Bölüm 6

İşitsel ve Görsel Verilerle Ruhsal Bozuklukların Hesaplamalı Analizinde Veri İşleme Hatları, Öznitelik Çıkarımı ve Çok-Kipli Füzyon

Uygar Aydin[1]

İnci Zaim Gökbay[2]

Özet

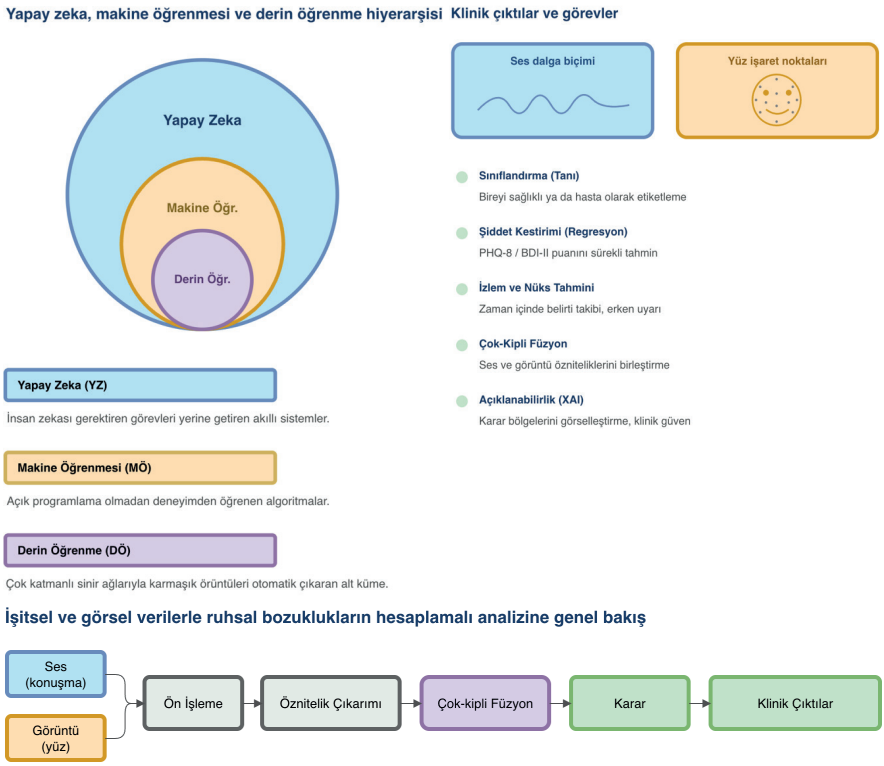
Ruhsal bozuklukların erken tanısı, tedavisi ve izlenmesi, belirtilerin öznel doğası ve geleneksel klinik yöntemlerin ölçüm sınırlılıkları nedeniyle güçtür. Bu bölüm, işitsel, görsel ve çok-kipli verilerden nesnel biyobelirteçler elde etmeyi amaçlayan hesaplamalı yaklaşımları sinyal işleme ve makine öğrenmesi perspektifinden ele almaktadır. Veri kaynakları, klinik değerlendirme ölçekleri, ön işleme adımları, ses ve görüntü verileri için öznitelik çıkarım yöntemleri, sınıflandırma mimarileri ve çok-kipli füzyon stratejileri teknik düzeyde incelenmektedir. Bu teknik çerçeve son yıllarda belirli yöntemler ve hedefler çevresinde yoğunlaşan ve özellikle COVID-19 sonrası dönemde hızla genişleyen kapsamlı bir literatüre dayanmaktadır. Bu yöntem ve hedef yoğunlaşması içinde depresyon tespiti baskın araştırma hedefi olarak öne çıkmış, evrişimli sinir ağları (CNN) ise temel mimari haline gelmiştir. Öznitelik düzeyinde gerçekleştirilen ses-görüntü füzyonu, tek-kipli çözümlere kıyasla kayda değer doğruluk kazanımları sağlamıştır. Bölümde ayrıca temsili yüz ve ses tanıma mimarileri, temel ve yardımcı sınıflandırma yöntemleri arasındaki ayrım ile uzaktan ve sürekli değerlendirme, mahremiyeti koruyan girişimsel olmayan izlem ve kaynak erişiminin sınırlı olduğu koşullara uyarlanabilirlik gibi gerçek yaşam uygulamaları tartışılmaktadır. Bununla birlikte kültürel ve dilsel çeşitlilikten yoksun veri kümeleri, tanı ile tedavi arasındaki boşluk ve modellerin yorumlanabilirlik eksikliği alandaki yeniliklerin klinik etkiye dönüşmesinin önündeki başlıca engeller olmayı sürdürmektedir. Dolayısıyla bu

1 İstanbul Üniversitesi, Fen Bilimleri Enstitüsü, Enformatik Anabilim Dalı, İstanbul; uygaraydin1@ogr.iu.edu.tr; https://orcid.org/0000-0002-9052-3512

2 İstanbul Üniversitesi, Bilgisayar ve Bilişim Teknolojileri Fakültesi, Yapay Zeka ve Veri Mühendisliği Bölümü, Yapay Zeka ve Veri Mühendisliği Anabilim Dalı, İstanbul; inci.gokbay@istanbul.edu.tr; https://orcid.org/0000-0002-4488-1642

https://doi.org/10.58830/ozgur.pub1351.c5539

engellerin aşılması alanın önceliklerini doğrudan şekillendirmektedir. Yapılan araştırmalar model doğruluğunu artırmanın ötesinde, bu alanda çeşitlilik içeren ve uzun vadeli veri kümeleri oluşturulmasının, modellerin tüm klinik sürecini kapsayacak şekilde genişletilmesinin ve klinik ihtiyaçları doğrudan karşılayan, yorumlanabilir sistemler geliştirilmesinin gerekli olduğunu ortaya koymuştur.



Grafik Özet. İşitsel-görsel verilerle ruhsal bozuklukların hesaplamalı analizine genel bakış, yöntem hiyerarşisi, klinik çıktılar ve görevler. (Özgün şekil; yazarlar tarafından oluşturulmuştur.)

1. Giriş

Dünya Sağlık Örgütü (WHO) ruh sağlığını, bireyin yeteneklerinin farkında olduğu, olağan yaşam stresleriyle baş edebildiği, verimli biçimde çalışabildiği ve topluluğuna katkı sunabildiği bir iyilik hali olarak tanımlar (WHO, 2022b). Buna karşın ruhsal sorunlar, fiziksel hastalıklar kadar görünür olmamaları nedeniyle giderek büyüyen bir halk sağlığı yüküne dönüşmüştür. Global Hastalık Yükü (Global Burden of Disease, GBD) 2019 analizine göre dünyada her sekiz kişiden biri (yaklaşık 970 milyon birey) bir ruhsal bozuklukla yaşamaktadır. Bunların yaklaşık 280 milyonu depresyon, 301

milyonu ise anksiyete bozukluğu tanısı taşımaktadır (GBD 2019 Mental Disorders Collaborators, 2022; WHO, 2022c). Ruh sağlığı sorunlarının oluşturduğu yükün ekonomik boyutu da çarpıcıdır. Nitekim İngiltere'de pandemi öncesinde ruhsal bozukluklar sebebiyle ortaya çıkan yıllık ekonomik ve toplumsal maliyetin 105 milyar sterline ulaştığı raporlanmıştır (Adcock ve Parkin, 2016). Bu maliyetin gerisinde, yalnızca bozukluğun kendisi değil, ona erişimi ve tedaviyi güçleştiren etkenler de yer almaktadır. Ruhsal bozukluklar bireyin iyilik hali, eğitim ve çalışma yaşamı ile sosyal ilişkileri üzerinde olumsuz etkiler doğurur. Tanı ve tedavi süreçlerine erişimdeki güçlükler ve uzun bekleme süreleri ise bu etkileri ağırlaştırmakta ve sağlık sistemleri üzerindeki yükü daha da artırmaktadır. Nesnel, ölçeklenebilir ve erişilebilir değerlendirme araçlarına duyulan ihtiyaç bu yapısal sorunlar çerçevesinde belirginleşmektedir.

COVID-19 salgını hem talebi nicel olarak artırmış hem de talebin doğasını değiştirmiştir. Salgının ilk yılında küresel anksiyete yaygınlığında %25,6 ve depresyon yaygınlığında %27,6 oranında bir artış bildirilmiştir (Santomauro vd., 2021; WHO, 2022a). Yüz yüze terapi modellerinin kesintiye uğraması, uzaktan erişilebilir ve ölçeklenebilir dijital çözümlere yönelik acil bir ihtiyaç doğurmuştur. Bu ihtiyaç, duygusal hesaplama (affective computing) araştırmalarının ulaştığı teknolojik olgunlukla birleşince, ses ve yüz temelli hesaplamalı yöntem araştırmalarını belirgin biçimde hızlandırmıştır (He, Niu, vd., 2022; Low vd., 2020). Bu dönemde ulusal ve uluslararası fon kuruluşlarının dijital ruh sağlığı teknolojilerine sağladığı büyük ölçekli kaynaklar, yapay zeka ve makine öğrenmesi uzmanlarının dikkatlerini bu alana çekmiştir. Böylece toplumsal talep, teknolojik olgunluk ve kurumsal teşvik alanın hızla genişlemesinin zeminini hazırlamıştır.

Geleneksel psikiyatrik değerlendirmenin temel kısıtı büyük ölçüde öznel belirti bildirimine ve klinik gözleme dayanmasıdır. Bu noktada konuşma sinyalleri ve yüz ifadeleri gibi gözlemlenebilir davranışsal ve fizyolojik ipuçlarından nesnel biyobelirteçler (biomarker) çıkarmayı amaçlayan hesaplamalı yöntemler önem kazanmıştır (He, Niu, vd., 2022; Low vd., 2020). Bu bölüm, alanı bir mühendislik problemi olarak ele alır ve üç temel eksende yapılandırır. İlk olarak yüz ve ses analizlerinin erken tanıya katkıları, ardından makine ve derin öğrenme tekniklerinin tanı ile tedaviyi desteklemedeki etkinliği, son olarak da bu tekniklerin tanı, tedavi ve takip sürecini iyileştirme yolları değerlendirilmektedir.

Bu bölümün ele aldığı yaklaşımlar salt teknik çözümler değil, daha geniş bir kuramsal dönüşümün parçasıdır. Yapay zeka ve özellikle derin öğrenme, elle tasarlanmış kurallar yerine temsilleri doğrudan veriden hiyerarşik biçimde öğrenmeyi mümkün kılarak görüntü, ses ve dil işlemede paradigma değiştirici

bir rol üstlenmiştir (LeCun vd., 2015). Bu modellerin gücü büyük ölçüde veri ölçeğine bağlı olduğundan yüksek hacim, hız ve çeşitlilik gösteren veri kümelerini niteleyen büyük veri (big data), sağlık ile davranış bilimlerinde giderek merkezi bir rol kazanmıştır (Monteith vd., 2015). Sağlık ve davranış verisinin dijitalleşmesiyle ortaya çıkan bu veri yığınları, öğrenme temelli yöntemlerin itici gücü haline gelmiştir. Bu öğrenme paradigması ve büyük veri birikimi, ruh sağlığı alanında iki gelişmede somutlaşır. Birincisi hesaplamalı psikiyatrinin (computational psychiatry) gelişimidir. Beyin ve davranışı matematiksel modellerle ele alan bu yaklaşım, nörobilim ile klinik uygulama arasında veri ve kuram temelli bir köprü kurar (Montague vd., 2012; Huys vd., 2016). İkincisi ise dijital fenotiplemedir (digital phenotyping). Akıllı telefon ve giyilebilir cihazların yaygınlaşması bireyin günlük davranışını sürekli ve nesnel biçimde ölçen büyük ölçekli veri akışları doğurarak bu kavramı ortaya çıkarmıştır (Insel, 2017). Büyük veri, öğrenme paradigması ve klinik gereksinimin kesişiminde olan makine öğrenmesi, psikiyatrik değerlendirmeyi öznel bildirimden nesnel ölçüme taşıma potansiyeli taşır (Dwyer vd., 2018). İşitsel ve görsel veriler, davranışsal bilgi açısından zengin ve en az girişimsel kaynaklar arasında yer aldığından bu bölümün odağını oluşturur.

Bu odak doğrultusunda bölümde veri kaynakları, veri işleme hattı, performans değerlendirmeleri ve füzyon stratejileri, gerçek yaşam uygulamaları ve klinik entegrasyon, tartışma ve sınırlılıklar ile sonuç ve öneriler derinlemesine incelenmiştir.

2. Veri Kaynakları ve Klinik Ölçekler

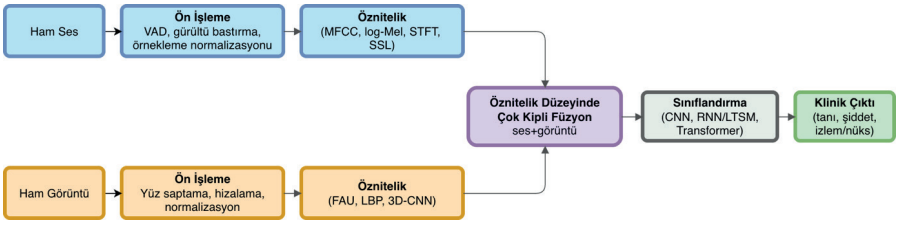
Bu alandaki çalışmalar hem açık kaynak hem de araştırmacıların kendileri tarafından toplanmış olan veri kümelerine dayanır. En sık kullanılan açık kaynak kümeleri klinik görüşme videoları içeren DAIC-WOZ (Gratch vd., 2014) ve AVEC serisidir (Valstar vd., 2013; Valstar vd., 2016). Bu kümeler çok-kipli ve doğrudan ruhsal bozukluklara odaklandıkları için tercih edilirler. DAIC-WOZ veri kümesi bir sanal görüşmecinin yürüttüğü yarı yapılandırılmış klinik görüşmelerden oluşur ve PHQ-8 etiketleriyle birlikte sunulur. AVEC serisi ise 2013'ten bu yana depresyon tanısını bir alt yarışma olarak ele almış ve alana standart bir kıyaslama zemini kazandırmıştır (Valstar vd., 2013; Valstar vd., 2016). Bunların yanında çok-kipli MODMA, etkileşimli duygusal IEMOCAP ve vlog temelli D-Vlog gibi kümeler de kullanılmış, yalnızca görüntü içeren CK+, yalnızca ses içeren ISED ve RAVDESS (Livingstone ve Russo, 2018) ile metin temelli ISEAR gibi daha özelleşmiş kaynaklar tekil çalışmalarda yer bulmuştur. Hastalık şiddeti ağırlıklı olarak Hasta Sağlığı Anketi (Patient Health Questionnaire, PHQ-8) ve Hamilton Depresyon Derecelendirme Ölçeği (Hamilton Depression Rating Scale, HAMD) gibi standart klinik ölçeklerle

etiketlenmiştir. Bazı çalışmalarda mevcut ölçeklerin tanı için yetersiz kaldığı gerekçesiyle Zung Kendini Değerlendirme Depresyon Ölçeği (Self-Rating Depression Scale, SDS) gibi alternatifler tercih edilmiştir (Xie vd., 2021).

Mevcut veri kümelerinin sunduğu zaman tasarrufu ve etik kolaylığa karşılık, bu kümeler özelleştirilemez ve belirli dillerle sınırlıdırlar. Bu nedenle bazı araştırmacılar psikiyatrik değerlendirme sırasında çekilen video kayıtlarından kendi kümelerini oluşturmuştur. Örnekler arasında yüksek çözünürlüklü gerçek zamanlı analiz için toplanan kümeler (Gilanie vd., 2022), duygusal uyaran sunumuna dayalı görevlerle desteklenen tasarımlar (Liu vd., 2024) ve farklı dil ve popülasyonlara yönelik kümeler (Hall vd., 2024; Kim vd., 2023; Mahayossanunt vd., 2023) yer alır. Bu eğilim, aşağıda tartışılacağı üzere, dilsel ve kültürel çeşitlilik açısından önemli bir sınırlılık yaratmaktadır. Kendi verisini toplayan araştırmacılar uyaran temelli görevler ya da yüksek çözünürlüklü kayıt düzenekleriyle depresif belirtileri daha belirgin biçimde ortaya çıkarmayı amaçlamıştır. Bu çabalar değerli olmakla birlikte, örneklem sınırlılığı ve standart dışı protokoller sonuçların karşılaştırılabilirliğini ve yeniden üretilebilirliğini sınırlandırmaktadır.

3. Veri İşleme Hattı

Hesaplamalı yöntemlerin büyük çoğunluğu ön işleme (preprocessing), öznitelik çıkarımı ve sınıflandırmadan oluşan üç aşamalı klasik bir hattı izler. Ses ve görüntü kollarının öznitelik düzeyinde birleştiği bu üç aşamalı işleme hattının genel görünümü Şekil 1'de gösterilmektedir.



Şekil 1. İşitsel-görsel ruhsal bozukluk analizinde veri işleme hattı; ses ve görüntü kolları öznitelik düzeyinde birleşir.

3.1. Ön İşleme

Ön işleme, ham ses ve video kayıtlarını öznitelik çıkarımına uygun, gürültüden arındırılmış ve normalize edilmiş bir temsile dönüştürür. Tipik adımlar konuşma etkinliği tespiti (Voice Activity Detection, VAD), yüz saptama ve hizalama, çerçeve seçimi ile ölçek ve aydınlatma normalizasyonunu içerir. Gerçek-zamanlı uygulamalarda yüksek video çözünürlüğü ince kas hareketlerinin

ve mikro-ifadelerin (micro-expression) daha belirgin yakalanmasına olanak tanır (Gilanie vd., 2022). Ses tarafında bu hazırlık gürültü bastırma, sessiz bölümlerin konuşma etkinliği tespitiyle ayıklanması ve örnekleme oranının standartlaştırılmasını kapsar. Görüntü tarafında ise yüz işaret noktalarına (facial landmarks) göre hizalama, ilgi bölgesinin kırpılması ve poz ile aydınlatma farklarının dengelenmesi öne çıkar. Bu adımlardaki bir hata sonraki katmanlara doğrudan taşındığından, ön işleme kalitesi nihai sınıflandırma başarısının görünmeyen ama belirleyici bir bileşenidir.

3.2. Öznitelik Çıkarımı

Öznitelik çıkarımı, el yapımı öznitelikler ve öğrenilmiş derin temsiller olmak üzere iki temel yaklaşıma dayanır. Görüntü tarafında öğrenilmiş derin temsil yaklaşımı kendi içinde ikiye ayrılır. Bir uçta ResNet, VGG, SeNet ve üç boyutlu evrişimli ağlar gibi mimariler ham görüntüden uçtan uca öznitelik öğrenir. Diğer uçta OpenFace ve FaceReader gibi araç takımları iki aşamalı bir hat kurar. Bu araçlar görüntüden önce yüz işaret noktaları, bakış yönü ve yüz hareket birimi yoğunlukları gibi yorumlanabilir orta düzey öznitelikler çıkarır, sonraki ağ ise bu zenginleştirilmiş temsiller üzerinde çalışır. Yorumlanabilir çıktıları nedeniyle OpenFace ve FaceReader en sık başvurulan araçlar arasında yer alır. Bu ikinci yol alanın ham pikselleri tek aşamada işlemekten çok aşamalı ve yüksek düzeyli temsillere yönelimini yansıtır. Aynı yönelimin bir uzantısı olarak derin mimarilere dikkat mekanizması (attention mechanism) ve çok ölçekli temsil üreten Özellik Piramit Ağları (Feature Pyramid Networks) gibi bileşenler eklenerek daha soyut ve göreve özgü temsiller hedeflenir (Xu vd., 2024). Uçtan uca öğrenme ile araç temelli yüksek düzeyli temsil çıkarımı arasındaki seçim, yorumlanabilirlik (interpretability) ile esneklik arasındaki dengeye ve veri büyüklüğüne göre yapılır. Uçtan uca derin ağlar daha esnek ancak yorumlanması güç temsiller üretirken, araç temelli betimleyiciler daha yorumlanabilir olup az veriyle daha kararlı sonuç verir.

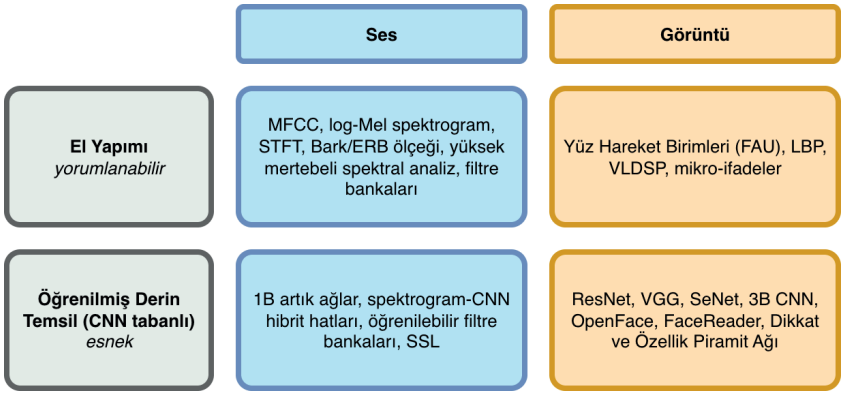
El yapımı görüntü betimleyicileri çoğunlukla bu yorumlanabilir uca konumlanır. Bunların başında, yüz ifadelerinin anatomik temelini doğrudan yansıttığı için yüksek açıklanabilirliğe sahip Yüz Hareket Birimleri (Facial Action Units, FAU) gelir. Nitekim OpenFace gibi araçların ürettiği temel çıktı da bu birimlerdir. Doku temelli Yerel İkili Örüntüler (Local Binary Patterns, LBP) ve dinamik bir betimleyici olan Hacimsel Yerel Yönelimli Yapısal Örüntü (Volume Local Directional Structural Pattern, VLDSP) (Uddin vd., 2022) ile mikro-ifade analizine dayalı yöntemler (Gilanie vd., 2022) bu aileyi tamamlar. Bu çeşitlilik, el yapımı betimleyicilerin yorumlanabilirliği ile öğrenilmiş temsillerin esnekliğini birleştiren hibrit tasarımlara olan ilgiyi yansıtır.

Ses tarafında benzer bir ayrım geçerlidir. Ham dalga biçimini uçtan uca işleyen 1B artık ağların yanı sıra, el yapımı akustik öznitelikleri girdi alan hibrit veri işleme hatları (pipeline) yaygındır ve bu hatlar daha az veriyle daha kararlı sonuç hedefler. El yapımı akustik öznitelikler arasında Mel-Frekansı Kepstral Katsayıları (Mel-Frequency Cepstral Coefficients, MFCC) en yaygın spektral temsildir ve sıklıkla CNN tarafından üretilen spektrogram öznitelikleriyle birleştirilir (Das ve Naskar, 2024). COVAREP gibi araç takımları standart akustik öznitelik kümeleri sağlarken (Degottex vd., 2014), öğrenilebilir zaman-uzamı filtre bankaları (learnable time-domain filterbanks) sabit betimleyiciler yerine veriden optimize edilen bir ara çözüm sunar (Yang vd., 2023). Bu çekirdek temsillerin yanında Log-Mel spektrogram, Bark ölçeği ve eşdeğer dikdörtgensel bant genişliği gibi düşük frekansa odaklı temsiller ile Kısa-Zamanlı Fourier Dönüşümü (Short-Time Fourier Transform), Kepstrum, Gabor dönüşümü ve yüksek mertebeli spektral analiz (Higher-Order Spectral Analysis) gibi dönüşüm temelli öznitelikler de kullanılır.

Tablo 1. Ses ve görüntü için başlıca öznitelik çıkarım aileleri ve temsil eden örnek çalışmalar.

Öznitelik ailesi	Başlıca yöntemler	Örnek çalışma
CNN tabanlı görüntü	ResNet, VGG, SeNet, 3B CNN (uçtan uca); OpenFace, FaceReader (iki aşamalı araç); dikkat ve Özellik Piramit Ağı (mimari bileşen)	(Xu vd., 2024)
CNN tabanlı ses	1B artık ağlar (ham dalga biçimi), spektrogram-CNN hibrit hatları, öğrenilebilir filtre bankaları	(Das ve Naskar, 2024)
El yapımı ses	MFCC, log-Mel spektrogram, STFT, filtre bankaları	(Yang vd., 2023)
El yapımı görüntü	Yüz Hareket Birimleri (FAU), LBP, VLDSP, mikro-ifadeler	(Gilanie vd., 2022)

Öznitelik ailelerinin temsil türü (el yapımı, öğrenilmiş) ve kip (ses, görüntü) ekseninde sınıflandırılması Şekil 2'de özetlenmektedir.



Şekil 2. Öznitelik çıkarımı taksonomisi: temsil türü (el yapımı / öğrenilmiş) × kip (ses / görüntü).

Bu öznitelik ailelerinin somut mimari karşılıkları literatürde belirgindir. Yüz dinamiklerini değerlendirmek için değişken çekirdek boyutları kullanan Çok-Ölçekli Uzamsal-Zamansal Ağ (Multiscale Spatiotemporal Network, MSN), tekdüze çekirdekli C3D gibi modellere kıyasla ince yüz değişimlerini daha verimli yakalamıştır (de Melo vd., 2020). DepNet ise video temelli analizde yüz ifadelerinin zamansal özniteliklerini modelleyerek doğruluğu artırmıştır (He, Guo, vd., 2022). Yorumlanabilirlik açısından kritik bir örnek, küresel ortalama havuzlama (global average pooling) katmanı içeren bir derin evrişimli ağ ile Depresyon Etkinleştirme Haritaları (Depression Activation Maps) üreten DepressNet'tir. Bu haritalar, depresyon şiddetine dair anlamlı bilgi taşıyan yüz bölgelerini işaretleyerek sonuçların klinik yorumunu güçlendirmiştir (Zhou vd., 2020). Ses tarafında sesli ve sessiz harf düzeyinde fonem temelli bir CNN mimarisi olan AudVowelConsNet (Muzammel vd., 2020) ile üç boyutlu evrişimli ağlar (Wang vd., 2021) konuşmanın klinik açıdan ayırt edici örüntülerini yakalamaya yönelik tamamlayıcı yaklaşımlardır.

3.3. Sınıflandırma Mimarileri

Sınıflandırma katmanında evrişimli sinir ağları (Convolutional Neural Network, CNN) baskın mimaridir. CNN, görüntü ve ses verisinde otomatik öznitelik çıkarımı ve sınıflandırma için temel araç haline gelmiştir. Dikkat mekanizmaları (attention mechanism), modelin odağını girdideki ayırt edici bölgelere yönlendirerek daha derin ve ayrıntılı temsiller üretir ve giderek daha sık tamamlayıcı bileşen olarak kullanılır (Othmani vd., 2022; Xu vd., 2024). Donmuş duygulanım veya konuşma hızındaki yavaşlama gibi zamansal dinamiklerin modellenmesinde ise yinelemeli ağlar (Recurrent Neural Network, RNN, LSTM, Bi-LSTM) öne çıkar (Uddin vd., 2022). Genel eğilim, statik

örüntüleri yakalayan temel modellerden, zamansal dinamikleri modelleyen ağlara ve nihayetinde kararını gerekçelendirebilen açıklanabilir yapay zeka (Explainable AI, XAI) yaklaşımlarına doğru bir evrimdir (Mahayossanunt vd., 2023; Xie vd., 2021). Bu mimari çeşitliliğin altında pratik bir gerekçe yatar. Evrişimli ağlar görece az veriyle güçlü uzamsal öznitelikler öğrenebildikleri için baskın omurgayı oluştururken, yinelemeli ağlar ardışık çerçeveler arasındaki bağımlılıkları modelleyerek konuşma temposu ve ifade geçişleri gibi zamansal ipuçlarını yakalar. Evrişimli ve yinelemeli ağların bu rol paylaşımı dışında kalan mimariler de belirli sınırlılıkları aşmak üzere denenmiştir. Çizge sinir ağları (Graph Neural Networks, GNN) ve karışım modelleri kipler arası ilişkileri daha esnek temsil edebildikleri için niş senaryolarda sınanmış, Transformer (dönüştürücü) mimariler ise yinelemeli ağların zorlandığı uzun menzilli bağımlılıkları dikkat ağırlıklarıyla modelleyerek yeni bir yön açmıştır. Her farklı mimari belirli bir ihtiyaca yanıt verdiğinden mimari seçimi tek bir ölçüte değil; veri büyüklüğü, yorumlanabilirlik beklentisi ve hesaplama bütçesi arasındaki dengeye dayanır.

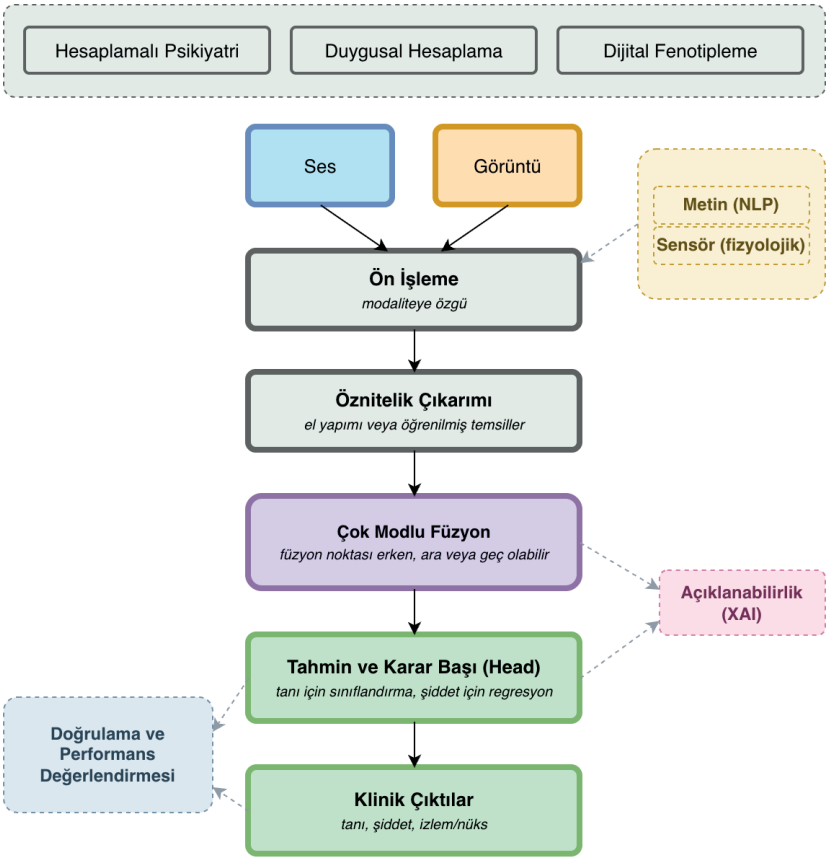
Bu mimariler tek tek ele alınmalarının yanı sıra, tanı sürecinde üstlendikleri role göre de sınıflandırılabilir. Tanı desteği sağlayan bu modeller yardımcı bir mekanizma kullanıp kullanmamalarına göre iki gruba ayrılabilir. Temel yaklaşımlar, tek başına bir CNN veya geleneksel makine öğrenmesi modeliyle doğrudan sınıflandırma yapar. Yardımcı yaklaşımlar ise bu omurgayı dikkat mekanizmaları, uzamsal-zamansal modüller veya çok-kipli füzyon (multimodal fusion) ile güçlendirir. Örneğin yüz ifadesi ile göz bebeği tepkisini bir öz-dikkat (self-attention) ağında birleştiren (Liu vd., 2024) ya da dikkat mekanizmasını bir öznitelik piramidiyle eşleştiren (Xu vd., 2024) çalışmalar, ayırt edici bölgelere odaklanarak başarımı artırmayı amaçlar. Söz konusu ayrım, model karmaşıklığı ile yorumlanabilirlik ve veri verimliliği arasındaki ödünleşimi de görünür kılar. Pratikte yardımcı mekanizmaların katkısı, eklenen karmaşıklığın getirdiği veri ve hesaplama maliyetiyle birlikte ve veri büyüklüğüne göre değerlendirilmelidir. Nitekim küçük örneklemlerde sade bir omurga, aşırı uyum riski nedeniyle daha güvenilir sonuç verebilir.

Ses temelli tanıma yaklaşımları da benzer bir çeşitlilik gösterir. Mel-Frekansı Kepstral Katsayıları ile CNN tarafından üretilen spektrogram özniteliklerini birleştiren melez modeller yaygın bir başlangıç noktasıdır (Das ve Naskar, 2024). Öğrenilebilir zaman-uzamı filtre bankaları ise sabit el yapımı filtreler yerine veriden öğrenilen, dikkatle yönlendirilen temsiller sunar (Yang vd., 2023). Fonem düzeyinde uzmanlaşmış mimariler (sesli ve sessiz harfleri ayrı ağlarla işleyen AudVowelConsNet gibi) konuşmanın ince akustik yapısını hedeflerken (Muzammel vd., 2020), üç boyutlu evrişimli ağlar zaman ve frekans eksenlerini birlikte modelleyerek depresif konuşma örüntülerini

yakalamaya çalışır (Wang vd., 2021). Bu yaklaşımların ortak amacı, depresif konuşmanın düşük perde, monoton tonlama ve yavaşlamış tempo gibi ayırt edici örüntülerini, dile ve konuşmacıya olabildiğince bağımsız bir biçimde temsil etmektir.

Bu CNN-merkezli manzara, son yıllarda öz-denetimli öğrenme (self-supervised learning, SSL) temelli konuşma temel modellerinin (foundation models) yükselişiyle değişmektedir. wav2vec 2.0 (Baevski vd., 2020), HuBERT (Hsu vd., 2021) ve WavLM (Chen vd., 2022) gibi, etiketsiz büyük konuşma veri kümeleri üzerinde önceden eğitilen modeller, sınırlı klinik veriyle dahi güçlü akustik temsiller sağladıkları için el yapımı özniteliklerin ve sıfırdan eğitilen CNN'lerin yerini hızla almaktadır. Bu modeller tipik olarak ince ayar (fine-tuning) veya dondurulmuş gömme (frozen embedding) çıkarımı yoluyla kullanılır ve veri kısıtı olan ortamlarda transfer öğrenme yoluyla başarımı artırdıkları gösterilmiştir (Wu vd., 2023; Zhang vd., 2024). Güncel yönelim, sabit spektral betimleyiciler yerine bu öğrenilmiş temsilleri CNN-Transformer melez mimarileriyle veya dikkat-havuzlamalı yinelemeli katmanlarla birleştirmektir. Dolayısıyla alandaki mimari özet, CNN'in baskın omurga olduğu bir enstantanenin ötesine geçerek, önceden eğitilmiş temel modellerin ve dikkat tabanlı mimarilerin giderek merkezi rol üstlendiği bir geçiş dönemini yansıtmaktadır.

Veri kaynaklarından ön işleme, öznitelik çıkarımı ve sınıflandırmaya uzanan bu işleme zinciri, paradigma katmanı ve açıklanabilirlik bileşenleriyle birlikte Şekil 3'te bütünleşik bir hesaplamalı psikiyatri çerçevesi olarak sunulmaktadır.



Şekil 3. İşitsel ve görsel verilerle ruhsal bozukluk analizinde bütünleşik hesaplamalı psikiyatri çerçevesi

4. Performans Değerlendirmesi ve Çok-Kipli Füzyon

Bir tanı sisteminin başarımı tek bir sayıya indirgenemeyecek kadar çok boyutlu olduğundan, değerlendirme aşamasında birbirini tamamlayan çeşitli ölçütlere başvurulur. Modeller, doğruluk (accuracy), kesinlik (precision), duyarlılık (recall) ve F1 skoru gibi karışıklık matrisi (confusion matrix) temelli ölçütlerin yanı sıra ortalama mutlak hata (Mean Absolute Error, MAE) ve kök ortalama kare hata (Root Mean Square Error, RMSE) ile değerlendirilir. Bunun yanında ROC eğrisi (ROC curve) ve AUC, k-katlamalı çapraz doğrulama (k-fold cross-validation) ile korelasyon katsayıları kullanılır. Klinik bağlamda özgüllük (specificity) özellikle kritiktir. Sağlıklı bir bireyin hasta olarak sınıflandırılması (yanlış pozitif), gereksiz ilaç kullanımına, psikolojik strese ve toplumsal damgalanmaya yol açabilir. Bu nedenle tanısal modellerin yalnızca duyarlılığı değil, yanlış pozitifleri azaltan ölçütleri de gözetmesi gerekir.

Çalışmaların çoğu, doğruluğun yanında özgüllüğü, ROC eğrisi altındaki alanı ve k-katlamalı çapraz doğrulama sonuçlarını birlikte raporlar. Sürekli çıktı üreten modellerde ise Uyum Korelasyon Katsayısı (Concordance Correlation Coefficient, CCC) ve Pearson korelasyonu tercih edilir. Metrik seçimindeki bu bilinçli çeşitlilik, modellerin yalnızca ortalama başarısını değil, farklı alt gruplardaki kararlılığını, genellenebilirliğini ve klinik güvenliğini de görünür kılmayı amaçlar.

Tablo 2. Bölümde atıflı seçilmiş çalışmaların raporlanan başarımı.

Çalışma	Veri kümesi (N)	Kip / Öznitelik	Hedef	Doğrulama	Raporlanan başarım
Othmani vd. (2022)	DAIC-WOZ (189)	Ses + görüntü füzyon (FAU + spektrogram)	İkili (nüks/ depresyon)	LOSO	Doğruluk %87,4; F1 %82,3
Muzammel vd. (2020)	DAIC-WOZ (189)	Ses (fonem-düzeyi CNN)	PHQ-8 ikili	Eğitim-test bölmesi	Doğruluk %86,06; AUC 0,83; ort. F1 %85,85
Das ve Naskar (2024)	DAIC-WOZ; MODMA	Ses (MFCC + CNN-spektrogram)	İkili	Çapraz doğrulama	Doğruluk >%90 (DAIC ve MODMA)
Kim vd. (2023)	Korece, kendi (318)	Ses (log-Mel CNN)	İkili (MDB/ kontrol)	10-katlı CV	Doğruluk %78,14; ort. AUC 0,86
Zhou vd. (2020)	AVEC2013/2014	Görüntü (çok-bölgeli ResNet)	BDI-II regresyon	AVEC protokolü	MAE 6,21; RMSE 8,39 (AVEC2014) †

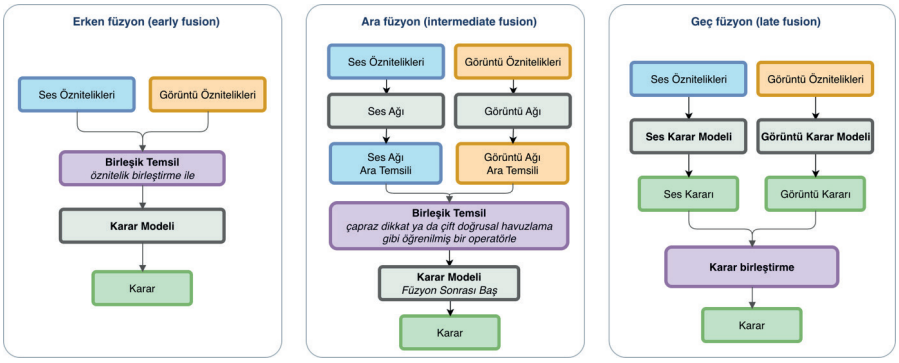
† Zhou vd. (2020) değerleri ikincil karşılaştırma kaynaklarından alınmıştır. Tablodaki çalışmalar farklı hedef, doğrulama stratejisi ve örneklem kullandığından doğrudan bir doğruluk sıralaması olarak okunmamalıdır.

Söz konusu ölçütler tek bir kipe dayalı modellerin ulaştığı sınırları görünür kıldıkça, araştırmacıları farklı veri kiplerini birleştiren füzyon temelli yaklaşımlara yöneltmiştir. Othmani ve arkadaşları, vokal ve görsel ipuçlarına dayalı yeni bir biyobelirteç tanımlayan ve majör depresif bozuklukta nüks olasılığını öngören bir Normallik Modeli (Model of Normality, MoN) çerçevesi geliştirmiştir. Videodan yüz hareket birimi öznitelikleri, konuşmadan ise spektrogram öznitelikleri çıkarılmış ve öznitelik düzeyinde gerçekleştirilen ses-görüntü füzyonu yalnızca sese dayalı modele kıyasla belirgin bir başarım artışı sağlamıştır. En iyi sonuç bir denek dışarıda bırakma (Leave-One-Subject-Out, LOSO) stratejisiyle %87,4 doğruluk ve %82,3 F1 skoru olarak raporlanmıştır (Othmani vd., 2022). Benzer biçimde Uddin ve arkadaşları,

biri ses (LSTM tabanlı) diğeri video için iki uzamsal-zamansal ağ tasarlamış, video ağında özel VLDSP betimleyicisini kullanmış ve öznitelikleri Zamansal Dikkatli Havuzlama ile özetleyip çok-kipli çarpanlara ayrılmış çift doğrusal havuzlama tekniğiyle birleştirmiştir (Uddin vd., 2022). Bu örnekler basit bir öznitelik birleştirmenin dahi çok-kipli modellerin doğruluğunu belirgin biçimde artırabildiğini ortaya koymaktadır. Tanıdan tedavi ve takibe uzanan örnekler sınırlı olmakla birlikte yön göstericidir. Psikotik bozukluklar üzerine yürütülen bir çalışmada yüz ifadeleri FaceReader tabanlı bir duygu tanıma algoritmasıyla çıkarılmış ve ifadelerin zaman içindeki geçişleri Grup Yinelemeli Çoklu Model Tahmini (Group Iterative Multiple Model Estimation, GIMME) ağ modelleriyle incelenmiştir. Psikotik grupta nötr ifadeden mutluluğa geçişlerin belirgin biçimde daha zayıf olduğu Cohen'in d değeriyle raporlanmıştır (Hall vd., 2024). Vokal ve görsel ipuçlarına dayalı Normallik Modeli çerçevesi ise nüks olasılığını öngörerek tanının ötesinde bir izlem perspektifi sunmuştur (Othmani vd., 2022). Bu çalışmalar, modellerin pasif tanı araçlarından dinamik klinik karar destek sistemlerine evrilme potansiyelini somutlaştırmaktadır.

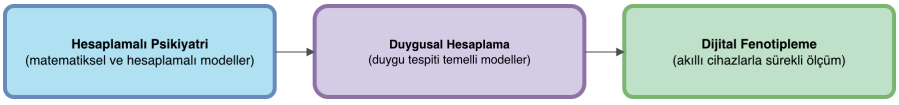
Çok-kipli modellerin bu kazanımlarının ardında kiplerin hangi aşamada birleştirildiğine ilişkin bir tasarım tercihi yatar. Çok-kipli makine öğrenmesinin klasik taksonomisi füzyonu, erken ve geç füzyon ile bu ikisini birleştiren melez (hybrid) füzyon olarak ayırır (Baltrušaitis vd., 2019). Derin öğrenme literatürü ise buna, kiplerin öğrenilmiş ara temsillerini ağın orta katmanlarında birleştiren ara (intermediate) füzyonu ekleyerek füzyonu erken, ara ve geç olmak üzere üç düzeyde ele alır (Stahlschmidt vd., 2022). Erken füzyon (early fusion) ham veriyi ya da düşük düzeyli öznitelikleri tek bir ortak temsilde birleştirip tek bir model eğitir. Kipler arası ince etkileşimleri yakalayabilir ancak zamansal hizalama ve boyut uyumsuzluğuna duyarlıdır. Geç füzyon (late fusion) her kip için ayrı modeller eğitip yalnızca karar düzeyinde, örneğin olasılıkların birleştirilmesiyle bütünleştirir. Kip kaybına karşı dayanıklı ve uygulaması kolaydır, fakat kipler arası tamamlayıcı bilgiyi büyük ölçüde göz ardı eder. Ara füzyon (intermediate fusion) ise kiplerin öğrenilmiş ara temsillerini ağın orta katmanlarında, çoğunlukla dikkat ya da çift doğrusal havuzlama gibi mekanizmalarla birleştirir ve iki uç arasında denge kurar. Othmani ve arkadaşlarının öznitelik düzeyinde birleştirme yaklaşımı erken füzyona, yüz ve göz bebeği temsillerini bir öz-dikkat ağında bütünleştiren tasarım (Liu vd., 2024) ile çok-kipli çift doğrusal havuzlama (Uddin vd., 2022) ise ara füzyona karşılık gelir. Geç füzyona örnek olarak, ses, görüntü ve metin kiplerinin her biri için ayrı eğitilen modellerin tahminlerini ağırlıklı ortalamayla birleştiren AVEC 2016 topluluk sistemi verilebilir (Williamson vd., 2016). Genel eğilim erken ya da geç füzyondan kipler arası etkileşimi öğrenen dikkat tabanlı ara

füzyona doğrudur. Erken, ara ve geç füzyonun şematik karşılaştırması Şekil 4'te verilmektedir.



Şekil 4. Çok-kipli füzyon stratejileri. Erken (öznitelik düzeyi), ara (öğrenilmiş ara temsil düzeyi) ve geç (karar düzeyi) füzyon.

Füzyon stratejilerindeki yönelim COVID-19 salgını sonrasında belirgin biçimde hızlanan ve alanın genelini saran büyüme ve olgunlaşmanın da bir parçasıdır. Bu büyüme alanın ulaştığı teknolojik olgunlukla da uyumludur (He, Niu, vd., 2022; Low vd., 2020). Bu olgunlaşmanın ardında ruh sağlığı bilimlerindeki üç aşamalı yapısal dönüşüm yatar. Bu süreç geleneksel ve belirti merkezli yaklaşımlardan matematiksel ve hesaplamalı modellerle çalışan Hesaplamalı Psikiyatri'ye (computational psychiatry) geçişle başlamıştır. Akabinde duygu tespitine dayalı Duygusal Hesaplama modellerinin yükselişiyle sürmüş ve son olarak akıllı cihazlarla sürekli veri toplamayı mümkün kılan Dijital Fenotipleme (digital phenotyping) çağına ulaşmıştır. Benzer eğilim AVEC yarışma serisinin on yıl içinde temel duygu tanımadan klinik açıdan daha anlamlı problemlere (depresyon ve bipolar bozukluk tanıma) evrilmesinde de görülür (Valstar vd., 2013; Valstar vd., 2016). Şekil 5'te görüldüğü gibi bu üç evre birbiriyle ilişkilidir. Günümüz çalışmaları çoğu zaman her üç paradigmanın araçlarını birlikte kullanır. Nitekim klinik görüşme verisinden öğrenen modeller ile akıllı cihaz akışlarını birleştiren hibrit tasarımlar hem laboratuvar denetimini hem de ekolojik geçerliliği aynı çatı altında toplama eğilimindedir.



Şekil 5. Hesaplamalı ruh sağlığında üç-aşamalı paradigma evrimi.

5. Gerçek-Yaşam Uygulamaları ve Klinik Entegrasyon

Hesaplamalı yöntemlerin önemi klinik akışa entegrasyonlarıyla ölçülür. Literatür üç pratik kullanım alanına işaret etmektedir.

5.1. Uzaktan ve Sürekli Değerlendirme

Yüz yüze görüşmenin pandemi veya coğrafi mesafe nedeniyle mümkün olmadığı durumlarda akıllı telefon ve web tabanlı uygulamalar hasta ile klinisyenin eşzamanlı veya eşzamansız etkileşimine olanak tanır (Uddin vd., 2022; Xie vd., 2021). Yüz ve ses teknolojileriyle sürekli izlem, bireyin ruhsal durumundaki ani değişimleri (intihar riski veya ani duygulanım dalgalanmaları gibi) erken saptayarak hızlı müdahaleye olanak verir (Othmani vd., 2022; Wang vd., 2021). Telefon tabanlı uygulamalar ise günlük ruh hali ve belirti takibini mümkün kılarak hastanın kendi tedavisine etkin katılımını sağlar (Prabhu vd., 2022). Böylece hastayı sürecin edilgen bir nesnesi olmaktan çıkarıp aktif bir paydaşa dönüştürerek tedaviye bağlılığı güçlendirir. Bu uygulamalarda psikiyatrik değerlendirme klinik dışına taşınır ve kişiselleştirilmiş, proaktif bir izlem paradigması doğar. Sürekli izlem klinik ziyaretler arasındaki uzun boşluklarda belirti dalgalanmalarının gözden kaçmasını engelleyerek erken uyarı üretebilir ve tedavi planının zamanında uyarlanmasına olanak tanır.

5.2. Girişimsel Olmayan İzlem ve Mahremiyet

Kamera ve mikrofon temelli yöntemler kan testi veya beyin görüntüleme gibi girişimsel yöntemlere kıyasla daha hızlı, düşük maliyetli ve mahremiyeti koruyan bir alternatif sunar (Gilanie vd., 2022; Prabhu vd., 2022). Bu yaklaşım, damgalanma kaygısı yaşayan hastalar için tedaviye erişimi kolaylaştırır ve geleneksel yöntemlerin yarattığı psikolojik engelleri azaltır. Ses ve video temelli sistemlerin klinik doğruluğu korumanın yanında mahremiyeti önceliklendirmesi, uzun vadeli izlem için güven, erişilebilirlik ve etik temelli yeni bir standardın habercisi olarak değerlendirilebilir (Gilanie vd., 2022). Girişimsel olmayan bu yaklaşım, özellikle düşük sosyoekonomik düzeydeki bölgelerde yüksek maliyetli laboratuvar altyapısına olan bağımlılığı azaltır. Bununla birlikte kamera ve mikrofon verisinin sürekli toplanması, rıza yönetimi ve veri güvenliği konularında dikkatli protokoller gerektirir. Aksi halde mahremiyet avantajı hızla bir risk kaynağına dönüşebilir.

5.3. Farklı Kültürler ve Kısıtlı Kaynaklar

Yüz ifadeleri dil bariyerinden büyük ölçüde bağımsız bir gösterge sunarken (Mahayossanunt vd., 2023), ses temelli analizle birleştirildiğinde farklı kültürel bağlamlarda daha etkili sonuçlar elde edilebilir (Das ve Naskar, 2024). Akıllı

telefon, düşük çözünürlüklü kamera ve temel mikrofon gibi yaygın cihazlarla çalışabilen modeller klinik uzman kaynağının kısıtlı olduğu bölgelerde ruh sağlığı hizmetlerine erişimi artırma potansiyeli taşır (Gilanie vd., 2022; Xie vd., 2021). Bu yönüyle yapay zeka destekli sistemler yalnızca gelişmiş sağlık altyapısına sahip bölgelere değil, kaynakları kısıtlı topluluklara da uyarlanabilir. Düşük bant genişliğine ve sınırlı donanıma uyum sağlayan hafif modeller bu erişilebilirliğin teknik önkoşuludur. Ayrıca yerel dillere ve kültürel ifade biçimlerine uyarlanmış modeller küresel ölçekte adil ve temsil edici bir hizmet sunabilmek için kritik öneme sahiptir.

5.4. Açıklanabilirlik ve Klinik Güven

Klinik kabul için modelin yalnızca doğru değil, gerekçesinin de denetlenebilir olması gerekir. Bu gereksinim, açıklanabilir yapay zeka (XAI) yöntemlerini klinik uygulamaların merkezine taşır (Tjoa ve Guan, 2021). XAI yöntemleri iki ana gruba ayrılır. Birinci grup, kararı sonradan açıklayan post-hoc yöntemlerdir. Bunlar arasında evrişimli ağların hangi görüntü bölgelerine baktığını ısı haritalarıyla gösteren sınıf etkinleştirme haritaları ve Grad-CAM (Selvaraju vd., 2017) ile her bir özniteliğin karara katkısını oyun teorisine dayanan Shapley değerleriyle niceleyen SHAP (Lundberg ve Lee, 2017) öne çıkar. Nitekim bu bölümde ele alınan DepressNet'in ürettiği Depresyon Etkinleştirme Haritaları bu sınıf etkinleştirme soyağacının ruh sağlığına uyarlanmış bir örneğidir (Zhou vd., 2020). İkinci grup, açıklamayı modelin yapısına gömen içsel (intrinsic) yöntemlerdir. Dikkat ağırlıklarının görselleştirilmesi ve bütünleşik gradyanlarla beslenen yorumlanabilir mimariler bu gruba girer ve depresyon tespitinde doğrudan uygulanmıştır (Mahayossanunt vd., 2023; Xie vd., 2021). Klinik bağlamda XAI'nin değeri güven, güvenlik ve düzenleyici olmak üzere üç düzeyde açıklanabilir. Güven düzeyinde, klinisyenin modelin kararını sorgulayıp doğrulamasına olanak tanır. Güvenlik düzeyinde, modelin klinik dışı yapay ipuçlarına dayanıp dayanmadığını ortaya çıkararak yanlılığı görünür kılar. Düzenleyici düzeyde ise biyometrik veriyle çalışan tanı sistemlerinin onay süreçleri için gerekli şeffaflığı sağlar. Bununla birlikte post-hoc açıklamalar, kararın gerçek nedenini her zaman sadık biçimde yansıtmayabileceği, dolayısıyla açıklama yöntemlerinin kendisinin de doğrulanması gerektiği unutulmamalıdır (Tjoa ve Guan, 2021).

6. Tartışma ve Sınırlılıklar

Literatürdeki en belirgin örüntü tematik yoğunlaşmadır. Araştırmalar ağırlıklı olarak depresyona odaklanmış; psikoz çok az çalışılmış (Hall vd., 2024), anksiyete bozuklukları ise neredeyse hiç ele alınmamıştır. Bunun nedenleri arasında depresyonun ses ve yüzde görece kolay yakalanabilen

izler bırakması (düşük vokal ton, monoton prozodi, yavaş konuşma hızı, azalmış mimik ve göz teması) ile AVEC (Valstar vd., 2013; Valstar vd., 2016) ve DAIC-WOZ (Gratch vd., 2014) gibi standart, erişilebilir kümelerin araştırmayı bu yöne çekmesi yer alır. Nitekim DAIC-WOZ anksiyete için de tasarlanmış olsa da pratikte yalnızca depresyon ve TSSB (travma sonrası stres bozukluğu) etiketleri yaygın biçimde kullanılmıştır. Bu yoğunlaşma, depresyon modellerini giderek olgunlaştırırken, anksiyete, bipolar bozukluk ve şizofreni gibi belirtileri kısmen örtüşen durumlar için hesaplamalı yöntemlerin gelişimini zorlaştırmaktadır. Oysa bu bozukluklarda da depresyondakine benzer davranışsal ve fizyolojik göstergeler bulunduğundan tanılar arası ortak örüntüleri yakalayabilen modellere ihtiyaç vardır. Depresyonla klinik açıdan en sık iç içe geçen bozukluklardan biri olan anksiyete, bu örüntülerin modellenmesi bakımından kendine özgü sorunlar barındırır.

Anksiyetenin otomatik tespitini güçleştiren bu sorunlar dört başlık altında toplanabilir:

- Anksiyete belirtileri bağlama bağlı ve epizodiktir. Örneğin, sosyal anksiyetesi olan bireyler düşük stresli ortamlarda normal görünüp yalnızca belirli tetikleyiciler altında klinik belirti gösterebilir. Bu da tek seanslık verinin tanısal geçerliliğini sınırlar.
- Aşırı endişe ve ruminasyon gibi çekirdek belirtiler büyük ölçüde içsel bilişsel süreçlerdir. Depresyonun daha belirgin somatik belirtilerinin aksine, bu gizli nitelikler otomatik tespiti zorlaştırır.
- Eş tanılı (komorbid) depresyon-anksiyete tablolarında, anksiyetenin perdeyi yükseltme eğilimi depresyonun perdeyi düşürme etkisini maskeleyebilir. Bu da algoritmaları yanıltan, karşılıklı örtüşen akustik sinyaller üretir.
- Kamera ve kayıt cihazları bir gözlemci paradoksu (observer paradox) yaratır. Kaydedilmek, sosyal anksiyetenin çekirdeğindeki 'değerlendirilme korkusunu' tetikleyerek ağır belirtili hastaların katılımını engelleyebilir ve veriyi yanlı hale getirebilir.

İkinci önemli sorun tanı-tedavi boşluğudur. Mevcut modeller ağırlıklı olarak tanıya odaklanırken, gerçek klinik ve ekonomik yük, uzun vadeli tedavi yönetimi, belirti izlemi ve nüks riski tahmininde yatar. Bu boşluğun kapatılması üç düzeyde engelin aşılmasını gerektirir. Teknolojik düzeyde, derin öğrenmenin 'kara kutu' niteliği klinisyenlerin kararların ardındaki sinyalleri anlamasını engelleyerek modellerin klinik kabulünü güçleştirir. Bu nedenle yorumlanabilir ve açıklanabilir mimariler kritik önemdedir (Mahayossanunt vd., 2023; Xie vd., 2021). Etik ve yasal düzeyde, yüz görüntüsü ve ses kaydı

gibi biyometrik veriler GDPR ve HIPAA kapsamında oldukça hassastır. Verileri toplama, saklama ve kurumlar arası paylaşım üzerindeki sınırlamalar ham veri paylaşımını kısıtlayarak metodolojik durağanlık riski yaratır. Yöntemsel düzeyde ise kesitsel tasarımlar tedavi yanıtı, belirti seyri ve nüks riski gibi boylamsal soruları yanıtlayamaz. Bunun yerine hastaları haftalar veya aylar boyunca izleyen zaman serisi verilerine ihtiyaç vardır. Anksiyete gibi az çalışılan alanlarda veri sınırlılığını aşmak için bağlama duyarlı protokoller, sanal gerçeklik temelli uyarım ve akıllı cihazlarla pasif algılama önerilmektedir. Depresyon-anksiyete eş tanısı için ise çoklu etiketli ve standartlaştırılmış modelleme çerçeveleri gereklidir.

Üçüncü sorun veri kümelerinde çeşitliliğin eksikliğidir. Yaygın kümeler büyük ölçüde İngilizce ve ağırlıklı olarak batılı popülasyonlardan oluşur. Oysa dil, etnik köken ve kültürün depresyon belirtilerinin görünümünü değiştirebildiği bilinmektedir. Bu boşluğa yanıt olarak kimi araştırmacılar Korece (Kim vd., 2023), Tayca (Mahayossanunt vd., 2023) ve Çince (Yang vd., 2023) veri kümeleri oluşturmuştur. Psikotik bozukluklar üzerine çalışan ekipler ise kendi popülasyonlarına özgü kümeler toplamak zorunda kalmıştır (Hall vd., 2024). Bu durum yetersiz temsil edilen bozukluklar için erişilebilir verinin ne denli sınırlı olduğunu açıkça göstermektedir. Bu açığı kapatmaya yönelik bir başka yaklaşım RAVDESS (Livingstone ve Russo, 2018), CREMA-D (Cao vd., 2014) ve eNTERFACE'05 (Martin vd., 2006) gibi profesyonel oyunculardan oluşturulan duygusal ifade kümelerini örnek alarak mahremiyet ve ham veri erişimi sorunlarını azaltan denetimli kümeler üretmektir. Dil, kültür ve etnik köken açısından dengeli, boylamsal ve çok bozukluklu kümelerin geliştirilmesi modellerin genellenebilirliği için belirleyici olabilir.

Bu sınırlılıkların önemli bir kısmı, alanın büyük veri ve temel model (Foundation Model) çağına geçişiyle yeniden çerçevelenmektedir. Temel modeller etiketsiz devasa veri kümeleri üzerinde öz-denetimli biçimde önceden eğitilen ve çok sayıda alt göreve uyarlanabilen genel amaçlı modellerdir (Bommasani vd., 2021). Tıpta bu yaklaşım, farklı kipleri tek bir modelde birleştiren genel tıbbi yapay zeka vizyonuna doğru ilerlemektedir (Moor vd., 2023). Ruh sağlığı için bu geçiş iki yönlü bir fırsat sunar. Bir yandan konuşma ve görüntü temel modelleri sınırlı klinik veriyle dahi güçlü temsiller sağlayarak veri kısıtı sorununu hafifletebilir ve dijital fenotipleme akışlarındaki büyük ölçekli, etiketsiz veriyi değerlendirilebilir kılabilir (Insel, 2017; Dwyer vd., 2018). Öte yandan bu modeller yeni riskleri de beraberinde getirir. Eğitim verisindeki demografik dengesizlikleri ölçekleyerek pekiştirilebilir, hesaplama maliyetleri nedeniyle kaynakları kısıtlı ortamlar için erişilemez hale gelebilir ve kararlarının yorumlanması güçleştiğinden açıklanabilirlik gereksinimi karşılanamayabilir. Dolayısıyla temel modeller çeşitlilik, mahremiyet ve

yorumlanabilirlik sorunlarını ortadan kaldırmaz. Aksine, bu sorunları daha büyük ölçekte yeniden üretme riski taşıdıklarından dikkatli bir değerlendirme gerektirir.

7. Sonuç ve Öneriler

Bu alandaki, özellikle depresyon tespitindeki hesaplamalı yöntemler kayda değer biçimde olgunlaşmıştır. Ancak bu birikimin klinik etkiye dönüşmesi için aşılması gereken yapısal güçlükler vardır. Bu güçlüklerin aşılmasında dört stratejik yönelim öne çıkmaktadır.

- Hedeflenmiş ve temsil gücü yüksek veri kümeleri geliştirilmelidir. Yalnızca daha çeşitli değil, anksiyete, psikoz ve bipolar bozukluk gibi yeterince temsil edilmeyen durumları hedefleyen, kültürler arası ve boylamsal kümelere öncelik verilmelidir. Federe öğrenme gibi mahremiyet koruyan yaklaşımlar ya da RAVDESS (Livingstone ve Russo, 2018), CREMA-D (Cao vd., 2014) ve eNTERFACE'05 (Martin vd., 2006) gibi kümeleri örnek alarak profesyonel oyuncularla üretilen kümeler etik kısıtlarla bilimsel ilerleme arasındaki gerilimi azaltabilir.
- Araştırma kapsamı klinik sürekliliğe genişletilmelidir. Tek seferlik tanı modellerinden tedavi yanıtını öngören, belirti şiddetini dinamik olarak izleyen ve nüks riskini değerlendiren modellere geçilmelidir (Othmani vd., 2022). Böylece bu teknolojiler kişiselleştirilmiş tedaviyi destekleyen aktif klinik karar destek sistemlerine dönüşebilir.
- Klinik kullanım için yorumlanabilirlik önceliklendirilmelidir. Bir modelin kliniğe geçişi yalnızca doğruluğuna değil, güvenilirliğine ve şeffaflığına da bağlıdır. Bu nedenle açıklanabilir yapay zeka yöntemleri standart bir uygulama olarak teşvik edilmelidir (Mahayossanunt vd., 2023; Xie vd., 2021).
- Çok-kipli temel modeller ve dönüştürücü mimariler, alanın umut vadeden gelecek yönelimlerinden biri olarak öne çıkmaktadır. Dikkat mekanizmasına dayanan dönüştürücü mimari (Vaswani vd., 2017) uzun menzilli zamansal bağımlılıkları ve kipler arası etkileşimleri tek bir çatıda modelleyebildiğinden ses, görüntü ve metin ortak bir temsil uzayında birleştiren çok-kipli dönüştürücüler (multimodal transformers) için doğal bir zemin oluşturur (Xu vd., 2023). Etiketsiz büyük veride önceden eğitilip klinik göreve uyarlanan temel modellerle birleştiğinde bu mimariler, veri kısıtı sorununu azaltma ve tek bir modelle birden çok ruhsal bozukluğu ve klinik aşamayı kapsama potansiyeli taşır (Bommasani vd., 2021; Moor vd., 2023). Ancak bu yönelimin yukarıda

vurgulanan yanlılık, hesaplama maliyeti ve açıklanabilirlik koşullarıyla birlikte ilerlemesi gerekir.

Sonuç olarak, bu alandaki bir sonraki büyük atılım yalnızca model doğruluğunu artırmaya değil, araştırma odağını daha dengeli kılmaya, metodolojik ve etik güçlüklere yaratıcı çözümler üretmeye ve en önemlisi, klinik ihtiyaçları doğrudan karşılayan güvenilir ve yorumlanabilir sistemler inşa etmeye dayanmalıdır. Bu dönüşüm yalnızca teknik bir ilerleme değil, mühendislik, klinik bilimler ve etik arasında sürdürülebilir bir iş birliği gerektiren disiplinler arası bir olgunlaşma sürecidir. Alanın kendi başarısını ölçme biçimini doğruluk odaklı ölçütlerden klinik fayda odaklı ölçütlere kaydırması, bu olgunlaşmanın belirleyici göstergesi olabilir. Nihayetinde işitsel ve görsel verilerden elde edilen nesnel biyobelirteçlerin değeri laboratuvar başarımıyla değil, gerçek hastaların tanı, tedavi ve izlem yolculuğuna kattığı somut iyileşmeyle ölçülebilir.

Kaynakça

Adcock, A., & Parkin, L. (2016). Report from the independent Mental Health Taskforce to the NHS in England. House of Commons Library.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 33, 12449–12460.

Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., … Liang, P. (2021). On the opportunities and risks of foundation models. arXiv. https://doi.org/10.48550/arXiv.2108.07258

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). CREMA-D: Crowd-sourced emotional multimodal actors dataset. IEEE Transactions on Affective Computing, 5(4), 377–390. https://doi.org/10.1109/TAFFC.2014.2336244

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., … Wei, F. (2022). WavLM: Large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing, 16(6), 1505–1518. https://doi.org/10.1109/JSTSP.2022.3188113

Das, A. K., & Naskar, R. (2024). A deep learning model for depression detection based on MFCC and CNN-generated spectrogram features. Biomedical Signal Processing and Control, 90, 105898. https://doi.org/10.1016/j.bspc.2023.105898

Degottex, G., Kane, J., Drugman, T., Raitio, T., & Scherer, S. (2014). COVAREP, a collaborative voice analysis repository for speech technologies. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 960–964).

de Melo, W. C., Granger, E., & Hadid, A. (2020). A deep multiscale spatiotemporal network for assessing depression from facial dynamics. IEEE Transactions on Affective Computing, 13(3), 1581–1592. https://doi.org/10.1109/TAFFC.2020.3021755

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. Annual Review of Clinical Psychology, 14, 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

GBD 2019 Mental Disorders Collaborators. (2022). Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019. The Lancet Psychiatry, 9(2), 137–150. https://doi.org/10.1016/S2215-0366(21)00395-3

Gilanie, G., Asghar, M., Qamar, A. M., Ullah, H., Khan, R. U., Aslam, N., & Khan, I. U. (2022). An automated and real-time approach of depression detection from facial micro-expressions. Computers, Materials & Continua, 73(2). https://doi.org/10.32604/cmc.2022.028229

Gratch, J., Artstein, R., Lucas, G. M., Stratou, G., Scherer, S., Nazarian, A., … Marsella, S. (2014). The Distress Analysis Interview Corpus of human and computer interviews (DAIC-WOZ). In Proceedings of LREC (Vol. 14, pp. 3123–3128).

Hall, N. T., Hallquist, M. N., Martin, E. A., Lian, W., Jonas, K. G., & Kotov, R. (2024). Automating the analysis of facial emotion expression dynamics: A computational framework and application in psychotic disorders. Proceedings of the National Academy of Sciences, 121(14), e2313665121. https://doi.org/10.1073/pnas.2313665121

He, L., Guo, C., Tiwari, P., Su, R., Pandey, H. M., & Dang, W. (2022). DepNet: An automated industrial intelligent system using deep learning for video-based depression analysis. International Journal of Intelligent Systems, 37(7), 3815–3835. https://doi.org/10.1002/int.22704

He, L., Niu, M., Tiwari, P., Marttinen, P., Su, R., Jiang, J., … Wang, Z. (2022). Deep learning for depression recognition with audiovisual cues: A review. Information Fusion, 80, 56–86. https://doi.org/10.1016/j.inffus.2021.10.012

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 3451–3460. https://doi.org/10.1109/TASLP.2021.3122291

Huys, Q. J. M., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. Nature Neuroscience, 19(3), 404–413. https://doi.org/10.1038/nn.4238

Insel, T. R. (2017). Digital phenotyping: Technology for a new science of behavior. JAMA, 318(13), 1215–1216. https://doi.org/10.1001/jama.2017.11295

Kim, A. Y., Jang, E. H., Lee, S.-H., Choi, K.-Y., Park, J. G., & Shin, H.-C. (2023). Automatic depression detection using smartphone-based speech signals: Deep CNN approach. Journal of Medical Internet Research, 25, e34474. https://doi.org/10.2196/34474

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

Liu, X., Shen, H., Li, H., Tao, Y., & Yang, M. (2024). Multimodal depression detection based on self-attention network with facial expression and pupil. IEEE Transactions on Computational Social Systems. https://doi.org/10.1109/TCSS.2024.3405949

Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). PLOS ONE, 13(5), e0196391. https://doi.org/10.1371/journal.pone.0196391

Low, D. M., Bentley, K. H., & Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. Laryngoscope Investigative Otolaryngology, 5(1), 96–116. https://doi.org/10.1002/lio2.354

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

Mahayossanunt, Y., Nupairoj, N., Hemrungrojn, S., & Vateekul, P. (2023). Explainable depression detection based on facial expression using LSTM on attentional intermediate feature fusion with label smoothing. Sensors, 23(23), 9402. https://doi.org/10.3390/s23239402

Martin, O., Kotsia, I., Macq, B., & Pitas, I. (2006). The eNTERFACE'05 audio-visual emotion database. In IEEE International Conference on Data Engineering Workshops (ICDEW) (p. 8). https://doi.org/10.1109/ICDEW.2006.145

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. Trends in Cognitive Sciences, 16(1), 72–80. https://doi.org/10.1016/j.tics.2011.11.018

Monteith, S., Glenn, T., Geddes, J., & Bauer, M. (2015). Big data are coming to psychiatry: A general introduction. International Journal of Bipolar Disorders, 3(1), 21. https://doi.org/10.1186/s40345-015-0038-9

Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. Nature, 616(7956), 259–265. https://doi.org/10.1038/s41586-023-05881-4

Muzammel, M., Salam, H., Hoffmann, Y., Chetouani, M., & Othmani, A. (2020). AudVowelConsNet: A phoneme-level based deep CNN architecture for clinical depression diagnosis. Machine Learning with Applications, 2, 100005. https://doi.org/10.1016/j.mlwa.2020.100005

Othmani, A., Zeghina, A.-O., & Muzammel, M. (2022). A model of normality inspired deep learning framework for depression relapse prediction using audiovisual data. Computer Methods and Programs in Biomedicine, 226, 107132. https://doi.org/10.1016/j.cmpb.2022.107132

Prabhu, S., Mittal, H., Varagani, R., Jha, S., & Singh, S. (2022). Harnessing emotions for depression detection. Pattern Analysis and Applications, 25(3), 537–547. https://doi.org/10.1007/s10044-021-01020-9

Santomauro, D. F., Mantilla Herrera, A. M., Shadid, J., Zheng, P., Ashbaugh, C., Pigott, D. M., … Ferrari, A. J. (2021). Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020

due to the COVID-19 pandemic. The Lancet, 398(10312), 1700–1712. https://doi.org/10.1016/S0140-6736(21)02143-7

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 618–626). https://doi.org/10.1109/ICCV.2017.74

Stahlschmidt, S. R., Ulfenborg, B., & Synnergren, J. (2022). Multimodal deep learning for biomedical data fusion: A review. Briefings in Bioinformatics, 23(2), bbab569. https://doi.org/10.1093/bib/bbab569

Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Transactions on Neural Networks and Learning Systems, 32(11), 4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314

Uddin, M. A., Joolee, J. B., & Sohn, K.-A. (2022). Deep multi-modal network based automated depression severity estimation. IEEE Transactions on Affective Computing, 14(3), 2153–2167. https://doi.org/10.1109/TAFFC.2022.3179478

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., … Pantic, M. (2013). AVEC 2013: The continuous audio/visual emotion and depression recognition challenge. In Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge.

Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., … Pantic, M. (2016). AVEC 2016: Depression, mood, and emotion recognition workshop and challenge. In Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

Wang, H., Liu, Y., Zhen, X., & Tu, X. (2021). Depression speech recognition with a three-dimensional convolutional network. Frontiers in Human Neuroscience, 15, 713823. https://doi.org/10.3389/fnhum.2021.713823

Williamson, J. R., Godoy, E., Cha, M., Schwarzentruber, A., Khorrami, P., Gwon, Y., Kung, H.-T., Dagli, C., & Quatieri, T. F. (2016). Detecting depression using vocal, facial and semantic communication cues. In Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge (pp. 11–18). https://doi.org/10.1145/2988257.2988263

World Health Organization. (2022a). COVID-19 pandemic triggers 25% increase in prevalence of anxiety and depression worldwide. WHO News.

World Health Organization. (2022b). Mental health: Strengthening our response. WHO Fact Sheet. https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response

World Health Organization. (2022c). Mental disorders. WHO Fact Sheet. https://www.who.int/news-room/fact-sheets/detail/mental-disorders

Wu, W., Zhang, C., & Woodland, P. C. (2023). Self-supervised representations in speech-based depression detection. In ICASSP 2023 - IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 1–5). IEEE. https://doi.org/10.1109/ICASSP49357.2023.10094910

Xie, W., Liang, L., Lu, Y., Wang, C., Shen, J., Luo, H., & Liu, X. (2021). Interpreting depression from question-wise long-term video recording of SDS evaluation. IEEE Journal of Biomedical and Health Informatics, 26(2), 865–875. https://doi.org/10.1109/JBHI.2021.3092628

Xu, N., Huo, H., Xu, J., Ma, L., & Wang, J. (2024). Automatic diagnosis of depression based on attention mechanism and feature pyramid model. PLoS One, 19(3), e0295051. https://doi.org/10.1371/journal.pone.0295051

Xu, P., Zhu, X., & Clifton, D. A. (2023). Multimodal learning with transformers: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(10), 12113–12132. https://doi.org/10.1109/TPAMI.2023.3275156

Yang, W., Liu, J., Cao, P., Zhu, R., Wang, Y., Liu, J. K., … Zhang, X. (2023). Attention guided learnable time-domain filterbanks for speech depression detection. Neural Networks, 165, 135–149. https://doi.org/10.1016/j.neunet.2023.05.041

Zhang, X., Zhang, X., Chen, W., Li, C., & Yu, C. (2024). Improving speech depression detection using transfer learning with wav2vec 2.0 in low-resource environments. Scientific Reports, 14, 9543. https://doi.org/10.1038/s41598-024-60278-1

Zhou, X., Jin, K., Shang, Y., & Guo, G. (2020). Visually interpretable representation learning for depression recognition from facial images. IEEE Transactions on Affective Computing, 11(3), 542–552. https://doi.org/10.1109/TAFFC.2018.2828819