri çağırma ölçütü açısından en yüksek değeri 0.964 ile histogram elde etmişken en küçük değeri 0.684 ile bölge büyütme elde etmiştir. Doğruluk ölçütü performans yönüyle ele alındığında U-Net 0.997 ile en yüksek başarıma sahiptir. En düşük başarım ise 0.717 ile aktif kontura aittir. Glottal alanın 100 ve 200'den büyük olduğu durumlar için yapılan karşılaştırmalarda da U-Net modeli başarımını sürdürmüştür. Diğer modellerin de başarımının arttığı gözlenmiştir. Özellikle aktif kontur modelinin başarımında hassasiyet ölçütü açısından %59.5 oranında artış gözlenmiştir. Aynı zamanda görüntü işlemede kullanımı yaygın olan dice skoru ölçütü ile modellerin başarımları incelenmiştir. Bu ölçütün doğru yorumlaması için kutu diyagramları üzerinden yorumlanması daha uygundur. Önceki ölçütlerde olduğu gibi Glottal alan büyüklüğüne bağlı olarak modellere üç farklı veri seti üzerinde tahmin yaptırılmıştır. Kutu diyagramları üzerinde dice skorlarının medyan değerleri klasik modeller için histogram 0.63-0.70-0.72, aktif kontur 0.60-0.69-0.75 ve bölge büyütme 0.68-0.72-0.74 şeklinde iken U-Net modeli ise 0.69-0.80-0.83 sonuçlarını elde etmiştir. Buna göre tüm modeller de glottal alan büyüklüğüne göre tahmin sonuçlarında farklılık olduğu söylenebilir. Glottal alanın her durumunda en iyi başarımı gösteren yine modern derin öğrenme yöntemi olan U-Net modeli olmuştur.

Ayrıca konuşma verisinden glottal alan tahmini yapabilen 3 farklı derin öğrenme modeli geliştirilmiştir. Modeller evrişim katmanları içeren oto kodlayıcılardır. Otokodlayıcı (AE), Gürültü Yok Eden Evrişimli (DnCNN) ve Evrişimli (CNN) olmak üzere 3 farklı model ile eğitim, doğrulama ve test işlemleri yapılmıştır.

Eğitim aşamasında evrişim katmanlarının farklı etkilerini gözlemlemek için çekirdek büyüklüğü, filtre büyüklüğü ve katman sayısı yönünden farklı parametre değerleri için sırasıyla 180, 100 ve 100 adet farklı eğitim yapılmıştır. Eğitilen modellerden doğrulama seti üzerinde en iyi başarımı gösteren modeller seçilerek test verisi üzerindeki performansları karşılaştırılmıştır. Modellerin performansları incelenirken ortalama karesel hata performans ölçütü olarak kullanılmıştır. Test setleri üzerindeki ortalama karesel hata başarımları AE, DnCNN ve CNN modeller için sırasıyla 0.000196, 0.0019063, 0.002085 şeklinde olmuştur.

Anahtar Kelimeler: U-Net, Görüntü bölütleme, Derin öğrenme, Glottis.

2023, 53 sayfa

ABSTRACT

M.Sc. Thesis

DEEP LEARNING BASED ESTIMATION OF GLOTTAL AREA FROM SPEECH AND VOCAL FOLDS IMAGES

Yaşar Said DERDİMAN

Süleyman Demirel University
Graduate School of Natural and Applied Sciences
Department of Electrical and Electronics Engineering

Supervisor: Asst. Prof. Dr. Turgay KOÇ

In this thesis, a U-Net based model is proposed for glottis detection. The proposed model was compared with the classical models and the performance of the model was compared with the performance of the traditional methods. In the comparison of the models, the effect of the glottal area size on the performance was also examined. U-Net has been tested on test data with three different classical models as histogram, region enlargement and active contour. In the case where the glottal area, including very small openings, is greater than zero, the highest performance in terms of sensitivity was obtained in the U-Net model with 0.867, while the Active Contour model achieved a poor result and remained at 0.389. In terms of recall criteria, the histogram obtained the highest value with 0.964, while the smallest value obtained 0.684 region enlargement. When the accuracy criterion is considered in terms of performance, U-Net has the highest performance with 0.997. The lowest performance belongs to the active contour with 0.717. The U-Net model continued to perform well in the comparisons made for cases where the glottal area is greater than 100 and 200. It was observed that the performance of other models increased as well. Especially in the performance of the active contour model, an increase of 59.5% was observed in terms of precision. At the same time, the performance of the models was examined with the dice score criterion, which is widely used in image processing. For the correct interpretation of this criterion, it is more appropriate to interpret it through box diagrams. As in the previous criteria, the models were predicted on three different data sets depending on the Glottal area size. The median values of the dice scores on the box diagrams were 0.63-0.70-0.72 in the histogram, 0.60-0.69-0.75 in the active contour and 0.68-0.72-0.74 in the region enlargement for the classical models, while the U-Net model obtained 0.69-0.80-0.83. Accordingly, it can be said that there is a difference in the estimation results according to the size of the glottal area in all models. The U-Net model, which is also a modern deep learning method, has shown the best performance in all cases of glottal area.

In addition, 3 different deep learning models have been developed that can make glottal area estimation from speech data. Models are autoencoders with convolution layers. Training, verification and testing processes were carried out

with 3 different models: Autoencoder(AE), Noise Canceling Convolutional(DnCNN) and Convolutional(CNN). In order to observe the different effects of convolution layers during the training phase, 180, 100 and 100 different trainings were conducted for different parameter values in terms of kernel size, filter size and number of layers, respectively. The models that show the best performance on the validation set from the trained models were selected and their performances on the test data were compared. While examining the performances of the models, the mean square error was used as a performance measure. The mean square error performances on the test sets were keras, 0.000196, 0.0019063, 0.002085 for noiseless and noisy models, respectively.

Keywords: U-Net, Image segmentation, Deep learning, Glottis.

2023, 53 pages

TEŞEKKÜR

Bu araştırma için beni yönlendiren, karşılaştığım zorlukları bilgi ve tecrübesi ile aşmamda yardımcı olan değerli Danışman Hocam Yrd. Doç. Dr. Turgay KOÇ'a teşekkürlerimi sunarım.

Bu araştırmada yer alan tüm/kısmi nümerik hesaplamalar TÜBİTAK ULAKBİM, Yüksek Başarım ve Grid Hesaplama Merkezi'nde (TRUBA kaynaklarında) gerçekleştirilmiştir. Çalışmalarımız sırasında TÜBİTAK ULAKBİM'e TRUBA kaynaklarını paylaştığı için teşekkür ederim.

Tezimin her aşamasında beni yalnız bırakmayan aileme sonsuz sevgi ve saygılarımı sunarım.

Yaşar Said DERDİMAN
ISPARTA, 2023

ŞEKİLLER DİZİNİ

Sayfa

ÇİZELGELER DİZİNİ

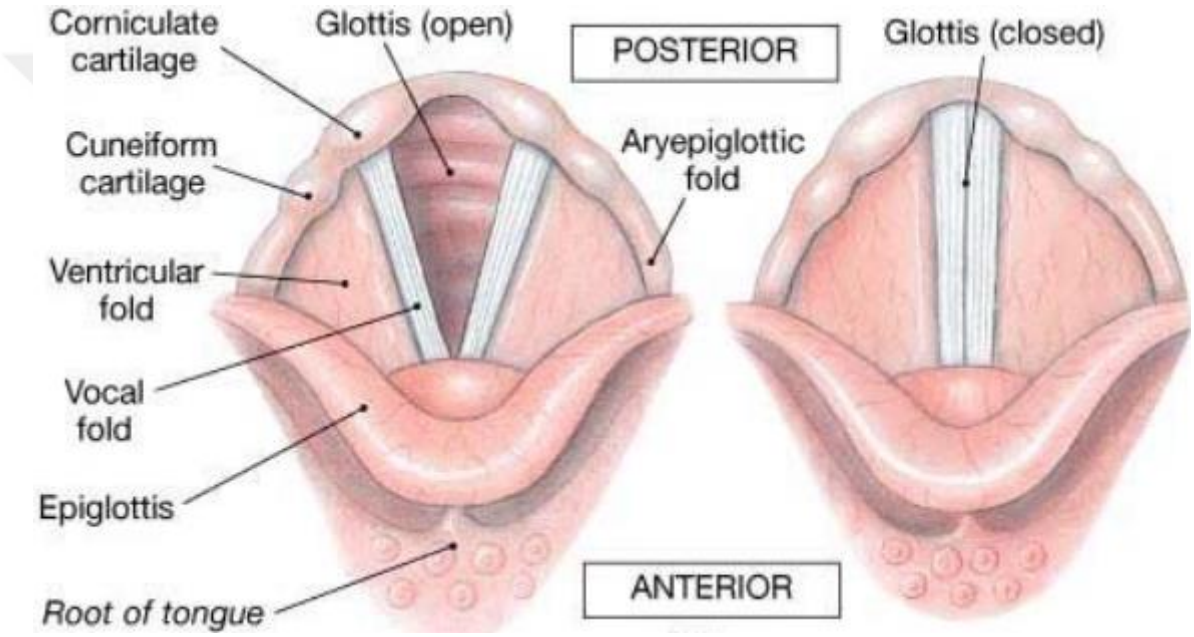
Sayfa

SİMGELER VE KISALTMALAR DİZİNİ

ASPP	Atrous spatial pyramid pooling havuzlama
AE	Otokodlayıcı
CNN	Evrişimli sinir ağı
D	Doğruluk performans ölçütü
DCS	Dice skoru
DnCNN	Gürültü Yok Eden Evrişimli Sinir Ağı
FN	Yanlış negatif değerleri
FP	Yanlış pozitif değerleri
FPS	Saniye/kare
GA	Glottal alan
HA	Hassasiyet
HSV	Yüksek hızlı video
IOU	Intersection over union (Jaccard indeksi)
MSE	Ortalama karesel hata
R	Geri çağırma
TN	Doğru negatif değerleri
TP	Doğru pozitif değerleri

1. GİRİŞ

Çevremizle sesli iletişim kurmamızı sağlayan yapıya konuşma denir. Akciğer ile başlayan bu yapı vokal kıvrımlar ile devam eder. Akciğerden gelen hava, vokal kıvrımlar arasında sıkışarak titreşimler oluşturur. Oluşan titreşimlerin ağız, dudak, dil ve burun yolunda şekil kazanması ile konuşma oluşur. Bu süreçte en önemli yapı vokal kıvrımlardır. Sağ ve sol vokal kıvrımlar arasındaki bölgeye glottis denir. Vokal kıvrımların titreşimi sırasında glottisin genişliği değişmektedir. (Şekil 1.1)



Şekil 1.1. Vokal Kordun Açık(sol) ve kapalı(sağ) durumlarında glottis

Konuşma ile ilgili önemli sorunlar glottisin yetersiz kapanmasından ya da kapanamamasından kaynaklıdır. Bu duruma çeşitli kistler, kanserler veya hastalıklar sebep olabilir. Bu patolojik bozukluklar sebebi ile konuşma sistemi işlevini ya büyük oranda ya da tamamen kaybetmektedir. İnsanlara özgü olan ve en önemli iletişim aracı olan konuşma gelişen teknoloji ile doğal olarak pekçok uygulama ve sektörde yaygın bir şekilde kullanılmaktadır. Konuşma analizi gerçekleştirilerek konuşmanın modellenmesi bu uygulamalarda performansı doğrudan etkilemektedir.

Ses üretim mekanizmasını ve ses bozukluklarının klinik teşhisini anlamayı amaçlayan çalışmalarda öncelik, ses telleri (vokal kord) titreşiminin özelliklerini iyi tanımak ve bu özelliklerin doğru biçimde yorumlanabilmesidir. Ek olarak bu çalışmalarda alınan görüntüler konuşma analizi içinde kaynak olarak kullanılmaktadır. Ses tellerinin vibrasyonunu incelemek için yapılan çalışmalarda son yıllarda yüksek hızlı endoskopik kameralarla alınan görüntüler kullanılmaktadır (Drioli ve Foresti, 2020, Khairuddin vd., 2020).

Glottis görüntülerinin alınabilmesi için 4000-20000 kare/saniye(fps) arasında çekim yapabilen HSV cihazlar kullanılmaktadır. Görüntüler alındıktan sonra manuel bölütleme teknikleri ile analiz edilir. Bu analizler görüntülerin titreşimli yapısı, glottisin kapalı olduğu veya çok küçük olduğu durumların bulunması gibi nedenlerden dolayı anlamlı sonuçlar sağlamakta zorlanmaktadır. Bu da güvenilirliği düşürmektedir.

Glottis tespitinde piksel tabanlı sınıflandırma yapılmaktadır. Bu işlem için HSV görüntülerinde glottisin diğer bölgelerden daha düşük piksel yoğunluğuna sahip olma özelliği kullanılır. Bu özellik sayesinde glottis bölgesi ve arka plan bölgesi olmak üzere 2 etiketli bir görüntü yapılabilmektedir. Fakat görüntülerin yanlış açı ile elde edilmesi, karanlık çekimler, ses tellerinin net görünür olmaması vb. sorunlar nedeni ile yanlış bölütleme sonuçları elde edilebilmektedir.

Ses tellerinin lokasyonunu tespit edebilmek görüntülerin işlenmesinde önemli bir aşamadır. Son yıllarda görüntü işleme uygulamalarında en çok tercih edilen yöntem derin öğrenmedir. Ancak bu uygulamaların ses telleri üzerine uygulaması çok azdır. Özellikle konuşma verileri kullanılarak glottis tespiti yapılmamıştır. Bundan dolayı, diğer yöntemlerin aksine doğrudan konuşma verisi kullanılarak hem glottis tespitinin hem de glottis bölütlemesinin yapılması planlanmıştır.

Glottis bölütlemesi için kullanılan yöntemler literatürde otomatik veya yarı otomatik olmak üzere iki başlık altında uygulanmaktadır. Otomatik yöntemlerde görüntüler bir sistem tarafından analiz edilerek ilgi bölgesi belirlenir ve bölütleme işlemi yapılır. Daha önce yapılan çalışmalarda ilgi bölgesinin otomatik

belirlenmesi için toplam yoğunluk değişimi, hareket kestirimi gibi yöntemler kullanılmıştır. İlgi bölgelerinin belirlenmesinin ardından glottis bölütlemesi ise Aktif Kontur (AC), Bölge Büyütme (RG), Histogram (H) gibi farklı geleneksel yöntemlere ek olarak İleri Beslemeli Sinir Ağları, Evrişimsel Sinir Ağları ve Gauss Karışım Modelleri de kullanılmıştır. Yarı otomatik bir yöntemde ise, araştırmacı ortaya çıkan sorunlarda sisteme ihtiyaç kadar müdahale edebilir. Yarı otomatik yöntemlerin uygulanabilmesi için video içerisinden belli bir görüntü aralığı seçilir. Uygunluğu onaylanan bu görüntü aralığındaki görüntülerin her biri glottisin minimum ve maksimum açık olduğu anların temsil görüntülerini içermelidir. Daha sonra bu görüntülerden glottal alanını (GA) içeren bir ilgi bölgesi belirlenir. İlgi bölgesi üzerinde glottis bölütlemesi yapılır.

Ses telleri (vokal kord) titreşiminin özelliklerini tanımlamak, ses üretim mekanizmasını ve ses bozukluklarının klinik teşhisini anlamayı amaçlayan çalışmalar için gereklidir. Bu çalışmalara yüksek hızlı dijital görüntüleme tekniklerinin uygulanması, bir hastanın gerçek vokal kord titreşimlerini çözebilecek bir frekansta titreşimli vokal kord görüntü dizilerinin yakalanmasını mümkün kılar (Koç, 2014). Ayrıca son zamanlarda görüntü alanında elde ettiği üstün başarılar sayesinde derin öğrenmenin temel mimarilerinden olan evrişimsel sinir ağları (ESA) medikal bölütleme alanında kullanılmaya başlanmıştır. Bu işlem için geliştirilen kodlayıcı ve kod çözücü yapılarından oluşan U-Net ve varyantı derin sinir ağı modelleri sayesinde ses telleri görüntülerinde bölütleme işleminde önemli ilerleme kaydedilmiştir. EGG, fonasyon sırasında göreceli ses kordonu temas alanındaki (VFCA) değişiklikleri ölçmek için yaygın olarak kullanılan, invazif olmayan bir yöntemdir (Hampala,2015). Ses telleri görüntüleri yanında konuşma ve elektroglottograf sinyallerinin analiz edilmesi sayesinde invazif olmayan yöntemler de kullanılmış olacaktır, bu sayede daha fazla özniteliğin elde edilmesi sağlanarak glotal bölgenin kestiriminin daha net bir şekilde yapılması amaçlanmaktadır. Bu çalışmanın amacı, ses telleri görüntülerinin yanı sıra konuşmadan elde edilen konuşma ve elektroglottograf sinyalleri de kullanılarak daha detaylı bilgi verecek ve medikal uygulamalarda daha verimli kullanılabilecek bir sistem geliştirmesini sağlamaktır. Konuşma analizi için derin öğrenme modelleri kullanılarak ve yeni

modeller geliştirilerek görüntü, konuşma ve elektroglottograf sinyallerinden özniteliklerin çıkarılması ve bu öznitelikler sayesinde ses telleri için bir tahmin yapılması ve ses tellerindeki herhangi bir problemin tespitinin daha hızlı, daha az maliyetli ve daha erken teşhis edilebilir olmasını sağlamak amaçlanmaktadır.

Bu çalışmada Konuşma sinyalinden glottis bölütlemesi yapabilen bir sistem geliştirilmiştir. Bu sayede yüksek maliyetli ve ulaşımı zor olan HSV kameralar yerine, hastanın konuşmasının işlenmesi ile glottis bölütlemesi yapılması ve tanı konulabilmesi amaçlanmaktadır. Klasik modellerin yanında özellikle biyomedikal görüntü bölütlemesinde başarısı kanıtlanmış olan U-net tabanlı derin öğrenme modellerinin geliştirilmesi, saniyede 4000 kare gibi yüksek hızlardaki görüntü üzerinde işlem yükünün minimize edilerek gerçek zamanlı uygulanabilirliği araştırılacaktır.

Ayrıca konuşma analizi için literatürdeki yöntemlerden farklı olarak ses bilgisinin yanında ses telleri görüntüleri kullanılması ve bu sayede daha gelişmiş modeller elde edilmesi amaçlanmaktadır. Konuşma ve elektroglottograf (EGG), gibi sinyaller birlikte konuşma analizi için kullanılarak konuşma üretim sistemi hakkında daha detaylı bilgi verecek yeni modeller geliştirilmeye çalışılacaktır. Ses ve görüntünün birlikte kullanılarak ses telleri için tahmin oluşturma işlemi önemli bir çalışma olduğu gibi medikal alanda bazı kolaylıklara da imkan sağlayabilecektir. Ses tellerinde görüntü inceleme imkânı veren medikal cihazların yüksek maliyetli olması ve bulunma güçlüğü gibi nedenlerden dolayı hastadan sadece ses kaydı alınarak ses tellerindeki herhangi bir problem invazif olmayan yöntemlerle tahmin edilebilmesi medikal alanda büyük kolaylıklar sağlayacaktır.

2. KAYNAK ÖZETLERİ

Manuel, yarı otomatik veya otomatik olmak üzere üç farklı şekilde glottis bölütleme yapılmaktadır. Her bir yönteme ait literatürde farklı çalışmalar bulunmaktadır.

Manuel sistemlerde genellikle eşikleme, histogram, aktif kontur, bölge büyütme vb. algoritmalar ile bölgenin bulunması sağlanır (Yan vd., 2012, Zhang vd., 2010, Aghlmandi ve Faez,2012). Yarı otomatik modellerde kullanıcının istenen bir anda elle müdahale edebilmesi mümkün olduğundan avantajlı modeller oldukları söylenebilir. Kök noktaları, hareket kestirimi, toplam yoğunluk değişimi gibi algoritmalar yarı otomatik modellerde tercih edilen yöntemlerdir (Pinheiro vd., 2014, Andrade-Miranda,2017). Tam otomatik yöntemlerde ise kullanıcı tarafından herhangi bir müdahale yapılmamaktadır. Günümüzde makine öğrenmesi yöntemleri ve en yaygın olarak da derin öğrenme algoritmaları ile tam otomatik bölütleme yapılmaktadır (Rao vd., 2018, Schenk vd., 2015, Kopczynski, vd., 2018, Hamad vd., 2019).

Alku ve diğerleri (2019), sesli konuşmanın kaynağının tespitinde kullanılan gırtlaksı ters filtreleme (GIF) tekniklerinin doğruluğunu ölçmek için kullanılabilen ve 4 farklı sistem kullanılarak oluşturulmuş 4 adet veri setinden oluşan OPENGLOT platformunu tanıtmışlardır. Bu veri setlerinden uygulanan GIF algoritmasının çalışmasına uygun olanını kullanarak doğruluk sonuçlarını elde etmek mümkündür. Bu platform GIF algoritmalarının doğruluğunu ölçmek için kullanılan ortak bir platform eksikliğinden hareketle oluşturulmuştur. Rao ve diğerleri (2018), Glottal titreşim şeklinin ana adımlarından biri olan ve glottis lokalizasyonu ile glottis bölütlemesi olarak iki ana kısma ayrılan otomatik glottis bölütleme için derin sinir ağı (DNN) tabanlı bir lokalizasyon ve bölütleme şeması önermişlerdir. Her piksel ve komşuluğundaki diğer piksellerin glottis ve background şeklinde iki sınıfa ayrılması şeklinde bir sınıflandırma problemi ortaya koymuşlardır. Problemde 3 konuşma dili patoloğu tarafından işaretlenmiş 18 deneğin stroboskopik videolarından oluşan bir veri seti kullanılmıştır.

Önerilen DNN modeli ile %65.33'lük bir lokalizasyon performansı ve 0.74 bölütleme dice skoru elde etmişlerdir.

Son zamanlarda derin öğrenme medikal görüntü işleme alanında büyük bir yol kat etmiştir. 2015 yılında geliştirilen U-Net derin öğrenme modeli sayesinde biyomedikal görüntülerin bölütlenmesinde yüksek bir doğruluk oranına ulaşılmıştır. Model içerisinde barındırdığı kodlayıcı kısmında görüntüden öznitelikleri çıkarırken, kod çözücü kısmında çıkardığı öznitelikler ile görüntüyü birleştirerek bölütlenmiş görüntüyü elde etmemizi sağlar.

Jha ve diğerleri (2020), U-Net modeli baz alınarak yeni pek çok model geliştirilmiştir. Bu modellerden biri DOUBLE U-Net modelidir. Model U-Net modelini geliştirmeyi amaçlayan literatürdeki farklı modellerden biridir. Double U-Net, CVC-ClinicDB veri seti üzerinde ve Küçük, orta ve büyük ölçekli deri lezyonlarının sınırlarının belirlenmesi görevlerinde test edilmiştir.

Belagali ve diğerleri (2020), iki adımlı bir evrişimsel sinir ağı (CNN) modeli önermişlerdir. 1. Adım CNN(CNN-1) lokalizasyon işlemi için kullanılırken 2. Adım CNN(CNN-2) bölütleme işlemi gerçekleştirir. Temel yapısı itibariyle U-Net varyantı olan bu önerilen model sayesinde önceden Rao ve diğerleri (2018) tarafından önerilen DNN modelinde %65.33 olan lokalizasyon doğruluğu %90 seviyelerine çıkartılmış ve dice skoru %65 'e çıkartılarak CNN modellerin bölütlemede daha iyi olduğu gösterilmiştir.

SA-UNet modeli de U-Net modeli baz alınarak geliştirilen modellerden biridir. Modelin avantajlarından biri eğitilebilir parametre sayısının diğer modellere göre ciddi oranda düşük olmasıdır. Bu sayede hem eğitim maliyeti hem de eğitim süresi çok düşük olmaktadır. Modeli diğerlerinden ayıran mimarisinde barındırdığı Spatial Attention Module (SAM) adlı yapıdır. Yapı sinir ağının kodlayıcısının son katmanından aldığı verileri ayrı ayrı maksimum havuzlama ve ortalama havuzlama katmanlarından geçirerek birleştirir.

Guo ve diğerleri (202), tarafından göz damarları bölütleme işleminde kullanılan SA-UNet 2 farklı veri seti üzerinde test edilmiş ve ikisinde de karşılaştırıldığı modellerden daha iyi sonuç vermiştir. SA-UNet veri setlerinden birinde 0.8263 F1 skoru elde ederken diğer veri setinde 0.8153 F1 skoru elde ederek başarım sağlamıştır.

3. MATERYAL VE YÖNTEM

3.1. Veri Seti

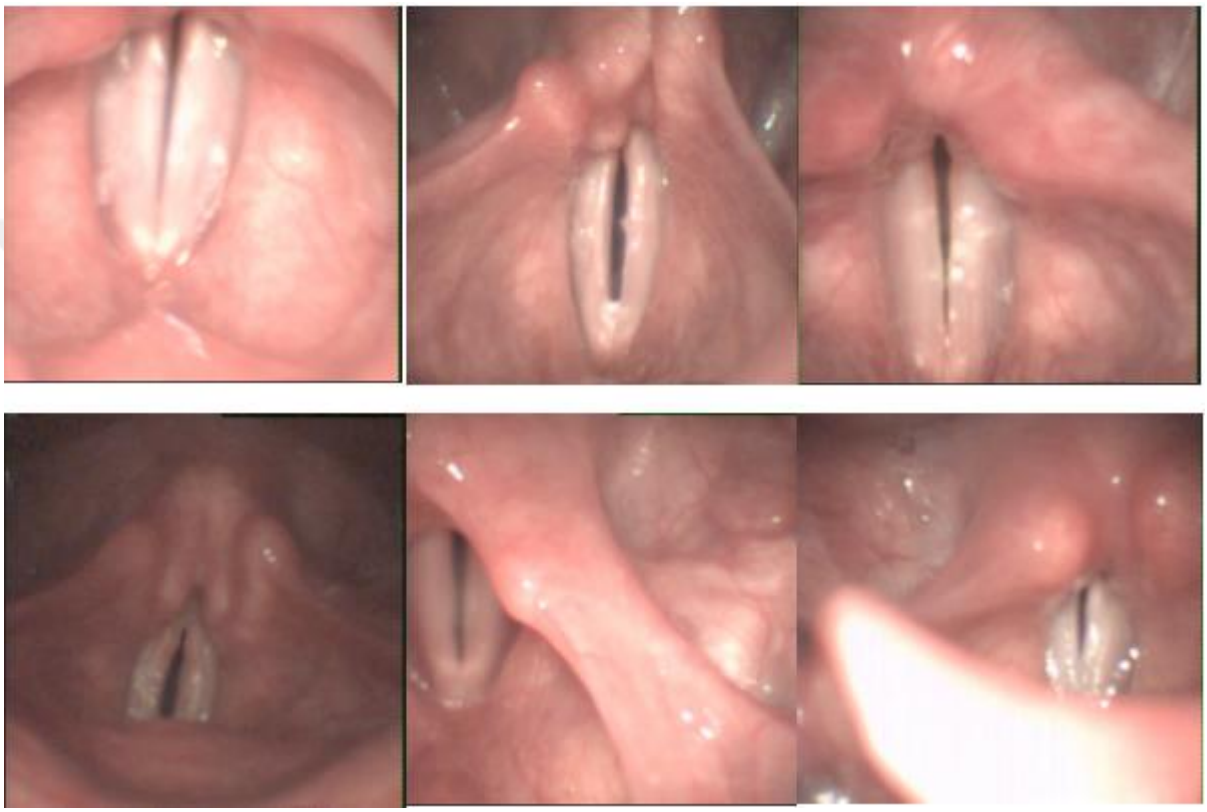
Yapılacak olan ses ve konuşma analizinin farklı modelleri için farklı veri setleri kullanılmıştır. Veri seti çeşitliliği modellerin genelleştirme başarımını artırmak için önemlidir. Veri setleri eğitim, test ve doğrulama şeklinde üç parçaya ayrılır. Eğitim seti modeli eğitmek için kullanılırken doğrulama seti modelin parametrelerini güncellemek için kullanılır. Test seti ise modelin genelleme olarak adlandırılan, hiç görmediği veriler üzerindeki başarımını test etmek için kullanılır. Veri setinin bölünmesi ile elde edilen üç farklı veri setinin rastgele biçimde dağılması önemlidir. Eğitim setinin tamamı benzer örnekler içerirse ve doğrulama veya test setinde örneklerin benzerliği az ise modelin parametreleri ve test değerlendirme sonuçları uygun sonuç vermeyecektir. Bu sebepten dolayı veri setleri ayrılırken düzgün dağılımlı bir şekilde seçilmeye çalışılmıştır. Düzgün dağılım için veri setinin özellikle uç noktaları olan glottisin maksimum ve minimum açıklık anları belirlenip eğitim, doğrulama ve test setlerine yüzdelik miktarlarına uygun bir şekilde dağıtılmalıdır.

3.1.1. Ircam veri seti

Yapılan çalışmada IRCAM HSV veri tabanından 3000 adet 256x256 boyutlarında görüntü kullanılmıştır (Degottex ve Bianco, 2010). Bu veri setinde kullanılan görüntüler ses telleri netliği, glottis konumu, açıklık-kapalılık durumu değişkenlik gösteren görüntülerdir. Veri setinden örnek görüntüler Şekil *3.1*'de görülmektedir. Bu veri seti görüntüleri genellikle aydınlık seviyesi yüksek görüntülerdir.

Glottis tespitini zorlaştıran en büyük etkenlerden biri, görüntüdeki glottis alanının arka plan alanından çok daha küçük olmasıdır. Glottis bölgesinin küçüklüğü ile birlikte glottisin tam kapalı olduğu görüntülerde hiç glottis pikselinin bulunmaması görüntüler üzerinde çalışılmasını zorlaştıran etkenlerden birisidir. Bu veri setinde bulunan 3000 görüntüden 1481 tanesi

hiçbir glottis pikseli içermeyen görüntülerdir. Bu 1481 adet glottis içermeyen görüntü eğitim, test ve doğrulama setlerine düzgün dağılmazsa modelin ya aşırı uyumuna ya da eksik uyumuna sebep olacaktır. Her iki durumda da genelleştirme skoru düşük kalacağı için modelin verimli ve düzgün bir model olması beklenemez. Yapılan çalışmada glottis içermeyen görüntülerin veri setlerine yüzdelik oranlarına göre dağıtılması sağlanmıştır.

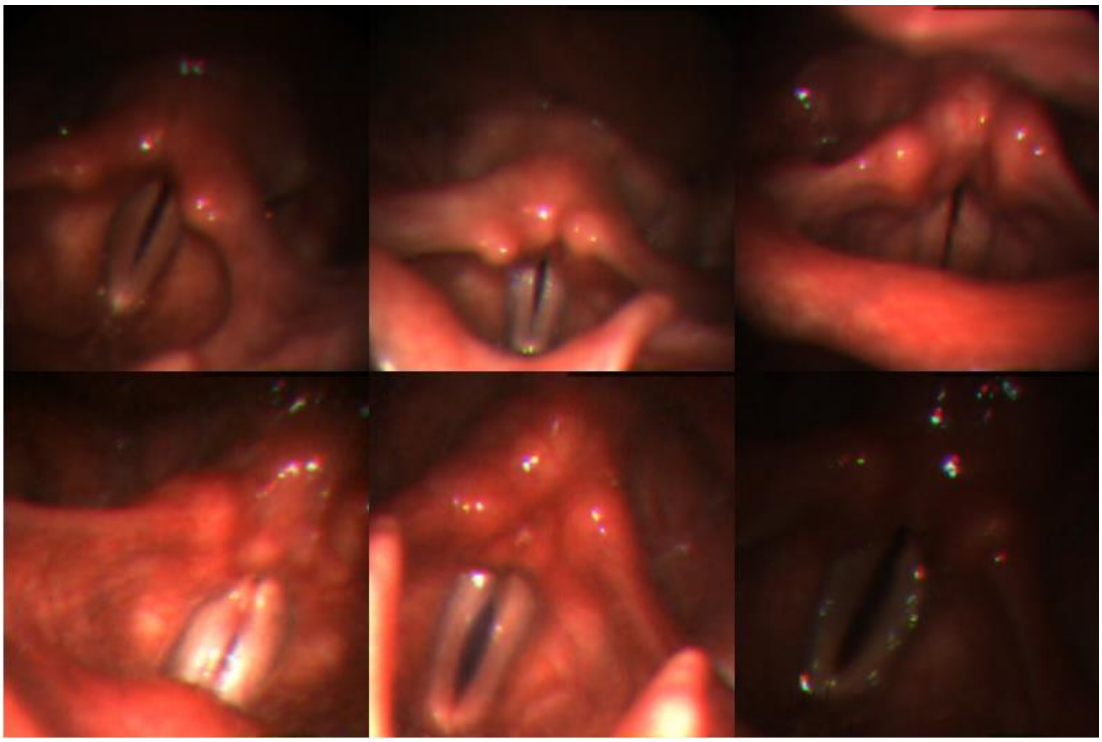


Şekil 3.1. IRCAM veri setinden elde edilen örnek görüntüler

3.1.2. Fehling veri seti

Yapılan çalışmada Fehling vd. (2020), tarafından oluşturulan diğer bir veri seti kullanılmıştır. Bu veri seti modelin genelleştirme durumunu incelemek ve tahmin edilecek sınıflar hakkında öznitelik öğrenmek gibi çeşitli sebepler için kullanılmıştır. Bu veri setinde 13 HSV videosundan elde edilen toplam 13000 görüntü bulunmaktadır. Görüntüler IRCAM veri setindeki ile benzer şekilde 256x256x3 boyutundadır. Şekil *3.2*'de veri setinden örnekler verilmiştir. Görüntülerden de anlaşıldığı üzere bu veri setindeki görüntüler IRCAM veri

setindeki görüntülerden çok daha fazla karanlık piksel içermektedir. Karanlık pikselin fazla olması glottis olarak tanımlanan bölgelerin yanlış yerlerde oluşması sonucunu da beraberinde getirebilir. Buna engel olmak amacıyla görüntüler üzerinde aydınlatma teknikleri kullanılarak özellikle kenarlarda yoğunlaşan karanlık piksellerin aydınlatılması sağlanabilir.

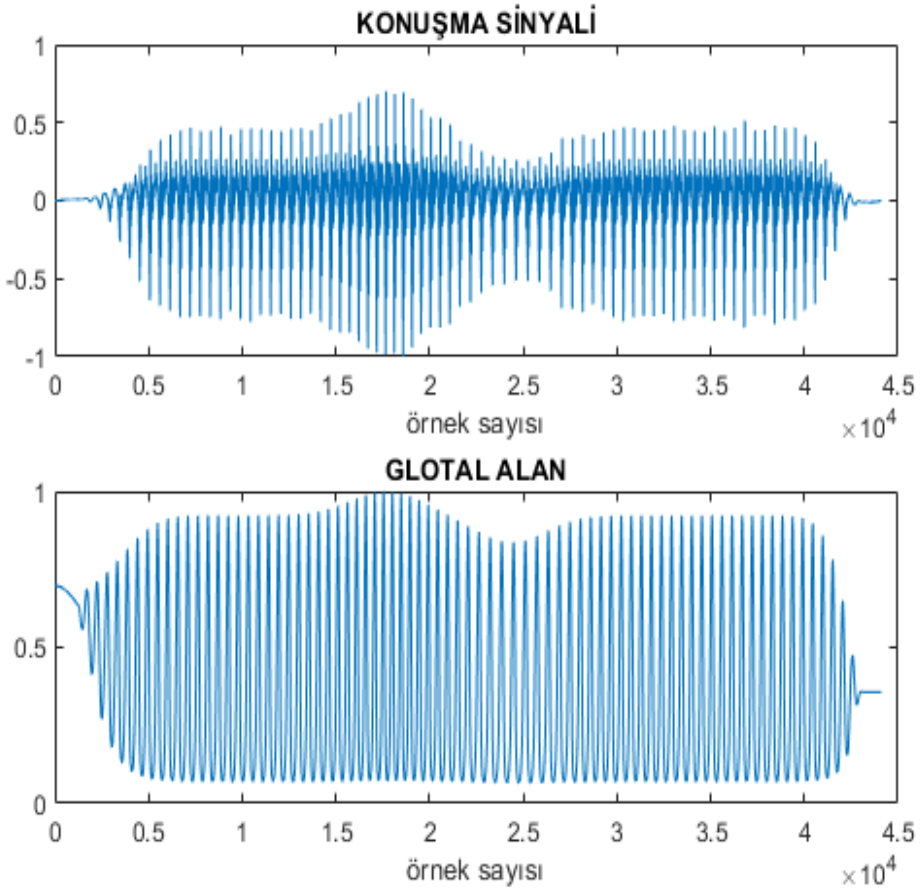


Şekil 3.2. Fehling tarafından oluşturulan veri setinden elde edilen örnek görüntüler (Fehling vd.,2020)

Bu veri seti içerisindeki görüntüler sağlıklı denekler ile fonksiyonel bozuklukları ve organik lezyonları olan patolojik deneklere ait görüntülerdir. Veri seti oluşturulurken hastalık tespiti için görüntü analizi özelinde model oluşturulduğu için her bir görüntü arka plan, sol ses teli, sağ ses teli ve glottis olmak üzere 4 farklı etiket ile etiketlenmiştir. Bu çalışmada ise sadece glottis tespiti yapılması amaçlanmaktadır. Dolayısı ile veri setinin çalışmada kullanılabilmesi için sol ve sağ ses tellerinin etiketleri arka plan olarak değiştirilmiştir. Böylelikle glottis ve arka plan şeklinde 2 etiketli görüntüler elde edilmiştir.

3.1.3. Openglot veri seti

Ses ve konuşma analizinde önemli bir aşama olan konuşma verisinden glottal alanın tespiti çalışmasında Alku vd., (2018) tarafından oluşturulmuş olan openglot veri seti kullanılmıştır. Veri seti 4 farklı havuz şeklinde oluşturulmuştur. Bu havuzlardan 2 numaralı havuz içerisinde aynı kişiye ait konuşma, glottal flow ve glottal alan verileri bulunmaktadır. Çalışmada kullanılan veri seti 48 erkek ve 48 kadın denekten alınan 96 adet veriden oluşmaktadır. Şekil 3.3'te openglot veri setinden örnek bir veri gösterilmiştir.



Şekil 3.3. Openglot veri setinden elde edilen bir sinyal. (Üstte konuşma sinyali, altta glotal alan)

Veri setinin farklı sınıflandırma türleri de bulunmaktadır. Fonasyon, addüksiyon açısı, cinsiyet ve temel frekans olmak üzere her veride 4 etiket vardır. Verilerdeki konuşma sinyalleri 44100 örnek içermektedir. Bu örnek sayısı verilerin derin öğrenmede işlenmesini çok zorlaştığı için alt örnekleme yapılması gerekmiştir.

3.2. Yapay Sinir Ağları ve Derin Öğrenme

Temel amacı insanların yapabileceği zeka gerektiren işleri gerçekleştirmektir. İlk yıllarda bilgisayar biliminde öncü isimlerin ortaya attığı "bilgisayarın düşünebilmesi" meselesi düşünsel işlevleri taklit eden yapılara yapay zeka adının verilmesini sağlamıştır. Chollet (2018) yapay zekayı entelektüel görevlerin insanlar tarafından yapılması yerine, otomatikleştirilmesi olarak tanımlamıştır.

Yapay sinir ağları 1950'li yıllarda başlayan yapay zeka uygulamalarından biri olmasına rağmen yapay zekadan daha önce uygulanmaya başlamıştır. 1944'te ilk olarak önerilen sinir ağı sisteminin ilk eğitilebilir örneği olan The Perceptron (eğitilebilir tek bir sinir ağı) Frank Rosenblatt tarafından 1957'de literatüre dahil edilmiştir. Modern zamanlarda da kullanılmaya devam eden sinir ağı modeli Şekil'te gösterilmiştir. İlk sinir ağından bugüne aktivasyon(activation) fonksiyonları ve geri bildirim algoritmaları gibi eğitim kalitesini artıran parametrelerin çeşitliliği artmıştır. Derin öğrenme günümüzde en başarılı sinir ağı modellerini barındıran yapay zekanın alt dallarından biri olarak pek çok alanda kullanılmaya ve yüksek başarım elde etmeye devam etmektedir.

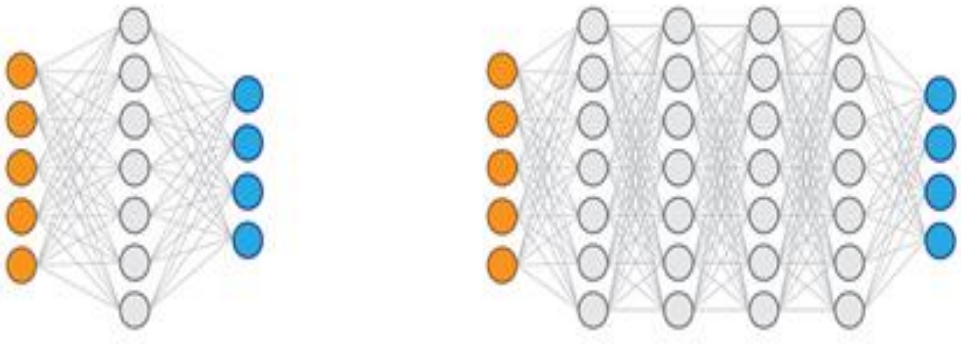
Yapay zeka, Makine öğrenmesi ve derin öğrenme alanlarını da kapsayan geniş bir alandır. Yapay zekanın temel amacı insanların normal olarak gerçekleştirebildiği görevleri otonom olarak yapabilecek makinelerin gerçekleştirmesini sağlamak iken, Makine öğrenmesinde amaç, bilgisayarlara bir oluşum veya görev sürecinin nasıl işlediğini öğretmektir.

Klasik programlamada Şekil'teki gibi kurallar ve veriler girdi olarak kullanılıp, çıktı olarak cevaplar elde edilirken, Makine öğrenmesinde klasik programlamadan farklı olarak veri ve cevaplar girdi olarak kullanılır ve veri ile cevaplar arasındaki kuralların öğrenilmesi sağlanır. Kurallar öğrenildikten sonra yeni veri girişleri ile yeni verilere karşılık gelen cevaplar üretilebilir.

Şekil 3.4. Klasik programlama ile makine öğrenmesinin probleme yaklaşım biçimleri (chollet,2018)

3.2.1. Yapay sinir ağları

Bir adet giriş katmanı, bir adet çıkış katmanı ve bir veya daha fazla gizli katmandan oluşan ve bilgi işleyen modele yapay sinir ağı denir. Bu modeller eğer bir adet gizli katman içeriyorsa sığ yapay sinir ağı modeli diye adlandırılır. İki veya daha fazla gizli katmandan oluşan modellere ise derin yapay sinir ağı modeli veya daha yaygın bir kullanımla derin öğrenme modeli denir. Sığ ve derin sinir ağlarına karşılık gelen yapılar Şekil'te gösterilmiştir. Yapay sinir ağlarında katman sayısının artması sonucu işlem sayısı ve hesaplama süresi artmaktadır. Sınıf sayısının veya veri setinde veri miktarının arttığı durumlarda çoğunlukla nöron sayısı ve katman sayısı yüksek olan modeller tercih edilir (Nielsen, 2015, Aggarwal, 2018).



Şekil 3.5. Sığ sinir ağı modeli(sol) ve derin öğrenme modeli(sağ)

Giriş katmanından modele gönderilen veriler gizli katmanlarda işlenerek öznitelikler elde edilir. Bu özniteliklerden elde edilen sonuçlar tahmin verisi

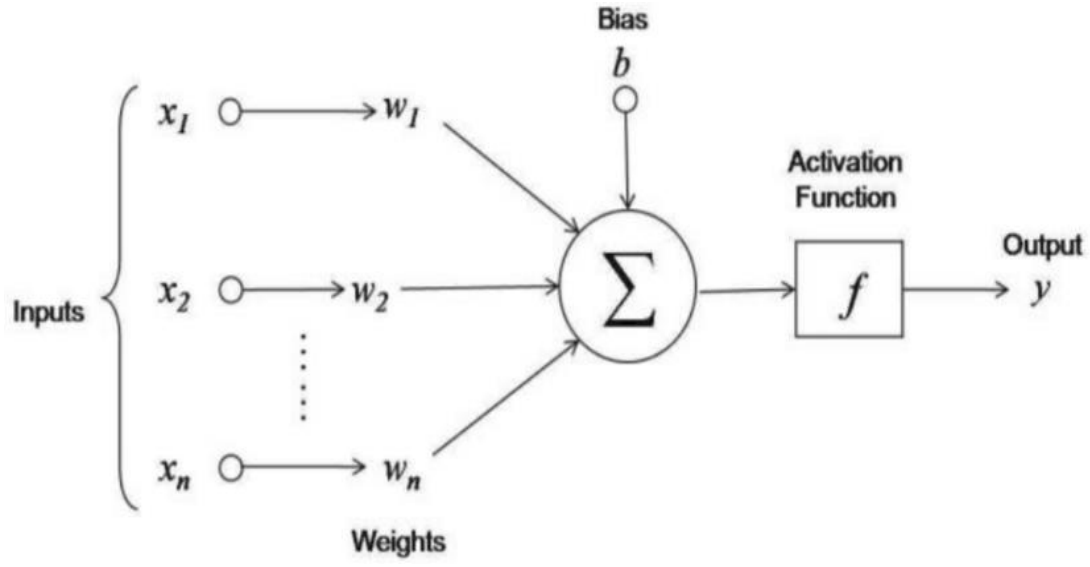
olarak çıktı katmanına gönderilir. Çıktı katmanının sonucu bir kayıp fonksiyonu ile işlenerek tahmin verisi ile gerçek veri arasındaki doğruluk skoru elde edilir. Geri yayılım algoritmaları sayesinde skor gizli katmanlara tekrar gönderilerek modelin gizli katmanlarındaki ağırlık verilerinin yeniden ayarlanması sağlanır. Bu sayede doğruluk skoru iyileştirilmiş olur.

Yapay sinir ağları denklem 3.1'deki gibi tanımlanır:

$$y = f(b + \sum_{i=1}^{n} x_n * w_n) \quad (3.1)$$

Yukarıdaki denklemde x_n giriş verileri (Inputs) ve y çıkış verisidir(output). B bias yani önyargı değişkenini ifade ederken w_n ise ağırlıkları ifade eder. Denklemde f fonksiyonu modelin çıktısını elde etmemizi sağlayan ve çıktıyı genellikle 0 ve 1 arasına sıkıştıran aktivasyon fonksiyonudur.

Formülden de anlaşılacağı üzere geri yayılım ile değiştirilebilen değişkenler önyargı ve ağırlıklardır. Şekil'te yapay sinir ağları denkleminin şematik olarak gösterimi verilmiştir.

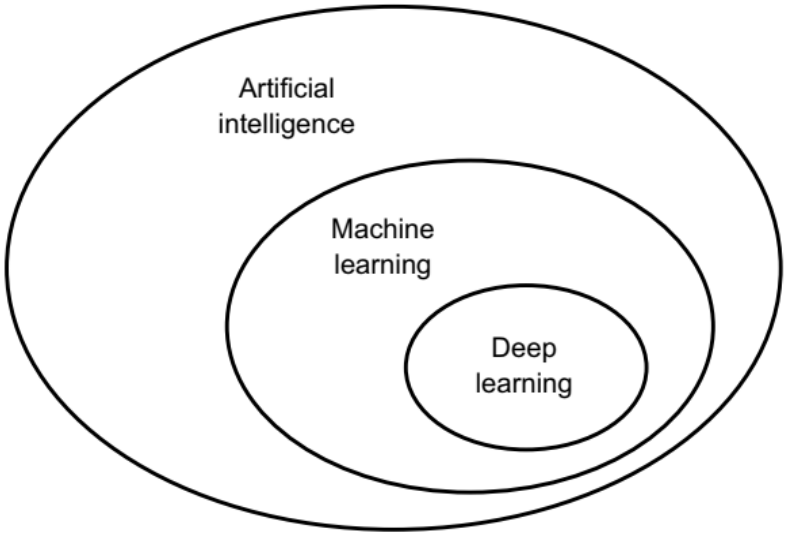


Şekil 3.6. Bir yapay sinir ağı örneği

Sinir ağları, düğümlerle birbirine bağlanmış girift yapıda bir modeldir. Düğümler nöron katmanının tamamını birbirine bağlar. Düğümlerin her birinin bilgi işlenirken sahip olduğu bir ağırlık değeri vardır. Aktivasyon fonksiyonu düğümde hesaplanan değeri eşik değer aralığına sıkıştırarak bir sonraki katmana iletir. Aynı zamanda sinir ağlarının doğrusallığını bozarak ve doğrusal olmayan özellikler kazandırarak sinir ağlarının daha zorlu problemler ile baş edebilmesini sağlar. Aktivasyon fonksiyonlarının en yaygın kullanılanları RELU, softmax ve sigmoid fonksiyonlarıdır. (Derin öğrenme, goodfellow).

3.2.2. Derin öğrenme

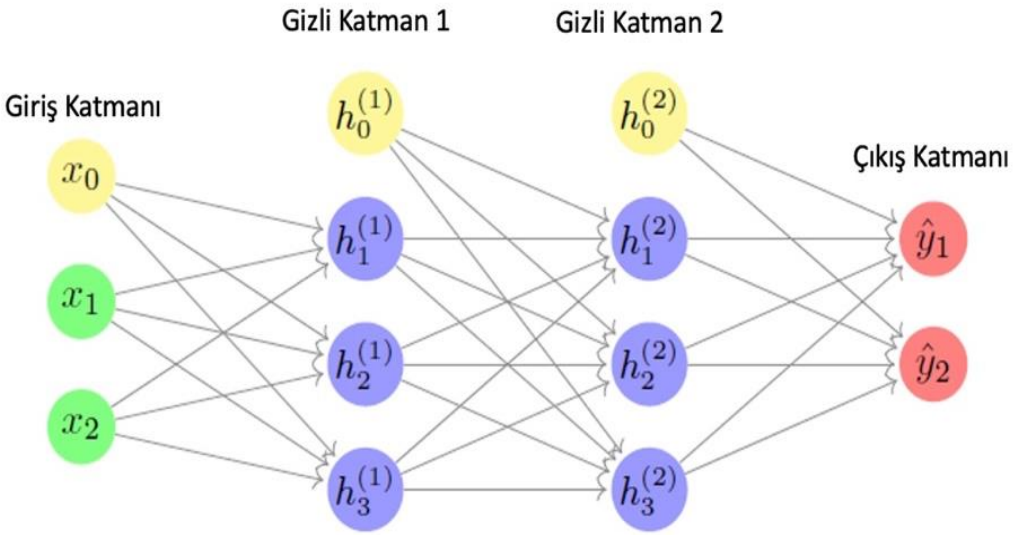
Son yıllarda elde ettiği başarılar sayesinde yapay zeka (YZ) pek çok bilimsel araştırmada kullanılır hale gelmiştir. Yapay zeka, normal şartlarda insanlar tarafından yapılan entelektüel görevleri otomatikleştirme çabasıdır. Bu nedenle YZ, makine öğrenimi ve derin öğrenmeyi de kapsayan geniş bir alandır. İlk YZ algoritmaları 1950li yıllarda uygulanmıştır. 1950'li yıllardan beri gelişimini sürdüren yapay zekanın önemli alt dallarından biri olan makine öğrenmesinin alt dalları bulunmaktadır. Bu alt dallar Denetimli öğrenme, denetimsiz öğrenme, yarı denetimli öğrenme ve pekiştirmeli öğrenme olmak üzere dört ana başlıkta incelenir. Bu çalışmada denetimli öğrenme yapılmıştır. Denetimli öğrenme, verilerin üzerinde etiket bilgilerinin yer alması ve bu etiketlere uygun gelen girdi ve çıktıların belirtilerek modelin eğitiminin yapılmasıdır. Çalışmada kullanılan denetimli öğrenme için derin öğrenme modelleri tercih edilmiştir. Derin öğrenme 2014'te ImageNet yarışmasında elde ettiği başarıdan bu yana pek çok alanda öncü olarak kullanılan bir makine öğrenmesi uygulaması olmayı başarmıştır. Yapay zeka, makine öğrenmesi ve derin öğrenme arasındaki ilişki Şekil 3.'da gösterilmiştir.



Şekil 3.7. Yapay zeka, makine öğrenmesi ve derin öğrenme arasındaki ilişki

Derin öğrenme, makine öğrenmesinin alt dalı olan yapay sinir ağlarının çok katmanlı olarak geliştirilen modellerini ifade eder. Bu modeller arasında sıklıkla kullanılanlar evrişimsel sinir ağları (ESA), yinelemeli sinir ağları, çekişmeli üretici ağlardır. Yapay sinir ağı katmanlarından oluşan derin öğrenme modelleri giderek daha anlamlı temsilleri öğrenmeyi vurgulayan, verilerden öğrenme yöntemlerine yeni bir bakış açısıdır (Chollet, 2018). Derin öğrenmede kullanılan öğrenme kelimesi makine öğrenmesi kelimesi ile aynı manayı taşımaktadır. Derin öğrenme modellerinin, makine öğrenmesi algoritmalarından farkı daha anlamlı kurallar üretebilmesidir.

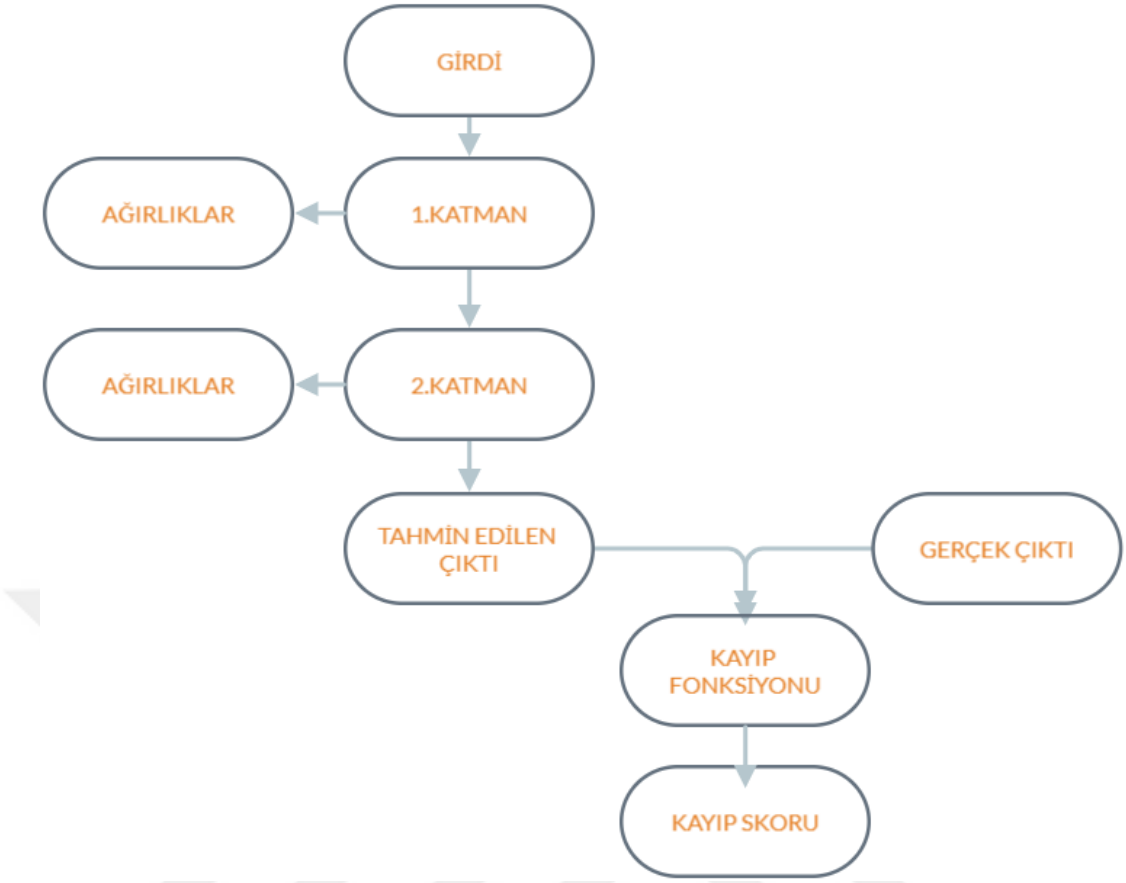
Bir derin öğrenme modeli klasik programlama gibi doğrudan programlanmak yerine, girilen veriler vasıtası ile birtakım parametreler üzerine eğitilir. Modele öğretilmek istenen görevle ilgili optimum sayıda örnek sokulur ve parametreler ayarlanarak sistemin otomatik olarak en uygun sonuçları bulacak şekilde eğitilmesi sağlanır. Sesten, glottal alanı tespit eden bir sistem eğitilmek istendiğinde sisteme ses verileri ve ses verilerine karşılık glottal alanlar modele girdi olarak verilir. Bu verileri kullanan model sesten glottal alan tahminini en iyi şekilde yapan parametreleri öğrenerek tahmin işini otomatik hale getirmeyi amaçlar. Derin öğrenme modelleri bir adet giriş katmanı, bir adet çıkış katmanı ve farklı fonksiyonlar içeren birden fazla gizli katmandan oluşur (Şekil *3.*).



Şekil 3.8. Bir derin öğrenme modeli örneği

3.2.3. Kayıp fonksiyonu ve en iyileme algoritmaları

Derin öğrenme modellerinde önemli adımlardan biri kayıp skorunu elde etmektir. Girdi verileri katmanlara rastgele olarak verilen ağırlık değerleri ile eğitilir. Eğitim sonucunda çıktı olarak verilen tahmin değeri ile gerçek değer bir kayıp fonksiyonunda işlenerek kayıp skoru elde edilebilir (Şekil *3.*). Kayıp skoru farklı fonksiyonlar ile ölçülebilir. İki etiketli sınıflandırma işlemlerinde genellikle binary cross entropy tercih edilirken iki etiketten daha fazla etiket sınıfı içeren veri setlerinin kayıp skorlarını hesaplamak için categorical cross entropy tercih edilmektedir. Yüksek örnekleme sahip verilerde ise Mean Square Error (MSE), Mean Absolute Error (MAE) gibi fonksiyonlar tercih edilir. Bu farklı kullanım yerlerine bakarak kullanılan fonksiyonun kayıp skorunu düzgün tahmin etmede ve genelleştirmenin artmasında önemli olduğu söylenebilir. Yani en küçük kayıp skorunu sağlayabilen kayıp fonksiyonu her zaman doğru seçim olmayabilir.



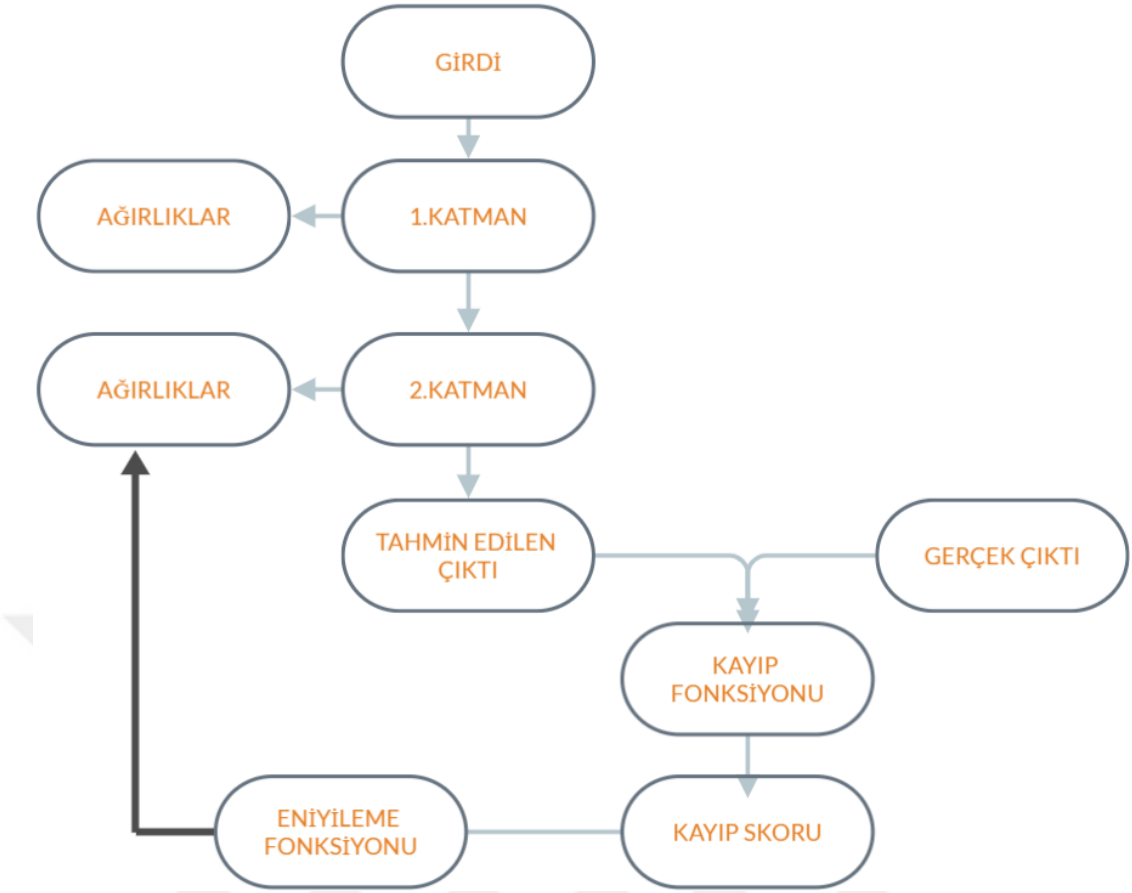
Şekil 3.9. Kayıp fonksiyonu ile bir modelin kayıp skorunun belirlenmesinin akış diyagramı

Katman ağırlıkları başlangıçta rastgele atandığı için model performansını gösteren kayıp skoru yüksek çıkacaktır. Kayıp skorunu düzeltmek için sistemin bir geribesleme ile skor azaltması sağlanmalıdır. Bu geribesleme yöntemi ağırlıkları optimize edip kayıp skorunu düzeltene kadar eğitim işlemini tekrarlamalıdır. Geri yayılım ile eniyileme fonksiyonunun Ağrılıkları güncellemesinin akış diyagramı Şekil 3.10'da gösterilmiştir. Şekilde ağırlıkların optimizasyonunu yaparak ağırlıkların yeniden hesaplanmasını sağlayan blok diyagram gösterilmiştir. Başlangıçta rastgele olan ağırlıklar, optimizasyon sayesinde uygun aralığa çekilir ve kayıp skoru olabildiğince azaltılmış olur. Kayıp skorunun azalması neticesinde en doğru tahmin verileri elde edilir. En iyi çıktıyı elde etmeyi amaçlayan bu algoritmaya eğitim döngüsü denir. Eğitim döngüsü adımları şu şekildedir:

1. Yeterli sayıda örnek içeren veri seti ve veri setine karşılık gelen çıktı değerler tanımlanır.

2. Yapay sinir ağı tasarlanır ve veri seti ağa gönderilerek çıktılar elde edilir.
3. Elde edilen çıktılar ile gerçek çıktı değerleri arasındaki kayıp miktarı kayıp fonksiyonları ile hesaplanır.
4. Her bir katmanın ağırlıkları skoru iyileştirmeye çalışacak şekilde optimizasyon algoritması ile yeniden güncellenir.

Bu şekilde süren bir eğitim sürecinde her hesaplama sonunda daha iyi ağırlıkların elde edilmesi beklenir. Bu öğrenim sürecinde ortaya iki sorun çıkabilir. Bu sorunlardan biri modele çok fazla iterasyon yaptırmanın sonucu olarak giriş verilerini ezberlemesidir. Bu duruma Aşırı öğrenme denir. Aşırı öğrenilmiş model, eğitim verileri üzerinde çok başarılıdır ancak test verileri veya eğitim verileri haricindeki veriler üzerinde başarımı çok düşüktür. Bu durumda elde edilen model düşük skorlu olduğu için kullanılamayacaktır. Bir diğer sorun ise modelin çok basit olması veya problemi karşılayacak sayıda veri olmaması sonucunda ortaya çıkan eksik öğrenmedir. Eksik öğrenme durumunda hem eğitim verilerinde hem de eğitim verileri hariç tüm verilerde kayıp skoru yüksek ve model başarımı düşüktür. Bu durumlardan kurtulmak için modelin genelleştirme performansının artırılmasını sağlayan en uygun iterasyon sayısı manuel olarak seçilmelidir. Doğrulama ve test işlemlerinin skorları sonucunda uygun iterasyon sayısına ulaşılabilir.



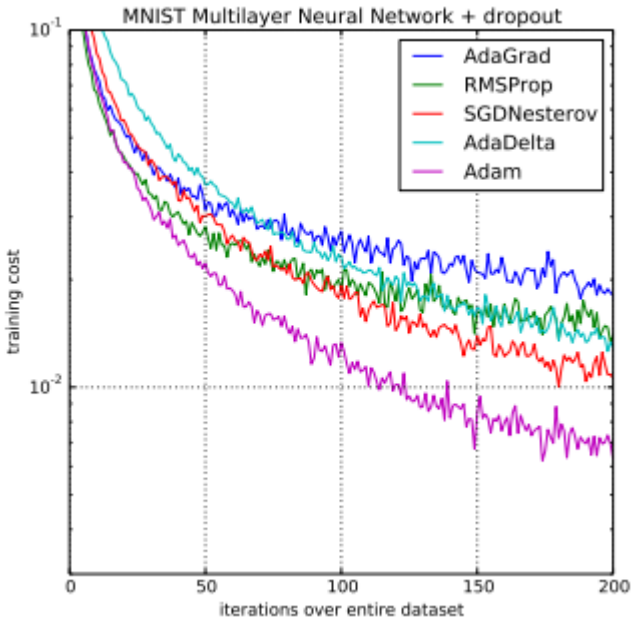
Şekil 3.10. Eniyileme fonksiyonunun ağırlıkları yeniden güncellemesi (Chollet,2018)

Yapay sinir ağlarında, ağırlıkların değerlerini ayarlamak için geribesleme sisteminde optimize edilmesi gereken değer olarak kayıp skorları kullanılır. Bu ayarlama Stokastik Gradyan İnişi (SGD), Geri Yayılım Algoritması, Adam gibi geri yayılım algoritmaları tarafından yapılır.

$$w_{t+1} = w_t - a * \left(\frac{\partial L}{\partial wt}\right) \tag{3.2}$$

Yukarıdaki denklemde ∂L / ∂wt Mevcut gradyanı, a öğrenme katsayısını wt ise mevcut ağırlığı göstermektedir. Denklem derin öğrenmenin ilk zamanlarında en çok kullanılan ve derin öğrenmenin gelişimine en çok katkısı olan SGD algoritmasına aittir (Ruder vd., 2016). SGD algoritmasının eğitim esnasında optimum noktayı ararken aşırı salınım yapmasından dolayı Momentum algoritması geliştirilmiştir. Bu algoritma salınımı azaltarak eğitim maliyetini azaltmaktadır. Bu şekilde ilerlemesini sürdüren optimizasyon algoritmalarından

ADAM algoritması günümüzde en çok tercih edilen algoritmadır. Momentum ve Rmsprop algoritmalarının avantajlı yönleri kullanılarak ortaya çıkarılmıştır. (Kingma ve Ba, 2014). Bu çalışmada eniyileme algoritması olarak ADAM algoritması kullanılmıştır. Şekil *3.*'de MNIST veri seti üzerinde Adam algoritmasının yaygın olarak tercih edilen diğer algoritmalardan daha düşük bir kayıp skoruna ulaşabildiği görülmektedir.



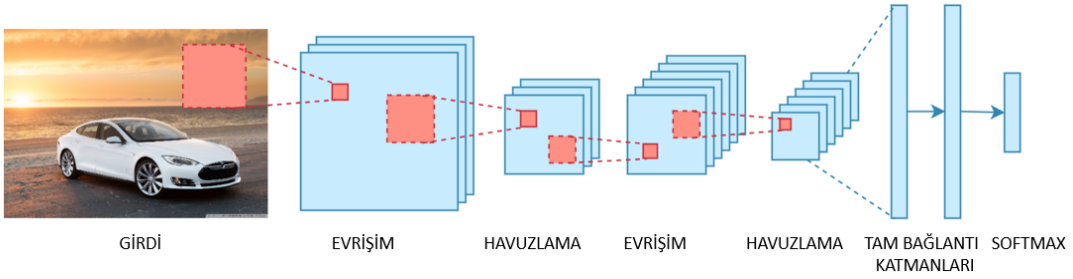
Şekil 3.11. MNIST veri seti üzerinde Adam algoritmasının başarımı (Kingma ve Ba, 2014)

3.2.4. Evrişimli sinir ağları

Derin öğrenmede kullanılan farklı sinir ağı mimarileri vardır. Günümüzde en yaygın kullanılan çeşitlerden biri de evrişimsel sinir ağlarıdır (ESA).

Bilgisayarlı görü ve görüntü işleme alanlarında derin ağların büyük olasılıkla en önemli bileşeni eğitilebilir çok katmanlı evrişimlerdir (Szeliski, 2022). Gradyan azalması ile eğitilen çok katmanlı derin ağların, geniş veri setlerinden karmaşık, yüksek boyutlu, doğrusal olmayan bağlamları öğrenme yeteneği, onları görüntü tanıma problemleri için önemli bir aday haline getirir (Szeliski, 2022). Şekilde evrişimli sinir ağı içeren bir derin öğrenme modeli görülmektedir. Evrişimli sinir ağı içeren modellerde, Giriş katmanı ile veri seti modele girdi olarak verilir. Sonra

veriler evrişim (konvolüsyon) katmanında verilerden farklı özniteliklere göre işlenir ve sınıflandırılmış veriler çıkış katmanından elde edilir. Elde edilen sonuç ve gerçek sonuç arasında hesaplanan kayıp skoru geri yayılım algoritması ile katmanlara gönderilerek ağırlıklar güncellenir. Her bir iterasyonda kayıp skoru azaltılarak hedeflenen sonuca en yakın değer elde edilir. İlerleyen bölümlerde her bir katmanın özellikleri anlatılacaktır.



Şekil 3.12. Bir evrişimli sinir ağı modeli

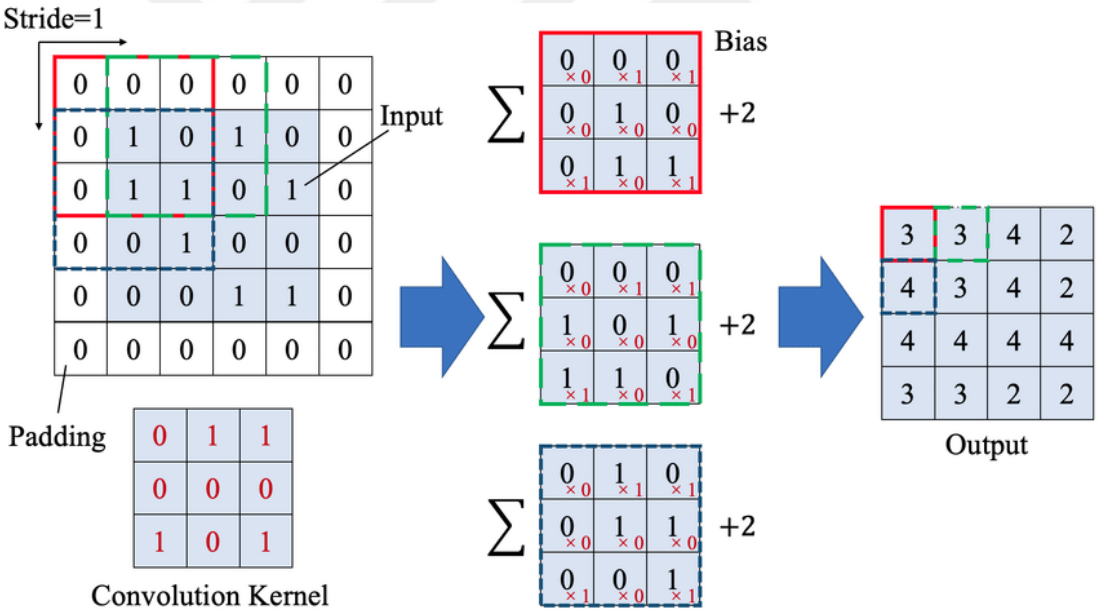
3.2.4.1. Giriş katmanı

Giriş katmanı diğer modellerde de olduğu gibi evrişimli sinir ağı modellerinin ilk katmanıdır. Bu katmanda veri seti girdi olarak ağa gönderilir. Veri setindeki veri sayısının fazla olması genellikle iyi bir durumdur ve modelin doğruluğunu artırır ancak eğitim süresi ve maliyetin artması da olumsuz yönler olarak karşımıza çıkar. Optimum sayıda veriden daha yüksek veri ile çalışmak ise modeli aşırı öğrenmeye götürebilir. Optimum sayı altındaki veri setlerinde eğitim süresi iyileştirilebilir ancak bu durum da az öğrenme problemine sebep olabilir.

3.2.4.2. Evrişim katmanı

Evrişim katmanı, Evrişim işleminin yapıldığı katmandır. Evrişimli sinir ağlarına adını vermesinden de anlaşıldığı üzere en önemli katmandır. Genellikle ilk evrişim katmanları kıvrımlar, dokular, kenarlar vb. genel öznitelikleri yakalarken, son katmanlar daha öznel olan nesne şekli, nesne boyu vb. öznitelikleri yakalar.

Bu katmanlarda çeşitli filtreler kullanılarak farklı özniteliklerin yakalanması sağlanır. Bu işlemi yapan filtreler özel olarak tasarlanabilir. Filtrelerin sayısı, boyutları, çekirdek büyüklüğü gibi özellikleri model mimarisi oluşturulurken belirlenebilen özelliklerdir. ESA eğitilirken yapılan geri besleme ile filtrelerin katsayılarının yeniden belirlenmesi sağlanarak doğruluğun artırılması amaçlanır. Şekil *3.*'te gösterilen bir evrişim işlemidir. Uygun filtre seçimini yapmak için yapılan iterasyonlar sonucunda farklı çıktılar elde edilebilir. Filtre seçimi ise eniyileme algoritması ile yapılır. Evrişim işleminde adım büyüklüğü değişkenine göre çıktı boyutu ayarlanabilir. Adım büyüklüğü filtrenin girdi üzerinde nasıl ilerleyeceğini belirleyen büyüklüktür. Adım büyüklüğü şekildeki gibi bir olarak ayarlandığı takdirde girdi ile çıktı aynı boyutta olur. Ancak birden büyük adım büyüklükleri seçildiği takdirde çıktı boyutu girdi boyutundan daha büyük olacaktır.

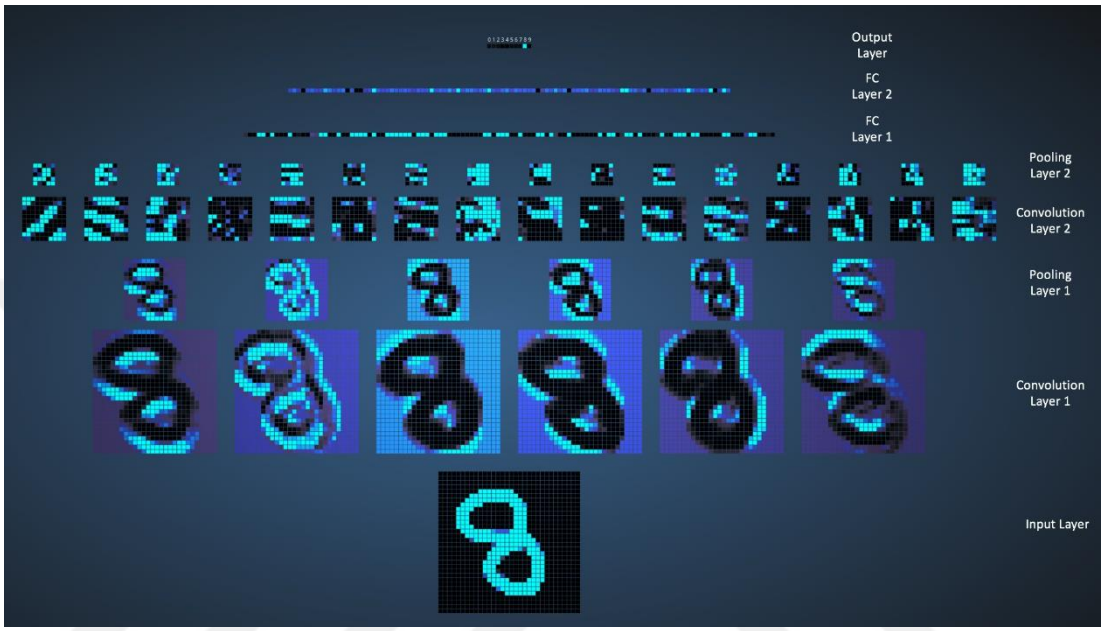


Şekil 3.13. Bir sinir ağında girdiye uygulanan evrişim işlemi

Burada gösterilen padding(doldurma) işlemi ise görüntünün çıktıyı uygun boyuta getirmek için sıfırlarla doldurulması işlemidir. Bu işlem modelin ezberleme oranını düşürmek için de yapılabilir.

Şekilde bir sekiz sayısı görüntüsü girdi olarak verildiğinde farklı evrişim katmanlarının bulduğu öznitelikler görülmektedir. Bu öznitelikler sonucunda

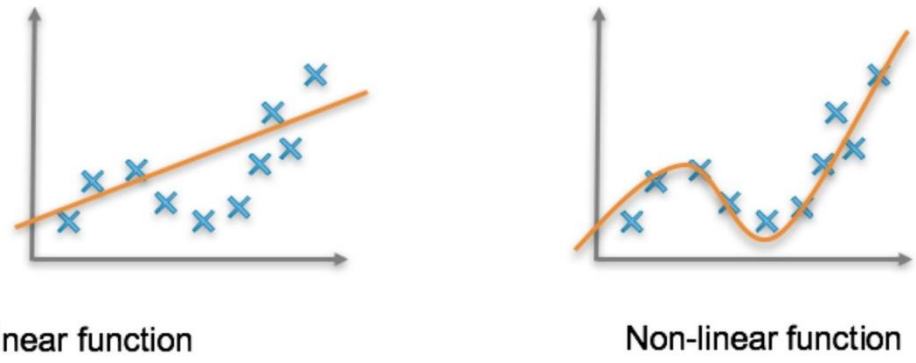
sekizin neye benzediğini anlayabilen evrişimli sinir ağı modeli sınıf etiketleri için belli olasılıklar üretir. Bu durumda etiketlerin rakamlar olması gerekir ve 10 farklı sınıf için olasılık sonuçları alınır. Bu olasılıklar içerisinden en yüksek olanı eşik değer ile belirlenerek sonuçta rakamları tanıyabilen bir model oluştuğu gözlenebilir.



Şekil 3.14. 8 sayısının evrişim katmanları ile öznitelik haritasının oluşturulması

3.2.4.3. Aktivasyon katmanı

Yapay sinir ağlarının doğrusallığı karmaşık işlemlerin üstesinden gelemez. Evrişimli modellerde de aynı sorun yaşanmaktadır. Bu sorundan kurtulmak amacıyla evrişim katmanının çıktısı bir aktivasyon fonksiyonu yardımıyla doğrusallıktan kurtarılır. Şekil *3.*' te doğrusallığın (linearity) aktivasyon fonksiyonu ile bozulmasına örnek verilmiştir. Şekle göre verilerin doğrusal bir denklem tarafından düzgün tahmin edilmesinin mümkün olmadığı açıktır. Bu durumda doğrusallığı bozmak için bir aktivasyon fonksiyonu kullanmak bir zorunluluktur. Farklı aktivasyon fonksiyonlarının model doğruluğu üzerine etkileri farklı farklı olacağından doğru aktivasyon fonksiyonu seçimi önem arz etmektedir.



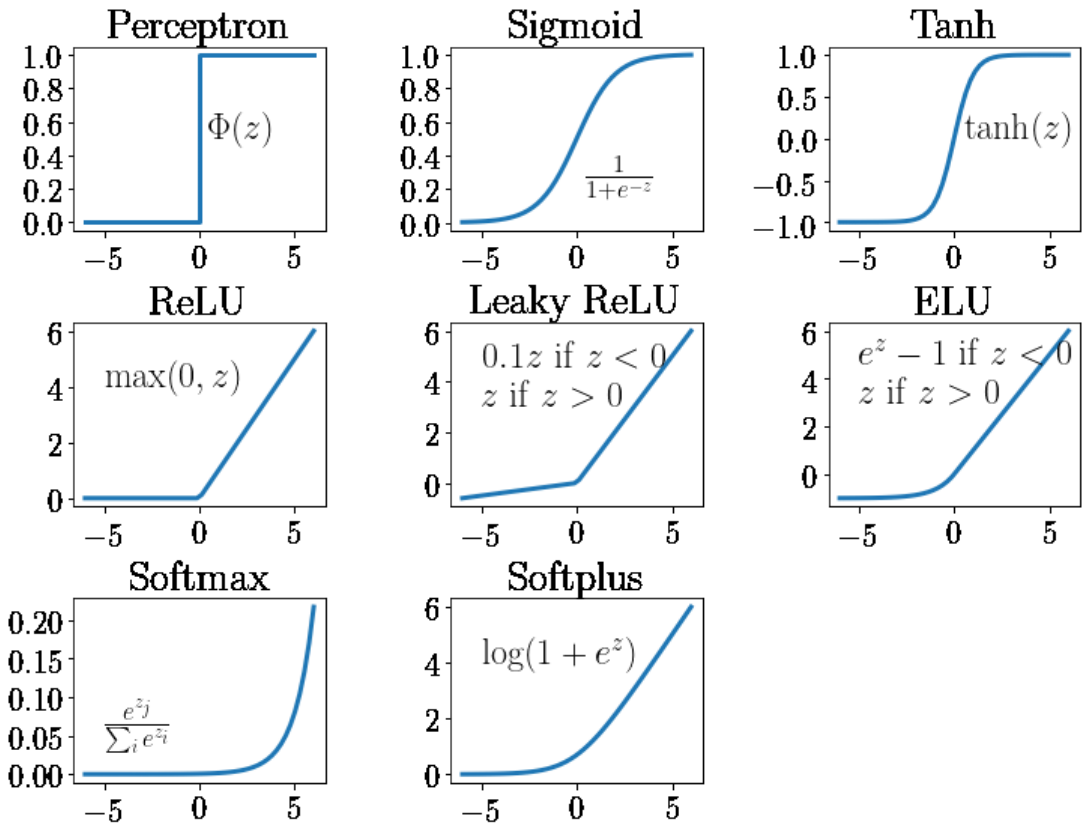
Şekil 3.15. Aktivasyon fonksiyonu yardımı ile doğrusallığın bozulması (Bellamkonda, 2019)

Aktivasyon fonksiyonlarından sıklıkla tercih edilenler Basamak fonksiyonu, doğrusal fonksiyon, sigmoid fonksiyonu, ReLU fonksiyonu, sızıntı ReLU fonksiyonu, hiperbolik tanjant fonksiyonu, softmax fonksiyonu ve swish fonksiyonudur. Şekil 3.'da fonksiyonlara ait denklemler verilmiştir. Denklemlere bakıldığında fonksiyonların farklı durumlarda diğerlerine üstünlük sağladığı anlaşılabilir. Dolayısı ile bu çalışmada genelleştirme başarımı yüksek olan ve benzer uygulamalarda kullanılan hiperbolik tanjant fonksiyonu ve sigmoid fonksiyonu kullanılmıştır.

AKTİVASYON FONKSİYON	DENKLEM	ARALIK
Doğrusal Fonksiyon	$f(x) = x$	$(-\infty, \infty)$
Basamak Fonksiyonu	$f(x) = \begin{cases} 0 \; için \; x < 0 \\ 1 \; için \; x \geq 0 \end{cases}$	$\{0, 1\}$
Sigmoid Fonksiyon	$f(x) = \sigma(x) = \frac{1}{1 + e^{-x}}$	$(0, 1)$
Hiperbolik Tanjant Fonksiyonu	$f(x) = \tanh(x) = \frac{(e^{x} - e^{-x})}{(e^{x} + e^{-x})}$	$(-1, 1)$
ReLU	$f(x) = \begin{cases} 0 \; için \; x < 0 \\ x \; için \; x \geq 0 \end{cases}$	$[0, \infty)$
Leaky (Sızıntı) ReLU	$f(x) = \begin{cases} 0.01 \; için \; x < 0 \\ x \; için \; x \geq 0 \end{cases}$	$(-\infty, \infty)$
Swish Fonksiyonu	$f(x) = 2x\sigma(\beta x) = \begin{cases} \beta = 0 \; için \; f(x) = x \\ \beta \to \infty \; için \; f(x) = 2\max(0, x) \end{cases}$	$(-\infty, \infty)$

Şekil 3.16. Çeşitli aktivasyon fonksiyonlarına ait denklemler (kızrak, 2019)

Şekil 3.'de kullanılan fonksiyonlardan bazılarının grafikleri verilmiştir. Verilen fonksiyonların doğrusal olmadığı görülmektedir ve bu doğrusal olmama (nonlinear) durumu sayesinde doğrusal girdi olarak verilen evrişim katmanı çıktılarına doğrusal olmayan nitelikler kazandırmaları mümkün olmaktadır.

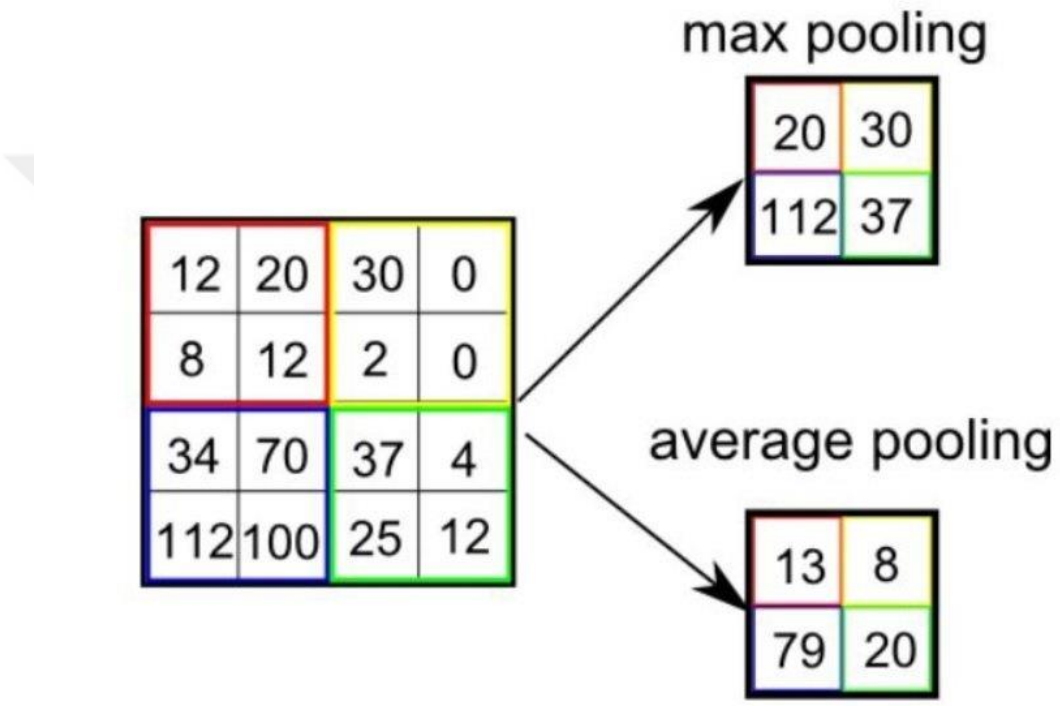


Şekil 3.17. Aktivasyon fonksiyonlarına ait grafikler

3.2.4.4. Havuzlama katmanı

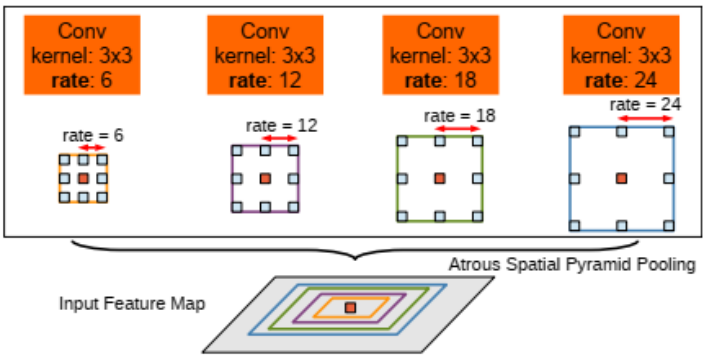
Havuzlama katmanında, evrişim katmanına benzer şekilde filtreleme yapılır ancak evrişim işleminden farklı bir işlem yapılır. Havuzlama katmanının görevi öznitelik çıkarmak değil giriş verisinin boyutunu değiştirmektir. Bu değişiklik sayesinde sonraki evrişim katmanının daha spesifik özellikler keşfetmesi sağlanır. Yaygın olarak kullanılan iki tür havuzlama vardır. Maksimum havuzlama ve ortalama havuzlama. Maksimum havuzlama işleminde matrisin seçilen bölgesinden maksimum değer alınarak tek piksel şeklinde çıktıya verilir. Böylece

belli bölgelerdeki maksimum değerlerden oluşan yeni bir matris elde edilir. Ortalama havuzlamada ise matristen seçilen bölgedeki tüm piksellerin ortalaması alınır. Böylece belli bölgelerdeki ortalama değerlerden oluşan yeni bir matris elde edilmiş olur. Matrisin boyutu da yeniden ayarlanmış olur. Şekil *3.*'de 4x4 boyutunda bir matris giriş olarak verilerek ortalama ve maksimum havuzlama yapılarak boyutunun 2x2 şekline getirilmesi gösterilmiştir.



Şekil 3.18. Maksimum(üst) ve ortalama(alt) havuzlama (Islam,2020)

Bu çalışmada maksimum ve ortalama havuzlamaya ek olarak Atrous Spatial Pyramid Pooling (ASPP) isimli havuzlama metodu da kullanılmıştır. Bu havuzlama çeşidinde veriler matrisin genelinden belli bir dağılım oranına göre seçilmektedir. Şekil *3.*'da ASPP türü havuzlamanın nasıl yapıldığı gösterilmiştir. Bir pikselin etrafındaki 3x3'lük matris oluşturan pikselleri belirtilen oranlarda genişleterek evrişim işlemi uygulanır. Daha sonra tüm sonuçlar birleştirilip tekrar evrişim işlemi uygulanır ve havuzlama işlemi yapılmış olur.



Şekil 3.19. Atrous Spatial Pyramid Pooling (ASPP) havuzlamanın uygulanışı (Jha vd., 2020)

3.2.4.5. Bırakma (Drop out) katmanı

Bu katmanda modelin aşırı uyumunun önüne geçmek amaçlanır. Aşırı uyumu önlemek için henüz matematiksel olarak net bir yöntem yoktur. Ancak eğitim verilerini artırmak ve çeşitlendirmek, düzenleme (regularization), erken durdurma, en iyi modeli kaydetme gibi çeşitli çözümler bulunmaktadır (Ying,2019). Bırakma katmanı da aşırı uyumu önlemekte başarım sağlayan bir uygulamadır. Bu katman kendisine giriş olan verilerden belirtilen orandaki kadar veriyi sıfıra eşitler. Bununla aşırı uyumun önüne geçilerek modelin genelleme kabiliyetinin artması hedeflenir (Srivastava vd.,2014).

3.2.4.6. Çıktı katmanı

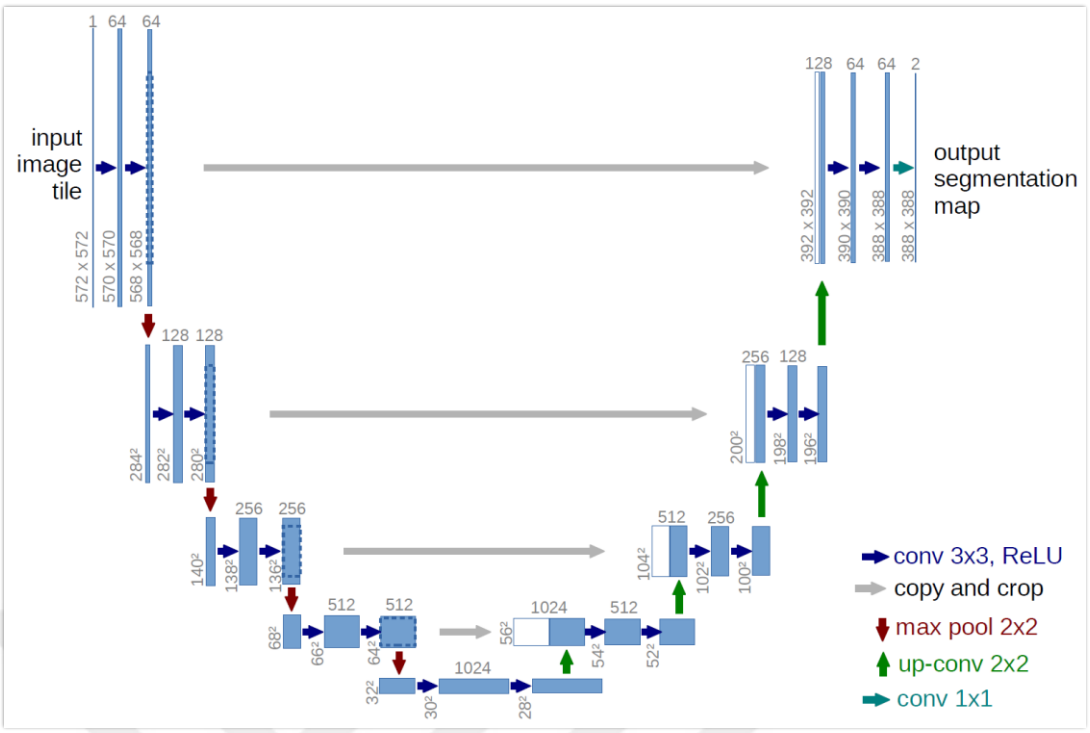
Modelin çıktısının elde edildiği katmandır. Son katman olduğu için tahmin etmek istediğimiz veriyi bize veren katmandır diyebiliriz. Eğer sınıflandırma yapılıyorsa sınıflara ait olasılık oranları bu katmandan elde edilir. Eğer kodlayıcı ve kod çözücü yapısı ile görüntü matrisinden yine görüntü matrisi elde edilecekse bu katmanın çıktısından elde edilir. Bu çalışmada 3 boyutlu görüntülerden 3 boyutlu görüntü tahminleri ve 1 boyutlu konuşma vektöründen glottal alana ait 1 boyutlu vektör tahmini yapılmıştır.

3.3. Biyomedikal Görüntü Bölütlemede Derin Öğrenme Kullanımı

Son zamanlarda derin öğrenme medikal görüntü işleme alanında büyük bir yol kat etmiştir. Derin öğrenme modellerinin evrişimli katmanlar kullanması sayesinde görüntüler üzerinden farklı özniteliklerin çıkarılması mümkün hale gelmiş ve bu sayede biyomedikal görüntülerin doğruluklarında önemli artışlar gözlenmiştir. Ayrıca kodlayıcı – kod çözücü içeren farklı sinir ağı mimarileri ile tahmin edilen görüntünün bölütlenmiş görüntü olması sağlanabilmektedir. Bu alanda kullanılan modellerden bazıları alt başlıklarda verilmiştir.

3.3.1. U-Net

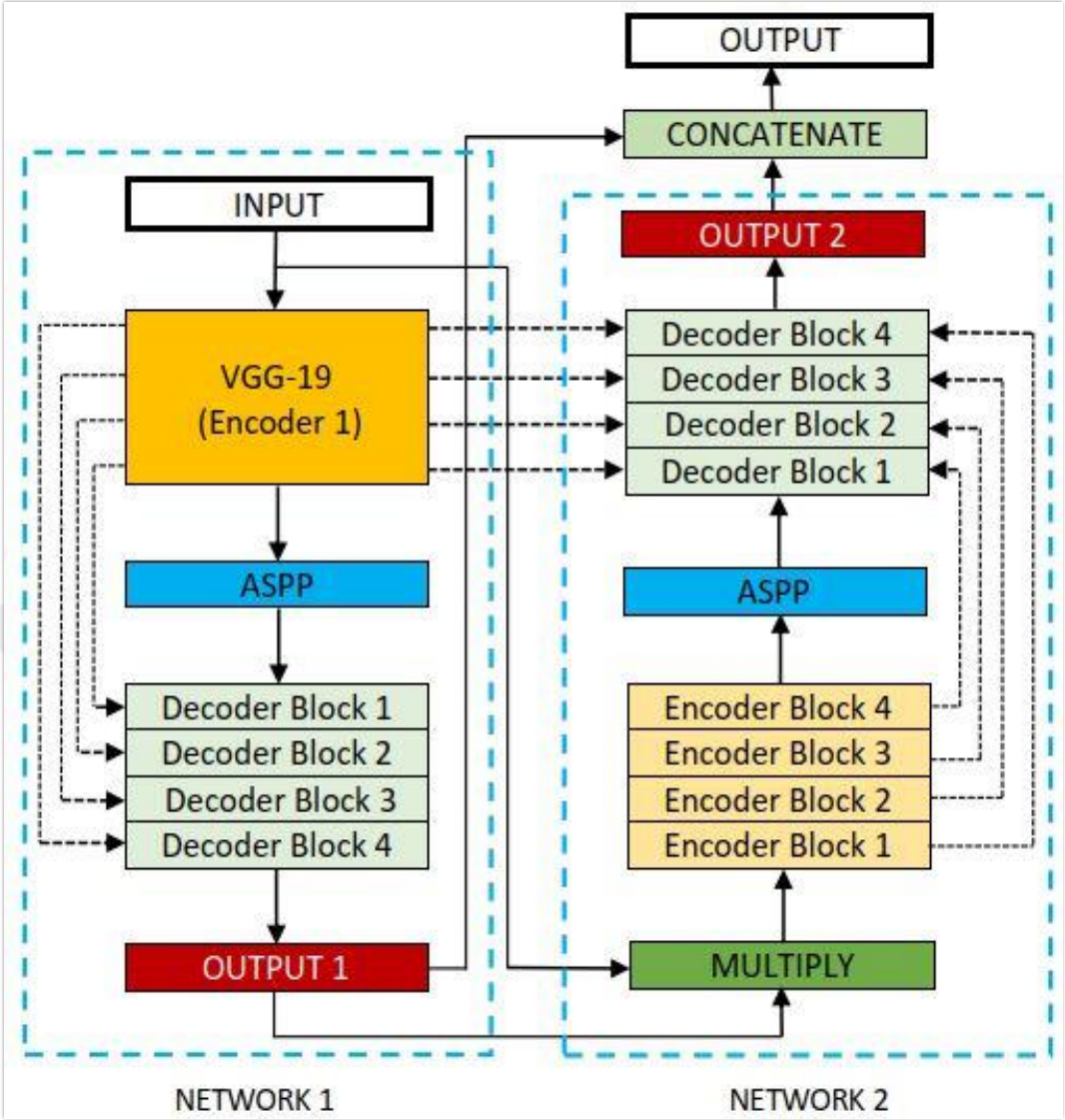
2015 yılında geliştirilen U-Net derin öğrenme modeli sayesinde biyomedikal görüntülerin bölütlenmesinde yüksek bir doğruluk oranına ulaşılmıştır. Şekil 2.1'de mimarisi gösterilen model içerisinde barındırdığı kodlayıcı kısmında görüntüden öznitelikleri çıkarırken, kod çözücü kısmında çıkardığı öznitelikler ile görüntüyü birleştirerek bölütlenmiş görüntüyü elde etmemizi sağlar. Model girdi olarak 256x256x3 boyutunda görüntüleri kullanıp, çıktı olarak 256x256x1 boyutunda bölütlenmiş görüntüyü verebilmektedir. U-Net modeli, Hela hücrelerinin tespiti görevinde 0.9203 IOU skoru ile rakip modelleri geride bırakmayı başarmıştır.



Şekil 3.20. U-Net mimarisi (Ronneberger vd., 2015)

3.3.2. Double U-Net

U-Net modeli baz alınarak yeni pek çok model geliştirilmiştir. Bu modellerden biri DOUBLE U-Net modelidir. Model U-Net modelini geliştirmeyi amaçlayan literatürdeki farklı modellerden biridir. Bu model 2 kodlayıcı ve 2 kod çözücü yapısı içermesi nedeni ile daha karmaşık öznitelikleri çözümleme kabiliyetine sahiptir. Derin öğrenmenin avantajlarından biri de önceden başka bir veri seti üzerinde eğitilmiş olan modelin ağırlıkların yeni geliştirile modellerde kullanılabilmesidir. Bu avantaj sayesinde DOUBLE U-Net'in birinci kodlayıcısında kullanılmıştır. DOUBLE U-Net modelinin birinci kodlayıcı yapısı VGG-19 modelini içermektedir. Birinci kodlayıcının çıktısı Atrous Spatial Pyramid Pooling (ASPP) ile havuzlama yapıldıktan sonra birinci kod çözücüye aktarılır. Birinci kod çözücü çıktısı ile giriş görüntüsü çarpılarak, ikinci kodlayıcıya aktarılır. İkinci kodlayıcı katmanlarının çıktısı yeniden ASPP ile havuzlama yapılarak ikinci kod çözücüye aktarılır. Sonuçta ise her iki kod çözücünün çıktıları birleştirilerek sonuç çıktı elde edilir.



Şekil 3.21. Double U-Net Mimarisi (Jha vd.,2020)

Double U-Net, CVC-ClinicDB veri seti üzerinde ve Küçük, orta ve büyük ölçekli deri lezyonlarının sınırlarının belirlenmesi görevlerinde test edilmiş ve U-Netten daha iyi sonuçlar elde edilmiştir. Otomatik polip tespiti üzerine eğitilen DOUBLE U-Net, dice skorunda (DCS) 0.7649 puan elde ederek U-Net'i 0,4729 puan farkla geçmeyi başarmıştır. mIOU skorunda da 0.6255 puan elde eden DOUBLE U-Net, U-Net'i 0,4496 puan farkla geçmiştir.

3.3.3. SA-UNet

SA-UNet modeli de U-Net modeli baz alınarak geliştirilen modellerden biridir. Modelin avantajlarından biri eğitilebilir parametre sayısının diğer modellere

göre ciddi oranda düşük olmasıdır. Bu sayede hem eğitim maliyeti hem de eğitim süresi çok düşük olmaktadır. Modeli diğerlerinden ayıran mimarisinde barındırdığı Spatial Attention Module (SAM) adlı yapıdır. Bu yapı havuzlama (pooling) işlemi yapan bir havuzlama katmanıdır. Yapı sinir ağının kodlayıcısının son katmanından aldığı verileri ayrı ayrı maksimum havuzlama ve ortalama havuzlama katmanlarından geçirerek birleştirir. Ortaya çıkan yeni veriyi bir evrişim katmanından geçirdikten sonra sigmoid aktivasyon fonksiyonuna gönderir. Ortaya çıkan yeni veri ile kodlayıcının son katmanındaki veri element bazlı çarpma işleminin ardından kod çözücünün ilk katmanına gönderilir.

3.4. Performans Ölçütleri

Derin öğrenme modelleri ile elde edilen modellerin doğrulama seti ve test seti üzerindeki başarımlarını gözlemlemek için farklı ölçütler kullanılmaktadır. Ölçütlerin kullanımına dikkat edilmelidir. Her veri seti için her ölçüt doğru sonuç vermeyecektir. Veri seti içerisinde her sınıfa ait eşit miktarda etikete olmaması gibi durumlarda bazı ölçütler hatalı sonuçlara sebep olacaktır.

Performans ölçütleri eğitim aşamasında eğitim seti ve doğrulama seti üzerinde kullanılarak modelin optimum uyumu yakalaması sağlanır. Uygun model elde edildikten sonra ölçütler test verisine uygulanarak modelin genelleştirme başarımı hakkında bilgi sahibi olunur.

İkili sınıflandırmada kullanılan ölçütler için aşağıdaki tanımlar gereklidir:

Doğru Pozitif (True Positive - TP) : Sınıflandırmada bulunması gereken pozitif değerlerden kaç tanesinin pozitif olarak bulunduğunu gösterir.

Doğru Negatif (True Negative - TN) : Sınıflandırmada bulunması gereken negatif değerlerden kaç tanesinin negatif olarak bulunduğunu gösterir.

Yanlış Pozitif (False Positive – FP) : Sınıflandırmada bulunması gereken pozitif değerlerden kaç tanesinin negatif olarak bulunduğunu gösterir.

Yanlış Negatif (False Negative – FN) : Sınıflandırmada bulunması gereken negatif değerlerden kaç tanesinin pozitif olarak bulunduğunu gösterir.

3.4.1. Doğruluk (D)

Derin öğrenme sıklıkla tercih edilen ölçütlerden birisidir. Modelin doğru tahmin ettiği kısmın tüm veriye oranını ifade eder. Doğruluk, verinin bir kısmının tamamına oranı olduğu için 0 ve 1 arasında bir değer alır. 0 değeri alması tamamen yanlış bir tahmin yapıldığına işaret ederken, 1 şeklinde bir tahmin yapılması tamamen doğru bir tahmin yapıldığı anlamına gelir.

Bu ölçüt her durumda düzgün çalışmaz. Görüntünün %98'inin arkaplan olduğu glottis bölütleme işlemi gibi durumlarda glottis tamamen arka plan olarak tahmin edilse bile doğruluk oranı %98 çıkabilir. Bu şekilde eşit dağılımlı etiketlere sahip olmayan veri setlerinde tercih edilmemesi gereken bir ölçüttür. Denklem 3.3'te doğruluk ölçütünün hesaplanışı verilmiştir.

$$ACC = \frac{TP+TN}{FP+TP+TN+FN} \tag{3.3}$$

3.4.2. Hassasiyet (HA)

Glottis olarak tahmin edilen bölgelerin gerçekte ne kadar glottis olabileceği ihtimalini verir. Denklem 3.4 hassasiyet formülünü göstermektedir.

$$HA = \frac{TP}{FP+TP} \tag{3.4}$$

Hassasiyet ne kadar yüksekse glottis o kadar doğru tahmin edilmiş demektir. Hassasiyetin düşmesi ise glottis bölgesinin arka plan üzerindeki bazı bölgelerde de tahmin edildiği anlaşılır.

3.4.3. Geri çağırma (R)

Glottis olduğu tahmin edilen piksellerin kaç tanesinin doğru tahmin edildiğinin tüm glottis tahminlerine oranıdır. Denklem 3.5'teki gibi hesaplanır.

$$R = \frac{TP}{FN+TP} \tag{3.5}$$

R glottis içinde arka planların oluşmasının ölçütüdür. R 1'e ne kadar yakınsa glottis o derece düzgün çıkar. R ne kadar 0'a yakınsa glottis o kadar arka planla kaplıdır. Dikkat edilirse HA ve R birbirine bağlıdır. Dolayısıyla iyi bir modelde hem R hem de HA 1'e yakın olmalıdır.

3.4.4. F1 skoru

HA ve R değerlerinin bir tanesi 1'e çok yakınken diğerinin 0'a çok yakın olması gibi sorunlarla karşılaşılmaktadır. Bu nedenle HA ve R tek başına doğru yorumlamaya olanak vermeyebilmektedir. F1 skoru, HA ve R'nin etkilerini birleştirmeyi amaçlar. Denklem 3.6'teki gibi hesaplanır. Formülden anlaşıldığı üzere HA ve R'nin harmonik ortalamasıdır.

$$F1 = 2 * \frac{PR*R}{PR+R} \tag{3.6}$$

F1'in model performansı ile ilgili önemli bilgiler verebilmesi için HA veya R'nin çok küçük olmaması gerekir.

3.4.5. Jaccard indeksi (Intersection over union)

İki farklı kümenin birbirine ne kadar benzediğinin ölçüsüdür. Basitçe kesişim kümesi eleman sayısının birleşim kümesi eleman sayısına bölünmesi ile elde edilir. Denklem 3.7'deki gibi ifade edilir.

$$IOU = \frac{A \cap B}{A \cup B} = \frac{TP}{FP + TP + FN} \tag{3.7}$$

Intersection over union (IOU) görüntü analizinde sıklıkla kullanılır. Bölütleme çalışmalarında lokalizasyonun ne kadar doğru olduğunun ölçüsünü ifade eder. IOU, 0 ve 1 arasında değer alır. 1'e ne kadar yakınsa glottisin görüntü üzerindeki yeri o kadar doğru tahmin edilmiş demektir.

3.4.6. Dice skoru (DCS)

DCS ölçütü ile iki görüntünün birbirine benzerlik oranı elde edilir. F1 skora benzeyen bir ölçüttür. Görüntü bölütleme uygulamalarında sıklıkla kullanılır. Denklem 3.8'deki gibi hesaplanır.

$$\text{DCS} = 2 * \frac{PR*R}{PR+R} \tag{3.8}$$

Bu çalışmada hem eğitim aşamasında hem doğrulama aşamasında hem de test aşamasında kullanılmıştır. Bölütleme uygulamasında glottis tahmininin yorumlanmasında fayda sağlamıştır.

3.4.7. Ortalama karesel hata(MSE)

MSE sayesinde tahmin edilen çıktıların gerçek çıktılardan ne kadar farklı olduğu hakkında bilgi edinilebilir. MSE 0 ve 1 arasında çıkmak zorunda değildir. Tahmin verileri gerçek değerlerden çok farklı ise hatanın karesini almasından dolayı çok yüksek sonuçlar (pozitif ve negatif sonsuza kadar) verebilir. MSE 0'a yaklaştıkça modelin kalitesi artıyor demektir. Farkların karesini alması sayesinde hatalara karşı hassasiyeti yüksektir. Denklem 3.9'daki gibi hesaplanır.

$$MSE = \frac{1}{n}\sum_{i=0}^{n}(y_i - y'_i)^2 \tag{3.9}$$

Yukarıdaki denklemde n veri setindeki veri sayısını, y gerçek çıktıyı, y' ise tahmin edilen çıktıyı ifade eder.

4. ARAŞTIRMA BULGULARI VE TARTIŞMA

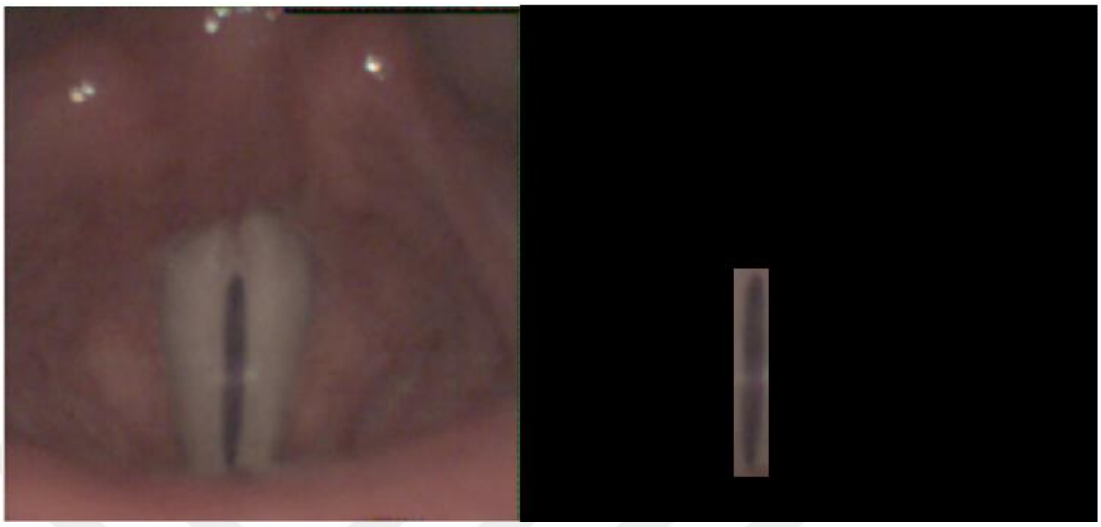
Bu çalışmada ses ve konuşma analizi için üç farklı çalışma yapılmıştır. Bu çalışmalardan biri U-Net ile bölütleme tahmini yapılması ve klasik modellerin performansı ile karşılaştırılması, diğeri U-Net'e benzer modellerin performans karşılaştırması olmak üzere 3 boyutlu görüntüler üzerinde yapılan analizlerdir. Üçüncü çalışma ise ses verileri kullanılarak 2 farklı oto kodlayıcı yapısı ile glottal alanın tahmin edilmesidir.

4.1. U-Net Kullanılarak Görüntü Bölütleme İşleminin Klasik Yöntemlerle Karşılaştırılması

Bu çalışmada U-Net performansı Histogram, Aktif Kontur ve bölge büyütme ile karşılaştırılmıştır. Klasik yöntemler olan Histogram, Aktif Kontur ve bölge büyütme glottisin bölütlenmesi işlemi için ses tellerinin ilgi bölgesinin bulunmasını gerektirir. Derin öğrenme modeli olan ve oto kodlayıcı yapısı içeren U-Net ise direk olarak 256x256 boyutundaki 3 kanal RGB görüntü verileri üzerinde bölütleme işlemi yapabilir. Geleneksel yöntemlerin çalışma performansını artırmak için ilgi bölgesi manuel olarak tespit edilip veri setinin düzenlenmesi sağlanmıştır. Manuel olarak düzenlenen ilgi bölgelerinden bir örnek Şekil 4.1'de verilmiştir.

Bu uygulamada veri seti olarak ırcam veri seti kullanılmıştır. Klasik yöntem ile derin öğrenme tekniklerinin performanslarını karşılaştırmak için bazı problemler ile başa çıkmak gerekir. Klasik yöntemlerin eğitime ihtiyacı olmadığı için tüm veri seti üzerindeki performansı incelenebilir ancak makine öğrenmesi yaklaşımları tüm veriler üzerinde test edilemez. Verinin eğitim ve test veri seti şeklinde ikiye ayrılması gerekir. Bu nedenle veri setindeki 3000 adet görüntü rastgele olarak ikiye bölünmüştür. Tüm görüntüler üzerinde performans incelemesi yapılabilmesi için modelin hem test grubundaki hem de eğitim grubundaki verileri görmesi gerekir. Bunun için 2 farklı eğitim yapılmıştır. Bir eğitimde eğitim grubundaki veri seti ve test aşamasında test veri seti kullanılırken, diğer eğitimde test veri seti eğitim aşamasında ve eğitim veri seti

test aşamasında kullanılmıştır. Böylece tüm veri seti için tahmin performansı elde edilmiştir.



Şekil 4.1. Orijinal görüntü(sol) ve maskeli görüntü(sağ)

Bu uygulamada veri seti olarak ırcam veri seti kullanılmıştır.

Veri seti düzenleme işlemi için 10 video önce birleştirilip daha sonra boyutları ayarlanmıştır. Tüm veri setini içeren tensör 3000x256x256x3 boyutundadır. U-Net bu görüntülerin her birini 256x256x3 boyutunda eğitime sokar. Sonuç çıktı olarak da 256x256x1 boyutunda veriler döndürür. U-Net bu işlemi yaparken oto kodlayıcı kullanır. Oto kodlayıcı yapısında kodlayıcı ve kod çözücü bulundurur. Kodlayıcı tarafında görüntü boyutunu gittikçe küçülterek 32x32 boyutunda bir veri elde eder. Daha sonra kod çözücü bölgesinde her bir katmanda o katmana karşılık gelen kodlayıcı verisi de eklenerek evrişim ve üst örnekleme yapılır. Görüntü kod çözücü sonunda eski boyutuna gelmiş olur.

Bu çalışmada kullanılan U-Net modeli keras kütüphanesi yardımı ile oluşturulmuştur. U-Net mimarisinde 15 adet 2D ESA, 3 adet havuzlama katmanı, 3 adet üst örneklem katmanı, 3 birleştirme katmanı, 2 bırakma katmanına ek olarak giriş ve çıkış katmanlarıyla beraber toplamda 27 katman bulunmaktadır. Aktivasyon için relu fonksiyonu tercih edilmiştir. Son katmanda farklı olarak sigmoid fonksiyonu kullanılmıştır. Evrişim filtresinin boyutları 3x3 olarak

ayarlanmıştır. Son katmanda tek kanallı görüntü elde edebilmek için 1x1 boyutunda filtre tercih edilmiştir. Binary crossentropy ile kayıp skoru hesaplanmıştır. Performans ölçümünde doğruluk ölçütü kullanılmıştır. Modeli eniyilemek için adam tercih edilmiştir.

Sistem performansının incelenmesinde glottis açıklığının kaç piksel büyüklüğünde olduğu düşünülerek bir karşılaştırma yapılmıştır. Klasik modellerde de U-Net'te de modelin GA kapalı görüntülerde başarımı daha düşüktür. Bunun sebebi özellikle klasik modellerin daha büyük alanları daha kolay ayırt edebilecek şekilde formülize edilmiş olmasıdır. Dolayısıyla modellerin performansı GA büyüklüğünden bağımsız düşünülemez. Performansları GA' ya bağlı olarak incelemek için GA> 0, GA> 100, GA> 200 durumlarındaki skorlar ayrı ayrı hesaplanıp incelenmiştir. GA> 100 için görüntü içerisinde en az 101 piksel olan görüntüler tercih edilmiştir. Bu veri setindeki 3000 görüntüden 1814 tanesi GA> 100 için kullanılmıştır. GA> 200 için görüntü içerisinde en az 201 piksel bulunan görüntüler kullanılmıştır. Bu durumda da veri setindeki görüntülerden 1519 tanesi GA> 200 için kullanılmıştır.

Çizelge 4.1'de GA'nın büyüklüğüne bağlı olarak 4 modelin de HA, R ve D için performansları verilmiştir. GA>0 durumuna bakıldığında HA yönünden en yüksek başarım 0.867 ile U-Net modelinde elde edilirken, Aktif Kontur modeli kötü bir sonuç elde ederek 0.389'da kalmıştır. R ölçütü açısından en yüksek değeri 0.964 ile histogram elde etmişken en küçük değeri 0.684 ile bölge büyütme elde etmiştir. D ölçütü performans yönüyle ele alındığında U-Net 0.997 ile en yüksek başarıma sahiptir. En düşük başarım ise 0.717 ile aktif kontura aittir.

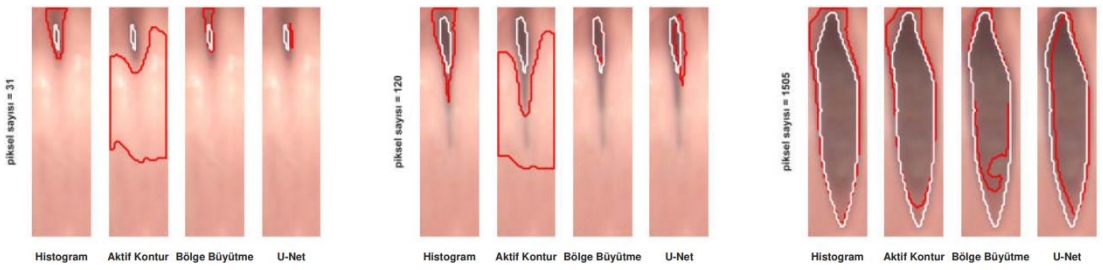
Çizelgede HA sütunları incelendiğinde U-Net GA'ya göre ciddi bir değişim göstermemiştir. Değişim 0.867'den 0.883'e çıkmıştır. Ancak klasik modeller GA büyüklüğü arttıkça daha fazla başarım elde etmişlerdir. Klasik modellerden aktif kontur ise başarımını 0.388'den 0.619'a yükselterek %59,5 oranında başarımını artırmıştır. R sütunları incelendiğinde ise aktif kontur başarımı neredeyse hiç değişmemiştir. Değişim sadece 0,03 puan olurken histogram, bölge büyütme ve U-Net için %1'in altında bir değişim söz konusudur. D sütunları incelendiğinde

histogram, bölge büyütme ve U-Net modellerinde değişim %1'in altında kalmıştır. Ancak aktif kontur 0.12 puan değerinde bir başarım artışı göstermiştir.

Çizelge 4.1. Histogram, aktif kontur, bölge büyütme ve U-Net'in GA büyüklüğüne göre HA, R ve D açısından performansları

	GA > 0			**GA >100**			**GA >200**		
	HA	**R**	**D**	**HA**	**R**	**D**	**HA**	**R**	**D**
Histogram	0.5572	0.9635	0.8514	0.5956	0.9635	0.8422	0.6277	0.9637	0.8400
Aktif Kontur	0.3885	0.9176	0.7172	0.5051	0.9278	0.7759	0.6188	0.9464	0.8325
Bölge Büyütme	0.7881	0.6843	0.9074	0.8487	0.6820	0.8996	0.8673	0.6820	0.8889
U-Net	0.8668	0.7786	0.9972	0.8701	0.7889	0.9965	0.8832	0.7987	0.9962

Şekil 4.2'de farklı GA büyüklüklerine sahip üç adet görüntü üzerinde dört modelin başarımları gösterilmiştir. Görüntüler GA içerisinde 31, 120, 1505 adet piksel bulundurmaktadır. GA küçük iken yani 31 ve 120 piksel içerdiği durumlar için aktif kontur ciddi derecede sorunlu tahminler yaparken GA'nın 1505 piksel içerdiği durumda tahmin başarımını çok yükseltmiştir. Tahmin başarımının histogram ve bölge büyütme içinde GA'ya bağlı olarak değiştiği gözlenmektedir.

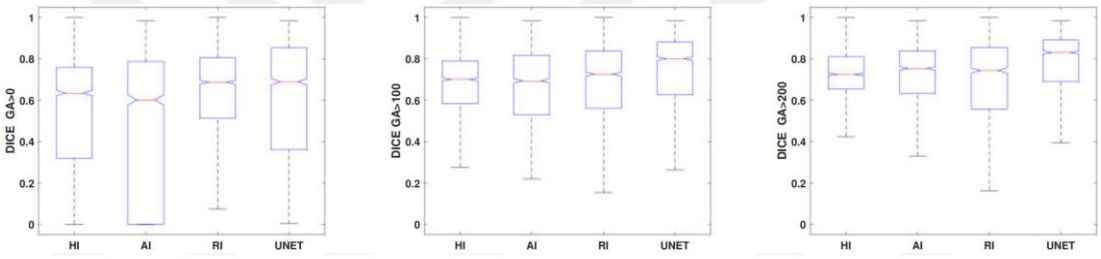


Şekil 4.2. GA'ya bağlı olarak dört modelin görüntü tahminleri

Bölütleme performansları GA büyüklüğüne bağlı olarak DCS ölçütü ile de incelenmiştir. Şekil 4.3'te gösterilen kutu diyagramları DCS skorlarının dağılımını göstermektedir. Kutu diyagramında GA>0 durumu için gözlem yapıldığında aktif konturun 0 ve 1 arasında geniş bir skalada dağıldığı görülmektedir. En dar aralıkta ise bölge büyütme bulunmaktadır. U-Net'in dağılımı daha geniş aralıkta olsa da medyan değeri bölge büyütmeden bir miktar daha büyüktür. Genel olarak baktığımızda ise tüm modellerin GA büyüdükçe daha dar aralıkta değişim gösterdiği görülmektedir. Aynı zamanda tüm modeller için başarımın arttığı da

göze çarpmaktadır. Aktif konturun dağılımında GA arttıkça ciddi derecede daralma gözlenmiştir. Şekilde HI histogram, RI bölge büyütme, AI aktif kontur ve U-Net de U-Net'i ifade etmektedir.

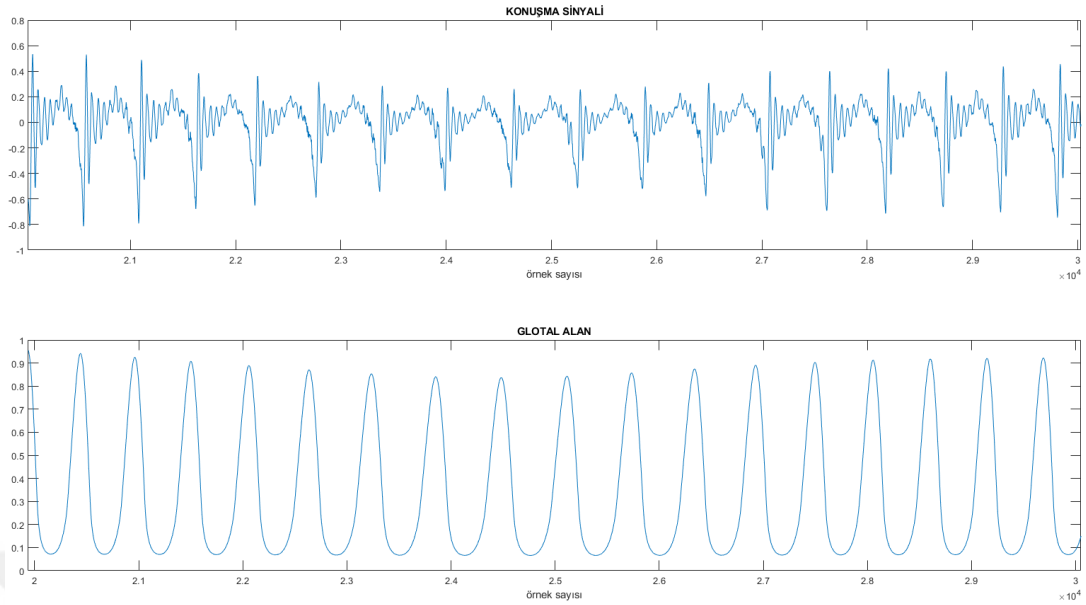
Tüm kutu diyagramlarına baktığımızda histogram, aktif kontur ve U-Net için GA'nın performansı etkileyen bir faktör olduğu gözlenmektedir. Ancak bölge büyütmede aynı etki gözlenmemiştir. Dağılım tüm GA durumları için neredeyse aynıdır. Diyagramlar üzerindeki DCS skorlarının medyanlarına bakıldığında histogram 0.63-0.70-0.72, aktif kontur 0.60-0.69-0.75 ve bölge büyütme 0.68-0.72-0.74 değerlerine sahiptir. U-Net modelinin ise 0.69-0.80-0.83 medyan değerleri ile aynı şekilde GA'ya göre değişim göstermiştir. GA'nın her durumunda en iyi başarımı sağlayan U-Net olmuştur.



Şekil 4.3. GA büyüklüğüne bağlı olarak DCS skorlarının kutu diyagramları

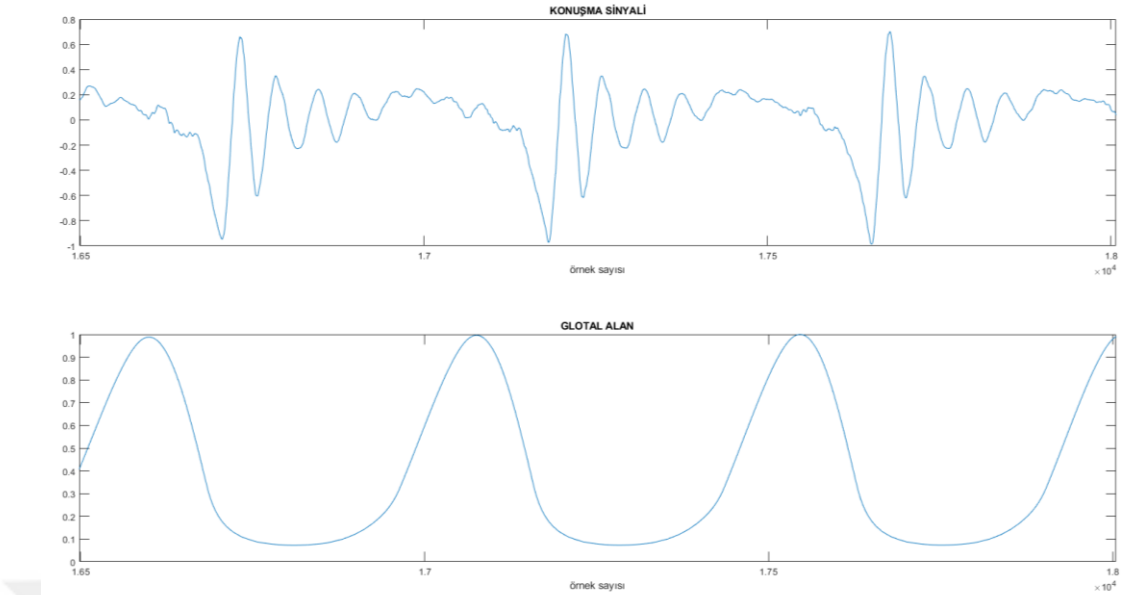
4.2. Konuşma Verisinden Glottal Alan Tahmini İçin Geliştirilen Modellerin Karşılaştırılması

Glottal alanın belirlenmesi çok önemli olmasına rağmen görüntü bölütleme için gereken HSV görüntülerin elde edilmesi çok maliyetlidir. Bu nedenle elde edilmesi çok daha kolay olan konuşma verisi kullanılarak GA tahmini yapabilmek süreci hem daha ucuz hem de daha kolay bir hale getirebilir. Bu çalışmada openglot veri setinden elde edilen veriler, iki farklı model üzerinde test edilerek performansları incelenecektir. Openglot veri setinden bir görüntü Şekil 4.4'te verilmiştir. Veri setinde hem konuşma hem de glottal alanın bulunması sayesinde bu veri seti hem eğitim hem de test için kullanılabilecektir. Şekil 4.4'te üstteki sinyal konuşma sinyalidir ve modelimizde girdi olarak kullanılmıştır. Altta ise GA bulunmaktadır. GA ise modelimizde çıktı olarak kullanılmıştır.



Şekil 4.4. Openglot veri setine ait bir sinyalin belli bir aralıktaki dalga formu

Veri setinde 96 adet sinyal bulunmaktadır. Veri setindeki veri miktarı az görünmesine rağmen her bir sinyal 44000 örnek içerdiği için yeterli miktarda örnek bulunduğu söylenebilir. 44000 örnek modelin aşırı uyumuna sebep olabileceğinden veri setinde bazı düzenlemeler yapılmalıdır. Veriler altıda birine düşürülerek alt örnekleme yapılmıştır. Bu alt örnekleme sonucunda her bir veri seti 7350 örnek bulundurur hale gelmiştir. Şekil 4.5'te üstteki sinyalin alt örneklenmiş hali verilmiştir. Bu alt örnekleme işleminden sonra verilerin modelde girdi olabilecek biçimde ayarlanması sağlanmalıdır. Bunun için sinyaller 256 örnek içeren daha küçük parçalara ayrılmıştır. Böylece modele girdi olarak verilecek veri seti boyutu nx256 şeklinde olmuştur. Eğitim, doğrulama ve test işlemleri için 5378x256 boyutunda bir veri seti elde edilmiştir. Buna göre 5378 adet örnek veri elde edilmiştir. Daha sonra veriler eğitim, test, doğrulama şeklinde üç parçaya ayrılmıştır. Verilerin %80'ine denk gelen 4305 adet sinyal eğitim seti, verilerin %10'una denk gelen 533 veri doğrulama ve kalan %10'luk kısımdaki 538 veri ise test verisi olarak ayrılmıştır.



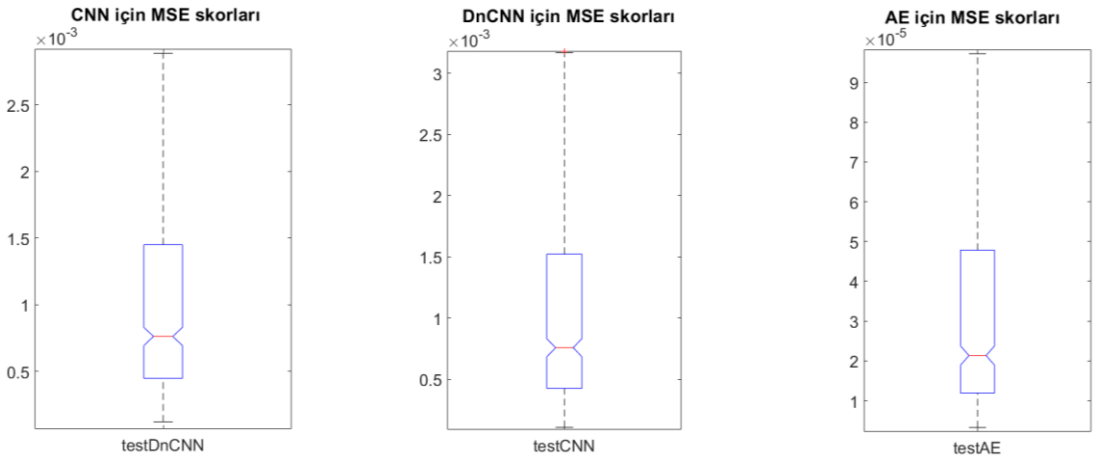
Şekil 4.5. Şekil 4.4'teki sinyalin alt örneklenmiş sinyali

Keras kütüphanesinin oluşturduğu oto kodlayıcı mimarisi ve Zhang vd., (2018) tarafından kullanılan gürültü yok edici bir oto kodlayıcı mimarisi ile üç farklı model oluşturulmuştur. Modellerde 2 boyutlu sinyale uygun olarak 1B evrişim katmanları kullanılmıştır. Gürültü yok eden mimariden 2 farklı model çıkarılmıştır. Bu modellerden biri gürültü yok eden katman içerirken diğeri içermemektedir. Gürültü yok eden katman içeren modelden gürültü yok eden evrişimli (DnCNN) ve Gürültü yok eden katman içermeyen modelden evrişimli (CNN) model olarak bahsedilmiştir. Üçüncü modelden ise otokodlayıcı (AE) modeli olarak bahsedilmiştir. Tahminlerin başarımları üzerine katman sayısı, filtre büyüklüğü ve çekirdek büyüklüğünün araştırılması için çeşitli eğitimler yapılmıştır. Python yazılımı ile yapılan eğitimlerde bir döngü şeklinde AE modeli için 180 eğitim ve diğer iki model için 100 farklı eğitim yapılmıştır. Her eğitim sonunda 3 ayrı veri seti için de MSE başarımları ölçülerek en iyi modeller seçilmiştir. En iyi modeller AE için 9 evrişim katmanı, filtre büyüklüğü 512 ve çekirdek büyüklüğü 3, gürültüsüz model için 5 evrişim katmanı, 512 filtre ve çekirdek büyüklüğü 12, gürültülü model için 5 evrişim katmanı, 512 filtre ve çekirdek büyüklüğü 12 şeklinde değerler alan modeller olmuştur. Seçilen modeller karşılaştırılırken yine MSE ölçütü kullanılmıştır. Seçilen modeller

doğrulama seti için MSE ölçütü ile bakıldığında AE, DnCNN ve CNN modeller sırasıyla 0.000324, 0.002701, 0.002831 sonuçlarını vermişlerdir.

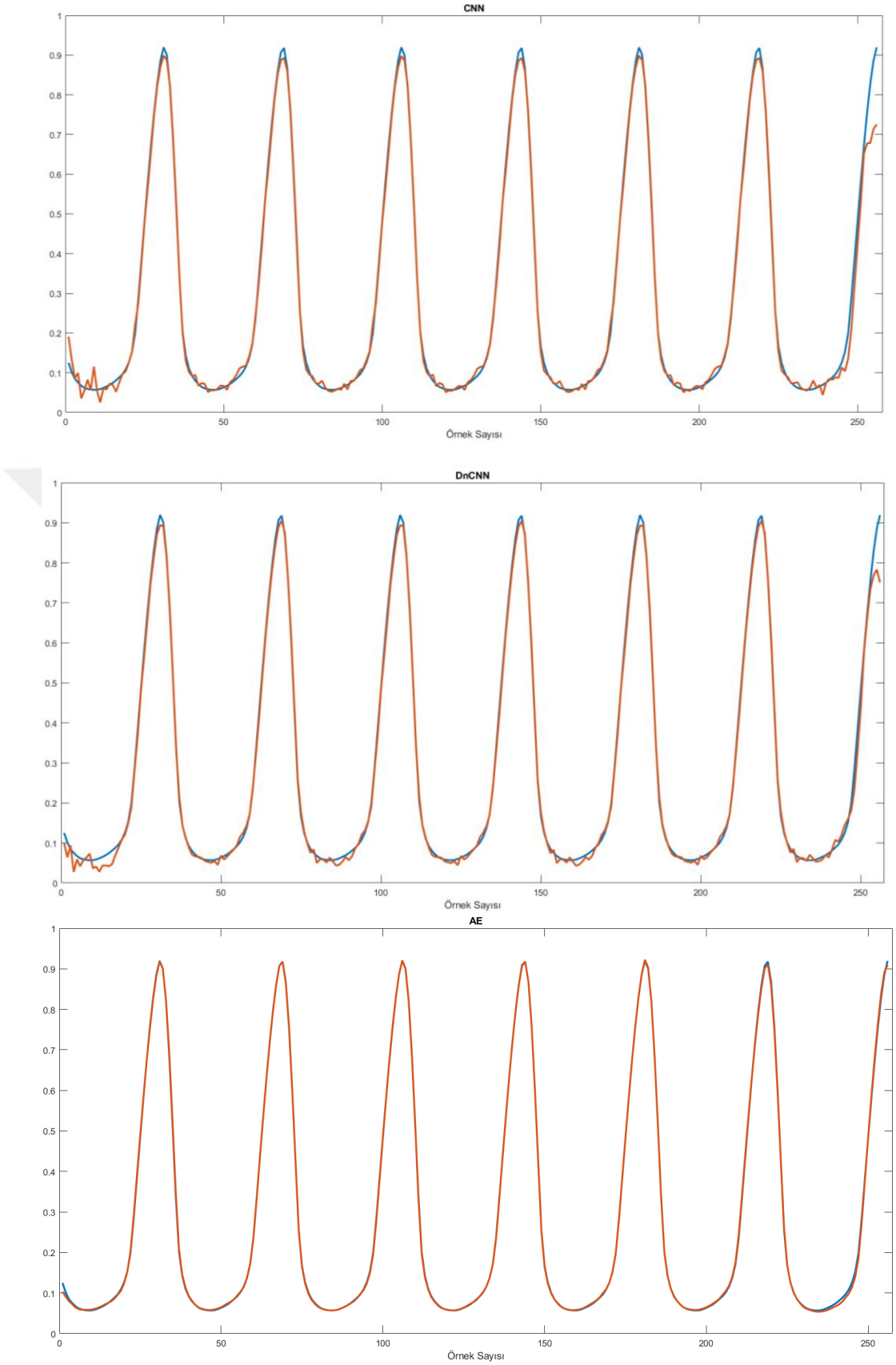
Modellerin performanslarının karşılaştırılması için test seti üzerindeki performansları incelenmiştir. Test setleri üzerindeki MSE başarımları AE, DnCNN ve CNN modeller için sırasıyla 0.000196, 0.0019063, 0.002085 şeklinde sonuçlar alınmıştır.

Modellerin tahmin sonuçlarının daha sağlıklı incelenebilmesi için kutu diyagramı oluşturulmuştur (Şekil 4.6). Veri setindeki 538 adet 256 boyutundaki vektörler için ayrı ayrı MSE skorları hesaplanmıştır. Şekilde en soldaki kutu diyagramı DnCNN modele aittir ve MSE medyanı 0.000762 olarak bulunmuştur. Ortadaki diyagram CNN modele aittir ve MSE medyanı 0.000758 bulunmuştur. Sağdaki diyagram ise AE'ye aittir ve MSE medyanı 0.0000213 olarak bulunmuştur. MSE sonuçlarına göre AE üstün bir başarı elde etmiştir ancak DnCNN ve CNN modelin karşılaştırılması mümkün görünmemektedir. Diyagramlara bakıldığı zaman gürültüsüz modelin biraz daha önde olduğu görülmektedir.



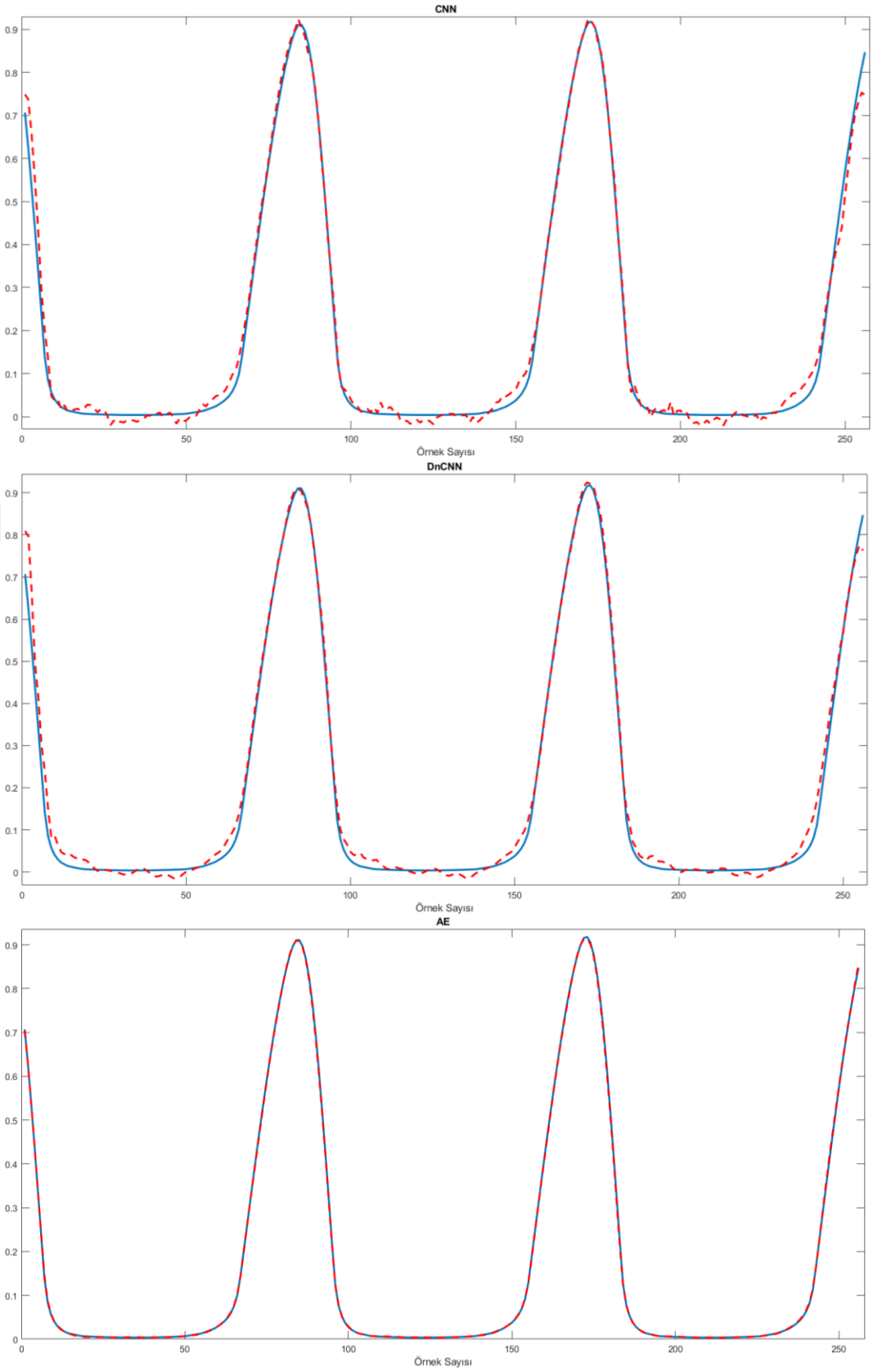
Şekil 4.6. Test verileri için MSE ölçüm sonuçlarına göre kutu diyagramları (solda DnCNN, ortada CNN ve sağda AE için mse skorları)

Modellerin test verileri üzerinde yaptığı tahminlerden iki örnek Şekil 4.7 ve Şekil 4.8'de verilmiştir. Grafiklerde mavi renkli sinyaller gerçek GA ve kırmızı renkli sinyaller ise tahmin edilen GA değerleridir. Şekil 4.7'de 538 adet test verisinden ilki için sonuçlar verilmiştir. Sonuçlara bakıldığında keras modelinin en iyi performansı sergilediği görülmekle beraber tüm modellerin başarım sağladığı söylenebilir.



Şekil 4.7. Modellerin test verisinin 1 numaralı verisi için GA tahminleri (mavi renk gerçek sinyal, kırmızı renk tahmin edilen sinyal)

Şekil 4.8’ e bakıldığında modellerin hiçbirinin gerçek GA’nın yanından geçemediği görülmektedir. Veri seti içinden hata payının yüksek olduğu 233 numaralı veri seçilmiştir. Bu veride modellerin MSE’leri CNN ve DnCNN için 0.006 ve AE için MSE 0.0005 olarak ölçülmüştür. Bu da MSE ölçütünün çok çok düşük olması gerekliliğini ortaya koymaktadır.



Şekil 4.8. Modellerin test verisinin 233 numaralı verisi için GA tahminleri (mavi gerçek sinyal, kırmızı tahmin edilen sinyal)

5. TARTIŞMA VE SONUÇLAR

Yapay zekanın hızlı gelişimi ile gelişen dijital sistemler sayesinde ses tellerinin yüksek hızlı görüntülerinin ses ve konuşma analizinde kullanılması mümkün hale gelmiştir. Ses tellerinin titreşimi ile yapılacak detaylı çalışmalarda HSV görüntülerinden glottis bölgesinin, görüntü bölütleme ile tespit edilmesi ve glottal alanın zamana bağlı değişiminin elde edilmesi ciddi önem taşımaktadır. Glottis bölütlemesinde, ilgi bölgesinin belirlenmesi bölütleme performansını direk olarak etkilemektedir.

Glottis bölütleme performansında klasik yöntemler tüm görüntü üzerinde çalışmakta zorlanmaktadır. Bu tezde, biyomedikal görüntü bölütleme alanında başarım elde etmiş olan U-Net derin öğrenme mimarisinden glottis bölütlemesi için özel bir model geliştirilmiştir. Daha sonra bu modelin başarımı hassasiyet, geri çağırma, doğruluk ve dice skoru açısından klasik modeller ile karşılaştırılmıştır. Geliştirilen U-Net modeli doğruluk, geri çağırma, hassasiyet ölçütleri açısından diğer modellere göre daha kararlı sonuçlar elde ederek doğrulukta 0.99, hassasiyette 0.88 ve geri çağırmada 0.78 skorlarını elde etmiştir. Bu başarımlara gör hem hassasiyet hem de doğruluk açısından diğer modellere üstün gelmeyi başarmıştır. Ayrıca dice skorları açısından karşılaştırma yapılmıştır. Dice katsayıları ile inceleme yapmak için kutu diyagramları kullanılmıştır. Bu incelemede tüm modellerin başarımlarını GA büyüdükçe artırdığı görülmüştür. Modellerden en başarılı olanı ise U-Net modeli olmuştur. U-Net performansının da GA büyüklüğünden etkilenebildiği anlaşılmıştır. Bahusus ses tellerinin temas etmesi ile glottis pikselinin sıfır olduğu anlarda aktif konturun başarımında ciddi düşüşlerin olduğu anlaşılmıştır.

Ses tellerinin titreşimini incelemek amacıyla HSV görüntüleri incelemek maliyetli bir iştir ve her zaman mümkün olmamaktadır. Bu nedenle bu tez çalışması kapsamında konuşma verisinden glottal alanın tahmini için üç farklı derin öğrenme modeli geliştirilerek performansları karşılaştırılmıştır.

Konuşma sinyalleri girdi olarak kullanılıp derin öğrenmede sıklıkla tercih edilen evrişim katmanları ile oluşturulan modellerin Glottal alanı tahmin etmesi sağlanmıştır. Tahmin işleminde zorluklardan biri veri büyüklüğünün modelin verimini ciddi oranda değiştirmesi olduğu için veriler önce alt örnekleme ile sonra da 256 vektörlük parçalara bölünmüştür. Yapılan işlem sayesinde modellerin veriyi yorumlaması daha kolay hale gelmiştir.

Modellerin daha iyi sonuç verebilmesi için katman sayısı, filtre büyüklüğü ve çekirdek büyüklüğü yönünden farklı durumlarla eğitilmesi sağlanmıştır. Bu eğitimler sonucunda modellerin performanslarının bu parametrelere büyük oranda bağlı olduğu görülmüştür. Yapılan eğitimler sonucunda en iyi modeller doğrulama seti performanslarına göre seçilerek test verisi üzerinde performanslara bakılmıştır. Performans ölçütü olarak MSE kullanılmıştır. Yapılan incelemeler sonucunda MSE'nin çok çok küçük çıkmadığı takdirde anlamlı sonuçların elde edilemeyeceği anlaşılmıştır. AE modelinin DnCNN ve CNN modellerden daha iyi olduğu ve daha kararlı çalıştığı gözlenmiştir. Alınan MSE sonuçları DnCNN, CNN ve AE modelleri için sırasıyla 0.000762, 0.000758, 0.0000213 şeklinde bulunmuştur. Sonuçlara bakılarak AE'nin yüksek performans gösterdiği anlaşılmaktadır.

KAYNAKLAR

Aggarwal, C. C. (2018). Neural networks and deep learning. Springer,.

Bellamkonda Praneeth, (2019). Deep Learning: Activation Functions Erişim Tarihi: 02.12.2022 https://ibelieveai.github.io/ActivationFunctions/#

Bishop, C. M. (2006). Pattern recognition and machine learning. springer.

Chollet, F. (2018). Deep Learning mit Python und Keras: Das Praxis-Handbuch vom Entwickler der Keras-Bibliothek. MITP-Verlags GmbH & Co. KG.

C. Guo, M. Szemenyei, Y. Yi, W. Wang, B. Chen and C. Fan, "SA-UNet: Spatial Attention U-Net for Retinal Vessel Segmentation," 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 2021, pp. 1236-1242, doi: 10.1109/ICPR48806.2021.9413346.

Degottex, G., & Bianco, E. (2010). IRCAM Databases of High Speed Videoendoscopy. UPMC-Ircam, France.

Derdiman, Y. S., & Koc, T. (2021, June). Deep Learning Model Development with U-net Architecture for Glottis Segmentation. In 2021 29th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.

Francois, C. (2017). Deep learning with Python.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning. Cambridge: MIT press.

Graves, A., Fernández, S., & Schmidhuber, J. (2005). Bidirectional LSTM networks for improved phoneme classification and recognition. In Artificial Neural Networks: Formal Models and Their Applications–ICANN 2005: 15th International Conference, Warsaw, Poland, September 11-15, 2005. Proceedings, Part II 15 (pp. 799-804). Springer Berlin Heidelberg.

Jha, D., Riegler, M. A., Johansen, D., Halvorsen, P., & Johansen, H. D. (2020, July). Doubleu-net: A deep convolutional neural network for medical image segmentation. In 2020 IEEE 33rd International symposium on computer-based medical systems (CBMS) (pp. 558-564). IEEE.

Islam, M. A. (2020). Reduced Dataset Neural Network Model for Manuscript Character Recognition.

Keras API reference Erişim Tarihi:01.09.2022 https://keras.io/api/

Kızrak Ayyüce, (2019). Derin Öğrenme İçin Aktivasyon Fonksiyonlarının Karşılaştırılması Erişim Tarihi:01.12.2022. https://ayyucekizrak.medium.com/derin-öğrenme-için-aktivasyon-fonksiyonlarının-karşılaştırılması-cee17fd1d9cd

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Koç, T., & Çiloğlu, T. (2014). Automatic segmentation of high speed video images of vocal folds. Journal of Applied Mathematics, 2014.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

Mehta, D.D.; Hillman, R.E. The Evolution of Methods for Imaging Vocal Fold Phonatory Function. SIG 5 Perspect. Speech Sci. Orofac. Disord. 2012, 22, 5–13.

Moccia, S., De Momi, E., Guarnaschelli, M., Savazzi, M., Laborai, A., Guastini, L., ... & Mattos, L. S. (2017). Confident texture-based laryngeal tissue classification for early stage diagnosis support. Journal of Medical Imaging, 4(3), 034502.

Nielsen, M. A. (2015). Neural networks and deep learning. San Francisco, CA: Determination press.

Pinheiro, A. P., Dajer, M. E., Hachiya, A., Montagnoli, A. N., & Tsuji, D. (2014). Graphical evaluation of vocal fold vibratory patterns by high-speed videolaryngoscopy. Journal of Voice, 28(1), 106-111.

Schenk, F., Aichinger, P., Roesner, I., & Urschler, M. (2015). Automatic high-speed video glottis segmentation using salient regions and 3D geodesic active contours. Annals of the British Machine Vision Association, 2015(1), 1-15.

Sharma, S. (2017). Activation functions in neural networks. Towards Data Science, 6.

Wang, J., & Jo, C. (2007, August). Vocal folds disorder detection using pattern recognition methods. In 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (pp. 3253-3256). IEEE.

Verikas, A., Gelzinis, A., Bacauskiene, M., & Uloza, V. (2006). Towards a computer- aided diagnosis system for vocal cord diseases. Artificial Intelligence in Medicine, 36(1), 71-84.

Verikas, A., Gelzinis, A., Bacauskiene, M., Hållander, M., Uloza, V., & Kaseta, M. (2010). Combining image, voice, and the patient's questionnaire data to

categorize laryngeal disorders. Artificial intelligence in medicine, 49(1), 43-50.

Rao, M. A., Krishnamurthy, R., Gopikishore, P., Priyadharshini, V., & Ghosh, P. K. (2018, January). Automatic Glottis Localization and Segmentation in Stroboscopic Videos Using Deep Neural Network. In INTERSPEECH (pp. 3007-3011).

Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

Yılmaz, A., Derdiman, Y. S., & KOÇ, Turgay. Ses Telleri Görüntülerinde Otomatik Piksel Tabanlı Sınıflandırma için Performans Ölçütlerinin İncelenmesi. Avrupa Bilim ve Teknoloji Dergisi, 103-110.

Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. Neural computation, 31(7), 1235-1270.