ACTA INFOLOGICA 2019;3(2):60-82
dergipark.org.tr/acin

İSTANBUL UNIVERSITY PRESS

DOI: 10.26650/acin.590690

REVIEW ARTICLE

Çoklu Ortam Sistemleri İçin Siber Güvenlik Kapsamında Derin Öğrenme Kullanarak Ses Sahne ve Olaylarının Tespiti

Sound Scene and Events Detection using Deep Learning in the Scope of Cyber Security for Multimedia Systems

Bahadır Karasulu[1]

ÖZ

Günümüzde doğadaki birçok doğal ses kaynağı yanısıra sentetik sesler de çoklu ortam sistemlerinde kullanılmaktadır. Bu seslerin bulunduğu ortamlar (sahneler) biyometrik yetkilendirme, güvenlik isterleri ve gürbüz/güvenli sesli/görüntülü iletişim için önem arz etmektedir. Konuşma/konuşmacı tanıma, doğrulama gibi özel kısıtlara sahip ses biçemleri haricinde çoklu seslerin ayrıştırılması, gürültü giderilmesi, ses sahnesi/olaylarının tespiti ve ses etiketleme işlemleri siber güvenlik açısından daha güvenli bilişim sistemleri oluşturulması adına gün geçtikçe önem kazanmaktadır. Derin öğrenme katmanlı altyapısı gereği oldukça iyi bir biçimde ham verideki özniteliklerin ve anlamsal ilişkinin elde edilmesine olanak sunmasından dolayı son yılllarda siber güvenlik alanında da tercih edilir olmuştur. Bu çalışmada siber güvenlik kapsamında çoklu ortam verisi olarak ses (veya konuşma) analizi ve sınıflandırma/tahminleme ve tespit için derin öğrenme mimari modellerinin kullanımı irdelenmiştir. Çalışmamızda 2015 ilâ 2019 yılları arasındaki yayınlarda öne çıkan modeller olan derin sinir ağları, evrişimli sinir ağları, tekrarlayıcı sinir ağları, kısıtlanmış Boltzmann makinesi ve derin inanç ağları sistematik olarak incelenmiştir. Böylece siber güvenlikte ses/konuşma işleme, sesle aldatmayı önleme, tutarlı ve yüksek başarımlı sonuçları elde etmeye dair literatürdeki yönelim kırkı aşkın çalışma üzerinden bilimsel bulgulara dayanan tartışma ve yorumlarla açıkça ortaya konulmaktadır.

Anahtar kelimeler: Derin Öğrenme, Konuşmacı Tanıma, Ses Olayı Tespiti, Siber Güvenlik

[1]Çanakkale Onsekiz Mart Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Çanakkale, Türkiye

ORCID: B.K. 0000-0001-8524-874X

Corresponding author:
Bahadır Karasulu,
Çanakkale Onsekiz Mart Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Çanakkale, Türkiye
Telephone: +90 286 218 00 18 / 2356
E-mail address: bahadirkarasulu@comu.edu.tr

Submitted: 11.07.2019
Revision Requested: 14.11.2019
Last Revision Received: 26.11.2019
Accepted: 02.12.2019

Citation: Karasulu, B. (2019). Sound scene and events detection using deep learning in the scope of cyber security for multimedia systems. *Acta Infologica, 3*(2), 60-82. https://doi.org/10.26650/acin.590690

ABSTRACT

In addition to many natural sound sources in nature, synthetic sounds are also used in the multimedia systems of our modern world. Environments (i.e., sound scenes) with these sounds are important for biometric authorization, security requirements and robust/safer voice/video communication. Apart from audio formats that have special constraints such as speech/speaker recognition and verification, the separation of polyphonic sounds, noise reduction, detection of sound scenes/events and voice tagging processes are gaining importance in order to create safer information systems in terms of cyber security. In recent years deep learning has been preferred in the field of cyber security due to its layered infrastructure, which enables the easy extraction of attributes and semantic relationships in the raw data. In this study, the use of deep learning architecture models for voice (or speech) analysis and classification/prediction and detection as multimedia data in cyber security coverage is examined. In our study, deep neural networks, convolutional neural networks, recurrent neural networks, restricted Boltzmann machine and deep belief networks are systematically reviewed as prominent models in the publications between 2015 and 2019. Therefore, the orientation in the literature on voice/speech processing in cyber security, prevention of voice spoofing, and achieving consistent and high performance results is clearly demonstrated through discussions and comments based on scientific findings over fourty studies.

Keywords: Deep Learning, Speaker Recognition, Sound Event Detection, Cyber Security.

1. GİRİŞ

Bilişim sistemlerinin bir türü olarak çoklu ortam sistemleri günümüzde gittikçe daha fazla veri ile işlem yapılması nedeniyle oldukça önem kazanmaktadır. Özellikle çoklu ortam veri tipleri olan ses, görüntü, vidyo ve metin bileşenleri kullanılarak yapılan içerik tabanlı bilgi elde etme (content-based information retrieval) sayesinde literatürdeki çoğu çalışmada insan ile makine etkileşimindeki anlamsal çözümlemede makinenin (bilgisayar-tabanlı bilişim sistemi) sensör bilgisiyle anlamsal bilgi arasındaki oluşan boşluk (semantic gap) böylece doldurulmaya çalışılmaktadır. Bir bilgi elde etme süreci için ele alınan çoklu ortam bileşenindeki ayırt edici özniteliklerin elde edilmesiyle (feature extraction) bilişim sisteminde kullanılır biçime dönüştürülmesi anlamsal bir yargının oluşarak nihai kararın verilebilmesine dayanak oluşturmaktadır. Çoklu ortam sistemlerinde endeksleme, özetleme (abstraction) ve sorgu sonucuyla eşleştirme (matching) gibi olguların daha az maliyetle yapılabilmesi için özniteliklerin iyi ve eksiksiz bir biçimde tanımlanması gerekir. Veri işleme yoluyla düşük seviyeden öznitelikler (daha somut) ve orta seviyeden öznitelikler kullanılarak üst seviye özniteliklere dayanan (daha soyut) anlamlar çıkarmak mümkündür. Veri işlemeyi destekleyici öznitelik seçme ve azaltma gibi bazı ön işlemlerden sonra bunlar sıklıkla bir makine öğrenmesi veya derin öğrenme modeli şeklindeki sınıflandırıcı/tahminleyici altyapıya geçirilerek genel sistemin daha fazla soyutlamaya dayanacak üst seviyeden nihai kararlarına zemin oluşturmaktadır (Goodfellow, Bengio ve Courville, 2016).

Siber güvenlik (Cyber security) yıllardır bilişim sistemleri alanında çalışılan önemli bir araştırma alanıdır. Siber olaylar veya saldırılar, sıklıkla bilgisayar tabanlı sistemlere sızma yoluyla bilgi elde etme, onları etkisiz hale getirme amaçlı kullanılır gibi algılansa da Nesnelerin Interneti (Internet of Things, IoT) ve büyük verinin gündeme gelmesiyle Endüstri 4.0 alanında tehdit unsuru olabilecek diğer olgular da ön plana çıkmaya başlamıştır. Yapay zekâya veya makine öğrenmesine dayanan sistemlerin (akıllı sistemler olarak da adlandırılır) gelişmesiyle hem donanım hem yazılım alanında siber olay çeşitliliği artmıştır. Özellikle pahalı donanımların içerildiği sistemlerde donanımların işlevsiz hale getirilmesi önemli bir tehdit unsurudur. Bu tip saldırılar yoluyla bir hizmetin teslimatı geciktirilebilmekte, değiştirilebilmekte ve bertaraf edilebilmektedir. Veri analizinde en uygun modellerden birisi de derin öğrenmedir. Ses/konuşma tanıma gibi çoklu ortam verilerini taban alan işlemlerde, öznitelik öğrenmede ve çevreye/verilen probleme özgü çözümlere uyum sağlayabilen sistemler derin öğrenme yoluyla oluşturulabilmektedir. Ayrıca sahtecilik (fraud) ve suistimal (abuse) yapılan çeşitli durumlarda derin öğrenme daha etkin çözüme yönelik sonuçlar verebilmektedir. Özellikle anlamsal (semantic) veriler sayesinde anlamsal kurallarla çıkarsama (inference/reasoning) yapan akıllı sistemlerin kullanımı gün geçtikçe artmaktadır (Sağıroğlu ve Koç, 2017). Kritik öneme haiz sistemler olarak ülke/şehir enerji altyapısı, su dağıtım şebekesi, iletişim altyapısı gibi sistemler hem bilişim öğelerini hem de fiziksel öğeleri birlikte içermekte olması ve siber saldırılara karşı çokça zayıf noktaları bulunduğundan oldukça fazla saldırıya maruz kalabilmektedirler. Böyle durumlarda oluşan hasarın maliyeti, bilgi kaybının yaşatacağı zarar ve diğer olumsuzluklar nedeniyle bu tip sistemlerin iyi bir biçimde korunması da gerekmektedir (Muratoğlu, Okul, Aydın ve Bilge, 2018). Bu nedenlerle siber riskler analiz edilmeli ve buna karşın önlemler alınmalıdır. Özellikle büyük veri (big data), akıllı veri (smart data) ve anlamsal veri (semantic data) kullanabilen vidyo gözetleme (video surveillance), vidyo veritabanı (database) ve sesli/görüntülü kimlik doğrulama/geçiş (authentication) ile ilgili hizmet sunan bilişim sistemleri böyle çoklu ortam verilerini işlemekte ve sıklıkla bu tip siber olay/saldırılara açık bir şekilde bulunduklarından çeşitli saldırılara maruz kalmaktadırlar (Sağıroğlu ve Koç, 2017).

Ses, doğada insan algısını oluşturan temel bileşenlerdendir. Sesin yönü (doğrultusu), şiddeti, süresi ve çevresel olayları tarif etmemizde sunduğu zengin bilgi hacmi nedeniyle insan tarafından fiziksel olarak gözlemlenemeyen çeşitli olayları bile yorumlamamızda (soyutlama) ve anlamsal çıkarımlarda bulunmamızda büyük rol oynar. Ses olayları, birbirlerinden ayırt edilebilmesine olanak sunan bazı belirgin özniteliklere sahip çeşitli bireysel veya çok sesli (polyphonic) olan ve genellikle harmonik sinyallerden meydana gelir. Havada yayılan sinyal insan kulağı tarafından alınarak beyinde sinirsel belirli işlemlerden geçirilmekte ve yorumlanmaktadır. Bu mekanizmaya benzer bir yoldan ses işleme ve ses olayları tespit etme yaklaşımları yıllar boyunca çeşitli çalışmalarda geliştirilmiştir. Bir ses olayı, belirli bir fiziksel ses kaynağı tarafından üretilen belirli bir ses olarak tarif edilir. Örneğin; kuş tarafından ötme sesinin oluşturulması, araba tarafından motor sesinin oluşturulması gibi sesler sayılabilir. Bu tarz olaylar iyi bir biçimde tanımlanmış ve görece kısa süreli olmaktadır. Birçok olayın veya sadece bir tane ses olayının farklı durumlarını içerebilen bir ses sahnesi, çeşitli ses kaynaklarından birlikte

oluşmuş bir bütün halindeki ses kompozisyonudur. Bu sahneyi iç mekân ve dış mekân olmak üzere iki farklı ortam tipine ayırabiliriz. Buna göre iç mekânda örneğin radyoda müzik çalarken o esnada çamaşır makinesi sesi de işitilebilmektedir. Dış mekânda ise örneğin insanlar sohbet ederken arkaplanda bir arabanın geçmekte olduğu işitilebilmektedir. Ses analizi yapılırken ses sinyali çeşitli aşamalardan geçirilmektedir. Sınıflandırma aşamasında ses sinyali içeren ses kayıtları önceden belirlenmiş bir kategori etiketleri kümesiyle karşılaştırılmak üzere sınıflandırma işlemine tabi tutulur. Ses olayı tespitinde ise, belirli bir sesin veya seslerin zaman içerisinde hangi zamansal noktada oluştuğunun bulunmaya çalışıldığı bir işlem gerçekleştirilir. Buradaki her bir ses örneğinin ya oluştuğu anda tespit edilmesine ya da sesin aktif olduğu her bir zamansal yerleşimin bulunmasına çalışılmaktadır. Bu yolla, iki ayrı ses kaydının aynı ses sahnesine ait olup olmadığı veya farklı ses sahnelerinden alınıp alınmadıkları da tespit edilebilmektedir (Virtanen, Plumbley ve Ellis, 2018).

Ses işleme ve analizi sayesinde çoklu ortam sistemleri için hesaplamalı ses analizi yoluyla otomatik konuşma tanıma (automatic speech recognition, ASR), ses içerik analizi (sound/audio content analysis) ve müzikten bilgi elde etme (music information retrieval, MIR) kapsamında birçok çalışma literatürde bulunmaktadır. Makine öğrenmesi, derin öğrenme ve istatistiksel yaklaşımlar sayesinde sınıflandırma ve analiz aşamaları ile ses olay tespiti, ses bölütleme (segmentation) ve çevresel ses algılama (environmental sound perception) gibi konular yüksek doğrulukta bir başarımla gerçekleştirilebilmektedir. Derin öğrenme ham verilerden öznitelikleri oldukça iyi bir biçimde öğrenmektedir. Bunu katmanlı sinir ağı yapısına borçludur. Ayrıca makine sensör bilgisi ile insanın algıladığı bilgi arasındaki anlamsal boşluğu (semantic gap) doldurmaya yönelik olarak düşük seviyeli öznitelikler ile yüksek seviyeli öznitelikler arasındaki ilişkilerin ortaya konulmasına katkı sunarak nihai karar aşamasının başarımının artırılmasına olanak sunmaktadır (Goodfellow ve ark., 2016; Patterson ve Gibson, 2016).

Bu makalede ses/konuşma analizi, sınıflandırması ve ses olayı tespitinde derin öğrenme mimarilerini ve makine öğrenmesi yöntemlerini/tekniklerini kullanan literatürdeki çalışmalar incelenerek kıyaslama yoluyla farklılıkları ve benzer yönleri sistematik biçimde ortaya konulmaktadır. Siber olaylarda da sesin/konuşmanın analiz edilmesini gerektiren durumlara sıklıkla rastlanılmaktadır. Bu nedenle önleyici tedbirler kapsamında ses/konuşma analizi ve ses olayı tespiti yoluyla siber saldırılara maruz kalmadan veya sese yönelik bir siber olay/saldırı başlangıcını takiben en kısa sürede buna tepki verilmesi adına makine öğrenmesi/derin öğrenme kullanımı bu makalede tartışılmaktadır. Bu kapsamda bu makalede sese yönelik siber saldırıların tespitinde sesin işlenmesine dair dikkat edilecek temel hususlar irdelenmiştir.

Bu makale dört bölümden oluşmaktadır. İlk bölümde ses sahnesi, ses olayı ve ses olay tespiti hakkında temel bazı bilgiler verilerek ses olay tespiti yapılmasının çoklu ortam sistemleri için ses analizi temelinde böyle bir analizin neden önemli olabileceği üzerinde durulmaktadır. İkinci bölümde ses verisinin temsilinin nasıl olduğu, ses özniteliklerinin neler olabileceği ve nasıl elde edilip ses sınıflandırma ve ses olayı tespitinde kullanıldığına ilişkin temel bilgilere yer verilmektedir. Ayrıca aynı bölümde derin öğrenme mimari modelleri hakkında bilgiler de bulunmaktadır. Üçüncü bölümde kapsamlı bir literatür incelemesi içerisinde makine öğrenmesi, derin öğrenme gibi konularda yapılan çalışmaların detaylarına bakılmakta ve bu çalışmaların öne çıkan yanları üzerinden birbirlerine göre farklılık ve benzerliklerine değinilmektedir. Bu yolla elde edilen bilimsel bulgular ışığında ses olay tespitinin siber güvenlikle alanında kullanımına dair değerlendirme yapılmaktadır. Dördüncü bölümde literatürdeki çalışmalardan edinilen bilimsel bulgular ve sonuçlara dayanan bir tartışma ve konuyla ilgili çeşitli yorumlar bulunmaktadır.

2. SESİN İŞLENMESİ VE SES OLAYI TESPİTİ

Beş duyu ile algıladığımız dünyamızdaki çevremizde cereyan eden olayların büyük bir kısmını işitme duyumuz yoluyla ses olayları olarak algılarız. Çevremizdeki çoğu ses anlamsal olarak aynı hissi uyandırabilmekte, fakat birbirinden farklı sese dair akustik özelliklere sahip olabilmektedir. Hesaplamalı ses analizi kapsamında bir ses olayını tespit etmede sadece bir sınıf (sesi ifade eden etiket) ile diğer sınıfları (diğer ses etiketleri) karşılaştırarak bir sonuca ulaşılmaya çalışır, bunun için uygun hesaplama gereksinimleriyle (işlemci gücü, bellek ve depolama miktarı) işlem yapılır. Bu bakış açısıyla sesin işlenmesi, ses olay tespiti olguları taban alınarak makine öğrenmesi veya derin öğrenmenin kullanımıyla uygun sonuç elde etmeyi literatürdeki birçok çalışmada görmek mümkündür. Sınıf sayısı arttıkça ses örneğine/örneklerine uygun doğru sınıfların

tespitinin oranı düşebilmekte, bahsi geçen anlamsal benzerlikler nedeniyle daha belirgin özniteliklerin seçilmesiyle ayrıştırıcı bir sınıflandırma elde edilmeye çalışılmaktadır.

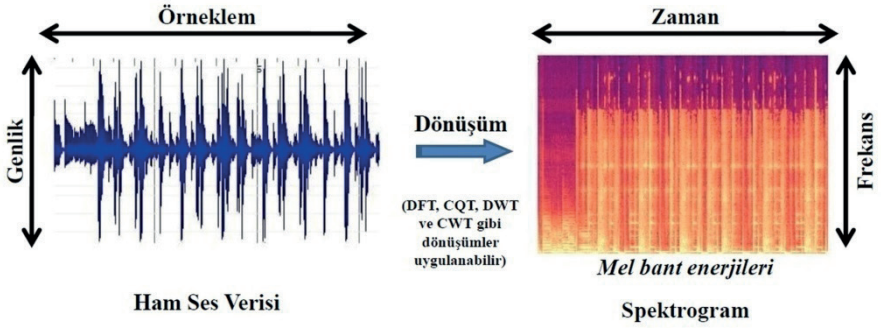
Sesin düzgünce, belirli biçemlerde işlenebilmesi için önce veri edinimi (data acquisition) aşamasının gerçekleştirilmesi gerekir. Ses verisi edinimi; akustik verinin tipini, kayıt ortamını, eğer metabilgi ile oluşturulan bir veri topluluğu ise kayıta ait metabilgiyi içermekte, böylece örnekleme oranı ve frekans (Hz) değerlerindeki ses sinyallerinden oluşturulmuş dosyaları içeren veri kümesinin elde edilmesini sağlamaktadır. Veri edinimi başarıyla gerçekleştikten ve ses kayıt dosyaları uygun ortamda depolandıktan sonra ses işleme (audio processing) aşamasına geçilir. Özellikle ses işleme aşamasının yüksek başarısı makine öğrenmesi/derin öğrenme ile yapılacak tespit ve sınıflandırma başarımını da iyileştirmektedir. Ses işlemede ön işlem (preprocessing) oldukça önemli bir adımdır, çünkü sesteki gürültünün giderilmesi veya ses düzeyinin yükseltilmesiyle hedef ses olayının daha belirgin hale gelmesi sağlanabilmetedir. Daha sonra ön işlemden geçirilen ses sinyallerinden ses öznitelikleri elde edilir (audio feature extraction). Ses ön işleme sırasında gerekiyorsa birden çok kaynaktan gelerek üst üste örtüşme (overlapping) yapan sinyaller ayrıştırılabilmektedir. Bu bahsi geçen ses (akustik) öznitelikleri ses analizinin temel taşı olarak kullanılmaktadır (Virtanen ve ark., 2018).

2.1. Ses Analizi Ve Öznitelik Elde Etme

Ses olayı tespit ve ses/konuşmacı tanıma için aynı sınıfa atanan ses örneklerinden elde edilen ses özniteliklerinin birbirleriyle çok az farklı olması hatta mümkünse farklı sınıfa atanan ses örneklerinin ses özniteliklerinin ise birbirlerinden çok daha farklı olması istenmektedir. Böylece sınıflar arası ayrıştırma veya benzerliklerin bulunması işlemi verimli yapılabilmektedir. Bu bakış açısıyla, ses analizi ses sinyalinden ses özniteliklerinin sesi temsil etmesi için elde edilmesine dayanmaktadır. Makine öğrenmesi/derin öğrenmenin daha az maliyetli bir biçimde gerçekleştirilmesi de ses özniteliklerinin (akustik özellikler) düzgünce temsil edilmesi (bilgisayar ortamına aktarılması) yoluyla olmaktadır. Böyle bir ses veri temsili, öznitelik olarak ifade edilen sayısal değerlerden oluşmuş haldeki ses sinyali içeriğidir. Ses öznitelikleri vektör veri tipi olarak çoğunlukla çalışmalarda kullanılmaktadır. Bu öznitelikler; sinyal enerjisini, frekans dağılımını ve zamandaki sinyal değişimini gösteren sayısal değerler olabilmekte ve çoğu öznitelik ses sinyalini çerçevelere, pencerelere bölünmesiyle ve spektrum (tayf) analizi yapılarak veya sinyalin istatistiksel bilgisinin (frekans bileşenleri büyüklüğüne bakılması) elde edilmesiyle oluşturulmaktadır.

Çerçeveye bölme işlemi, ses sinyalinin sabit uzunluklu çerçevelere belirli bir zaman adımı ile değişecek şekilde (örneğin 50 milisaniye gibi) bölünmesine denilmektedir. Sinyal enerjisinin frekans domenine (domain) veya zaman-frekans gösterimine dönüştürülmesi öznitelik temsilinde önemlidir. En çok kullanılan dönüşüm (transformation) tipleri ayrık Fourier dönüşümü (discrete Fourier transform, DFT), sabit-Q-dönüşümü (constant-Q-transform, CQT), ayrık dalgacık dönüşümü (discrete wavelet transform, DWT) ve sürekli dalgacık dönüşümü (continuous wavelet transform, CWT) olarak literatürdeki çalışmalarda yer almaktadır.

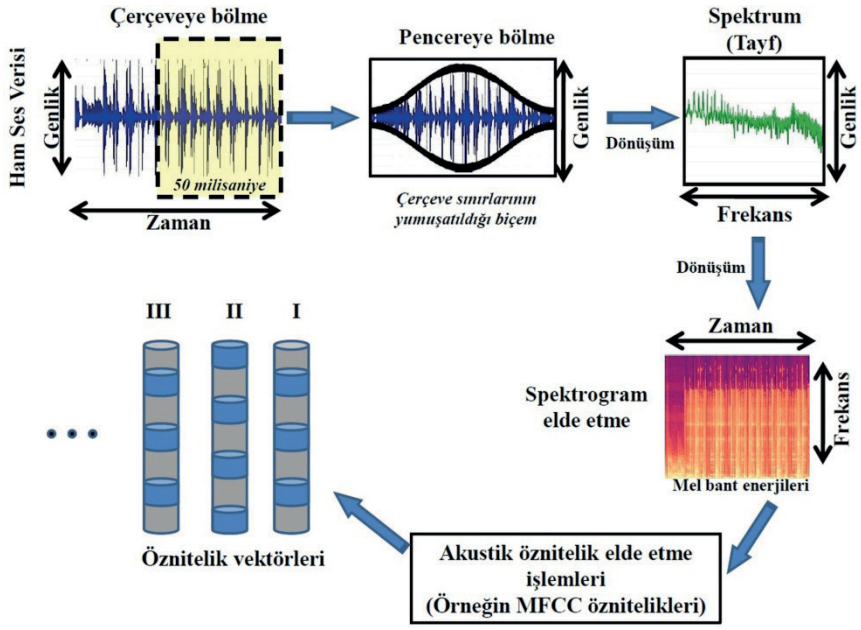
Pencere bölmede, pencere fonksiyonu ile ses çerçeveleri arası ani geçişler ve spektrumu (tayfı) bozan çerçeve sınırları yumuşatılarak daha düzgün bir yapı elde edilir. Bu sayede spektrum analizi daha doğru yapılabilmektedir (Virtanen ve ark., 2018). Şekil 1 'de ham (raw) ses sinyalinden verinin elde edilmesi ve buna karşılık gelen zaman-frekans gösteriminde mel bant enerjilerini içeren bir spektrogram gösterilmektedir.



Şekil 1. Ham ses sinyalinden elde edilen veri ve dönüşümler yapıldığında buna karşılık gelen spektrogram.

Ses analizinde çokça kullanılan bir ses veri temsil biçemi de spektrogram'dır. Bir ses sinyali'nin spektrogramı, zaman-frekans domenindeki (time-frequency domain) öznitelik matrisi olarak tanımlanmıştır. Bu matris, bir ses kaydındaki ardışık zaman çerçeveleri için frekans domeni öznitelik vektörlerinin uç uca birbirlerine eklenmesiyle oluşturulur (Çakır, 2019). İnsan kulağı algısal olarak sinyalin düşük frekanstaki değişimlerine karşı daha fazla hassastır. Mel ölçeği doğrusal olmayan bir frekans ölçeğidir, buna göre ses perdeleri insan tarafından eşit aralıkta algılanacak şekilde ayarlanmaktadır. Bu ölçeğe uygun olarak mel spektrogramı bir matris olarak mel bantı enerji özniteliklerinin vektörlerinden oluşur. Üçgen filtreler içerek şekilde mel ölçeğini kullanan mel filtre bankası çeşitli filtreler için merkezi frekanslarla bandgenişliğini genişletir (Çakır, 2019).

Şekil 2'de akustik (ses) öznitelik vektörlerinin elde edilmesi işleminde uygulanan çerçeveye bölme, pencereye bölme, spektrum oluşturma, spektrogram oluşturma ve çeşitli dönüşüm işlemlerinden oluşan ses özniteliklerinin elde edilmesi (feature extraction) süreci gösterilmektedir.



Şekil 2. Akustik (ses) özniteliklerini elde etme süreci.

Ses işlemedeki temel bazı öznitelikler tipleri; *zamansal* (temporal), *tayfsal* (spectral) ve *prozodi* (prosodic) olarak literatürdeki çalışmalarda verilmektedir. Zamansal öznitelikler olarak; "*kısa-dönemli enerji*" özniteliği ile sesin yüksekliği belirlenirken, "*sıfır geçiş oranı*" (zero-crossing rate, ZCR) özniteliği ise bir ses sinyalinin işaretinin değiştiğini veya bir ses çerçevesi için ses sinyalinin belirli zaman biriminde sıfır sinyalini kaç kez geçtiğini belirten sayısal değerdir. Böylece kaç kez sinyal işaret değiştirdiyse bunu ses çerçeve sayısına bölerek ZCR değeri elde edilmiş olur (Korkmaz ve Boyacı, 2018; Babaee, Anuar, Wahab, Shamshirband ve Chronopoulos, 2017).

Tayfsal (spectral) öznitelikler olarak; "*mel-frekans kepstral katsayıları*" (mel-frequency cepstral coefficients, MFCCs) bir ses çerçevesinden MFCC'lerin elde edilmesinde ayrık kosinüs dönüşümü (discrete cosine transform, DCT) kullanılarak elde edilen mel ölçeği filtre bankasının çıktılarına dair kepstral katsayılarından oluşmaktadır. MFCC öznitelik vektörü çoğunlukla 12 ilâ 15 kadar en düşük DCT katsayısından oluşur. Ayrıca *delta MFCC* ile ardışık zaman çerçeveleri için MFCC'ler arası farka bakılabilir.

"*Tayfsal orta nokta*" (Spectral centroid) özniteliği ise sinyal dağılımı boyunca tayfsal enerjinin orta noktasındaki değere denilmektedir. Bu öznitelik sayesinde gürültüye karşı daha gürbüz (robust) bir hesaplama yapılabilir, çünkü en baskın sinyal frekansını zaman boyunca temsil etmektedir. "*Tayfsal kayma*" (Spectral rolloff) ses sinyalinin biçimine ait eğriliğinin (skewness) frekansının hesaplamasında kullanılan bir özniteliktir. "*Tayfsal akı*" (Spectral flux) özniteliği ise ses sinyalinin ani değişimler/ataklar gösterdiği noktaların tespitinde kullanılır. "*Tayfsal entropi*" (Spectral entropy) özniteliği ses bilgi

kaynağındaki belirsizliğin ölçütü olarak bilgi miktarını ölçmede kullanılmaktadır. Bunların haricinde diğer bazı tayfsal öznitelikler de literatürde mevcuttur (Babaee ve ark., 2017).

Prozodi (Prosodic) öznitelikleri, frekans bazlı özniteliklerin algısal biçimlerine dayanmaktadır. Buna göre, insanın algısal olarak bir anlam atfettiği haliyle ortamdan dinlediği ses sinyalinin içeriğindeki bilgiyi ifade eden prozodi öznitelikleridir. Bunlardan; "*Perde (Pitch) ya da temel frekans*" özniteliği ses perdelerinin zaman içerisinde tüm sinyal hakkında istatistik oluşturulmasında kullanımını temel alır. "*Ses yüksekliği/yoğunluğu*" ses sinyal enerjisinin insan kulağı tarafından ses genliğinin algılanmasıyla hesap edilmesine dayanır. "*Ritim/süre*" özniteliği ses kaydındaki *sesli* (voiced) ve *sessiz* (unvoiced) kısımların tespiti, zamansal değişimlerin anlaşılması gibi olguları içermektedir (Babaee ve ark., 2017).

Burada bahsi geçen ve diğer bazı özniteliklerin kullanımı sayesinde ses veri kümesi üzerinde daha ileri işlemleri içeren (danışmanlı, danışmansız öğrenme vs.) aşamalar literatürdeki çalışmalarda gerçekleştirilmektedir.

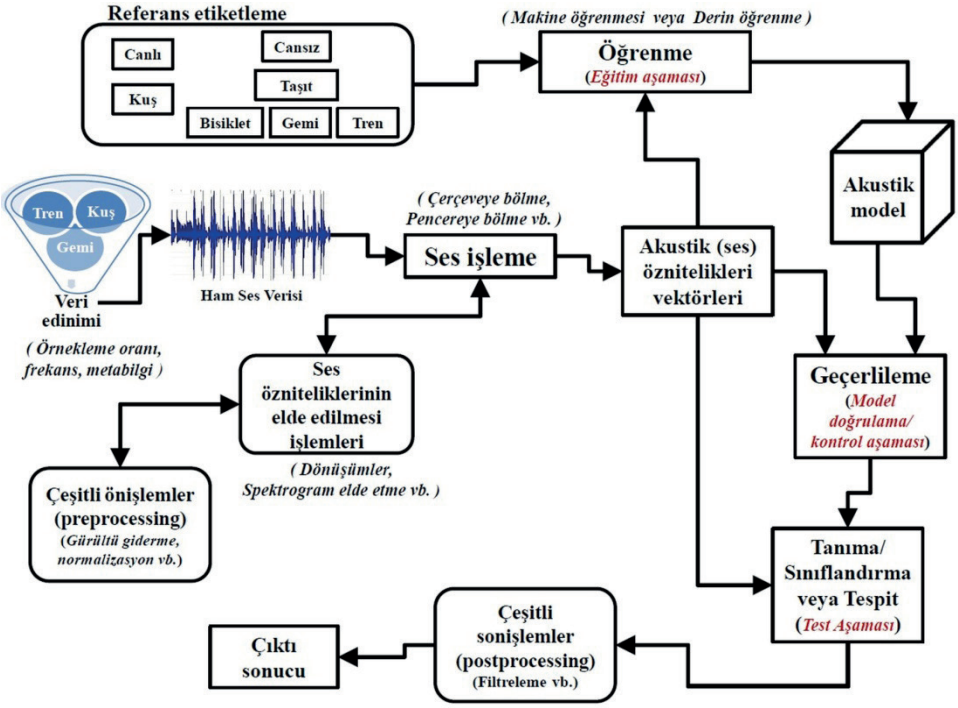
2.2. Ses Olayı Tespiti İçin Sınıflandırma Ve Derin Öğrenme

Makine öğrenmesi yoluyla bilgisayar biliminde kümeleme ve sınıflandırma yapılabilmektedir. Ses olay tespiti temelinde makine öğrenmesi, veri madenciliği, uygulamalı matematik ve istatistik gibi bilimlerin kullanıldığı, sınıflandırmanın sıklıkla önceden belirlenmiş etiketlere karşın eğitimle sonuca varılan ve belirli bir süre alan bir biçimde karşılaştırma yapılmasına dayandırıldığı bir bakış açısına sahiptir (Çakır, 2019; Babaee ve ark., 2017). Eğitim aşamasının yoğunluğuna göre sınıflandırma işlemi için gerekli süre uzamaktadır. Bunun yanı sıra, gerçek zamanlı (çok küçük veya tolere edilebilir süreler zarfında) cevap üreten sistemler de mevcuttur. Makine öğrenmesiyle (derin öğrenmede de kullanılan bazı çeşitleri ile) sınıflandırma; çeşitli yayınlarda denetimli olarak da ifade edilen *danışmanlı öğrenme* (supervised learning), denetimsiz olarak da ifade edilen *danışmansız öğrenme* (unsupervised learning), *yarı-danışmanlı* (semi-supervised) veya pekiştirmeli olarak da ifade edilen *destekleyici öğrenme* (reinforcement learning) çeşitlerinde yapılabilmektedir (Babaee ve ark., 2017; Goodfellow ve ark., 2016).

Danışmanlı öğrenme tipinde önceden tanımlı etiketlerle ses sınıfları belirtilmekte bu yolla yüksek doğrulukta bir sınıflandırma eğitim sonrası ses kayıtlarının doğru sınıflara gruplandırılması ile sonuçlanmaktadır. Danışmansız öğrenme tipinde önceden tanımlı etiketler olmadığından tüm veri üzerinde ilgili özelliklere göz atarak ses kayıtlarına dair benzerlikler aranır ve buna istinaden bir sınıflandırma sonucu elde edilir. Yarı-danışmanlı öğrenmede hem önceden etiketlenmiş hem de etiketlenmemiş veriyi içeren bir bileşik veri kullanılır. Destekleyici (pekiştirmeli) öğrenmede bir etmen (agent) ait olduğu çevrede kendi aldığı aksiyonların sonucuna göre elde ettiği ödülü en büyüklemeye veya ceza puanını en küçüklemeye çalışmakta, bu yolla öğrenme gerçekleşmektedir (Babaee ve ark., 2017; Goodfellow ve ark., 2016).

Ses kayıtlarının topluluğuna dair ses olayı etiketleri ve akustik öznitelikler kullanılarak yapılan sınıflandırmada, ses öznitelikleri ses çerçevelerinden elde edilmektedir. Elde edilen bu özniteliklerle girdi matrisi oluşturulup sınıflandırma aracına (çoğunlukla derin öğrenme özelinde yapay sinir ağına) verilerek eğitilmesi sağlanır. Eğitim ile elde edilen sonuçlar, referans etiketlemesi yapılmış çıktı matrisi ile karşılaştırılarak genel başarım ölçülür. Bu karşılaştırma işlemi için bilgi elde etme (information retrieval) teorisindeki çeşitli başarım ölçütleri de mevcuttur (Çakır, 2019). Bu nesnel başarım ölçütlerinin en sık kullanılanları; anma (recall) veya gerçek pozitif oranı (true positive rate, TPR), duyarlılık (precision), gerçek negatif oranı (true negative rate, TNR), yanlış negatif oranı (false negative rate, FNR), yanlış pozitif oranı (false positive rate, FPR) ve doğruluk (accuracy) oranıdır. F-ölçüsü ise anma ile duyarlılık ölçütlerinin harmonik ortalaması alınarak hesaplanılır (Babaee ve ark., 2017). Özellikle bir sınıflandırıcının/tahminleyicinin yada tespitin ne kadar başarılı bir sonuç ürettiğinin anlaşılmasında bu ölçütler, istatistiksel hata oranları gibi basit değerlerin yanı sıra oldukça kullanışlı birer indikatör vazifesi de görmektedir. Konuşmacı tanıma özelinde, denk hata oranı (equal error rate, EER), kelime hata oranı (word error rate, WER) (Zeyer, Doetsch, Voigtlaender, Schlüter ve Ney, 2017), konuşma kalitesinin algısal değerlendirmesi (perceptual evaluation of speech quality, PESQ), kısa-dönemli nesnel anlaşılırlık (short-time objective intelligibility, STOI) ve konuşma bozunma oranı (speech to distortion ratio, SDR) gibi ölçütler literatürdeki çalışmalarda kullanılmaktadır (Kang, Shin ve Kim, 2018; Samui, Chakrabarti ve Ghosh, 2017). Tüm bunların yanı sıra alıcı işletim karakteristiği (receiver operating characteristics, ROC) eğrisi ve ROC alanı altında kalan alan (area under curve, AUC) değeriyle başarım değerlendirmesi

yapılması TPR ve FPR değerlerine dayandıkları için çalışmalarda sıklıkla kullanılan başarım değerlendirme (performance evaluation) yaklaşımlarındandır. Şekil 3'te öğrenme (eğitim) aşaması, geçerlileme (model doğrulama/kontrol aşaması) ve tanıma/sınıflandırma/tespit (test) aşamasından oluşan temel ses analizi sistemi gösterilmektedir. Şekil 3'te veri edinimi (data acquisition) yoluyla analog ses sinyalleri sayısal sinyale ve böylece veriye dönüştürüldükten sonra ses işleme adımında ses (akustik) özniteliklerinin elde edilmesi işlemleri yapılmaktadır. Gerekli durumlarda önişlem (preprocessing) sayesinde sinyaldeki gürültülerin giderilmesi ve normalizasyon gibi işlemler yapılırken, son işlem (postprocessing) sayesinde ise filtreleme işlemleri yapılabilmektedir. Ses analizi sürecinde ilk önce; öğrenme, geçerlileme ve test aşamalarında kullanılacak akustik (ses) öznitelikleri oluşturulmakta, buna uygun olarak tüm süreç ilerlemektedir. Öğrenme süreci sayesinde oluşturulan akustik model, makine öğrenmesi veya derin öğrenme modellerinden herhangi biri ile oluşturulabilmekte ve test sonucunda elde edilecek çıktı sonucunun sınıflandırma/tespit veya tahmin başarımı bu akustik modelin ne kadar iyi kurulabildiğine bağlı olarak değişmektedir. Bu nedenle akustik modelin oldukça iyi bir biçimde kurulması ve modelin doğruluğunun sınanması belirli bir süre alan hesap yoğun bir işlemdir.



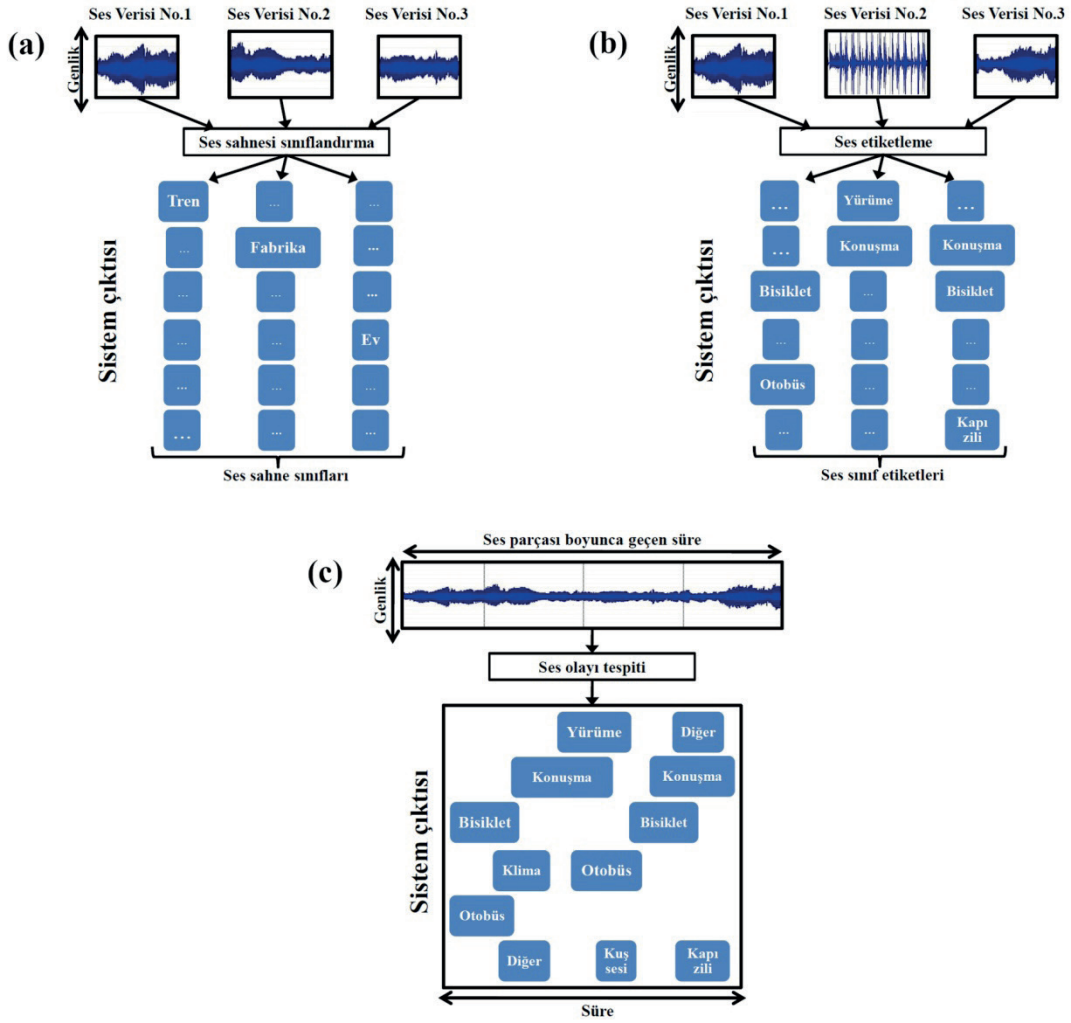
Şekil 3. Temel ses analiz sistemi

Eğitim aşamasında, sınıflandırıcı sistem tarafından bir fonksiyon öğrenilmeye çalışılır. Bu fonksiyon bir model olarak verilen girdi verisi ile hedef çıktısını birbirine eşlemeye çalışmakta, sınıflandırıcı sistem parametrelerin kullanımı yoluyla girdi ile hesaplanmış çıktı arasındaki farkı (hata payı) en aza indirgenmeye uğraşmaktadır. Bu hata payı ne kadar az olursa, bahsi geçen eşleşme o kadar iyi olacak ve genel sınıflandırma başarımı yüksek olacaktır. Test aşamasında hesaplanan çıktılarla gerçek çıktının (hedef çıktının) ne kadar eşleşip eşleşmediği kontrol edilir. Tüm ses kayıt verisinin hepsini eğitim veya test için kullanmak yerine, belirli bir miktarını eğitim aşamasında, belirli bir miktarını geçerlileme (eğitimin akustik model üzerinden düzgünce yapılıp yapılmadığının kontrolü) için kullanmak, daha sonra da verinin geriye kalan belirli miktarı ile de test işlemi gerçekleştirilmesi uygun olmaktadır. Test verisinin eğitim aşamasında kullanılmamış olması nedeniyle gerçek hayattan gelen ses kayıtlarına göre ses olayı tespitinin ne kadar doğru yapılabildiğini de kanıtlamaya yardımcı olmaktadır. Ses olayları, tekli (monophonic) veya çoklu (polyphonic) seslerden oluşabildiği için eğitim aşamasında buna dair detaylara da dikkat edilir. Derin öğrenmenin bazı mimari modellerinde ses verisini zaman serisi gibi ele alarak girdi verisi halinde kullanmak da gerekebilmektedir (Çakır, 2019).

Çok karmaşık kavramların daha basit temel kavramlar baz alınarak sıradüzensel biçimde bir çizge (graph) haline getirilmesiyle, bilgisayarca daha kolay (daha az maliyetli) temsil edilebilmesine dayanan bir yaklaşım literatürdeki çalışmalarda ortaya

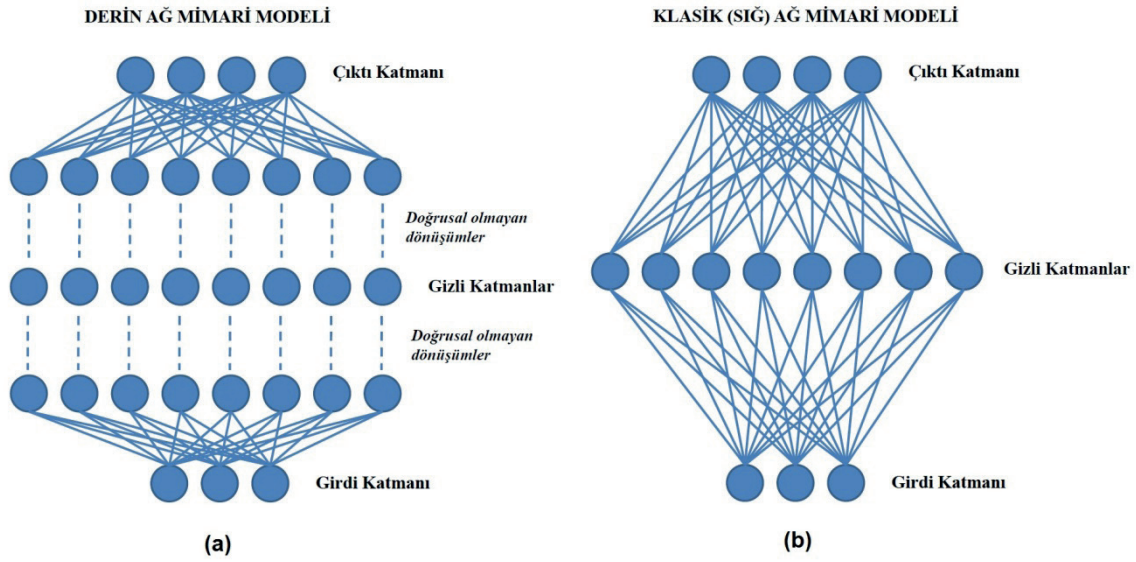
atılmıştır. Buna göre çok katmanlı böyle bir kavramlar haritası şeklindeki hesaplamalı çizge (computational graph) sayesinde kavramlar sıradüzeni uygun mimari modeller yardımıyla öğrenilebilmektedir. Klasik makine öğrenmesinde de kullanıldığı haliyle sadece eniyileme yapmak böyle bir durumda yeterli olmamakta, katmanlı yapının girdi verilerini nasıl temsil etmesi gerektiği ve öznitelik elde etmenin çeşitli yolları sayesinde *derin öğrenme* yaklaşımı ortaya çıkmıştır (Goodfellow ve ark., 2016; Patterson ve Gibson, 2016).

Klasik makine öğrenmesinden farklı olarak derin öğrenmede elle tasarlanmış özniteliklerin yerine basit özniteliklerden daha soyut özniteliklere geçiş için daha çok katmanı içeren bir mimarinin kullanılması, böylece temsili öğrenme (representative learning) yapılması fikri temel alınmıştır. Derin öğrenme için eğitim verisi miktarı (modele bağlı olarak) artırıldıkça veya model katmanları sayısı/boyutları değiştirildikçe öznitelik öğrenimine bağlı olarak daha yüksek başarım da elde edilebilmektedir. Günümüzde büyük verinin kullanımının artması ve bilgisayar donanımının gelişmesi derin öğrenmenin de yaygınlaşmasında etkili olmuştur (Goodfellow ve ark., 2016). Şekil 4'te çeşitli ses analiz sistemleri görülmektedir. Şekil 4 (a) ses sahnesi sınıflandırma sisteminde ve Şekil 4 (b) ses etiketleme sisteminde birden çok ses verisi girdi olarak alındıktan sonra Şekil 3'te belirtilen temel ses analiz sistemindeki izlenen sürece benzer bir yoldan çıktı sonuçları üretilmektedir. Buna göre girdi uzayı uygun çıktı sınıflarına eşleştirilmiş olmaktadır. Şekil 4 (c) şıkkında ise belirli bir ses parçası boyunca geçen süre zarfında oluşan çeşitli ses olayları zaman içerisindeki uygun noktalara sistemin yaptığı tespit sonucunda yerleştirilmektedir. Sistem tüm bu süre boyunca hangi olayların hangi zaman aralığında gerçekleştiğini ve birbirleriyle olan ilişkilerini ortaya koyabilmektedir (Virtanen ve ark., 2018).



Şekil 4. Ses analiz sistemi çeşitleri, (a) ses sahnesi sınıflandırma, (b) ses etiketleme, (c) ses olayı tespiti (Virtannen, Plumbley ve Ellis, 2018)

Literatürde sıklıkla bazı derin öğrenme mimari modelleri kullanılmaktadır. Ele alınan ses kaydının daha iyi temsil edilebilmesi için yüksek çözünürlüklü bir spektrogram alanı birçok ses çerçevesinin yan yana getirilmesiyle oluşturulur. Bu spektrogram alanı girdi olarak derin öğrenme mimari modellerine verilebilmektedir. Derin öğrenme modelleri olarak derin sinir ağları (deep neural network, DNN) birden çok gizli katmanı bulunan ve bu katmanlarda doğrusal olmayan (non-linear) birçok gizli birime (sinir düğümü) sahip, bu düğümlere ait çeşitli ağırlıkları da mevcut olan ve çıktı katmanında oldukça fazla düğüme sahip olabilen modeller olarak önerilmiştir. Modelden modele farklılıklar olsa da mimari de genellikle girdi katmanı, gizli katmanlar ve çıktı katmanı olacak şekilde, girdiye karşılık aktivasyon fonksiyonu ile ateşleme yapılarak çıktı üretilen bir yapısal ağ biçimi mevcuttur. Sıklıkla üstüste yığılan (stacked) katmanlardan oluşmakta ve böyle bir ağın derinliği katman ve düğüm sayısıyla ilişkili olmaktadır. Şekil 5 'de basit bir derin öğrenme mimari modeli olarak derin sinir ağları (DNN) ile klasik (sığ) sinir ağı modelinin klasik makine öğrenmesinde kullanılan biçemiyle çok katmanlı algılayıcı (multilayer perceptron, MLP) karşılaştırılması şematik olarak verilmektedir. Derin öğrenme modeli olan derin sinir ağları (DNN'ler) genel tasarım olarak ileri beslemeli (feed-forward) yapıda olan, girdi katmanı ile çıktı katmanı arasında belirli bir aktivasyon fonksiyonuna uygun çalışan, çoğunlukla da geri yayılımın (backpropagation) tercih edildiği ve sonuçta elde ettiği güncel çıktı (actual output) ile hedeflenen çıktı (target output) arasındaki farkı bir hata hesabı gibi alarak maliyet fonksiyonunu (cost function) hesaplayan bir modeldir. Aşırı öğrenme (overfitting) ve az öğrenme (underfitting) durumlarından kaçınmak için ön eğitim (pretraining) yapılabilmekte böylece ince ayar (fine tuning) yapılarak ağ düğüm ağırlıkları güncel olarak ayarlanabilmektedir (Hinton, Deng, Yu, Dahl, Mohamed, Jaitly, Senior, Vanhoucke, Nguyen, Sainath ve Kingsburry,, 2012). DNN'ler çoğunlukla konuşma tanıma, ses sınıflandırma ve tekrar oluşturma/sentezleme gibi konularda kullanılmaktadır. Sınıflandırma başarımı özniteliklerin elde edilmesi aşamasına çokça bağımlı olduğu için doğru ve yeterli (gerekli) özniteliklerin belirlenmesi ve buna göre eğitim için bu öznitelik değerlerinin kullanılması önemlidir.

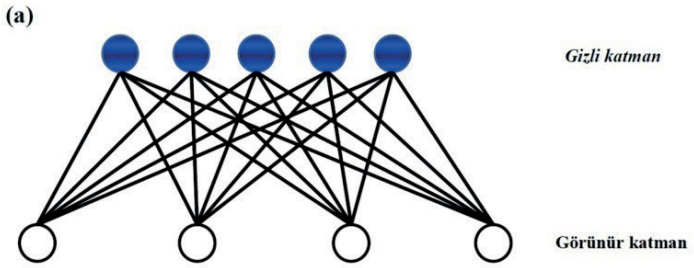


Şekil 5. Derin öğrenme ile sığ öğrenme modelinin karşılaştırması, (a) Derin model (DNN), (b) Klasik (sığ) model (MLP).

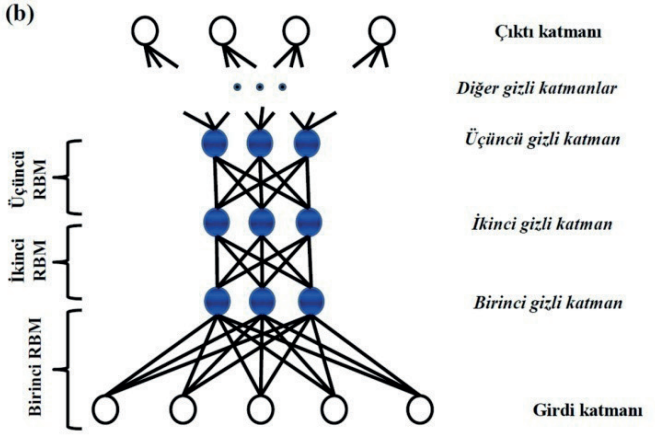
Üretken ağlar (generative networks) şeklinde anılan hesaplamalı çizge modeline örnek olarak derin inanç ağı (deep belief network, DBN) tipindeki ağları verebiliriz. Bunlar üstüste yığılmış kısıtlanmış Boltzmann makinesi (restricted Boltzmann machine, RBM) alt ağlarının (katmanlarının) derinlik oluşturulmasına dayanmaktadır. Bu ağlar çoğunlukla gürültü giderme ve sinyal temizleme, sınıfları ayrıştırma, konuşma tanıma gibi danışmasız veya yarı-danışmanlı bir eğitim ile çalışmaktadır. Ayrıca diğer derin öğrenme modelleri için ön eğitim veya ince ayar amaçlı RBM'ler de kullanılabilmektedir. Buna benzer olarak, üretken yığılmış oto kodlayıcılar (generative auto-encoders, GSAE) tipi derin öğrenme mimari modelinde bir aktivasyon fonksiyonu sıklıkla sigmoid tipinde olmakta ve gerekli dönüşüm doğrusal olmayan bir biçimde gerçekleşmektedir (Qian, Chen ve Yu, 2016). Üretken ağ olan DBN gibi çoğunlukla enerji tabanlı fonksiyonlara dayalı (eniyileme çoğunlukla dahili olarak yapılır) biçimde öznitelik öğrenimi yapan ağların eğitim süreleri diğer ağlara göre uzun olsa da sınıflandırma başarımları görece diğer ağ modellerine göre

yüksek olabilmektedir. Bu nedenlerle daha kısa sürede eğitim yapabilen bazı modellere ve DBN'lerin modifikasyonlarına da literatürdeki çalışmalarda yer verilmiştir. Bunun yanı sıra bu mimari modeller girdilerin birbirleri ile arasındaki ilişkiyi de (öznitelikler arası bağımlılıklar olarak) ortaya çıkarabilmektedir (Espi, Fujimoto, Kinoshita ve Nakatani, 2015). Şekil 6'da bir kısıtlanmış Boltzmann makinesi (RBM) ve bunu kendi katmanları olarak kullanan bir derin inanç ağı (DBN) görülmektedir. RBM'nin özelliği olarak biri gizli biri görünür katmanlardaki düğümler arasındaki bağlantılar tam bağlantılı (fully connected) olarak kurulurken, görünür katman içerisindeki düğümler arasında veya gizli katman içerisindeki düğümler arasında bağlantı kurulmamaktadır.

Evrişimli sinir ağları (Convolutional neural network, CNN) öncelikli olarak görüntü işleme alanında kullanılsa da yerel özniteliklerin elde edilmesindeki başarısı nedeniyle belirli küçük parçalar olarak spektrogram alanı yamalarının (patch) yardımıyla özniteliklerin filtrelerle işlenmesini sağlamaktadır. Böylece bu küçük alanlardan (yamalardan) öznitelikler öğrenilmekte ve ağın eğitim aşamasında kullanılmaktadır. CNN'ler evrişim ve biriktirme (convolution-and-pooling) işlemlerini yapan çok katmanlı algılayıcılar (multilayer perceptron, MLP) olarak çalışırlar. Bu yolla katmanlar diğer derin sinir ağlarındaki gibi birden çok gizli katman ve düğüm ile oluşturulmaktadır. Çok boyutlu evrişimli öznitelik haritası eğitim amacıyla katman yapısına uygun olarak tam bağlantılı (fully connected) katmanlar arasında katmandan katmana geçirilir. CNN'ler akustik sahne sınıflandırmada, transfer öğrenmede (bir veri alanından öğrenilen sınıflandırma bilgisini başka alana uygulamak) ve alana adaptasyonda yoğunlukla kullanılmaktadır (Lecun, Bottou, Bengio ve Haffner, 1998; Sainath, Kingsburry, Saon, Soltau, Mohamed, Dahl ve Ramabhadran, 2015; Espi ve ark., 2015; Zhou, Bai ve Du, 2018). Şekil 7'de ses işleme yoluyla sınıflandırma yapmakta kullanılan iki evrişim (convolution) katmanı, iki tane biriktirme (pooling) katmanı ve bir tane tam bağlantılı (fully connected) katmandan oluşan temel bir CNN ağı görülmektedir (Kiranyaz, Avci, Abdeljaber, Ince, Gabbouj ve Inman, 2019).

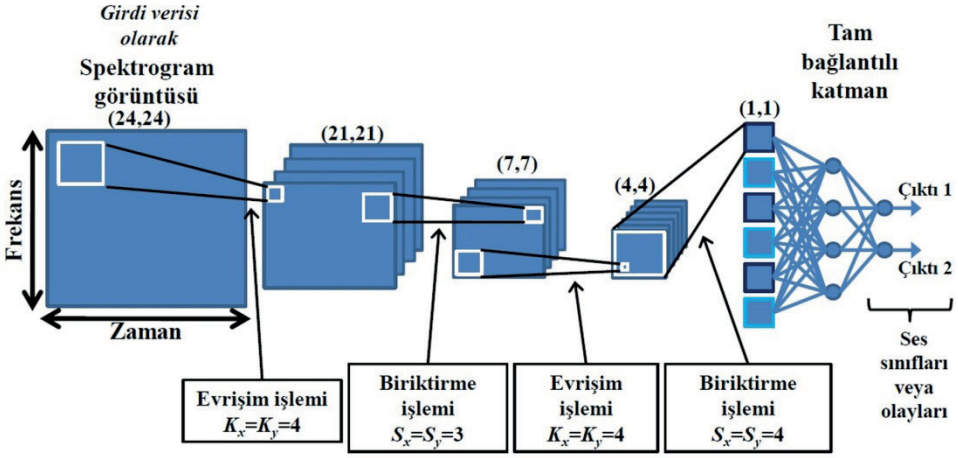


İki katmanlı (biri gizli biri görünür katman) bir kısıtlanmış Boltzmann Makinesi (RBM)



İkiden fazla katmanı bulunan (RBM biçeminde katmanları olan) bir derin inanç ağı (DBN)

Şekil 6. Üretken ağlar, (a) kısıtlanmış Boltzman makinesi (RBM), (b) derin inanç ağı (DBN).



Şekil 7. Evrişimli sinir ağı (CNN)

Şekil 7'de gösterilen büyük matrislerin (mavi renkli) üzerindeki beyaz kareler, ilgili CNN ağı katmanında kullanılan yama (patch) şeklinde bu büyük matrislerden alınıp filtrelenerek (convolution-and-pooling) işlenen matris alt parçalarını göstermektedir. Burada *K* terimi çekirdek (kernel) ve *S* terimi alt örnekleme faktörü (subsampling factor) deyimlerinin gösterimidir. Bu büyük matrisler ise öznitelik haritaları (feature maps) olarak CNN tarafından katmanlar arası işlemler ile sağlanan bilgi akışıyla oluşturulmaktadır. Uygun her katmandan sonra uygun filtreleme işlemleri yapılırken, uygun katmanlarda da aktivasyon işlemi yapılmaktadır. Aktivasyon fonksiyonu sayesinde çok katmanlı algılayıcı (MLP) benzeri ateşleme bu sefer tüm ağ yerine ilgili katman bazında yapılmaktadır. Şekil 7'de ilk girdi olarak alınan spektrogram görüntüsü (ses öznitelikleri/bilgisi içeren bir gri tonlamalı görüntü olarak) önce (24,24) büyüklüğündeki bir matris haline getirilip ilk katmanda kullanılmakta ve ağın sonraki katmanlarında uygun evrişim (convolution) ve biriktirme (pooling) işlemleri kullanılarak işlenen matris büyüklüğü (öznitelik haritası) katmandan katmana gittikçe azaltılmaktadır (indirgenmektedir). İkinci katmandaki (ikinci matrisin oluşturulduğu ilk evrişim katmanı) öznitelik haritası evrişim işlemi çekirdek büyüklüğü *x* ve *y* eksenleri için sırasıyla *Kx=Ky*=4 olarak alınarak işlenmiş halde (21,21) büyüklüğündeki matris bir öznitelik haritası haline getirilmektedir. Sonraki katmanda (ilk biriktirme katmanı) biriktirme işlemi alt örnekleme faktörleri *x* ve *y* eksenleri için sırasıyla *Sx=Sy*=3 olarak alınarak işlenmiş halde matris (7,7) büyüklüğüne indirgenerek ilgili öznitelik haritası haline getirilir. Bir sonraki katmanda (ikinci evrişim katmanı) yeni bir evrişim işlemi *x* ve *y* eksenleri için sırasıyla *Kx=Ky*=4 olarak alınarak işlenmesiyle matris (4,4) büyüklüğüne indirgenmekte ve ilgili öznitelik haritası oluşturulmaktadır. Sonraki katmandaki biriktirme işlemi (ikinci biriktirme katmanı) sayesinde *x* ve *y* eksenleri için sırasıyla *Sx=Sy*=4 için (1,1) büyüklüğüne indirgenen matris bir öznitelik vektörü halini alır (düzleştirme). Tam bağlantılı katman aynen bir MLP ağının çalışmasına benzer olarak uygun aktivasyon fonksiyonu kullanılmasıyla çıktı katmanının ses sınıfı çıktı sonuçlarını üretmesini sağlar. CNN ağının bu şekilde eğitim yineleme adımları boyunca ağırlıkların güncellendiği bir MLP ağına benzer bir yapıya dönüşerek uygun aktivasyon fonksiyonu ile katmandan katmana öznitelik haritasının aktarıldığı, bu yolla özniteliklerin en iyi şekilde öğrenildiği ve düşük hata oranlarıyla sınıflandırma yapabildiği bir biçemde olduğu görülmektedir (Kiranyaz, Avci, Abdeljaber, Ince, Gabbouj ve Inman, 2019). CNN'lerin derin veya sığ olması gibi olguların yanı sıra CNN'lerin düğüm sayısı, eğitim yineleme adım sayısı, hangi regülarizasyonun kullanılacağı, biriktirme (pooling) katman sayısı, filtre sayısı ve biçimi gibi üstün parametrelerinin (hyperparameters) düzgünce belirlenmesi/seçilmesi ve kayıp (maliyet) fonksiyonu tipi (çapraz entropi vs.), eniyileme algoritma cinsi (adaptif momentler (ADAM), Nesterov momenti ve RMSProp vs.), iletim sönümü (dropout) miktarının uygun olarak seçilmesi ve verilen veriye uygun bir eğitimin yapılması sınıflandırma başarımının artmasında önemli rol oynamaktadır (Bhatt, Gupta, Arora ve Raman, 2018).

CNN'ler büyükçe veri kümeleri için oldukça iyi sonuçlar verebilmekte, çevresel ses sınıflandırmada kullanılabilmektedirler. Fakat görece küçük veri kümelerinde eğitim yapıldıktan sonra ağ'a hiç gösterilmemiş veri ile test yapıldığında sınıflandırma başarımının düşük olabildiği durumlarla da karşılaşılmaktadır. Veri artırma (data augmentation) bu probleme uygun bir

çözümdür. Buna göre veri kümesindeki eğitim örnekleri belirli deformasyona ve bazı işlemlere tabi tutularak değiştirilir fakat bu işlemler eğitim verisine dair etiketlerin anlamsallığını değiştirmez. Anlamsallığın korunduğu böyle deformasyonlar sonuçta ses alanına uygulandığında veri miktarı artar fakat anlamsal olarak hiç bir değişiklik olmaz, bunun yanı sıra ses sınıflandırma başarımı da böylece artmaktadır (Salamon ve Bello, 2017).

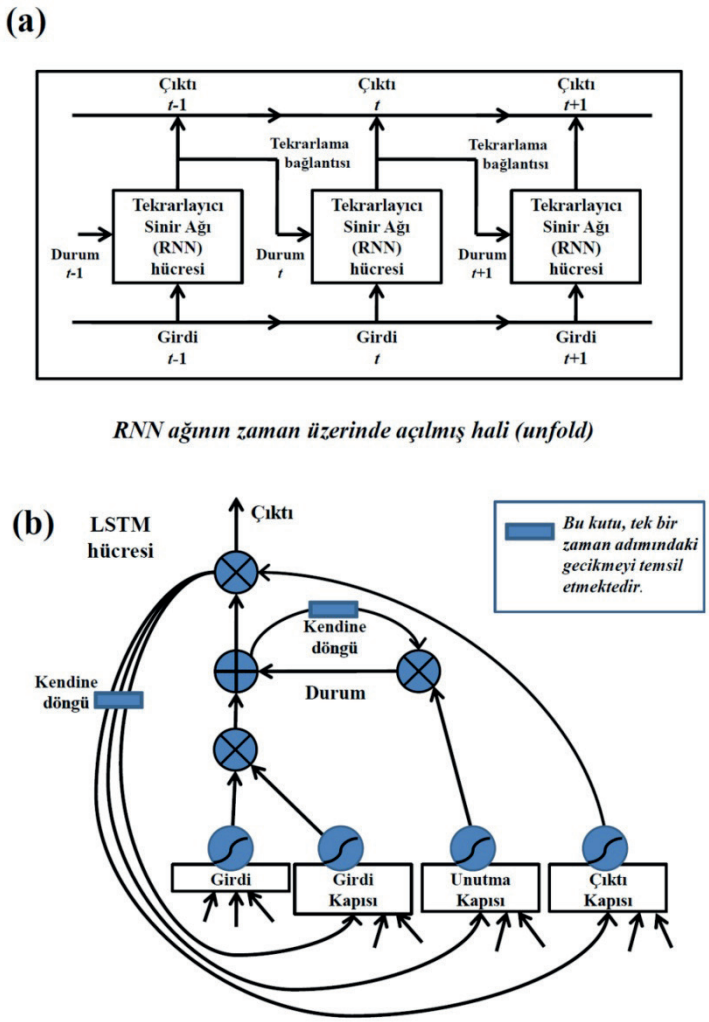
CNN tipi ağlar genellikle ham veri üzerinden üç farklı seviyede soyutlamayı (abstraction) sağlamaktadır. Düşük, orta ve yüksek seviyeli bu soyutlamalar, sırasıyla düşük seviyeli özniteliklerin elde edilmesiyle oluşan daha genel bilgilerden (örneğin görüntü verisi için renk, doku, biçim) katmanlı yapının kullanımı sayesinde orta seviyeli ve yüksek seviyeli özniteliklerin ifade ettiği bilgilere (örüntü ve yapısal bilgiler gibi) kadar bir sıradüzeni (hiyerarşi) oluşturur. Bu yolla ince ayara (fine-tuning) pek ihtiyaç duymadan verideki uzamsal bilgiler ve anlamsal içeriklerin belirlenmesi ve sınıflandırmada/tahminlemede kullanılması sağlanır. Bu nedenle CNN kullanımı sayesinde anlamsal boşluk iyi bir biçimde doldurulmaya çalışılır ve rafine sonuçlar elde edilir.

Tekrarlayıcı sinir ağları (Recurrent neural network, RNN), sıralı biçimde verilen verileri (çoğunlukla zaman serisi gibi ilişki barındıran veriler) işlemek için özelleştirilmiş hesaplamalı çizge modeli olarak verilen sinir ağı mimari modelleridir. CNN'ler görüntülerin büyükçe verileriyle uğraşırken, RNN'ler ise büyükçe verilerin sıralanmış dizileri ve bunların ölçeklenmiş halleriyle uğraşmaktadır. Her iki mimari tipinin de kendine göre avantajlı olduğu yanları mevcuttur. Hesaplamalı çizgeler ile bir olay zinciri (ilişkilerin mekanizması) ifade edilebilmekte, bu çizgeyi açarak (unfolding) derin öğrenme ağ mimarisinde yapısal olarak ilgili parametrelerin paylaşımı da sağlanabilmektedir. Bu yolla girdiler ve ilgili parametreler çıktılara (maliyet fonksiyonu olarak alınan kayıp fonksiyonunun da hesabıyla) eşlenmektedir (Goodfellow ve ark., 2016).

Özellikle RNN tipinde mimari modeller olarak, uzun-kısa süreli bellek (long-short term memory, LSTM) ve kapılı tekrarlayan birim (gated recurrent unit, GRU) tipi ağlar RNN'lerin çokça kullanılan modelleridir. RNN'ler verilen sıralı dizinin başından sonuna doğru işletildiğinde tek yönlü olarak adlandırılırlar. Eğer dizinin başından sonuna (ileriye doğru) bir RNN'yi işletir ve dizinin sonundan başlayıp başına doğru da (geriye doğru) başka bir RNN işletilirse, bunun yanı sıra bu iki yönden elde edilen genel sonuç da birleştirilirse buna çift yönlü RNN (bidirectional RNN) denilmektedir.

Şekil 8'de RNN ağı yapısı gösterilmektedir. Şekil 8 (a) ile zaman üzerinde açılmış (unfold) halde bir RNN ağı verilirken, Şekil 8 (b) ile bir LSTM hücresi blok şeması ile gösterilmektedir. Şekil 8 (b)'deki LSTM hücresinin girdileri ve girdi, unutma ve çıktı kapıları için sigmoid eğriliği benzeri bir yapı kullanılmakta, durum (state) birimi ise ağırlığı unutma kapısı aracılığıyla kontrol edilmekte olan doğrusal bir kendine döngüyle sürülmektedir. LSTM hücre çıktısı, çıktı kapısı tarafından kapatılabileceği gibi, durum birimi ise ek bir girdi şeklinde diğer kapı birimleri için de kullanılabilmektedir. Dikdörtgen kutu ile gösterilen ise tek bir zaman adımı için LSTM hücresi tarafından oluşturulan gecikmeyi ifade etmektedir (Chollet, 2017).

Bu mimari modeller ve diğer çeşitli tipteki modeller literatürde konuşma/konuşmacı tanıma (speech/speaker recognition), ses taklit etme (voice mimicry), ses sentezleme (sound synthesis) ve otomatik konuşmacı doğrulama sistemlerini aldatmanın tespiti (spoofing detection for automatic speaker verification) gibi konulardaki çalışmalarda kullanılmaktadır. Literatürdeki çalışmalarda oldukça fazla sayıda ve çeşitte veri kümeleri ve veri tabanları kullanılmıştır. İlgili yayınlardan bunların detaylarına ulaşılabilir. Sayfa sınırı nedeniyle bu makalede her bir veri küme veya veri tabanının teknik detayına yer verilmemiştir. Genel olarak bu tip veri topluluklarında belirli bir mücadele/yarışma usulüne göre en iyi başarıma ulaşan yöntem ve teknikler bir liste aracılığıyla bu alanda çalışan araştırmacılara ilan edilmekte ve bu veri kümeleri veya veri tabanlarıyla elde edilen sonuçları bir adım daha ileri götürerek iyileştirecek yeni sonuçlar elde etmeleri beklenmektedir.



Şekil 8. Tekrarlayıcı sinir ağı (RNN), (a) RNN ağının zaman üzerinde açılmış hali (Chollet, 2017), (b) uzun-kısa süreli bellek (LSTM) hücresi (Goodfellow ve ark., 2016)

Sonraki bölümde bu alandaki birçok çalışmaya dair yapılan incelemeye yer verilmektedir. Ayrıca bu çalışmalardan elde edilen bilimsel bulgulara dayanan tartışma ve yorumlar da makalenin en son bölümünde verilmektedir.

3. BULGULAR

Literatürde siber güvenlik alanında çoklu ortam sistemlerini kullanan biyometrik yetkilendirmeyi ihlal etmeye, sahtecilik veya suistimal yapılmasına yönelik veya siber saldırı yapılması nedeniyle biyometrik sistemlerin sağlıklı işleyişini bozmaya yönelik birçok çalışma bulunmaktadır. Konuşma/konuşmacı tanıma özellikle derin öğrenme mimarilerinin gelişimi ile önem kazanan alanlardan olmuştur. Konuşmanın bilgisayar tarafından tanınabilmesi için bir dizi halinde akustik sinyalin konuşmacı tarafından üretilmesi ve bunların da doğal dildeki kelimelere dönüştürülmesi gerekir. Konuşma/konuşmacıyı tanıma sistemleri özel olarak tasarlanan öznitelikleri kullanan girdilerle eğitim yapılmasına dayanarak sonuç üretir. Bunun için girdileri ön işlemden (preprocessing) geçirmektedirler. Bunun aksine çoğu derin öğrenme mimari modeliyle böyle bir ön işleme gerek kalmadan ham (raw) verilerle eğitim yapılabilmektedir. Verilen bir girdi dizisi şeklindeki çıktı dizisine (en olası dil dizisine) eşleyen bir fonksiyonun (ilişkileri bulan fonksiyon) oluşturulmasına otomatik konuşma tanıma (automatic speech recognition, ASR) denilmektedir (Goodfellow ve ark., 2016). ASR ile doğal dil işleme (natural language processing, NLP) aynı anlama gelmemektedir. NLP, bilgisayar tarafından bir dildeki yapının çözümlenerek (fonetik yapı, dilbilgisi, sözdizimi ve anlamsallık olarak) kullanılması veya başka bir dile çevrilmesi anlamındadır. NLP çoğunlukla ASR konusunu da alt alan olarak içerebilmektedir. Sesli yanıt sistemleri, kişisel asistan servisi, metinden sese veya sesten metine çeviri, konuşma etiketleme

gibi kullanım alanları mevcuttur. Otomatik konuşmacı doğrulama (automatic speaker verification, ASV) sistemleri belirli kimlik bilgileri ve konuşma örneklerini girdi olarak almakta ve biyometrik tabanlı bir çoklu ortam sistemine giriş yapılmasını onaylayacak veya red edecek şekilde çalışmaktadır. Metine bağımlı veya metinden bağımsız ASV türleri mevcuttur. Bu tip sistemler yardımıyla çevrimiçi ödeme işlemleri yapılabilmekte, otomobil veya telefon kilitleme/açma sistemleri kullanılabilmektedir. Bunlara yapılan siber saldırılar ise aldatma saldırıları olarak bilinirler. İki temel türde aldatma saldırısı mevcuttur. Bunlar, önceki bir bilgiye dayanmayan değiştirilmiş/oluşturulmuş konuşma ile doğrudan sisteme yapılan saldırılar veya sistem bileşenlerinde değişikliğe neden olabilecek sistem seviyesi erişimle yapılan doğrudan olmayan (endirekt) biçimdeki saldırılardır (Qian ve ark., 2016).

Akustik sahne sınıflandırma (acoustic scene classification, ASC) hesaplamalı ses sahnesi analizi tabanlı olarak çeşitli ses kaynaklarından alınan ses sinyallerini (otobüs, klima, insanlar, kuşlar vb.) nerede bulunduğuna dair (hava alanı, otopark, ev, metro istasyonu vb.) çeşitli öntanımlı sınıflara eşleyerek bulmayı amaçlamaktadır (Zhou ve ark., 2018). Spektrogramların akustik sahnelerin zaman-frekans temsili olarak kullanıldığı bir biçimde çoğunlukla CNN gibi derin öğrenme mimari modellerinde bu tip veri temsilleri kullanılmaktadır. Zamansal ve konumsal bilgi spektrogramdan elde edilmekte, fakat frekans çözünürlüğü ile zaman çözünürlüğü arasında birbirleriyle çelişen durumlar olabilmektedir. Bu nedenle CNN modelindeki filtre biçimi ve sayısı gibi parametreler bu çözünürlüklerden etkilenmektedir (Zheng, Mo, Xing ve Zhao, 2015). Bunun yanı sıra CNN ve diğer RNN tipleri hem bireysel olarak hem de birlikte (CNN+LSTM gibi) sinerji oluşturacak biçimde kullanılabilmektedir.

Çevresel ses etiketleme (environmental audio tagging, EAT) ilgilenilen akustik sahnedeki belirli akustik olayların var olup olmadığının tahmin edilmesine uğraşmaktadır. Buna göre akıllı mobil cihazlar ile ses bölütleme, ses içerik sınıflandırma ve otomatik ses etiketleme yapılabilmekte, arkaplandaki gürültülerden ayıklanan ses, konuşma veya müzik için belirli etiketler belirlenebilmekte veya tahmin edilebilmektedir. Makine öğrenmesi veya derin öğrenme yoluyla bağlamsal bilgi içerikle beraber ilgili ilişkilerin farklı ses olayı sınıfları arasında ortaya çıkarılmasını da sağlamaktadır (Xu, Huang, Wang, Foster, Sigtia, Jackson ve Pumbley, 2017). Akustik olay tespiti (acoustic event detection, AED) konuşma olmayan (non-speech) akustik sinyallerin tespit edilip sınıflandırılması yoluyla sürekli bir akustik sinyalin işlenerek başlangıç ve bitiş süreleriyle ilişkilendirilmiş olay etiketlerinin bir dizisi haline çevrilmesini sağlamaktadır (Espi ve ark., 2015). Böylece ses olay tespiti sayesinde konuşma iletişim dökümü, ses sahnesinin çözümlenmesi, ses/konuşma iyileştirme (enhancement) ve ASR işlemleri gürültüden arındırılmış olması için gerekli işlemler de kullanılarak yapılabilmektedir.

Otomatik dil belirleme/tespit (automatic language identification, ALI) sayesinde verilen bir konuşma örneğinin hangi dile ait olduğunun otomatik olarak belirlenmesine çalışılmaktadır. Bu yolla insan-bilgisayar etkileşimi, sesli yanıt cihazları ve konuşmacı tanıma/doğrulama işlemleri farklı dildeki sesler için derin öğrenme kullanılarak yapılabilmektedir (Lopez-Moreno, Gonzalez-Dominguez, Martinez, Plchot, Gonzalez-Rodriguez ve Moreno, 2016). Ayrıca morfo-sözdizimsel (morfosyntactic) değişimlerle doğal dil damgalama (watermarking) şemaları kullanımı sayesinde çoklu ortam dokümanları da (özellikle ses parçaları göz önüne alındığında) daha güvenli hale getirilebilmektedir (Meral, Sankur, Özsoy, Güngor ve Sevinç, 2009). Bu bakış açısıyla ses parçalarının güvenliğinin sağlanması, konuşmacı doğrulamanın konuşma sentezlemeyi kullanan aldatma saldırılarına karşı daha gürbüz olmasının gerekliliği, ses/konuşmacı doğrulama (speaker verification) sistemlerine olan ihtiyacı artırmıştır. Böyle sistemlerde hem istatistiksel hem ses özniteliklerine dayanan bilgi harmanlanmakta hem de ses taklidi (voice mimicry) olup olmadığı ortaya çıkarılabilmektedir (Hautamäki, Kinnunen, Hautamäki ve Laukkanen, 2015; Khodabakhsh, Mohammadi ve Demiroğlu, 2017). Bu nedenlerle otomatik aldatma tespiti (automatic spoofing detection, ASD) ve konuşmacı doğrulama sistemleri önem kazanmakta ve derin öğrenme, alan adaptasyonu, anti aldatma (biyometrik güvenlik), sentetik konuşma tespiti konularında çalışmalar yapılmaktadır (Hanilçi, Kinnunen, Sahidullah ve Sizov, 2016; Todisco, Delgado ve Evans, 2017; Himawan, Villavicencio, Sridharan ve Fookes, 2019; Qian ve ark., 2016).

Aldatmayı tespit edip önlemek adına karmaşık ses ortamlarında ses iyileştirmesi yapılması önemli bir işlemdir. Bunun için birçok çalışma literatürdeki derin öğrenme mimari modelleri kullanılarak yapılmıştır (Li, Liu, Shi, Dong ve Cui, 2016; Kang ve ark., 2018; Samui ve ark., 2019). Ses iyileştirme sayesinde gürültü giderme (denoising) yapılması da çalışmalarda derin öğrenme mimari modellerinin kullanılmasıyla başarıyla yapılmıştır. Özellikle tayfsal (spektral) eşleştirme derin öğrenme

sayesinde daha temiz konuşmaya dair logaritmik büyüklük spektrogramının hesaplanmasında kullanılmıştır (Han, Wang, DeLiang, Woods, Merks ve Zhang, 2015).

Literatürdeki çalışmaları siber saldırıların yapılabildiği alanlarda ses işleme, ses olayı tespiti, derin öğrenme ile bu tip saldırıların tespit edilmesine olanak sunacak gürbüz ve yüksek başarımlı altyapıların oluşturulması özelinde inceleyecek olursak; ASR için oluşturulan bir akustik modeli için çift yönlü LSTM kullanarak kelime hata oranı ölçütünün değerini çok düşük olarak elde eden bir derin öğrenme çözümü Zeyer ve arkadaşları (2017) çalışmasında mevcuttur. Gizli Markov modeli (Hidden Markov model, HMM) ve çift yönlü LSTM kullanarak akustik model oluşturulup büyük ölçekli konuşma tanıma çözümü öneren Chen, Yan ve Huo (2015) yayını literatürdeki önemli bir çalışmadır. Ayrıca, 2017 yılındaki Qian, Evanini, Wang, Lee ve Mulholland (2017) çalışmasında, çocuk konuşmasının yerli (anadilde) veya yerli olmadığını ayırt edecek LSTM-RNN tabanlı bir ASR sistemi oluşturulmuştur. Gürültü giderme yoluyla konuşmayı ayrıştırmada derin RNN ile birlikte tekrarlayıcı-zamansal RBM kullanan bir derin öğrenme modeli ise konuşma iyileştirme işlemi için 2017 yılındaki Samui ve arkadaşları (2017) çalışmasında önerilmiştir.

Sharan ve Moir (2017) çalışmasında çeşitli ses öznitelikleri kullanılarak sesli gözetleme (surveillance) uygulamalarındaki gürültülü ortama karşın gürbüz (gürültüden daha az etkilenen) bir ses sınıflandırma yapmak için derin öğrenme mimari modeli olarak RBM katmanlarından oluşan bir yapı kullanılmıştır. 2019 yılındaki Samui, Chakrabarti ve Ghosh (2019) çalışmasında bulanık mantık tabanlı derin inanç ağı (fuzzy deep belief network, FDBN) modeli kullanılarak zaman-frekans maskeleme yoluyla danışmanlı bir konuşma iyileştirme işlemi yapılmıştır. Mel ölçeği ile ağırlıklandırılmış ortalama karesel hata ile zamansal ve tayfsal değişimleri dikkate alan bir amaç fonksiyonunu kullanan derin öğrenme tabanlı algoritma sayesinde konuşma iyileştirme Kang ve arkadaşları (2018) çalışmasında gerçekleştirilmiştir. DBN kullanarak konuşma iyileştirme yapılmasında iyileştirilmiş en küçük kareler uyarlanır filtreleme tekniğinin kullanıldığı Li ve arkadaşları (2016) çalışmasında, ağırlıklandırılmış gürültü giderici otokodlayıcıya göre daha iyi sonuçlar elde edilmiştir. 2016 yılındaki Qian ve arkadaşları (2016) çalışmasında derin öznitelikler kullanılarak otomatik aldatma tespiti yapılmıştır. Buna göre çalışmada önerilen DNN ve RNN tabanlı öznitelikler sayesinde denk hata oranı (EER) oldukça az çıkarak deneylerde başarılı sonuçlar elde edilmiştir. DNN olarak ileri beslemeli modeller kullanılırken, RNN için LSTM ve çift yönlü LSTM tipi ağ modelleri kullanılmıştır.

Zhou ve arkadaşları (2018) çalışmasında özellikle akustik sahne sınıflandırma (ASC) için MFCC özniteliklerini girdi olarak alan bir CNN ile, mel spektrogramını da bir diğer tarafta girdi olarak alan başka bir CNN arasında transfer öğrenme yoluyla görüntü ve ses özniteliklerinin birlikte kullanıldığı bir yapı ortaya konulmuştur. Özer, Özer ve Fındık (2018) çalışmasında gürültüye karşın gürbüz ses olay sınıflandırma yapmak için CNN kullanılmış, bunun için görüntü biçemindeki spektrogramlar girdi olarak alınıp doğrusal nicemlendirilmiş görüntüler haline getirilmiş ve böylece öznitelikler elde edilmiştir. Oldukça başarılı olan deneysel sonuçlara göre CNN ağının düşük gürültü oranlarına (dB cinsinden) ulaştığı görülmektedir. Akustik sahne sınıflandırmada (ASC) çoklu spektrogram füzyonunu ve sınıf etiket genişletmeyi kullanan CNN ağı tabanlı bir çözüm Zheng ve arkadaşları (2015) çalışmasında önerilmiştir. Buna göre bir CNN birden çok spektrogram üzerinde öznitelik elde etme amacıyla kullanılmış ve sonra bu öznitelikler füzyon yoluyla birleştirilip akustik sahneleri ifade etmek için kullanılmıştır. Sınıflar arası benzerlikler göz önüne alınarak sınıf etiketlerinin genişletilmesi sayesinde süper (üst) sınıflar oluşturulabilmiştir. Bu yolla yüksek başarımda (yüksek doğruluk oranı) sınflandırma sonuçları çalışmada elde edilmiştir.

Valenti, Squartini, Diment, Parascandolo ve Virtanen (2017) çalışmasında ASC işlemi için CNN kullanılmış, düşük seviyeli özniteliklerin ilişkili kullanımı sayesinde sınıflandırma başarımının iyi sonuçlar oluşturduğu görülmüştür. Huzaifah (Huzaifah, 2017) çalışmasında ses sınıflandırması ve konuşma tanıma için sinyal işlemede kullanılabilecek şekilde kısa dönemli Fourier dönüşümü (short-time Fourier transform, STFT), CQT, CWT dönüşümlerinden öznitelikler ve MFCC ile oluşturulan ses öznitelikleri elde edilmiştir. Bunun yanı sıra, sinir ağı yapısına bakıldığında ise katman sayısı olarak biri daha derin biri daha sığ yapıda olan CNN tipi derin öğrenme modellerine bu özniteliklerin verilmesiyle eğitim gerçekleştirilmiştir. Eğitim iki farklı veri kümesi üzerinden yapılarak sınıflandırma başarımları kıyaslanmış, ses karakteristiklerinin CNN mimarisinin sağladığı avantajlarla daha doğru tespit edildiği anlaşılmıştır.

Farhadipour, Veisi, Asgari ve Keyvanrad (2018) çalışmasında, beyin fonksiyonlarının bozulmasından kaynaklanan dizartri (dysarthric) tipi konuşma bozukluğuna sahip hastalardan alınan konuşma sinyallerinden sese dair (akustik) özniteliklerin elde edilmesinde DBN oto-kodlayıcı kullanımı önerilmiştir. İlgili çalışmada akustik öznitelikler olarak MFCC değerleri bir DBN tarafından kodlanmış ve bu öznitelikler derin ağ mimarisine uygun olarak çok katmanlı algılayıcıya (multilayer perceptron, MLP) aktarılmıştır. Elde edilen deneysel sonuçlara göre metne bağlı kalarak veya metne bağlı kalmaksızın yapılan konuşmaların olduğu durumlarda yüksek doğrulukta konuşmacı tanıma oranlarına ulaşılmıştır. Konuşmacı belirleme ile ilgili bu çalışmanın literatüre katkısı özniteliklerin gürbüz bir biçimde temsil edilmesini sağlayarak var olan sinir ağının başarımını daha iyi hale getirmiş olmasıdır.

Chung, Park ve Jung (2019) çalışmasında bir ileri beslemeli DNN tabanlı akustik modelin gürbüzlüğünü artırmak için sırayla ağırlıklandırılmış (rank-weighted) rekonstrüksiyon öznitelik yaklaşımı önerilmiştir. Buna göre, girdi özniteliği yapısı için sıraya ve geçersizliğe bakıldıktan sonra bir sırayla ağırlıklandırılmış rekonstrüksiyon yöntemi uygulanmakta ve böylece konuşma bilgisine dair bileşenler korunmakta, önemsiz bileşenlerin sayısı azaltılmaktadır. Görece uzun bir bağlam (context) penceresi kullanıldığında bu yaklaşım ile birlikte oluşturulan ileri beslemeli DNN ağ modeli oldukça düşük hata oranları elde edilmesiyle iyi sonuçlar ortaya koymaktadır. İlgili çalışmanın literatüre katkısı bu açıdan yüksek boyutla düşük sırada verilen ağırlıkların birleştirilmesine dayanan bir kazanımı DNN ağ yapısıyla tümleştirip konuşmacı/konuşmayı tanıma işlemini düşük hata oranında yüksek doğrulukla gerçekleştirmesidir.

Alisamir, Ahadi ve Seyedin (2018) çalışmasında tam bağlantılı DNN ağı kullanılarak İran dilindeki (Farsça) konuşmayı tanıma ve ham konuşma sinyali ile işlemde hem bu ağ modelini hem de bu ağ modelinin yanı sıra kapılı tekrarlayan birim (GRU) ve iletim sönümü (dropout) kullanımını fonem/konuşma tanıma için denemişlerdir. Deneylerde kullandıkları ilk veri kümesi FarsDat isimli Acemce konuşma veri kümesidir. Bu küme üzerinde MFCC özniteliklerini temel alarak uzun-kısa süreli bellek (LSTM) ve DNN ağını içeren bir tümleşik model denemişlerdir. Ayrıca aynı veri kümesi ile CNN, LSTM ve DNN tümleştirmesine dayanan bir modeli de denemişlerdir. Bu iki modelle de düşük hata oranları ile konuşma tanıma sonuçları elde etmişlerdir. Aynı modeller ve benzer yapıları farklı veri kümesinde de deneyerek yüksek başarım elde ettiklerini kanıtlamışlardır. Bu açıdan ilgili çalışmanın literatüre katkısı dilden bağımsız olabilecek bir yoldan konuşma tanımayı genelleştirebilecek DNN tabanlı bir çözüm öneriyor oluşudur.

Anand, Singh, Srivastava ve Lall (2019) çalışmasında oldukça kısa süreli konuşma ifadeleri kullanarak ve az sayıda konuşmacının verisinin bulunduğu durumlarda CNN ağı kullanımı sayesinde konuşmacı tanımanın spektrogram görüntüsünden elde edilen özniteliklerle yapılabileceğini gösterilmiştir. Ayrıca, kapsül ağı (capsule network) isimli CNN ağı ve DNN ağındaki skaler çıktı sinir düğümü yerine bir kapsülü (vektör çıktı sinir düğümü) kullanan yeni bir derin öğrenme mimari modelini konuşmacı tanımada kullanmışlardır. Deneyleri iki farklı veri kümesi üzerinden gerçekleştirerek, kapsül ağını iki farklı tipte CNN ağ modeliyle de karşılaştırmışlardır. Deneylerde önerilen yaklaşımın daha başarılı sonuçlar üretmediği fakat daha etkin (daha az parametre sayısı ile diğer ağlara yakın) sonuçlar ürettiğini göstermişlerdir. İlgili çalışmanın literatüre katkısı derin öğrenme modeli olan kapsül ağını etkin bir biçimde kısa süreli konuşmalara ve az sayıda konuşmacının olduğu durumlara uygulayarak diğer ağ modellerine yakın konuşma/konuşmacı tanıma sonuçları elde etmiş olmasıdır.

Kong, Xu, Sobieraj, Wang ve Plumbley (2019) çalışmasında zayıf biçimde etiketlenmiş veri üzerinden ses olayı tespiti yapmada zaman-frekans bölütlemesi yapılması için CNN ağı kullanılması ve sınıflandırma eşleştirmesinde de genel ağırlıklandırılmış sıra biriktirme yapılmasını içeren bir yaklaşım önerilmiştir. Buna göre, CNN ağı zaman-frekans bölütleme maskesini zamandaki olaylar için zayıf biçimde etiketlenmiş veriden öğrenebilmekte, bu iş için ise logaritmik mel spektrogram görüntüsünü kullanmaktadır. İlgili çalışmanın literatüre katkısı zayıf biçimde etiketlenmiş ses olayları verisiyle yüksek doğrulukta ses olayı tespitini yapan bir derin öğrenme yaklaşımını öneriyor oluşudur.

Chung, Nagrani ve Zisserman (2018) çalışmasında büyük ölçekli bir konuşmacı doğrulama veri kümesi (VoxCeleb2) oluşturularak buna dair milyondan fazla adet kısa süreli konuşma ifadelerinin CNN ağı sayesinde farklı koşullar altındaki deneylerde kullanılarak konuşmacı tanımanın yüksek doğrulukta yapılması sağlanmıştır. İlgili çalışmanın literatüre katkısı özellikle büyük ölçekli bir görsel-işitsel (audio-visual) konuşmacı tanıma veri kümesi sunuyor oluşu ve bu veri kümesi

üzerindeki konuşmacı tanıma işlemlerini CNN tipinde derin öğrenme ağ modelleri ile yapıyor oluşudur. Bu açıdan ilgili çalışmanın literatüre katkısı, kısa süreli konuşma ifadelerini temel alan bir derin öğrenme yaklaşımının önerilmesidir. Bu yaklaşımın deneylerde başarıyla uygulandığı da anlaşılmaktadır.

Etienne, Fidanza, Petrovskii, Devillers ve Schmauch (2018) çalışmasında konuşma/konuşmacı duygu tanımayı CNN ve çift yönlü LSTM ağ katmanlarından oluşan bir derin öğrenme mimari modeli kullanarak yapmışlardır. Ayrıca, veri artırma (data augmentation) kullanarak duygu tanıma başarımı iyileştirilmiştir. Ham halde verilen spektrogram görüntülerinden yüksek seviyeli öznitelikler CNN katmanları ile elde edilirken, uzun süreli zamansal bağımlılıklarsa LSTM katmanları ile çift yönlü olarak ortaya çıkartılmıştır. İlgili çalışmanın literatüre katkısı CNN ve çift yönlü LSTM ile oluşturulan melez mimarinin konuşmacı duygusu tanımada başarılı bir şekilde kullanıldığı bir yaklaşımın önerilmesidir.

Morfi ve Stowell (2019) çalışmasında ses olayı tespitini bir tane sınıf üzerinden yapacak ve ses etiketlemesi oluşturacak iki ayrı işlemi derin öğrenme mimari modelleri kullanacak şekilde önermişlerdir. Bu işlemler sayesinde ses içeriğinin dökümünü (audio transcription) çoklu işlemlere atanmış olarak düşük kaynaklı veri kümelerinde eğitim başarımını artırmayı hedeflemişlerdir. Bu iki işlem sırasıyla "*Ne zaman*" (When) ve "*Kim*" (Who) şeklinde isimlendirilmiş ve birbirinden farklı iki boyutlu (2D) CNN ağları ile kurulan yapıda işlemlerdir. "*Ne zaman*" isimli işlem için önerilen derin sinir ağı modeli CNN ve çift yönlü GRU katmanları içerirken, "*Kim*" isimli işlem için önerilen model ise sadece CNN katmanlarından oluşmaktadır. Bu CNN katmanları; evrişim, biriktirme ve yığın normalizasyonu biçimindeki katmanlardır. İlgili çalışmanın literatüre katkısı ilk işlem ile ses olayı tespitini "*Ne zaman*" isimli ağ modeliyle yapması, ikinci işlem ile ses etiketlemeyi "*Kim*" isimli ağ modeliyle yaparak yüksek başarımlı sonuçların elde edilmesini sağlamasıdır. Bu iki işleme ait iki ağ modeli ayrı ayrı ve birlikte birçok deneyde kullanılmıştır.

Zazo, Lozano-Diez, Gonzalez-Dominguez, Toledano ve Gonzalez-Rodriguez (2016) çalışmasında LSTM ağ modeli kullanılarak otomatik dil belirleme işlemi çeşitli akustik (ses) öznitelikler kullanılarak yapılmıştır. Deneylerde oldukça kısa süreli konuşma ifadeleri içeren veri kümeleriyle sekiz farklı dilin ayırt edilmesinde LSTM ağının sağladığı kazanımlar ve başarım sonuçları ortaya konulmuştur. Deney sonuçlarına göre sistemin konuşulan dili belirleme doğruluğu klasik yöntemlere göre oldukça yüksek olurken, denk hata oranı ise oldukça düşük olarak elde edilmiştir. İlgili çalışmanın literatüre katkısı otomatik dil belirlemede LSTM tipi ağ modelinin kullanılmasıyla kısa süreli konuşma ifadelerinden bile herhangi bir dili yüksek doğrulukla otomatik olarak belirleyebildiğinin gösterilmesidir.

Jayalakshmi, Chandrakala ve Nedunchelian (2018) çalışmasında klasik makine öğrenmesi yöntemlerinden olan destek vektör makinesi (support vector machine, SVM) sınıflandırıcının kullanıldığı bir ses olayı sınıflandırma yaklaşımında özlü bir genel istatistiksel öznitelik tabanlı temsil kullanılarak MFCC tabanlı akustik (ses) özniteliklerinin daha etkin kullanımı önerilmiştir. Bu yaklaşımın sınıflandırma başarımı gizli Markov modeli (hidden Markov model, HMM) ve Gauss karışım modeli (Gaussian Mixture Model, GMM) yaklaşımlarıyla çeşitli ses veri kümeleri üzerinden kıyaslandığında genel istatistiksel öznitelik tabanlı yaklaşımın literatüre katkı olarak daha yüksek başarım ile çok değişkenli biçimde verilen değişken uzunluktaki akustik olayları daha düşük (indirgenmiş) boyutta ifade edebildiği ortaya çıkmıştır. Ayrıca sınıflandırma başarımının diğer yaklaşımlara gore iyileştirilebilir olduğu da ifade edilmiştir.

Literatür taramasından da anlaşıldığı üzere birçok farklı alanda birçok derin öğrenme mimari modeli farklı problemlerin çözümünde oldukça verimli olarak kullanılmış ve yüksek başarımda sınıflandırma veya tespit sonuçları elde edilmiştir. Tablo 1'den görülebileceği gibi 2015 ilâ 2019 yılları arasında yapılan literatürdeki 36 adet çalışma yoğunlukla derin öğrenme modellerinden DNN, CNN, LSTM, RBM ve DBN kullanımına dayanan, otomatik sahne sınıflandırma, otomatik konuşma/konuşmacı tanıma/sınıflandırma, otomatik ses olayı tespiti/sınıflandırma, anti sesle aldatma (anti spoofing) ve ses iyileştirme gibi çalışmaları içermektedir. Tablo 1'de en çok yayın mecrası olarak uluslararası dergilerin tercih edildiği ve mevcut yayınların çoğunluğunda CNN tipinde derin öğrenme modeli kullanıldığı görülmektedir. Ayrıca, derin öğrenme haricinde çokça tercih edilen makine öğrenmesi yöntem ve tekniklerinden Gauss karışım modeli (Gaussian Mixture Model, GMM), doğrusal regresyon ve destek vektör makinesi (support vector machine, SVM) ile yapılan beş adet çalışmaya da Tablo 1'de yer verilmiştir. Tablo 1'de incelenen çalışmaların literatüre katkıları kısa cümlelerle özetlenmiş olarak Tablo 2'de verilmiştir.

Tablo 2'de incelenen çalışmalar ilgili çalışma alanı kategorilerine göre gösterilmektedir.

Tablo 1

Yıllara ve kategorilere göre literatürdeki incelenen çalışmalar

Yayın ismi	Yılı / Yayın mecrası	Çalışma alanı	Kullanılan model/teknik
Konuşma İyileştirme			
Kang ve arkadaşları (2018)	2018/ Uluslararası dergi	Konuşma iyileştirme	Derin öğrenme (DNN)
Samui ve arkadaşları (2019)	2019/ Uluslararası dergi	Konuşma iyileştirme	Derin öğrenme (Bulanık DBN)
Li ve arkadaşları (2016)	2016/ Uluslararası dergi	Konuşma iyileştirme, gürültü sınıflandırma	Derin öğrenme (DNN)
Qian ve arkadaşları (2017)	2017/ Uluslararası konferans	Konuşma iyileştime	Derin öğrenme (Çift yönlü LSTM+RNN)
Han ve arkadaşları (2015)	2015/ Uluslararası konferans	Konuşmada yankılanma giderme, konuşmada gürültü giderme	Derin öğrenme (DNN)
Sahne/Çevresel Ses Sınıflandırma			
Zhou ve arkadaşları (2018)	2018/ Uluslararası sempozyum	Otomatik sahne sınıflandırma	Derin öğrenme (Transfer öğrenme ve CNN)
Zheng ve arkadaşları (2015)	2015/ Çevrimiçi uluslararası arXiv yayını	Otomatik sahne sınıflandırma	Derin öğrenme (CNN)
Valenti ve arkadaşları (2017)	2017/ Uluslararası konferans	Otomatik sahne sınıflandırma	Derin öğrenme (CNN)
Salamon ve Bello (2017)	2017/ Uluslararası dergi	Çevresel ses sınıflandırma	Derin öğrenme (CNN)
Konuşmacı Doğrulama			
Himawan ve arkadaşları (2019)	2019/ Uluslararası dergi	Konuşmacı doğrulama	Derin öğrenme (CNN)
Todisco ve arkadaşları (2017)	2017/ Uluslararası dergi	Otomatik konuşmacı doğrulama	Makine öğrenmesi (GMM)
Hautamaki ve arkadaşları (2015)	2015/ Uluslararası dergi	Konuşmacı doğrulama	Makine öğrenmesi (Evrensel arkaplan modelli GMM)
Ses Olayı Sınıflandırma			
Özer ve arkadaşları (2018)	2018/ Uluslararası dergi	Ses olayı sınıflandırma	Derin öğrenme (CNN)
Espi ve arkadaşları (2015)	2015/ Uluslararası dergi	Otomatik ses olayı tespiti	Derin öğrenme (CNN)
Sharan ve Moir (2017)	2017/ Uluslararası dergi	Otomatik ses olayı sınıflandırma	Derin öğrenme (DNN)
Morfi ve Stowell (2019)	2019/ Uluslararası dergi	Ses olayı tespiti ve Ses etiketleme	Ses olayı tespti için Derin öğrenme (CNN+Çift yönlü GRU) Ses etiketleme için Derin öğrenme (CNN)
Kong ve arkadaşları (2019)	2019/ Uluslararası dergi	Ses olayı tespiti	Derin öğrenme (CNN)
Jayalakshmi ve arkadaşları (2018)	2018/ Uluslararası dergi	Ses olayı sınıflandırma	Makine öğrenmesi (SVM)
Ses Olayı Etiketleme			
Xu ve arkadaşları (2017)	2017/ Uluslararası dergi	Çevresel ses etiketleme	Derin öğrenme (Küçültülen DNN)
Konuşma/Konuşmacı Tanıma/Ayrıştırma			
Zeyer ve arkadaşları (2017)	2017/ Uluslararası konferans	Konuşma tanıma	Derin öğrenme (Çift yönlü LSTM)
Samui ve arkadaşları (2017)	2017/ Uluslararası konferans	Konuşma ayrıştırma	Derin öğrenme (Tekrarlayıcı zamansal RBM)
Sainath ve arkadaşları (2015)	2015/ Uluslararası dergi	Konuşma tanıma	Derin öğrenme (CNN)
Farhadipour ve arkadaşları (2018)	2018/ Uluslararası dergi	Konuşmacı tanıma/belirleme	Derin öğrenme (DBN+MLP)
Alisamir ve arkadaşları (2018)	2018/ Uluslararası dergi	Konuşmacı/Konuşma tanıma	Derin öğrenme (DNN+LSTM)
Anand ve arkadaşları (2019)	2019/ Çevrimiçi uluslararası arXiv yayını	Konuşmacı/Konuşma tanıma	Derin öğrenme (CNN ve Kapsül ağı)
Chung ve arkadaşları (2018)	2018/ Çevrimiçi uluslararası arXiv yayını	Konuşmacı tanıma	Derin öğrenme (CNN)
Etienne ve arkadaşları (2018)	2018/ Uluslararası dergi	Konuşma/Konuşmacı duygusu tanıma	Derin öğrenme (CNN+Çift yönlü LSTM)
Konuşulan dili belirleme			
Lopez-Moreno ve arkadaşları (2016)	2016/ Uluslararası dergi	Otomatik konuşulan dil belirleme	Derin öğrenme (DNN)
Zazo ve arkadaşları (2016)	2016/ Uluslararası dergi	Otomatik konuşulan dil belirleme	Derin öğrenme (LSTM)
Sesle Aldatma Tespiti/Doğrulama			
Qian ve arkadaşları (2016)	2016/ Uluslararası dergi	Sesle aldatma tespiti	Derin öğrenme (DNN ve RNN)
Hanilçi ve arkadaşları (2016)	2016/ Uluslararası dergi	Sentetik sesi/konuşmayı tespit etme	Makine öğrenmesi (GMM)
Khodabakhsh ve arkadaşları (2017)	2017/ Uluslararası dergi	Aldatıcı sesi doğrulama	Makine öğrenmesi (Doğrusal regresyon)
Ses Karakteristiği Belirleme/Akustik Model Oluşturma			
Huzaifah (2017)	2017/ Çevrimiçi uluslararası arXiv yayını	Ses olayı/konuşma sınıflandırma için ses karakteristiği tespiti	Derin öğrenme (CNN)
Bhatt ve arkadaşları (2018)	2018/ Uluslararası çalıştay	Çok kanallı derin akustik özniteliklerin birleştirilmesi	Derin öğrenme (CNN)
Chen ve arkadaşları (2015)	2015/ Uluslararası konferans	Akustik model oluşturma	Derin öğrenme (Derin çift yönlü LSTM)
Chung ve arkadaşları (2019)	2019/ Uluslararası dergi	Akustik model oluşturma	Derin öğrenme (İleri beslemeli DNN)

Tablo 2

İncelenen çalışmaların literatüre katkıları.

İncelenen çalışmanın literatüre katkısı	
Konuşma İyileştirme	
Kang ve arkadaşları (2018)	DNN ağında kullanılacak mel ölçeği ile ağılıklandırılmış ortalama karesel hata ile zamansal ve taysal değişimleri kapsayan bir amaç fonksiyonu konuşma iyileştirme için önerilmiştir.
Samui ve arkadaşları (2019)	Zaman-frekans maskeleme yoluyla danışmanlı bir bulanık mantık tabanlı DBN konuşma iyileştirme için önerilmiştir.
Li ve arkadaşları (2016)	DBN için en küçük kareler uyarlanır filtreleme tekniğinin kullanımı konuşma iyileştirme için önerilmiştir.
Qian ve arkadaşları (2017)	Çocuk konuşmasının yerli (anadilde) veya yerli olmadığını uzun bir bağlamsal yapıdan başarıyla ayırt edebilmeyi sağlayan LSTM-RNN tabanlı bir ASR sistemi biçimindeki yapı önerilmiştir.
Han ve arkadaşları (2015)	Daha temiz konuşmanın elde edilmesinde tayfsal eşleştirme yoluyla derin öğrenme sayesinde spektrogram oluşturulması önerilmiştir.
Sahne/Çevresel Ses Sınıflandırma	
Zhou ve arkadaşları (2018)	CNN ağının MFCC öznitelikleri ile eğitiminden edinilen öğrenim bilgisinin öğrenme transferiyle diğer bir CNN ağının mel spektrogram ile eğitimi sürecine bütünleştirilmesi önerilmiştir.
Zheng ve arkadaşları (2015)	Otomatik sahne sınıflandırmada CNN ağının kullandığı filtre biçimi ve sayısı göz önüne alınarak çoklu spektrogram füzyonu ve sınıf etiket genişletme yaklaşımı önerilmiştir.
Valenti ve arkadaşları (2017)	Düşük seviyeli özniteliklerin kullanımı ile CNN ağının otomatik sahne sınıflandırma başarımını artırdığı bir yaklaşım önerilmiştir.
Salamon ve Bello (2017)	Küçük veri kümeleri için test başarımının artması adına veri artırma işlemi sayesinde başarımı yüksek bir derin CNN ağı kullanarak çevresel ses sınıflandırma yapılmasına dair yaklaşım önerilmiştir.
Konuşmacı Doğrulama	
Himawan ve arkadaşları (2019)	Konuşmacı doğrulama sistemlerinde anti aldatma için derin öğrenme tabanlı alan adaptasyonu önerilmiştir.
Todisco ve arkadaşları (2017)	Sabit Q kepstral katsayıları kullanıp otomatik konuşmacı doğrulamada makine öğrenmesi tabanlı GMM modeli kullanımı önerilmiştir.
Hautamaki ve arkadaşları (2015)	Ses taklitinin anlaşılmasıyla otomatik konuşmacı doğrulama yapılabilmesi için makine öğrenmesi tabanlı evrensel arkaplan kullanan GMM önerilmiştir.
Ses Olayı Sınıflandırma	
Özer ve arkadaşları (2018)	Düşük gürültü oranlarına ulaşan CNN ağı için spektrogram görüntüsünden öznitelik elde edilmesiyle gürbüz bir ses olay sınıflandırma yaklaşımı önerilmiştir.
Espi ve arkadaşları (2015)	Spektro-zamansal yerelliğin akustik (ses) olayı tespitinde derin öğrenme yaklaşımıyla işlenmesiyle tüm sinyalin olay etiketlerinin bir dizisi haline getirilmesi yaklaşımı önerilmiştir.
Sharan ve Moir (2017)	RBM katmanları ile oluşturulan derin öğrenme modeli ile gürültülü ortama karşın gürbüz bir ses sınıflandırma yapılmasında çeşitli ses özniteliklerini içeren bir yaklaşım önerilmiştir.
Morfi ve Stowell (2019)	Ses olayı tespitini CNN ve çift yönlü GRU katmanları içeren "*Ne zaman*" isimli ağ modeliyle ve ses etiketlemeyi ise sadece CNN katmanlarından oluşan "*Kim*" isimli ağ modeliyle yapan bir yaklaşım önerilmiştir.
Kong ve arkadaşları (2019)	Zayıf biçimde etiketlenmiş ses olayları verisiyle yüksek doğrulukta ses olayı tespitini yapan bir yaklaşım önerilmiştir.
Jayalakshmi ve arkadaşları (2018)	Çok değişkenli biçimde verilen değişken uzunluktaki akustik (ses) olayların daha düşük (indirgenmiş) boyutta ifade edilmesiyle SVM kullanarak bu olayları sınıflandıran bir yaklaşım önerilmiştir.
Ses Olayı Etiketleme	
Xu ve arkadaşları (2017)	Çevresel ses etiketleme için derin öğrenme ile danışmansız biçimde öznitelik öğrenilmesiyle bağlamsal bilgiye dayanarak farklı ses olayları arasındaki ilişkilerin ortaya çıkartıldığı bir yaklaşım önerilmiştir.
Konuşma/Konuşmacı Tanıma/Ayrıştırma	
Zeyer ve arkadaşları (2017)	Çift yönlü LSTM ağı kullanılarak kelime hata oranının oldukça düşük olduğu bir derin öğrenme yaklaşımı önerilmiştir.
Samui ve arkadaşları (2017)	Konuşma iyileştirmede konuşma sinyalinden gürültüyü ayrıştırmada derin RNN ile tekrarlayıcı-zamansal RBM kullanan bir yaklaşım önerilmiştir.
Sainath ve arkadaşları (2015)	Sürekli yapıda verilen büyük ölçekli konuşma tanıma için CNN ağı ve doğrultulmuş doğrusal birim (rectified linear unit, ReLU) kullanımını temel alan bir derin öğrenme yaklaşımı önerilmiştir.
Farhadipour ve arkadaşları (2018)	Konuşmacı belirlemede DBN sayesinde özniteliklerin daha gürbüz bir biçimde temsil edilmesini sağlayarak MLP ağının başarımını daha iyi hale getiren bir derin öğrenme yaklaşım önerilmiştir.
Alisamir ve arkadaşları (2018)	Konuşulan dilden bağımsız olacak şekilde konuşma tanımayı genelleştirebilecek DNN tabanlı bir yaklaşım önerilmiştir.
Anand ve arkadaşları (2019)	Kapsül ağının etkin biçimde kısa süreli konuşmalara ve az sayıda konuşmacı olan durumlara uygulanması önerilmiştir.
Chung ve arkadaşları (2018)	Kısa süreli konuşma ifadelerini işleyerek yüksek başarımlı CNN ağı tabanlı bir derin öğrenme yaklaşımı önerilmiştir.
Etienne ve arkadaşları (2018)	CNN ve çift yönlü LSTM ile oluşturulan melez mimarinin konuşmacı duygusu tanımada kullanımı önerilmiştir.
Konuşulan dili belirleme	
Lopez-Moreno ve arkadaşları (2016)	Derin ileri beslemeli ağları kullanarak konuşma sinyalinden ayrıştırıcı dil bilgisini öğrenmede DNN tabanlı akustik model oluşturulması önerilmiştir.
Zazo ve arkadaşları (2016)	Otomatik dil belirlemede LSTM tipi ağ modelinin kullanılmasıyla kısa süreli konuşma ifadelerinden bile herhangi bir dili yüksek doğrulukla belirleyen bir yaklaşım önerilmiştir.
Sesle Aldatma Tespiti/Doğrulama	
Qian ve arkadaşları (2016)	Otomatik aldatma tespiti için DNN ve RNN tabanlı öznitelikleri kullanan denk hata oranı düşük bir yaklaşım önerilmiştir.
Hanilçi ve arkadaşları (2016)	Gürültü ortamda sentetik sesin (konuşma) tespitinde makine öğrenmesi tabanlı GMM ile aldatma tespitinin yapılabildiği bir yaklaşım önerilmiştir.
Khodabakhsh ve arkadaşları (2017)	Aldatıcı sesi doğrulamada sınırlı adaptasyon verisi kullanarak istatistiksel konuşma sentezleme yapılmasında makine öğrenmesi tabanlı doğrusal regresyon kullanımı önerilmiştir.
Ses Karakteristiği Belirleme/Akustik Model Oluşturma	
Huzaifah (2017)	Ses iyileştirme için CNN ağı ile çeşitli ses öznitelikleri kullanarak ses karakteristiğini belirleyen bir yaklaşım önerilmiştir.
Bhatt ve arkadaşları (2018)	Çok kanallı derin öğrenme mimarisi (CNN) kullanılarak akustik öznitelik füzyonuna dayanan bir yaklaşım önerilmiştir.
Chen ve arkadaşları (2015)	Akustik model oluşturulmasında HMM ve derin çift yönlü LSTM ağını kullanan büyük ölçekli konuşma tanıma için bir yaklaşım önerilmiştir.
Chung ve arkadaşları (2019)	Konuşma/Konuşmacıyı tanımada yüksek boyutla düşük sırada verilen ağırlıkların birleştirilmesine dayanan bir kazanımı DNN ağıyla tümleştiren bir yaklaşım önerilmiştir.

Bu alandaki çalışmalara bakıldığında; siber güvenlik için gerekli/alınacak tedbirlere taban oluşturacak biçimde ham verilerin işlenerek kategorize edilmesi ve ayrıştırılması işlemlerinde yüksek doğruluk başarısı sağlanması adına derin öğrenme ve makine öğrenmesinin yıllar geçtikçe vazgeçilmez bir hale gelmekte olduğu anlaşılmaktadır.

4. TARTIŞMA VE SONUÇLAR

Başta biyometrik yetkilendirme olmak üzere çeşitli alanlardaki siber saldırı gün geçtikçe artmakta ve buna karşın tedbir alınması önem kazanmaktadır. Bu bakış açısına göre; ses işleme çoklu ortam bileşenlerinin biri veya birçoğunu birlikte içeren senaryolarda verilen verilerin işlenmesi, anlamlandırılması ve buna dayanarak belirli kararların verilmesinde sıklıkla kullanılmaktadır. Ayrıca, makine öğrenmesine veya derin öğrenmeye son yıllarda olan rağbetin temel nedeni, yarı otomatik veya tam otomatik sınıflandırma işleminin danışmanlı (denetimli) veya danışmansız (denetimsiz) bir yoldan yapılmasıyla başarımı yüksek sınıflandırma, tespit ve doğrulama oranlarına ulaşılmasındandır. Özellikle bu çalışmamız özelinde değerlendirildiğinde; sese dayalı siber saldırıların tespit edilmesi, önlenebilmesi ve saldırıya karşın telafi, sorun giderme yordamlarının işletilebilmesi için öncelikle doğru biçimde ses analizi, ses olay tespiti, ses sahnesi sınıflandırması, gürültü temizleme, ses iyileştirme, ses bölütleme gibi işlemlerin yapılması gerekmektedir. Buna göre, derin öğrenmenin anlamsal çıkarımların yüksek seviyeli özniteliklere olan ihtiyacını düşük ve orta seviyeli özniteliklerden yola çıkarak öznitelik öğrenmeye dayalı bir biçimde doldurmaya çalışması; sınıflandırma, tespit ve tahminleme başarımının yüksek oluşu, önümüzdeki senelerde büyük veri, akıllı veri gibi verilerin kullanılmasıyla verinin hacimsel ölçüm birimi olarak büyük ölçeklere ulaştığı (zetabyte boyutları da dahil) problemlerin çözümünde de tercih edilir olacağını göstermektedir.

4.1. Derin Öğrenmenin Sağladığı Kazanımlar

Derin öğrenme insan algısıyla makine algısını görece birbirine yaklaştırmayı ve bu iki olgu arasındaki anlamsal boşluğu kapatmayı amaçlayan bir yaklaşımdır. Makine öğrenmesinden yola çıkılarak oluşturulan bir yaklaşım olduğu için örneğin danışmanlı öğrenmede de olduğu gibi yeterince eğitim verisinin verilmesi durumunda (insan uzman tarafından etiketlemesi yapılmış veri) insan tarafından elde edilebilecek algısal veriye dayanan tahmin/sınıflandırma sonucuna oldukça yakın (oldukça az hataya sahip) sonuçlara ulaşılması mümkündür. Derin öğrenmeye yakından baktığımızda, her verinin öncelikle vektörlere dönüştürüldüğü ve girdi uzayı ile çıktı uzayının birbirine bir fonksiyon üzerinden eşleştirilmeye çalışıldığı bir öğrenme tipini temel alan derin öğrenmenin elde etmek istediği anlamsallığı böylece karşılıklı ilişkilerden (uzaylar arası ilişkileri tutan vektörler sayesinde) hesaplamalı olarak oluşturabildiği görülmektedir. Temsili öğrenme ile karmaşıklık azaltılarak daha alt seviyeden ilişkiler daha üst seviyeden anlamlı ilişkilere ve böylece sınıflandırma/tespit veya tahmin nihai kararlarına dönüştürülebilmektedir. Sinir ağı modeli haline getirilmiş hesaplamalı çizgeler üzerinde bu tip ilişkilerin kodlanarak tutulması ve işlenmesi bu anlamsallığın sistematik olarak oluşturulabilmesine imkan sunmaktadır (Chollet, 2017). Bu bakış açısıyla derin öğrenme sayesinde, etkin veri temsili yapılabilmekte, girdi uzayı ile çıktı uzayı arasında güçlü bir ilişki modeli oluşturulabilmekte ve verilen sinir ağı modelinin parametreler aracılığıyla ayarları esnek olarak değiştirilebilmektedir.

Çalışmamızda incelediğimiz, CNN tabanlı çalışmalarda yüksek seviyeden anlamsal bilgiyi ses spektrogramı gibi düşük seviyeden bilgi (öznitelikler) içeren bir veriden elde etmenin klasik makine öğrenmesi yöntem ve modellerine göre hem daha kolay hem de daha yüksek başarıma sahip olduğu görülmektedir. Örneğin, Özer ve arkadaşları (2018) çalışmasında görüntü biçeminde ses spektrogramları girdi olarak alınarak nicemlendirilmiş halde öznitelik vektörü biçemine getirildikten sonra sinir ağının eğitim aşamasında kullanılmıştır. Zheng ve arkadaşları (2015) çalışmasında da benzer bir yaklaşımla akustik sahneleri ifade etmek için ses spektrogramı girdi olarak alınarak öznitelik vektörüne çevirildikten sonra CNN ağının eğitiminde kullanımı tercih edilmiştir. Ses verisinin zaman serisi gibi uzun bir bağlam bilgisine (context) sahip olduğu biçemde sinir ağına verildiği çalışmalarda tekrarlayıcı sinir ağları (RNN) kullanılarak hem uzamsal hem de zamansal ilişkiler girdi uzayı ile çıktı uzayı arasında düşük hata oranıyla elde edilebilmiştir. Çalışmalarda özellikle LSTM kullanımı dikkat çekmektedir. Qian ve arkadaşları (2017) çalışmasında RNN ağlarının bir tipi olan LSTM ağı tabanlı otomatik konuşma tanıma sistemi oluşturularak oldukça uzun bağlam bilgisine sahip konuşmanın ana dilde (yerli) bir konuşma olup olmadığı ayırt edilebilmiştir. Ayrıca, çift yönlü LSTM kullanılarak akustik model oluşturulması sayesinde büyük ölçekli (uzun bir bağlam bilgisi içeren) konuşma tanıma işlemi de Chen ve arkadaşları (2015) çalışmasında yapılmıştır.

Birçok aktivasyon fonksiyonu (sigmoid, softmax vb.) ile çalışabilen derin öğrenme modelleri sayesinde girdi uzayı ile çıktı uzayı arasındaki doğrusal olmayan karmaşık ilişkiler rahatlıkla kurulabilmektedir. Bunu sağlarken verilen sinir ağının böyle güçlü bir ilişki modelini oluşturmasını kolaylaştıracak üstün parametre (hyper parameter) kümesinin de doğru değerlerle belirlenmesi gerekmektedir. Bu açıdan üstün parametrelerin değerlerine dair ayarlar istendiğinde esnek olarak değiştirilebilmelidir. Literatürdeki çalışmalarda üstün parametrelerin eniyilenmesi (optimization) adına ADAM, Nestorov momenti ve RMSProp gibi eniyileme algoritmaları sıklıkla kullanılmaktadır. Bu sayede derin öğrenmenin daha az hata oranıyla daha yüksek sınıflandırma/tespit ve tahmin başarımı değerlerine ulaşması ve klasik makine öğrenmesine göre üstün yönlerini ortaya çıkarması mümkün olmaktadır.

Konuya daha üstten bakacak olursak; derin öğrenmenin ses olayı tespiti, ses sahnesi sınıflandırma, konuşma/konuşmacıyı tanıma ve sesle aldatma tespiti gibi konularda klasik makine öğrenmesine nazaran sağladığı kazanımlar arasında; veri temsiline dayalı öğrenme yaparak elle öznitelik belirleme yerine otomatik öznitelik elde etmede sağladığı avantajlar ve ağ modelinin parametreler üzerinden esnek olarak (katman sayısı artışı, eniyilemeye olan yatkınlık vb.) yapılabilen model modifikasyonu sayılabilir. Ayrıca, derin öğrenmenin verinin içerisinde gömülü ilişkilerin güçlü bir biçimde ortaya çıkarılarak anlamsallığın derecesinin artırılmasına sunduğu katkı en önemli kazanımlarından birisidir.

4.2. Güncel Yönelimler

Derin öğrenmenin ses analizinde kullanımı önümüzdeki yıllarda sesli asistan ve yanıt sistemlerinin tam otomatik olarak hayatımıza girmesiyle insan bilgisayar etkileşimindeki fiziksel girdi cihazlarının (klavye, fare ve dokunmatik ekran vb.) terkedilerek sesli ve görsel iletişime dayalı yazılım geliştirme, sesli yönlendirmeyle bilgisayar ve akıllı telefon kullanımına ve sanayide endüstriyel tasarım, ürün üretimi gibi alanların içeriklerinde zenginleşmeye de neden olacaktır. Sanal gerçeklik ve artırılmış gerçeklik alanları ile etkileşimli olarak ses analizine dayalı derin öğrenme uygulamalarının tercih edilebilme potansiyeli önümüzdeki günler için giderek artmaktadır.

Teknoloji geliştikçe, donanım imkânları artmaktadır. Özellikle klasik seri bilgisayarlar yerini günümüzde büyük çaplı paralel (eşzamanlı) veya dağıtık (distributed) çalışan bilgisayarlara/programlara, onlar da ileriki senelerde kuantum bilgisayarlara bırakacaktır. Bu açıdan çoklu ortam verilerinin yönetilebilirliği, depolanması, aktarılması ve işlenilmesi, bu yolla var olan siber saldırıların hızlı ve daha az maliyetle tespiti ve bertaraf edilmesi ülkemiz adına ve evrensel olarak gün geçtikçe daha çok önem arz etmektedir. İleriki çalışmalara yön verecek güncel araştırmanın şu anki yönelimi (trend), derin öğrenme/makine öğrenmesi ve kuantum bilgisayarların/hesaplamanın sinerjisi üzerine kurulan çok disiplinli çalışmalardır.

Finansal Destek: Yazar bu çalışma için finansal destek almamıştır.

KAYNAKLAR

Alisamir, S., Ahadi, S. M., & Seyedin, S. (2018). An end-to-end deep learning model to recognize Farsi speech from raw input. In *Proceedings of IEEE 4th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)* (pp. 1–5). Tehran, Iran: IEEE. http://dx.doi.org/10.1109/ICSPIS.2018.8700538

Anand, P., Singh, A. K., Srivastava, S., & Lall, B. (2019). Few shot speaker recognition using deep neural networks. *Electrical Engineering and Systems Science, Audio and Speech Processing(eess.AS), ArXiv.* 1–5. Retrieved from https://arxiv.org/abs/1904.08775

Babaee, E., Anuar, N. B., Wahab, A. W. A., Shamshirband, S., & Chronopoulos, A. T. (2017). An overview of audio event detection methods from feature extraction to classification, *Applied Artificial Intelligence, 31*(9–10), 661–714. http://dx.doi.org/10.1080/08839514.2018.1430469.

Bhatt, G., Gupta, A., Arora, A., & Raman, B. (2018). Acoustic features fusion using attentive multi-channel deep architecture. *Proceedings of CHIME 2018 Workshop on Speeech Processing in Everyday Environments*, Hyderabad, India, 30–34. http://dx.doi.org/10.21437/CHiME.2018-7

Boddapati, V., Petef, A., Rasmusson, J., & Lundberg, L. (2017). Classifying environmental sounds using image recognition networks. *Procedia Computer Science, 112*, 2048–2056. http://dx.doi.org/10.1016/j.procs.2017.08.250.

Chen, K., Yan, Z-J., & Huo, Q. (2015). Training deep bidirectional LSTM acoustic model for LVCSR by a context-sensitive-chunk BPTT approach. In *Proceedings of the INTERSPEECH 2015 16th Annual Conference of the International Speech Communication Association*: Vol 1-5 (pp.3600–3604). Dresden, Germany: ISCA archive. Retrieved from https://www.isca-speech.org/archive/interspeech_2015/i15_3600.html

Chollet, F. (2017). *Deep learning with python*. Newyork, NY: Manning Publication.

Chung, H., Park, J. G., & Jung, H.-Y. (2019). Rank-weighted reconstruction feature for a robust deep neural network-based acoustic model. *ETRI Journal*, *41*(2), 235–241. http://dx.doi.org/10.4218/etrij.2018-0189

Chung, J. S., Nagrani, A., & Zisserman, A. (2018). VoxCeleb2: Deep speaker recognition. *Computer Science, Sound (cs.SD), Electrical Engineering and Systems Science, Audio and Speech Processing(eess.AS). ArXiv.* 1–6. Retrieved from https://arxiv.org/abs/1806.05622v2.

Çakır, E. (2019). *Deep neural networks for sound event detection*. (Doctoral Dissertation, Tampere University, Finland). Retrieved from https://tutcris.tut.fi/portal/files/17626487/cakir_12.pdf

Espi, M., Fujimoto, M., Kinoshita, K., & Nakatani, T. (2015). Exploiting spectro-temporal locality in deep learning based acoustic event detection. *EURASIP Journal On Audio Speech And Music Proccessing, 2015*(26), 1–12. http://dx.doi.org/10.1186/s13636-015-0069-2.

Etienne, C., Fidanza, G., Petrovskii, A., Devillers, L., & Schmauch, B. (2018*).* CNN+LSTM architecture for speech emotion recognition with data augmentation. In *Proceedings of the INTERSPEECH 2018 Workshop on Speech, Music and Mind* (pp.21–25). Hyderabad, India:ISCA archive. http://dx.doi.org/10.21437/SMM.2018-5.

Farhadipour, A., Veisi, H., Asgari, M., & Keyvanrad, M. A. (2018). Dysarthric speaker identification with different degrees of dysarthria severity using deep belief networks, *ETRI Journal (Electronics and Telecommunications Research Institute)*, *40*(5), 643–652. http://dx.doi.org/10.4218/etrij.2017-0260

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning.* Adaptive Computation and Machine Learning Series. Cambridge, MA: MIT Press.

Han, K., Wang, Y., Wang, D, Woods, W. S., Merks, I., & Zhang, T. (2015). Learning spectral mapping for speech dereverberation and denoising. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23*(6), 982–992. http://dx.doi.org/10.1109/TASLP.2015.2416653.

Hanilçi, C., Kinnunen, T., Sahidullah, M., & Sizov, A. (2016). Spoofing detection goes noisy: An analysis of synthetic speech detection in the presence of additive noise. *Speech Communication, 85*, 83–97. http://dx.doi.org/10.1016/j.specom.2016.10.002

Hautamäki, R. G., Kinnunen, T., Hautamäki, V., & Laukkanen, A. M. (2015). Automatic versus human speaker verification: The case of voice mimicry. *Speech Communication, 72*, 13–31. http://dx.doi.org/10.1016/j.specom.2015.05.002

Himawan, I., Villavicencio, F., Sridharan, S., & Fookes, C. (2019). Deep domain adaptation for anti-spoofing in speaker verification systems. *Computer Speech & Language, 58*, 377–402. http://dx.doi.org/ 10.1016/j.csl.2019.05.007

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsburry, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine, 29*(6), 82–97. http://dx.doi.org/10.1109/MSP.2012.2205597.

Huzaifah, M. (2017). Comparison of time-frequency representations for environmental sound classification using convolutional neural networks. *Computing Research Repository (CoRR)*, *ArXiv.* 1–5. Retrieved from https://arxiv.org/abs/1706.07156v1.

Jayalakshmi, S. L., Chandrakala, S., & Nedunchelian, R. (2018). Global statistical features-based approach for acoustic event detection. *Applied Acoustics*, *139*, 113–118. http://dx.doi.org/10.1016/j.apacoust.2018.04.026.

Kang, T. G., Shin, J. W., & Kim, N. S. (2018). DNN-based monaural speech enhancement with temporal and spectral variations equalization. *Digital Signal Processing, 74*, 102–110. http://dx.doi.org/10.1016/j.dsp.2017.12.002

Khodabakhsh, A., Mohammadi, A., & Demiroglu, C. (2017). Spoofing voice verification systems with statistical speech synthesis using limited adaptation data. *Computer Speech & Language, 42*, 20–37. http://dx.doi.org/10.1016/j.csl.2016.08.004

Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., & Inman, D.J. (2019). 1D convolutional neural networks and applications: A survey. *Computing Research Repository (CoRR)*, *ArXiv.* 1–20. Retrieved from https://arxiv.org/abs/1905.03554v1.

Kong. Q., Xu, Y., Sobieraj, I., Wang, W., & Plumbley, M. D. (2019). Sound event detection and time–frequency segmentation from weakly labelled data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(4), 777–787. http://dx.doi.org/10.1109/TASLP.2019.2895254.

Korkmaz, Y. ve Boyacı, A. (2018). Adli bilişim açısından ses incelemeleri. *Fırat Üniversitesi Mühendislik Bilimleri Dergisi, 30*, 329–343.

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324. http://dx.doi.org/10.1109/5.726791.

Li, R., Liu, Y., Shi, Y., Dong, L., & Cui, W. (2016). ILMSAF based speech enhancement with DNN and noise classification. *Speech Communication, 85*, 53-70. http://dx.doi.org/10.1016/j.specom.2016.10.008

Lopez-Moreno, I., Gonzalez-Dominguez, J., Martinez, D., Plchot, O., Gonzalez-Rodriguez, J., & Moreno, P. J. (2016). On the use of deep feedforward neural networks for automatic language identification. *Computer Speech & Language, 40*, 46–59. http://dx.doi.org/10.1016/j.csl.2016.03.001

Meral, H. M., Sankur, B., Özsoy, A. S., Güngör, T., & Sevinç, E. (2009). Natural language watermarking via morphosyntactic alterations. *Computer Speech & Language, 23*(1), 107–125. http://dx.doi.org/10.1016/j.csl.2008.04.001

Morfi, V., & Stowell, D. (2018). Deep learning for audio event detection and tagging on low-resource datasets. *Applied Sciences*, *8*(8):1397, 1–16. http://dx.doi.org/10.3390/app8081397.

Muratoğlu, O., Okul, Ş. & Aydın, M. A. & Bilge, H. S. (2018, September). Review on cyber risks relating to security management in smart cars. *Proceedings 3rd International Conference on Computer Science and Engineering (UBMK18)*, Sarajevo, Bosnia and Herzegovina, 406–409. http://dx.doi.org/10.1109/UBMK.2018.8566569.

Özer, İ., Özer, Z., & Fındık, O. (2018). Noise robust sound event classification with convolutional neural network. *Neurocomputing, 272*, 505–512. http://dx.doi.org/10.1016/j.neucom.2017.07.021

Patterson, J. & Gibson, A. (2016). *Deep learning.* Sebastopol, CA: O'Reilly Media, Inc.

Qian, Y., Chen, N., & Yu, K. (2016). Deep features for automatic spoofing detection. *Speech Communication, 85*, 43–52, http://dx.doi.org/10.1016/j.specom.2016.10.007

Qian, Y., Evanini, K., Wang, X., Lee, C. M., & Mulholland, M., (2017). Bidirectional LSTM-RNN for improving automated assessment of non-native children's speech. In *Proceedings of the INTERSPEECH 2017 18th Annual Conference of the International Speech Communication Association* (pp.1417–1421). Stockholm, Sweden:ISCA Archive. http://dx.doi.org/10.21437/Interspeech.2017-250

Sağıroğlu Ş. ve Koç, O. (2017). *Büyük veri ve açık veri analitiği: Yöntemler ve uygulamalar.* Ankara: Grafiker Yayınları.

Sainath, T. N., Kingsbury, B., Saon, G., Soltau, H., Mohamed, A.-R., Dahl, G., & Ramabhadran, B. (2015). Deep convolutional neural networks for large-scale speech tasks. *Neural Networks, 64*, 39–48. http://dx.doi.org/10.1016/j.neunet.2014.08.005

Salamon, J., & Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters, 24*(3), 279–283. http://dx.doi.org/ 10.1109/LSP.2017.2657381.

Samui, S., Chakrabarti, I., & Ghosh, S. K. (2017, August). Deep recurrent neural network based monaural speech separation using recurrent temporal restricted boltzmann machines. In *Proceedings Interspeech 2017 18th Annual Conference of the International Speech Communication Association* (pp.3622–3626). Stockholm, Sweden:ISCA archive. http://dx.doi.org/10.21437/Interspeech.2017-57.

Samui, S., Chakrabarti, I., & Ghosh, S. K. (2019). Time–frequency masking based supervised speech enhancement framework using fuzzy deep belief network. *Applied Soft Computing, 74*, 583–602. http://dx.doi.org/10.1016/j.asoc.2018.10.031.

Sharan, R. V., & Moir, T. J. (2017). Robust acoustic event classification using deep neural networks. *Information Sciences, 396*, 24–32. http://dx.doi.org/10.1016/j.ins.2017.02.013.

Todisco, M., Delgado, H., & Evans, N. (2017). Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Computer Speech & Language, 45*, 516–535. http://dx.doi.org/10.1016/j.csl.2017.01.001.

Valenti, M., Squartini, S., Diment, A., Parascandolo, G., & Virtanen, T. (2017, May). A convolutional neural network approach for acoustic scene classification. In *Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN)* (pp.1547–1554). Anchorage, AK. http://dx.doi.org/10.1109/IJCNN.2017.7966035.

Virtanen, T., Plumbley, M. D., Ellis, D. (Eds.). (2018). *Computational analysis of sound scenes and events*. Cham, Switzerland: Springer International Publishing AG. http://dx.doi.org/10.1007/978-3-319-63450-0.

Xu, Y., Huang, Q., Wang, W., Foster, P., Sigtia, S., Jackson, P. J. B., & Plumbley, M. D. (2017). Unsupervised feature learning based on deep models for environmental audio tagging. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25*(6), 1230–1241. http://dx.doi.org/10.1109/TASLP.2017.2690563.

Zazo, R., Lozano-Diez, A., Gonzalez-Dominguez, J., Toledano, D. T., & Gonzalez-Rodriguez, J. (2016). Language identification in short utterances using long short-term memory (LSTM) recurrent neural networks. *PLoS ONE*, *11*(1):e0146917, 1–17. http://dx.doi.org/10.1371/journal.pone.0146917.

Zeyer, A., Doetsch, P., Voigtlaender, P., Schlüter, R., & Ney, H. (2017). A comprehensive study of deep bidirectional LSTM RNNS for acoustic modeling in speech recognition. In *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp.2462–2466). New Orleans, LA:IEEE. http://dx.doi.org/10.1109/ICASSP.2017.7952599.

Zheng, W., Mo, Z., Xing, X., & Zhao, G. (2018). CNNs-based acoustic scene classification using multi-spectrogram fusion and label expansions. *Computing Research Repository (CoRR)*, *ArXiv.* 1–7. Retrieved from https://arxiv.org/abs/1809.01543v1.

Zhou, H., Bai, X., & Du, J. (2018, November). An investigation of transfer learning mechanism for acoustic scene classification. *Proceedings of 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)* (pp. 404-408). Taipei City, Taiwan. http://dx.doi.org/10.1109/ISCSLP.2018.8706712