Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı, S. 103-110, Kasım 2020

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue, pp. 103-110, November 2020

Research Article

Ses Telleri Görüntülerinde Otomatik Piksel Tabanlı Sınıflandırma için Performans Ölçütlerinin İncelenmesi

Ayşenur Yılmaz[1*], Yaşar Said Derdiman[2], Turgay Koç[3]

[1] Süleyman Demirel Üniversitesi, Mühendislik Fakültesi, Elektrik-Elektronik Mühendisliği Bölümü, Isparta, Türkiye (ORCID: 0000-0003-2858-2412)
[2] Süleyman Demirel Üniversitesi, Mühendislik Fakültesi, Elektrik-Elektronik Mühendisliği Bölümü, Isparta, Türkiye (ORCID: 0000-0002-7266-0417)
[3] Süleyman Demirel Üniversitesi, Mühendislik Fakültesi, Elektrik-Elektronik Mühendisliği Bölümü, Isparta, Türkiye (ORCID: 0000-0002-4846-7772)

(International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) 2020 – 22-24 Ekim 2020)

(**DOI:** 10.31590/ejosat.819463)

ATIF/REFERENCE: Yılmaz, A., Derdiman, Y. S. & Koç, T. (2020). Ses Telleri Görüntülerinde Otomatik Piksel Tabanlı Sınıflandırma için Performans Ölçütlerinin İncelenmesi. *Avrupa Bilim ve Teknoloji Dergisi*, (Özel Sayı), 103-110.

Öz

Son yıllarda yapılan konuşma sistemi ile ilgili sorunların tespit edilmesinde ve konuşma analizinde gelişen teknolojinin getirdiği imkanlar sayesinde ses tellerinin yüksek hızlı görüntüleri yaygın olarak kullanılmaya başlanmıştır. Bu yüksek hızlı görüntüler konuşmacının ses tellerinin vibrasyonuyla ilgili detaylı bilgiler içerir. Fakat verinin büyüklüğü göz önüne alındığında bu görüntülerin manuel olarak işlenmesi mümkün görünmemektedir. Bu nedenle son yıllarda geliştirilen otomatik görüntü işleme algoritmaları ile ses telleri görüntülerinden glottis tespiti ve bölütlenmesi popüler hale gelmiştir. Bu çalışmada literatürdeki çalışmalardan farklı olarak ses telleri görüntülerinin piksel tabanlı otomatik sınıflandırılabilmesi için kullanılabilecek olan doğruluk, keskinlik (hassasiyet), geri çağırma, F1 skoru ve eşit hata oranı performans ölçütleri incelenmiştir. Bununla birlikte literatürdeki piksel tabanlı sınıflandırma modeli olan derin yapay sinir ağı temel sistem olarak alınarak yeni önerilen Gauss Karışım Modeli tabanlı sistem ile kıyaslanmıştır. Boyutları 256x256 olan manuel olarak bölütlenmiş 3000 adet yüksek hızlı endoskopik kamera görüntüsü rasgele olarak eğitim, geliştirme ve değerlendirme veri setlerini oluşturmak için kullanılmıştır. Veri seti ile eğitilen modellerin, geliştirme ve değerlendirme setleri ile yapılan çalışmalar sonucunda ikili sınıflandırmada yaygın olarak kullanılan doğruluk, keskinlik, geri çağırma ve F1 skoru ölçütlerinin modelden modele yaklaşık sadece %1 oranında değiştiği ve bu sonuçların sistem performansını yansıtma konusunda, aynı durumda % 22 değişim gösterebilen eşit hata oranı kadar etkili olmadığını göstermiştir. Bu çalışmanın sonucunda sistemlerin doğruluk değerleri aynı kalsa bile eşit hata oranı farkları değişebilmekte, bu nedenle aşırı uydurulmuş sistemlerin daha doğru kestirilebildiği gösterilmektedir. Temel sistem ile önerilen modeller karşılaştırıldığında, önerilen sistem 4096 karışımlı Gauss Karışım Modeli, kullanılan bütün performans ölçütleri için en iyi sonucu vermiş olup, değerlendirme setindeki eşit hata oranı için %22'lik bir performans iyileştirmesi göstermiştir.

Anahtar Kelimeler: Veri işleme ve tanıma, Konuşma işleme, Makine öğrenmesi.

Analysis of Performance Metrics for Automatic Pixel-Based Classification in Vocal Cord Images

Abstract

In recently years, thanks to the opportunities brought by the developing technology, high-speed images of the vocal cords have been started to widely use in detection of problems with the speech system and analysis of speech. These high-speed images contain detailed information about the vibration of the speaker's vocal cords. However, considering the size of the image data, it does not seem possible to manually process these images. For this reason, glottis detection and segmentation from vocal cord images has become popular with the development of automatic image processing algorithms in recent years. Unlike the other literature studies, in this study, the accuracy, precision (sensitivity), recall, F1-score and equal error rate performance criteria are examined used to automatically classify vocal cord images based on pixels. In addition to this, deep artificial neural network, that pixel classification based model in the literature, has been compared with the newly proposed model Gaussian Mixture Model. 3000 high speed endoscopic camera images manually segmented with dimensions 256x256 pixels were used to generate training, development and evaluation data sets of randomly. As a result of the studies conducted with the validation and evaluation sets of models trained with the data set, the accuracy, precision, recall and F1 score

[*] Sorumlu Yazar: Süleyman Demirel Üniversitesi, Mühendislik Fakültesi, Elektrik-Elektronik Mühendisliği Bölümü, Isparta, Türkiye ORCID: 0000-0003-2858-2412, aysenrylmaz357@gmail.com

criteria, which are commonly used in binary classification, changed only by 1% from model to model. And this result has shown that other performance metrics are not as effective as equal error rate that reflecting the system 22% change in the same situation. As a result of this study, even if the accuracy values of the systems remain the same, equal error rate differences may change, therefore it has been shown that overfitted systems can be predicted more accurately. Comparing the models proposed with the based system, the proposed system gave the best result for all performance criteria using the 4096 component Gaussian Mixture Model, and it is showed a performance improvement of 22% for the equal error rate in the evaluation set.

Keywords: Data processing and recognition, Speech processing, Machine learning.

1. Giriş

İnsanın en temel iletişim aracı olan konuşma üretim sisteminin analizi ve konuşmadaki bozuklukların incelenmesinde ses tellerindeki vibrasyonun değişiminin önemli bir yeri vardır. Ses tellerinin vibrasyonundaki bozukluklar, fonksiyonel disfoni, parezi (ses telleri kısmi felci), vokal polip ve nodüller, gırtlak kanseri ve larenjit (ses telleri iltihabı) gibi hastalıklardan kaynaklanabilir.[1-3] 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. Ayrıca bu görüntüler konuşma analizi içinde kaynak olarak kullanılmaktadır. Ses tellerinin vibrasyonunu incelemek için yapılan çalışmalarda son yıllarda yüksek hızlı endoskopik kameralarla alınan görüntüler kullanılmaktadır.[4-6] Bu görüntüler bir hastanın ses tellerinin titreşiminin bir periyodunun yakalanmasını ve ses telleri titreşimlerinin ölçümünü sağlar.[7] Ses telleri normal konuşma esnasında 70-400 Hz frekans aralığında titreşirler. Klinik ses analizinde kullanılan HSV sistemleri bir periyot içerisinde ses tellerinin frekansına göre her bir periyotta en az 10 kare görüntüleme yapabilir ve özellikle düşük frekansta olan erkek ses tellerinin bir periyodunu detaylı yakalayabilmek için yeterli hıza sahiptir.[8,9]

Genellikle HSV görüntüleri incelenirken ilk aşamada ilgi bölgesi (İB) belirlenir, ardından İB içerisinde bölütleme işlemi gerçekleştirilir. [7] İB belirlenmesinde literatürde ilk başvurulan yöntem manuel yapılması, el ile seçilmesidir. [10-12] Diğer bir yöntem ise başlangıç bölgesinin manuel olarak seçilip, sonraki analizler için bölgenin otomatik adapte edilmesidir. Bu yöntemlerde İB otomatik izlenmesi için hareket kestirimi, kök noktaları, toplam yoğunluk değişimi gibi algoritmalar kullanmaktadır. [13-14] İB'nin otomatik olarak belirlenmesi için yapılan çalışmalarda ise toplam yoğunluk değişimi, hareket kestirimi gibi yöntemlerinin yanında son zamanda derin yapay sinir ağları tabanlı yöntemler kullanılmaya başlanmıştır. [15-18] Ses tellerinin vibrasyonu sırasında ses tellerinin olduğu bölgedeki yoğunluk değerlerinin zamandaki değişimi genellikle büyük olmaktadır. Bu değişimin mutlak olarak ölçülüp ortalaması alınarak İB bölgesinde yüksek değerler elde edip daha sonra elde edilen bu iki boyutlu değişim haritası veya diğer ismiyle Toplam Değişim Resmi (Total Variation Image) otomatik bulunan eşik değerler ile ikili resme dönüştürülerek ilgi bölgesinin kestirimi gerçekleştirilebilmektedir. [19] Fakat bu yöntem, ses tellerinin sürekli olarak vibrasyon yaptığı durumlarda kullanılabilir. Ses tellerinin hareket etmediği durumlarda İB'yi belirleyebilmek için makine öğrenmesi ve derin öğrenme yöntemleri kullanılabilir. Makine öğrenmesi kullanılarak yapılan bir yöntemde İB' nin belirlenmesi için glottisteki yerel renk ve şekil bilgisi kullanılmıştır. Eğitim, tanıma ve bölütleme modüllerinden oluşan bu sistemde 60 farklı glottis şekli manuel olarak tespit edilmiş ve bir dizi tanımlayıcılar hesaplanarak glottis bölgesi tespit edilmesinde kullanılmıştır. Elde edilen bölge glottisin sonraki framelerde İB içinde kalması için otomatik izlenmiştir. Bu yöntemin çok sayıda blok içermesi ve glottis için farklı şekiller belirlemede kullanıcı üzerindeki yükünden dolayı karmaşık ve getirdiği iş yükü yüksektir. [20] Daha farklı olarak yakın zamanda glottis lokasyonunu otomatik belirlemek ve bölütlemek için yapılan çalışmada RGB renk kanallarını öznitelik çıkarımında kullanarak, çok katmanlı derin yapay sinir ağı eğitip ilgi bölgesi bir sınıflandırma problemi olarak ele alınmıştır. Glottal bölge renginin ve çevresinin glottal bölgenin dışından farklı bir desene sahip olduğu bilgisi kullanılarak HSV görüntüleri 3 × 3' lük parçalara ayrılıp her bir parçanın orta noktasındaki pikselin glottis olup olmamasına göre bir etiket bilgisi oluşturulup RGB kanallarından gelen her biri 9x1'lik vektörlere dönüştürülmüş 3x3' lük görüntü verisinden 27x1'lik öznitelik vektörü oluşturularak 4 katman 128 nöronlu bir DNN (Deep Neural Network) modeli Doğruluk (Accuracy) metriği esas alınarak eğitilmiş ve %65 performans elde edilmiştir.[15]

Bu çalışmanın amacı piksel tabanlı sınıflandırma ile HSV görüntü işlemede Doğruluk (Accuracy) yanında Eşit Hata Oranı (Equal Error Rate), Hassasiyet (Precision), Geri Çağırma (Recall) ve F1- Skoru ölçütlerinin lokalizasyon performansı açısından kullanımını incelemek ve DNN ile önerilen Gauss Karışım Modeli (Gaussian Mixture Model-GMM)' ni kıyaslamaktır. Bu makale dört bölüm şeklinde düzenlenmiştir. Bölüm 2' de çalışmanın yöntemi, oluşturulan öznitelikler, kullanılan sınıflandırma modelleri ve performans ölçütleri hakkında detaylı bilgiler verilecektir. Bölüm 3' de, veri setinin hazırlanması ve yapay sinir ağı modellerinin eğitim aşamasından bahsedilecektir. Bölüm 4 'de eğitim, geliştirme ve değerlendirme aşamalarından elde edilen performans ölçütleri baz alınarak yapay sinir ağı modelleri ve Gauss karışım modeli kıyaslaması yapılacaktır. Bölüm 5' de ise çalışma sonuçlarına göre bir yapay sinir ağı modelinin performans başarısında esas alınması gereken performans ölçütü üzerinde durulacaktır.

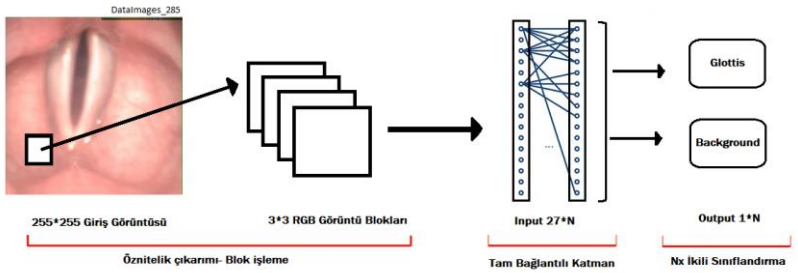
2. Yöntem

Bu çalışmada HSV görüntü işleme için piksel sınıflandırma tabanlı bir sistem kullanılmıştır. Sistemin ilk aşamasında HSV görüntülerinden öznitelik çıkarılmış, ardından HSV görüntüleri öznitelikleri kullanılarak glottis veya arka plan olacak biçimde ikili sınıflandırılmıştır.

2.1. Öznitelik Çıkarımı

Bu çalışmada kullanılan öznitelikler RGB renk kanallarından elde edilmiştir. [15] 'de yapılan çalışmada olduğu gibi HSV görüntüleri her bir 3x3'lük parçalardan elde edilmiş 1x9'luk RGB bileşenlerinin birleştirilerek 1x27'lik öznitelik vektörü oluşturulmuştur. Birbirleriyle örtüşmeyen görüntü üzerindeki 3x3'lük bütün parçalar Nx27'lik bir veri seti oluşturacak biçimde birleştirilir. Öznitelik çıkarımı Şekil 1'de sol tarafta gösterilmiştir. Veri setinde kullanılan bütün pikseller için 3x3'lük parçanın orta

noktasındaki pikselin etiketi glottis ise 1 arka plan ise 0 olacak biçimde oluşturulmuştur. Bu öznitelik ve etiket bilgisi DNN ve GKM modellerinin eğitimi ve değerlendirilmesi için kullanılmıştır.



Şekil. 1 Önerilen çerçeve. Glottis lokalizasyonu için izlenen adımların gösterimi: ön işleme, bloklama, eğitim aşaması, test

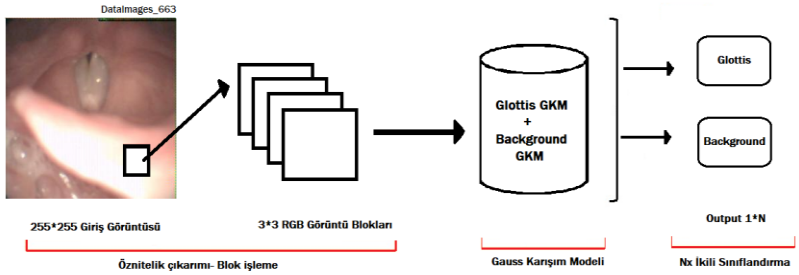
2.2. Sınıflandırma Modelleri

2.2.1. Yapay Sinir Ağı Modelleri

Bu çalışmada [15]' de Rao vd., tarafından kullanılmış 4 katmanlı-128 nöronlu-relu aktivasyonlu ve sigmoid çıkış fonksiyonlu derin yapay sinir ağı modeli (DNN) temel sistem olarak seçilmiştir. Temel sistem 27 boyutlu öznitelik vektörünü giriş olarak alıp çıkışta ilgili pikselin glottis mi arka plan mı olduğuna dair 0 ile 1 arasında skor üretmektedir. Temel sistemin diğer kendi alt kümesi olan 1-2-3 gizli katmanlı, 32-64-128-256 nöronlu ve aktivasyon için relu, selu, tanh kullanan [21] toplamda temel sistem ile birlikte 37 sinir ağı modeli kullanılmıştır. Bu modeller içerisinde eğitim ve geliştirme setlerinde başarımı en yüksek olanlar 1 gizli katman-128 nöron-selu modeli temel DNN sistem ile beraber çalışmanın devamında ölçütler ile birlikte analiz için kullanılmıştır. Şekil 1'de öznitelik vektörüyle yapay sinir ağının etkileşimi gösterilmektedir.

2.2.2. Gauss Karışım Modeli

HSV görüntülerinde glottis ve arka plandan elde edilen öznitelik vektörlerinin olasılık dağılımı Gauss karışımı olarak düşünülebilir. Bu yaklaşımla öznitelik vektörleri sınıflarına göre ayrılarak glottis ve arka plan için ayrı ayrı 1024, 2048 ve 4096 karışım bileşenine sahip GKM modelleri kullanılmıştır. GKM' ler için gerekli olan ortalama vektörü ve kovaryans matrisleri EM(Expectation Maximisation) algoritması ile eğitim seti kullanılarak belirlenir.[22] Elde edilen GKM modellerinin çıktıları geliştirme ve değerlendirme setleri ile detaylı performans analizi için kullanılır. Önerilen GKM modeli Şekil 2' de gösterilmektedir.



Şekil. 2 Önerilen Gauss Karışım Modeli ile HSV sınıflandırma sistemi

2.3. Performans Ölçütleri

Bu çalışmada sistemlerin performanslarını kıyaslamak amacıyla ikili sınıflandırma problemlerinde sıklıkla kullanılan ölçütlerden Doğruluk (Accuracy-ACC), Hassasiyet (Precision-PR), Geri Çağırma (Recall- R), F1- Skoru ve Eşit Hata Oranı (Equal Error Rate-EER) kullanılmıştır. Bu ölçütlerin hesaplanmasında ikili sınıflandırma sistemi sonucunda meydana gelen şu durumlar kullanılır.

Doğru Pozitif (True Positive-TP) : Sınıflandırma sonucunda referans pozitif piksellerin kaç tanesinin doğru bulunduğunu belirtir. Bu çalışmada pozitif piksellerin glottise ait olduğu kabul edilmiştir.

Doğru Negatif (True Negative-TN) : Sınıflandırma sonucunda referans arka plana ait piksellerin kaç tanesinin doğru bulunduğunu belirtir.

Yanlış Pozitif (False Positive-FP) : Sınıflandırma sonucunda referans arka plan piksellerin kaç tanesinin glottis olarak bulunduğunu belirtir.

Yanlış Negatif (False Negative-FN) : Sınıflandırma sonucunda referans pozitif piksellerin kaç tanesinin arka plan olarak bulunduğunu belirtir.

Referanstaki toplam glottis piksel sayısı TP+FN, arka plan piksel sayısı TN+FP olarak ifade edilebilir.

2.3.1. Doğruluk

Doğruluk arka plan ve glottis piksellerinin ne kadarının doğru sınıflandırıldığını gösterir ve şu şekilde hesaplanır:

$$ACC=\frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

ACC 0 ile 1 aralığında değer alır. ACC 1'e yaklaştıkça doğru sınıflandırılmış piksel sayısı artar. ACC ölçütü verisetinin dengeli dağılımı durumunda sistem performansını daha iyi yansıtmaktadır. Fakat dengesiz veri setinde yani sınıf örnek sayıları arasında büyük farklar olması durumunda ACC ölçütü bölütleme performansını doğrudan temsil edemeyebilir. Örneğin HSV görüntülerinde genellikle glottis bölgesi arka plana göre çok daha küçüktür. Eğer glottis bölgesi resmin onda birini oluşturursa ve sistem glottis piksellerinin hepsini yanlış tahmin ederse ACC 0.9 gibi çok büyük bir değer alabilir. Bu gibi durumlarda hassasiyet ve geri çağırma ölçütleri bölütleme performansını daha doğru yansıtmaktadır.

2.3.2. Hassasiyet

Bu çalışmada glottis olarak sınıflandırılan piksellerin gerçekte glottis olma ihtimalini gösterir. Aşağıdaki gibi denklem ile ifade edilir:

$$PR=\frac{TP}{TP+FP} \quad (2)$$

PR 'nin artması arka plan piksellerinin glottis olarak sınıflandırılmasını önler, FP 'nin azalmasını sağlar. Bu nedenle düşük PR' a sahip HSV görüntülerinde arka plan üzerinde sahte glottis bölgeleri belirebilir. Daha yüksek performans için PR' nin daha büyük olması istenir.

2.3.3. Geri Çağırma

Bu çalışmada glottis olarak sınıflandırılan piksellerin ne kadarının doğru olarak tahmin edildiğini gösterir. Aşağıdaki denklem ile ifade edilir:

$$R=\frac{TP}{TP+FN} \quad (3)$$

Geri çağırma glottis bölgesi içinde yanlış tahmin edilen glottis piksellerinin büyüklüğü hakkında bilgi verir. R oranının 1 olması glottisin tamamen doğru olarak sınıflandırıldığını gösterir ancak PR düşük ise glottisin haricinde farklı yerlerde sahte glottis bölgeleri oluşur. Bu nedenle PR ve R birbirinden bağımsız değerlendirilemez. İdeal bir sistem için PR ve R 1 olmalıdır.

2.3.4. F1 Skoru

F1 skoru, **PR** ve R'nin etkisini birleştirmek amacıyla, PR ve R 'nin harmonik ortalaması olarak tanımlanmıştır. F1-skor ölçütü, Denklem (4) ile ifade edilir.

$$F1=2\frac{PR.R}{PR+R} \quad (4)$$

PR ve R değerlerinden biri çok küçük değil ise F1 bölütleme performansıyla ilgili önemli bilgi verir. F1 değeri 0-1 aralığında olup 1' e yaklaştıkça performansın arttığını gösterir.

2.3.5. Eşit Hata Oranı

Eşit hata oranı (EER) sistem çıkışında glottis için yapılan hata olasılığı, sahte glottis piksel olasılığı (Pfa) ile arka plan için yapılan hata oranının (PM) birbirine eşit olması durumunu ifade eder. Bu amaçla glottis ve arka plan pikselleri için sistemlerin verdikleri skorlar bir eşik değer belirlenerek değerlendirilir. Glottis için yapılan hata olasılığı $Pfa(\boldsymbol{\theta}) = \mathbf{1}$ - $R(\boldsymbol{\theta})$ ile arka plan için yapılan hata olasılığı $PM(\boldsymbol{\theta}) = FP(\boldsymbol{\theta}) / (FP(\boldsymbol{\theta}) + TN(\boldsymbol{\theta}))$' nin birbirine eşit olduğu eşik değeri, $\boldsymbol{\theta}$, optimizasyon ile belirlenir. Burada elde edilen $\boldsymbol{\theta}$ değeri HSV görüntüsünün ikili forma dönüştürülmesinde kullanılır. EER hatalı glottis olasılığını belirlemektedir. Hatalı glottis olasılığı EER küçüldüğünde sıfıra yaklaşır dolayısıyla PR 1'e yaklaşır. Aynı şekilde hatalı arka plan olasılığı da sıfıra yaklaştığı için R' de 1'e yaklaşır. Bu nedenle en küçük EER'ye sahip sistem daha başarılı performans gösterebilir.

3. Yapılan Çalışmalar

3.1 Veri Setinin Hazırlanması

Bu çalışmada IRCAM HSV görüntü veritabanı kullanılmıştır. [23] Veri tabanında 256x256 boyutunda 10 farklı videodan her birinden eşit olmak üzere 3000 görüntü seçilmiş ve geliştirilen uygulama ile manuel olarak bölütlenmiştir[19]. Veri tabanı 3x3 'lük RGB parçalarına ayrılabilmesi için genellikle arka plan olan 1. satır ve 1. sütuna ait RGB değerleri çıkarılarak 255x255 'lik boyuta indirilmiştir. Her kanala ait 255x255'lik görüntüler 3x3'lük parçalara ayrıldıktan sonra her bir resim için 7225x27 boyutunda öznitelik verisi elde edilmiştir. Aynı anda referans bölütlenmiş görüntü kullanılarak 7225x1'lik etiket vektörü oluşturulmuştur. HSV görüntülerinde glottis büyüklüğü arka plana göre genellikle çok küçük olduğu için veri seti dengeli değildir. Bu nedenle eldeki veri seti dengeli eğitim ve geliştirme seti yapılabilmesi için veri tabanından 70000 glottis piksele ait öznitelik vektörü ile 70000 arka plan piksele ait öznitelik vektörü rasgele seçilerek eğitim seti oluşturuldu. Benzer biçimde 35000 glottis için 35000 arka plan için rasgele seçilen vektörler geliştirme seti için kullanıldı. Kalan diğer verilerden rasgele seçilen daha dengesiz 1 milyon piksele ait veri değerlendirme setini oluşturmak için kullanıldı.

3.2 Yapay Sinir Ağı Modellerinin Eğitilmesi

Eğitim için hazırlanan veri seti derin ve sığ yapay sinir ağı modellerinin eğitiminde kullanılmıştır. Derin ve yapay öğrenme modellerinin eğitiminde keras- tensorflow kütüphanesi kullanılmıştır.[24] Eğitilen modellerin derlenmesinde kaybı küçültmek için ikili etiketli verisetlerinde tercih edilen ikili çapraz entropi (binary_crossentropy) fonksiyonu tercih edilmiştir. Sinir ağı modelleri geri yayılım algoritması ile eğitildi. Optimizasyon algoritması olarak öğrenme oranı 1e-08 olarak ayarlanan "ADAM" kullanılmıştır.[25] Eğitim ve geliştirme süreçleri boyunca izlenecek ölçüt olarak ACC kullanılmıştır. Eğitim esnasında modele verilen her 32 verilik yığın için tahmin edilen değer ile gerçek değer arasındaki fark hesaplanarak modelin ağırlıkları öğrenilmiştir. Eğitim verisetinin eğitimi tamamlanınca ağırlıklar modelin daha önce hiç görmediği geliştirme veriseti üzerinde test edilerek aşırı uydurma olup olmadığı incelenmiştir. Modeller aşırı uydurma olmayacak biçimde seçilmiştir. Sistem çıkışı $\boldsymbol{\theta}$ ile eşiklenerek glottis için 1 arka plan için 0 olacak biçimde belirlenmiştir.

3.3 GKM Modellerinin Eğitilmesi

Gauss karışım modellerinin eğitiminde, eğitim setinde bulunan glottis ve arka plan örnekleri ayrı ayrı kullanılmıştır. Glottis öznitelikleri ile glottise ait bir GKM, arka plan öznitelikleri ile arka plana ait bir GKM, EM algoritması ile eğitilmiştir. GKM'deki karışım sayısı olarak 1024, 2048 ve 4096 seçilmiştir. Seçilen bu Gauss dağılımların kovaryans matrisi için tam, küresel ve köşegen matris ayrı ayrı seçilmiştir. Eğitilen bu sistemlerin verdikleri olasılık değerlerinin logaritması alınmış ve bunların farkı alınarak log ihtimal oranı (log likelihood ratio-LLR) hesaplanmıştır. Sistem çıkışı LLR> $\boldsymbol{\theta}$ için 1 diğer durumlar için ise 0 olarak karar verilmiştir. Sistemin performansı $\boldsymbol{\theta}$ ile değişmektedir. Bu çalışmada EER'yi sağlayan $\boldsymbol{\theta}$ değeri kullanılmıştır.

4. Çıktılar ve Tartışma

Yapay sinir ağı tabanlı modeller eğitim seti ile eğitilmiş ve geliştirme seti üzerinde doğruluklarına bakılarak performansları ölçülmüştür. Eğitilen 37 modelden en başarılı modeller olan temel sistem 4-128-relu DNN modeli ve 1-128-selu olan SNN sığ modeli ileri analiz için seçilmiştir.

GKM tabanlı sistemler eğitim seti ile eğitilmiş bu sistemlerden tam kovaryans matrisine sahip olanlar, aşırı uydurma sebebiyle düşük performans göstermiştir. Bu nedenle ileri analiz için sadece küresel kovaryans GKMs ve köşegen kovaryans GKMd modelleri seçilmiştir.

Eğitim, geliştirme ve değerlendirme setleri ile seçilen GKM, DNN ve SNN modellerine giriş olarak verildi. Sistem çıktıları EER' yi sağlayan $\boldsymbol{\theta}$ değerini belirlemek için analiz edildi. Belirlenen $\boldsymbol{\theta}$ değerinde veri setleri sınıflandırılarak EER, PR, R, F1 ve ACC ölçütleri hesaplanarak sonuçlar elde edildi. Eğitim, geliştirme ve değelendirme setleri üzerindeki performans ölçütlerine ait değerler sırasıyla Tablo 1, Tablo 2 ve Tablo 3'te gösterilmiştir.

Tablo 1 incelendiğinde modeller eğitim seti üzerinde 0.95 ile 0.98 aralığında doğruluk değerine sahiptir. Benzer biçimde PR, R, F1 ölçütleri de aynı şekilde oluşmuştur. Burada bu ölçütler arasında yüzde yüz bir korelasyon görülmektedir. Diğer taraftan sistemlerin EER değerleri daha farklı bir biçimde oluşmuş ve DNN için en yüksek değer olan 4.92, GKMd-4096 için en düşük değer olan 1.60 değeri gözlenmiştir. Tablo 1'deki sonuçlara göre PR, R, F1 ve ACC ölçütlerinden sadece bir tanesini performans incelemek için kullanmak yeterlidir. Ancak dikkat edilirse sistem performanslarındaki değişim bu ölçütlerin her birinde aynı olup minimum düzeydedir. Elde edilen 0.95-0.98 değer aralığı model seçimi için 1.60 ile 4.92 arasındaki EER ile kıyaslandığında önemli derecede küçüktür. EER glottis ve arka plan için sistemin eşit hata yaptığı durumu ifade eder. Bu nedenle EER küçüldükçe ACC, PR, R ve F1 değerleri doğru orantılı olarak yükselecektir. Bu sebepten EER diğer ölçütlerinde bilgilerini taşımaktadır. En küçük EER' ye sahip sistem daha başarılı performans gösterir.

Tablo 1 - Eğitim Seti Performansı					
YÖNTEM	% EER	Kesinlik (PR)	Geri Çağırma (R)	F1 Skoru	Doğruluk (ACC)
SNN	4.55	0.95	0.95	0.95	0.95
DNN	4.92	0.95	0.95	0.95	0.95
GKMs-1024	4.08	0.96	0.96	0.96	0.96
GKMd-1024	3.58	0.96	0.96	0.96	0.96
GKMs- 2048	3.16	0.97	0.97	0.97	0.97
GKMd-2048	2.53	0.97	0.97	0.97	0.97
GKMs-4096	2.09	0.98	0.98	0.98	0.98
GKMd-4096	1.60	0.98	0.98	0.98	0.98

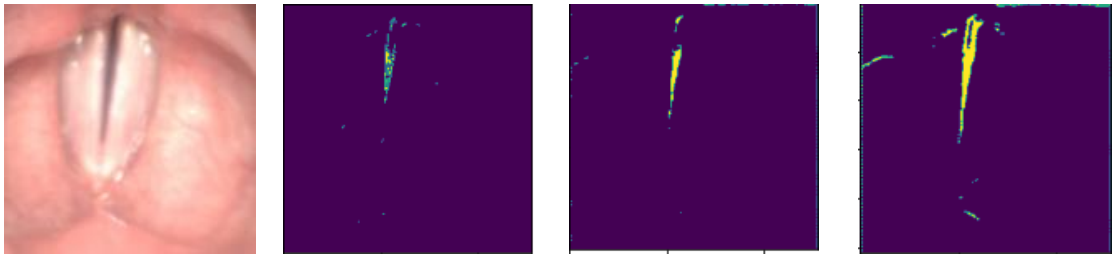
Sistemlerin geliştirme seti performansı Tablo 2' de gösterilmiştir. Geliştirme seti eğitim setine göre daha küçük olup, sistemlerin performanslarını aşırı uydurma için izleme amacıyla kullanılmıştır. Geliştirme setindeki sistem performansları Tablo 1' de verilen eğitim seti performanslarıyla benzerlik içermektedir. ACC, F1, R ve PR'nin her model için aynı değere eşit olup 0.95 veya 0.96 olarak değişmektedir. EER değerleri ise en yüksek 4.98 ile DNN, en düşük 3.83 ile GKMs-4096 olarak bulunmuştur. Eğitim setinde olduğu gibi geliştirme setinde de EER daha büyük dinamik aralığa sahiptir. ACC, F1, R, PR gibi ölçütlere göre sistem performanslarında en fazla %1'lik bir değişim gözlenirken EER'de bu fark %23 civarındadır.

Tablo 2 - Geliştirme Seti Performansı					
YÖNTEM	% EER	Kesinlik (PR)	Geri Çağırma (R)	F1 Skoru	Doğruluk (ACC)
SNN	4.56	0.95	0.95	0.95	0.95
DNN	4.98	0.95	0.95	0.95	0.95
GKMs-1024	4.92	0.95	0.95	0.95	0.95
GKMd-1024	4.65	0.95	0.95	0.95	0.95
GKMs- 2048	4.48	0.96	0.96	0.96	0.96
GKMd-2048	4.20	0.96	0.96	0.96	0.96
GKMs-4096	3.83	0.96	0.96	0.96	0.96
GKMd-4096	4.11	0.96	0.96	0.96	0.96

Modellerin eğitim aşamasında hiç görmediği değerlendirme seti üzerindeki performans değerleri Tablo-3'te verilmiştir. Değerlendirme seti üzerindeki değişim eğitim ve geliştirme setlerine göre ACC ve R için benzerlik gösterirken PR ve F1 için farklılık göstermektedir. ACC ve R diğer setlerde olduğu gibi 0.95- 0.96 değerlerine sahiptir. Bu R değeri glottisin %95' inin doğru sınıflandırıldığını göstermektedir. Diğer yandan PR değeri daha önceki durumların aksine daha geniş bir dinamik aralıkta değer almakta olup sistem performansını yansıtmakta ACC ve R'ye göre daha belirleyicidir. Örneğin aynı R değerine sahip ilk dört model içinde SNN 0.45 PR değeriyle bu modeller arasında en az yanlış pozitif (FP) üreten sistem olup EER' si 4.55 ile en küçük olanıdır. DNN ve SNN, eğitim ve geliştirme setlerinde aynı ACC, PR ve F1'e sahip iken SNN daha düşük EER' ye sahiptir. Veri sayısının daha çok olduğu değerlendirme setinde PR değeri açısından SNN lehine %2' lik bir fark oluşmuştur. Bu fark ayrıca EER' yede yansıyarak %10' luk bir EER azalması gözlenmiştir. Diğer setlerde de %10'luk bu fark açıkça görülebilmektedir. Veri sayısı arttığı için EER' deki fark PR sonuçlarına yansımıştır. EER ve PR arasındaki ilişki ile uyumluluk gözlemlenmektedir. Yapay sinir ağı eğitiminde kullanılan ACC ölçütüne göre SNN ve DNN aynı performansı göstermesine rağmen EER'ye göre SNN'nin daha başarılı olması modelin katman sayısı arttırılarak her zaman daha iyi sonuç elde edilemeyeceği karmaşıklık sebebiyle aşırı uydurmaya sebep olabileceği görülmektedir.

Tablo 3 - Değerlendirme Seti Performansı					
YÖNTEM	% EER	Kesinlik (PR)	Geri Çağırma (R)	F1 Skoru	Doğruluk (ACC)
SNN	4.55	0.45	0.95	0.61	0.95
DNN	4.98	0.43	0.95	0.59	0.95
GKMs-1024	4.96	0.43	0.95	0.59	0.95
GKMd-1024	4.66	0.44	0.95	0.61	0.95
GKMs- 2048	4.46	0.46	0.96	0.62	0.96
GKMd-2048	4.11	0.48	0.96	0.64	0.96
GKMs-4096	3.87	0.49	0.96	0.65	0.96
GKMd-4096	4.04	0.48	0.96	0.64	0.96

GKM modellerinin performansı yakından incelendiğinde eğitim seti üzerinde karışım sayısı 1024'den 2048'e ve 2048'den 4096' ya arttırıldığında EER değerleri yaklaşık olarak 1 puan azalmıştır. Diğer yandan geliştirme ve değerlendirme setlerinde GKMs aynı karışım sayısı 1024'den 2048'e ve 2048'den 4096' ya geçerken 0.5'lik bir iyileşme gözlenirken GKMd' de 1024'den 2048'e geçerken yaklaşık 0.5, 2048'den 4096'ya geçerken yaklaşık 0.1'lik bir iyileşme görülmüştür. Buradan GKMd'nin 4096 karışıma çıktığında aşırı uydurmanın meydana gelmeye başladığı gözlenebilir. Bütün modeller birlikte değerlendirildiğinde GKMs-4096 en düşük EER, 3.87, en yüksek PR, 0.49, en yüksek F1, 0.65' e sahip olup temel sistem DNN' e göre %22 daha az EER ile en iyi performansı gösteren sistem olarak görülmektedir. Ölçütlerin dinamik aralıkları dikkate alındığında geliştirme setinde olduğu gibi değerlendirme setinde de %22 dinamik aralığı ile EER ölçütü sistem performanslarını diğer ölçütlere göre daha iyi yansıtmaktadır.



Şekil 3 HSV Piksel Sınıflandırma Sonuçları

Soldan sağa doğru Orijinal, PR 0.62 - R 0.21, PR 0.41 - R 0.38, PR 0.31 R 0.72,

Bir HSV karesinin sınıflandırmasında 0.99 ACC değerine sahip üç farklı sınıflandırma sonucu Şekil 3'te gösterilmektedir. Görüntülerde sol taraftan sağa doğru gidilirken PR azalıp, R artmaktadır. En solda bulunan orijinal görüntü farklı eşik değerler kullanılarak sınıflandırıldığında sırasıyla PR 0.62 - R 0.21, PR 0.41 - R 0.38 ve en sağda görünen PR 0.31 R 0.72 olan sonuç elde edilmiştir. ACC değerleri çok yüksek ve aynı olmasına karşın sınıflandırma sonuçlarında büyük farklılık görülmektedir. PR 0.62-R-0.21 durumunda yüksek PR, düşük R sebebiyle arka planda hatalı glottis bölgeleri oluşmamış ancak gerçek glottis bölgesinin % 79'u tespit edilememiştir. PR 0.41 – R 0.38 olduğu durumda glottis önceki duruma göre daha iyi tespit edilebilmiş ancak % 62 civarında kayıp görülmüştür. Son olarak PR 0.31 R 0.72 için glottis bölgesindeki kayıp, keskinliğin biraz daha azalmasıyla % 38'e düşmüş ancak arka planda hatalı glottis pikselleri oluşmuştur. ACC bu üç durumu birbirinden ayırt edemediği için yeterli bilgiyi sağlayamamaktadır. En doğru sonuç için PR ve R'nin 1 e yaklaşması gerekmektedir.

5. Sonuç

Bu çalışmada öncelikle ses tellerinin yüksek hızlı görüntülerinin piksel sınıflandırma yöntemi ile işlenmesi için olası performans ölçütleri araştırılmıştır. Bununla birlikte literatürde bulunan bir derin yapay sinir ağı (DNN) modeli temel alınarak, diğer yapay sinir ağı modelleri ile kıyaslanmış, ayrıca bu sınıflandırma problemi için Gauss Karışım Modeli (GKM) tabanlı bir sistem önerilmiştir. Performans metrikleri olarak literatürde yaygın olarak kullanılan doğruluk, keskinlik (hassasiyet), geri çağırma, F1 skoru gibi ölçütler ile birlikte eşit hata oranı metriği bu konuda ilk defa kullanılmıştır. Yapılan çalışmanın sonucunda doğruluk, keskinlik, geri çağırma ve F1 ölçütleri eğitim ve geliştirme setlerinde birbirinin benzeri skorlar üretmiş modelden modele yaklaşık %1'lik bir değişim göstermiştir. Daha büyük miktarda dengesiz veri bulunan değerlendirme setinde ise sadece keskinlik ve F1 skorunda değişim gözlenmiştir. Eşit hata oranı ise bütün veri setlerinde ortalamada %22'nin üzerinde dinamik aralığa sahip olmuş olup, sistem performanslarının ölçülmesinde daha detaylı bilgi sağlamıştır. Yapay sinir ağlarıyla yapılan çalışma sonucunda doğruluğa dayalı olarak eğitim yapılması durumunda benzer performans göstermelerine karşın sistem karmaşıklığı arttıkça eşit hata oranının arttığı gözlenmiştir. Bu nedenle eğitim için doğruluk ölçütü kullanmak hatalara sebep olmuştur. Genel olarak önerilen Gauss Karışım Modeli, piksel sınıflandırmasında temel sisteme göre yaklaşık olarak %22'lik daha iyi bir performans göstermiştir.

6. Teşekkür

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 ederiz.

Kaynakça

[1] Cen, Q., Pan, Z., Li, Y., & Ding, H. (2019, January). Laryngeal Tumor Detection in Endoscopic Images Based on Convolutional Neural Network. In *2019 IEEE 2nd International Conference on Electronic Information and Communication Technology (ICEICT)* (pp. 604-608). IEEE.

[2] Turkmen, H. I., Karsligil, M. E., & Kocak, I. (2015). Classification of laryngeal disorders based on shape and vascular defects of vocal folds. Computers in biology and medicine, 62, 76-85.

[3] Aubreville, M., Knipfer, C., Oetter, N., Jaremenko, C., Rodner, E., Denzler, J., ... & Maier, A. (2017). Automatic classification of cancerous tissue in laserendomicroscopy images of the oral cavity using deep learning. Scientific reports, 7(1), 1-10.

[4] Drioli, C., & Foresti, G. L. (2020). Fitting a biomechanical model of the folds to high-speed video data through bayesian estimation. Informatics in Medicine Unlocked, 20, 100373.

[5] Khairuddin, K. A. M., Ahmad, K., Ibrahim, H. M., & Yan, Y. (2020). Description of the Features and Vibratory Behaviors of the Nyquist Plot Analyzed From Laryngeal High-Speed Videoendoscopy Images. Journal of Voice.

[6] Fehling, M. K., Grosch, F., Schuster, M. E., Schick, B., & Lohscheller, J. (2020). Fully automatic segmentation of glottis and vocal folds in endoscopic laryngeal high-speed videos using a deep Convolutional LSTM Network. Plos one, 15(2), e0227791.

[7] Andrade-Miranda, G., Stylianou, Y., Deliyski, D. D., Godino-Llorente, J. I., & Henrich Bernardoni, N. (2020). Laryngeal Image Processing of Vocal Folds Motion. Applied Sciences, 10(5), 1556.

[8] Deliyski, D. D., Powell, M. E., Zacharias, S. R., Gerlach, T. T., & de Alarcon, A. (2015). Experimental investigation on minimum frame rate requirements of high-speed videoendoscopy for clinical voice assessment. Biomedical Signal Processing and Control, 17, 21-28.

[9] Ogutcen, M. Y. Koc, T., (2019). Yüksek Hızlı Ses Telleri Görüntülerinin Düzlemsel Aydınlatma Modeli ile Aktif Kontur Tabanlı Segmentasyonu, EEMKON 2019, Elektrik Elektronik Mühendisliği Kongresi, p.427-431.

[10] Yan, Y., Chen, X., & Bless, D. (2006). Automatic tracing of vocal-fold motion from high-speed digital images. IEEE Transactions on Biomedical Engineering, 53(7), 1394-1400.

[11] Zhang, Y., Bieging, E., Tsui, H., & Jiang, J. J. (2010). Efficient and effective extraction of vocal fold vibratory patterns from high-speed digital imaging. Journal of Voice, 24(1), 21-29.

[12] Yan, Y., Du, G., Zhu, C., & Marriott, G. (2012, March). Snake based automatic tracing of vocal-fold motion from high-speed digital images. In 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 593-596). IEEE.

[13] Andrade-Miranda, G., & Godino-Llorente, J. I. (2017). Glottal Gap tracking by a continuous background modeling using inpainting. Medical & Biological Engineering & Computing, 55(12), 2123-2141.

[14] 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.

[15] 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).

[16] 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.

[17] Kopczynski, B., Strumillo, P., Just, M., & Niebudek-Bogusz, E. (2018, November). Acoustic Based Method for Automatic Segmentation of Images of Objects in Periodic Motion: Detection of vocal folds edges case study. In 2018 Eighth International Conference on Image Processing Theory, Tools and Applications (IPTA) (pp. 1-6). IEEE.

[18] Hamad, A., Haney, M., Lever, T. E., & Bunyak, F. (2019). Automated Segmentation of the Vocal Folds in Laryngeal Endoscopy Videos Using Deep Convolutional Regression Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 0-0).

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

[20] Gloger, O., Lehnert, B., Schrade, A., & Völzke, H. (2014). Fully automated glottis segmentation in endoscopic videos using local color and shape features of glottal regions. IEEE Transactions on Biomedical Engineering, 62(3), 795-806.

[21] Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning (Vol. 1, p. 2). Cambridge: MIT press.

[22] Kasapoğlu, B., & Turgay, K. O. Ç. (2020) Sentetik ve Dönüştürülmüş Konuşmaların Tespitinde Genlik ve Faz Tabanlı Spektral Özniteliklerin Kullanılması. Avrupa Bilim ve Teknoloji Dergisi, 398-406.

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

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

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