T.C.
KONYA TEKNİK ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

EVRİŞİMSEL SİNİR AĞLARI KULLANILARAK VİDEO TABANLI İZOLE İŞARET DİLİ TANIMA

Ali AKDAĞ

DOKTORA TEZİ

Bilgisayar Mühendisliği Anabilim Dalı

Haziran-2024
KONYA

TEZ KABUL VE ONAYI

Ali AKDAĞ tarafından hazırlanan "**Evrişimsel Sinir Ağları Kullanılarak Video Tabanlı İzole İşaret Dili Tanıma**" adlı tez çalışması 26/06/2024 tarihinde aşağıdaki jüri tarafından oy birliği ile Konya Teknik Üniversitesi Lisansüstü Eğitim Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda DOKTORA TEZİ olarak kabul edilmiştir.

Jüri Üyeleri	**İmza**
Başkan Prof. Dr. Murat CEYLAN (Konya Teknik Üniversitesi)	……..
Danışman Doç. Dr. Ömer Kaan BAYKAN (Konya Teknik Üniversitesi)	……..
Üye Prof. Dr. Erkan ÜLKER (Konya Teknik Üniversitesi)	……..
Üye Doç. Dr. Gür Emre GÜRAKSIN (Afyon Kocatepe Üniversitesi)	……..
Üye Doç. Dr. Tahir SAĞ (Selçuk Üniversitesi)	……..

Yukarıdaki sonucu onaylarım.

Prof. Dr. Mevlüt UYAN
Enstitü Müdürü

TEZ BİLDİRİMİ

Bu tezdeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edildiğini ve tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

DECLARATION PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Ali AKDAĞ

Tarih:

ÖZET

DOKTORA TEZİ

EVRİŞİMSEL SİNİR AĞLARI KULLANILARAK VİDEO TABANLI İZOLE İŞARET DİLİ TANIMA

Ali AKDAĞ

Konya Teknik Üniversitesi
Lisansüstü Eğitim Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Doç. Dr. Ömer Kaan BAYKAN

2024, 120 Sayfa

Jüri
Doç. Dr. Ömer Kaan BAYKAN
Prof. Dr. Murat CEYLAN
Prof. Dr. Erkan ÜLKER
Doç. Dr. Gür Emre GÜRAKSIN
Doç. Dr. Tahir SAĞ

İşaret dili dünya genelinde milyonlarca işitme engelli birey için temel bir iletişim aracıdır. Ancak, işaret dilini anlamak ve kullanmak, işitenler arasında yaygın bir beceri değildir, bu da işitme engelli bireyler arasında sosyal izolasyon riskini artırır. Bu tez, kelime tabanlı İşaret Dili Tanıma (İDT - Sign Language Recognition, SLR) teknolojilerindeki mevcut kısıtlamaları ele alarak, bu alandaki algılama doğruluğunu ve genellenebilirliğini artırmayı hedeflemektedir. Bu kapsamda üç ana çalışma üzerinden, işaret dilinin manuel ve manuel olmayan unsurları kapsamlı bir şekilde analiz edilerek, derin öğrenme tabanlı sistemler sunulmuştur.

İlk çalışmada, R3D ve R(2+1)D evrişim bloklarının avantajlarını birleştiren R3(2+1)D-SLR ağı önerilmiştir. Bu ağ, uzamsal ve zamansal özellikleri etkili bir şekilde çıkararak, işaret dili tanımada yüksek doğruluk ve sağlamlık sunar. R3(2+1)D-SLR tabanlı geliştirilen işaret dili tanıma sistemi, işaretçinin vücut, el ve yüzünden elde edilen verileri bir araya getirerek, Destek Vektör Makinesi (DVM) kullanımıyla sınıflandırma yapmaktadır. Önerilen sistemde RGB verileri yerine görsel poz verileri kullanılmasıyla arka plan çeşitliliğine karşı doğruluk ve sağlamlıkta önemli iyileştirmeler sağlandığı gösterilmiştir. Bu sistem BosphorusSign22k-genel ve LSA64 veri kümelerinde işaretçiden bağımsız değerlendirmelerde %94,52 ve %98,53 test doğruluğu elde etmiştir.

İkinci çalışma, izole İDT görevi için yenilikçi bir yaklaşım sunmaktadır, bu yaklaşım poz verilerini, bu verilerden türetilen Hareket Tarihçesi Görüntüleri (HTG) ile entegre etmeye odaklanır. Araştırma, vücut, el ve yüz pozlarından elde edilen uzamsal bilgileri işaretin zamansal dinamiklerini yansıtan üç kanallı HTG verileriyle bütünleştirir. Özellikle, geliştirilen parmak pozu tabanlı HTG özelliği, İDT'deki mevcut yaklaşımlara göre parmak hareketlerinin ve jestlerin nüanslarını daha başarılı bir şekilde yakalamaktadır. Bu özellik, işaret dilinin zengin detaylarını daha doğru bir şekilde işleyerek sistemin doğruluğunu ve güvenilirliğini artırmaktadır. Ek olarak, doğrusal enterpolasyon kullanılarak eksik poz verilerinin tamamlanması genel model performansını iyileştirmiştir. Rastgele Sızdıran Düzeltilmiş Doğrusal Birim (RReLU) ile güçlendirilmiş ResNet-18 modeli temelinde elde edilen özelliklerin birleşimi ve DVM ile sınıflandırma yoluyla manuel ve manuel olmayan özellikler arasındaki etkileşim başarıyla ele alınmıştır. Bu entegre yöntem, BosphorusSign22k-genel, BosphorusSign22k, LSA64 ve GSL veri kümelerinde yapılan deneylerde sırasıyla %96,94, %94,87, %98,68 ve %95,14 doğruluk elde ederek mevcut metodolojilere kıyasla rekabetçi ve üstün sonuçlar göstermiştir.

Üçüncü çalışma, işaret dili tanımada parmakların özelliklerine ve konfigürasyonlarına odaklanarak yenilikçi bir, çok kanallı yaklaşım sunmaktadır. Ayrı kanallarda işlenen görsel parmak pozu verilerine dayanan bu yaklaşım, parmak hareketlerinin detaylı analizini sağlamak üzere tasarlanmıştır. Önerilen Çok-Kanallı MobileNetV2 modeli, parmaklara dair çok kanallı verileri kullanarak işaret dili tanıma sürecinde yüksek doğruluk ve hassasiyet sunmaktadır. Çalışma ayrıca, poz verilerinden elde edilen vücut ve yüz bilgilerinin işlenmesiyle, işaret dilinin manuel olmayan özelliklerini de entegre etmektedir. Önerilen sistem, BosphorusSign22k-genel, BosphorusSign22k, LSA64 ve GSL veri kümeleri üzerinde sırasıyla %97,15, %95,13, %98,93 ve %95,37 gibi kayda değer doğruluk oranları elde etmiştir. Bu sonuçlar, önerilen yöntemin genellenebilirliğini ve uyarlanabilirliğini vurgulayarak, işaret dili tanıma literatüründeki mevcut çalışmalara göre rekabet üstünlüğünü kanıtlamaktadır.

Bu tez, işaret dili tanıma teknolojilerindeki yenilikçi yaklaşımların, işaret dilinin zenginliğini ve ince ayrıntılarını daha doğru bir şekilde yakalayarak iletişim engellerini azaltma potansiyeline işaret etmektedir. Her üç çalışma da farklı veri kümelerinde yüksek doğruluk oranları elde ederek, pratik uygulamalarda İDT sistemlerinin etkinliğini ve güvenilirliğini artırmıştır.

Anahtar Kelimeler: Derin Öğrenme, Evrişimsel Sinir Ağları, İnsan-Bilgisayar Etkileşimi, İşaret Dili Tanıma, Makine Öğrenmesi, Özellik Füzyonu

ABSTRACT

PhD THESIS

VIDEO-BASED ISOLATED SIGN LANGUAGE RECOGNITION USING CONVOLUTIONAL NEURAL NETWORKS

Ali AKDAĞ

Konya Technical University
Institute of Graduate Studies
Department of Computer Engineering

Advisor: Assoc. Prof. Dr. Ömer Kaan BAYKAN

2024, 120 Pages

Jury
Assoc. Prof. Dr. Ömer Kaan BAYKAN
Prof. Dr. Murat CEYLAN
Prof. Dr. Erkan ÜLKER
Assoc. Prof. Dr. Gür Emre GÜRAKSIN
Assoc. Prof. Dr. Tahir SAĞ

Sign language is a fundamental communication tool for millions of hearing-impaired individuals worldwide. However, understanding and using sign language is not a common skill among hearing individuals, which increases the risk of social isolation for the hearing-impaired. This thesis addresses the current limitations in word-based Sign Language Recognition (SLR) technologies, aiming to enhance the accuracy and generalizability of detection in this field. In this context, through three main studies, both the manual and non-manual elements of sign language are comprehensively analyzed, presenting deep learning-based systems.

In the first study, the R3(2+1)D-SLR network, which combines the advantages of R3D and R(2+1)D convolutional blocks, was proposed. This network effectively extracts spatial and temporal features, providing high accuracy and robustness in sign language recognition. The sign language recognition system developed based on the R3(2+1)D-SLR architecture integrates data obtained from the signer's body, hands, and face, and classifies it using Support Vector Machines (SVM). The proposed system demonstrates significant improvements in accuracy and robustness against background variability by using visual pose data instead of RGB data. This system achieved test accuracies of 94,52% and 98,53% in signer-independent evaluations on the BosphorusSign22k-general and LSA64 datasets, respectively.

The second study presents an innovative approach for the task of isolated SLR, focusing on integrating pose data with derived Motion History Images (MHI). The research combines spatial information obtained from body, hand, and face poses with three-channel MHI data that reflect the temporal dynamics of the sign. Notably, the developed finger-pose-based MHI feature captures the nuances of finger movements and gestures more successfully than current approaches in SLR. This feature enhances the system's accuracy and reliability by more accurately processing the rich details of sign language. Additionally, the use of linear interpolation to complete missing pose data has improved overall model performance. The combination of features obtained from the ResNet-18 model enhanced with Randomized Leaky Rectified Linear Units (RReLU) and classification through SVM has successfully addressed the interaction between manual and non-manual features. This integrated method has demonstrated competitive and superior results compared to existing methodologies, achieving accuracies of 96,94%, 94,87%, 98,68%, and 95,14% in experiments conducted on the BosphorusSign22k-general, BosphorusSign22k, LSA64, and GSL datasets, respectively.

The third study introduces an innovative multi-channel approach focusing on the characteristics and configurations of fingers in sign language recognition. Based on visual finger pose data processed in separate channels, this approach is designed to provide detailed analysis of finger movements. The proposed Multi-Channel MobileNetV2 model utilizes multi-channel data on fingers to offer high accuracy and precision in the sign language recognition process. Additionally, the study incorporates non-manual features of sign language by processing body and face information derived from pose data. The proposed system has achieved notable accuracy rates of 97,15%, 95,13%, 98,93%, and 95,37% on the BosphorusSign22k-general, BosphorusSign22k, LSA64, and GSL datasets, respectively. These results highlight the generalizability and adaptability of the proposed method, proving its competitive superiority over existing studies in the sign language recognition literature.

This thesis indicates that innovative approaches in sign language recognition technology have the potential to reduce communication barriers by more accurately capturing the richness and subtle details of sign language. All three studies have achieved high accuracy rates across different datasets, enhancing the effectiveness and reliability of SLR systems in practical applications.

Keywords: Convolutional Neural Networks, Deep Learning, Feature Fusion, Human-Computer Interaction, Machine Learning, Sign Language Recognition

ÖNSÖZ

Tez çalışmam süresince gösterdiği üstün ilgi, değerli bilgileri ve teşvik edici yönlendirmeleriyle bana her zaman destek olan tez danışmanım Doç. Dr. Ömer Kaan BAYKAN'a en içten teşekkürlerimi sunarım.

Tez izleme komitesi üyesi olarak çalışmalarıma yön veren ve bilgi birikimleriyle beni aydınlatan değerli hocalarım, Prof. Dr. Erkan ÜLKER ve Prof. Dr. Murat CEYLAN'a teşekkür ederim.

Tokat Gaziosmanpaşa Üniversitesi ve Konya Teknik Üniversitesi Bilgisayar Mühendisliği bölümlerinde görev yapan değerli öğretim elemanlarına akademik gelişimim süresince verdikleri katkılar için teşekkür ediyorum.

Son olarak, emeklerini hiçbir zaman esirgemeyen ve her zaman bana destek ve güç veren aileme sonsuz teşekkürlerimi sunuyorum.

Ali AKDAĞ
KONYA-2024

İÇİNDEKİLER

SİMGELER VE KISALTMALAR

Simgeler

δ	:	Hareket Tarihçesi Görüntüsünde bozunma parametresi
$*$	:	Evrişim işlemi
Δ	:	Genişleme miktarı
∇_θ	:	Gradyan
$\in$	:	Elemanıdır sembolü
a	:	Rastgele örneklenen bir katsayı
C	:	Kanal sayısı
D	:	İki ardışık video çerçevesi arasındaki fark
D^k	:	Manhattan uzaklığı
e	:	Euler sayısı
f_{1D}	:	1D temel evrişim fonksiyonu
f_{2D}	:	2D temel evrişim fonksiyonu
f_{3D}	:	3D temel evrişim fonksiyonu
H	:	Yükseklik
H_τ	:	Hareket Tarihçesi Görüntüsü
I	:	Görüntü matrisi
J	:	Kayıp fonksiyonudur
K	:	Evrişim çekirdeği veya filtresi
K_{1D}	:	1D evrişim çekirdeği
K_{2D}	:	2D evrişim çekirdeği
K_{3D}	:	3D evrişim çekirdeği
L	:	İşaret noktalar kümesi
l	:	Bir poz işaret noktası
M	:	MediaPipe Holistic
O	:	Evrişim işlemi sonucu oluşan çıkış matrisi
T	:	Boş siyah tuval
t	:	Zaman
V	:	Video
v	:	Öz vektör
v_t	:	Gradyan güncellemelerinin birikimi
W	:	Genişlik
w	:	Hiperdüzlem normal vektörü
X	:	Veri matrisi
Y	:	Gerçek değerler
$\hat{Y}$	:	Tahmin edilen değerler
Z	:	Özellik vektörü
β	:	Yığın normalleştirmede kaydırma parametresi
γ	:	Yığın normalleştirmede ölçekleme parametresi
η	:	Öğrenme oranı
θ	:	Model parametreleri
λ	:	Özdeğer
ξ	:	Eşik değeri
τ	:	Hareket Tarihçesi Görüntüsünde bir karenin hareket süresi
ψ	:	Hareketin varlığını tanımlayan güncelleme fonksiyonu

Kısaltmalar

BCE	:	İkili Çapraz Entropi (Binary Cross-Entropy)
CAMSHIFT	:	Sürekli Uyarlanabilir Ortalama Kaydırma (Continuously Adaptive Mean Shift)
CPU	:	Merkezi İşlem Birimi (Central Processing Unit)
DVM	:	Destek Vektör Makinesi (Support Vector Machine)
ESA	:	Evrişimsel Sinir Ağları (Convolutional Neural Networks)
FN	:	Yanlış Negatifler (False Negatives)
FP	:	Yanlış Pozitifler (False Positives)
GPU	:	Grafik İşlem Birimi (Graphics Processing Unit)
GRU	:	Geçitli Tekrarlayan Birim (Gated Recurrent Unit)
HMM	:	Gizli Markov Modeli (Hidden Markov Model)
HOF	:	Optik Akış Histogramı (Histogram of Optical Flow)
HOG	:	Yönlendirilmiş Gradyanların Histogramı (Histogram Of Oriented Gradients)
HTG	:	Hareket Tarihçesi Görüntüsü (Motion History Image)
I3D	:	Genişletilmiş 3D Evrişimsel Ağ (Inflated 3D Conv. Network)
IDT	:	Geliştirilmiş Yoğun Yörünge (Improved Dense Trajectory)
İDT	:	İşaret Dili Tanıma (Sign Language Recognition)
LR	:	Öğrenme Oranı (Learning Rate)
LRS	:	Öğrenme Oranı Programı (Learning Rate Schedule)
MC3	:	Karma Evrişimsel 3D Ağ (Mixed Convolutional 3D Network)
MCE	:	Çok Sınıflı Çapraz Entropi (Multiclass Cross-Entropy)
MEMP	:	Çoklu Çıkarma ve Çoklu Tahmin (Multiple Extraction And Multiple Prediction)
MSE	:	Ortalama Kare Hata (Mean Squared Error)
ORB	:	Yönlendirilmiş FAST ve Döndürülmüş BRIEF (Oriented FAST And Rotated BRIEF)
RBF	:	Radyal Taban Fonksiyonu (Radial Basis Function)
ReLU	:	Düzeltici Doğrusal Birim (Rectified Linear Unit)
RGB	:	Kırmızı, Yeşil, Mavi (Red, Green, Blue)
RGB-HTG	:	RGB Hareket Tarihçesi Görüntüsü
RReLU	:	Rastgele Sızdıran Düzeltilmiş Doğrusal Birim (Randomized Leaky Rectified Linear Unit)
SGD	:	Stokastik Gradyan İnişi (Stochastic Gradient Descent)
SIFT	:	Ölçekle Değişmeyen Özellik Dönüşümü (Scale-Invariant Feature Transform)
SL-GCN	:	Yapılandırılmış Öğrenmeli Çizge Evrişim Ağı (Structured Learning Graph Convolutional Network)
ST-GCN	:	Uzamsal-Zamansal Çizge Evrişim Ağı (Spatial Temporal Graph Convolutional Network)
SURF	:	Hızlandırılmış Sağlam Özellikler (Speeded-Up Robust Features)
TAF	:	Zamansal Birikimli Özellikler (Temporal Accumulated Features)
TBA	:	Temel Bileşen Analizi (Principal Component Analysis)
TN	:	Doğru Negatifler (True Negatives)
TP	:	Doğru Pozitifler (True Positives)
TSSI	:	Ağaç Yapısı İskelet Görüntüsü (Tree Structure Skeleton Image)
UKSB	:	Uzun Kısa Süreli Bellek (Long Short-Term Memory)
YN	:	Yığın Normalleştirme (Batch Normalization)

1. GİRİŞ

İnsanlar, düşüncelerini, duygularını ve niyetlerini genellikle sözlü iletişim yoluyla ifade ederler. Ancak, işitme engeli olan veya işitme güçlüğü çeken birçok birey, konuşulanları anlamakta ve karşılık vermekte zorluk çekmektedir. Dünya İşitme Engelliler Federasyonu'na göre, dünya çapında 70 milyondan fazla işitme engelli insan bulunmakta ve bunların %80'den fazlası gelişmekte olan ülkelerde yaşamaktadır. (Anonymous, 2023). İşitme engelli insanların çoğu, birbirleriyle iletişim kurmak için sözlü kelimeler veya işitsel ifadeler yerine işaret dili olarak bilinen görsel işaretleri kullanmaktadır (Sreemathy ve diğ., 2023). Bu görsel işaretler, el şekli, duruşu, pozisyonu ve hareketi gibi manuel özellikler ile baş ve vücut duruşu, yüz ifadeleri ve dudak hareketleri gibi manuel olmayan özelliklerin karışımından oluşur (Mukushev ve diğ., 2020). Bölgeler ve kültürler arasında farklılık gösteren, karmaşık bir dil yapısına sahip olan işaret dili dünya çapında çok sayıda işitme engelli birey tarafından kullanılmaktadır (Zahid ve diğ., 2022). Her işaret dili, tıpkı sözel diller gibi, kendine özgü dilbilgisi yapılarına ve kurallarına sahiptir, bu yapılar ve kurallar dilin anlaşılmasını ve doğru kullanımını sağlar (Erten ve Arıcı, 2022).

İşaret dilini öğrenmek ve kullanmak işitme engelli bireylerin sosyal entegrasyonunda kritik bir rol oynamaktadır; ancak genel nüfusun işaret dili hakkındaki bilgi eksikliği ve işaret dili tercümanlarının azlığı iletişim engellerine sebep olmaktadır. Sonuç olarak, birçok işitme engelli birey sosyal ilişkiler kurmada zorlanmakta, eğitim, sağlık ve istihdam alanlarında çeşitli engellerle karşılaşmaktadır. Bu durum, sosyal izolasyon ve yalnızlık hissine yol açabilmektedir. Teknolojik gelişmeler, bu sorunu ele almayı amaçlayan İşaret Dili Tanıma (İDT - Sign Language Recognition, SLR) çalışmalarının geliştirilmesini teşvik etmektedir. Bilgisayarlı görme ve makine öğrenimi disiplinleri, işaret dilini otonom olarak metinsel veya işitsel formatlara dönüştürme konusunda yeni yaklaşımlar aramaktadır. Bu çerçevede geliştirilen İDT sistemlerinin temel amacı, işareti doğru bir şekilde tanımak için algoritmalar ve yöntemler geliştirmektir.

İDT sistemleri, bu teknolojik çabanın bir parçası olarak geliştirilmekte ve işitme engelliler ile işaret dili bilmeyenler arasında etkili bir köprü görevi görmeyi amaçlamaktadır. İşitme engelli insanların hareketleri konuşma veya yazı diline çevrildiğinde, işiten insanlar bunları anlayabilir, böylece işiten çoğunluk ile aradaki dilsel engel ortadan kalkarak işitme engelli bireylerin topluma entegrasyonu sağlanır (Rastgoo

ve diğ., 2021). Bununla beraber, İDT çalışmaları yalnızca iletişimi kolaylaştırmak için değil, aynı zamanda erişilebilir eğitim araçları, oyunlar ve işaret dilinin görsel sözlüklerini oluşturmak gibi girişimler yoluyla işitme engelli topluluğun kullanabileceği içeriği artırmak için de oldukça önemlidir (Das ve diğ., 2022).

İDT üzerine yapılan çalışmalar, kullanılan girdi verisinin türüne göre iki ana gruba ayrılabilir: sensör tabanlı sistemler ve görüş tabanlı sistemler (Ahmed ve diğ., 2018). Sensör tabanlı sistemler, genellikle bükülme (flexion), ivmeölçer, yakınlık sensörleri vb. ile donatılmış eldivenlerin kullanımını içeririr (Oz ve Leu, 2011; Gupta ve Kumar, 2021; Liu ve diğ., 2023). Bu yaklaşımın bir sınırlaması, kullanıcıların bileklerine veya ellerine birden fazla sensörle donatılmış elektronik eldiven takmaları gerekliliğidir. Bu durum, kullanıcıların sorunsuz ve ergonomik bir insan-bilgisayar etkileşimi kurmalarını engellemektedir. Ayrıca, sensör tabanlı yaklaşımlar genellikle işaret diline eşlik eden yüz ifadeleri, dudak hareketleri, göz hareketleri ve vücut dili gibi manuel olmayan jestsel bileşenleri dikkate almamaktadır (Ibrahim ve diğ., 2019). Diğer yandan, görüntü tabanlı sistemler, temel verilerin kameralar aracılığıyla elde edilmesini kolaylaştırır ve bu sayede duyusal eldivenlerdeki sensör ihtiyacını ortadan kaldırır. Bu sistemler, işaret dilinin oluşumuna katkıda bulunan manuel olmayan özelliklerin değerlendirilmesine de olanak tanır.

Başka bir açıdan, İDT sistemleri kullanılan girdi verilerinin biçimine göre iki ana kategoriye ayrılabilir: statik ve dinamik. Statik girdi tipik olarak bir işaretin tek ve hareketsiz bir görüntüsünden oluşurken, dinamik girdiler işaret dilinin akışkan ve değişken doğasını yakalamak için video kayıtlarını veya hareketli işaretlerin görüntü dizilerini kullanır. Bu da işaretin hem uzamsal hem de zamansal olarak değerlendirilmesini gerektirir. Dinamik girdiler, ayrıca “izole” (Sarhan ve Frintrop, 2023) ve "sürekli" (Aloysius ve Geetha, 2020) olmak üzere iki alt kategoriye ayrılır. “İzole” girdi, genellikle tek bir işareti veya kelimeyi ayrı ayrı sunar; her işaretin açıkça belirlenmiş başlangıç ve bitiş noktaları vardır. Öte yandan, “sürekli” girdi, işaretler arasındaki geçişler ve dilin doğal akışı ile işaretlerin doğal bir işaret dili akışı içinde sunulmasını kapsar (Rastgoo ve diğ., 2021).

Bu tez çalışması, video tabanlı izole işaret dili tanıma görevi özelinde, işitme engelli bireylerin toplumla etkileşimini kolaylaştırmayı amaçlayan İDT sistemlerinin geliştirilmesine yönelik üç ayrı çalışmayı kapsamaktadır. Bu çalışmalar, işaret dilinin karmaşık ve dinamik yapısını daha iyi anlamak ve doğru bir şekilde çevirebilmek için derin öğrenme teknikleri ve gelişmiş görüntü işleme yöntemlerini kullanmıştır. İşaret

dilini, yüz ifadeleri, vücut hareketleri, el ve parmak jestleri gibi manuel ve manuel olmayan unsurları içeren geniş bir spektrumda ele almak, bu tezin temelini oluşturmaktadır. Çalışmalarda önerilen her bir İDT sistemi, poz verileri kullanarak, Türk İşaret Dili başta olmak üzere, çeşitli işaret dili veri kümeleri üzerinde işaretçiden bağımsız değerlendirmelerde yüksek test doğrulukları elde etmiştir. Bu sistemler, işaret dilini anlama ve çeviri süreçlerini otomatize ederek, işitme engelli bireylerin sosyal entegrasyonunu destekleme potansiyeline sahiptir. Sonuç olarak, bu çalışmalar işaret dili kullanan bireylerle işiten toplum arasındaki iletişim engellerini aşmada önemli adımlar atmaktadır.

1.1. Tezin Amacı ve Önemi

Mevcut İDT yöntemlerinde kaydedilen önemli ilerlemelere rağmen, hâlâ çözülmesi gereken önemli zorluklar mevcuttur. Bu zorlukların başında, farklı işaretçiler arasındaki çeşitlilik gelir; bu çeşitlilik, tanıma süreçlerini önemli ölçüde karmaşıklaştırmakta ve sistemin işaretçiden bağımsız olarak yüksek doğruluk oranlarında performans göstermesini zorunlu kılmaktadır (Podder ve diğ., 2023). Etkin bir İDT sistemi, işaret dili unsurlarını -yüz ifadeleri (Kumar ve diğ., 2018; Irasiak ve diğ., 2023), vücut dili ve el hareketleri- kapsamlı olarak değerlendirmelidir, çünkü bu unsurların bütünsel analizi, eksiksiz bir İDT deneyimi için elzemdir (Javaid ve Rizvi, 2023; Özdemir ve diğ., 2023). İşaret dilinin hem uzamsal hem de zamansal boyutları göz önünde bulundurulduğunda, bu bilgileri entegre edebilen etkili modellerin geliştirilmesi zorunlu hale gelir. Ayrıca, değişken arka planlar, çeşitli aydınlatma koşulları ve görsel karmaşalar gibi faktörler, algoritmaların işaret sözcüklerini doğru bir şekilde tanıyıp ayırt etmesi için ek zorluklar oluşturur (Hamada ve diğ., 2004; Tian ve diğ., 2023). Gerçek zamanlı uygulamaların geliştirilmesi ise, yüksek hesaplama gücü talepleri nedeniyle, algoritmaların hızlı ve doğru çalışmasını sağlamak için önemli çabalar gerektirir (Hori ve Yamamoto, 2024). Bu tür zorluklar, işaret dili tanıma alanında derinlemesine araştırmaları ve yenilikçi gelişmeleri zorunlu kılmıştır. Bu sayede, işaret dili kullanan bireyler ile genel toplum arasındaki iletişim engelleri aşılabilir ve karşılıklı anlayış ile etkileşim olanakları artırılabilir.

Bu tez çalışmasının temel motivasyonu, İDT teknolojisindeki yukarıda bahsedilen zorlukları aşmak ve geliştirilen sistemler aracılığıyla işaret dili hareketlerini metin veya konuşmaya dönüştürerek işitme engelli bireylerin toplumdaki iletişim engellerini

kaldırmaktır. Bu süreç, İDT'nin doğruluğunu artırarak işaret dili kullanan bireyler ile işiten bireyler arasında etkili iletişim köprüleri kurmayı hedefler; böylece her iki grubun birbirini daha iyi anlamasına ve etkileşimde bulunmasına olanak tanınmış olur. Bu tez çalışması, aynı zamanda gelecekteki araştırmalara da yol gösterici olacak şekilde, işaret dili tanıma sistemlerinin uygulanabilirliğini ve etkinliğini artırmaya yönelik yöntemler ve çözümler sunmayı amaçlamaktadır.

1.2. Tezin Literatüre Katkısı

Bu tez çalışması, video tabanlı izole İDT alanına yönelik yöntemler geliştirerek literatüre katkıda bulunmuştur. Üç adet Science Citation Index Expanded makalesi kapsamında sunulan bu yeniliklerin, işaret dili tanıma alanında önemli katkılar sağlaması beklenmektedir.

"Enhancing Signer-Independent Recognition of Isolated Sign Language through Advanced Deep Learning Techniques and Feature Fusion" (Akdag ve Baykan, 2024a) isimli çalışmada, işaret dilini hem uzamsal hem de zamansal boyutlarda analiz edebilen yenilikçi bir R3(2+1)D-SLR ağı geliştirilmiştir. Bu çalışmanın literatüre katkıları şöyle sıralanabilir:

- R(2+1)D ve R3D evrişim bloklarını birleştiren ve işaret dilinin hem uzamsal hem de zamansal özelliklerinin etkin bir şekilde öğrenilmesini sağlayan yeni bir R3(2+1)D-SLR ağı sunulmuştur.
- MediaPipe Holistic kullanılarak elde edilen yüksek hassasiyetli poz verileri ile, ayrıntılı uzamsal ve zamansal özellikler çıkarılmıştır.
- Yüksek tanıma doğruluğu için, eğitilmiş R3(2+1)D-SLR modellerinden elde edilen vücut, el ve yüze ait karmaşık özelliklerin birleştirilip DVM ile sınıflandırıldığı bir yöntem geliştirilmiştir.
- Test veri kümesine gerçek dünya arka plan görüntüleri dahil edilerek, İDT'deki kritik bir zorluğu ele alan sistemin, farklı ortamlara uyarlanabilirliği ve güvenilirliği analiz edilmiş ve değişken arka planlara karşı poz görüntüleri kullanılarak, pratik ve gerçek dünya çözümleri için sistem sağlamlığı artırılmıştır.

İkinci çalışmada, görsel poz verileri ve bu verilerden elde edilen Hareket Tarihçesi Görüntüleri (HTG) kullanılarak eğitilen ResNet-18 modellerine dayalı bir İDT sistemi

önerilmiştir. Bu çalışma, işaret dili hareketlerinin daha detaylı ve doğru bir şekilde anlaşılmasına olanak tanıyan yenilikçi yöntemler sunmuştur. “Isolated Sign Language Recognition through Integrating Pose Data and Motion History Images” (Akdağ ve Baykan, 2024) isimli makalede sunulan çalışmanın literatüre katkıları şöyle sıralanabilir:

- MediaPipe Holistic kullanılarak elde edilen detaylı poz verileri ve eksik poz verilerinin doğrusal enterpolasyon yoluyla tamamlanmasıyla işaret dili tanımada iyileşme sağlamıştır.
- HTG’ler poz verilerine uygulanarak, işaret dili hareketlerinin hem uzamsal hem de zamansal yönlerini kapsayan bir temsil sağlanmıştır.
- Parmak pozu tabanlı hareket tarihçesi görüntü (PARMAK-POZU-HTG) özelliği, parmak hareketlerinin ayrıntılı analizini mümkün kılarak işaret dilinin daha incelikli yönlerinin anlaşılmasına katkıda bulunmuştur.
- Poz verilerinin ve bu poz verilerinden türetilen HTG'lerin RReLU aktivasyon fonksiyonu ile geliştirilmiş ResNet-18 modellerinden elde edilen özelliklerini birleştiren ve DVM ile sınıflandıran bir İDT sistemi geliştirilmiş ve yüksek tanıma doğruluğu elde edilmiştir.

Üçüncü çalışma, İDT’deki parmak hareketlerinin ve pozisyonlarının daha detaylı analizine odaklanmıştır. Çalışma, parmakların sağladığı nüanslı yorumlamanın, doğru iletişim için ne kadar önemli olduğunu vurgulamaktadır. “Multi-Stream Isolated Sign Language Recognition Based on Finger Features Derived from Pose Data” (Akdag ve Baykan, 2024b) isimli makalede sunulan çalışmanın literatüre katkıları şu şekilde sıralanabilir:

- Çalışmada, parmak pozlarının ayrı kanallarda görselleştirilmesiyle elde edilen PARMAK-POZU, parmak bölgelerinin kırpılması ve birleştirilmesiyle daha da detaylandırılan BİRLEŞTİRİLMİŞ-PARMAK-POZU ve PARMAK-POZU verilerine uygulanan Çerçeve Farkı Yöntemiyle zamansal değişiklikleri yakalayan ÇF-PARMAK-POZU özellikleri geliştirilmiştir. Bu özellikler, İDT sürecindeki parmak hareketlerinin ve konfigürasyonlarının ayrıntılı analizini sağlayarak, işaret dilinin ince nüanslarının daha iyi anlaşılmasını sağlamıştır.
- MobileNetV2 mimarisine dayanan ve çok kanallı parmak tabanlı özellikleri verimli bir şekilde işleyebilen Çok Kanallı-MobileNetV2 modeli önerilmiştir.

- Önerilen sistem, poz verilerinden elde edilen vücut ve yüz görüntülerini ve bu görüntülere Çerçeve Farkı Yöntemi uygulanarak oluşturulan görselleri de kullanmıştır. Bu görsel veriler MobileNetV2 modelinin eğitiminde kullanılmış ve manuel olmayan ifadelerin önemi vurgulanmıştır.
- Genişletilmiş yaklaşım, eğitilmiş Çok Kanallı-MobileNetV2 ve MobileNetV2 modellerinden çıkarılan parmak, vücut ve yüzle ilgili özelliklerin boyutlarını Temel Bileşen Analizi (TBA) ile azaltmıştır. Daha sonra, bu özellikleri birleştirerek daha kapsamlı bir İDT sistemi oluşturmak için DVM ile sınıflandırmıştır. Bu entegre yaklaşım, İDT sürecinde doğruluğu ve kapsamlılığı en üst düzeye çıkararak bu alanda önemli bir yenilik getirmiştir.

1.3. Tezin Organizasyonu

Bölüm 1'de, İDT çalışmalarının önemi ve bu alanın işitme engelli bireylerin sosyal entegrasyonuna katkısı üzerine genel bir bakış sunulmuştur. Ayrıca, tezin amacı, önemi ve literatüre katkıları detaylandırılmıştır. Bölüm 2'de, İDT teknolojilerinin evrimi ve metodolojik ilerlemeleri incelenmiştir. İDT'nin geleneksel makine öğrenimi tekniklerinden derin öğrenme yaklaşımlarına geçişi ele alınarak, bu alandaki ilgili çalışmalar üzerinden önemli gelişmeler özetlenmiştir. Bölüm 3'te, kullanılan veri kümeleri, MediaPipe Holistic ile poz verilerinin elde edilmesi, Doğrusal Enterpolasyon, Hareket Tarihçesi Görüntüsü ve Çerçeve Farklı Yöntemi gibi veri ön işleme yöntemleri hakkında bilgi verilmiş ve Evrişimsel Sinir Ağları'nın temel yapısı, Destek Vektör Makineleri gibi diğer metodolojik süreçler anlatılmıştır. Bölüm 4'te, önerilen İDT yöntemleri, veri ön işleme adımından başlanarak metodolojik tasarıma kadar ayrıntılı bir şekilde açıklanmış, her bir yöntemin geliştirilme sürecindeki deneysel tasarım aşamaları detaylandırılmıştır. Ayrıca, bu yöntemler literatürdeki benzer çalışmalarla karşılaştırmalı olarak analiz edilerek, önerilen yaklaşımların mevcut çalışmalara göre avantajları ve yenilikleri vurgulanmıştır. Bölüm 5'te ise, izole İDT görevi için gerçekleştirilen çalışmalardan elde edilen sonuçlar özetlenerek gelecek çalışmalara yönelik öneriler sunulmuştur.

2. KAYNAK ARAŞTIRMASI

İDT teknolojilerinin gelişimi, başlangıçta el yapımı özelliklere ve klasik makine öğrenimi tekniklerine dayanan ilk sürümlerden, özellik çıkarma ve sınıflandırmada önemli gelişmeler sağlayan derin öğrenme modellerine kadar uzanmaktadır. Bu bölüm, Türk İşaret Dili'ne ait BosphorusSign22k veri kümesi (Camgoz ve diğ., 2016; Özdemir ve diğ., 2020) ile birlikte, Arjantin İşaret Dili'nden LSA64 (Ronchetti ve diğ., 2016b) ve Yunan İşaret Dili olan GSL (Greek Sign Language) (Adaloglou ve diğ., 2020) gibi veri kümelerini kullanan çalışmalara özellikle odaklanarak, izole İDT'deki son gelişmelere ve kullanılan metodolojilere genel bir bakış sağlamayı amaçlamaktadır. Bu dillerin tercih edilmesinin sebebi, daha az temsil edilen dil yapılarına ışık tutarak, önerilen İDT yöntemlerinin çeşitli dil yapılarına ne kadar iyi uyum sağlayabildiğini göstermektir.

Görme tabanlı İDT sistemlerinin ilk versiyonları temel olarak el yapımı özelliklere ve makine öğrenimi yöntemlerine dayanıyordu. Bu özelliklere örnek olarak, Ölçekle-Değişmeyen Özellik Dönüşümü (Scale-Invariant Feature Transform - SIFT) (Yasir ve diğ., 2016), Hızlandırılmış Sağlam Özellikler (Speeded-Up Robust Features - SURF) (Yang ve Peng, 2014; Kour ve Mathew, 2017) Yönlendirilmiş Gradyanların Histogramı (Histogram of Oriented Gradients - HOG) (Raj ve Jasuja, 2018) ve Temel Bileşen Analizi (TBA) (Principal Component Analysis - PCA) (Gweth ve diğ., 2012; Mohana ve Reddy, 2014) verilebilir. Bunlar gibi, işaret dili harflerinden, sayılarından veya kelimelerinden çıkarılan özellikler, sınıflandırma için çeşitli makine öğrenimi algoritmaları tarafından ele alınmaktadır. Bu çalışmalar ağırlıklı olarak harfler ve sayılar gibi statik işaretlerin tanınmasına odaklanmıştır. Ancak izole İDT görevi için, bu temel yaklaşımlar üzerine kurulan daha gelişmiş teknikler kullanan çalışmalar da geliştirilmiştir.

Madani ve Nahvi (2013) izleme için CAMSHIFT (Continuously Adaptive Mean Shift) algoritmasını ve özellik çıkarımı için Radon dönüşümünü kullanmış, elde edilen özellikleri minimum mesafe sınıflandırma yöntemiyle sınıflandırmıştır. Bu yöntem, 20 izole Pakistan İşaret Dili kelimesini %95,56 doğrulukla doğru bir şekilde tanımlayabilmiştir. Li ve diğ. (2016), 11 Tayvan İşaret Dili kelimesini tanımak için sınıflandırıcı olarak Gizli Markov Model (Hidden Markov Model - HMM) kullanmış, temel bilgileri kaybetmeden gereksiz boyutları ortadan kaldırmak için entropi tabanlı bir K-Ortalamalar Algoritması ile TBA'yı entegre etmiş, ayrıca optimizasyon için Yapay Arı Kolonisi Algoritması (YAKA) ve Baum-Welch algoritmalarını dahil ederek %91,30'luk

bir tanıma doğruluğu elde etmiştir. Ahmed ve diğ. (2017) Dinamik Zaman Bükme (Dynamic Time Warping -DTW) kullanan görüntü tabanlı bir el hareketi tanıma sistemi tanıtmıştır. El hareketlerini izleyerek, aynı zamanda hem ellerin hem de yüzün özelliklerini analiz ederek, önerdikleri sistem Hint İşaret Dili'nden 24 işareti tanımada %90 doğruluk oranına ulaşmıştır. Ronchetti ve diğ. (2016a) konum, hareket ve el şekli gibi çeşitli özelliklere dayanan olasılıksal bir model önermiştir. Bu model, LSA64 veri kümesinde (64 sınıflı) %97'lik bir doğruluk elde etmiş, ayrıca işaretçiden bağımsız testlerde ortalama %91,7'lik bir doğruluk göstermiştir. Rodríguez ve Martínez (2018) kümülatif şekil farkı ve Destek Vektör Makinesi (DVM) yöntemine dayalı bir model tasarlamış ve LSA64 veri kümesinde %85 doğruluk elde etmiştir. Özdemir ve diğ. (2020), işaret dilini tanıma için, Geliştirilmiş Yoğun Yörüngeler (Improved Dense Trajectory - IDT) metodunu kullanmışlardır. Bu yöntem, video kareleri arasındaki hareketleri izleyerek HOG (Histogram of Oriented Gradients), HOF (Histogram of Optical Flow), ve MBH (Motion Boundary Histogram) gibi özellikleri çıkarır. Bu özellikler, işaret dilinin karmaşık el hareketleri ve vücut dili gibi unsurlarını analiz etmekte kullanılmıştır. Çalışma, Türk İşaret Dili'ndeki 744 farklı işareti içeren BosphorusSign22k veri setinde %88,53 doğruluk oranı ile başarılı sonuçlar elde etmiştir. Geleneksel makine öğrenimi tekniklerine dayanan bu yöntemler İDT konusunda önemli adımlar atmış olsa da elde edilen sonuçlar, sınıflandırılan kelime sayısına göre nispeten düşük kalmaktadır. Bunun nedeni, bu yöntemlerin işaret dilinin karmaşık ve dinamik doğasını tam olarak modelleyememesidir. Ayrıca, geleneksel makine öğrenimi algoritmaları genellikle karmaşık özellik mühendisliği gerektirir ve bu da alana özgü uzmanlık gerektirir. Bu süreç, makine öğrenimi algoritmasına devam etmek için en uygun özellikler seçilmeden önce ayrıntılı veri analizi yoluyla güçlü el yapımı özelliklerin çıkarılmasını gerektirir. Ayrıca, veri kümesinin çeşitliliği arttıkça, yeni ve daha sofistike yöntemlerin geliştirilmesi gerekli hale gelebilir. Zaman içinde, teknolojik gelişmeler ve makine öğreniminin evrimi ile derin öğrenme teknikleri özellik çıkarma sürecini otomatikleştirerek manuel özellik mühendisliği ihtiyacını önemli ölçüde azaltmıştır.

Derin öğrenme yöntemlerinin İDT alanına getirdiği yenilikler, sistemlerin daha karmaşık ve çeşitli işaretleri anlama becerisini önemli ölçüde geliştirmiştir. Derin öğrenmeye dayalı mevcut teknikler genellikle Evrişimsel Sinir Ağlarına (ESA) dayanmaktadır. 2 boyutlu (2D) ESA'lar, genellikle zamansal bilgi içermeyen ve statik görüntülerden oluşan rakam ve harf işaretleri için sıklıkla kullanılmıştır. Ezel (2018) tarafından Türk İşaret Dili alfabesini sınıflandırmak amacıyla gerçekleştirilen çalışmada,

renkli (RGB) görüntülerden oluşan ve sadece elleri içeren bir veri seti kullanılmış. Veri artırma teknikleri uygulanarak genişletilen bu veri seti, AlexNet modeli kullanılarak sınıflandırılmış ve %99,61 gibi yüksek bir doğruluk oranı elde edilmiştir. Damaneh ve diğ. (2023) statik el hareketlerini tanımak için ESA, Gabor Filtresi ve ORB (Oriented FAST and Rotated BRIEF) özellik tanımlayıcısını kullanan hibrit bir yöntem önermiş; Massey (Barczak ve diğ., 2011), Amerikan İşaret Dili alfabesi (Sharma ve diğ., 2020) ve Amerikan İşaret Dili (Rahim ve diğ., 2020) veri kümeleri üzerinde sırasıyla %99,92, %99,8 ve %99,80 doğruluk oranlarına ulaşmıştır. Das ve diğ. (2023) Bangladeş İşaret Dilindeki rakamları ve karakterleri, ESA ve Rastgele Orman sınıflandırıcılarını birleştiren hibrit bir yöntem kullanarak sırasıyla %97,33 ve %91,67 doğrulukla tahmin etmiştir. Aldhahri ve diğ. (2023) Arap İşaret Dilindeki harfleri %94,46 doğrulukla tanımlayan bir ESA mimarisi geliştirmiştir. Ma ve diğ. (2022), J ve Z harfleri gibi dinamik hareketler içeren veri kümelerinde ardışık iki zaman çerçevesi arasındaki özellik ifadesinin korelasyonunu iyileştirmek için İki Akışlı Karma Evrişimsel Sinir Ağını kullanmıştır. Önerdikleri İki Akışlı Karma ResNet50 modeli Amerikan İşaret Dili Alfabesi veri kümesinde %97,57 doğruluk elde etmiştir. Alsharif ve diğ. (2023) AlexNet (Krizhevsky ve diğ., 2012), ConvNeXt (Liu ve diğ., 2022), EfficientNet (Tan ve Le, 2019), ResNet-50 (He ve diğ., 2016) ve Vision Transformer (Dosovitskiy ve diğ., 2020) yöntemlerini kullanarak Amerikan İşaret Dili alfabesini sınıflandırmada sırasıyla %99,50, %99,51, %99,95, %99,98 ve %88,59 doğruluk elde etmiştir.

2D-ESA'lar, harf ve rakam gibi statik işaret görüntülerini tanımada önemli başarılar elde etmiş olsa da zamansal özellikleri yakalamada bazı sınırlamaları vardır. Halbuki kelimelerin ve cümlelerin temsili, işaretlerin doğası gereği hem uzamsal hem de zamansal özelliklere sahiptir. Bu durum araştırmacıları, kelimelerin ve cümlelerin anlamını çıkarmak için yalnızca hareketlerin ve jestlerin uzamsal konumunu değil, aynı zamanda bu hareketlerin zaman içindeki sırasını ve dinamiklerini de tespit etmek için sofistike yöntemler geliştirmeye yöneltmiştir. Bu temel üzerine inşa edilen derin öğrenme tabanlı yaklaşımların geliştirilmesi ve uygulanması hem uzamsal hem de zamansal boyutları entegre eden modellere doğru önemli bir kaymaya işaret ederek İDT alanında derin ilerlemelere yol açmıştır. Örneğin, Masood ve diğ. (2018), ESA kullanarak elde ettikleri uzamsal özellikleri, Tekrarlayan Sinir Ağına (TSA) girdi olarak vererek, zamansal özellikleri elde ettikleri bir model tasarlamıştır. Geliştirdikleri yöntem, LSA64 veri kümesinin bir alt kümesinde (46 sınıf) %95,2 test doğruluğu elde etmiştir. Zhang ve Li (2019), dinamik videoların zamansal ve uzamsal özelliklerini çıkarmak ve

değerlendirmek için bir 3 boyutlu (3D) ESA ve Evrişimsel Uzun Kısa Süreli Belleği (Evrişimsel-UKSB) birleştiren Çoklu Çıkarma ve Çoklu Tahmin (Multiple extraction and Multiple prediction - MEMP) ağı adı verilen bir yapay sinir ağı tasarlamıştır. Geliştirdikleri yapı LSA64 veri kümesi üzerinde %99,063'lük bir test başarısı göstermiştir. Imran ve Raman (2020), Gareket Tarihçesi Görüntülerini (HTG), Dinamik Görüntüler (DG) ve RGB Hareket Görüntü (RGB-HG) şablonları gibi bir videoyu tek bir görüntüde temsil edebilen çeşitli veri türlerini kullanarak önceden eğitilmiş üç ESA modeline ince ayar yapmıştır. Bu üç modelin sonuçları, çekirdek tabanlı bir aşırı öğrenme makinesi kullanılarak yenilikçi bir füzyon metodolojisi ile birleştirilmiş ve LSA64 üzerinde %97,81'lik bir doğruluk elde edilmiştir. Özdemir ve diğ. (2020) tarafından yapılan çalışmada hem 2D hem de 3D evrişim katmanlarını birleştiren bir yapıya sahip olan Karma Evrişimsel 3D Kalıntı Ağı (Mixed Convolutional Neural Network - MC3) yöntemi, BosphorusSign22k veri kümesi üzerinde %78,85 doğruluk elde etmiştir. Adaloglou ve diğ. (2022) 310 farklı sınıftan oluşan GSL veri kümesindeki işaretleri sınıflandırmak için GoogLeNet+TConvs (Cui ve diğ., 2019), 3D-ResNet (Pu ve diğ., 2019) ve Genişletilmiş (Inflated) 3D Evrişimsel Ağ (I3D) (Szegedy ve diğ., 2016) modellerini kullanmıştır. Bu modellerle, işaretçiden bağımsız testlerde sırasıyla %86,03, %86,23 ve %89,74 doğruluk oranlarına ulaşılmıştır. Wang ve diğ. (2021a) daha yüksek doğruluk elde edebilen (2+1)D evrişim tabanlı bir model olan (2+1)D-SLR ağını önermiştir. Bu ağ, LSA-64 veri kümesi üzerinde %98,7 doğruluk elde etmiştir. Sincan ve Keles (2021), I3D modelini RGB videoları kullanarak eğitmiş ve ResNet-50 modelini RGB hareket tarihçesi görüntü verileriyle eğitmiştir. BosphorusSign22k veri kümesi üzerinde modellerden elde edilen özellikleri geç füzyon tekniği ile birleştirdikleri yöntemi kullanarak %94,83 doğruluk elde etmişlerdir.

Bu çalışmalar önemli başarılar elde etmiş olsa da işaret dili videolarının karmaşıklığını ve çeşitliliğini tam olarak yakalayamamıştır. Özellikle, işaret dilinde anlamın büyük bir kısmını aktaran ellere ve yüz ifadelerine yeterince odaklanılmamıştır. Yüz ifadeleri ve jestlerin yanı sıra ellerin detaylı hareketleri ve pozisyonları da işaret dilinin anlamını güçlendiren kritik unsurlardır. Mevcut çalışmaların bu önemli bileşenleri göz ardı etmesi, İDT sistemlerini daha da geliştirmek için yeni yollar aranmasına neden olmuştur. Bu kapsamda Gökçe ve diğ. (2020) el, vücut ve yüz görüntülerini kullanarak 18 katmanlı Karma Evrişimsel 3D Kalıntı Ağlarını (MC3-18) eğitmiştir. Bu ağlardan elde edilen özellikleri skor seviyesinde birleştirerek BosphorusSign22k test verisi üzerinde %94,94 doğruluk elde etmişlerdir. Gündüz ve Polat (2021) Inception3D (3 Boyutlu

Başlangıç Ağı) model eğitiminde yüz, eller ve vücut bölgelerine ait görüntüler ile optik akış verilerini, UKSB ağ eğitiminde ise vücut ve el pozu verilerini kullanmıştır. Eğitilen modellerden elde edilen özellik akışları birleştirilerek iki katmanlı bir sinir ağının girişi olarak kullanılmış ve bu yöntemle, 174 sınıftan oluşan BosphorusSign22k-genel veri kümesi üzerinde %89,3'lük bir test doğruluğu elde edilmiştir. Jebali ve diğ. (2024) tarafından önerilen yöntemde ESA ve UKSB gibi derin öğrenme modelleri kullanılarak, hem el hareketlerinden hem de yüz ifadeleri gibi manuel olmayan bileşenlerden gelen bilgileri eş zamanlı olarak işleyebilen bir sistem geliştirilmiş, manuel olmayan özelliklerin kullanılmasıyla sistem performansında önemli bir iyileşme gözlemlenmiştir. Önerdikleri sistem, 450 ve 26 sınıflı veri kümeleri üzerinde sırasıyla %90,12 ve %94,87'lik test doğrulukları elde etmiştir.

İDT sistemlerini tasarlarken, tanıma doğruluğunun beraberinde dikkate alınması gereken başka bir nokta da arka plan ve işaretçi varyasyonlarına karşı dayanıklılıktır. Ayrıca, işaret dilinin ince nüanslarını daha iyi anlamak için daha ayrıntılı bir analiz yapmak önemlidir. Bu ihtiyaçlar, poz verilerine dayalı sistemlerin önemini vurgulamaktadır. Poz verileri, İDT sistemleri geliştirilirken bu sınırlamaların üstesinden gelme potansiyeline sahiptir. Bu veriler, arka plan varyasyonlarının ve işaretçi farklılığının sistemlerin genel performansı üzerindeki etkisini azaltabilir. Ellerin, yüzün ve vücudun konumlarını ve hareketlerini doğrudan yakalayan bu veriler, sistemlerin farklı koşullar altında bile tutarlı sonuçlar vermesini sağlayabilir. Bu da İDT sistemlerinin daha sağlam ve güvenilir olmasını sağlar. Bu bağlamda, (Konstantinidis ve diğ., 2018b) UKSB ağlarını eğitmek için RGB videolarından çıkarılan el ve vücut poz bilgilerini kullanan ve bu ağlardan elde edilen özellikleri geç füzyon tekniği ile birleştiren bir model önermiştir. Geliştirdikleri model LSA64 üzerinde %98,09 doğruluk oranına ulaşmıştır. Aynı yazarlar başka bir çalışmalarında (Konstantinidis ve diğ., 2018a) RGB video ve optik akış bilgilerini analiz eden ek akışlar ekleyerek yöntemlerinin performansını %99,84'e yükseltmişlerdir. Kindiroglu ve diğ. (2019) BosphorusSign22k-genel veri kümesinde, Zamansal Birikimli Özellikler (Temporal Accumulated Features - TAF) yöntemini kullanarak yaptıkları çalışmalarında %81,58'lik bir doğruluk oranı elde etmişlerdir. Bu yöntem, işaret dili videolarından poz bilgisinin biriktirilmesini ve ton tabanlı renklendirme ile temsil edilmesini içerir. Aynı yazarlar sonraki bir çalışmalarında (Kındıroglu ve diğ., 2023) değişken uzunluktaki videoları sabit uzunluktaki özelliklerle temsil etmek için poz tabanlı bir görsel temsil olan Hizalanmış Zamansal Birikimli Özellikler (Aligned Temporal Accumulated Features - ATAF) adlı yeni bir yöntem

önermiş ve bu yöntemi MC3-18 (Tran ve diğ., 2018) modelleriyle birleştirerek BosphorusSign22k veri kümesi üzerinde %94,9 doğruluk elde etmiştir. Selvaraj ve diğ. (2021), GSL veri kümesi üzerindeki deneylerde sırasıyla %86,6, %89,5, %93,5 ve %95,4 doğruluk oranlarına ulaşan UKSB, Dönüştürücü (Transformer), Uzamsal-Zamansal Çizge Evrişim Ağı (Spatial Temporal Graph Convolutional Network - ST-GCN) ve Yapılandırılmış Öğrenmeli Çizge Evrişim Ağı (Structured Learning Graph Convolutional Network - SL-GCN) modellerini kullanarak poz verilerine dayalı bir yaklaşım benimsemiştir. Samaan ve diğ. (2022) MediaPipe'tan elde edilen poz işaret noktalarını kullanarak on sınıflı bir veri kümesi üzerinde UKSB, iki yönlü UKSB ve Kapılı Tekrarlayan Birim (Gated Recurrent Unit – GRU) modellerinde sırasıyla %99,6, %99,3 ve %100 doğruluk elde etmiştir. Önerilen yöntemin test doğruluğu yüksek olmasına rağmen, sınıf sayısı genel işaret dili sözlüklerinde kullanılan kelime sayısına kıyasla oldukça düşüktür. Alyami ve diğ. (2023), Dönüştürücü tabanlı bir model kullanarak el ve yüz poz verilerini sınıflandırmıştır. LSA64 veri kümesinde, işaretçiye bağlı ve işaretçiden bağımsız testlerde sırasıyla %98,25 ve %91,01 doğruluk elde etmişlerdir. Özdemir ve diğ. (2023), ST-GCN ve Çoklu İşaret Uzun Kısa Süreli Bellek (Multi-Cue Long Short-Term Memory - MC-LSTM) modellerine dayalı bir yöntem tasarlamıştır. Bu yöntem, işaret dilini daha doğru bir şekilde tanımak için vücut pozu verileri ile yüz ve eller gibi çeşitli jestsel bilgileri de kullanır. Bu yöntem, BosphorusSign22k veri kümesi üzerinde yapılan testlerde %92,58'lik bir doğruluk elde etmiştir. de Castro ve diğ. (2023) özetlenmiş RGB karelerini, ellerin ve yüzün bölümlere ayrılmış bölgelerini, eklem mesafelerini ve yapay olarak oluşturulmuş derinlik verilerini bir 3D-ESA aracılığıyla işlemeyi içeren çok akışlı bir yaklaşım sunmuştur. Bu yöntemde, yapay derinlik haritalarının eklenmesinin farklı işaretçiler için genelleme kapasitesini artırdığı gösterilmiştir. Yöntemleri 20 sınıflı bir veri kümesi üzerinde %91 tanıma doğruluğu elde etmiştir. Hamza ve Wali (2023), her sınıfın çok az örnek içerdiği 80 sınıftan oluşan bir veri kümesi üzerinde dönüştürme ve döndürme gibi veri artırma tekniklerini kullanarak, 3D-ESA modeli ile %93,33 tanıma doğruluğu elde etmiştir. Bu çalışma, ayrıca sınırlı veri kümeleri üzerinde veri artırma yöntemlerinin etkinliğini göstermiştir. Laines ve diğ. (2023), poz verilerini RGB görüntüsüne dönüştüren ve İDT için iskelet tabanlı modellerin doğruluğunu artıran bir Ağaç Yapısı İskelet Görüntüsü (Tree Structure Skeleton Image - TSSI) temsili kullanarak izole İDT için yenilikçi bir yaklaşım sunmuştur. TSSI gösteriminde, sütunlar poz işaret noktaları, satırlar bu noktaların zaman içindeki değişimini, RGB kanalları ise noktaların (x, y, z) koordinatlarını temsil eder. Önerilen bu veriler DenseNet-121 modeli ile

sınıflandırılmış, 100, 226 ve 30 sınıflı veri kümeleri için sırasıyla %81,47, %93,13 ve %98 tanıma doğrulukları elde edilmiştir. Özellikle, yüz ifadeleri ve ayrıntılı el hareketleri gibi işaret dilinin kritik bileşenlerinin dikkate alınması, modelin doğruluğunu önemli ölçüde artırmıştır.

3. MATERYAL VE YÖNTEM

Bu bölümde, tez çalışmasında kullanılan temel materyaller ve metodoloji ayrıntılı bir şekilde anlatılmıştır. Öncelikle araştırmanın temelini oluşturan izole işaret dili veri kümeleri, poz tahmini için MediaPipe Holistic'in kullanımı, eksik veri noktalarını tamamlamak amacıyla uygulanan doğrusal enterpolasyon tekniği, görsel poz verilerinin hazırlanması, hareketin zamansal evrimini yansıtan Hareket Tarihçesi Görüntüleri ve Çerçeve Farkı Yöntemi ele alınmıştır. Ayrıca, Evrişimsel Sinir Ağlarının (ESA) temel yapısı ve bileşenleri, özellik füzyonu için kullanılan basit birleştirme tekniği ve özellik sınıflandırması için tercih edilen Destek Vektör Makineleri (DVM) ve diğer süreçler detaylandırılmıştır.

3.1. Kullanılan Veri Kümeleri

Bu tez çalışması kapsamında, önerilen yöntemlerin etkinliğini kapsamlı bir şekilde değerlendirmek için birden fazla işaret dili veri kümesi kullanılmıştır. Her bir veri kümesi, yöntem geliştirmeden literatürdeki çalışmalarla karşılaştırmalı analize kadar, deneysel tasarımda benzersiz bir amaca hizmet etmektedir. Bu veri kümeleri kullanılarak, önerilen yöntemlerin sağlamlığı, çok yönlülüğü, ölçeklenebilirliği ve İDT görevinde geniş uygulanabilirlik potansiyeli vurgulanarak kapsamlı bir doğrulama sağlanması amaçlanmıştır. Çalışmalarda kullanılan veri kümelerinin bir özeti Çizelge 3.1'de verilmiştir.

Çizelge 3.1. Tez çalışmasında kullanılan veri kümeleri

Veri Kümesi	Referans	İşaretçi Sayısı	Örnek Sayısı	Sınıf Sayısı
BosphorusSign22k-genel	(Camgoz ve diğ., 2016; Özdemir ve diğ., 2020)	6	5.788	174
BosphorusSign22k	(Camgoz ve diğ., 2016; Özdemir ve diğ., 2020)	6	22.542	744
LSA64	(Ronchetti ve diğ., 2016b)	10	3.200	64
GSL	(Adaloglou ve diğ., 2020)	7	40.826	310

3.1.1. BosphorusSign22k veri kümesi

BosphorusSign22k veri kümesi (Camgoz ve diğ., 2016; Özdemir ve diğ., 2020), İDT alanındaki bilimsel çalışmalara destek sağlamak üzere oluşturulmuş olup Türk İşaret Dili'nde sağlık, finans ve günlük yaşamda sık kullanılan kelimelerden oluşan video kayıtlarını içerir.

Veri kümesindeki tüm işaret dili videoları, Microsoft Kinect v2 kamerası kullanılarak 1920x1080 çözünürlüğünde kaydedilmiştir. Bu kayıtlar, işaretçilerin kamera karşısında belirli bir mesafede durduğu ve arkalarında yeşil perde teknolojisi kullanıldığı standart bir düzenle gerçekleştirilmiştir. Bu düzenleme, işaret dilindeki el hareketlerini ve mimikleri net bir şekilde kaydetmeyi amaçlamıştır. Veri kümesine ait örnek görüntü kareleri Şekil 3.1'de gösterilmiştir.

Şekil 3.1. BosphorusSign22k veri kümesinden örnek görüntü kareleri (Özdemir ve diğ., 2020)

BosphorusSign22k, sağlık sektöründe 428, finans sektöründe 163 ve genel (günlük) kullanımda 174 olmak üzere toplam 744 işaret dili kelimesini barındırmaktadır. Her bir video, renkli görüntü, derinlik haritası ve işaretçinin iskelet bilgileri gibi çeşitli veri türlerini içerir. Toplamda altı işaretçi ile oluşturulan bu veri kümesinde işaretçilerden biri, işaretçiden bağımsız değerlendirme için test veri kümesi olarak ayrılmıştır.

Deneysel çalışmalar, BosphorusSign22k veri kümesinin bir alt kümesi olan BosphorusSign22k-genel üzerinde gerçekleştirilmiştir. Bu alt küme, toplam 174 işaret sözcüğü sınıfı için 4.839 eğitim videosu ve 949 test videosu içermektedir. Bu alt küme önerilen yöntemlerin ilk testleri ve iyileştirmeleri için ideal bir platform sunmaktadır.

BosphorusSign22k veri kümesinin tamamı, 22.542 video ve 744 işaret sözcüğü sınıfından oluşmaktadır. Bu veri kümesi, yöntemlerin ölçeklenebilirliğini ve daha büyük bir işaret dili veri kümesi üzerindeki performansını değerlendirmek için kullanılmıştır.

3.1.2. LSA64 veri kümesi

Önerilen İDT sistemlerinin farklı işaret dillerinde uygulanabilirliğini doğrulamak için Arjantin İşaret Dili'nden 64 işaret sözcüğü içeren LSA64 veri kümesi de kullanılmıştır. Ronchetti ve diğ. (2016b) tarafından oluşturulan bu veri kümesi, her bir kelimesinin on işaretçi tarafından beş kez tekrarlanmasıyla elde edilen, toplam 3.200

video görüntüsü içermektedir. LSA64 veri kümesine ait örnek görseller Şekil 3.2'de gösterilmiştir.

Şekil 3.2. LSA64 veri kümesinden örnek görüntü kareleri (Ronchetti ve diğ., 2016b)

LSA64 veri kümesi açıkça eğitim ve test olarak ayrılmamıştır. Bu nedenle, literatürdeki çalışmalar temel alınarak üç farklı deney seti oluşturulmuştur:

Deney 1 (D1): Marais ve diğ. (2022b) tarafından yapılan çalışmada olduğu gibi, beşinci ve onuncu işaretçiler test verisi olarak, geri kalanlar ise eğitim verisi olarak kullanılmıştır.

Deney 2 (D2): On işaretçiden dokuzu eğitim için ayrılmış, kalan bir işaretleyici ise test verisi olarak kullanılmıştır. Bu prosedür, her seferinde test için farklı bir işaretçi seçilerek on kez yinelenmiştir.

Deney 3 (D3): Veri kümesi rastgele olarak, %80'i eğitim ve %20'si test amacıyla ayrılmıştır.

3.1.3. GSL veri kümesi

Tez çalışması kapsamında önerilen yöntemlerin değerlendirildiği bir diğer veri kümesi ise Adaloglou ve diğ. (2020) tarafından oluşturulan ve 310 Yunan İşaret Dili kelimesi içeren GSL (Greek Sign Language) veri kümesidir. Bu veri kümesi yedi farklı işaretçi tarafından oluşturulmuş toplam 40.826 örnek içermektedir. Bir işaretçi tarafından elde edilen veriler, işaretçiden bağımsız değerlendirmeler için eğitim kümesinden çıkarılmıştır. Eğitim kümesinden çıkarılan verilerden 2.331'i doğrulama verisi, 3.500'ü

ise test verisi olarak ayrılmıştır. Veri kümesinden örnek görüntü kareleri Şekil 3.3'te gösterilmiştir.

Şekil 3.3. GSL veri kümesinden örnek görüntü kareleri (Adaloglou ve diğ., 2020)

3.2. MediaPipe Kütüphanesi

"Poz tahmini" ve "işaret noktası" terimleri, bilgisayarla görme ve makine öğrenimi tekniklerinin temel kavramlarındandır. "Poz tahmini", bir görüntüdeki insan figürünün vücut duruşunu anlamak için kullanılan bir tekniktir ve bu işlem, insanın vücut bölümlerinin konumlarını belirleyerek, genellikle 2D veya 3D koordinatlar cinsinden, her bir vücut ekleminin pozisyonunu tahmin eder. Elde edilen veriler spor analizleri, insan-makine etkileşimi ve işaret dili tanıma gibi çeşitli uygulamalarda kullanılabilir. "İşaret noktası" ise, insan vücudunun belirli anatomik noktalarını ifade eder; bu noktalar genellikle vücut eklemleri gibi belirgin noktalardır ve her biri, görüntü üzerindeki x ve y koordinatları ile bazen z derinlik koordinatı içeren bir konum bilgisi taşır. Örnek bir işaret dili görüntü karesinden elde edilen poz işaret noktaları Şekil 3.4'te gösterilmiştir.

Şekil 3.4. Örnek bir video karesinden poz işaret noktalarının elde edilmesi

Bu tez çalışmasında, işaret dili hareketlerinin doğasında bulunan uzamsal ve zamansal özellikleri elde etmek için poz verilerinden yararlanılmıştır. Hızla ilerleyen bilgisayarlı görme ve makine öğrenimi alanlarında, videolardan doğru insan pozu çıkarımı, İDT gibi uygulamalar için oldukça önemlidir. Önde gelen kütüphanelerden biri olan MediaPipe'ın Holistic çözümü (Lugaresi ve diğ., 2019; Grishchenko ve Bazarevsky, 2020) derin öğrenme yöntemlerini kullanarak gerçek zamanlı doğru poz tahmini için yüz, el ve vücut konumlarını izleyerek verimli bir çözüm sunar. MediaPipe'ın tercih edilmesinin nedenleri arasında gerçek zamanlı işleme kabiliyeti, yüksek doğruluk oranları ve kapsamlı poz tahmini özellikleri yer almaktadır. MediaPipe, İDT sistemlerinde karşılaşılan incelikli gereksinimleri etkili bir şekilde ele alarak yüksek performans ve düşük gecikme süresi sunar. Ayrıca açık kaynaklı olması, geniş bir geliştirici topluluğu tarafından desteklenmesi ve sürekli güncellenmesi bu teknolojinin esnekliğini ve uyarlanabilirliğini artırmaktadır.

Çalışmalarda, veri ön işleme adımının ilk aşamasında MediaPipe Holistic çözümü kullanılarak her bir video karesinden vücut, el ve yüz pozu işaret noktaları elde edilmiştir. Örnek bir video (V), $I_1, I_2, \dots, I_N$ olarak gösterilen N çerçeveden oluşur. Her çerçeve $W \times H \times C$ boyutlarına sahiptir. W (weight) ve H (height) sırasıyla çerçevenin genişliği ve yüksekliğidir, C (channel) renk kanallarının sayısıdır. M olarak gösterilen MediaPipe Holistic modeli, yüz ($L_{yüz,n}$), eller ($L_{eller,n}$) ve vücut ($L_{vücut,n}$) için bir dizi L_n işaret noktası kümesi çıkarmak üzere her kareyi işler (Denklem 3.1):

$$L_n = M(I_n) = \{L_{yüz,n}, L_{eller,n}, L_{vücut,n}\} \quad (3.1)$$

$L_{yüz,n}$ 468 işaret noktası, $L_{eller,n}$ 42 işaret noktası (her el için 21) ve $L_{vücut,n}$ 33 işaret noktası içerir. Her bir işaret noktası $l \in L_n$ (x, y, z) şeklinde üçlü olarak temsil edilir; burada x ve y işaret noktasının görüntü düzlemindeki piksel koordinatlarıdır, z ise kamera düzlemine göre derinlik bilgisidir.

3.3. Doğrusal Enterpolasyon

Doğrusal enterpolasyon, bilinen iki nokta arasındaki değerleri tahmin etmek için kullanılan bir yöntemdir. Bu yöntem, iki nokta, yani (x_1, y_1) ve (x_2, y_2) arasında doğrusal bir ilişki varsayarak, bu iki nokta arasındaki herhangi bir x değeri için y değerini hesaplar.

Matematiksel olarak, doğrusal enterpolasyon formülü (Noor ve diğ., 2015) Denklem 3.2'deki gibi ifade edilir:

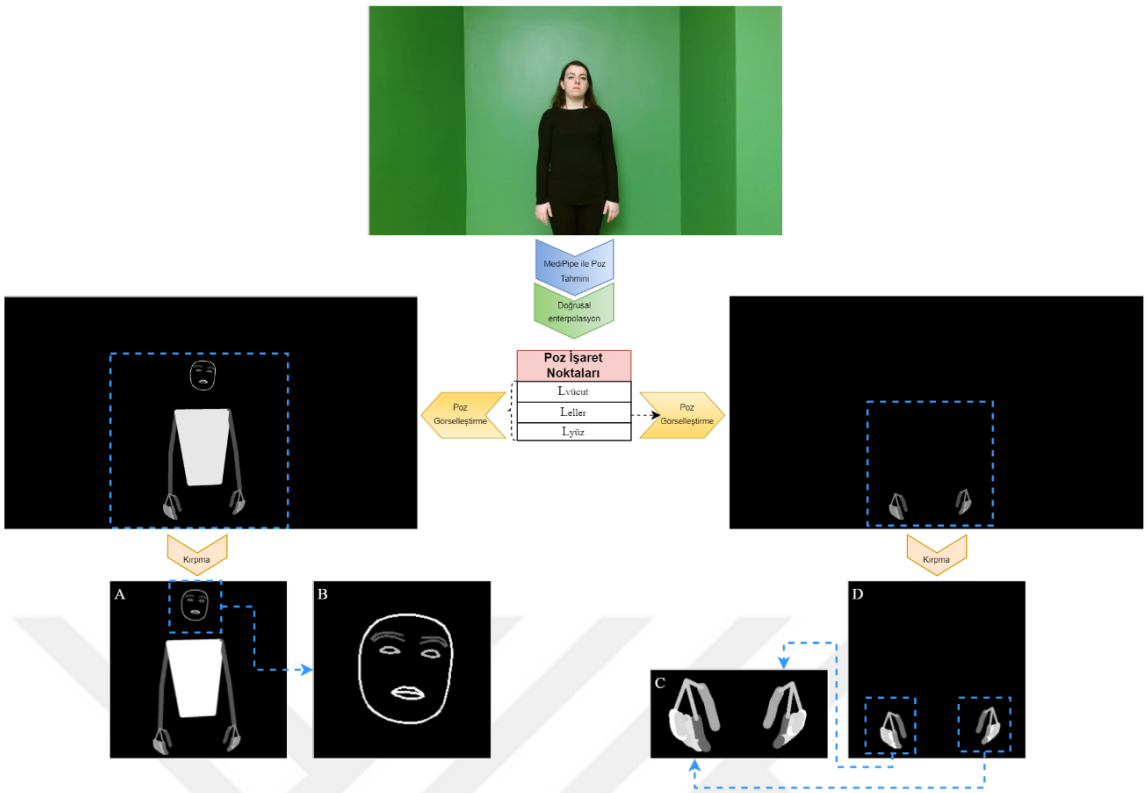
$$y = y_1 + \frac{(y_2 - y_1)}{(x_2 - x_1)} \times (x - x_1) \quad (3.2)$$

Bu denklem, x değerinin x_1 ile x_2 arasında olduğu durumlarda geçerlidir ve bu iki nokta arasındaki herhangi bir değeri tahmin etmek için kullanılır.

Doğrusal enterpolasyon yöntemi, basitliği ve az hesaplama gerektirmesi nedeniyle sıklıkla tercih edilir. Tez çalışması kapsamında bu yöntem eksik poz verilerinin tamamlanabilmesi için kullanılmıştır.

3.4. Poz Görselleştirme

MediaPipe Holistic çözümü kullanılarak her video karesinden vücut, eller ve yüz için işaret noktaları elde edildikten sonra, bu alt bölümde açıklanan adımlar kullanılarak görsel poz verileri oluşturulmuştur. Sayısal işaret noktalarının görsel temsillere dönüştürülmesi, İDT gibi alanlarda poz verilerinin yorumlanabilirliğini artırarak yakalanan pozların daha sezgisel bir şekilde analiz edilmesini sağlar. Bu görsel poz verileri Şekil 3.5'te gösterilmektedir.



Şekil 3.5. Bir görüntü karesinden görsel poz verilerinin elde edilmesi
(A) VÜCUT-POZU, (B) YÜZ-POZU, (C) BİRLEŞTİRİLMİŞ EL-POZU, (D) EL-POZU (Akdağ ve Baykan, 2024)

Şekil 3.5'te gösterilen VÜCUT-POZU görüntüsü vücudu, elleri ve yüzü tek bir karede görselleştiren kapsamlı bir temsil sağlar. Bu işlem, videonun (V) her bir karesi (I_n) için tüm işaret noktaları (L_n) ve bunlar arasındaki bağlantılar kullanılarak gerçekleştirilir. İlk olarak, her kare (I_n) için orijinal görüntü ile aynı boyutlarda ($W \times H$) boş bir siyah tuval (T_n) oluşturulur. Bu tuval, hem işaret noktalarını hem de aralarındaki bağlantıları çizmek için bir arka plan görevi görür. Her bir işaret noktası, $l \in L$ olmak üzere, bir işaret noktası $l = (x, y, z)$ şeklinde tanımlanabilir. Burada x ve y, işaret noktasının tuval üzerindeki piksel koordinatlarını, z ise derinlik bilgisini gösterir. Buradaki z bilgisi kullanılmamıştır, bu nedenle bir işaret noktasını $l = (x, y)$ olarak ifade edilmiştir. Her bir poz işaret noktası arasındaki bağlantı, belirli işaret nokta çiftleri arasındaki ilişkileri temsil eder. Her bir bağlantıya farklı bir renk ve kalınlık değeri atanabilir. Bağlantıların çizimi, başlangıç ve bitiş işaret noktaları arasında çizgiler çizilerek gerçekleştirilir. Çizim, $l_{başlangıç} = (x_{başlangıç}, y_{başlangıç})$ ve $l_{bitiş} = (x_{bitiş}, y_{bitiş})$ olarak tanımlanan başlangıç ve bitiş noktaları kullanılarak, her bir $bağlantı_k$, $bağlantı_k = (l_{başlangıç}, l_{bitiş})$ için yapılır. Çizim fonksiyonu aşağıda verilmiştir (Denklem 3.3):

$$\text{ÇizgiÇiz}(T_n, l_{\text{başlangıç}}, l_{\text{bitiş}}, \text{renk}_k, \text{kalınlık}_k). \quad (3.3)$$

ÇizgiÇiz fonksiyonu, tuval (T_n) üzerinde $l_{\text{başlangıç}}$ ve $l_{\text{bitiş}}$ noktaları arasında, belirtilen renk_k ve kalınlık_k ile bir çizgi çizer. Bu fonksiyon hem noktaların hem de aralarındaki bağlantıların görsel bir temsilini sağlar ve görselleştirme sürecinin temelini oluşturur.

Çalışmalar, vücut pozu görüntülerine ek olarak, işaret dilinin kritik bir bileşeni olan el hareketlerinin ayrıntılı analizine de odaklanmıştır. Bu amaçla, sadece elleri gösteren EL-POZU görüntüleri de oluşturulmuştur. Bu görüntüler, VÜCUT-POZU oluşturma adımlarına benzer adımlarla, yalnızca el işaret noktaları (L_{eller}) kullanılarak çizilmiştir.

VÜCUT-POZU ve EL-POZU verileri elde edildikten sonra bu görüntüler üzerinde odak alanı kırpma işlemi gerçekleştirilmiştir. Bu işlem, görsel verilerin gereksiz kısımlarını kırparak sonraki analiz aşamalarında kullanılmak üzere standart bir yapı oluşturmayı amaçlamaktadır. İlk olarak, kırpılacak veriler için sınırlayıcı kutular belirlenir. Bu sınırlayıcı kutular için, tüm video üzerinden elde edilen bir dizi L_n işaret noktası kullanılmıştır. Verilen L_n işaret noktalar kümesi için, yani her işaret noktası $l_i = (x_i, y_i)$ için, $i = 1,2,\ldots,N$ arasında x ve y'nin minimum ve maksimum değerlerini bulma işlemi matematiksel olarak aşağıdaki denklemlerde gösterilmiştir:

$$x_{min} = \min_{i=1}^{N} x_i, \quad (3.4)$$

$$y_{min} = \min_{i=1}^{N} y_i, \quad (3.5)$$

$$x_{max} = \max_{i=1}^{N} x_i, \quad (3.6)$$

$$y_{max} = \max_{i=1}^{N} y_i. \quad (3.7)$$

x ve y koordinatlarının minimum ve maksimum değerlerini bulduktan sonra, işaret noktalarının ve bağlantılarının görüntüde tamamen görünür olması için yeterli alan bırakmak üzere bir dolgu değeri eklenmiştir. Bu işlem sırasında, minimum değerlerin 0'dan küçük olmadığını ve maksimum değerlerin görüntünün yüksekliğinden (H) ve genişliğinden (W) büyük olmadığı kontrol edilmiştir. Bu düzeltmeler, sınırlayıcı kutunun görüntünün alanını aşmamasını ve ilgili tüm yer işaretlerini içermesini garanti eder. İşlemler aşağıdaki denklemlerde gösterilmiştir:

$$x'_{min} = \max(0, x_{min} - \text{dolgu}) \quad (3.8)$$

$$y'_{min} = \max(0, y_{min} - \text{dolgu}) \quad (3.9)$$

$$x'_{max} = \min(W, x_{max} + \text{dolgu}) \quad (3.10)$$

$$y'_{max} = \min(H, y_{max} + \text{dolgu}) \quad (3.11)$$

Bu işlemlerden sonra sınırlayıcı kutu, kare formatına yaklaştırılmıştır. Bunu yapmak için ilk olarak Denklem 3.12 ve Denklem 3.13'te gösterildiği gibi sınırlayıcı kutunun genişliği (W') ve yüksekliği (H') bulunur:

$$W' = x'_{max} - x'_{min}, \quad (3.12)$$

$$H' = y'_{max} - y'_{min}. \quad (3.13)$$

Sınırlayıcı kutunun genişliği yüksekliğinden büyükse, kutunun yükseklik ekseni etrafında genişletilmesi gerekir; aksi takdirde, kutunun genişlik ekseni etrafında genişletilmesi gerekir. Bu işlem için kullanılacak genişleme miktarı (Δ) Denklem 3.14 kullanılarak hesaplanır:

$$\Delta = \frac{|W' - H'|}{2} \quad (3.14)$$

Genişleme miktarı Δ bulunduğunda, sınırlayıcı bölgeyi genişletmek için Denklem 3.15 kullanılır:

$$\begin{cases} y'_{min} = \max(0, y_{min} - \Delta), & y'_{max} = \min(H, y_{max} + \Delta) \quad \text{eğer } W' > H', \\ x'_{min} = \max(0, x_{min} - \Delta), & x'_{max} = \min(W, x_{max} + \Delta) \quad \text{aksi halde} \end{cases} \quad (3.15)$$

Bu denklem, görüntünün fiziksel sınırlarını aşmadan sınırlayıcı bölgenin hem genişliğini hem de uzunluğunu orantılı olarak ayarlamak için gerekli koordinat bilgilerini sağlar. Denklem 3.16'da gösterilen kırpma işlemi, daha sonra videonun (V) her karesi için gerçekleştirilir:

$$I'_n = I_n[y'_{min} : y'_{max}, x'_{min} : x'_{max}]. \quad (3.16)$$

Bu işlem, videodaki (V) her bir kareye uygulandığında, tüm karelerin tutarlı ve odaklanmış bir şekilde kırpılmasını sağlar. Sonuç olarak, bu işlem, poz görüntülerinin ilgili alanına odaklanır ve gereksiz verileri ortadan kaldırır.

VÜCUT-POZU ve EL-POZU verileri elde edildikten sonra, bu veriler kullanılarak iki yeni özellik oluşturulmuştur: YÜZ-POZU ve BİRLEŞTİRİLMİŞ-EL-POZU. YÜZ-POZU, VÜCUT-POZU verisinden yüz bölgesinin $L_{yüz}$ işaret noktaları yardımıyla kırpılmasıyla elde edilir. BİRLEŞTİRİLMİŞ-EL-POZU, EL-POZU verisinden ellere karşılık gelen bölgelerin L_{eller} işaret noktaları aracılığıyla kırpılması ve ardından birleştirilmesiyle elde edilir. Bu kırpılmış özellikler, insan pozlarının inceliklerine daha yakından bakılmasını ve daha derinlemesine anlaşılmasını sağlar. YÜZ-POZU verilerinin oluşturulması, yüz ifadelerinin ve nüanslarının odaklanmış analizi için önemliyken, benzer şekilde, BİRLEŞTİRİLMİŞ-EL-POZU verileri, el hareketlerinin ve konumlarının ayrıntılı yorumlanması için daha zengin bir bağlam sağlar. Bu özelliklerin İDT sistemindeki tanıma doğruluğunu etkilemesi beklenmektedir.

Bu bölümde MediaPipe Holistic ile elde edilen poz işaret noktalarının, görsel poz verilerine dönüştürülme süreci genel olarak anlatılmıştır. Tez kapsamında gerçekleştirilen, her bir çalışmada oluşturulan diğer görsel poz verileri Önerilen Yöntemler ve Deneysel Çalışmalar bölümünde, ilgili önerilen yöntem altında anlatılmıştır.

3.5. Zamansal Video Analizi Teknikleri

3.5.1. Hareket Tarihçesi Görüntüsü (HTG)

Hareket Tarihçesi Görüntüsü (HTG), video karelerindeki hareket bilgisini tek bir görüntüde özetlemek için kullanılan bir temsil yöntemidir (Bobick ve Davis, 2001). HTG'ler hareket analizi ve hareket tabanlı sınıflandırma uygulamalarında yaygın olarak kullanılmaktadır. Özellikle insan etkileşimleri, jestler ve İDT gibi alanlarda önemlidirler (Zhang ve diğ., 2019; Naeem ve diğ., 2020; Chandragiri ve Ijjina, 2021; Sincan ve Keles, 2021; Ghosh ve diğ., 2023).

HTG'lerin altında yatan temel kavram, bir dizi video karesi içindeki hareketin zamansal ilerlemesini görsel olarak temsil etmektir. Bunun için hareketin başlangıcından sonuna kadar her karenin ağırlıklı bir kombinasyonu gereklidir. HTG algoritmasının

sonucu, her pikselin yoğunluğunun o pikseldeki hareketin zamansal yakınlığına göre hesaplandığı bir görüntüdür (Ahad, 2013).

HTG'nin hesaplanması, video karelerinin ardışık farklarının alınmasıyla başlar. Bu farklar hareketin nerede ve ne zaman meydana geldiğini gösterir. Ardından, bu hareket bilgisi tüm karelerde biriken bir HTG'de görüntülenir. HTG yalnızca hareketin yeri ve zamanı hakkında bilgi sağlamakla kalmaz, aynı zamanda hareketin yönü ve hızı gibi özellikleri de yakalar. Bu, hareketin dinamik özelliklerinin sınıflandırma için önemli olduğu durumlarda kullanışlıdır (Ahad ve diğ., 2012).

HTG'nin matematiksel formülü Denklem 3.17'de ifade edilmektedir:

$$H_\tau(x,y,t) = \begin{cases} \tau & \text{if } \psi(x,y,t) = 1, \\ \max(0, H_\tau(x,y,t-1) - \delta) & \text{aksi halde} \end{cases} \tag{3.17}$$

Bu denklemde, (x, y) piksel konumlarını ve t zamanı temsil eder. τ parametresi bir karenin hareket süresini ifade eder, δ bozunma parametresi olarak adlandırılır. ψ hareketin varlığını tanımlayan güncelleme fonksiyonudur. Bu güncelleme fonksiyonu Denklem 3.18'de her yeni video karesi analiz edildiğinde uygulanır:

$$\psi(x,y,t) = \begin{cases} 1 & \text{eğer } D(x,y,t) > \xi, \\ 0 & \text{aksi halde} \end{cases} \tag{3.18}$$

Bu denklemde ξ bir eşik parametresidir. $D(x,y,t)$, Denklem 3.19'da tanımlanan iki ardışık çerçeve arasındaki farkı temsil eder:

$$D(x,y,t) = |I(x,y,t) - I(x,y,t-1)| \tag{3.19}$$

Bu işlem sonucunda, yakın zamanda hareket etmiş olan pikseller daha parlak bir görüntü oluştururken, daha eski hareketler daha koyu bir görüntü oluşturur. Sonuç olarak HTG, bir video dizisindeki hareketi görselleştiren gri tonlamalı bir görüntüdür.

Gri tonlamalı HTG'nin ötesinde, RGB-HTG (Sincan ve Keles, 2021) HTG'nin temelleri üzerine inşa edilir ve zaman içinde hareket dinamiklerinin daha zengin ve daha ayrıntılı bir temsilini sağlamak üzere geleneksel HTG'yi genişletir. Bir RGB-HTG, video dizisini üç eşit parçaya bölerek, her biri için ayrı hareket geçmişleri hesaplayarak oluşturulur. Her renk kanalı (kırmızı, yeşil ve mavi) belirli zaman aralıklarında hareket

bilgisini temsil etmek üzere atanır. RGB-HTG hesaplaması, HTG hesaplamasının bir uzantısı olarak düşünülebilir ve RGB kanalları boyunca hareket verilerini işlemek ve derlemek için uyarlanmıştır.

3.5.2. Çerçeve Farkı Yöntemi

Çerçeve Farkı Yöntemi video analizinde yaygın olarak kullanılan bir hareket algılama yöntemidir (Zhan ve diğ., 2007; Singla, 2014; Husein ve diğ., 2019). Bu yöntem poz verilerinden elde edilen videolardan zamansal bilgi çıkarmak için kullanılmıştır. Bu yöntem, hareketin yoğunluğunu ve zamansal dağılımını etkili bir şekilde görselleştirdiği için hareket odaklı iletişim biçimlerini incelemek için idealdir (Husein ve diğ., 2019).

Çerçeve Farkı Yöntemi uygulanırken, ilk olarak ardışık çerçeveler arasındaki piksel farkının mutlak değeri hesaplanır. Formülasyon Denklem 3.20'de gösterilmiştir:

$$D(x, y, t) = |I(x, y, t) - I(x, y, t-1)| \tag{3.20}$$

Burada $I(x, y, t)$ t zamanında (x, y) koordinatındaki video akışının piksel değerini temsil eder ve $I(x, y, t-1)$ önceki $t-1$ zamanında (x, y) koordinatındaki piksel değeridir. Ortaya çıkan $D(x, y, t)$ t zamanındaki çerçeve farkını temsil eder, bu da hareketin bir göstergesi olarak kabul edilir. Hareket haritası daha sonra tüm video boyunca kare farklarının toplanmasıyla elde edilir. Burada ekstra olarak, kare farklarını tüm video üzerinden toplamak yerine videoyu üç eşit parçaya bölerek her parça için ayrı hareket haritaları elde edilmiştir, böylece her parça için elde edilen hareket haritası oluşturulan görüntünün bir kanalını temsil eder. Her parça için hareket haritası $F_k(x, y)$ (burada k kanal numarasıdır) o parçadaki tüm çerçeve farklarının toplamı olarak hesaplanır (Denklem 3.21):

$$F_k(x, y) = \sum_{t=\text{başlangıç}}^{\text{bitiş}} D(x, y, t) \tag{3.21}$$

Burada "başlangıç" ve "bitiş" her kanal için başlangıç ve bitiş kare numaralarıdır. Her kanalın hareket haritası $F_k(x, y)$, daha sonra normalleştirilir ve 0 ila 255 arasında ölçeklendirilir. Bu işlem, verileri görsel analiz için uygun hale getirir. Normalleştirme işlemi Denklem 3.22 kullanılarak gerçekleştirilir:

$$F_{norm,k}(x,y) = 255 \times \frac{F_k(x,y) - \min(F_k)}{\max(F_k) - \min(F_k)} \quad (3.22)$$

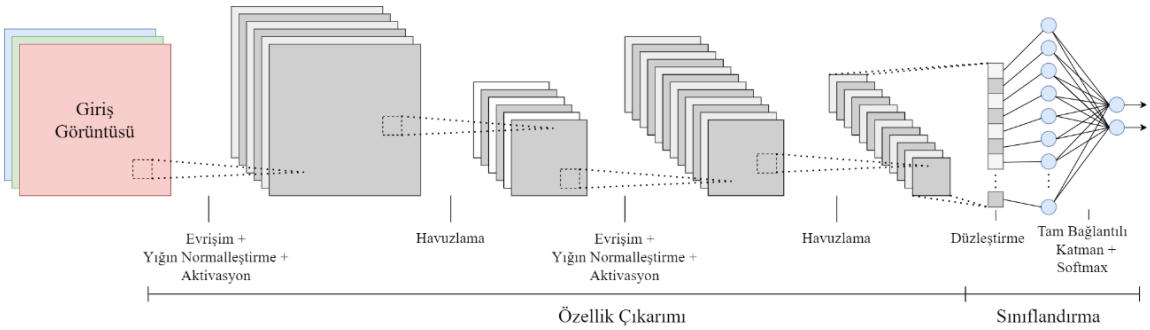
3.6. Evrişimsel Sinir Ağları (ESA)

Evrişimsel Sinir Ağları (ESA), bilgisayar bilimindeki derin öğrenme yöntemlerinin önemli bir alt dalıdır ve özellikle görüntü tabanlı uygulamalar için tasarlanmış güçlü bir sinir ağı türüdür. Bu ağlar, biyolojik görme sisteminden ilham alarak, verilerdeki desenleri tanımak için tasarlanmıştır. Bu nedenle, özellikle bilgisayarlı görü alanında birçok görevde önemli bir rol oynamaktadır.

ESA'lar modern anlamdaki temellerinin atıldığı, 1980'lerin sonlarından itibaren önemli gelişmeler kaydetmiştir. Bu dönemin en dikkat çekici gelişmesi, LeCun ve diğ. (1989) tarafından gerçekleştirilen, el yazısı posta kodlarını tanımak için tasarlanmış LeNet ağıdır. Geri yayılım algoritmasını kullanarak ağırlıkları otomatik olarak ayarlayabilen bu mimari, modern ESA'ların atası sayılabilir. 2000'lerin başında, ESA'lar, bilgisayar donanımlarının gelişimi ve büyük ölçekli veri setlerinin ortaya çıkışı ile yeniden popülerlik kazanmıştır. Bu dönemde, ESA'ların eğitimi için gerekli olan hesaplama gücünü sağlayan GPU'lar sayesinde, derin öğrenme algoritmalarının potansiyeli daha geniş çapta anlaşılmaya başlanmıştır. 2012 yılında, Krizhevsky ve diğ. (2012) tarafından geliştirilen AlexNet modelinin ImageNet yarışmasında elde ettiği başarı, ESA'lar için önemli bir dönüm noktası olmuştur. Derin yapısı, ReLU aktivasyon fonksiyonu ve etkili eğitim stratejileri sayesinde, AlexNet, ESA'ların kullanımını yeni bir seviyeye taşımıştır. Bu başarı ile birlikte, ESA'lar akademik ve endüstriyel araştırmalarda daha fazla kullanılmaya başlanmıştır.

Klasik makine öğrenimi ve geleneksel görüntü işleme yöntemleri, özellik çıkarımı ve özellik seçimi gibi önemli adımları gerçekleştirmek için genellikle alanında uzman kişilerin katkısına ihtiyaç duyar. Fakat ESA'lar, barındırdığı evrişim katmanları ile girdi olarak sunulan ham veriden doğrudan gerekli özellikleri çıkararak, bu süreçleri otomatize ederler. Evrişim katmanları, veri üzerinde lokal bir analiz yaparak, piksel gruplarından zengin özellikler çıkarır. Bu özellikler, daha sonraki aşamalarda nesne tanıma veya sınıflandırma gibi işlemler için kullanılır. Böylece ESA'lar, geleneksel metotlara kıyasla daha hızlı ve az müdahale ile daha güvenilir sonuçlar sunar. Bu özellikler, veriyi işleme sürecini hızlandırır ve insan kaynaklı hataları azaltır.

ESA mimarisi, iki ana bölümden oluşur: özellik çıkarımı ve sınıflandırma. Temel bir ESA mimarisinin yapısı Şekil 3.6'da gösterilmiştir. Devam eden alt bölümlerde burada yer alan bileşenler açıklanacaktır.



Şekil 3.6. Temel ESA mimarisi

3.6.1. Özellik çıkarımı

3.6.1.1. Giriş Katmanı

ESA'nın ilk katmanı olan Giriş Katmanı, modelin verilerle ilk etkileşime girdiği yerdir. Bu katman, modelin işleyeceği görüntülerin boyutunu ve türünü tanımlar; örneğin, renkli bir görüntü işleniyorsa üç kanal (RGB) alacak şekilde, gri seviye bir görüntü için ise tek kanal kullanılacak şekilde yapılandırılır.

Giriş görüntü boyutunun seçimi, ESA'nın eğitim sürecine büyük etki yapar. Büyük görüntü boyutları, modelin daha detaylı özellikleri algılamasına olanak tanırken, aynı zamanda bellek ihtiyacını, işlem süresini ve gereksinim duyulan hesaplama kaynaklarını artırır. Diğer yandan, daha küçük görüntü boyutları, hafıza kullanımını ve eğitim süresini azaltır fakat modelin genel başarısını olumsuz yönde etkileyebilir. Bu yüzden, uygun giriş görüntü boyutunun seçilmesi, ESA'nın donanımsal gereksinimleri ve performans hedefleri arasında denge kurulmasını gerektirir (Inik ve Ülker, 2017).

Giriş Katmanında yapılan seçimler, ESA'nın özellik çıkarımı ve sınıflandırma yeteneklerini doğrudan etkileyeceği için, bu katmanın tasarımı üzerinde özenle durulması önemlidir. Doğru boyutlandırma ve veri formatlaması, ESA'nın daha sonraki katmanlarda veriyi etkin bir şekilde işlemesini ve yüksek doğrulukla sonuçlar üretmesini sağlar.

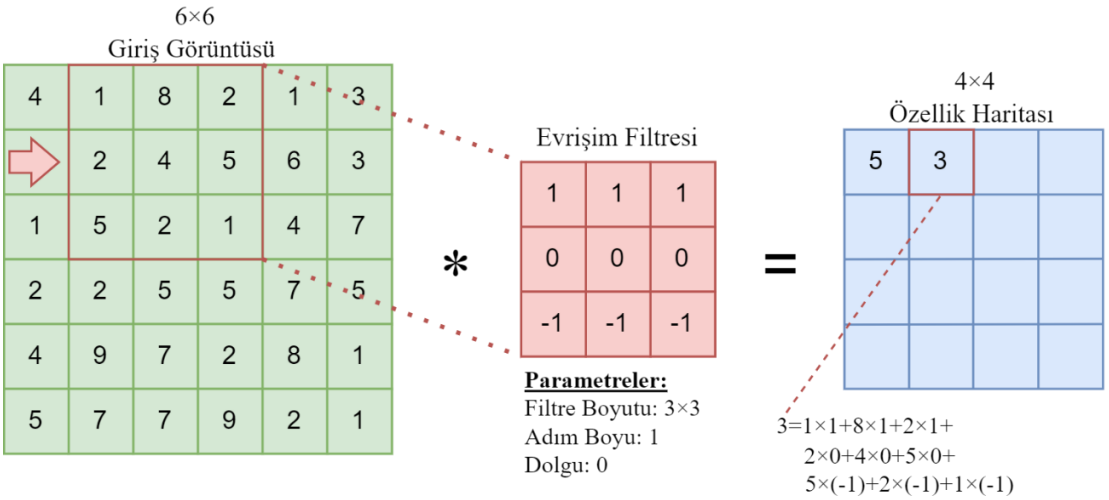
3.6.1.2. Evrişim (Konvolüsyon - Convolution) Katmanı

Evrişim Katmanı, ESA'nın görüntü gibi veri türlerinden karmaşık özellikleri çıkarmasını sağlayan temel bileşendir. Bu katman, belirli özellikleri yakalamak üzere tasarlanmış çok sayıda filtreyi veya çekirdeği (kernel) giriş verisi üzerinde kaydırarak uygular. Her bir filtre, verideki belirli bir özelliği -örneğin kenarlar, köşeler veya doku gibi- algılamak için tasarlanmıştır. Evrişim işlemi sırasında, her filtre giriş verisinin üzerinde gezdirilir ve her pozisyonda filtre ile giriş verisinin yerel bölgesi arasında bir çarpım toplamı hesaplanır. Bu işlem matematiksel olarak Denklem 3.23'teki gibi ifade edilir:

$$I * K = O(x, y) = \sum_{i=-k}^{k} \sum_{j=-k}^{k} I(x + i, y + j) \cdot K(i, j) \quad (3.23)$$

Burada, I giriş verisi (örneğin bir görüntü) matrisidir. K evrişim çekirdeği veya filtresidir. Bu bölümde ele alınan evrişim işlemleri için K sembolü 2 boyutlu evrişim çekirdeğini ifade eder. İlerleyen bölümlerde, üç boyutlu evrişim çekirdeklerinden ayırt edebilmek için K sembolü yerine K_{2D} sembolü kullanılacaktır. "$*$" matematiksel olarak evrişim işlemini temsil eder. $O(x, y)$ çıktı matrisi olup, giriş verisindeki her noktadaki özellik çıkarımının sonucunu gösterir. i ve j, filtre matrisinin sırasıyla yatay ve dikey indeksleridir. k, filtre boyutunun yarısıdır ve genellikle $k = 1$ için 3×3, $k = 2$ için 5×5 boyutunda bir filtre kullanıldığını gösterir.

Evrişim işleminde, filtre (K), giriş matrisi (I) üzerindeki her bir konumda uygulanır. Filtre merkezi (x, y) noktasına getirilir ve çevresindeki $2k + 1$ boyutundaki bölgeyle çarpımlar yapılır. Her (x, y) konumu için, filtre K'nın elemanları ile giriş matrisinin karşılık gelen elemanları çarpılır ve bu çarpımların toplamı $O(x, y)$ değerini verir. Bu işlem, filtre boyunca yinelemeli olarak yapılarak tüm görüntü boyunca özellik haritaları oluşturulur. Şekil 3.7'de örnek bir evrişim işlemi gösterilmiştir.



Şekil 3.7. Evrişim işlemi

Bu işlem sonucunda elde edilen özellik haritaları (feature maps), giriş verisinin farklı yönlerdeki veya özelliklerdeki bilgilerini temsil eder. Bu özellik haritaları, ağın daha sonraki katmanlarında daha karmaşık özellikleri tespit etmek ve nihayetinde girdi verisinin ne olduğunu anlamak için kullanılır. Her bir filtre, belirli bir deseni veya özelliği algılamak için özelleştirilmiş olabilir. Çeşitli filtrelerin kombinasyonu, giriş verisinin kapsamlı bir analizini sağlar.

Evrişim işlemi sırasında kullanılan iki önemli parametre adım boyu (stride) ve dolgudur (padding). Adım boyu, filtrenin her hareketinde kaç piksel kaydırılacağını belirler. Büyük bir adım boyu değeri, daha az özellik haritası boyutu üretir ve hesaplama yükünü azaltabilir, ancak aynı zamanda özellik çıkarma kabiliyetini sınırlayabilir. Dolgu, giriş verisinin etrafına eklenen sıfırlar (veya başka bir değer) ile filtrenin tam olarak uygulanabilmesi için giriş matrisinin boyutunu yapay olarak büyütür. Bu, özellikle giriş verisinin kenarlarında bulunan bilgilerin kaybolmasını önlemek için kullanılır.

Bu şekilde, Evrişim Katmanı, ESA'ların temel yapı taşlarından biri olarak görev yapar. Bu katman sayesinde, derin öğrenme modelleri, görüntü ve diğer benzeri veri türlerinden önemli özellikleri başarılı bir şekilde çıkarabilir ve bu özellikler üzerinden yüksek düzeyde sınıflandırma veya tanıma işlemleri gerçekleştirebilir. Evrişim katmanının doğru bir şekilde tasarlanması ve parametrelerinin ayarlanması, modelin genel başarımının ve doğruluğunun anahtarıdır (Maitra ve diğ., 2018).

3.6.1.3. 3D ve (2+1)D evrişimler

ESA'lar, karmaşık görüntü verilerinden özellikler çıkarma ve bu özellikleri kullanarak doğru sınıflandırmalar yapma yetenekleri sayesinde birçok görüntü işleme görevinde bir standart haline gelmiştir. Özellikle, İki Boyutlu Evrişimsel Sinir Ağları (2D-ESA'lar) yapısal basitlikleri ve verimli öğrenme algoritmalarıyla yaygın olarak kabul görmüştür (LeCun ve diğ., 2015). Bir 2D-ESA'da kullanılan temel evrişim fonksiyonu Denklem 3.24'teki gibi ifade edilebilir:

$$f_{2D}(I) = I * K_{2D} \tag{3.24}$$

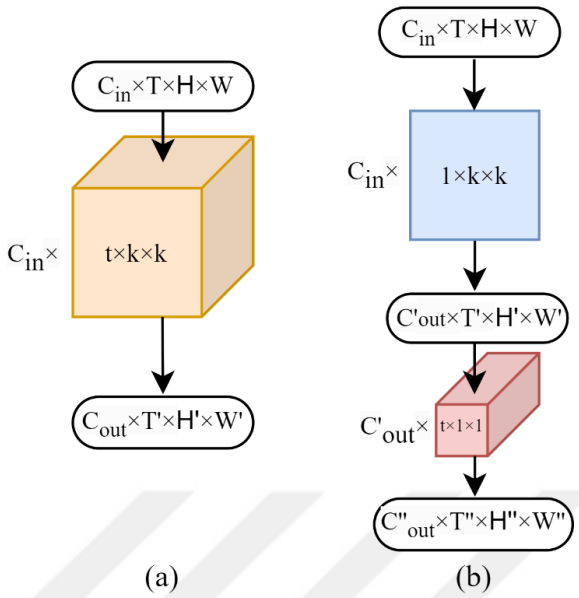
Burada, I bir görüntü matrisi, K_{2D} bir 2D evrişim çekirdeği ve "*" evrişim işlemini temsil eder. Bu işlem görüntülerdeki desen, kenar ve doku gibi özellikleri tespit etmek için kullanılır. Ancak, işaret dili kelimeleri gibi birbirini izleyen kareler arasındaki ilişkinin önemli olduğu video verilerinden zamansal özelliklerin çıkarılması için yetersizdir. Bu sorun, videodan hem zamansal hem de uzamsal bilgileri öğrenmek için 2D evrişime üçüncü bir boyut eklenerek elde edilen 3D evrişim filtreleri kullanılarak çözülmüştür (Tran ve diğ., 2015; Carreira ve Zisserman, 2017; Köpüklü ve diğ., 2021).

3D-ESA'lar (Taylor ve diğ., 2010; Ji ve diğ., 2013) olarak adlandırılan derin öğrenme ağları, videolar ve 3D tıbbi görüntüler dahil olmak üzere üç boyutlu verileri işleyebilen ağlardır. Bu ağlar, geleneksel ESA'lar tarafından kullanılan 2D filtrelerin aksine, verilerden hem uzamsal hem de zamansal özellikleri çıkarmak için 3D evrişim filtreleri kullanır. Bu yaklaşım, videolardaki nesnelerin hareketini ve dinamiklerini ve 3D görüntülerdeki yapılar arasındaki uzamsal ilişkileri etkili bir şekilde tespit etmeyi mümkün kılar. Bir video (V) üzerindeki 3D evrişim işlemi Denklem 3.25 ile gösterilebilir:

$$f_{3D}(V) = V * K_{3D} \tag{3.25}$$

3D evrişim işlemleri giriş verilerinin tüm kanallarını kapsayacak şekilde tasarlanmıştır; bu da her bir evrişim çekirdeğinin (K_{3D}) $C_{in} \times t \times k \times k$ boyutunda olduğu anlamına gelir. Burada, C_{in} giriş kanallarının sayısını, t zamansal derinliği ve k çekirdek boyutunu temsil etmektedir (Şekil 3.8.a). Bu yapı, çekirdeğin giriş verilerinin çok kanallı yapısını idare edebileceği ve tüm kanalları aynı anda işleyeceği anlamına gelir. Bu işlem sonucunda, çekirdek tarafından çıkarılan özellikler tüm kanalların bilgilerini içerir ve

C_{out} özellik haritaları ortaya çıkar. Bu yaklaşım, 3D evrişimin gelişmiş uzamsal-zamansal özellik çıkarma kabiliyetinin temelini oluşturur.



Şekil 3.8. Evrişim türleri: (a) 3D evrişim; (b) (2+1)D evrişim (Akdag ve Baykan, 2024a)

Ardışık kareler arasındaki zamansal bilgileri çıkarmak için kullanılan bir diğer yapı da (2+1)D evrişimdir (Tran ve diğ., 2018). (2+1)D evrişim, geleneksel 3D evrişim işlemini uzamsal ve zamansal boyutlara ayırarak, önce her video karesindeki uzamsal bilgileri 2D evrişimler aracılığıyla çıkarmakta, ardından elde edilen uzamsal özellik haritaları üzerinde 1D evrişimler aracılığıyla zamansal özellikleri işlemektedir (Şekil 3.8.b). Bir videonun (V) (2+1)D evrişimi Denklem 3.26 ile ifade edilebilir.

$$f_{1D}(f_{2D}(V)) = (V * K_{2D}) * K_{1D} \tag{3.26}$$

Bu ayrıştırılmış yaklaşım, modelin hem uzamsal ayrıntıları hem de zamansal varyasyonları daha etkili bir şekilde öğrenmesini sağlar (Huang ve diğ., 2021). Ayrıca, 2D ve 1D evrişim işlemleri arasına aktivasyon fonksiyonlarının eklenmesi ağın karmaşıklığını ve doğrusal olmayan öğrenme kapasitesini artırmaktadır. Bu strateji, uzamsal ve zamansal boyutlar arasındaki ilişkileri daha iyi yakalamak için modelin doğrusal olmayan özelliklerini güçlendirir, böylece modelin genel performansını artırır (Tran ve diğ., 2018).

3.6.1.4. Yığın Normalleştirme (Batch Normalization)

Yığın Normalleştirme (Ioffe ve Szegedy, 2015), Evrişimsel Sinir Ağlarında sıklıkla kullanılan bir tekniktir. Bu teknik, her mini-yığın için özelliklerin ortalamasını ve varyansını normalleştirerek, katmanların giriş dağılımlarını düzenler. Bu düzenleme, modelin eğitim sürecinde karşılaşabileceği içsel uyumlu değişiklikleri azaltır ve daha hızlı öğrenmeyi sağlar.

Yığın Normalleştirmenin işleyişinde, ilk adımda, mini-yığının her bir özelliği için ortalama değerler hesaplanır. Ardından, bu özelliklerin varyansı bulunur. Elde edilen bu değerler kullanılarak, her özellik sıfır ortalama ve birim varyansa sahip olacak şekilde normalleştirilir. Son olarak, modelin öğrenme sürecine katkıda bulunmak için normalleştirilmiş özelliklere ölçekleme ve kaydırma işlemleri uygulanır. Bu ölçekleme ve kaydırma parametreleri, model tarafından öğrenilen değerlerdir ve her özellik grubunun veri setindeki farklılıklara adaptasyonunu sağlar. Denklem 3'teki formül Yığın Normalleştirme (YN) sürecini göstermektedir:

$$\mathrm{YN}(x) = \gamma\left(\frac{x-\mu_B}{\sqrt{\sigma_B^2+\epsilon}}\right) + \beta \tag{3.27}$$

Bu denklemde μ_B ve σ_B^2 sırasıyla mini yığının ortalaması ve varyansıdır. γ ve β ise ölçekleme ve kaydırma için öğrenilebilir parametrelerdir.

Yığın Normalleştirme, eğitim sürecini hızlandırarak daha yüksek öğrenme oranlarının kullanılmasına olanak tanır. Bu da modelin daha çabuk ve etkin bir şekilde eğitilmesine yardımcı olur. Ayrıca, modelin başlangıç parametrelerine olan duyarlılığını azaltır, bu da farklı başlangıç ağırlıklarıyla modelin daha tutarlı sonuçlar üretmesine imkan tanır. Normalleştirme işlemi, aynı zamanda aşırı uyuma karşı bir koruma sağlayarak modelin genelleme yeteneğini artırır (Ioffe ve Christian, 2017).

Yığın Normalleştirme, özellikle derin öğrenme modellerinin performansını ve stabilitesini önemli ölçüde artırabilen temel bir tekniktir. Bu yöntem, karmaşık sinir ağlarının eğitiminde karşılaşılan zorlukların üstesinden gelmede kritik bir rol oynar (Ioffe ve Szegedy, 2015).

3.6.1.5. Aktivasyon fonksiyonu

Aktivasyon fonksiyonları, ESA içinde, evrişim katmanları tarafından üretilen özellik haritalarına uygulanır. Bu fonksiyonlar, ağın doğrusal olmayan özellikleri işleyerek karmaşık fonksiyonları modellemesini sağlar ve modelin eğitim sürecinde verideki doğrusal olmayan ilişkileri öğrenmesine yardımcı olur (Kılıçarslan ve diğ., 2021). Yaygın olarak kullanılan bazı aktivasyon fonksiyonları şunlardır:

Sigmoid: Girişi 0 ile 1 arasında bir değere sıkıştırır ve genellikle ikili sınıflandırma görevlerinin çıktı katmanında kullanılır. Matematiksel formülü Denklem 3.28'de verilmiştir.

$$\text{sigmoid}(x) = \frac{1}{1+e^{-x}} \tag{3.28}$$

Hiperpolik Tanjant Fonksiyonu (tanh): Bu fonksiyon, girişi -1 ile 1 arasında bir değere sıkıştırır ve genellikle sigmoid fonksiyonuna göre daha iyi performans gösterir. Matematiksel formülüne Denklem 3.29'da yer verilmiştir.

$$\tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{3.29}$$

ReLU (Rectified Linear Unit): Bu aktivasyon fonksiyonu ESA'larda en yaygın kullanılan aktivasyon fonksiyonlarının başında gelir, Denklem 3.30'daki gibi tanımlanabilir:

$$\text{ReLU}(x) = \max(0, x) \tag{3.30}$$

ReLU aktivasyon fonksiyonu herhangi bir pozitif girdiyi olduğu gibi tutarken, herhangi bir negatif girdiyi sıfıra dönüştürerek çalışır. Bu basitlik sadece hesaplama verimliliğini kolaylaştırmakla kalmaz, aynı zamanda kaybolan gradyan problemini hafifletmeye yardımcı olur ve böylece öğrenme sürecini hızlandırır.

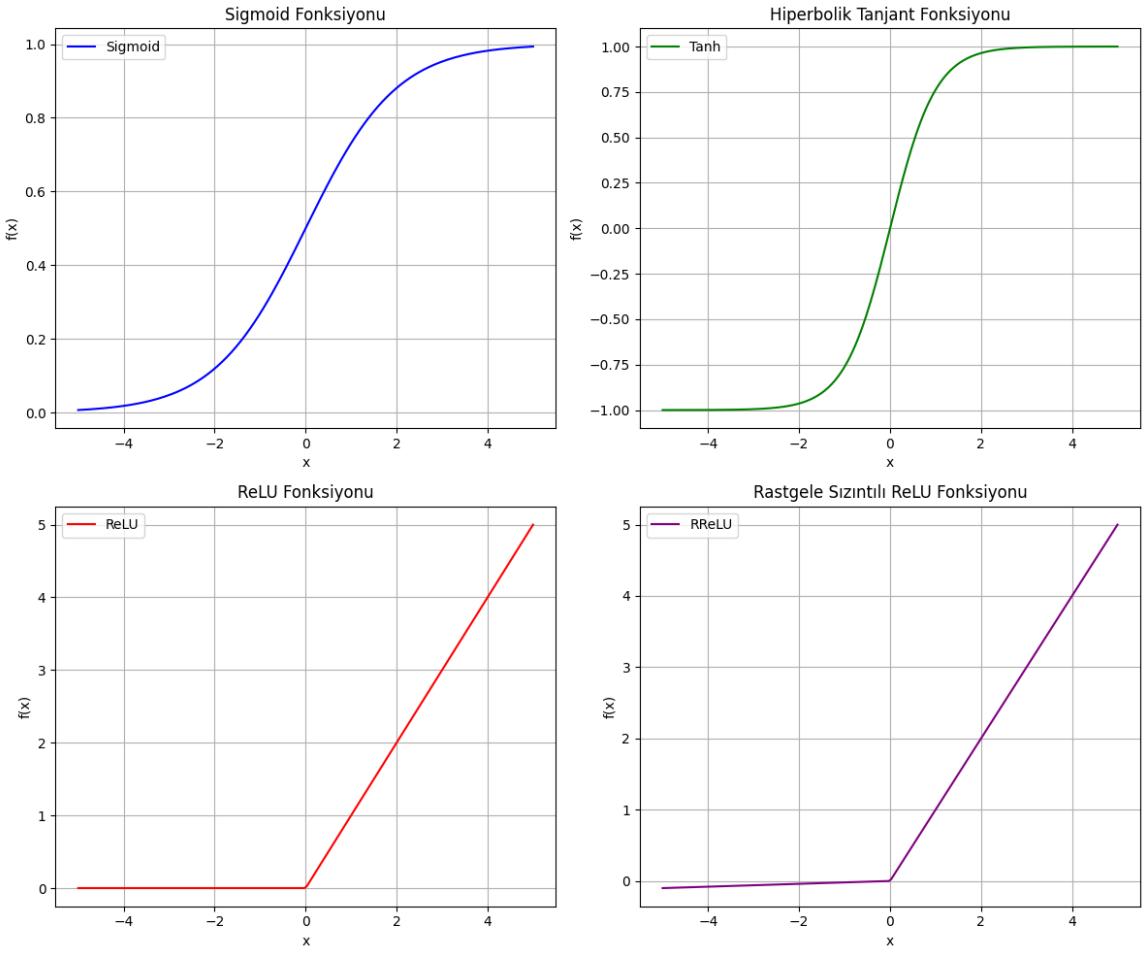
Randomized Leaky Rectified Linear Unit (RReLU): Önceden tanımlanmış alt ve üst sınırlar dahilinde negatif aktivasyon değerleri için rastgele bir eğim seçerek

parametrik olmayan bir davranış sergiler. Bu yaklaşım, modelin öğrenme ve genelleme kapasitesini artırarak aşırı uyuma karşı sağlamlığını önemli ölçüde geliştirir (Xu ve diğ., 2015). RReLU aktivasyon fonksiyonu matematiksel olarak şu şekilde gösterilebilir:

$$\text{RReLU}(x) = \begin{cases} x & \text{eğer } x \geq 0 \\ ax & \text{aksi halde} \end{cases} \tag{3.31}$$

Bu denklemde a, eğitim aşamasında U(alt, üst) tekdüze dağılımından rastgele örneklenen bir katsayıdır. Varsayılan olarak alt sınır 1/8 ve üst sınır 1/3 olarak ayarlanmıştır. Bu rastgelelik aktivasyona değişkenlik katarak modelin daha iyi genelleme yapabilmesine ve aşırı uyumu engellemesine katkıda bulunur. Ancak değerlendirme veya test aşamasında a, alt ve üst sınırların ortalaması olarak hesaplanan sabit bir değere ayarlanır.

Bu aktivasyon fonksiyonlarının grafikleri Şekil 3.9'da gösterilmiştir.



Şekil 3.9. Aktivasyon fonksiyonları

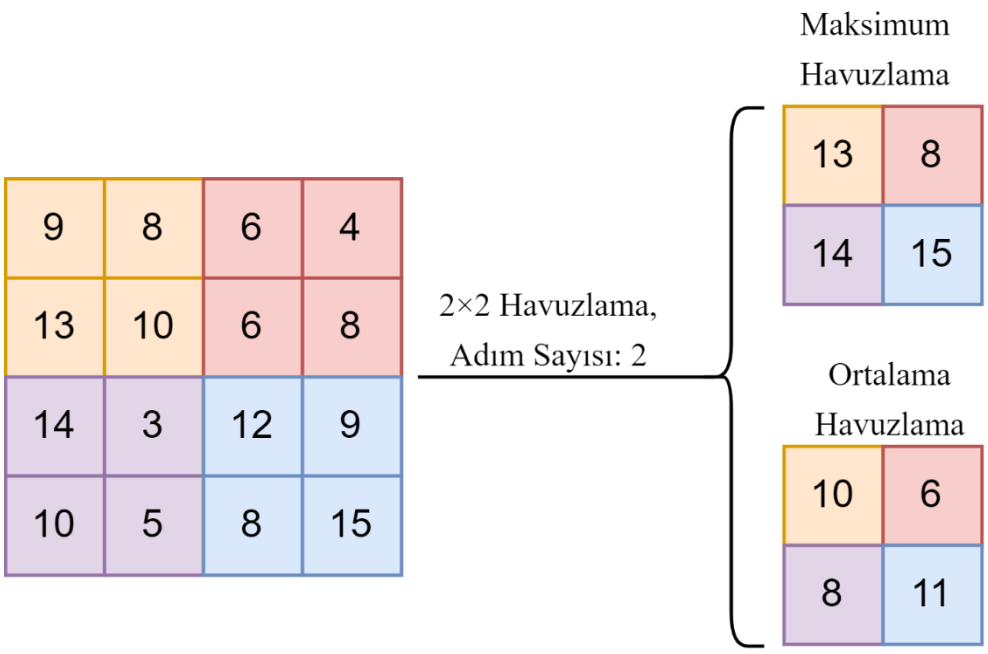
3.6.1.6. Havuzlama Katmanı (Pooling Layer)

Havuzlama katmanı, ESA içerisinde özellik haritalarının boyutunu azaltmak ve önemli özellikleri koruyarak hesaplama yükünü düşürmek için kullanılır. Bu katman, özellik haritalarındaki uzamsal boyutların sıkıştırılmasına yardımcı olur ve bu sayede modelin aşırı uyum (overfitting) riskini azaltırken, özelliklerin konumlarındaki küçük değişikliklere karşı daha dayanıklı hale gelmesini sağlar. Havuzlama işlemi genellikle maksimum veya ortalama havuzlama olarak iki farklı şekilde yapılır (Li ve diğ., 2019):

Maksimum Havuzlama: Bu yöntemde, belirlenen bir pencere boyutu içindeki en büyük değer, özellik haritasından seçilir. Bu, özellik haritalarının en belirgin özelliklerini korurken gereksiz bilgilerin filtreden geçirilmesini önler. Genellikle görsel özelliklerin dikkate alınması gereken görevlerde tercih edilir.

Ortalama Havuzlama: Bu yöntemde ise pencere boyutu içindeki tüm değerlerin ortalaması alınır. Bu, özelliklerin daha genel bir temsilini sağlar ve daha yumuşak bir bilgi sıkıştırma sunar.

Havuzlama işlemlerine ait örnek bir gösterim Şekil 3.10'da gösterilmiştir.

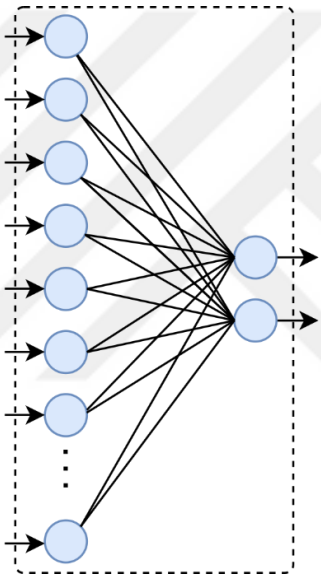


Şekil 3.10. Maksimum ve Ortalama Havuzlama örneği

3.6.2. Sınıflandırma

3.6.2.1. Tam Bağlantılı Katman (Fully Connected Layer)

Tam Bağlantılı Katman, ESA ve diğer derin öğrenme modellerinde sıklıkla kullanılan bir katmandır. Bu katman genellikle ağın çıktı katmanına yakın yer alır ve modelin nihai kararlarını şekillendirmede önemli bir role sahiptir (Scabini ve Bruno, 2023). Örneğin sınıflandırma görevlerinde, Tam Bağlantılı Katmanlar, evrişim katmanlarından gelen zengin özellik haritalarını alır ve bu bilgileri belirli sınıflarla ilişkilendirir. Örnek bir temsil Şekil 3.11'de gösterilmiştir.



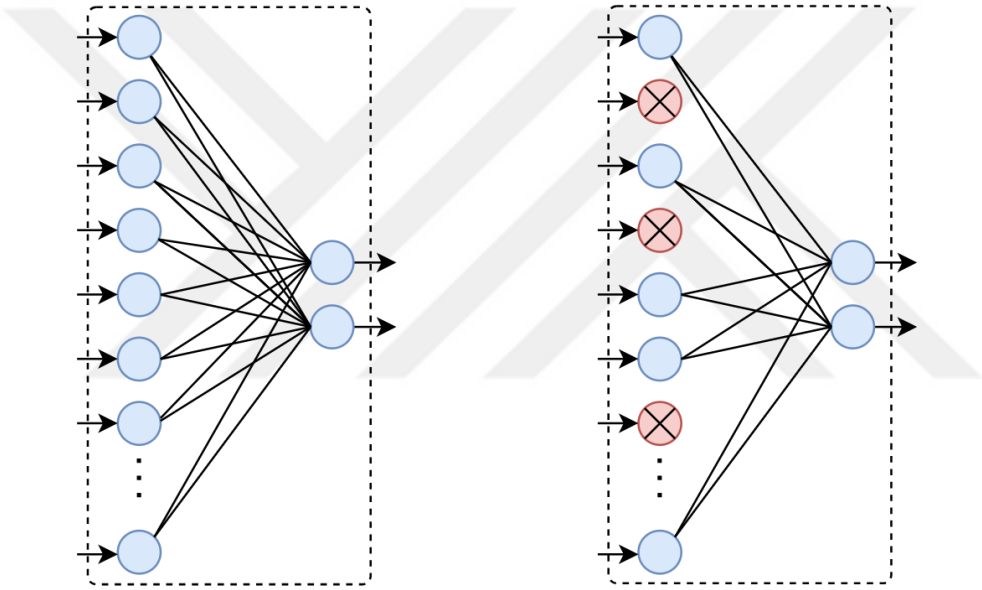
Şekil 3.11. Tam Bağlantılı Katman

Tam Bağlantı Katmanları, modelin öğrenme kapasitesini artırır, çünkü bu katmanlar yüksek düzeyde parametre içerirler. Bu parametreler, veri setindeki karmaşık ilişkileri ve düzenlilikleri modellemek için ayarlanabilir.

Tam Bağlantı Katmanının tasarımı, modelin genel mimarisine ve çözülmesi gereken problemin doğasına bağlı olarak değişiklik gösterir. Aşırı uyum riskini azaltmak için düzenlileştirme teknikleri (örneğin, Düğüm Seyreltme) sıkça kullanılır. Ayrıca, modelin eğitim süresini ve hesaplama gereksinimlerini optimize etmek amacıyla katman sayısının ve nöron sayısının dikkatlice seçilmesi gerekir.

3.6.2.2. Düğüm Seyreltme (Dropout)

Düğüm Seyreltme (Dropout), yapay sinir ağlarının eğitim sürecinde kullanılan bir düzenlileştirme tekniğidir ve özellikle derin öğrenme modellerinde aşırı uyum sorununun önlenmesine yardımcı olur. Bu teknik, rastgele bir şekilde belirli nöronların (ve bunların bağlantılarının) eğitim sürecinin her bir iterasyonunda aktif olmamalarını sağlayarak çalışır. Bu işlem, ağın farklı nöron alt kümeleri kullanarak eğitim yapmasını ve dolayısıyla daha sağlam, genelleyici bir model oluşturmasını teşvik eder (Salehin ve Kang, 2023). Şekil 3.12'de Tam Bağlantılı Katman örneğine Düğüm Seyreltme işleminin uygulanması gösterilmiştir.



Şekil 3.12. Düğüm Seyreltme işleminin gösterimi

Düğüm Seyreltme, modelin sadece belirli bir grup nörona bağımlı hale gelmesini engeller ve böylece farklı nöron kombinasyonlarının öğrenilmesine olanak tanır. Eğitim sürecinde Düğüm Seyreltme uygulanırken, test aşamasında tüm nöronlar kullanılır ve her nöronun çıktısı, eğitim sırasında aktif bırakılma olasılığına göre ölçeklenir. Bu yöntem, sinir ağlarının daha dayanıklı hale gelmesine ve görülmemiş veriler üzerinde daha iyi performans göstermesine katkıda bulunur. Düğüm Seyreltme, basit ve etkili bir yöntem olarak kabul edilir ve pek çok modern sinir ağı mimarisinde standart bir bileşen olarak yer alır (Srivastava ve diğ., 2014).

3.6.2.3. Softmax Katmanı

Softmax Katmanı, ESA içinde, sınıflandırma aşamasının sonunda yer alan, sınıflandırma görevlerinde kullanılan bir bileşendir. Bu katman, önceki katmanlardan gelen çıktıları, olasılık dağılımına dönüştürerek her bir sınıfın tahmin olasılığını hesaplar. Çok sınıflı sınıflandırma görevlerinde özellikle önemlidir çünkü modelin tahminlerini, gerçek sınıflara göre normalize edilmiş bir biçimde sunar (Ther ve diğ., 2023). Softmax fonksiyonunun matematiksel ifadesi şu şekildedir:

$$\text{softmax}(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \tag{3.32}$$

Burada x_i ifadesi, i indeksli sınıfa ait skorları (logitleri) temsil eder ve bu skorlar, model tarafından hesaplanan tahminlerdir. Eşitlikteki e^{x_i} terimi, i indeksli sınıfın skorunun üssünü alırken, paydada bulunan toplam terim ($\sum_{j=1}^{n} e^{x_j}$), tüm sınıf skorlarının üslerinin toplamıdır. Bu yapı, her bir sınıf için tahmin edilen değerlerin, tüm sınıflar arasında oransal bir biçimde dağılmasını sağlar ve çıktılar 0 ile 1 arasında bir değer alır. Böylece, tüm sınıfların olasılıklarının toplamı 1 olur (Ther ve diğ., 2023).

Softmax Katmanının işlevselliği, sınıflandırma görevinde son derece yararlıdır. Çünkü bu katman sayesinde, çeşitli sınıflara ait olasılıklar arasındaki farkları belirginleştirerek, modelin karar verme sürecini optimize eder. Dolayısıyla modelin genel performansını ve tahminlerinin doğruluğunu artırmada önemli bir faktördür.

3.6.3. Eğitim süreci ve parametreler

3.6.3.1. Kayıp fonksiyonları (loss functions)

Kayıp fonksiyonları, ESA gibi derin öğrenme modellerinde, modelin eğitimi sırasında ne kadar iyi performans gösterdiğini ölçen kritik bir bileşendir. Modelin tahminleri ile gerçek değerler arasındaki farkı matematiksel olarak ifade eden bu fonksiyonlar, modelin öğrenme sürecinin temelini oluşturur (Tian ve diğ., 2022). Kayıp fonksiyonları eğitim süresince, modelin hatalarını geriye yayarak (backpropagation) öğrenmesini sağlar. Bu fonksiyonlar sayesinde, ağın ağırlıkları gerçek çıktılara

yaklaşacak şekilde güncellenir (Terven ve diğ., 2023). Bu süreç, modelin eğitimi boyunca devam eder. ESA için en sık kullanılan kayıp fonksiyonları şunlardır:

Ortalama Kare Hata (Mean Squared Error - MSE): Gerçek değerler ile tahmin edilen değerler arasındaki farkların kareleri toplamının ortalaması alınarak hesaplanır. Formülü Denklem 3.33'te verilmiştir (Kerkhof ve diğ., 2023):

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{3.33}$$

Burada Y_i gerçek değerleri, $\hat{Y}_i$ ise modelin tahminlerini ifade eder ve n örnek sayısını belirtir. MSE, her bir tahminin gerçek değerden ne kadar uzak olduğunu kareleyerek hesaplar, böylece büyük hatalara daha fazla ceza verir (Bishop ve Nasrabadi, 2006).

Çapraz Entropi Kaybı (Cross-Entropy Loss): Sınıflandırma görevlerinde kullanılan bu kayıp fonksiyonu, modelin çıktılarının ve gerçek etiketlerin olasılık dağılımları arasındaki farkı ölçer. İki sınıflı (binary) ve çok sınıflı (multiclass) sınıflandırma problemlerinde etkilidir (Murphy, 2012; Goodfellow ve diğ., 2016).

İkili Çapraz Entropi (Binary Cross-Entropy – BCE):

$$\mathrm{BCE} = -\frac{1}{n}\sum_{i=1}^{n}[Y_i \log(\hat{Y}_i) + (1 - Y_i)\log(1 - \hat{Y}_i)] \tag{3.34}$$

Burada Y_i gerçek sınıf etiketleri (0 veya 1), $\hat{Y}_i$ ise modelin tahmin ettiği olasılıkları ifade eder.

Çok Sınıflı Cross-Entropy (Multiclass Cross-Entropy – MCE):

$$\mathrm{MCE} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{C} Y_{i,c}\log(\hat{Y}_{i,c}) \tag{3.35}$$

Burada, n toplam örnek sayısını, C sınıf sayısını, $Y_{i,c}$ ise i'inci örneğin c'inci sınıfa ait olma olasılığını (gerçek), $\hat{Y}_{i,c}$ ise modelin tahmin ettiği olasılığı ifade eder.

3.6.3.2. Optimizasyon algoritmaları

Optimizasyon algoritmaları, ESA ve diğer derin öğrenme modellerinin eğitimi sırasında kullanılan temel araçlardır (Goodfellow ve diğ., 2016). Bu algoritmalar, modelin kayıp fonksiyonunu minimize ederek, model parametrelerini en iyi şekilde ayarlar. Optimizasyon süreci, modelin öğrenme verimliliğini ve sonuçlarının doğruluğunu doğrudan etkiler (Bishop ve Nasrabadi, 2006).

Aşağıda, ESA modellerinde yaygın olarak kullanılan bazı temel optimizasyon algoritmaları ve bunların özellikleri yer almaktadır:

Gradyan İnişi (Gradient Descent): Bu temel optimizasyon tekniği, kayıp fonksiyonunun gradyanını hesaplar ve gradyanın işaretine göre parametreleri günceller. Bu yöntem, her iterasyonda tüm veri seti üzerinden gradyanı hesaplar; bu nedenle, büyük veri setleri için hesaplama açısından pahalı olabilir (Bottou, 2010). Bu yöntem Denklem 3.36'da gösterilmiştir:

$$\theta = \theta - \eta \nabla_{\theta} J(\theta) \tag{3.36}$$

Burada θ, model parametreleri, η öğrenme oranı, $J(\theta)$ ise kayıp fonksiyonudur.

Stokastik Gradyan İnişi (Stochastic Gradient Descent - SGD): Gradyan iniş yönteminin bir varyasyonu olan SGD, daha hızlı ve daha verimli bir şekilde yakınsama sağlamak için her iterasyonda rastgele seçilen bir veri ya da veri grubu üzerinden gradyanı hesaplar (Bottou, 2010). Bu yaklaşım, her bir iterasyonda tüm veri setini kullanmak yerine, rastgele seçilmiş bir mini-yığın kullanarak hesaplama yükünü azaltır ve modelin genelleme kabiliyetini artırır (Murphy, 2012). Bu yöntem, pratik uygulamalarda sıklıkla tercih edilen bir yaklaşımdır çünkü hem hızlı yakınsama hem de hesaplama verimliliği sağlar. Denklem 3.37, $(x^{(i)}, y^{(i)})$ rastgele seçilen veri ya da veri grubunu göstermek üzere, SGD'nin matematiksel denklemini göstermektedir.

$$\theta = \theta - \eta \nabla_{\theta} J(\theta; x^{(i)}, y^{(i)}) \tag{3.37}$$

Uyarlanabilir öğrenme oranı yöntemleri: Bu yöntemler (örneğin, AdaGrad, RMSprop, Adam), her bir parametre için öğrenme oranını adaptif olarak ayarlar.

Parametrelerin her biri farklı hızlarda güncellenir, bu da farklı özelliklerin önemine bağlı olarak daha hızlı yakınsama sağlar (Kingma ve Ba, 2014).

Optimizasyon algoritmalarının doğru seçimi, ESA modellerinin eğitim sürecinin başarısını büyük ölçüde etkiler. Modelin hızlı ve etkili bir şekilde eğitilmesi, uygun bir optimizasyon algoritması kullanılarak sağlanabilir. Bu algoritmalar, kayıp fonksiyonunun minimuma indirilmesinde kritik bir rol oynar ve modelin genel performansını ve tahmin doğruluğunu doğrudan etkiler (Ruder, 2016).

3.6.3.3. Eğitim parametreleri

Yığın boyutu (batch size): Yığın boyutu, derin öğrenme ve özellikle ESA modellerinin eğitimi sırasında her bir iterasyonda ağa beslenen eğitim örneklerinin sayısını belirler (Goodfellow ve diğ., 2016). Yığın boyutu seçimi, modelin eğitim sürecinde kullanılan bellek miktarı, eğitim süresi ve modelin genel performansı üzerinde doğrudan bir etkiye sahiptir.

Farklı boyutlardaki yığınlar, modelin öğrenme dinamiklerini ve sonuçta elde edilen modelin genelleme kabiliyetini farklı şekillerde etkileyebilir. Küçük yığın boyutları, modelin eğitim süresince ağırlıkları daha sık güncellemesine olanak tanır, bu da potansiyel olarak daha hızlı yakınsama sağlayabilir (Masters ve Luschi, 2018). Ayrıca, küçük yığınlar genellikle modelin yerel minimumlara takılma olasılığını azaltır ve daha iyi genelleme performansı sunar (Keskar ve diğ., 2016). Ancak, çok küçük yığın boyutları kullanmak, her bir güncellemede gürültülü gradyan tahminleri üretebilir ve bu da modelin eğitim sürecinin istikrarsızlaşmasına neden olabilir. Ayrıca, donanım kaynaklarını verimsiz kullanabilir çünkü her seferinde yalnızca birkaç örnek üzerinden hesap yapılır (Goodfellow ve diğ., 2016).

Büyük yığınlar, her bir iterasyonda daha fazla örnek üzerinden gradyan hesaplaması yaparak daha kararlı ve güvenilir gradyan tahminleri sağlar. Bu, özellikle büyük veri setlerinde hesaplama açısından daha verimli olabilir, paralel işleme ve vektörleştirme avantajlarından faydalanılabilir (Loukili ve diğ., 2022). Ancak, modelin yerel minimumlardan çıkmasını zorlaştırabilir ve bazı durumlarda modelin global minimuma ulaşmasını engelleyebilir. Ayrıca, büyük yığın boyutları daha fazla bellek gerektirir (Piao ve diğ., 2023).

Yığın boyutunun seçimi, genellikle veri setinin büyüklüğü, modelin karmaşıklığı ve kullanılan donanımın kapasitesine bağlı olarak yapılır. Deneme-yanılma yoluyla en

uygun yığın boyutunu belirlemek, çoğu makine öğrenmesi uygulamasının bir parçasıdır. Eğitim verilerinin çeşitliliği ve dağılımı göz önünde bulundurulduğunda, optimal yığın boyutu genellikle en iyi dengeyi sağlayacak şekilde seçilir (Perrone ve diğ., 2019; Usmani ve diğ., 2023).

Sonuç olarak, yığın boyutu, modelin eğitim sürecini ve sonuçta elde edilen modelin kalitesini doğrudan etkileyen önemli bir hiperparametredir. Uygun bir yığın boyutu seçmek, modelin hem eğitim verimliliğini hem de sonuçlarının doğruluğunu optimize etmek için kritik öneme sahiptir (Ruder, 2016; Lin, 2022; Usmani ve diğ., 2023).

Maksimum devir sayısı (epok): Maksimum devir sayısı, derin öğrenme modellerinin eğitim sürecinde eğitim veri setinin model tarafından kaç kez tamamen işlendiğini belirten bir parametredir. Maksimum devir sayısı, modelin veri seti üzerindeki öğrenme sürecinin derinliğini ve süresini doğrudan etkiler. Yeterli maksimum devir sayısı, modelin eğitim verilerinden gerekli tüm özellikleri öğrenmesine ve genelleme yapabilmesine olanak tanırken, yüksek maksimum devir sayısı kullanmak modelin eğitim verilerine aşırı uyum sağlamasına (overfitting) yol açabilir, bu da modelin yeni verilere karşı genelleme yapma kabiliyetini azaltır (Goodfellow ve diğ., 2016).

Maksimum devir sayısını belirlerken, modelin eğitim sürecinde erken durdurma (early stopping) gibi teknikler kullanılabilir. Erken durdurma, doğrulama kaybı belirli bir sayıda maksimum devir sayısı boyunca iyileşmediğinde eğitimi otomatik olarak durdurur. Bu, hem aşırı uyumu önler hem de gereksiz hesaplama kaynaklarının israfını engeller (Prechelt, 1998).

Öğrenme Oranı (Learning Rate- LR): Öğrenme oranı, derin öğrenme ve ESA modellerinde, model parametrelerinin her iterasyonda ne kadar güncelleneceğini belirleyen kritik bir hiperparametredir. Bu değer, optimizasyon sürecinde kayıp fonksiyonunun gradyanına göre parametrelerin ne kadar "hızlı" veya "yavaş" ayarlanacağını kontrol eder (Goodfellow ve diğ., 2016).

Öğrenme oranı, modelin eğitim sürecindeki performansını büyük ölçüde etkiler. Uygun bir öğrenme oranı seçimi, modelin veri setindeki özellikleri etkili bir şekilde öğrenmesini ve optimizasyon sürecinin verimli bir şekilde ilerlemesini sağlar. Çok yüksek bir öğrenme oranı, modelin optimum çözüme yakınsamasını engelleyebilir ve aşırı dalgalanmalara neden olabilir. Öte yandan, çok düşük bir öğrenme oranı, modelin çok yavaş öğrenmesine ve hatta eğitim sürecinin pratik bir zaman diliminde

tamamlanamamasına yol açabilir. Bu nedenle uygun bir öğrenme oranı seçimi, modelin verimli bir şekilde eğitilmesi ve yüksek performansla çalışması için oldukça önemlidir (Bottou, 2010).

Öğrenme Oranı Programı (Learning Rate Schedule-LRS): Öğrenme oranı programı, derin öğrenme modellerinin eğitimi sırasında öğrenme oranının zaman içinde nasıl ayarlanacağını belirleyen bir stratejidir. Bu program, modelin eğitim sürecinin farklı aşamalarında öğrenme hızını değiştirerek, hem hızlı yakınsamayı teşvik eder hem de modelin aşırı uyum riskini azaltır (Goodfellow ve diğ., 2016).

Öğrenme oranı, modelin eğitim verilerinden öğrenme yeteneğini ve genelleme performansını büyük ölçüde etkileyen bir faktördür. Sabit bir öğrenme oranı kullanmak bazı durumlarda yetersiz olabilir, çünkü modelin eğitim sürecinin farklı aşamaları farklı öğrenme ihtiyaçlarına sahip olabilir. Öğrenme oranı programı, bu ihtiyaçları dinamik bir şekilde karşılayarak modelin daha etkili bir şekilde eğitilmesini sağlar (Smith, 2017). Yaygın bazı öğrenme oranı programları şunlardır:

Zaman Tabanlı Azalma: Bu yöntem, belirli bir oranda öğrenme oranını sürekli azaltır. Genellikle, her epok sonunda öğrenme oranı, başlangıç değerinin bir yüzdesi kadar düşürülür. Bu sürekli düşüş, uzun eğitim süreçlerinde faydalı olabilir.

Adım Azalması (Step Decay): Bu tez çalışmasında da kullanılan bu yöntem, belirli aralıklarla öğrenme oranını kademeli olarak azaltır. Bu yöntem, modelin başlangıçta hızlı bir şekilde öğrenmesine olanak tanırken, eğitim sürecinin ilerleyen aşamalarında daha küçük adımlarla ince ayar yapılmasına yardımcı olur.

Döngüsel Öğrenme Oranı: Öğrenme oranı minimum ve maksimum değerler arasında döngüsel olarak değişir. Bu yaklaşım, yerel minimumlardan kaçınmayı ve genel minimuma ulaşmayı kolaylaştırır, çeşitli eğitim aşamalarında modelin farklı bölgeleri keşfetmesine olanak tanır.

Eğimli İyileştirme (Cosine Annealing): Öğrenme oranı, bir kosinüs fonksiyonunun eğrisi takip edilerek yavaşça azaltılır, bu da öğrenme sürecinin sonlarında daha ince ayar yapılmasını sağlar ve genellikle uzun süreli eğitimlerde tercih edilir.

Momentum katsayısı: Momentum katsayısı, derin öğrenme ve ESA modellerinde, optimizasyon sürecini hızlandırmak ve daha etkin bir yakınsama sağlamak için kullanılan bir hiperparametredir. Bu katsayı, öğrenme algoritmasının geçmiş gradyan güncellemelerini hesaba katarak mevcut güncelleme adımını etkiler (Sutskever ve diğ.,

2013). Denklem 3.38 ve Denklem 3.39 momentumlu bir gradyan inişi yöntemini göstermektedir.

$$v_t = \gamma v_t - \eta \nabla_\theta J(\theta) \quad (3.38)$$

$$\theta = \theta - v_t \quad (3.39)$$

Burada γ momentum terimidir (genellikle 0.9 gibi değerler alır), v_t önceki gradyan güncellemelerinin birikimini temsil eden momentum vektörüdür. Momentum, özellikle derin öğrenme modellerinde, algoritmanın yerel minimumlarda takılmasını önlemeye yardımcı olur ve daha hızlı yakınsamayı teşvik eder. Bu sayede, algoritma daha istikrarlı ve hızlı bir şekilde optimuma yaklaşabilir.

3.7. Temel Bileşen Analizi (TBA)

Temel Bileşen Analizi (TBA), yüksek boyutlu veri setlerinin boyutunu azaltırken bilgi kaybını en aza indiren güçlü bir istatistiksel tekniktir (Jolliffe, 2002) (Jolliffe & Cadima, 2016). TBA'nın temel çalışma prensibi, verilerin varyansını en üst düzeye çıkaran temel bileşenleri bulmaktır (Jolliffe ve Cadima, 2016). TBA, verilerin temel bileşenlerini çıkararak boyutunu azaltır ve bu sayede veri setindeki gürültüyü de azaltır, veri analizini daha verimli hale getirir (Abdi ve Williams, 2010).

TBA'nın uygulanması, veri matrisi üzerinde ortalama merkezleme ile başlar. Bu adım, her bir değişkenin ortalamasının çıkarılmasıyla gerçekleştirilir, böylece veri setinin merkezi orijine kaydırılır. Daha sonra, bu merkezi verilere özdeğer ayrışımı uygulanır. Bu matematiksel işlemler, veri setindeki ana bileşenleri (özvektörler) ve bu bileşenlere karşılık gelen varyans miktarlarını (özdeğerler) belirler. Özdeğerlerin büyükten küçüğe sıralanması, her bir temel bileşenin veri kümesindeki toplam varyansı ne kadar açıkladığını gösterir (Abdi ve Williams, 2010).

TBA, boyut indirgeme işlemi sırasında bilgi kaybını en aza indirir. Bu, veri kümesindeki en büyük varyansı açıklayan bileşenlerin seçilmesiyle gerçekleştirilir. Örneğin, veri kümesinde toplam varyansın %90'ını açıklayan ilk birkaç bileşen seçilerek, orijinal veri setinin boyutu önemli ölçüde azaltılabilir. Bu işlem, veri analizini daha yönetilebilir hale getirir ve aynı zamanda veri kümesinin temel yapısını korur (Jolliffe & Cadima, 2016).

3.8. Özellik Füzyonu

Özellik füzyonu, İDT sistemlerindeki performansı iyileştirmek için kritik bir tekniktir. Bu süreç, çeşitli derin öğrenme modelleri kullanılarak farklı kaynaklardan elde edilen özelliklerin tek bir temsilde bütünleştirilmesini içerir. Bu entegrasyon, modelin daha kapsamlı bilgi setleri üzerinden öğrenmesini ve daha doğru tahminler yapabilmesini sağlar (Baltrušaitis ve diğ., 2018).

Bu tez çalışması kapsamında özellikle vücut, eller ve yüz gibi farklı kaynaklardan gelen veriler için ayrı ayrı derin öğrenme modelleri eğitilmiştir. Her bir modelden elde edilen özellik vektörleri, verilerin daha geniş bir perspektiften değerlendirilmesini sağlamak amacıyla bir araya getirilmiştir. Bu özellikler, her bir özellik kümesinin bütünlüğünü ve ayırt ediciliğini korumadaki basitliği ve etkinliği nedeniyle seçilen bir yöntem olan basit birleştirme işlemi yoluyla gerçekleştirilmiştir (Ramachandram ve Taylor, 2017). Bu yöntem kullanılarak özellik vektörleri tek bir uzun vektör oluşturmak üzere uç uca eklenir (Denklem 3.40).

$$Z_{\text{birleştirilmiş}} = [Z_{\text{vücut}} \,||\, Z_{\text{el}} \,||\, Z_{\text{yüz}}] \tag{3.40}$$

Burada || operatörü birleştirme işlemini göstermektedir. Bu $Z_{\text{birleştirilmiş}}$ vektörü, her bir özellik kümesinin ayrı ayrı güçlü yönlerini bir araya getirerek farklı jest verilerinin hem kapsamlı hem de ayrıntılı bütünsel bir temsilini sağlar. Bu özellik füzyonu stratejisinin temel avantajı, çeşitli veri kaynaklarından elde edilen özelliklerin entegrasyonu aracılığıyla sınıflandırıcının mevcut verileri anlama ve yorumlama kapasitesini daha etkin bir şekilde kullanmasına olanak tanımasıdır. İşaret dili hareketlerinin çeşitli yönlerinden elde edilen özelliklerin bir araya getirilmesiyle elde edilen birleştirilmiş özellik vektörü, verilerin daha zengin ve daha incelikli bir temsilini sunar. Böylece, sınıflandırıcı, işaret dili hareketlerini ayırt etmek için daha geniş bir bilgi yelpazesinden yararlanabilir. Bu geniş bilgi yelpazesi, sınıflandırıcının daha doğru ve güvenilir performans sergilemesini sağlar.

3.9. Destek Vektör Makinesi (DVM)

Destek Vektör Makinesi (DVM), özellik uzayında sınıfları ayırmak için en uygun hiperdüzlemi bulmayı amaçlayan bir denetimli öğrenme algoritmasıdır. Bu hiper düzlem, farklı sınıflara ait veri noktalarını ayıran bir sınır olarak işlev görür ve modelin eğitim sürecinde, yüksek boyutlu özellik uzayında bu hiper düzlemleri oluşturarak eğitim veri setini etkili bir şekilde farklı sınıflara ayırması esastır (Cortes ve Vapnik, 1995).

Bu sürecin somut bir örneği olarak, verilen n adet eğitim örneği $(x_i, y_i)_{i=1}^{n}$ ile ele alındığında, her bir x_i özelliği d-boyutlu bir vektör olarak temsil edilir ve her örneğe bir sınıf etiketi (y_i) atanır. Bu d-boyutlu özellik vektörleri ve ilişkili sınıf etiketleri kullanılarak, DVM modeli özellik uzayında sınıfları ayıran bir hiper düzlem oluşturur. Bu hiper düzlem matematiksel olarak Denklem 3.41'deki gibi ifade edilir (Cortes ve Vapnik, 1995):

$$w \cdot x + b = 0 \tag{3.41}$$

Bu denklemde, w hiperdüzlem normal vektörüdür ve hiperdüzleme dik yönü belirtir. x bir özellik vektörünü temsil ederken, b bias terimi olarak hiperdüzlemin konumunu ayarlar. Verilen bir hiperdüzlem (w, b), veri noktalarını ayırır ve eğitim verilerini doğru şekilde sınıflandırmak için kullanılan bir fonksiyon sağlar (Cortes ve Vapnik, 1995) (Denklem 3.42):

$$f(x) = \text{sign}(w \cdot x + b) \tag{3.42}$$

Bu fonksiyon, x özellik vektörünün w ile noktasal çarpımı ve b bias teriminin toplamının işaretine göre sınıflandırma yapar. Özellik seti doğrusal olarak ayrılabilir olmadığında, Radyal Taban Fonksiyonu (Radial Basis Function - RBF) çekirdeği gibi bir çekirdek fonksiyonu, yani Denklem 3.43 kullanılarak veri, daha yüksek boyutlu bir uzaya taşınır (Amari ve Wu, 1999):

$$K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2} \tag{3.43}$$

Bu denklemde, $K(x_i, x_j)$ iki özellik vektörü x_i ve x_j arasındaki benzerliği ifade eden RBF çekirdek fonksiyonudur. γ, çekirdeğin genişliğini kontrol eden bir parametredir ve $\|x_i - x_j\|^2$, iki vektör arasındaki Öklid mesafesinin karesidir.

DVM'nin optimizasyon problemi, hedef fonksiyonun (Denklem 3.44) minimizasyonu ile veri noktalarını ayıran ve aynı zamanda marjı maksimize eden hiper düzlemi bulmayı amaçlar. Bu işlem, Denklem 3.45'te gösterilen kısıtlamalar altında gerçekleştirilir (Vapnik ve Vapnik, 1998):

$$\text{Minimizasyon: } \frac{1}{2}\|w\|^2 \tag{3.44}$$

$$y_i(x_i \cdot w + b) \geq 1, \quad \forall i \tag{3.45}$$

Denklem 3.44, hedef fonksiyon olarak tanımlanır ve DVM modelinin optimizasyon probleminde minimizasyonu gereklidir. Bu denklem, hiper düzlemin normal vektörü w'nin normunun karesinin yarısını ifade eder. Bu minimizasyon işlemi, sınıflar arası marjı maksimize ederek modelin genelleme kabiliyetini artırır. Pratikte, w'nin normunu küçültmek, destek vektörleri ile hiperdüzlem arasındaki marjı artırır, çünkü marjın büyüklüğü $\frac{1}{\|w\|}$ ile ters orantılıdır. Denklem 3.45, verilen kısıtlamaları temsil eder ve her bir eğitim örneğinin, ilişkili sınıf etiketiyle doğru bir şekilde sınıflandırılmasını sağlamak için hiper düzlem parametreleriyle uyumlu olmasını zorunlu kılar. Bu, hiper düzlem ile destek vektörleri arasında en az bir birim mesafenin korunmasını garanti eder (Vapnik ve Vapnik, 1998).

Optimizasyon sürecinde, Lagrange çarpanları kullanılarak hiper düzlem parametreleri, yani w ve b belirlenir, modelin sınıflandırma performansı en üst düzeye çıkarılır. Bu süreç sonunda, modelin karar sınırını belirleyen destek vektörleri elde edilir. Sonuç olarak, RBF çekirdeği ile geliştirilmiş DVM, özellikle doğrusal olmayan sınıflandırma problemlerinde güçlü ve esnek bir çözüm sunar (Murphy, 2012).

3.10. Değerlendirme Metrikleri

Yapay zeka ve makine öğrenimi sistemlerinin performansını doğru bir şekilde ölçmek, bu sistemlerin etkinliğini anlamak ve geliştirmek için kritik öneme sahiptir

(Sokolova ve Lapalme, 2009; Powers, 2020). Bu tez çalışması kapsamında da önerilen yöntemlerin etkinliğini ölçmek için çeşitli değerlendirme ölçütleri kullanmaktadır. Bu ölçütler aşağıda verilmiştir.

Gerçek pozitifler (True positives - TP): Modelin bir pozitif durumu doğru şekilde tahmin ettiği durumları ifade eder.

Yanlış pozitifler (False positives - FP): Modelin aslında pozitif olmayan bir durumu yanlışlıkla pozitif olarak tahmin ettiği durumları ifade eder.

Yanlış negatifler (False negatives - FN): Modelin aslında pozitif olan bir durumu yanlışlıkla negatif olarak tahmin ettiği durumları ifade eder.

Gerçek negatifler (True negatives - TN): Modelin bir negatif durumu doğru şekilde tahmin ettiği durumları ifade eder.

Modelin farklı kategorilerdeki genel etkinliği doğruluk (accuracy) ile ölçülür (Denklem 3.46). Bu, doğru tahmin edilen örneklerin sayısının değerlendirilen toplam örnek sayısına oranı olarak hesaplanır. Daha incelikli bir değerlendirme için kesinlik (precision) ve duyarlılık (recall) ölçütleri kullanılır. Kesinlik (Denklem 3.47), belirli bir sınıf için pozitif olarak tanımlanan tüm örnekler arasından doğru olarak tanımlanan pozitif örneklerin oranıdır. Duyarlılık (Denklem 3.48) modelin kaç tane gerçek pozitif örneği doğru tanımladığını ölçer. F1 skoru (Denklem 3.49), kesinlik ve duyarlılık ölçütlerinin harmonik ortalamasını temsil eden ve bu iki ölçütün dengeli bir görünümünü sunan birleşik bir metriktir (Bishop ve Nasrabadi, 2006). Bu değerlendirme metriklerine ait denklemler aşağıda verilmiştir:

$$\text{Accuracy} = \frac{\text{Toplam doğru tahminler}}{\text{Toplam tahminler}} \tag{3.46}$$

$$\text{Kesinlik} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3.47}$$

$$\text{Duyarlılık} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.48}$$

$$\text{F1 skoru} = 2 \times \frac{\text{Kesinlik} \times \text{Duyarlılık}}{\text{Kesinlik} + \text{Duyarlılık}} \tag{3.49}$$

Kapsamlı bir değerlendirme açısından, her sınıf için kesinlik, duyarlılık ve F1 skorunun makro ortalaması hesaplanmıştır. Bu yaklaşım, her kategorinin modelin genel etkinliğine katkısına eşit ağırlık vererek tüm sınıflar arasında dengeli bir değerlendirme yapılmasını sağlar.

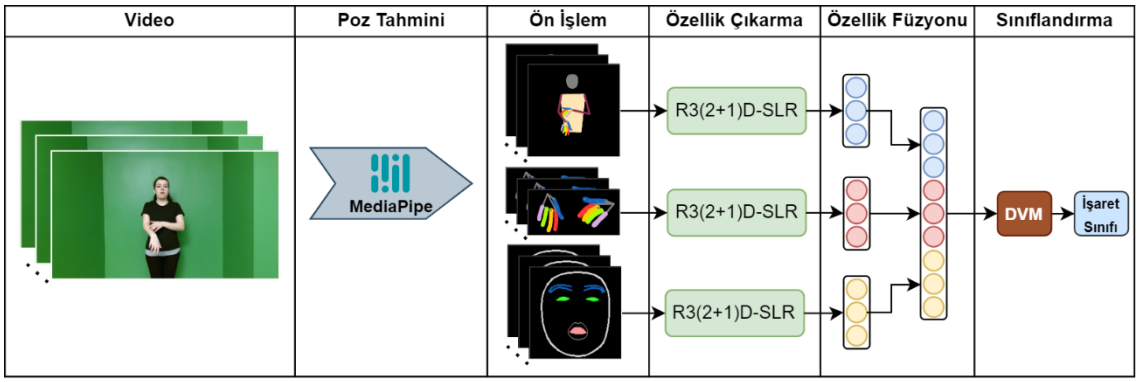
4. ÖNERİLEN YÖNTEMLER VE DENEYSEL ÇALIŞMALAR

Bu bölümde, tez kapsamında yürütülen çalışmalar detaylandırılmış ve önerilen İDT sistemlerinin deneysel sonuçları sunulmuştur. Çalışmalar, Intel Core i5-8400 işlemci, 16 GB RAM ve 12 GB NVIDIA GeForce GTX 1080 Ti GPU barındıran bir bilgisayar sistemi üzerinde gerçekleştirilmiştir. Bu süreçte, PyTorch, PyTorch-Lightning ve scikit-learn gibi gelişmiş kütüphanelerden yararlanılmıştır. Deneysel çalışmalar, öncelikle BosphorusSign22k-genel veri kümesi üzerinde yapılmış, daha sonra BosphorusSign22k, LSA64 ve GSL gibi farklı veri kümeleri ile testler genişletilmiştir.

4.1. Çalışma-1: R3(2+1)D-SLR Ağı ve Özellik Füzyonu ile İşaret Dili Tanıma

Bu bölümde sunulan çalışma, “Enhancing Signer-Independent Recognition of Isolated Sign Language through Advanced Deep Learning Techniques and Feature Fusion” (Akdag ve Baykan, 2024a) isimli makalede sunulmuştur. Bu çalışmada, hem kalıntı (residual) üç boyutlu (R3D) hem de zamansal olarak ayrılmış (R(2+1)D) evrişim bloklarının güçlü yanlarını birleştiren yenilikçi bir derin öğrenme modeli kullanılarak izole İDT görevi için yeni bir yaklaşım önerilmiştir. Bu birleştirilmiş evrişim blokları kullanılarak tasarlanan R3(2+1)D-SLR ağ mimarisi, doğru işaret tanıma için kritik olan karmaşık uzamsal ve zamansal özellikleri yakalama konusunda üstün bir yetenek sergilemiştir. Bu mimariye dayalı olarak geliştirilen İDT sistemi, işaretçinin vücudu, elleri ve yüzünden, R3(2+1)D-SLR modeli kullanılarak elde edilen özellikleri birleştirir ve sınıflandırma için DVM’ye iletir; böylece farklı vücut bölgelerinden elde edilen özellik bilgisi ile işaretlerin daha doğru bir şekilde tanımlanmasını sağlar. Bununla birlikte, önerilen sistemde ham RGB görüntü verileri yerine poz verilerinin kullanılmasının, çeşitli arka planlarda doğruluk ve sağlamlıkta önemli iyileştirmeler sağladığı gösterilmiştir.

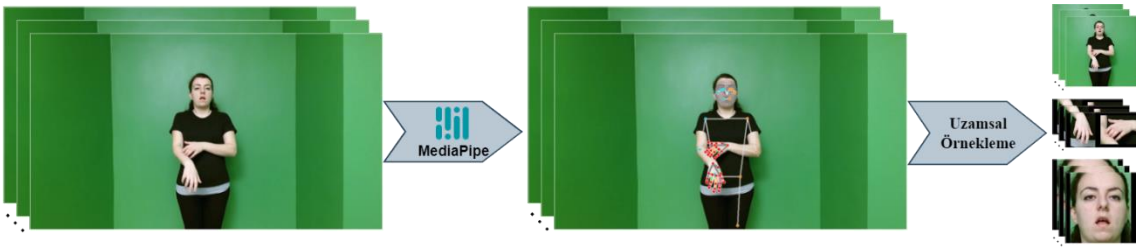
Önerilen nihai sistemin diyagramı Şekil 4.1'de gösterilmektedir. Bu bölümün devamında, önerilen sistemin tasarımında kullanılan veri ön işleme adımları, önerilen derin öğrenme mimarisi, kullanılan füzyon tekniği, sınıflandırma yöntemi ve elde edilen deneysel sonuçlar anlatılacaktır.



Şekil 4.1. Çalışma-1 için önerilen sistemin diyagramı (Akdag ve Baykan, 2024a)

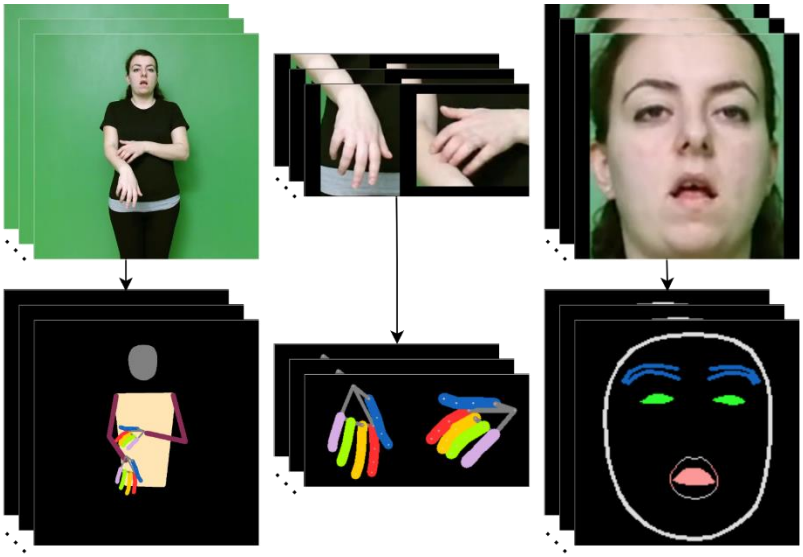
4.1.1. Veri ön işleme

Bu çalışmada, işaret dili videolarının içerdiği uzamsal ve zamansal özellikleri çıkarmak için MediaPipe kütüphanesinden elde edilen poz verilerinden yararlanılmıştır. İlk olarak, uzamsal bilgi elde etmek amacıyla MediaPipe Holistic yöntemi veri kümelerindeki her videonun tüm karelerine uygulanmıştır. İşaret noktalarının x ve y koordinatları kullanılarak görüntüler vücut, eller ve yüz bölgelerine ayrılmıştır. Bu adımdan sonra, derin öğrenme modelinin eğitim sürecinde kullanılmak üzere görüntüler yeniden boyutlandırılmıştır. Vücut ve yüz görüntüleri 112×112 çözünürlüğe, el görüntüleri ise 56×112 çözünürlüğe ayarlanmıştır. Bu işlemle, modelin her bir bileşeni ayrıntılı olarak analiz edebilmesi için odaklanmış veriler oluşturulmuştur. Elde edilen bu veriler Şekil 4.2'de gösterilmiştir.



Şekil 4.2. Uzamsal örnekleme (Akdag ve Baykan, 2024a)

Ayrıca, sistemin yüksek arka plan çeşitliliği ve karmaşıklığı koşullarında bile işaretleri yüksek doğruluk ve sağlamlıkla tanıması amaçlanmıştır. Bu nedenle, sistemde kullanılan geleneksel RGB video verisinin, MediaPipe Holistic ile elde edilen poz verisi tabanlı görüntü dizileriyle güncellenmesi gerçekleştirilmiştir. Bu kapsamda oluşturulan RGB video verilerine karşılık gelen poz görüntüleri Şekil 4.3'te gösterilmiştir.



Şekil 4.3. Poz görüntülerinin oluşturulması (Akdag ve Baykan, 2024a)

Ön işleme sırasında, bacaklar, işaret yorumlamasını etkilemedikleri için vücut pozu görüntülerinden çıkarılmıştır. Ayrıca, yüz de ayrı olarak değerlendirileceği için vücut pozu görüntüsünden çıkarılmıştır; bunun yerine yüz bölgesi, baş duruş pozisyonunu temsil etmek amacıyla düz bir renkle boyanmıştır.

El pozu görüntülerinde, el hareketleri ve işaret dilinin ince nüanslarını daha ayrıntılı analiz edebilmek için her parmağa farklı bir renk atanmıştır. Ayrıca, işaret dilinin nüanslarını yakalamada renkli ve renksiz el pozu görüntülerinin etkinliği karşılaştırılmıştır. Bu yaklaşım, renk kodlamasının modelin işaret algılama kapasitesi üzerindeki etkisini ayrıntılı olarak incelemeyi sağlamıştır.

Yüz pozu görüntülerinde, yüz ifadelerini ve duygusal ifadeleri daha net tespit etmek için kaşlar, gözler ve ağız farklı renklere boyanmıştır. Özellikle gözler ve ağzın içi, açma ve kapama hareketlerini vurgulamak için tamamen boyanmıştır. Bu ayrıntılı yaklaşımın, önerilen İDT sisteminin arka plan gürültüsüne karşı sağlamlığını artırması ve işaret dili bileşenlerinin kapsamlı bir analizi yoluyla doğruluk oranlarını iyileştirmesi beklenmektedir.

Derin öğrenme modelleri, eğitim sırasında sabit boyutlarda veri beklerler. Bu nedenle, elde edilen uzamsal örneklemelere ek olarak, işaret dili videolarının derin öğrenme modelleri tarafından etkili bir şekilde işlenebilmesi için zamansal örnekleme de gerçekleştirilmiştir. Bu süreçte, MediaPipe Holistic kullanılarak elde edilen her karenin işaret noktalarının x ve y koordinatları kullanılarak ardışık kareler arasındaki Manhattan

mesafesi hesaplanmıştır ve bu işlem, videodaki tüm kareler için tekrarlanmıştır. Formülasyon Denklem 4.1'de verilmiştir:

$$D^k = \sum_{i=1}^{N} (\left|x_i^k - x_i^{k+1}\right| + \left|y_i^k - y_i^{k+1}\right|) \quad (4.1)$$

Burada D^k, ardışık k ve $k + 1$ kareleri arasındaki toplam Manhattan uzaklığıdır ve N, her karedeki işaret noktalarının sayısını temsil eder. Bu yöntem, her bir özellik (vücut, eller, yüz) için kendi işaret noktaları kullanılarak hesaplanmıştır. Her bir video için, video boyunca elde edilen Manhattan mesafeleri göz önünde bulundurularak en yüksek varyasyonu gösteren 32 kare işareti temsil etmek için seçilmiştir. Eğer videoda 32'den az kare varsa, eksik kare sayısını tamamlamak için ilk ve son kareler eşit sayıda çoğaltılmıştır. Bu yaklaşım, işaret dilini anlamak için en yüksek hareket yoğunluğuna sahip kareleri seçmeye olanak tanır. Böylece, işaretleri en iyi temsil eden kareler, derin öğrenme modelinin eğitiminde kullanılmak üzere belirlenmiş olur.

4.1.1.1. Test veri kümesine arka plan görüntüleri ekleme

Gerçek dünya verilerinin farklı arka planları kapsadığı göz önüne alındığında, işaret dili videolarını tek tip bir arka plana karşı analiz etmek gerçekçi bir değerlendirme sağlamayabilir. Bu bağlamda, gerçek dünya koşullarını daha iyi taklit edebilmek için, veri kümelerindeki test videolarına iç ve dış mekanlardan seçilen 40 farklı arka plan görüntüsü özel olarak eklenmiştir. Bu yaklaşım, yöntemin tanıma performansını farklı arka planlarda değerlendirmek için test aşamasında kullanılmıştır. BosphorusSign22k-genel veri kümesinden alınmış, arka planı değiştirilmiş videolardan örnek kareler Şekil 4.4'te sunulmuştur.

Şekil 4.4. Arka plan eklenmiş örnek video kareleri

Bu prosedür sadece test seti için uygulanmıştır; eğitim setindeki videolarda herhangi bir arka plan değişikliği yapılmamıştır. Bu yaklaşımın nedeni, modelin eğitim sürecinde sabit bir arka plan üzerinde öğrenmesine izin vermek ve ardından test sürecinde modelin farklı arka planlar üzerindeki işaretleri ne kadar etkili bir şekilde tanıyabildiğini objektif olarak test etmektir.

4.1.2. Önerilen ESA modeli

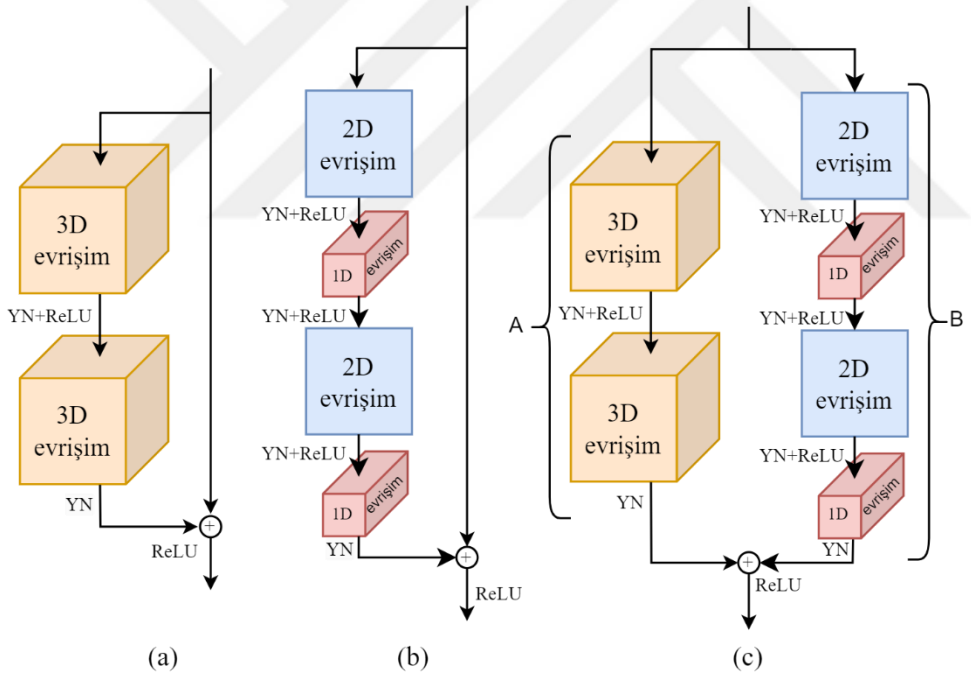
4.1.2.1. R3(2+1)D evrişim bloğu

Bir evrişim bloğu, tipik olarak evrişim katmanı, yığın normalizasyonu, doğrusal olmayan bir aktivasyon fonksiyonu ve havuzlama katmanı dahil olmak üzere birden fazla katmandan oluşan ESA'nın temel bir bileşenidir. Bir ESA mimarisi, daha derin bir ağ oluşturmak için istiflenmiş çoklu blok yapılardan oluşur. Her blok, giriş verilerinden daha karmaşık özellikler çıkarır ve son bloğun nihai çıktısı sınıflandırma veya diğer görevler için kullanılır.

R3D bloğu iki 3D evrişim katmanından (Şekil 4.5.a), R(2+1)D bloğu ise iki (2+1)D evrişim katmanından (Şekil 4.5.b) oluşmaktadır. Her katmandan sonra, ağ kararlılığını ve düzenliliğini artırmak için 3D Yığın Normalizasyonu (YN) uygulanır, ardından ise ReLU aktivasyon fonksiyonu uygulanır. Evrişim katmanlarından sonra, giriş verileri kalıntı bağlantı olarak çıkış değeri ile toplanır. i'inci kalıntı bloğunun çıktısı Denklem 4.2 ile ifade edilebilir. Elde edilen bu nihai değer yine bir ReLU fonksiyonundan geçirilerek sonraki katmanlara iletilir.

$$x_i = x_{i-1} + F(x_{i-1}; \theta_i) \tag{4.2}$$

Burada x_{i-1}, i katmanının giriş verisidir; $F(;\theta_i)$, θ_i ağırlıkları ile hesaplanan iki evrişim katmanını, 3D Yığın Normalizasyonunu ve ReLU fonksiyonlarının uygulanmasını ifade eder.



Şekil 4.5. Evrişimsel blok türleri: (a) R3D evrişim bloğu; (b) R(2+1)D evrişim bloğu; (c) R3(2+1)D evrişim bloğu (Akdag ve Baykan, 2024a)

Bu çalışmada, R3D ve R(2+1)D evrişim bloklarını birleştirerek oluşturulan yeni bir blok yapısı olan R3(2+1)D bloğu tanıtılmıştır. Bu yapı Şekil 4.5.c'de gösterilmektedir; burada giriş verileri hem R3D hem de R(2+1)D blokları aracılığıyla işlenmekte ve sonuçlar blokların çıkışlarında toplanmaktadır. Bu yaklaşımda, R3D evrişim bloğu A ve

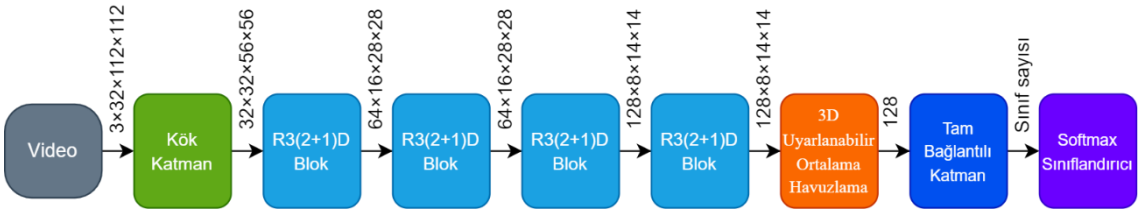
R(2+1)D evrişim bloğu B ile gösterilmek üzere, i'inci kalıntı blok için çıktı Denklem 4.3'teki gibi formüle edilebilir:

$$x_i = \mathrm{ReLU}(A(x_{i-1}) + B(x_{i-1})) \quad (4.3)$$

Önerilen R3(2+1)D-SLR ağındaki blok yapılarının bu füzyonu, video veri işlemenin iyileştirilmesine yönelik yenilikçi bir adımı temsil etmektedir. Bu mimari, R3D bloklarının derinliğini ve karmaşıklığını R(2+1)D bloklarının sunduğu rafine uzamsal-zamansal ayrıştırma ile birleştirerek, 3D verilerden dengeli bir özellik çıkarımı elde etmeyi amaçlamaktadır. Bir sonraki bölümde, R3(2+1)D blok yapısını kullanarak İDT görevi için tasarlanan ESA mimarisi tanıtılacaktır.

4.1.2.2. Önerilen R3(2+1)D-SLR ağı

R3(2+1)D evrişim bloğunu kullanan ve İDT görevi için tasarlanan R3(2+1)D-SLR ağ mimarisi Çizelge 4.1 ve Şekil 4.6'da sunulmaktadır. Mimari, bir kök katman ve ardından her biri bir R3(2+1)D bloğu içeren dört katmandan oluşur. R3(2+1)D bloğuna benzer yapıya sahip kök katman, 3D ve (2+1)D evrişim katmanlarının birleşiminden oluşur. Ancak R3(2+1)D bloğundan farklı olarak, kök katmanda bir 3D ve bir (2+1)D evrişim katmanı bulunur. Kök katmandaki evrişim blokları 7×7 boyutunda 2 adımlı 32 evrişim çekirdeğine sahiptir ve 3×32×112×112 boyutlarındaki (burada 3, RGB kanallarının sayısını; 32, toplam video karesi sayısını; 112×112 ise görüntünün genişliğini ve yüksekliğini temsil eder) gelen verileri işleyerek, 32×32×56×56 boyutlarında bir çıktı üretir. Bu veriler daha sonra bir sonraki katmana iletilir. Kök katmanını takip eden diğer katmanlar sırasıyla 64, 64, 128 ve 128 evrişim çekirdeği içerir. Birinci ve üçüncü katmanlarda adım boyu 2'ye ayarlanarak alt örnekleme gerçekleştirilir.



Şekil 4.6. Önerilen R3(2+1)D-SLR ağ yapısı (Akdag ve Baykan, 2024a)

Çizelge 4.1. Önerilen R3(2+1)D-SLR ağ mimarisi (Akdag ve Baykan, 2024a)

Katman	Giriş Boyutu	Çıkış Boyutu	R3(2+1)D-SLR	
Kök katmanı	$3 \times 32 \times 112 \times 112$	$32 \times 32 \times 56 \times 56$	$3 \times \left[\begin{pmatrix} 1 \times 7 \times 7 \text{ stride } 1,2,2 \\ 3 \times 1 \times 1 \text{ stride } 1,1,1 \end{pmatrix}\right], 32$	$3 \times [(3 \times 7 \times 7 \text{ stride } 1,2,2)], 32$
Katman 1	$32 \times 32 \times 56 \times 56$	$64 \times 16 \times 28 \times 28$	$32 \times \left[\begin{pmatrix} 1 \times 3 \times 3 \text{ stride } 1,2,2 \\ 3 \times 1 \times 1 \text{ stride } 2,1,1 \end{pmatrix}\right], 64$ $64 \times \left[\begin{pmatrix} 1 \times 3 \times 3 \text{ stride } 1,1,1 \\ 3 \times 1 \times 1 \text{ stride } 1,1,1 \end{pmatrix}\right], 64$	$32 \times [(3 \times 3 \times 3 \text{ stride } 2,2,2)], 64$ $64 \times [(3 \times 3 \times 3 \text{ stride } 1,1,1)], 64$
Katman 2	$64 \times 16 \times 28 \times 28$	$64 \times 16 \times 28 \times 28$	$64 \times \left[\begin{pmatrix} 1 \times 3 \times 3 \text{ stride } 1,1,1 \\ 3 \times 1 \times 1 \text{ stride } 1,1,1 \end{pmatrix}\right], 64$ $64 \times \left[\begin{pmatrix} 1 \times 3 \times 3 \text{ stride } 1,1,1 \\ 3 \times 1 \times 1 \text{ stride } 1,1,1 \end{pmatrix}\right], 64$	$64 \times [(3 \times 3 \times 3 \text{ stride } 1,1,1)], 64$ $64 \times [(3 \times 3 \times 3 \text{ stride } 1,1,1)], 64$
Katman 3	$64 \times 16 \times 28 \times 28$	$128 \times 8 \times 14 \times 14$	$64 \times \left[\begin{pmatrix} 1 \times 3 \times 3 \text{ stride } 1,2,2 \\ 3 \times 1 \times 1 \text{ stride } 2,1,1 \end{pmatrix}\right], 128$ $128 \times \left[\begin{pmatrix} 1 \times 3 \times 3 \text{ stride } 1,1,1 \\ 3 \times 1 \times 1 \text{ stride } 1,1,1 \end{pmatrix}\right], 128$	$64 \times [(3 \times 3 \times 3 \text{ stride } 2,2,2)], 128$ $128 \times [(3 \times 3 \times 3 \text{ stride } 1,1,1)], 128$
Katman 4	$128 \times 8 \times 14 \times 14$	$128 \times 8 \times 14 \times 14$	$128 \times \left[\begin{pmatrix} 1 \times 3 \times 3 \text{ stride } 1,1,1 \\ 3 \times 1 \times 1 \text{ stride } 1,1,1 \end{pmatrix}\right], 128$ $128 \times \left[\begin{pmatrix} 1 \times 3 \times 3 \text{ stride } 1,1,1 \\ 3 \times 1 \times 1 \text{ stride } 1,1,1 \end{pmatrix}\right], 128$	$128 \times [(3 \times 3 \times 3 \text{ stride } 1,1,1)], 128$ $128 \times [(3 \times 3 \times 3 \text{ stride } 1,1,1)], 128$
OrtHavuz	$128 \times 8 \times 14 \times 14$	128	3D Uyarlanabilir Ortalama Havuzlama	
TB	128	Sınıfı Sayısı	Tam Bağlantılı Katman	
Softmax	Sınıfı Sayısı	Sınıfı Sayısı	Softmax sınıflandırıcı	

Veriler, R3(2+1)D blokları aracılığıyla işlendikten sonra, son evrişim katmanını takip eden 3D Uyarlanabilir Ortalama Havuzlama işlemi ile 128 boyutlu bir vektöre indirgenir. Bu işlem, verileri daha yönetilebilir bir formata indirgerken temel özellikleri korumada etkilidir. Elde edilen 128 boyutlu vektör, sınıflandırma katmanına bir köprü görevi gören Tam Bağlantılı Katmana beslenir. Sınıflandırma işlemini tamamlamak için, Tam Bağlantılı Katmanın çıkışına bir Softmax işlevi uygulanır ve 128 boyutlu vektör, hedeflenen sınıf sayısına karşılık gelen olasılık puanlarına dönüştürülür. En yüksek olasılık puanına sahip sınıf, modelin nihai tahmini olarak seçilir ve İDT görevi için net ve anlaşılır bir sonuç sağlar. Bu mimari genel olarak, işaret dili video verilerinden uzamsal ve zamansal bilgilerin dengeli bir şekilde çıkarılmasını sağlayarak derinlik ve karmaşıklık açısından zengin bir özellik kümesi sunar.

4.1.3. Deneysel sonuçlar

Bu çalışmadaki, tüm derin öğrenme modelleri SGD optimizasyon algoritması kullanılarak, 0,9 momentum katsayısı ile 35 epok boyunca eğitilmiştir. Modeller için yığın boyutu 8 olarak belirlenmiş, öğrenme oranı başlangıçta 0,01 olarak ayarlanmış ve ardından her 15 epokta onda bir oranında azaltılarak modelin daha hassas ayarlamalarla optimize edilmesi sağlanmıştır.

4.1.3.1. Modellerin eğitimi

Önerilen R3(2+1)D-SLR modelinin performansını değerlendirmek için, aynı derinlikte yalnızca R(2+1)D bloklarını içeren R(2+1)D-10 mimarisi ve yalnızca R3D bloklarını içeren R3D-10 mimarisi geliştirilmiştir. Bunun yanında, daha fazla katmana sahip R3D-18 ve R(2+1)D-18 (Tran ve diğ., 2018) modellerinin ön eğitimli olmayan versiyonları da karşılaştırma amacıyla kullanılmıştır. Model performanslarının değerlendirilmesinde RGB vücut verileri kullanılmıştır. Bu modellerin eğitim sonrası test performansları Çizelge 4.2'de karşılaştırılmıştır.

Çizelge 4.2. Modellerin karşılaştırılması (BosphorusSign22k-genel veri kümesi) (Akdag ve Baykan, 2024a)

Modeller	Test Doğruluğu (%)	Eğitim Süresi (sn/yığın)	Çıkarım Süresi (ms)	Parametre Sayısı (milyon)
R3D-18	71,33	0,58	372	33
R(2+1)D-18	73,12	1,01	617	31
R3D-10	54,79	0,12	54	1,9
R(2+1)D-10	72,18	0,14	79	1,8
R3(2+1)D-SLR	79,66	0,24	116	3,8

R3(2+1)D-SLR mimarisi, BosphorusSign22k-genel veri kümesinde diğer modellere göre önemli bir performans avantajı göstermiştir. R3(2+1)D-SLR, İDT görevinde hem test doğruluğu (%79,66) hem de çıkarım süresi (116 ms) açısından dengeli bir verimlilik sunmuştur. Özellikle, modelin test doğruluğu, aynı derinlikte yalnızca R(2+1)D veya R3D blokları içeren modellerden önemli ölçüde daha yüksektir. Bu da blok düzeyindeki füzyon yaklaşımının modellerin genelleme yeteneğini artırarak performansın nasıl geliştirebileceğini göstermektedir.

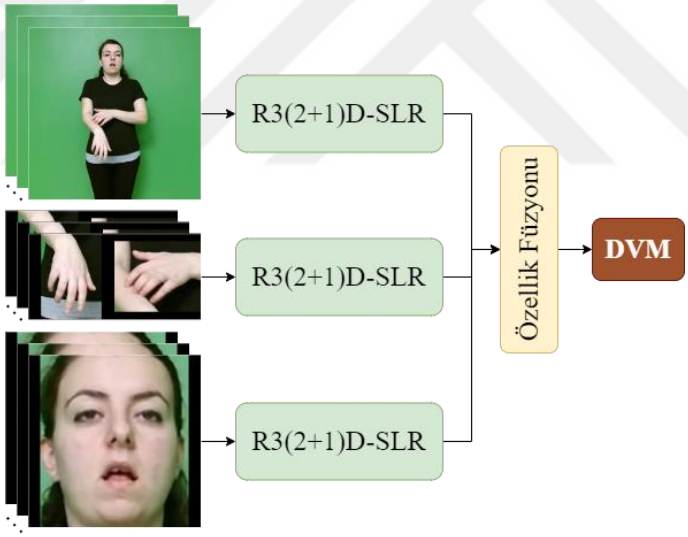
Eğitim ve çıkarım süresi metrikleri, R3(2+1)D-SLR modelinin karmaşıklık ve performans arasında uygun bir denge sağladığını göstermektedir. Bu denge, modelin gerçek zamanlı uygulamalar için pratik bir çözüm olabileceğini vurgulamakta ve uygulanabilirliğini artırmaktadır. Ayrıca, modelin 3,8 milyon parametre sayısı, hesaplama yükünü makul seviyelerde tutarken yüksek bir test doğruluğu elde edebilmektedir. Bu da modelin verimliliğini ve uygulanabilirliğini daha da artırmaktadır.

Sonuç olarak, R3(2+1)D-SLR modeli tarafından sağlanan dengeli performans profili, onu İDT görevleri için tercih edilebilir bir seçim olarak konumlandırmaktadır. Bu nedenle, bu çalışmanın (çalışma-1) ilerleyen bölümlerinde İDT sistem tasarımı için yalnızca önerilen R3(2+1)D-SLR ağı kullanılmıştır.

4.1.3.2. Ham verilere dayalı İDT sistemi

Tam bir vücut görüntüsü, işareti oluşturan tüm bileşenleri içeriyor olsa da çözünürlüğün 112×112'ye düşürülmesi, önemli ölçüde ayrıntı kaybına neden olur. Bu zorluk, vücut görüntülerine ek olarak, işareti temsil eden yüz ve el görüntülerinin de ayrı ayrı analiz edilmesiyle ele alınmıştır. Her biri için ayrı modeller eğitilerek, bu verilerin benzersiz temsil kabiliyetlerinden tam olarak yararlanılmıştır.

Çalışma kapsamındaki işaret tahmin stratejisi, eğitilmiş her modelin Tam Bağlantılı Katmanlarından önce çıkarılan 128 boyutlu özellik vektörlerini birleştirerek 384 boyutlu kapsamlı bir özellik vektörünün oluşturulmasına dayanır. Bu bileşik vektörü daha sonra, işaret dili kelimelerinin bütünsel bir temsilini yakalamak amacıyla DVM sınıflandırıcısı için girdi olarak kullanılmıştır. Şekil 4.7 önerilen sistemin genel diyagramını göstermektedir. Eğitilen modellerden ve füzyon + DVM yönteminden elde edilen sonuçlar Çizelge 4.3'te gösterilmiştir.



Şekil 4.7. Ham veri tabanlı sistemin diyagramı (Akdag ve Baykan, 2024a)

Çizelge 4.3. Ham veri tabanlı sistemin performansı (BosphorusSign22k-genel veri kümesi)

Özellik	Doğruluk (%)	Kesinlik (%)	Duyarlılık (%)	F1 skoru (%)
Vücut	79,66	79,35	81,65	77,88
Eller	83,24	82,15	85,8	81,16
Yüz	22,23	20,53	21,03	18,31
Füzyon + DVM	91,78	91,05	92,65	90,57

Çizelge 4.3'te belirtildiği gibi, el görüntüleri tam vücut görüntülerinden tutarlı bir şekilde daha iyi performans göstermiş ve %83,24'lük bir test doğruluğu ile ayırt edici özellikleri yakalamıştır. Yüz görüntüleri ise %22,23 ile daha düşük bir test doğruluğu elde

etse de 174 kelime sınıfında benzersiz özelliklerin belirlenmesinde etkili olmuştur. Eğitilen R3(2+1)D-SLR modellerinden elde edilen vücut, eller ve yüze dair özelliklerin entegrasyonunun DVM ile sınıflandırılması, test doğruluğunu %91,78'e yükseltmiştir. Bu yaklaşım, %91,05 kesinlik ve %92,65 duyarlılık değerleriyle sınıflandırma başarısını önemli ölçüde artırmış, ayrıca %90,57'lik F1 skoru ile bu başarının sadece doğrulukla sınırlı olmadığını, aynı zamanda modelin tutarlılık ve güvenilirlik açısından da iyi düzeyde performans sergilediğini göstermiştir. Bu sonuçlar, özellik füzyonu tabanlı yaklaşımın etkinliğini ortaya koymaktadır.

Farklı koşullar altında sistemin sağlamlığını değerlendirmek amacıyla arka planlar eklenerek değiştirilen test kümesiyle de analizler yapılmıştır. Elde edilen sonuçlar Çizelge 4.4'te gösterilmiştir.

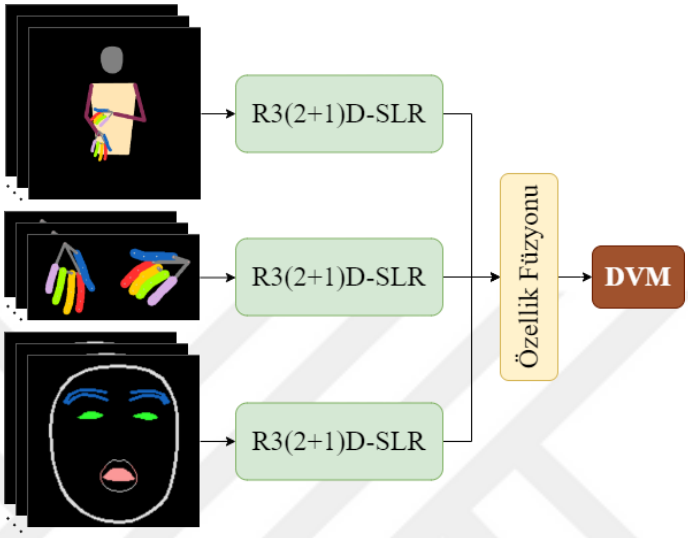
Çizelge 4.4. Değişken arka planlı test kümesi için ham veri tabanlı sistemin performansı (BosphorusSign22k-genel veri kümesi)

Özellik	Doğruluk (%)	Kesinlik (%)	Duyarlılık (%)	F1 skoru (%)
Vücut	16,12	15,87	21,64	13,93
Eller	56,69	54,81	59,81	51,08
Yüz	19,81	18,53	17,52	15,91
Füzyon + DVM	72,39	71,4	79,41	70,81

Test verilerine farklı arka planların eklenmesi, modellerin gerçek dünya koşulları altında nasıl performans gösterdiğinin anlaşılması açısından önemli bir adım olmuştur. Bu değişiklikle birlikte, test doğruluklarında belirgin bir düşüş yaşanmış ve arka plan karmaşıklığının model performansı üzerindeki etkisi net bir şekilde ortaya konulmuştur. Örneğin, vücut görüntülerinde doğruluk oranı %79,66'dan %16,12'ye düşmüştür. El görüntüleri, %83,24'ten %56,69'a düşen doğruluk oranı ile daha dirençli bir performans sergilemiş olsa da önemli ölçüde etkilenmiştir. Yüz görüntülerinin performansı ise %19,81 doğrulukla zaten düşük olan performansını nispeten korumuştur. Bu durum, arka plan görsellerinin görüntüde kapladığı alanla orantılı olarak performansın düştüğü şeklinde yorumlanabilir. Özellikle dikkat çeken bir durum, birleştirilmiş özelliklerin DVM ile sınıflandırılması ile test edilen yöntemin doğruluk oranının %91,78'den %72,39'a düşmesi olmuştur. Bu strateji, %71,4 kesinlik, %79,41 duyarlılık ve %70,81'lik F1 skoru ile nispeten yüksek değerler sunarak önerilen bütünleşik sistemin arka plan değişikliklerine karşı sağlamlığını göstermiştir.

4.1.3.3. Poz verilerine dayalı İDT sistemi

Değişken arka plan görüntülerine karşı daha dayanıklı bir İDT sistemi oluşturmak için, modellerin eğitimi görsel poz verileriyle tekrarlanmıştır (Şekil 4.8). Modellerin test doğruluk performansları Çizelge 4.5, Çizelge 4.6 ve Çizelge 4.7'de gösterilmiştir.



Şekil 4.8. Poz tabanlı sistemin diyagramı (Akdag ve Baykan, 2024a)

Çizelge 4.5. Poz tabanlı sistemin performansı (BosphorusSign22k-genel veri kümesi)

Özellik	Doğruluk (%)	Kesinlik (%)	Duyarlılık (%)	F1 skoru (%)
Vücut	79,97	79,04	81,06	77,94
Eller	88,40	87,32	89,04	86,41
Yüz	10,64	9,5	8,41	7,17
Füzyon + DVM	94,52	94,04	95,72	93,72

Çizelge 4.5, ham görüntülere kıyasla hem vücut hem de el pozu görüntüleri için performansın arttığını göstermektedir. Özellikle el pozlarının kullanılması, modelin performansını belirgin şekilde iyileştirmiştir; doğruluk oranı %5,16'lık bir artışla %88,40'a ulaşarak modelin el pozu görüntülerini ne kadar etkin bir şekilde işlediğini ortaya koymaktadır. Bu gelişme, modelin ele dair detayları ve karakteristik özellikleri daha başarılı bir şekilde tanıyıp sınıflandırabildiğini gösterir.

El pozu görüntülerinden elde edilen bu sonuçlar doğrultusunda, parmak rengi farklılaştırmasının etkisini daha detaylı incelemek amacıyla, BosphorusSign22k-genel veri setinden hem renkli hem de renksiz el pozu görüntüleri kullanılarak bir karşılaştırma yapılmıştır. Şekil 4.9 ham el görüntüleri de dahil olmak üzere, renksiz ve renkli el pozu görüntülerini içermektedir. Çizelge 4.6, eğitilen modellerin, bu üç farklı veri sürümü üzerinde elde ettiği sonuçları sunmaktadır.

(a) (b) (c)

Şekil 4.9. El verilerinin tüm versiyonları
(a) ham veri; (b) renksiz el pozu; (c) renkli el pozu (Akdag ve Baykan, 2024a)

Çizelge 4.6. El özelliklerinin karşılaştırılması

Özellik	Test Doğruluğu (%)
Ham el görüntüsü	83,24
Renksiz el pozu görüntüsü	83,14
Renkli el pozu görüntüsü	88,40

El parmaklarının farklı renklerle boyanması, test doğruluğunu renksiz versiyona göre %5,26, ham görüntü versiyonuna göre ise %5,16 oranında artırmıştır. Bunun nedeni, her parmağın farklı bir renkle renklendirilmesiyle parmağın yöneliminin, konumunun ve şeklinin daha iyi ayırt edilmesidir. Dolayısıyla, renkli parmaklara sahip el pozu görüntüsü hem ham el görüntüsünden hem de renksiz el pozu görüntüsünden daha iyi performans göstermiştir.

Yüz pozu görüntüleri için ham verilere kıyasla doğrulukta bir düşüş olmuştur. Bunun nedeni, yüz pozu görüntüsünde kaşlar, gözler ve dudaklar belirgin olsa da işareti temsil edebilecek birçok detayın kaybolmasıdır. Yüz pozu görüntülerindeki bu düşüşe rağmen, poz görüntülerinden çıkarılan özelliklerin birleştirilmesi ve DVM ile sınıflandırılması ile yapılan testte %91,78'den %94,52'ye bir artış olmuştur.

Çizelge 4.7. Değişken arka planlı test kümesi için poz tabanlı sistemin performansı (BosphorusSign22k-genel veri kümesi)

Özellik	Doğruluk (%)	Kesinlik (%)	Duyarlılık (%)	F1 skoru (%)
Vücut	76,29	75,34	78,0	73,8
Eller	86,51	85,32	87,63	84,46
Yüz	9,79	8,96	8,87	6,93
Füzyon + DVM	93,25	92,71	93,83	92,04

Çizelge 4.7'de, farklı arka planlar eklenerek modifiye edilmiş test verileri üzerinden gerçekleştirilen değerlendirmeler sunulmuştur. Bu değerlendirmelerde, ham görüntülerde görülen düşüşe göre oldukça sınırlı bir performans kaybı yaşanmıştır. Ham görüntülere kıyasla, poz görüntülerinde gözlemlenen performans artışı vücut görüntüleri için %16,12'den %76,29'a çıkarken, el görüntüleri için %56,69'dan %86,51'e yükselmiştir. Yüz görüntüleri ise arka plan değişikliklerinden minimal derecede etkilendiğinden dolayı bu kategoride bir performans artışı kaydedilmemiştir. Özellikle

dikkat çekici olan, füzyon + DVM sistem performansının %72,39'dan %93,25'e büyük bir sıçrama yapmasıdır. Elde edilen sonuçlar, sistemin farklı arka planlarda bile yüksek doğruluk ve kesinlikte sonuçlar verebildiğini göstermektedir, bu da önerilen İDT sisteminin sağlamlığını ve adaptasyon kabiliyetini ön plana çıkarmaktadır.

4.1.3.4. Diğer çalışmalarla karşılaştırmalar

Bu bölümde BosphorusSign22k-genel ve LSA64 veri kümeleri kullanılarak literatürde yapılan çalışmalar ile çalışma-1 kapsamında önerilen nihai İDT sisteminin performansları karşılaştırılmıştır. Karşılaştırma için yalnızca orijinal test verileri ile elde edilen sonuçlara yer verilmiştir. BosphorusSign22k-genel veri kümesini kullanan çalışmalar ve önerilen sistemin test sonuçları Çizelge 4.8'de gösterilmiştir.

Çizelge 4.8. Çalışma-1 kapsamında önerilen yöntemin BosphorusSign22k-genel veri kümesi üzerindeki literatür sonuçlarıyla karşılaştırılması

Çalışmalar	Girdi modalitesi	Odak Alanları	Test Doğruluğu (%)
Kindiroglu ve diğ. (2019)	Poz	Tam çerçeve	81,58
Gündüz ve Polat (2021)	RGB, poz, optik akış	Vücut, el, yüz	89,35
Çalışma-1 (önerilen yöntem)	RGB	Vücut, el, yüz	91,78
	Poz	Vücut, el, yüz	94,52

Çizelge 4.8'den de görülebileceği gibi, Kindiroglu ve diğ. (2019) tarafından sunulan, poz görüntülerine dayanan çalışma %81,58'lik bir doğruluk elde ederken, Gündüz ve Polat (2021) çok modlu verileri (RGB, poz, optik akış) kullanarak; vücut, eller ve yüz odak alanlarını birlikte değerlendirerek %89,35'lik daha yüksek bir doğruluk oranı elde etmiştir. Önerilen yöntem, ham görüntülerle %91,78, poz tabanlı girdilerle %94,52 doğruluk elde ederek mevcut çalışmaları geride bırakmıştır. Ayrıca, önerilen R3(2+1)D-SLR ağı, renkli el pozu görüntüsü ile %88,40 test doğruluğu elde etmiştir. Bu sonuçlar, önerilen modelin hem ham hem de poz görüntülerinden zengin uzamsal ve zamansal özellikleri etkili bir şekilde çıkarabildiğini göstermektedir. Vücuttan, ellerden ve yüzden çıkarılan bu özelliklerin birleştirilmesiyle tanıma doğruluğu önemli ölçüde artırılmıştır.

LSA64 veri kümesi için, literatürdeki çalışmalarla karşılaştırma yapmak amacıyla işaretçiye bağlı ve işaretçiden bağımsız değerlendirmeler yapılmıştır. Sonuçlar Çizelge 4.9'da gösterilmiştir.

Çizelge 4.9. Çalışma-1 kapsamında önerilen yöntemin LSA64 veri kümesi üzerindeki literatür sonuçlarıyla karşılaştırılması

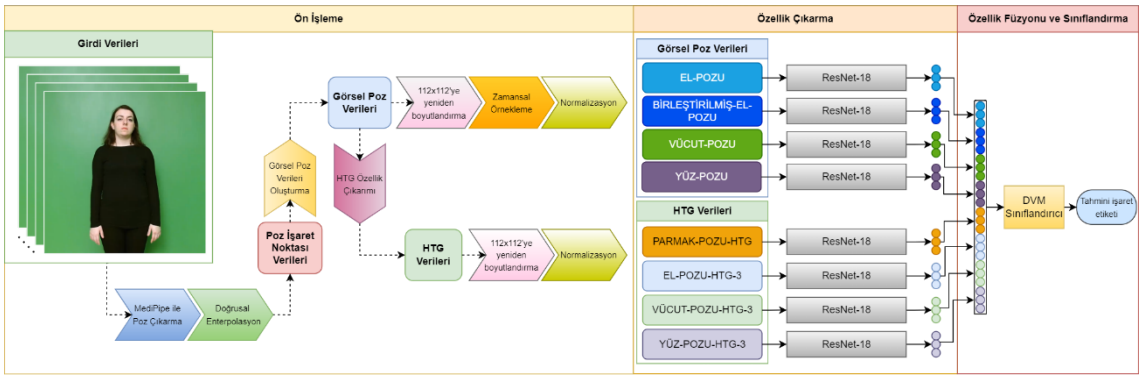
	Çalışmalar	Girdi modalitesi	Odak Alanları	Test Doğruluğu (%)
D1	Marais ve diğ. (2022b)	RGB	Tam çerçeve	74,22
	Çalışma-1 (önerilen yöntem)	RGB	Vücut, el, yüz	99,37
		Poz	Vücut, el, yüz	96,56
D2	Ronchetti ve diğ. (2016a)	RGB	El	91,70
	Rodríguez ve Martínez (2018)	RGB	Tam çerçeve	85
	Alyami ve diğ. (2023)	Poz	El, yüz	91,09
	Çalışma-1 (önerilen yöntem)	RGB	Vücut, el, yüz	99,53
		Poz	Vücut, el, yüz	98,53
D3	Ronchetti ve diğ. (2016a)	RGB	El	97
	Konstantinidis ve diğ. (2018b)	Poz	Vücut, el	98,09
	Konstantinidis ve diğ. (2018a)	RGB, poz, optik akış	Vücut, el, yüz	99,84
	Zhang ve Li (2019)	RGB	Tam çerçeve	99,063
	Imran ve Raman (2020)	HTG, DG, RGB-HG	Tam çerçeve	97,81
	Marais ve diğ. (2022a)	RGB	Tam çerçeve	95,50
	Bohacek ve Hruz (2022)	Poz	Vücut, el	100
	Alyami ve diğ. (2023)	Poz	El, yüz	98,25
	Çalışma-1 (önerilen yöntem)	RGB	Vücut, el, yüz	99,84
		Poz	Vücut, el, yüz	100

Çizelge 4.9'dan da görülebileceği gibi, önerilen özellik füzyonu tabanlı yöntem, LSA64 veri kümesini kullanarak hem işaretçi bağımsız değerlendirmeler olan D1 ve D2'de hem de işaretçiye bağlı bir değerlendirme olan D3'te literatürdeki mevcut çalışmalardan daha iyi performans göstermiştir. Beşinci ve onuncu işaretçinin test verisi olarak ayrıldığı D1 ortamında, önerilen yöntem Marais ve diğ. (2022b) tarafından InceptionV3-GRU modeline dayanan ve tam bir çerçeve görüntüsünü kullanan yöntemle karşılaştırıldığında %99,37 tanıma doğruluğu elde ederek uzamsal ve zamansal özellik çıkarımında üstünlüğünü göstermiştir. İşaretçi bağımsız bir başka deney ortamı olan D2'de, Alyami ve diğ. (2023) tarafından elde edilen %91,09 doğruluk oranı dönüştürücü modellerin güçlü yönlerini gösterirken, önerilen yöntem ham görüntü girdisi için %99,53, poz tabanlı girdi için %98,53 doğruluk elde ederek tanıma doğruluğunda önemli bir artış sağlamıştır. İşaretçi bağımlı bir değerlendirme olan D3'te, Konstantinidis ve diğ. (2018a) tarafından sunulan, RGB, poz ve optik akış verilerine dayalı çok modlu bir yaklaşım kullanarak vücut, eller ve yüzü kapsamlı bir şekilde ele alan VGG-16 ve UKSB tabanlı bir yöntem ile elde ettikleri %99,84'lük doğruluk oranı İDT'de kayda değer bir başarıdır. Bununla birlikte, önerilen yöntem ham görüntü tabanlı girdi için %99,84, poz tabanlı girdi için %100 doğruluk elde ederek üstünlük sağlamıştır. Özetle çalışma-1 kapsamında önerilen yöntem LSA64 veri kümesini kullanan literatürdeki çalışmaların performanslarını geçen üstün bir başarı elde etmiştir.

4.2. Çalışma-2: Poz ve HTG Verilerinin Entegrasyonu Yoluyla İşaret Dili Tanıma

Bu çalışmada, bir önceki bölümde anlatılan görsel poz verilerinin kullanımının ve özellik füzyonu tabanlı önerilen yöntemin etkinliğine dayanarak, yine poz verilerine ve bu verilerden türetilen Hareket Tarihçesi Görüntülerinin (HTG) entegrasyonuna dayalı bir İDT sistemi geliştirilmiştir. Bu sistemde, vücut, el ve yüz pozlarından elde edilen uzamsal bilgiler, işaretin zamansal dinamiklerine ilişkin üç kanallı HTG verilerinin sağladığı bilgilerle birleştirilmektedir. Özellikle, geliştirilen parmak PARMAK-POZU-HTG, İDT'deki yaklaşımların aksine, parmak hareketlerinin ve jestlerin ince nüanslarını yakalayarak tanıma başarısını önemli ölçüde artırmıştır. Ayrıca, bu çalışmada doğrusal enterpolasyon yoluyla eksik poz işaret noktaları tahmin edilerek genel tanıma doğruluğunda iyileşmeler sağlanmıştır. RReLU ile geliştirilmiş ResNet-18 modeline dayanan bu çalışmada, çıkarılan özelliklerin füzyonu ve bir DVM ile sınıflandırma yoluyla manuel ve manuel olmayan özellikler arasındaki etkileşim başarıyla ele alınmıştır. Bu bölümde anlatılan çalışma "Isolated Sign Language Recognition through Integrating Pose Data and Motion History Images" (Akdağ ve Baykan, 2024) isimli makalede sunulmuştur.

Şekil 4.10'daki diyagram, önerilen İDT sisteminin genel yapısını ve temel bileşenler arasındaki etkileşimi görsel olarak özetlemektedir. Aşağıdaki alt bölümler, sistematik bir yaklaşım kullanarak araştırmanın kapsamlı bir açıklamasını sunmaktadır.



Şekil 4.10. Çalışma-2 için önerilen sistemin diyagramı (Akdağ ve Baykan, 2024)

4.2.1. Veri ön işleme

4.2.1.1. Poz işaret noktalarının çıkarılması ve eksik poz verilerinin tahmini

Çalışmada ilk olarak, MediaPipe'ın Holistic çözümünden yararlanarak poz işaret noktaları çıkarılmıştır. Fakat özellikle el kamera açısına göre uygun bir konumda değilse, kameranın görüş alanının dışına çıkıyorsa, videodaki hareketler çok hızlıysa veya video sırasında el, bir nesne veya vücut tarafından kısmen kapatılmışsa, bazı karelerde el işaret noktaları elde edilemeyebilir. Nitekim, toplamda 492.541 çerçeveden oluşan BosphorusSign22k-genel veri kümesi ile yapılan deneylerde, bu çerçevelerin 21.551'inde sağ el pozu işaret noktası, 46.088'inde ise sol el pozu işaret noktası eksiktir. Bu eksik işaret noktaları sınıflandırma performansını etkileyebilir. Bu çalışmada, eksik el pozu işaret noktalarını tahmin etmek için doğrusal enterpolasyon yöntemi kullanılmıştır. Eksik bir el pozu işaret noktasının değeri, $L_{el,eksik,n}$, Denklem 4.4 kullanılarak tahmin edilebilir:

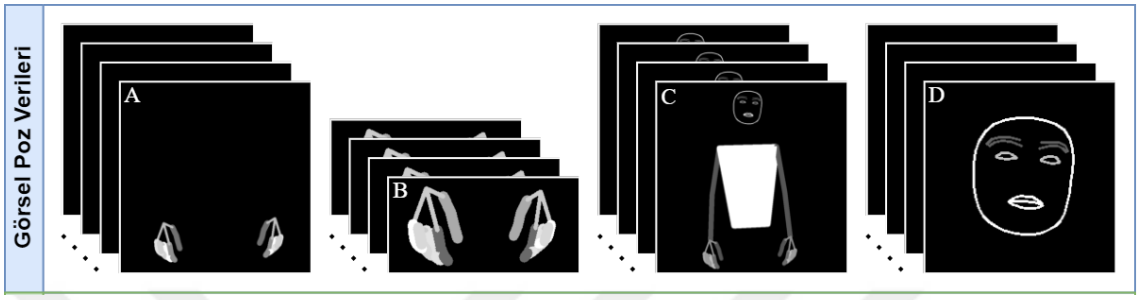
$$L_{el,eksik,n} = L_{el,bilinen,n-1} + i.\frac{(L_{el,bilinen,n+1} - L_{el,bilinen,n-1})}{(n+1)}. \quad (4.4)$$

Bu denklem, $L_{el,bilinen,n-1}$ ve $L_{el,bilinen,n+1}$ olarak gösterilen eksik veriden hemen önceki ve sonraki el pozu işaret noktalarının bilinen konumlarını kullanarak eksik el pozu işaret noktalarının konumunu doğrusal olarak enterpole eder. Eksik poz işaret noktalarının tespit edildiği bu yöntemin, örnek bir el pozu görüntüsünden elde edilen HTG verisine etkisi Şekil 4.11'de gösterilmiştir.

Şekil 4.11. Eksik poz tamamlamanın EL-POZU-HTG-1 üzerindeki etkisi (Akdağ ve Baykan, 2024)

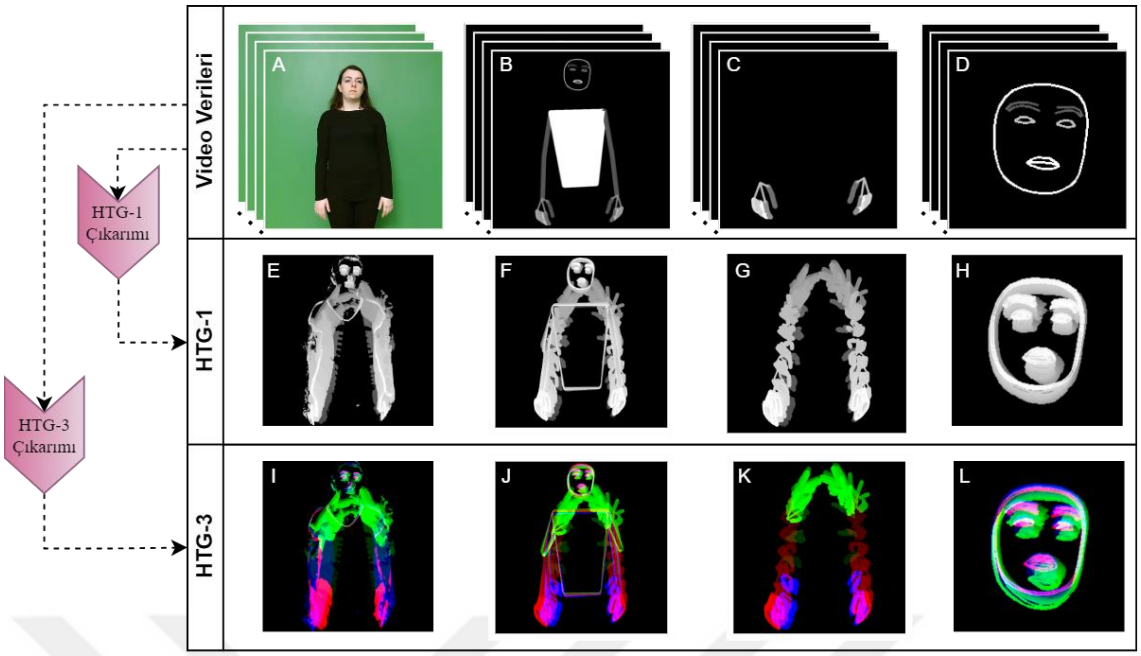
4.2.1.2. Görsel poz verileri ve Hareket Tarihçesi Görüntülerinin oluşturulması

MediaPipe Holistic çözümü kullanılarak her video karesinden vücut, eller ve yüz için işaret noktaları elde edildikten sonra, Bölüm 3.4'te açıklanan adımlar kullanılarak poz görüntü dizileri oluşturulmuştur. Bu görüntüler Şekil 4.12'de gösterilmektedir.



Şekil 4.12. Örnek bir işaret dili videosundan elde edilen görsel poz verileri
(A) EL-POZU, (B) BİRLEŞTİRİLMİŞ-EL-POZU, (C) VÜCUT-POZU, (D) YÜZ-POZU

Poz görüntülerinin oluşturulmasının ardından zamansal bilgilerin elde edilmesi için, bu poz görüntülerine HTG ve 3 kanallı HTG yöntemleri uygulanmıştır. 3 kanallı HTG (HTG-3) video dizisini üç eşit parçaya böler ve bu parçalardan her biri için ayrı HTG hesaplayarak oluşturulur. Böylece her renk kanalı (kırmızı, yeşil ve mavi) belirli zaman aralıklarındaki hareket bilgisini temsil eder. Örnek bir işaret dili videosu ve bu videodan elde edilen poz görüntülerinden HTG-1 ve HTG-3 verilerinin elde edilmesi Şekil 4.13'te gösterilmiştir.



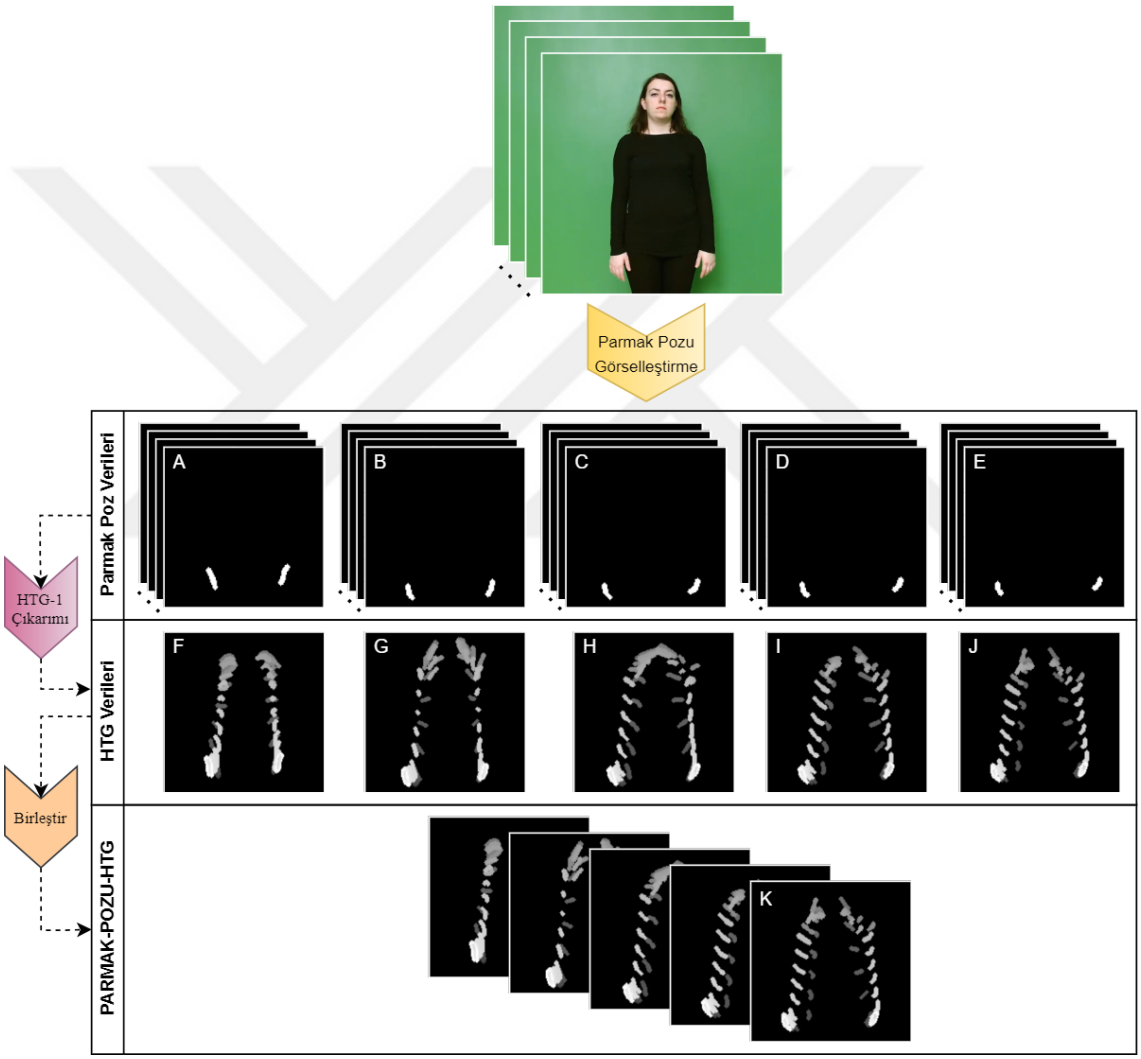
Şekil 4.13. Ham video ve poz verilerinden HTG-1 ve HTG-3 özelliklerinin çıkarılması (A) HAM veri, (B) VÜCUT-POZU, C) EL-POZU, (D) YÜZ-POZU, (E) HAM-HTG-1, (F) VÜCUT-POZU-HTG-1, (G) EL-POZU-HTG-1, (H) YÜZ-POZU-HTG-1, (I) HAM-HTG-3, (J) VÜCUT-POZU-HTG-3, (K) EL-POZU-HTG-3, (L) YÜZ-POZU-HTG-3 (Akdağ ve Baykan, 2024)

4.2.1.3. Parmak pozu tabanlı HTG

Bu bölümde, İDT sistemlerinde parmak hareketlerinin zamansal dinamiklerini yakalamayı amaçlayan ve bu çalışmanın yeni bir katkısı olan parmak pozu tabanlı hareket tarihçesi görüntüsünün (PARMAK-POZU-HTG) geliştirilmesi açıklanmaktadır. Başlangıçta, Bölüm 3.4'te açıklanan prosedürler izlenerek, $L_{eller,n}$ içindeki her parmağa özgü işaret noktaları kullanılarak, her bir parmak için poz görüntüleri oluşturulmuştur. Bu metodolojik yaklaşım, her biri bir video karesi boyunca parmak hareketlerinin bir tasvirini sağlayan beş farklı görüntü dizisinin oluşturulmasıyla sonuçlanmıştır. Daha sonra, HTG yöntemi bu dizilere ayrı ayrı uygulanarak her parmak için bir HTG verisi elde edilmiştir. Bu verilerin bütünleştirilmesi, parmak hareket dinamiklerinin özünü yansıtan kapsamlı bir beş kanallı HTG'nin oluşturulmasıyla sonuçlanmıştır. Bu yeni özelliğin matematiksel formülasyonu Denklem 4.5'te tanımlanmıştır:

$$\text{PARMAK} - \text{POZU} - \text{HTG} = \text{Birleştir}\left(H_\tau^{\text{baş\_parmak}}, H_\tau^{\text{işaret\_parmağı}}, H_\tau^{\text{orta\_parmak}}, H_\tau^{\text{yüzük\_parmağı}}, H_\tau^{\text{serçe\_parmak}}\right) \quad (4.5)$$

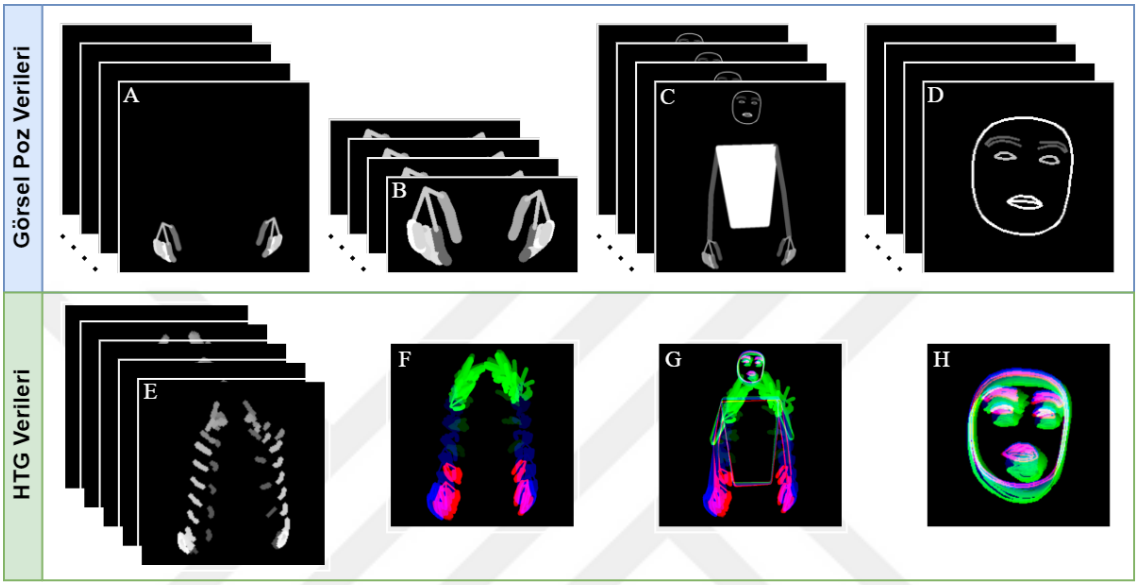
Bu denklem, her bir parmak için HTG'yi temsil eden H_τ^f işlevini kullanır. Burada f belirli bir parmağı (baş parmak, işaret parmağı, orta parmak, yüzük parmağı, serçe parmak) simgeler. Her H_τ^f işlevi x, y piksel konumundaki ilgili parmağın zaman (t) içerisindeki hareketlerinin birikimini ifade eder. "Birleştir" işlevi, hareket tarihçesi görüntülerini birleştirerek, beş farklı parmağın hareketinin zaman içindeki birikimini sunan 5×x×y boyutunda görsel bir veri matrisi oluşturur. PARMAK-POZU-HTG'nin oluşturulmasına ilişkin bir örnek Şekil 4.14'te gösterilmektedir.



Şekil 4.14. Parmak pozu görüntülerinden PARMAK-POZU-HTG'nin elde edilmesi (A) BAŞPARMAK POZU, (B) IŞARET PARMAĞI POZU, (C) ORTA PARMAK POZU, (D) YÜZÜK PARMAĞI ÇANTASI, (E) SERÇE PARMAK POZU, (F) BAŞPARMAK POZU-HTG, (G) İŞARET PARMAĞI POZU-HTG, (H) ORTA PARMAK POZU-HTG, (I) YÜZÜK PARMAĞI POZU-HTG, (J) SERÇE PARMAK POZU-HTG, (K) PARMAK-POZU-HTG (Akdağ ve Baykan, 2024)

4.2.1.4. Diğer ön işlem adımları

Yukarıda açıklanan süreçlere dayalı olarak nihai İDT sistemi için kullanılacak verilerin örnek bir gösterimi Şekil 4.15'te gösterilmiştir. Bu veriler ResNet-18 modellerine girdi olarak verilmeden önce bir dizi ön işleme adımından daha geçmektedir.



Şekil 4.15. Nihai İDT sisteminde kullanılacak veriler
(A) EL-POZU, (B) BİRLEŞTİRİLMİŞ-EL-POZU, (C) VÜCUT-POZU, (D) YÜZ-POZU, (E) PARMAK-POZU-HTG, (F) EL-POZU-HTG-3, (G) VÜCUT-POZU-HTG-3, (H) YÜZ-POZU-HTG-3 (Akdağ ve Baykan, 2024)

Öncelikle, çalışmada kullanılan derin öğrenme modellerinin verimli bir şekilde eğitilebilmesi ve doğru tahminler yapabilmesi için girdi verilerinin belirli bir format ve boyutta olması gerekmektedir. Bu nedenle tüm veriler, modelin girdi olarak beklediği boyut olan 112×112 çözünürlüğe indirgenmiştir. Ancak BİRLEŞTİRİLMİŞ-EL-POZU verileri her iki elin birleşik görüntüsünü içerdiğinden boyut 56x112 olarak ayarlanmıştır.

İkinci olarak, görsel poz verileri, hareketin zaman içindeki gelişimini yakalayan birden fazla kareden oluşan görüntü dizileridir. Bu nedenle, ResNet-18 modelinin bu görüntüleri etkili bir şekilde ele alabilmesi için, modelin işleyebileceği belirli sayıda kare seçilmesi gerekir. Model eğitiminin çeşitliliğini ve genelleştirilebilirliğini artırmak için, her eğitim döngüsünde bir video dizisinden rastgele 32 kare seçilmiştir. Bu şekilde, model, farklı zaman dilimlerindeki hareketleri öğrenebilir ve genel hareket modelini tanıyabilir. Bununla birlikte, veri setinde zamansal tutarlılığı ve karşılaştırılabilirliği sağlamak için modelin performansının değerlendirilmesi ve test edilmesi aşamalarında

her video dizisinden eşit aralıklı kareler seçilmiştir. Bu yöntem, modelin performansını, farklı videolar arasında daha adil ve tutarlı bir şekilde değerlendirilmesini sağlar.

Son aşamada normalizasyon işlemi gerçekleştirilmiştir. Bu işlem Denklem 4.6'da gösterilmiştir:

$$\text{Normalizasyon}(x) = \frac{x-\mu}{\sigma} \tag{4.6}$$

Burada x orijinal piksel değerini, μ ortalamayı ve σ standart sapmayı temsil eder. Normalleştirme işlemi, giriş verilerini ölçeklendirmek ve bu verilerin model tarafından daha verimli bir şekilde işlenmesini sağlamak için gerçekleştirilmiştir.

Nihai İDT sitemi için ResNet-18 modellerine girdi olarak verilecek tüm ön işleme adımlarından sonraki veri boyutları Çizelge 4.10'da gösterilmektedir.

Çizelge 4.10. Tüm ön işleme adımlarından sonra elde edilen veriler (Akdağ ve Baykan, 2024)

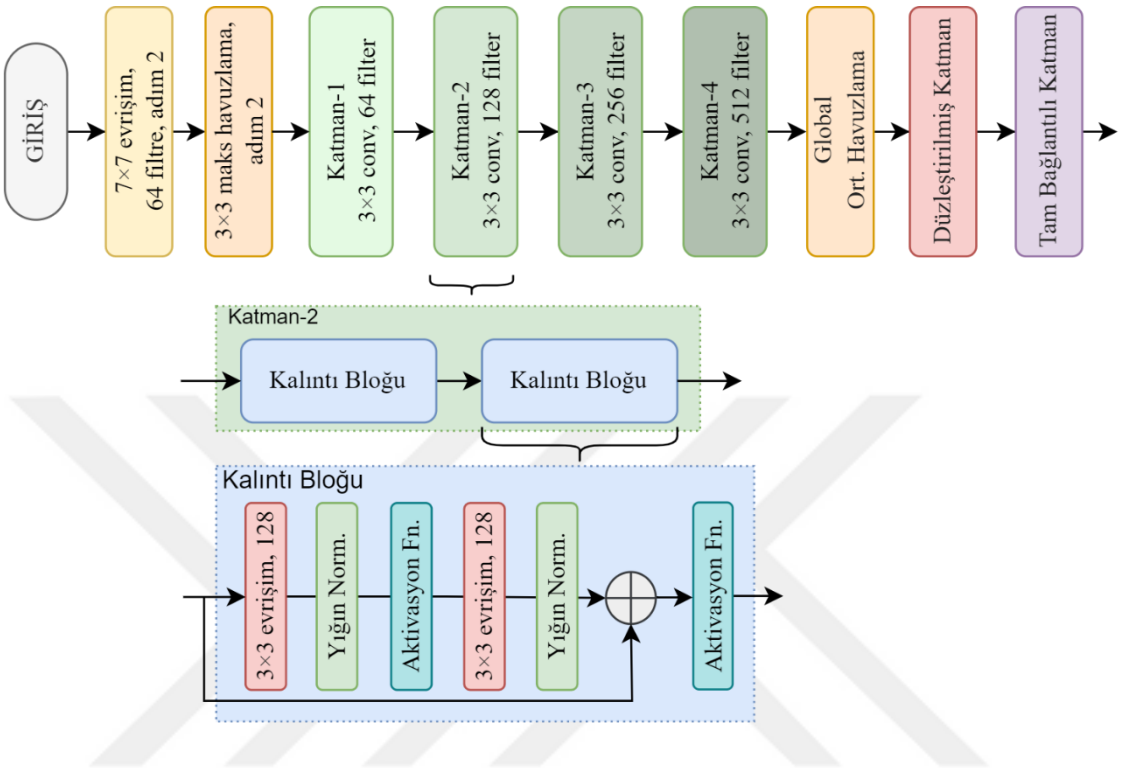
Veri	Boyut
EL-POZU	32x112x112
BİRLEŞTİRİLMİŞ-EL-POZU	32x56x112
VÜCUT-POZU	32x112x112
YÜZ-POZU	32x112x112
PARMAK-POZU-HTG	5x112x112
EL-POZU-HTG-3	3x112x112
VÜCUT-POZU-HTG-3	3x112x112
YÜZ-POZU-HTG-3	3x112x112

4.2.2. ResNet-18

Bu çalışma kapsamında, ResNet-18 modeli (He ve diğ., 2016) önerilen İDT sistemi için temel model olarak seçilmiştir. Bu seçim, ResNet-18'in kısa sürede etkili eğitim sağlayabilmesi ve yapısal özelliklerinin nispeten düşük hesaplama maliyeti gerektirmesi nedeniyle yapılmıştır.

ResNet-18 mimarisi, girdi boyutu 3 (RGB görüntü için), filtre boyutu 7×7 ve adım sayısı 2 olan 64 filtreli bir evrişim katmanı ile başlar. Daha sonra filtre boyutu 3×3 ve adım sayısı 2 olan bir Maksimum Havuzlama Katmanı gelir. Ardından her katmanda 2 kalıntı bloğu olacak şekilde dört katman üst üste yığılır. Katmanların filtre sayıları sırasıyla 64, 128, 256 ve 512'dir. Her bir kalıntı bloğu, filtre boyutu 3×3 olan iki evrişim katmanı içerir. Dört katmanın tamamlanmasının ardından, ağ mimarisindeki sonraki aşamalar Global Ortalama Havuzlama Katmanı (Lin ve diğ., 2013) ve ardından

sınıflandırma amacıyla Tam Bağlantılı Katmanın (Scabini ve Bruno, 2023) kullanımından oluşmaktadır. ResNet-18 model mimarisi ve bileşenleri Şekil 4.16'da gösterilmiştir.



Şekil 4.16. ResNet-18 model mimarisi ve temel bileşenleri

ResNet-18'in temel yapı taşı olan kalıntı blokları, derin ağların öğrenme yeteneklerini artırmak ve aşırı uyum sorunlarını azaltmak için tasarlanmıştır (He ve diğ., 2016). Kalıntı blokların temel fikri, giriş verilerini doğrudan bir sonraki katmana aktarmaktır. Bu, ağın öğrenme kapasitesini artırır ve derin yapıların eğitimini kolaylaştırır. Kalıntı bloğu Denklem 4.7'deki gibi tanımlanabilir:

$$x_i = x_{i-1} + F(x_{i-1}; \theta_i) \tag{4.7}$$

Burada x_{i-1} i katmanının giriş verisidir, $F(;\theta_i)$ θ_i ağırlıkları ile hesaplanan iki evrişimin kombinasyonunu, toplu normalizasyonu ve aktivasyon fonksiyonlarının uygulanmasını ifade eder.

ResNet-18'in İDT görevi için uyarlanması sırasında, ImageNet veri kümesi üzerinde eğitilmiş olan önceden tanımlanmış ağırlıklar kullanılarak modelin mimarisine ince ayar yapılmıştır. Transfer öğrenme yaklaşımı, modelin işaret dili veri kümemize

hızla adapte olmasını sağlayarak eğitim sürecini önemli ölçüde hızlandırır (Shaha ve Pawar, 2018). Bu metodoloji, yeni görevdeki performansı iyileştirmeyi ve böylece işaret dili sınıflandırmasının doğruluğunu artırmayı amaçlamaktadır.

Modelin yapısı, girdi katmanında RGB görüntüler için 3 kanallı (kırmızı, yeşil, mavi) 64 filtreli bir evrişim katmanı ile başlamasına rağman, bu çalışmada kullanılan verilerden bazıları modele girdi olarak verilebilecek uygun boyutta değildir: HTG-1 için bir kanal (1D), HTG-3 için üç kanal (3D), PARMAK-POZU-HTG için beş kanal (5D) ve video pozu verileri için otuz iki kanal (32D). Bu verilerin boyutları ResNet-18'in varsayılan giriş boyutlarıyla uyumlu olmadığından, uygun bir ön işleme adımı gerçekleştirilmiştir. Spesifik olarak, model mimarisinin başına veri türüne bağlı olarak değişen bir giriş boyutu, 3 çıkış boyutu, 3x3 çekirdek boyutu ve 1 adım boyutu ile bir evrişim katmanı eklenmiştir. Bu ön işleme adımı, çeşitli boyutlardaki verileri ResNet-18 modeli tarafından etkili bir şekilde işlenebilecek standart bir formata dönüştürerek, modelin özellik çıkarma için önceden eğitilmiş ağırlıklarından yararlanma yeteneğini engellemeden verilerin standart boyutlarının korumasını sağlar.

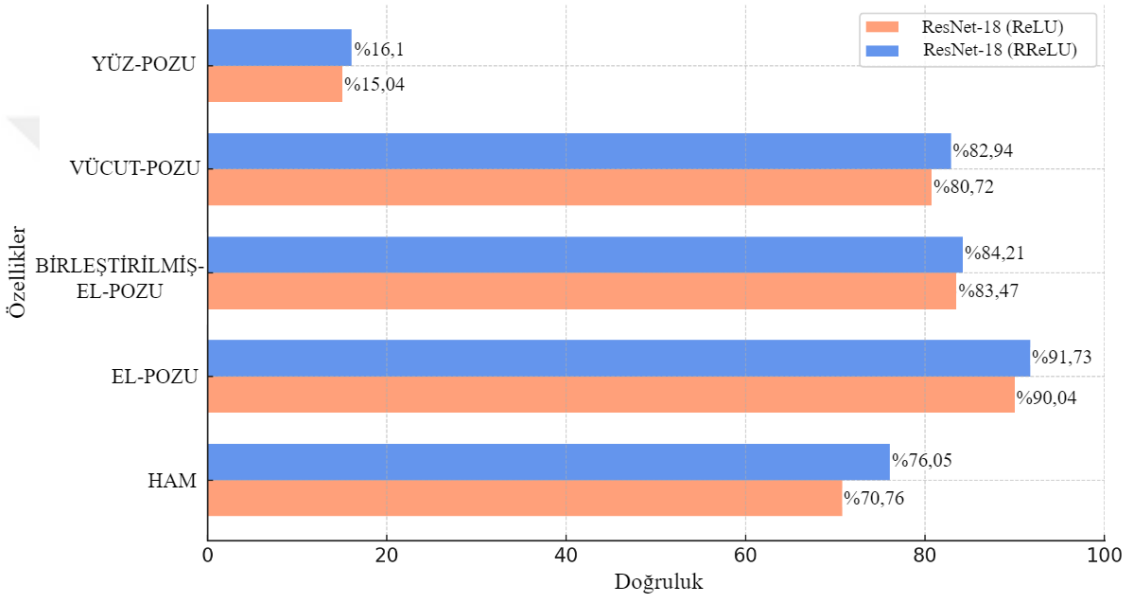
ResNet-18 model mimarisinde varsayılan olarak ReLU aktivasyon fonksiyonu kullanılır. Ancak, bu çalışma bağlamında, ResNet-18 mimarisindeki varsayılan aktivasyon fonksiyonu olan ReLU, RReLU ile değiştirilmiştir. RReLU, önceden tanımlanmış alt ve üst sınırlar dahilinde negatif aktivasyon değerleri için rastgele bir eğim seçerek parametrik olmayan bir davranış sunar. “Deneysel Sonuçlar” bölümünde, ReLU yerine RReLU aktivasyon fonksiyonunun kullanılmasının etkisi de analiz edilmiştir.

4.2.3. Deneysel sonuçlar

Bu çalışmada, ResNet-18 modelleri, aşırı uyumu önlemek amacıyla, Tam Bağlantılı Katmandan hemen önce yer alan 0,4 olasılıklı bir Düğüm Seyreltme katmanı ile donatılmıştır. Modeller için yığın boyutu 8 olarak ayarlanmış, öğrenme oranı ise başlangıçta 0,001 olarak belirlenmiş ve her 15 epokta bir, onda bir oranında azaltılmıştır. Modellerin eğitimi momentum katsayısı 0,9 olarak belirlenerek Stokastik Gradyan İnişi (SGD) optimizasyon algoritması ile 35 epok boyunca gerçekleştirilmiştir. Kullanılan video verileri için dinamik olarak rastgele seçilen 32 çerçeve üzerinde çalışılmış, ancak ek bir veri artırma tekniği uygulanmamıştır.

4.2.3.1. Poz görsellerinin sınıflandırma sonuçları

Başlangıçta, ResNet-18 modelleri ham video görüntüleri ve bu görüntülerden oluşturulan dört farklı poz görüntüsü kullanılarak eğitilmiştir. Özellikle bu deneyde, her bir özellik için İDT'nin doğruluğunu karşılaştırmak amacıyla ham video görüntüleriyle eğitilen modellerin performansı ile poz görüntüleriyle eğitilen modellerin performansı karşılaştırılmıştır. ReLU yerine RReLU aktivasyon fonksiyonunun kullanılmasının etkisi de incelenmiştir. Şekil 4.17'deki grafikte, elde edilen sonuçlar sunulmuştur.



Şekil 4.17. Ham ve poz görüntülerinin ResNet-18 ile sınıflandırılmasının sonuçları (BosphorusSign22k-genel veri kümesi) (Akdağ ve Baykan, 2024)

Şekil 4.17, model aktivasyon fonksiyonları arasındaki karşılaştırmaya göre, RReLU ile eğitilen ResNet-18 modellerinin tümünün ReLU ile eğitilen modellerden daha yüksek bir sınıflandırma doğruluğu sağladığını göstermektedir. Bu sonuçlar, aktivasyon fonksiyonunun derin öğrenme modellerinin performansında önemli bir rol oynayabileceğini doğrulamaktadır.

Elde edilen sınıflandırma sonuçları incelendiğinde, EL-POZU görüntüsüyle diğerlerine göre daha yüksek doğruluk oranları yakalanmıştır. Bu belirgin başarı, işaret dilinin özünün büyük ölçüde ellerle yapılan jestlere dayandığını ve ele dair bilgilerin işaret dili sınıflandırmasındaki önemli rolünü göstermektedir.

BİRLEŞTİRİLMİŞ-EL-POZU, el şekli, yönü ve benzeri konularda daha fazla ayrıntı içerir. Bu detay, ana görüntüden el bölgelerinin kırpılması ve birleştirilmesiyle

elde edilmiştir. Bu işlem, normalde genel görüntüde daha küçük ve belirsiz olan el bölgelerinin görsel olarak daha büyük ve daha ayrıntılı bir yapıda temsil edilmesini sağlamıştır. Ancak bu detaylı bilgiye rağmen, birleştirilmiş el pozu verileri, ellerin pozisyon bilgisini içermemektedir, bu da sınıflandırma sürecinde önemli bir eksikliktir. Ellerin pozisyon bilgisi, bazı işaretler pozisyon değişiklikleri yoluyla iletildiği için izole işaret dili kelimelerinin sınıflandırılmasında önemli bir rol oynamaktadır. Dolayısıyla, bu bilginin BİRLEŞTİRİLMİŞ-EL-POZU verileri tarafından kaybedilmesi, sınıflandırma doğruluğunda önemli bir düşüşe yol açmaktadır. Bu bulgular, İDT sistemlerinde kullanılacak özelliklerin seçiminde el pozisyonu bilgisinin ne kadar kritik olduğunu göstermektedir.

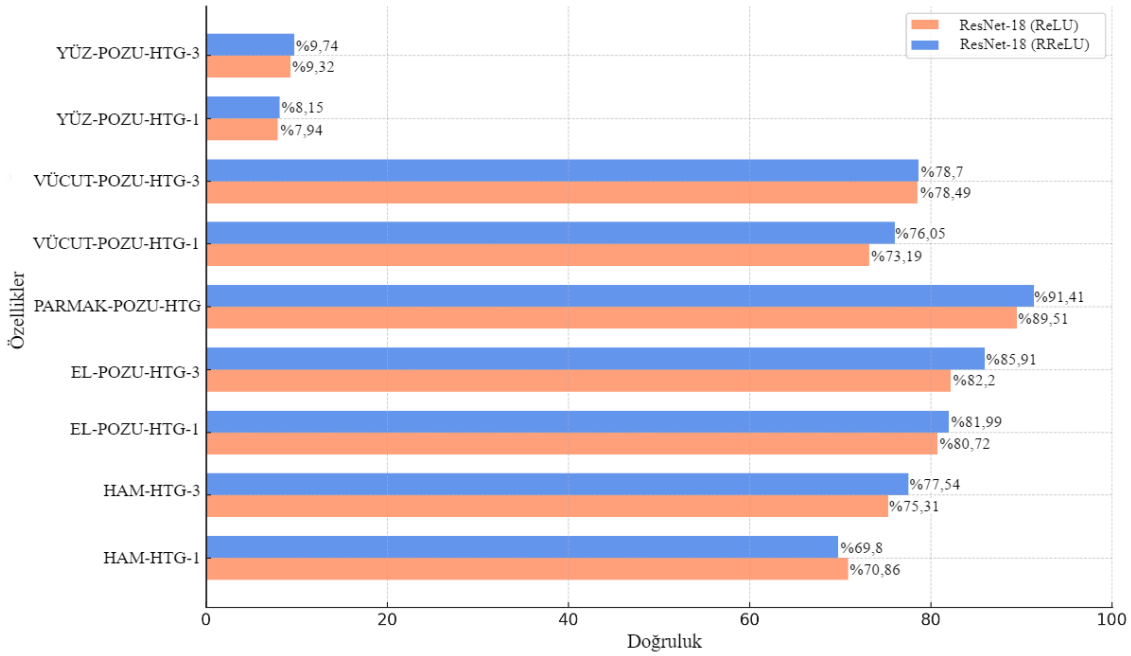
VÜCUT-POZU görüntüsü de işaret dili sınıflandırması için kritik bir öneme sahiptir. Bu, vücudun genel hareketlerini ve pozisyonlarını ve vücudun diğer bölümlerinin el hareketleriyle nasıl etkileşime girdiğini kapsar. Bununla birlikte, VÜCUT-POZU ile elde edilen doğruluk, EL-POZU ve BİRLEŞTİRİLMİŞ-EL-POZU görüntüleriyle elde edilen doğruluktan biraz daha düşüktür. Bunun temel nedeni, el hareketlerinin ve pozisyonlarının işaret dili kelimelerinin anlamını belirlemede daha belirleyici bir role sahip olmasıdır. Ellerin detaylı hareketleri birçok işaret dili kelimesinin anlamını doğrudan etkileyebilirken, genel vücut hareketleri bu yorumlamada dolaylı bir role sahiptir. Ayrıca, VÜCUT-POZU vücudun daha geniş bir alanını kapsadığından, ayrıntılı el hareketleri bu geniş alanda daha az belirgindir. Bu da VÜCUT-POZU özelliklerinin sınıflandırma doğruluğunu etkileyen bir başka faktör olabilir.

HAM video görüntüsünün doğrudan poz temsili olan VÜCUT-POZU, HAM video görüntülerine göre daha yüksek bir sınıflandırma doğruluğu göstermiştir. Bu, poz görüntülerinin ham görüntülere göre avantajını göstermektedir. Bunun nedeni, poz görüntülerinin yalnızca harekete veya jeste odaklanması; arka plandan ve işaretçinin saç, sakal/bıyık, kıyafet ve giyim tarzı gibi fiziksel özelliklerinden kaynaklanabilecek varyasyonları büyük ölçüde ortadan kaldırmasıdır. Dolayısıyla bu özellikler poz görüntülerinin genellenebilirliğini artırmakta ve sınıflandırmada ham video görüntüsünden daha üstün kılmaktadır.

YÜZ-POZU ile elde edilen düşük sınıflandırma doğruluğu, işaret dili sınıflandırması için yüze dair bilgilerin tek başına tam olarak yeterli olmadığını göstermektedir. Ancak işaret dili sadece el hareketleriyle sınırlı değildir; yüz ifadeleri ve vücut dili de önemli bir rol oynar. Bu nedenle, yüz özelliklerinin modelde bütünleşik olarak kullanılması daha doğru sınıflandırmalar elde etmek açısından önemlidir.

4.2.3.2. HTG verilerinin sınıflandırma sonuçları

Bu alt bölümde, ResNet-18'in sınıflandırma performansı farklı HTG verileriyle değerlendirilmiştir. Şekil 4.18'de sunulan bulgular, çeşitli HTG verileriyle elde edilen test doğruluklarını yansıtmaktadır. Bu araştırma, önceki alt bölümde değerlendirilen uzamsal özelliklerin yanı sıra HTG özelliklerinin işaret dili videolarındaki zamansal bilgileri ne kadar etkili bir şekilde yakalayabileceğini keşfetme hedefinin bir parçasıdır.



Şekil 4.18. HTG tabanlı verilerin ResNet-18 ile sınıflandırılmasının sonuçları (BosphorusSign22k-genel veri kümesi) (Akdağ ve Baykan, 2024)

Şekil 4.18'de sunulan bulgular, RReLU aktivasyon fonksiyonu kullanılarak eğitilen ResNet-18 modellerinin, ReLU aktivasyon fonksiyonu ile eğitilen modellere kıyasla HTG verileri için daha üstün doğruluk oranları sergilediğini göstermektedir. Elde edilen bulgular, RReLU'nun modelin öğrenme kapasitesini artırarak daha genelleştirilebilir ve tanınabilir özellikler elde ettiğinin göstergesidir ve bu durum poz görüntülerinin analizinde gözlemlenen sonuçları desteklemektedir.

Sonuçlar, VÜCUT-POZU-HTG-1, VÜCUT-POZU-HTG-3, EL-POZU-HTG-1, EL-POZU-HTG-3 ve PARMAK-POZU-HTG dahil olmak üzere poz tabanlı HTG verilerinin, işlenmemiş video görüntülerinden türetilen HTG verilerine, yani HAM-HTG-1 ve HAM-HTG-3 'e kıyasla daha yüksek doğruluk oranları sergilediğini göstermektedir. İDT bağlamında poz tabanlı HTG verilerinin kullanımıyla ilişkili potansiyel faydalar bu

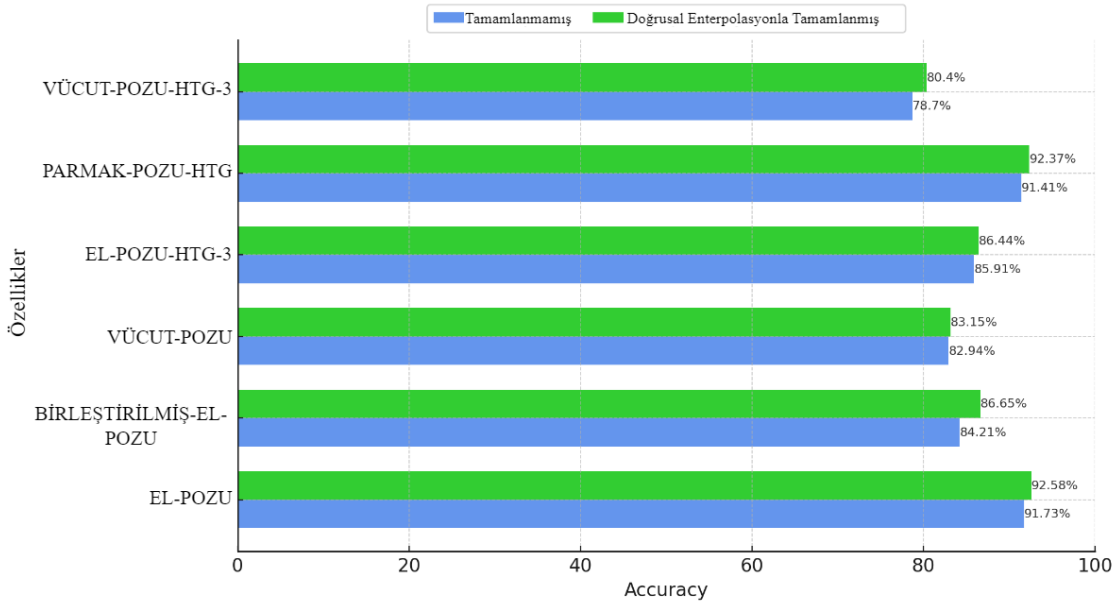
çalışmanın bulgularıyla gösterilmiştir. Ayrıca bu araştırma, tek kanallı ile üç kanallı HTG verileri arasında karşılaştırmalı bir analiz yapma fırsatı sunmaktadır. Bulgulara dayanarak, üç kanallı HTG (HTG-3) kullanılmasının, tek kanallı HTG (HTG-1) uygulanmasına kıyasla daha yüksek bir tanıma oranı verdiği sonucuna varılmıştır. HTG-1 tekniği tüm video süresi boyunca tüm hareketi kapsayan tek bir temsil oluştururken, HTG-3 stratejisi zamansal aralığı üç farklı aşamaya böler. Böylece, işaret dilinin sınıflandırılması ve analizi için mevcut verilerin ayrıntı düzeyi artırılmış olur.

PARMAK-POZU-HTG, incelenen diğer verilere kıyasla %91,41'lik bir tanıma oranıyla en yüksek doğruluğa ulaşmıştır. Bu üstün performans, PARMAK-POZU-HTG'nin işaret dili kelimeleri için ayırt edici bilgi sağlama kapasitesinden kaynaklanmaktadır. İşaret dilinin iletişim sürecinde parmak pozisyonlarının ve hareketlerinin rolü son derece önemlidir. Bu durum, parmak hareketlerinin ve pozisyonlarının doğru bir şekilde tanınması ve sınıflandırılmasının genel İDT performansını önemli ölçüde artırabileceğini göstermektedir.

Deneysel sonuçlardan yola çıkarak, çalışmayı takip eden bölümlerde, önerilen İDT sistemi için RReLU aktivasyon fonksiyonu ile geliştirilmiş ResNet-18 modeli kullanılmıştır. Bununla birlikte görsel poz verileri ve verilerden elde edilen HTG-3 ve önerilen PARMAK-POZU-HTG verileri kullanılmıştır.

4.2.3.3. Eksik pozların doğrusal enterpolasyon ile tamamlanmasının etkisi

MediaPipe Holistic'in performansı genel olarak tatmin edici olsa da hızlı el hareketleri veya oklüzyon nedeniyle bazı poz işaret noktaları doğru bir şekilde tahmin edilemeyebilir. Bu tür eksikliklerin üstesinden gelmek için, el pozlarına dayalı özellik çıkarımı gerçekleştirilirken doğrusal enterpolasyon kullanarak eksik poz işaret noktaları tamamlanmıştır. Tamamlanan poz işaret noktalarına dayalı sınıflandırma sonuçları Şekil 4.19'da sunulmuştur. Ayrıca doğrusal enterpolasyonun etkinliğini değerlendirmek için eksik poz verileriyle elde edilen performans sonuçları da aynı şekilde sunulmuştur.



Şekil 4.19. Eksik pozların doğrusal enterpolasyon ile tamamlanmasının etkisi (BosphorusSign22k-genel veri kümesi) (Akdağ ve Baykan, 2024)

Şekil 4.19'daki sonuçlara göre, doğrusal enterpolasyon kullanılarak eksik poz verilerinin tamamlanmasının sınıflandırma performansında genel bir artışa yol açtığı görülebilir. Eksik veriler, özellikle dinamik hareketleri ve hızlı değişimleri analiz ederken modelin başarısını olumsuz etkileyebilir. Bu nedenle, doğrusal enterpolasyon kullanarak eksik poz işaret noktalarını bulmak, modelin bu tür hareketleri daha doğru bir şekilde sınıflandırmasını sağlar. Ayrıca sonuçlar, doğrusal enterpolasyon gibi basit ve hesaplama açısından verimli yöntemlerin bile eksik veri sorununu çözmede ne kadar etkili olabileceğini ortaya koymaktadır. Bu tarz yöntemler, özellikle büyük veri kümeleri ve karmaşık model yapılarıyla çalışırken önemlidir, çünkü hem zamandan hem de kaynaklardan tasarruf sağlayabilir.

4.2.3.4. El özelliklerinin füzyonu ve sınıflandırma sonuçları

Bu alt bölümde, EL-POZU, BİRLEŞTİRİLMİŞ-EL-POZU, EL-POZU-HTG-3 ve PARMAK-POZU-HTG kullanılarak eğitilen ResNet-18 modellerinin düzleştirilmiş katmanlarından çıkarılan özellikler çeşitli kombinasyonlarda birleştirilip DVM ile sınıflandırılmıştır. Bu işlemlerin sonuçları Çizelge 4.11'de sunulmuştur. Bu yaklaşım, el pozu verileri ve HTG'ler aracılığıyla yakalanan dinamik hareketlerin birleştirilmesiyle, daha sağlam bir özellik kümesi sağlayarak İDT doğruluğunu potansiyel olarak artırabileceği hipotezine dayanmaktadır. Böyle bir füzyona duyulan ihtiyaç, zengin

çeşitlilikteki el hareketlerini doğru bir şekilde yorumlamak için genellikle çok yönlü bir özellik seti gerektiren işaret dilinin karmaşık doğasından kaynaklanmaktadır.

Çizelge 4.11. El özelliklerinin farklı kombinasyonlarda birleştirilmesi ve DVM ile sınıflandırma sonuçları (BosphorusSign22k-genel veri kümesi) (Akdağ ve Baykan, 2024)

EL-POZU	EL-POZU-HTG-3	BİRLEŞTİRİLMİŞ-EL-POZU	PARMAK-POZU-HTG	Doğruluk (%)	Kesinlik (%)	Duyarlılık (%)	F1 skoru (%)
✓	✓			92,83	92,24	93,22	91,85
✓		✓		94,09	93,42	93,48	92,68
✓			✓	93,78	93,10	93,96	92,65
	✓	✓		93,46	92,68	93,17	91,97
	✓		✓	92,93	92,20	93,09	91,73
		✓	✓	94,09	93,42	93,54	92,64
✓	✓	✓		94,62	93,92	94,25	93,26
✓	✓		✓	94,41	93,73	94,45	93,38
✓		✓	✓	94,73	94,09	94,47	93,47
	✓	✓	✓	94,83	94,19	94,92	93,70
✓	✓	✓	✓	95,36	94,77	95,44	94,29

Çizelge 4.11, farklı el özellikleri kombinasyonlarının yalnızca sınıflandırma doğruluğu değil, aynı zamanda kesinlik, duyarlılık ve F1 skoru üzerindeki etkisini de göstermektedir. Bazı el özellikleri tek başına yüksek sınıflandırma doğruluğu gösterirken, bu özelliklerin birleştirilmesi genel performansın yükseltilmesinde çok önemli bir rol oynamaktadır. Özellikle, EL-POZU ve EL-POZU-HTG-3'ün birleştirilmesiyle elde edilen %92,83'lük doğruluk, BİRLEŞTİRİLMİŞ-EL-POZU'nun eklenmesiyle %94,62'ye önemli ölçüde artmıştır. Bununla birlikte, PARMAK-POZU-HTG'lerin dahil edilmesinin performansı daha da iyileştirmede belirleyici olduğu kanıtlanmış ve %95,36'lık bir sınıflandırma doğruluğunun yanı sıra kesinlik, duyarlılık ve F1 skorunda karşılık gelen artışlarla sonuçlanmıştır. Bu gelişme, PARMAK-POZU-HTG'lerin diğer el özellikleriyle entegre edilmesinin önemini vurgulamakta, sınıflandırma doğruluğunu ve modelin genel sağlamlığını belirgin bir şekilde artırmaktadır.

El özelliklerinin derinlemesine incelenmesi, İDT görevinde büyük önem taşımaktadır. Bu araştırmadan elde edilen bulgular, ele dair özelliklerin ve bunların çeşitli kombinasyonlarının, işaret dili sınıflandırmasında kullanılan algoritmaların doğruluğu, kesinliği, duyarlılığı ve F1 skoru üzerinde önemli bir etkiye sahip olduğu fikrini desteklemekte ve etkili yorumlama için kapsamlı bir özellik setinin önemini vurgulamaktadır.

4.2.3.5. Manuel olmayan özelliklerin sınıflandırma performansı üzerindeki etkisi

İşaret dilinde sadece ellerin pozisyonu ve hareketleri değil, vücut ve yüz hareketleri de iletişimin önemli bir parçasıdır. İletişim alanında, yüz ifadeleri ve beden dili sadece tamamlayıcı değil, anlamların iletilmesinde ve kavramların kategorize edilmesinde çok önemli bir rol oynamaktadır. Bu çalışma el özelliklerinin vücut ve yüz ifadeleriyle entegrasyonunu ele almakta ve bu kombinasyonların sınıflandırmanın doğruluğu üzerindeki etkisini incelemektedir. El özelliklerinin birleştirilmesinden elde edilen sonuçlar bir önceki bölümde tartışılmıştır. Burada, VÜCUT-ÖZELLİKLERİ ve YÜZ-ÖZELLİKLERİ için analiz gerçekleştirilecektir. VÜCUT-ÖZELLİKLERİ, VÜCUT-POZU ve VÜCUT-POZU-HTG-3 özelliklerinin birleştirilmesiyle, YÜZ-ÖZELLİKLERİ ise YÜZ-POZU ve YÜZ-POZU-HTG-3 özelliklerinin birleştirilmesiyle elde edilmektedir. Ayrıca bir önceki bölümde tartışılan, ele dair tüm özelliklerin birleştirilmesiyle oluşsan özellik kümesi de EL-ÖZELLİKLERİ olarak adlandırılmıştır.

Çizelge 4.12, BosphorusSign22k-genel veri kümesinde manuel olmayan özelliklerin EL-ÖZELLİKLERİ ile füzyonunu ve DVM kullanılarak sınıflandırma sonuçlarını göstermektedir. Çizelgede farklı özellik kombinasyonları için test doğruluğu, kesinlik, duyarlılık ve F1 skor değerleri yer almaktadır.

Çizelge 4.12. Manuel olmayan özelliklerin el özellikleriyle birleştirilmesi ve DVM ile sınıflandırma sonuçları (BosphorusSign22k-genel veri kümesi) (Akdağ ve Baykan, 2024)

EL-ÖZELLİKLERİ	VÜCUT-ÖZELLİKLERİ	YÜZ-ÖZELLİKLERİ	Doğruluk (%)	Kesinlik (%)	Duyarlılık (%)	F1 skoru (%)
✓			95,36	94,77	95,44	94,29
	✓		88,61	87,96	90,70	87,54
		✓	14,33	14,38	18,98	13,18
✓	✓		95,89	95,30	96,30	95,07
✓		✓	95,57	94,98	95,99	94,63
	✓	✓	86,93	86,09	90,01	85,69
✓	✓	✓	96,94	96,43	96,99	96,31

Çizelgenin analizi, tek başına EL-ÖZELLİKLERİ için sınıflandırma doğruluğunun %95,36 olduğunu ve buna karşılık gelen kesinlik, duyarlılık ve F1 skor değerlerinin sırasıyla %94,77, %95,44 ve %94,29 olduğunu ortaya koymaktadır. VÜCUT-ÖZELLİKLERİ ve YÜZ-ÖZELLİKLERİ bağımsız olarak kullanıldığında, doğrulukları %88,61 ve %14,33'tür; EL ve VÜCUT özelliklerinin birleştirilmesi doğruluğu %95,89'a çıkarır. Bu kombinasyon, %95,57 doğruluk, %94,98 kesinlik, %95,99 duyarlılık ve %94,63 F1 puanı ile sonuçlanan EL ve YÜZ özellik

kombinasyonuna kıyasla biraz daha yüksek doğruluk, kesinlik, duyarlılık ve F1 skoru elde eder. Yalnızca VÜCUT ve YÜZ özelliklerinin kullanılması %86,93'lük bir doğruluk sağlar ki bu, EL-ÖZELLİKLERİ kullanılmadan elde edilen en yüksek doğruluktur. En yüksek sınıflandırma doğruluğu olan %96,94, %96,43 kesinlik, %96,99 duyarlılık ve %96,31 F1 skoru, EL, VÜCUT ve YÜZ özelliklerinin tümünün birleştirilmesiyle elde edilmiştir.

EL-ÖZELLİKLERİ'nin VÜCUT ve YÜZ özellikleriyle birleştirilmesi, işaret dili sınıflandırmasında daha yüksek bir doğruluk elde etmek için önemlidir. Bununla birlikte, YÜZ-ÖZELLİKLERİ'nin tek başına tanıma doğruluğu düşüktür ancak diğer özelliklerle birleştirildiğinde performansı artırmıştır, bu da yüz hareketlerinin bağımsız olarak kullanılmak yerine diğer özelliklerle birleştirildiğinde daha etkili olduğunu göstermektedir.

4.2.3.6. Diğer çalışmalarla karşılaştırma

Bu alt bölümde, önerilen İDT yönteminin performansı diğer çalışmalarla karşılaştırılmıştır. Karşılaştırmalar, PARMAK-POZU-HTG ile elde edilen sonuçları ve tüm özelliklerin birleştirilmesiyle elde edilen nihai yöntemi içermektedir.

Çizelge 4.13. Çalışma-2 kapsamında önerilen yöntemin BosphorusSign22k-genel veri kümesi üzerindeki literatür sonuçlarıyla karşılaştırılması

Çalışmalar	Girdi modalitesi	Odak Alanları	Test Doğruluğu (%)
Kindiroglu ve diğ. (2019)	Poz	Tam çerçeve	81,58
Gündüz ve Polat (2021)	RGB, poz, optik akış	Vücut, el, yüz	89,35
Çalışma-2 (önerilen yöntem)	HTG	Parmak	92,37
	Poz, HTG	Vücut, el, yüz, parmak	96,94

Çizelge 4.13, BosphorusSign22k-genel veri kümesi üzerinde gerçekleştirilen deneysel çalışmanın literatürdeki benzer çalışmalarla karşılaştırmasını sunmaktadır. Önerilen PARMAK-POZU-HTG, Kindiroglu ve diğ. (2019) ve Gündüz ve Polat (2021)'ın yöntemlerine kıyasla %92,37'lik bir doğrulukla öne çıkmaktadır. Özellikle, Gündüz ve Polat (2021) tarafından sunulan vücut, el, yüz, optik akış ve poz verilerini kullanan Inception3D ve UKSB-TSA gibi modellere dayalı karmaşık sistemle karşılaştırıldığında, PARMAK-POZU-HTG kullanılarak eğitilen nispeten daha az karmaşık olan ResNet-18 modeli %4,02'lik bir doğruluk artışı sağlamıştır. Inception3D gibi hem uzamsal hem de zamansal özellikleri değerlendiren bir modele karşı elde edilen bu sonuç, basit modellerle bile, uygun özellik mühendisliği ve verimli algoritma tasarımı

ile yüksek performans elde edilebileceğini göstermektedir. Ayrıca, sistemdeki vücut, el ve yüze dair özelliklerin entegrasyonu, BosphorusSign22k-genel veri kümesinde %96,94'lük önemli bir test doğruluğu elde etmiştir. Bu, önerilen yöntemin İDT için mevcut literatürdeki diğer yöntemlere göre önemli bir gelişme sağladığını göstermektedir.

Çizelge 4.14. Çalışma-2 kapsamında önerilen yöntemin BosphorusSign22k veri kümesi üzerindeki literatür sonuçlarıyla karşılaştırılması

Çalışmalar	Girdi Modalitesi	Odak Alanları	Test Doğruluğu (%)
Özdemir ve diğ. (2020)	RGB	Tam çerçeve	78,85
Özdemir ve diğ. (2020)	RGB	Tam çerçeve	88,53
Gökçe ve diğ. (2020)	RGB	Vücut, el, yüz	94,94
Sincan ve Keles (2021)	RGB, RGB-HTG	Tam çerçeve	94,83
Kındıroglu ve diğ. (2023)	RGB, poz	Tam çerçeve	94,90
Özdemir ve diğ. (2023)	RGB, poz	Vücut, el, yüz	92,58
Çalışma-2 (önerilen yöntem)	HTG	Parmak	86,10
	Poz, HTG	Vücut, el, yüz, parmak	94,87

Çizelge 4.14, BosphorusSign22k veri kümesine uygulanan yöntemlerin karşılaştırmasını içermektedir. Bu veri kümesi işaret dili tanıma araştırmaları için zengin bir içeriğe sahiptir ve 744 farklı sınıfı kapsamaktadır. Bu veri kümesi üzerinde de 174 sınıflı BosphorusSign22k-genel veri kümesindeki yüksek sonuçların tekrarlanması, yöntemin etkinliğini ve ölçeklenebilirliğini göstermektedir. Sonuçlara bakıldığında, PARMAK-POZU-HTG'nin ResNet-18 modeliyle elde edilen %86,10 doğrulukla, Özdemir ve diğ. (2023) tarafından MC3 modelini kullanarak ham video verilerinden hem zamansal hem de uzamsal özellikleri çıkaran yaklaşıma kıyasla %7,25 daha yüksek bir test doğruluğu sağladığı görülmüştür. Bu sonuçlar, işaret dili tanımada derin öğrenme modellerinin karmaşıklığının yanı sıra doğru özellik seçiminin önemini bir kez daha vurgulamaktadır. Sincan ve Keles (2021), ham video görüntüler ve bu görüntülerden elde edilen 3 Kanallı Hareket Tarihçesi Görüntülerine (RGB-HTG veya HTG-3) dayalı olarak geliştirdikleri sistemde ise I3D ve ResNet-50 modellerini kullanarak %94,83 tanıma oranı elde etmiştir. Tez kapsamında önerilen yöntemde ise, görsel poz verileri ve bunlardan elde edilen HTG-3 verilerini kullanarak %94,87 oranında bir tanıma doğruluğu sağlanmış, farklı olarak, önerilen yöntemde, Sincan ve Keles (2021) tarafından kullanılan I3D modelinin aksine, uzamsal ve zamansal özellikleri aynı anda öğrenebilen bir model kullanılmamış, bunun yerine, zamansal bilginin çıkarımı doğrudan HTG verilerine bırakılmıştır. Elde edilen sonuçlar, işaret dili tanımada HTG verileriyle zamansal özelliklerin başarılı bir şekilde yakalanabildiğini göstermiştir. BosphorusSign22k veri

kümesi üzerinde elde en yüksek doğruluk oranı ise Gökçe ve diğ. (2020) tarafından yapılan çalışmadır. Bu çalışma, zamansal ve uzamsal bilgileri işleyebilen güçlü bir önceden eğitilmiş model olan MC3-18 modelinin kullanımının yanı sıra vücut, el ve yüz gibi çeşitli bileşenlerin entegrasyonunu da içermektedir. Bu çalışmaya benzer bir yaklaşım izleyerek önerilen entegre yöntem, MC3-18'den daha az karmaşık bir model olan ResNet-18 ile vücut, el, yüz ve parmak özelliklerini birleştirerek %94,87 doğrulukla rekabetçi bir sonuç elde etmiştir. Elde edilen sonuç, zamansal özelliklerin her bir poz verisinden çıkarılan üç kanallı HTG özellikleri tarafından önemli ölçüde yakalanabileceğini göstermektedir. Ayrıca, Gökçe ve diğ. (2020)'nin çalışmasında modelleri eğitmek için RGB verilerinin kullanılması ve veri kümesinin düz bir arka plan üzerine inşa edilmiş olması, sistemin farklı arka plan koşullarına karşı potansiyel olarak hassas olabileceğini düşündürmektedir. Arka planların çeşitliliği göz önüne alındığında, bu gerçek dünya uygulamalarında önemli bir sınırlama olabilir. Önerilen yöntem tamamen poz verilerine dayanmaktadır ve bu sayede sitemin arka plan değişikliklerine karşı daha dayanıklı olması beklenmektedir. Bu bağlamda, poz tahmin yönteminin başarısı, sistemin genel performansı için belirleyici bir faktördür. Etkili poz tahmini sağlandığında, İDT sistemi arka plandaki değişimlerden bağımsız olarak işaret dili bileşenlerini doğru bir şekilde tanıyabilir ve yorumlayabilir.

Çizelge 4.15. Çalışma-2 kapsamında önerilen yöntemin LSA64 veri kümesi üzerindeki literatür sonuçlarıyla karşılaştırılması

	Çalışmalar	Girdi modalitesi	Odak Alanları	Test Doğruluğu (%)
D1	Marais ve diğ. (2022b)	RGB	Tam çerçeve	74,22
	Çalışma-2 (önerilen yöntem)	HTG	Parmak	94,68
		Poz, HTG	Vücut, el, yüz, parmak	98,43
D2	Ronchetti ve diğ. (2016a)	RGB	El	91,70
	Rodríguez ve Martínez (2018)	RGB	Tam çerçeve	85
	Alyami ve diğ. (2023)	Poz	El, yüz	91,09
	Çalışma-2 (önerilen yöntem)	HTG	Parmak	94,99
		Poz, HTG	Vücut, el, yüz, parmak	98,68
D3	Ronchetti ve diğ. (2016a)	RGB	El	97
	Konstantinidis ve diğ. (2018b)	Poz	Vücut, el	98,09
	Konstantinidis ve diğ. (2018a)	RGB, poz, optik akış	Vücut, el, yüz	99,84
	Zhang ve Li (2019)	RGB	Tam çerçeve	99,063
	Imran ve Raman (2020)	HTG, DG, RGB-HG	Tam çerçeve	97,81
	Marais ve diğ. (2022a)	RGB	Tam çerçeve	95,50
	Bohacek ve Hruz (2022)	Poz	Vücut, el	100
	Alyami ve diğ. (2023)	Poz	El, yüz	98,25
	Çalışma-2 (önerilen yöntem)	HTG	Parmak	98,90
		Poz, HTG	Vücut, el, yüz, parmak	100

Çizelge 4.15, önerilen yöntemin farklı işaret dillerinde uygulanabilirliğini doğrulamak için 64 sınıftan oluşan LSA64 veri kümesi üzerinde gerçekleştirilen deneylerin sonuçlarını göstermektedir. Bu çizelge, veri kümesinin üç farklı eğitim ve test koşuluna bölündüğü deneylerin sonuçlarına odaklanmaktadır. D1 ve D2 işaretçiden bağımsız değerlendirmeleri gösterirken, D3 eğitim ve test kümesinin rastgele belirlendiği durumdaki sonuçları göstermektedir. Beş ve on numaralı işaretçilerin test verisi olarak ayrıldığı D1 deney setindeki değerlendirmede, Marais ve diğ. (2022b) tarafından elde edilen %74,22'lik doğruluk oranı, PARMAK-POZU-HTG ile %94,68'e, nihai yöntemle ise %98,43'e yükselmektedir. 10 işaretçinin her birinin ayrı bir test olarak kullanıldığı D2'de deneyler 10 kez tekrarlanarak ortalama doğruluklar elde edilmiş; PARMAK-POZU-HTG ve nihai yöntem ile elde edilen sonuçlar diğer tüm çalışmaları geride bırakmıştır. Bu analize dahil edilen en son çalışmalardan biri olan Alyami ve diğ. (2023)'nin işaretçinin el ve yüzünden alınan işaret noktalarının dönüştürücü modeli ile sınıflandırıldıldığı çalışma, PARMAK-POZU-HTG kullanılarak eğitilen ResNet-18 ile elde edilen sonucunun gerisinde kalmıştır. D3 için Bohacek ve Hruz (2022) tarafından elde edilen %100 doğruluk oranı önerilen yöntemlerle eşdeğerdir ve benzer şekilde her iki çalışmada da poz verileri kullanılmıştır. Bu da işaret dili tanımada poz verilerinin etkinliğini bir kez daha kanıtlamaktadır. D3 için, önerilen PARMAK-POZU-HTG ile %98,90 doğruluk elde edilirken, tüm özelliklerin birleştirildiği sistem ile %100 doğruluk elde edilmiştir. Genel olarak, bu ayrıntılı karşılaştırma, önerilen yöntemin hem işaretçiden bağımsız hem de rastgele bölünmüş test koşullarında işaret dili tanıma alanında rekabetçi bir performans sergilediğini ve hatta mevcut çalışmalardan daha iyi performans gösterdiğini göstermektedir.

Çizelge 4.16. Çalışma-2 kapsamında önerilen yöntemin GSL veri kümesi üzerindeki literatür sonuçlarıyla karşılaştırılması

Çalışmalar	Girdi modalitesi	Odak Alanları	Test Doğruluğu (%)
Adaloglou ve diğ. (2020)	RGB	Tam çerçeve	86,03
Adaloglou ve diğ. (2020)	RGB	Tam çerçeve	86,23
Adaloglou ve diğ. (2020)	RGB	Tam çerçeve	89,74
Selvaraj ve diğ. (2021)	Poz	Tam çerçeve	86,60
Selvaraj ve diğ. (2021)	Poz	Tam çerçeve	89,50
Selvaraj ve diğ. (2021)	Poz	Tam çerçeve	93,50
Selvaraj ve diğ. (2021)	Poz	Tam çerçeve	95,40
Fang ve diğ. (2023)	RGB, poz	Tam çerçeve	91,49
Çalışma-2 (önerilen yöntem)	HTG	Parmak	88,55
	Poz, HTG	Vücut, el, yüz, parmak	95,14

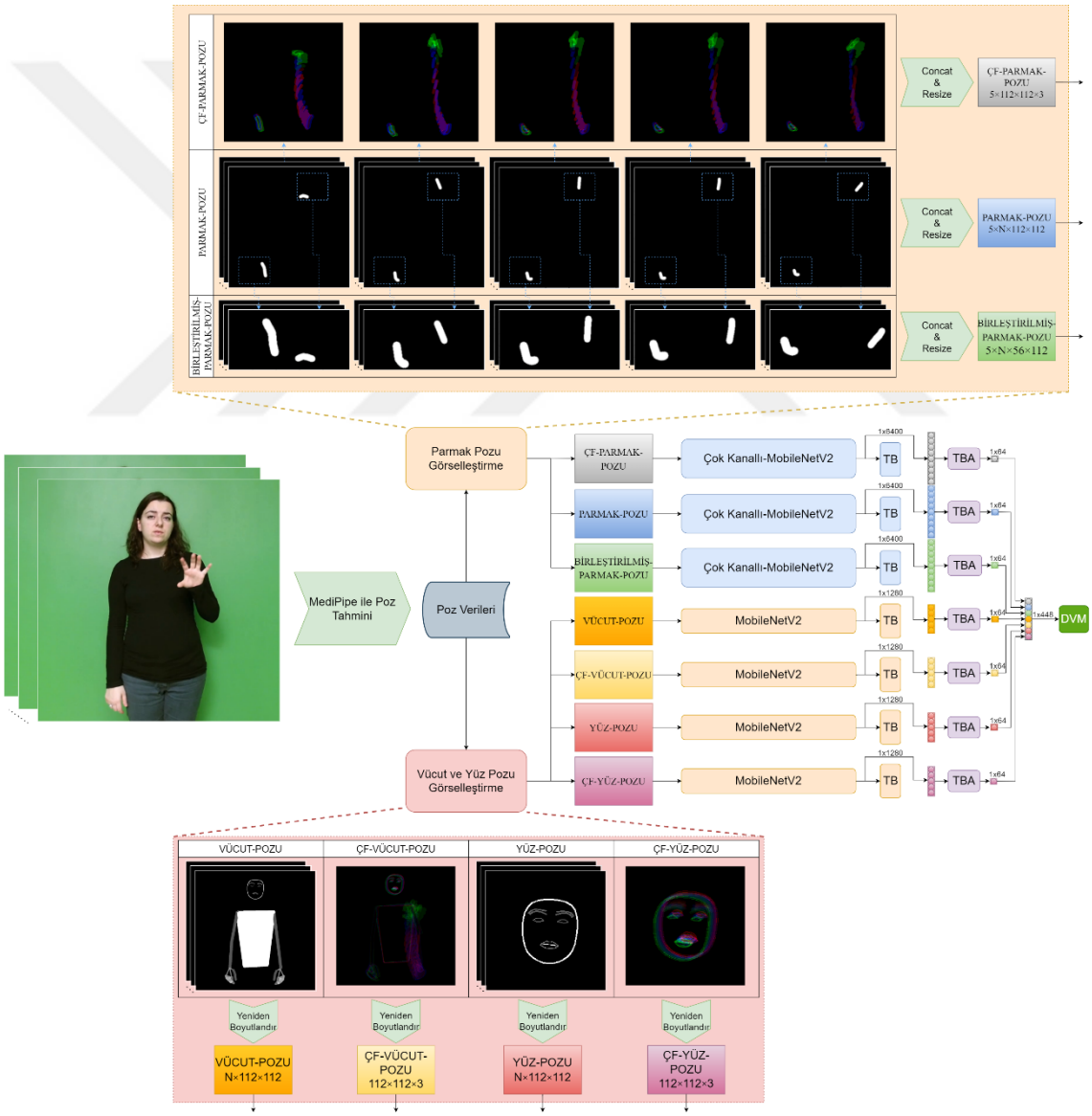
Önerilen yöntemin farklı işaret dillerine uygulanabilirliğini doğrulamak için kullanılan bir diğer veri kümesi de 310 farklı Yunan İşaret Dili kelimesi içeren GSL veri kümesidir. Çizelge 4.16 GSL veri kümesi üzerinde yapılan çalışmaları göstermektedir. Aynı yazar isimleri, aynı makaledeki farklı yöntemlerin sonuçlarını ifade etmektedir. GSL veri kümesi üzerinde yapılan deneylerde, önerilen PARMAK-POZU-HTG, özellikle son zamanlarda yapılan bazı çalışmalarla karşılaştırıldığında benzer bir doğruluk sağlamaktadır. Bununla birlikte, önerilen nihai yöntem, literatürdeki neredeyse tüm çalışmalardan daha yüksek bir doğruluk oranıyla öne çıkmaktadır. Sonuç olarak, GSL veri kümesi üzerinde gerçekleştirilen deneyler, önerilen yöntemlerin etkili olduğunu ve literatürdeki diğer yaklaşımlarla rekabet edebildiğini göstermektedir.

Özetle, Çizelge 4.13'ten Çizelge 4.16'ya kadar olan veriler incelendiğinde, çalışma-2 kapsamında önerilen yöntemlerin çok çeşitli veri kümeleri üzerinde literatürdeki yöntemlere kıyasla oldukça rekabetçi ve bazı durumlarda daha üstün sonuçlar sağladığı görülmüştür. Özellikle PARMAK-POZU-HTG çeşitli veri kümeleri üzerindeki performansıyla öne çıkmaktadır, bazı veri kümelerinde literatürdeki yöntemlere benzer hatta daha iyi sonuçlar elde etmiştir. Bu performans PARMAK-POZU-HTG'nin işaret dili sınıflandırmasında ne kadar değerli ve etkili olduğunu göstermektedir. Tüm özelliklerin füzyonunu içeren ve sınıflandırma için bir DVM kullanan yöntemden elde edilen doğruluk oranları ise, mevcut literatürde bildirilen en yüksek sonuçlarla karşılaştırılabilir veya bu sonuçları aşmaktadır. Bu durum, bu çalışmada kullanılan özellik çıkarma ve birleştirme yaklaşımının işaret dili sınıflandırma problemini ele almada oldukça etkili olduğunu göstermektedir. Ayrıca, deneysel sonuçların tutarlılığı yani kullanılan metodolojilerin farklı veri kümelerinde gösterdiği yüksek performans, önerilen yöntemin güvenilirlik ve sağlamlığa sahip olduğunu belirtmektedir.

4.3. Çalışma 3: Parmak Özelliklerine Dayalı Çok Akışlı İşaret Dili Tanıma

Tez kapsamında yapılan çalışmalarda önerilen parmak pozu tabanlı hareket tarihçesi görüntüsünün başarısına dayanarak, bu çalışmada, İDT görevinde parmakların özelliklerine ve konfigürasyonlarına daha çok odaklanan yenilikçi çok kanallı bir yaklaşım sunulmaktadır. Bu yaklaşımın temeli, MediaPipe Holistic kullanılarak elde edilen ve ayrı kanallarda işlenen parmak pozu verilerinden türetilen üç farklı veriye dayanmaktadır. Bu çok kanallı veriler kullanılarak, parmak hareketlerinin ayrıntılı bir

analizini sağlamak amacıyla Çok Kanallı-MobileNetV2 modeli önerilmiştir. Çalışmada, eğitilen tüm modellerden çıkarılan özellikler öncelikle Temel Bileşen Analizi (TBA) kullanılarak boyut indirgemeye tabi tutulmuş, daha sonra, bu işlenmiş özellikler birleştirilerek DVM ile sınıflandırılmıştır. Ardından, önerilen yöntem, her biri ayrı ayrı eğitilen MobileNetV2 modellerinden elde edilen tam vücut ve yüz verilerinin özellikleri dahil edilerek genişletilmiş ve daha sağlam hale getirilmiştir. Bu çalışma "Multi-Stream Isolated Sign Language Recognition Based on Finger Features Derived from Pose Data" (Akdag ve Baykan, 2024b) isimli makalede sunulmuş, önerilen yöntemin genel diyagramı ise Şekil 4.20'de gösterilmiştir.

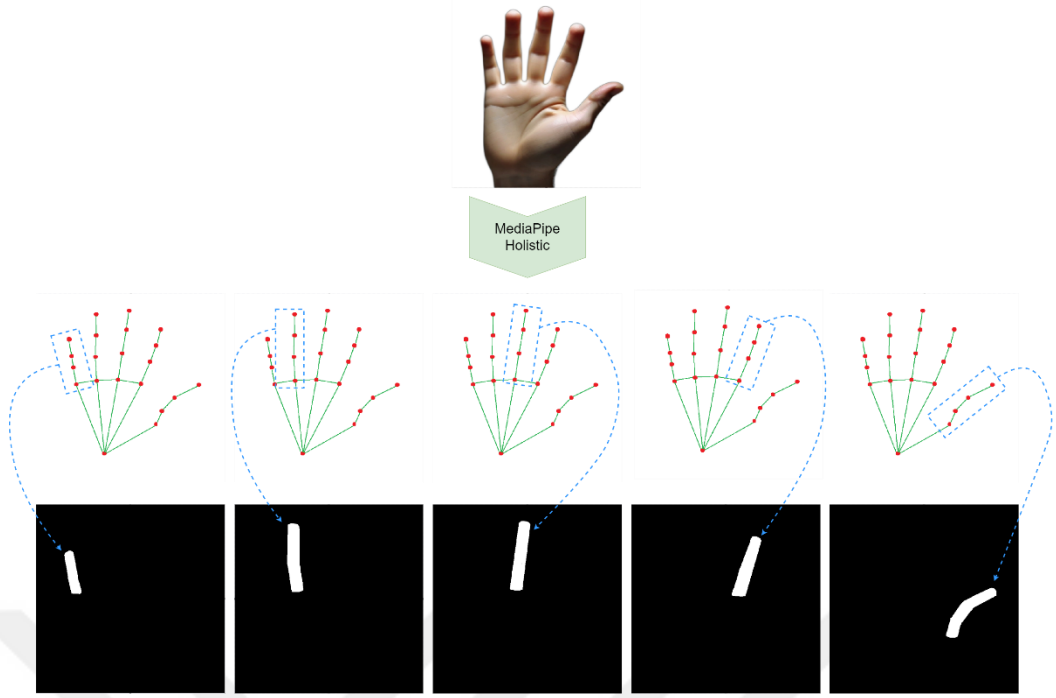


Şekil 4.20. Çalışma-3 için önerilen sistemin diyagramı (Akdag ve Baykan, 2024b)

Bu bölümün devamında, önerilen sistemin tasarımında kullanılan veri ön işleme teknikleri ve önerilen Çok Kanallı-MobileNetV2 modelinden sınıflandırma sonuçlarına kadar olan süreçler anlatılacaktır.

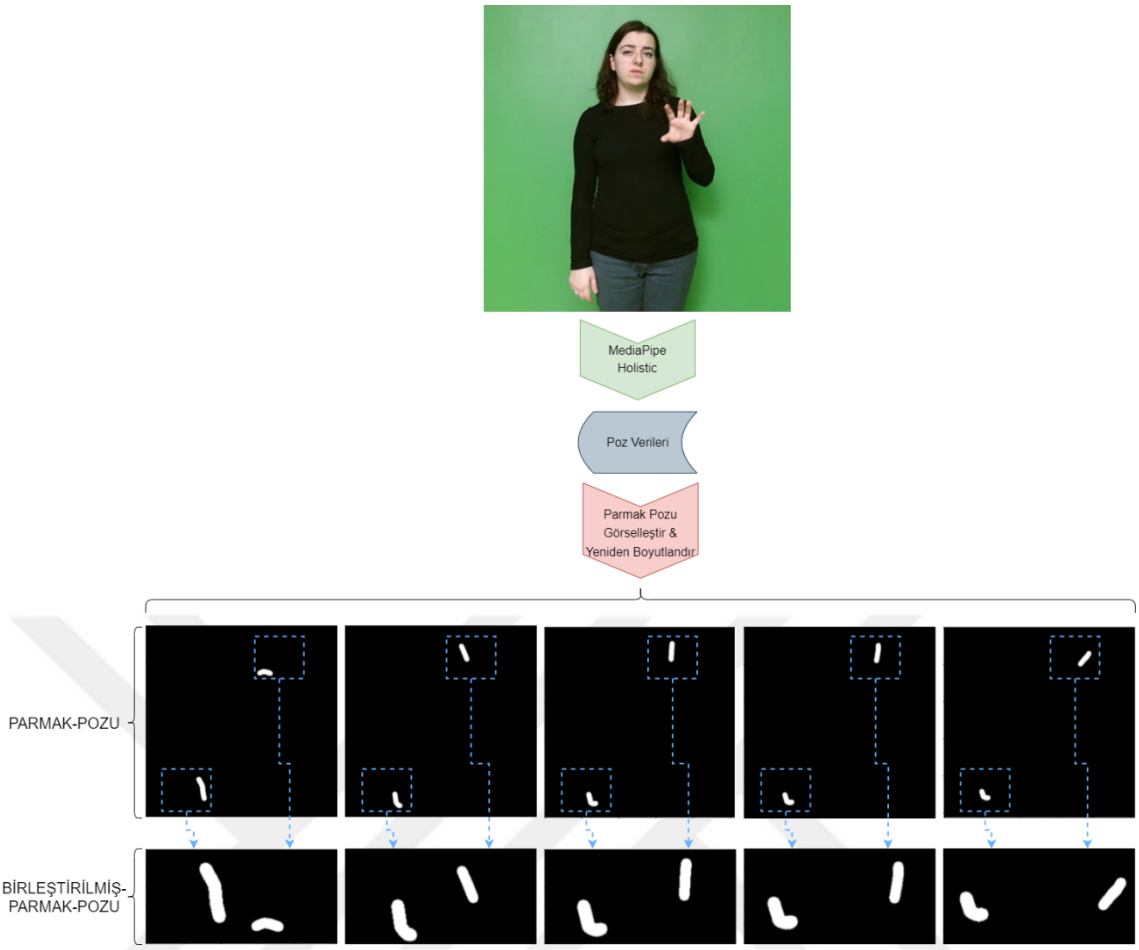
4.3.1. Veri ön işleme

Bu çalışmada da parmak pozu görüntülerini elde etmek için MediaPipe Holistic çözümü kullanılmıştır. Öncelikle poz işaret noktaları çıkarılmış, ardından eksik işaret noktası verilerini tamamlamak için doğrusal enterpolasyon yöntemi uygulanmıştır. Daha sonra, poz işaret noktaları Bölüm 3.4'te anlatılan yöntemle görsel poz verilerine dönüştürülmüştür. Bu dönüştürme işlemi, parmaklara karşılık gelen işaret noktaları kullanılarak, her parmağın görüntüsü ayrı bir kanalda tasvir edilecek şekilde gerçekleştirilmiştir. Elde edilen görüntüler parmakların konumlarını, hareketlerini ve birbirleriyle olan ilişkilerini ayrıntılı olarak göstermektedir. Bu ayrıntılı görselleştirme, parmak hareketlerini ve konumlarını incelemeyi kolaylaştırarak daha doğru İDT için gerekli ince ayrıntıları ortaya çıkarmıştır. El içeren bir görüntüden Parmak Pozu görüntülerinin elde edilmesine ilişkin bir gösterim Şekil 4.21'de sunulmuştur. Bu şekilde, ilk olarak MediaPipe Holistic ile elde edilen poz işaret noktaları temsili olarak gösterilmiştir. Ardından, bu poz işaret noktalarını kullanarak her bir parmağın siyah arka plan üzerinde beyaz olarak ayrıntılı bir şekilde oluşturulmasını gösteren görseller sunulmuştur.



Şekil 4.21. Bir el görüntüsünden PARMAK-POZU görüntüsünün elde edilmesi (Akdag ve Baykan, 2024b)

İşaret dili videosundan, parmak pozu görüntülerini içeren çerçevelerin oluşturulmasının ardından, parmak görüntülerini içeren bölgeler kırpılmış ve birleştirilmiştir. PARMAK-POZU görüntülerine kıyasla, parmakların daha yakın ve detaylı bir görselini sağlayan ve BİRLEŞTİRİLMİŞ-PARMAK-POZU olarak adlandırılan bu görüntülerin işaret dilindeki karmaşık ve ince parmak hareketlerinin analizini kolaylaştırması beklenmektedir. Bir işaret dili video karesinden PARMAK-POZU ve BİRLEŞTİRİLMİŞ-PARMAK-POZU görüntü verilerinin oluşturulmasına ilişkin bir görselleştirme Şekil 4.22'de gösterilmektedir.

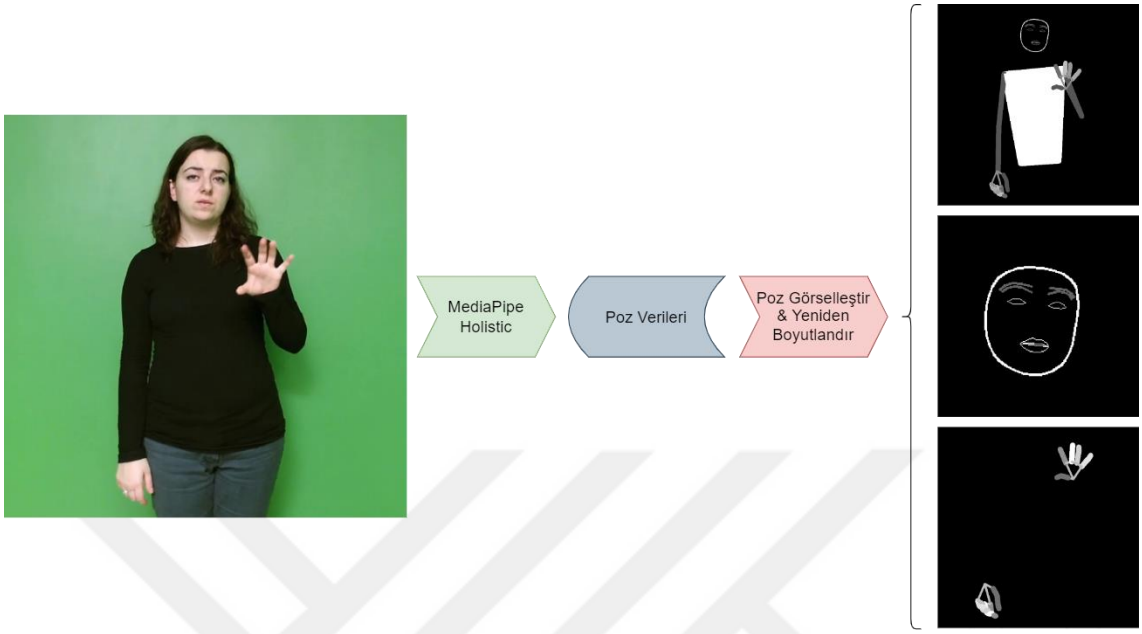


Şekil 4.22. Bir işaret dili video karesinden PARMAK-POZU ve BİRLEŞTİRİLMİŞ-PARMAK-POZU görüntülerinin oluşturulması (Akdag ve Baykan, 2024b)

PARMAK-POZU görüntülerinin ayrı kanallarda gösterilme stratejisi, parmakların birbirine çok yakın olduğu veya üst üste bindiği durumlarda bile her parmağın hareketini ve konumunu bağımsız olarak izlemeyi mümkün kılar. Bu sürecin sonunda, elde edilen bu çok kanallı görüntü verilerinin yükseklik ve genişlikleri, Çok Kanallı-MobilNetV2 modellerini eğitmek üzere sırasıyla 112×112 ve 56×112 boyutlarına yeniden boyutlandırılmıştır. Sonuç olarak N×5×112×112 ve N×5×56×112 boyutunda veriler oluşturulmuştur. Burada N toplam çerçeve sayısını, 5 ise kanal sayısını ifade eder.

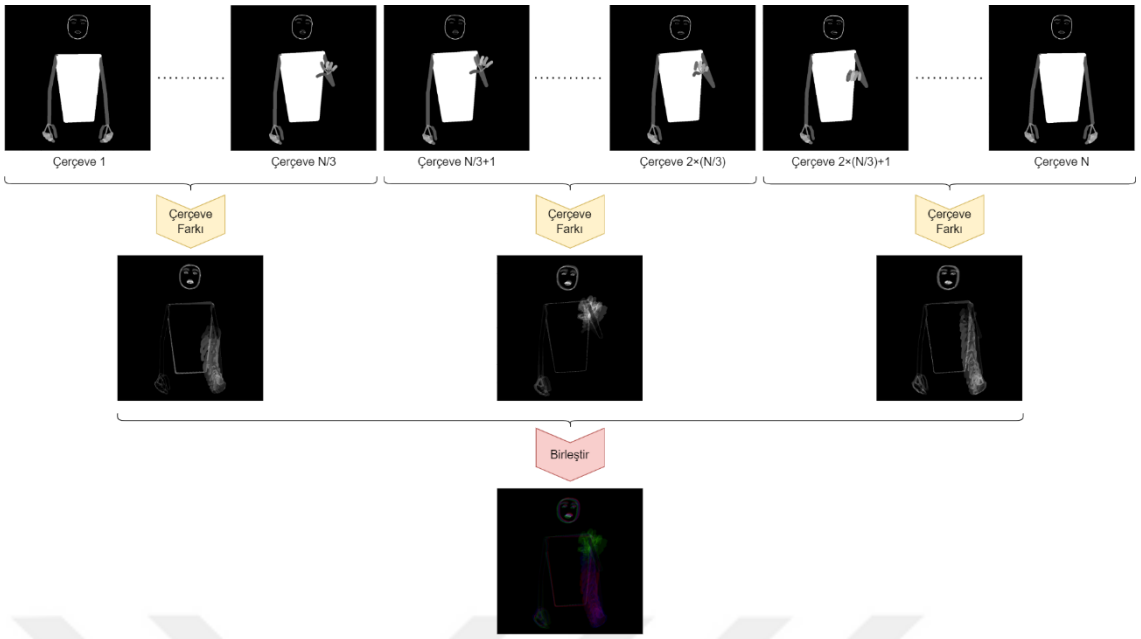
Sağlam bir İDT sistemi, manuel ve manuel olmayan özelliklerin birlikte değerlendirilmesini gerektirdiğinden dolayı, parmağa dair verilere ek olarak VÜCUT-POZU ve YÜZ-POZU da önerilen İDT sistemine dahil edilmiştir. PARMAK-POZU'nda, her bir parmağın ayrı bir kanalda temsil edilmesinin etkisini incelemek için EL-POZU görüntüleri de oluşturulmuştur. Bu görüntü verileri daha sonra 112×112 olarak yeniden boyutlandırılmıştır. Nihayetinde N, toplam çerçeve sayısı olmak üzere N×112×112

boyutunda görüntü verileri oluşturulmuştur. Örnek bir video karesinden oluşturulan bu görüntüler Şekil 4.23'te gösterilmiştir.



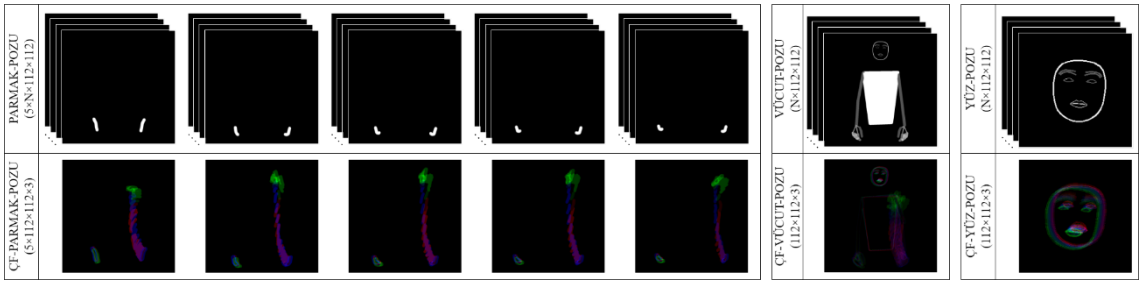
Şekil 4.23. Bir işaret dili video karesinden VÜCUT-POZU, YÜZ-POZU ve EL-POZU verilerinin oluşturulması (Akdag ve Baykan, 2024b)

Oluşturulan poz görüntü verilerinin ardından, bu verilerden zamansal bilgiyi elde edebilmek için her bir poz görüntü verisine Bölüm 3.5.2'de anlatılan üç kanallı Çerçeve Farkı (ÇF) Yöntemi uygulanmıştır. Çerçeve Farkı Yöntemi PARMAK-POZU'ndaki her bir kanal için uygulanmış ve tek bir PARMAK-POZU görüntü verisinden 5×112×112×3 boyutunda ÇF-PARMAK-POZU görüntüsü elde edilmiştir. Benzer şekilde, Çerçeve Farkı Yöntemi VÜCUT-POZU ve YÜZ-POZU verilerine de uygulanarak 112×112×3 boyutunda ÇF-VÜCUT-POZU ve ÇF-YÜZ-POZU görüntüleri oluşturulmuştur. Bu görüntüler parmak, vücut ve yüz hareketlerinden çıkarılan zamansal dinamikleri kapsayarak, zaman içindeki işaret dili hareketlerinin anlaşılmasını sağlar. ÇF-VÜCUT-POZU görüntülerinin VÜCUT-POZU görüntü verilerinden nasıl türetildiğine dair bir görselleştirme Şekil 4.24'te sunulmuştur. Bu gösterimde, N sayıda çerçeve içeren bir video örneğinin üç eşit parçaya bölündüğü gösterilmiştir. Çerçeve Farkı Yöntemi her bir parçaya ayrı ayrı uygulanır. Ardından, bu parçaların sonuçları birleştirilerek kompozit bir RGB görsel temsil oluşturulur.



Şekil 4.24. VÜCUT-POZU verilerinden ÇF-VÜCUT-POZU verilerinin oluşturulması (N: çerçeve sayısı) (Akdag ve Baykan, 2024b)

Çerçeve Farkı Yöntemiyle elde edilen zamansal bilgi temsili, İDT sistemi için oldukça önemlidir, çünkü hareketlerin sadece statik anlık görüntülerini değil, zaman içindeki dinamiklerini anlamaya olanak tanır. Şekil 4.25'te bu yöntemle elde edilen veri örnekleri gösterilmektedir; üstte PARMAK-POZU, VÜCUT-POZU ve YÜZ-POZU verileri, altta ise bunlara karşılık gelen çerçeve farkı versiyonları olan ÇF-PARMAK-POZU, ÇF-VÜCUT-POZU ve ÇF-YÜZ-POZU gösterilmektedir.

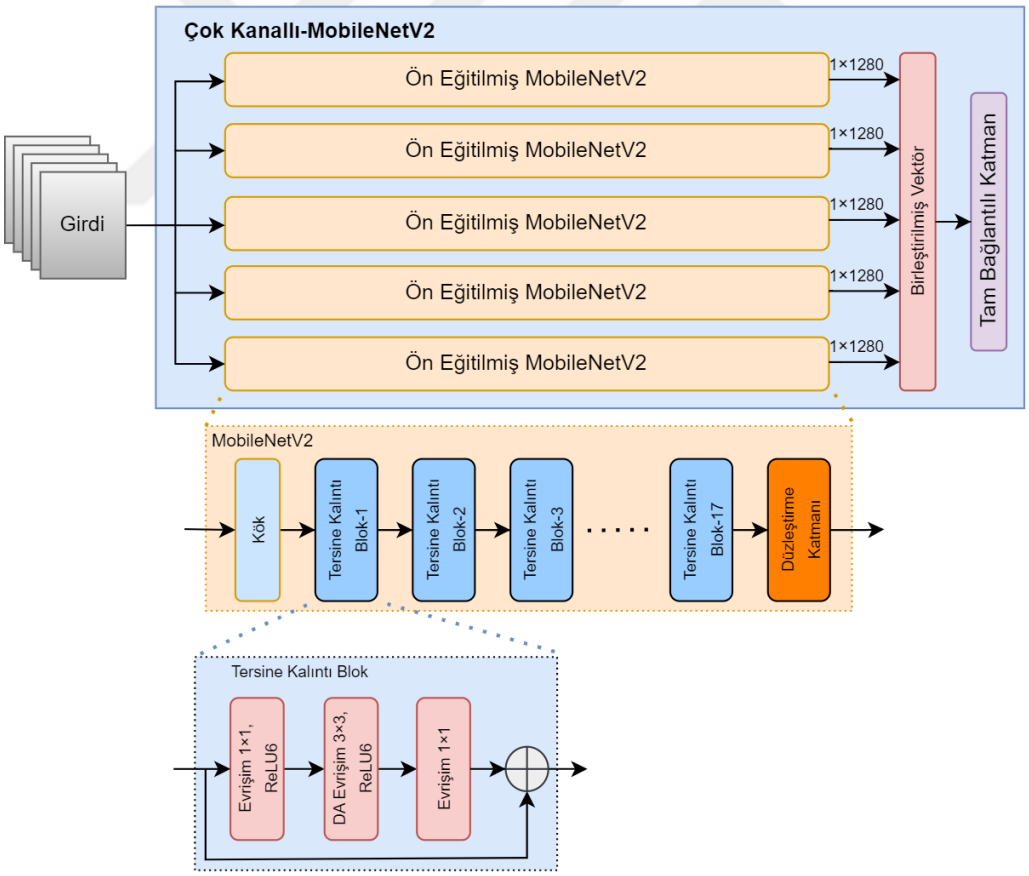


Şekil 4.25. ÇF-PARMAK-POZU, ÇF-VÜCUT-POZU ve ÇF-YÜZ-POZU görüntüleri PARMAK-POZU, VÜCUT-POZU ve YÜZ-POZU görüntü verilerinden oluşmaktadır (N: çerçeve sayısı) (Akdag ve Baykan, 2024b)

4.3.2. Önerilen ESA modeli

Bu çalışma kapsamında, derin öğrenme alanında büyük önem taşıyan MobileNetV2 mimarisinin yenilikçi kanallı bir uyarlaması olan Çok Kanallı-

MobileNetV2 modeli önerilmiştir. Sandler ve diğ. (2018) tarafından geliştirilen MobileNetV2, Derinlemesine Ayrılabilir (DA) evrişimler, Ters Kalıntı Blok (Inverted Residual Block) yapıları ve doğrusal darboğazlar gibi özelliklere sahip, mobil ve gömülü sistemler için uygun, hafif ve yüksek performanslı bir yapı sunmaktadır. Bu mimarinin temel avantajları arasında, düşük parametre sayısı, yüksek hesaplama verimliliği ve küçük model boyutu bulunmaktadır. Bu özellikler, MobileNetV2'yi çok kanallı bir yapıda kullanmak için ideal bir aday yapar. Çok Kanallı-MobileNetV2, farklı veri kanallarını ayrı ayrı işlemek için önceden eğitilmiş beş ayrı MobileNetV2 ağını birleştirecek şekilde inşa edilmiştir. Önceden eğitilmiş modellerin kullanılması, modelin eğitim sürecini hızlandırmakta ve kapsamlı veri kümeleri üzerinde eğitilmiş özellikleri tanıma kabiliyetini artırmaktadır. Önerilen Çok Kanallı-MobileNetV2 mimarisi, onun bir bileşeni olan MobileNetV2 mimarisi ve yine onun bir alt bileşeni olan Ters Kalıntı Blok yapısı Şekil 4.26'da gösterilmiştir.



Şekil 4.26. Önerilen Çok Kanallı-MobileNetV2 mimarisi ve alt bileşenleri (Akdag ve Baykan, 2024b)

Çok Kanallı-MobileNetV2 mimarisiyle elde edilen yaklaşım, her kanaldaki verilerin ayrı ve ayrıntılı analizine olanak tanır. Her bir alt model, modelin ilk katmanında yapılan özelleştirmelerle farklı bir kanalı işleyecek şekilde yapılandırılmıştır. İlk evrişim katmanının giriş boyutunun değiştirilmesini içeren bu yapılandırma, her modelin PARMAK-POZU ve BİRLEŞTİRİLMİŞ-PARMAK-POZU verileri için 32 kanallı girişleri ve ÇF-PARMAK-POZU verileri için 3 kanallı girişleri kabul etmesini sağlar, böylece farklı veri türlerinin işlenmesini kolaylaştırır.

MobileNetV2'nin Tam Bağlantılı Katmanından önceki düzleştirme katmanından 1280 özellik elde edilir. Çok Kanallı-MobileNetV2'te yer alan beş alt ağın düzleştirme katmanlarından elde edilen çıktılar tek bir çıktı vektörü (5×1280=6400) oluşturmak üzere birleştirilir. Birleştirilmiş vektör daha sonra modelin nihai çıktısını üretmek için Tam Bağlantılı bir katman tarafından işlenir; bu çıktı belirtilen sınıf sayısına göre boyutlandırılır ve belirli bir sınıflandırma görevi için kullanılabilir.

Bu çalışmada ayrıca, VÜCUT-POZU, ÇF-VÜCUT-POZU, YÜZ-POZU, ÇF-YÜZ-POZU ve EL-POZU görüntü verileri kullanılarak MobileNetV2 modelleri eğitilmiştir. VÜCUT-POZU, YÜZ-POZU ve EL-POZU verilerini kullanan modeller için, bu veriler eğitim sırasında rastgele 32 kare seçilerek kullanıldığından, ilk katmandaki evrişim işlemi 32 giriş boyutuna sahip olacak şekilde değiştirilmiştir. ÇF-VÜCUT-POZU ve ÇF-YÜZ-POZU verileri için modellerde ek bir değişiklik yapılmamıştır. Ayrıca, çalışmada kullanılan tüm derin öğrenme modellerinin çıkış katmanları, eğitim için kullanılan veri kümelerindeki sınıf sayısına göre yeniden yapılandırılmıştır.

4.3.3. Min-Max normalleştirme, TBA ile boyut azaltma

Her bir Çok Kanallı-MobileNetV2 modeli, PARMAK-POZU, BİRLEŞTİRİLMİŞ-PARMAK-POZU ve ÇF-PARMAK-POZU verileri üzerinde eğitildiğinde 6400 boyutlu bir özellik vektörü üretir. Bu özellikler sınıflandırma için birleştirildiğinde, her üç kaynaktan gelen verilerin füzyonuyla, toplam özellik boyutu üç katına çıkar. Bu büyük özellik seti sadece sınıflandırma ve analiz süreçlerindeki hesaplama yükünü artırmakla kalmaz, aynı zamanda boyutluluk laneti olarak adlandırılan modelin eğitim ve genelleme yeteneği üzerinde de olumsuz bir etkiye sahip olabilir (Altman ve Krzywinski, 2018). Temel Bileşen Analizi (TBA) temel özellikleri korurken veri kümesinin boyutunu anlamlı bir şekilde azaltmak için ideal bir yöntemdir (Aremu ve diğ., 2020). TBA, istatistiksel metodolojide etkili bir araç olarak işlev görür ve yüksek

boyutlu verilerin karmaşıklıklarını ele almak için özel olarak tasarlanmıştır. Bu teknik, mevcut özellikleri temel bileşenler olarak bilinen farklı bir değişkenler grubuna dönüştürerek çalışır. Tipik olarak, en önde gelen temel bileşenler en bilgilendirici olanlardır ve önemli bilgilerin büyük kısmını kapsarlar. Bu temel bileşenlerin ele alınmasıyla en kritik özellikler etkili bir şekilde korunurken verilerin boyutunda önemli bir azalma sağlanmış olur (Ringnér, 2008). Bu çalışmada, TBA farklı sayıda temel bileşenle uygulanarak elde edilen doğruluk oranları karşılaştırılmıştır. Böylece model performansını en üst düzeye çıkaracak en uygun temel bileşen sayısı belirlenmiştir.

TBA yöntemi kullanılmadan önce, özellik veri kümesine min-max normalizasyonu uygulanmıştır. Bu normalleştirme işlemi, veri kümesindeki her bir özelliği belirli bir aralıkta ölçeklendiren bir yöntemdir. Genellikle bu aralık [0, 1]'dir. Min-max normalizasyonu için formül Denklem 4.8'de ifade edilmiştir.

$$x_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{4.8}$$

Bu denklemde X_{norm} normalleştirme sonrası değeri, X orijinal değeri, X_{min} ve X_{max} ise sırasıyla özelliğin minimum ve maksimum değerlerini temsil etmektedir. Bu işlem, tüm özelliklerin aynı ölçekte olmasını sağlayarak TBA'nın daha etkili bir şekilde uygulanmasına katkıda bulunur.

Normalleştirme işlemi sonrası TBA kullanılarak özellik boyutunun azaltılmasının ardından, azaltılmış özellik setleri her veri örneği için birleşik bir özellik vektörü oluşturmak üzere birleştirilmiştir. Bu kapsamlı özellik vektörü hem uzamsal hem de zamansal dinamikleri etkili bir şekilde yakalayarak işaret dili hareketlerinin temel özelliklerini kapsar. Birleştirilmiş özellik vektörü daha sonra sınıflandırma adımında DVM için girdi olarak kullanılmıştır.

4.3.4. Deneysel sonuçlar

Bu çalışmada, derin öğrenme modelleri için başlangıç öğrenme oranı 0,001 olarak ayarlanmış ve her 15 epokta bir 0,1 oranında azaltılmıştır. Optimizasyon için, dalgalanmaları en aza indiren ve daha hızlı yakınsamayı destekleyen, 0,9 momentum değeri ile Stokastik Gradyan İnişi (SGD) kullanılmıştır. Daha iyi bellek kullanımı ve genelleme için mini parti boyutu 8 olarak belirlenmiş ve modeller toplam 35 epok

süresince eğitilmiştir. Ayrıca, -10 ile +10 derece arası rastgele rotasyon uygulanarak ve video verilerinden rastgele 32 kare seçilerek veri çeşitliliği artırılmıştır.

4.3.4.1. Çok Kanallı-MobileNetV2 modelinin sınıflandırma sonuçları

Bu çalışmanın ilk aşamasında, elde edilen PARMAK-POZU, BİRLEŞTİRİLMİŞ-PARMAK-POZU ve ÇF-PARMAK-POZU verileri kullanılarak Çok Kanallı-MobileNetV2 modelleri eğitilmiştir. Ek olarak, parmaklara dair poz görüntülerini ayrı kanallarda görselleştirmenin avantajlarını değerlendirmek için geleneksel bir MobileNetV2 modeli EL-POZU verileri kullanılarak eğitilmiştir. Bu modellerden elde edilen test analizlerinin performans metrikleri Çizelge 4.17'de listelenmiştir.

Çizelge 4.17. Parmak pozu ve el pozu görüntü verileri üzerinde eğitilen MultiChannel-MobileNetV2 ve MobileNetV2 modellerinin performans metrikleri (BosphorusSign22k-genel veri kümesi)

Model	Veri	Doğruluk (%)	Kesinlik (%)	Duyarlılık (%)	F1 skoru (%)
MultiChannel-MobileNetV2	PARMAK-POZU	94,52	93,94	94,49	93,40
MultiChannel-MobileNetV2	BİRL.-PARMAK-POZU	90,83	90,00	91,49	89,14
MultiChannel-MobileNetV2	ÇF-PARMAK-POZU	90,83	90,26	91,67	89,73
MobileNetV2	EL-POZU	90,41	89,70	91,18	88,90

Çok Kanallı-MobileNetV2 modeli, PARMAK-POZU verileriyle eğitildiğinde kayda değer bir başarı göstermiştir. Model %93,94'lük bir kesinlik ve %94,49'luk bir duyarlılık oranı sergilemiştir. Bu sonuçlar %93,40'lık F1 skoru ile birleştiğinde modelin tutarlılığını ve güvenilirliğini göstermektedir. Ayrıca, %94,52'lik doğruluk oranı modelin genel doğruluğunu vurgulamakta ve İDT'de parmak hareketlerinin ayrıntılı incelenmesinin öneminin altını çizmektedir.

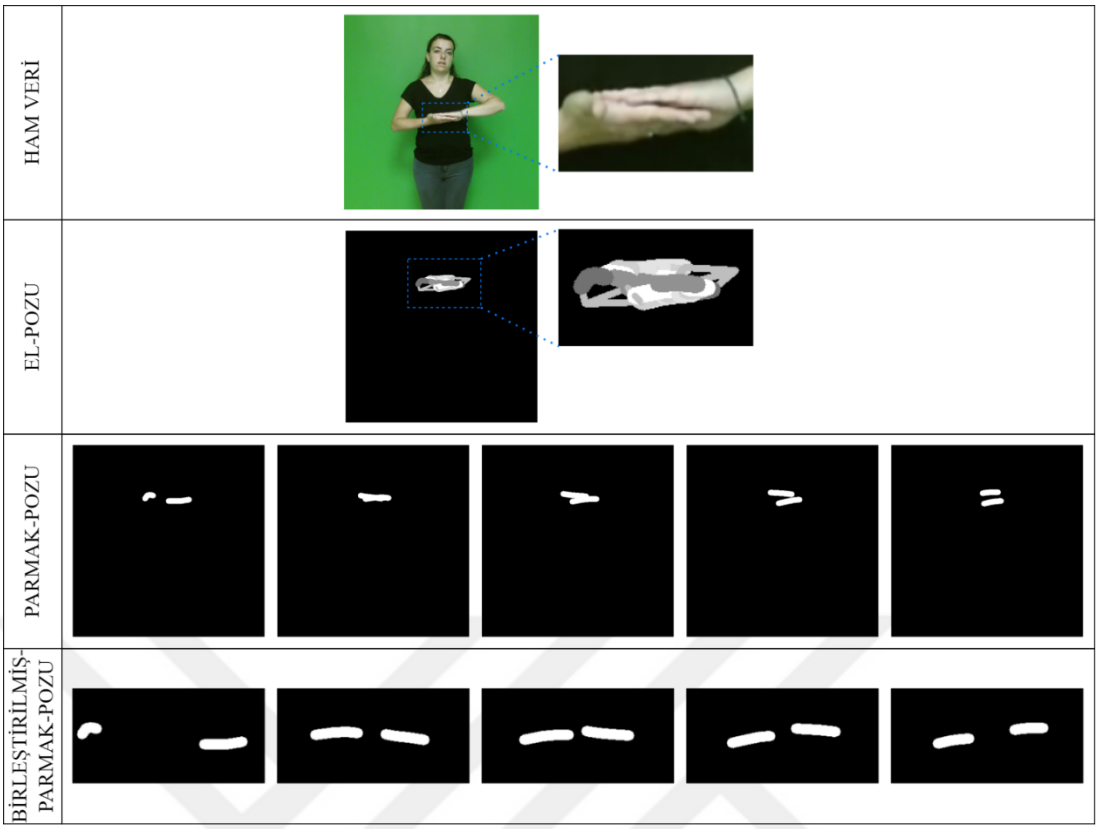
PARMAK-POZU'ndaki parmakları içeren bölgelerin kırpılmasıyla elde edilen BİRLEŞTİRİLMİŞ-PARMAK-POZU verisiyle eğitilen model, %90,00 kesinlik, %91,49 duyarlılık, %89,14 F1 skoru ve %90,83 doğruluk ile PARMAK-POZU verisine kıyasla biraz daha düşük değerler elde etmiştir. Bu durum, BİRLEŞTİRİLMİŞ-PARMAK-POZU'nda parmakların tam görüntü içindeki konumsal bilgilerinin kaybolmasına bağlanabilir. Bununla birlikte, %90,83 gibi yüksek doğruluk oranı, bu yaklaşımın İDT için değerli bir perspektif sunduğunu ve parmakların daha yakından incelenmesinin bazı ayrıntı kayıplarına rağmen hala önemli bilgiler sağladığını göstermektedir.

Çerçeve Farkı Yöntemini kullanarak, PARMAK-POZU verilerinden hareketin zamansal değişimlerini yakalayan ÇF-PARMAK-POZU verileriyle eğitilen model,

%90,26 kesinlik, %91,67 duyarlılık, %89,73 F1 skoru ve %90,83 doğruluk ile hareketin dinamiklerini başarılı bir şekilde yakalamıştır. Elde edilen yüksek doğruluk oranları, İDT'de zamansal bilginin kritik rolünü göstermektedir.

Son olarak, EL-POZU verileriyle eğitilen geleneksel MobileNetV2 modeli %90,41'lik doğruluk oranıyla diğer modellere kıyasla biraz daha düşük bir performans göstermiştir. Bu sonuç, parmakların detaylı analizinin İDT'de genel EL-POZU görüntülerine kıyasla daha fazla bilgi sağladığını ve daha etkili olduğunu göstermektedir.

Şekil 4.27, parmakların üst üste geldiği durumları görselleştirerek İDT'de karşılaşılan yaygın bir zorluğu göstermektedir. Parmakların üst üste binmesi, özellikle parmaklar birbirini tam olarak örttüğünde işaretlerin tanınmasını zorlaştırabilir. Bu sorun, poz bilgisi yoluyla elde edilen EL-POZU görüntü verileri için de geçerlidir. Ancak, her bir parmak konumunun ayrı kanallarda görselleştirilmesiyle elde edilen PARMAK-POZU görüntü verisi, örtüşme sorununa etkili bir çözüm sunmaktadır. Ayrıca, BİRLEŞTİRİLMİŞ-PARMAK-POZU verileri sadece parmak görüntülerinin daha yakından incelenmesini sağlamakla kalmaz, aynı zamanda her iki elin parmaklarını belirgin bir şekilde ayırarak ayrıntılı bir analize olanak tanır. Bu görsel içgörüler ve ilgili verilerden elde edilen sonuçlar, İDT'de parmak örtüşmesi gibi zorlukların üstesinden gelmede önerilen yöntemin etkinliğini ortaya koymaktadır. Her bir parmağı ayrı kanallarda işleyen Çok Kanallı-MobileNetV2 modelinin sağladığı detaylı analiz de, İDT sistemlerinin doğruluğunu ve hassasiyetini önemli ölçüde artırmaktadır.



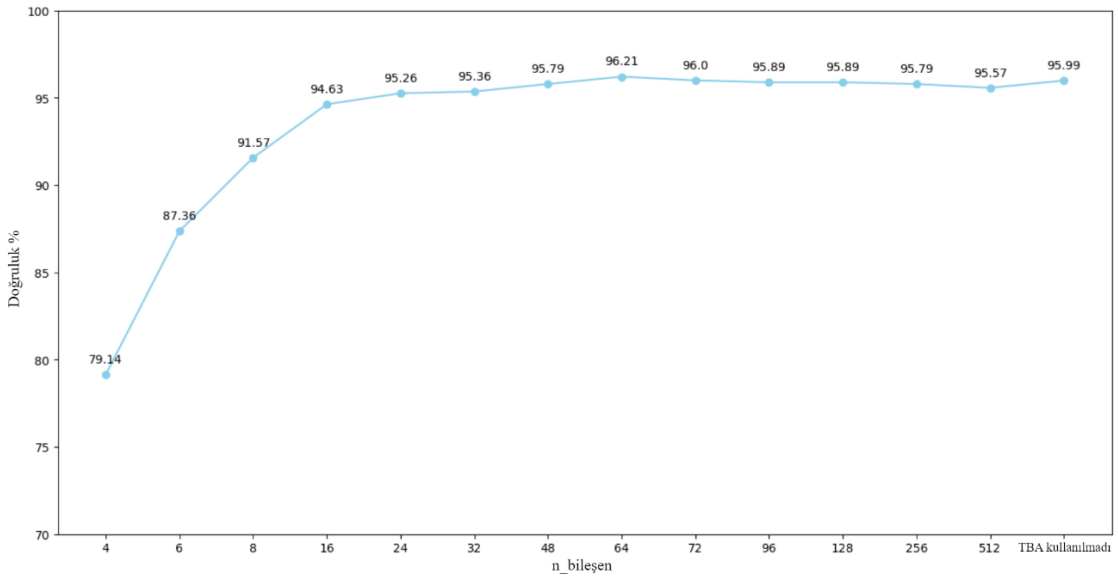
Şekil 4.27. Üst üste binen parmaklar sorununun gösterimi (Akdag ve Baykan, 2024b)

4.3.4.2. Parmak tabanlı özelliklerin füzyonu ve sınıflandırılması

Bu bölümde, eğitilmiş Çok Kanallı-MobileNetV2 modelleri aracılığıyla elde edilen özelliklerin birleştirilmesi ve bu birleşimin modelin genel doğruluğu üzerindeki etkisi araştırılmıştır. Çok Kanallı-MobileNetV2 modellerinden her biri, düzleştirme katmanı aracılığıyla her bir veri noktası için 6.400 boyutlu bir özellik vektörü üretir. Bu vektörler, verilerin karmaşık yapısını yansıtan zengin bir temsil sunmaktadır. PARMAK-POZU, BİRLEŞTİRİLMİŞ-PARMAK-POZU ve ÇF-PARMAK-POZU verilerinden elde edilen özellik vektörleri birleştirildiğinde (PARMAK-ÖZELLİKLERİ), veri noktası başına toplam 19.200 boyutlu bir özellik vektörü elde edilir. Ancak bu yüksek boyutluluk, özellikle hesaplama yükü ve veri yönetimi açısından zorluklara yol açabilir.

Yüksek boyutluluk sorununu çözmek için, her modelden elde edilen özellik kümelerine min-max normalizasyonu, ardından da TBA yöntemi uygulanmıştır. TBA yöntemiyle özellik vektörlerinin boyutu azaltılarak gereksiz bilgiler çıkarılmış ve verinin önemli yönleri korunmuştur. Her bir modele karşılık gelen, TBA sonrası indirgenmiş özellik setleri daha sonra birleştirilmiş ve DVM sınıflandırıcısı kullanılarak sınıflandırılmıştır. Ayrıca, en uygun bileşen sayısının belirlenmesi için bu süreç TBA'daki

farklı temel bileşen sayılarıyla tekrarlanmıştır. Bununla birlikte TBA'nın etkisini daha belirgin bir şekilde gözlemlemek için TBA uygulamadan da benzer bir sınıflandırma işlemi gerçekleştirilmiştir. TBA'da farklı temel bileşen sayıları ile yapılan deneysel çalışmanın sonuçları Şekil 4.28'deki grafikte gösterilmektedir.



Şekil 4.28. PARMAK-ÖZELLİKLERİ için önerilen yöntemin farklı temel bileşen sayıları için doğruluk değerleri (BosphorusSign22k-genel veri kümesi) (Akdag ve Baykan, 2024b)

Şekil 4.28'de yer alan grafikteki verilere göre, dört temel bileşen için %79,14 ile başlayan doğruluk oranı temel bileşen sayısı arttıkça önemli ölçüde artmıştır. Özellikle, 16, 24 ve 32 temel bileşen kullanıldığında, doğruluk oranlarında belirgin artışlar olmuştur (sırasıyla %94,63, %95,26 ve %95,36). Temel bileşen sayısı 64 olduğunda en iyi sınıflandırma sonucu elde edilmiştir. Ancak bu noktadan sonra temel bileşen sayısı arttırılmasına rağmen doğrulukta anlamlı bir değişme olmamıştır. Örneğin 72, 96 ve 128 temel bileşen ile elde edilen oranlar sırasıyla %96, %95,89 ve %95,89'dur. Bu durum, belirli bir noktanın ötesinde, ilave bileşenlerin modelin genel performansına katkıda bulunmadığını, hatta bu performansı azaltabileceğini göstermektedir. Bu analizden çıkarılan bir diğer sonuç ise TBA kullanılmadan elde edilen performansın (%95,99 doğruluk oranı) TBA ile elde edilen bazı sonuçlardan daha düşük olmasıdır. Bu durum, tüm özellik setinin kullanılmasının doğruluğu artırmadığını göstermektedir. Ancak, doğru temel bileşen sayısının, TBA'nın model performansını optimize edebileceği ve daha yüksek doğruluk oranlarına ulaşabileceği gösterilmiştir. Gerçekleştirilen deneysel analizlerde, optimum temel bileşen sayısı 64 olarak belirlenmiş ve bu sayı takip eden deneysel çalışmalarda kullanılmıştır.

PARMAK-POZU, BİRLEŞTİRİLMİŞ-PARMAK-POZU ve ÇF-PARMAK-POZU verileri parmak hareketlerini farklı açılardan yansıtmaktadır. Her biri, modelin sonuçları üzerindeki etkisini anlamak ve İDT sistemlerinin hassasiyetini ve verimliliğini artırmak için önemlidir. TBA yöntemi kullanılarak 64 temel bileşenle %96,21 doğruluk elde edilmesinin ardından, hangi verilerin sistemin performansını artırdığını ve hangilerinin daha az etkili olduğunu belirlemek için tüm olası özellik kombinasyonlarının füzyon işlemi gerçekleştirilmiştir. Bu analizin sonuçları Çizelge 4.18'de sunulmuştur.

Çizelge 4.18. Parmak özelliklerinin farklı kombinasyonlarda füzyonu ve sınıflandırmalarda sonuçları (BosphorusSign22k-genel veri kümesi) (Akdag ve Baykan, 2024b)

PARMAK-POZU	BİRLEŞTİRİLMİŞ-PARMAK-POZU	ÇF-PARMAK-POZU	Doğruluk (%)	Kesinlik (%)	Duyarlılık (%)	F1 skoru (%)
✓	✓		95,57	94,94	96,20	94,59
✓		✓	94,31	93,78	94,78	93,35
	✓	✓	95,57	94,98	95,89	94,67
✓	✓	✓	96,21	95,63	96,66	95,29

Çizelge 4.18'de sunulan sonuçlar, farklı özellik kombinasyonlarının önerilen İDT sisteminin performansı üzerindeki etkisini detaylandırmaktadır. Özellikle, tüm özellikler birleştirildiğinde %96,21'lik bir doğrulukla en yüksek performans elde edilmiştir. Bu durum, farklı özelliklerin entegre edilmesinin modelin hassasiyetini ve etkinliğini önemli ölçüde artırdığının altını çizmektedir. Özellik setlerinin bireysel performansları incelendiğinde, PARMAK-POZU verilerinin tek başına kullanımının %94,52'lik bir doğruluk oranıyla en yüksek performansı verdiği görülmektedir. Ancak bu bireysel performanslar, özellikler birleştirildiğinde elde edilen doğruluk oranlarından daha düşüktür. Örneğin, PARMAK-POZU ve BİRLEŞTİRİLMİŞ-PARMAK-POZU özelliklerinin birleştirilmesiyle elde edilen doğruluk oranı %95,57, her bir özellik türünün bağımsız olarak kullanılmasından daha yüksektir. Elde edilen bu sonuçlar, özelliklerin birleştirilmesinin sistemin genel performansını nasıl artırdığını ve her bir özellik kümesinin ayrı ayrı sağladığı bilgileri nasıl aştığını göstermektedir.

4.3.4.3. Manual olmayan özelliklerin sınıflandırma performansı

Sağlam bir İDT sistemi sadece parmakların değil, işareti oluşturan tüm bileşenlerin detaylı bir analizini kapsamalıdır. Bu nedenle çalışmanın devamında parmak özellikleri ile geliştirilen sisteme vücut ve yüz özellikleri de dahil edilmiştir. Öncelikle veri ön-işleme aşamasında elde edilen VÜCUT-POZU ve YÜZ-POZU verileri

kullanılmış, ek olarak, Çerçeve Farkı Yöntemi uygulanarak elde edilen ÇF-VÜCUT-POZU ve ÇF-YÜZ-POZU verileri de sisteme dahil edilmiştir. Tüm bu veri kümeleri, her biri için ayrı MobileNetV2 modellerinin eğitilmesinde kullanılmıştır. Eğitim sonucunda elde edilen performans metrik sonuçları Çizelge 4.19'da gösterilmiştir.

Çizelge 4.19. Vücut ve yüze dair veriler üzerinde eğitilen MobileNetV2 modellerinin performans metrikleri (BosphorusSign22k-genel veri kümesi için) (Akdag ve Baykan, 2024b)

Model	Data	Doğruluk (%)	Kesinlik (%)	Duyarlılık (%)	F1 skoru (%)
MobileNetV2	VÜCUT-POZU	80,08	79,69	82,81	78,44
MobileNetV2	ÇF-VÜCUT-POZU	85,14	84,40	86,33	83,23
MobileNetV2	YÜZ-POZU	15,70	16,32	16,91	13,04
MobileNetV2	ÇF-YÜZ-POZU	15,48	15,11	15,82	12,49

Çizelge 4.19, MobileNetV2 modelinin VÜCUT-POZU, ÇF-VÜCUT-POZU, YÜZ-POZU ve ÇF-YÜZ-POZU görüntü verileri üzerindeki performansını göstermektedir. Sonuçlara göre, VÜCUT-POZU ile model tarafından %80,08 doğruluk elde edilmiştir. VÜCUT-POZU'na Çerçeve Farkı Yönteminin uygulanmasıyla elde edilen ÇF-VÜCUT-POZU, %85,14'lük doğruluk oranıyla daha yüksek bir performans sergilemiştir. Bu sonuçlar, vücutla ilgili özelliklerin İDT görevinde önemli bileşenler olduğunu ve modellerin bu verileri etkili bir şekilde işleyebildiğini göstermektedir. Ayrıca, Çerçeve Farkı Yöntemiyle elde edilen ÇF-VÜCUT-POZU, hareket dinamikleri ve değişikliklerini yakalama konusunda iyi bir sonuç elde etmiştir. Öte yandan, YÜZ-POZU ve ÇF-YÜZ-POZU verilerinin performansı oldukça düşük kalmıştır (doğruluk oranları sırasıyla %15,70 ve %15,48). Ancak bu düşük skorlar, yüz özelliklerinin işareti bağımsız olarak temsil etmekten ziyade işaretin anlamını tamamlayıcı unsurlar olarak katkı sağlamasından kaynaklanmaktadır, diğer özelliklerle birleştirildiğinde tanıma doğruluğunda pozitif etki yapmıştır.

4.3.4.4. Tüm özelliklerin füzyonu ve sınıflandırılması

Çalışmanın bu aşamasında, eğitilmiş MobileNetV2 modellerinden elde edilen, VÜCUT-POZU, ÇF-VÜCUT-POZU, YÜZ-POZU ve ÇF-YÜZ-POZU verilerinden çıkarılan özelliklerin (1280 boyutlu) boyutu TBA yöntemi kullanılarak azaltılmıştır (n_bileşen = 64). Bu indirgenmiş özellikler daha sonra VÜCUT-ÖZELLİKLERİ (VÜCUT-POZU ve ÇF-VÜCUT-POZU verilerinden) ve YÜZ-ÖZELLİKLERİ (YÜZ-

POZU ve ÇF-YÜZ-POZU verilerinden) verilerini oluşturmak üzere birleştirilmiştir. Benzer şekilde PARMAK-ÖZELLİKLERİ'ni oluşturmak üzere eğitilmiş Çok Kanallı-MobileNetV2 modellerinden elde edilen özellikler, 64 temel bileşenli TBA ile indirgenmiş ve birleştirilmiştir. Daha sonra, VÜCUT-ÖZELLİKLERİ, YÜZ-ÖZELLİKLERİ, PARMAK-ÖZELLİKLERİ ve bunların olası tüm kombinasyonlarda birleştirilmesiyle elde edilen özellik kümeleri DVM ile sınıflandırılmıştır. Elde edilen sınıflandırma performansları Çizelge 4.20'de gösterilmiştir.

Çizelge 4.20. Tüm özelliklerin farklı kombinasyonlarda füzyonu ve sınıflandırma sonuçları (BosphorusSign22k-genel veri kümesi) (Akdag ve Baykan, 2024b)

PARMAK-ÖZELLİKLERİ	VÜCUT-ÖZELLİKLERİ	YÜZ-ÖZELLİKLERİ	Doğruluk (%)	Kesinlik (%)	Duyarlılık (%)	F1 skoru (%)
✓			96,21	95,63	96,66	95,29
	✓		87,88	87,38	90,66	87,03
		✓	17,07	16,73	20,34	14,78
✓	✓		96,31	95,80	96,77	95,54
✓		✓	96,73	96,21	96,97	95,92
	✓	✓	88,51	87,82	90,60	87,32
✓	✓	✓	97,15	96,65	97,29	96,45

Çizelge 4.20'deki sonuçlara göre, VÜCUT-POZU ve ÇF-VÜCUT-POZU özelliklerinin füzyonuyla elde edilen VÜCUT-ÖZELLİKLERİ %87,88, YÜZ-POZU ve ÇF-YÜZ-POZU özelliklerinin füzyonuyla elde edilen YÜZ-ÖZELLİKLERİ ise %17,07 doğruluk elde etmiştir. Hem VÜCUT-ÖZELLİKLERİ hem de YÜZ-ÖZELLİKLERİ'nden elde edilen sonuçlar, uzamsal özelliklerin Çerçeve Farkı Yöntemiyle temsil edilen zamansal özelliklerle birleştirilmesinin doğruluğu artırdığını göstermektedir.

PARMAK-ÖZELLİKLERİ setinin tek başına kullanımı %96,21'lik doğruluk oranıyla oldukça yüksek bir performans sergilemektedir. Bununla birlikte, PARMAK-ÖZELLİKLERİ ve YÜZ-ÖZELLİKLERİ kombinasyonunun, PARMAK-ÖZELLİKLERİ ve VÜCUT-ÖZELLİKLERİ kombinasyonuna kıyasla daha yüksek bir doğruluk oranına (%96,73) ulaştığı gözlemlenmiştir. Bu durum, yüz ifadelerinin tek başına işaretleri temsil etmede yetersiz kalmasına rağmen, işaret dilindeki duygusal tonları ve ince nüansları aktarma yeteneğine sahip olduğunu göstermektedir. Bu özelliklerin entegre edilmesi, İDT sistemlerinin hassasiyetini ve ayrıntılı anlayışını geliştirir.

En yüksek tanıma doğruluğu %97,15, tüm özellik setlerinin (PARMAK ÖZELLİKLERİ, VÜCUT ÖZELLİKLERİ ve YÜZ ÖZELLİKLERİ) birleştirilmesiyle

elde edilmiştir. Bu sonuç, sağlam bir İDT sistemi uygulamak için farklı özelliklerin entegrasyonunun ne kadar önemli olduğunu kanıtlamaktadır.

4.3.4.5. Diğer çalışmalarla karşılaştırma

Bu çalışma kapsamında gerçekleştirilen, diğer çalışmalarla karşılaştırma süreci önerilen iki ana yönteme odaklanmaktadır: birincisi parmak verilerinden (PARMAK-POZU, BİRLEŞTİRİLMİŞ-PARMAK-POZU ve ÇF-PARMAK-POZU) türetilen PARMAK-ÖZELLİKLERİ kümesine; ikincisi ise bu özelliklere ek olarak VÜCUT-ÖZELLİKLERİ (VÜCUT-POZU, ÇF-VÜCUT-POZU verilerinden) ve YÜZ-ÖZELLİKLERİ'ni (YÜZ-POZU, ÇF-YÜZ-POZU verilerinden) birleştiren TÜM ÖZELLİKLER kümesine dayanmaktadır. Bu iki yaklaşımın mevcut literatürle performans karşılaştırması Çizelge 4.21, Çizelge 4.22, Çizelge 4.23 ve Çizelge 4.24'te ayrıntılı olarak verilmiştir.

Çizelge 4.21. Çalışma-3 kapsamında önerilen yöntemin BosphorusSign22k-genel veri kümesi üzerindeki literatür sonuçlarıyla karşılaştırılması (Akdag ve Baykan, 2024b)

Çalışmalar	Girdi modalitesi	Odak Alanları	Test Doğruluğu (%)
Kindiroglu ve diğ. (2019)	Poz	Tam çerçeve	81,58
Gündüz ve Polat (2021)	RGB, poz, optik akış	Vücut, el, yüz	89,35
Çalışma-3 (önerilen yöntem)	Poz, Çerçeve Farkı	Parmak	96,21
	Poz, Çerçeve Farkı	Vücut, yüz, parmak	97,15

Çizelge 4.21'de BosphorusSign22k-genel veri kümesi üzerinde yapılan çalışmaların sonuçları karşılaştırılmaktadır. Gündüz ve Polat (2021) tarafından önerilen vücut, el, yüz, optik akış ve poz verilerini entegre eden Inception3D ve UKSB-TSA modellerine dayanan yöntem, bu karmaşık ve çok yönlü verileri kullanarak %89,35'lik bir doğruluk elde etmiştir. Bu yöntem, İDT alanında çeşitli model ve veri türlerini entegre etmenin potansiyel avantajını göstermiştir. Buna karşılık, önerilen PARMAK-ÖZELLİKLERİ yaklaşımı, işaret dilini anlamada parmak hareketlerinin kritik rolünü vurgulamakta ve yalnızca bu özelliklere odaklanarak %96,21'lik bir doğruluk elde etmektedir. Bu sonuç, parmağa dair verilerin derinlemesine analizinin İDT'nin doğruluğunu önemli ölçüde artırabileceğinin güçlü bir göstergesidir. Ayrıca, uzamsal ve zamansal bilgileri yakalama kapasitesine sahip Inception3D ve UKSB-TSA modellerine kıyasla, poz görüntü verileri (örn. VÜCUT-POZU) ve bu verilere uygulanan Çerçeve Farkı Yöntemiyle üretilen görüntülerin (örn. ÇF-VÜCUT-POZU) MobileNetV2 gibi

nispeten daha az karmaşık bir modelden elde edilen özelliklerinin birleştirilmesiyle %87,88 gibi dikkate değer bir doğruluk oranı sağlaması Çerçeve Farkı Yönteminin zamansal özellikleri tespit etmede oldukça etkili olduğunu göstermiştir. Bütün özelliklerin füzyonuyla gerçekleştirilen TÜM-ÖZELLİKLER yaklaşımıyla elde edilen %97,15 doğruluk oranı ise, İDT tanıma doğruluğunu daha da geliştirerek önemli bir katkı sağlamıştır.

Çizelge 4.22. Çalışma-3 kapsamında önerilen yöntemin BosphorusSign22k veri kümesi üzerindeki literatür sonuçlarıyla karşılaştırılması (Akdag ve Baykan, 2024b)

Çalışmalar	Girdi Modalitesi	Odak Alanları	Test Doğruluğu (%)
Özdemir ve diğ. (2020)	RGB	Tam çerçeve	78,85
Özdemir ve diğ. (2020)	RGB	Tam çerçeve	88,53
Gökçe ve diğ. (2020)	RGB	Vücut, el, yüz	94,94
Sincan ve Keles (2021)	RGB, RGB-HTG	Tam çerçeve	94,83
Kındıroglu ve diğ. (2023)	RGB, poz	Tam çerçeve	94,90
Özdemir ve diğ. (2023)	RGB, poz	Vücut, el, yüz	92,58
Çalışma-3 (önerilen yöntem)	Poz, Çerçeve Farkı	Parmak	92,79
	Poz, Çerçeve Farkı	Vücut, yüz, parmak	95,13

Çizelge 4.22, 744 sınıflı BosphorusSign22k veri kümesini kullanan çalışmaları göstermektedir. Geçmişteki en yüksek performanslar arasında Gökçe ve diğ. (2020) ve Kındıroglu ve diğ. (2023) uzamsal ve zamansal bilgileri işleyebilen güçlü bir model olan MC3-18 modelini kullanırken, Sincan ve Keles (2021) I3D modelinden yararlanmıştır. Önerilen yöntemde ise nispeten daha az karmaşık olan MobileNetV2 ve Çok Kanallı-MobileNetV2 modelleri kullanılmış ve zamansal bilgiler video verilerinden Çerçeve farkı Yöntemi kullanılarak elde edilmiştir. Gökçe ve diğ. (2020) tarafından vücut, el ve yüz bileşenlerini içeren yöntemle elde edilen %94,94'lük en yüksek tanıma doğruluğu, önerilen yöntemle (vücut, yüz ve parmağa dair özellikleri içeren) %95,13'lük bir değerle aşılmıştır. Ayrıca sadece parmak özelliklerine dayalı olarak geliştirilen yaklaşım %92,79 tanıma doğruluğu ile birçok çalışmayı geride bırakarak rekabetçi bir değer elde etmiştir.

Çizelge 4.23. Çalışma-3 kapsamında önerilen yöntemin LSA64 veri kümesi üzerindeki literatür sonuçlarıyla karşılaştırılması (Akdag ve Baykan, 2024b)

	Çalışmalar	Girdi modalitesi	Odak Alanları	Test Doğruluğu (%)
D1	Marais ve diğ. (2022b)	RGB	Tam çerçeve	74,22
	Çalışma-3 (önerilen yöntem)	Poz, Çerçeve Farkı	Parmak	95,93
		Poz, Çerçeve Farkı	Vücut, yüz, parmak	97,96
D2	Ronchetti ve diğ. (2016a)	RGB	El	91,70
	Rodríguez ve Martínez (2018)	RGB	Tam çerçeve	85
	Alyami ve diğ. (2023)	Poz	El, yüz	91,09
	Çalışma-3 (önerilen yöntem)	Poz, Çerçeve Farkı	Parmak	97,59
		Poz, Çerçeve Farkı	Vücut, yüz, parmak	98,93
D3	Ronchetti ve diğ. (2016a)	RGB	El	97
	Konstantinidis ve diğ. (2018b)	Poz	Vücut, el	98,09
	Konstantinidis ve diğ. (2018a)	RGB, poz, optik akış	Vücut, el, yüz	99,84
	Zhang ve Li (2019)	RGB	Tam çerçeve	99,063
	Imran ve Raman (2020)	HTG, DG, RGB-HG	Tam çerçeve	97,81
	Marais ve diğ. (2022a)	RGB	Tam çerçeve	95,50
	Bohacek ve Hruz (2022)	Poz	Vücut, el	100
	Alyami ve diğ. (2023)	Poz	El, yüz	98,25
	Çalışma-3 (önerilen yöntem)	Poz, Çerçeve Farkı	Parmak	99,09
		Poz, Çerçeve Farkı	Vücut, yüz, parmak	99,78

Çizelge 4.23'te 64 işaret sınıfı içeren LSA64 veri kümesi üzerinde yapılan çalışmalar sunulmaktadır. Bu değerlendirmede, LSA64 veri kümesinin farklı eğitim ve test bölümlemeleri için sonuçlar analiz edilmiştir. İşaretçiden bağımsın deney ortamları olan D1 ve D2 için PARMAK-ÖZELLİKLERİ ve TÜM-ÖZELLİKLER yaklaşımıyla elde edilen sonuçlar literatürdeki tüm çalışmalardan daha iyi performans göstermiştir. D1 için, PARMAK-ÖZELLİKLERİ ve TÜM-ÖZELLİKLER ile sırasıyla %95,93 ve %97,96 doğruluk oranları elde edilmiştir. Bu sonuçlar önerilen yöntemin üstün performansını ve uyarlanabilirliğini göstermektedir. D2 için ise, PARMAK-ÖZELLİKLERİ ve TÜM-ÖZELLİKLER ile sırasıyla %97,59 ve %98,93 doğruluk oranları elde edilmiş ve yöntemin genelleme kabiliyeti ve farklı işaretçiler tarafından üretilen işaret dillerini tanıma kapasitesinin oldukça yüksek olduğu kanıtlanmıştır. Veri kümesinin rastgele %80 eğitim ve %20 test olarak bölündüğü D3 için ise %99,78'lik doğruluk oranı literatürdeki en son çalışmalara çok yakın, rekabetçi bir sonuçtur. Bununla birlikte D3 setinden elde edilen sonuçların eğitim ve test verilerinin dağılımı açısından tamamen rastgele olduğunu ve rastgeleliğin karşılaştırılan tüm çalışmalarda değişebileceğini belirtmek gerekir.

Çizelge 4.24. Çalışma-3 kapsamında önerilen yöntemin GSL veri kümesi üzerindeki literatür sonuçlarıyla karşılaştırılması (Akdag ve Baykan, 2024b)

Çalışmalar	Girdi modalitesi	Odak Alanları	Test Doğruluğu (%)
Adaloglou ve diğ. (2020)	RGB	Tam çerçeve	86,03
Adaloglou ve diğ. (2020)	RGB	Tam çerçeve	86,23
Adaloglou ve diğ. (2020)	RGB	Tam çerçeve	89,74
Selvaraj ve diğ. (2021)	Poz	Tam çerçeve	86,60
Selvaraj ve diğ. (2021)	Poz	Tam çerçeve	89,50
Selvaraj ve diğ. (2021)	Poz	Tam çerçeve	93,50
Selvaraj ve diğ. (2021)	Poz	Tam çerçeve	95,40
Fang ve diğ. (2023)	RGB, poz	Tam çerçeve	91,49
Çalışma-3 (önerilen yöntem)	Poz, Çerçeve Farkı	Parmak	94,05
	Poz, Çerçeve Farkı	Vücut, yüz, parmak	95,37

Son olarak, önerilen yöntem, farklı işaret dillerine uygulanabilirliğini doğrulamak için 310 işaret sözcüğü içeren GSL veri kümesi üzerinde değerlendirilmiştir. Bu değerlendirmenin sonuçları Çizelge 4.24'te sunulmuştur. Adaloglou ve diğ. (2020) tarafından yapılan çalışmada, hem uzamsal hem de zamansal bilgiyi değerlendirebilen GoogLeNet + TConvs (Cui ve diğ., 2019) (%86,03), 3D-ResNet + BLSTM (Pu ve diğ., 2019) (%86,23) ve I3D + BLSTM (Carreira ve Zisserman, 2017) (%89,74) gibi yöntemler bu veri kümesi üzerinde kullanılmış ve belirtilen doğruluk oranlarına ulaşılmıştır. Buna karşın, önerilen yöntem sadece PARMAK-ÖZELLİKLERİ'ni kullanan yaklaşım ile %94,05 gibi dikkate değer bir tanıma doğruluğu elde etmiştir. Bu sonuç literatürdeki birçok çalışmayı geride bırakmakta ve parmak pozu tabanlı görüntülerin farklı kanallarda görselleştirilmesiyle elde edilen verilerin Çok Kanallı-MobileNetV2 modeliyle işlenmesinin ne kadar etkili olduğunu göstermektedir. TÜM-ÖZELLİKLER'i kullanan yaklaşım ile elde edilen %95,37'lik tanıma doğruluğu ise, Selvaraj ve diğ. (2021)'nin çalışmasında bulunan ve literatürde raporlanan en yüksek doğruluk oranı olan %95,40'a çok yaklaşmakta ve diğer tüm çalışmaları ise geride bırakmaktadır. Bu başarı, önerilen sistemin hem uzamsal hem de zamansal bilgileri etkin bir şekilde kullanarak İDT alanında sağlam ve güvenilir tanıma performansı sağlayabileceğini göstermektedir.

Özetle, bu çalışma kapsamında önerilen İDT yöntemi 64, 174, 310 ve 744 işaret sınıfı içeren çeşitli veri kümeleri üzerinde değerlendirilmiş ve literatürdeki mevcut çalışmalara kıyasla daha üstün veya rekabetçi sonuçlar elde ettiği görülmüştür. Özellikle, sadece parmakla ilgili özellikleri (PARMAK-ÖZELLİKLERİ) kullanan çok kanallı yaklaşım kayda değer bir performans elde etmiştir. Yüz (YÜZ-ÖZELLİKLERİ) ve vücut (VÜCUT-ÖZELLİKLERİ) ile ilgili özelliklerin entegrasyonu bu performansı daha da artırmıştır. Tüm veri kümelerinden elde edilen yüksek tanıma oranları, önerilen

metodolojilerin farklı veri kümeleri arasında güvenilir ve sağlam olduğunu göstermektedir.

4.5. Önerilen Yöntemlerin Karşılaştırılması

Bu bölümde, tez kapsamında gerçekleştirilen çalışmalar karşılaştırılmıştır. Karşılaştırmalar, çalışmalarda önerilen nihai İDT yöntemleri temel alınarak yapılmıştır. İlk olarak, önerilen yöntemlerin zaman ve bellek maliyet analizleri gerçekleştirilmiş, ardından çalışma kapsamında kullanılan tüm veri kümeleri üzerinden elde edilen tanıma doğrulukları incelenmiştir. Bu analizler Bölüm 4'ün ilk paragrafında açıklanan deneysel çalışmaların da gerçekleştirildiği sistem üzerinde GPU kullanılmadan yapılmıştır.

Zaman ve bellek maliyet analizleri, BosphorusSign22k-genel test veri kümesi kullanılarak gerçekleştirilmiştir. 949 örnek video içeren bu test kümesinde videoların toplam süresi 2.623 saniye, çerçeve sayısı ise 78.684'tür. Bu durum, bir videonun ortalama süresinin yaklaşık 2,76 saniye ve video başına düşen ortalama kare sayısının 82 olduğunu göstermektedir. Zamansal analiz, bu test verisinin tamamı üzerinde yapılmış ve ardından elde edilen değerlerin ortalaması alınmıştır. Bellek maliyeti analizi ise, veri setinin ortalama kare sayısı olan 82'ye eşit örnek bir video kullanılarak, tanımlama sürecinde kullanılan maksimum bellek miktarı esas alınarak gerçekleştirilmiştir. Sonuçlar, BosphorusSign22k-genel test veri kümesinden elde edilen doğruluk oranlarıyla birlikte Çizelge 4.25'te sunulmuştur.

Çizelge 4.25. Önerilen yöntemlerin zaman ve bellek maliyeti analizi

Yöntemler	Zamansal Analiz				Maksimum Bellek Kullanımı (MB)	Doğruluk (%)
	Veri Ön İşleme (ms)	Özellik Çıkarımı ve Sınıflandırma (ms)	Toplam Tanımlama Süresi (ms)	RTF		
Yöntem-1	4.851	294	5.146	1,86	2.268	94,52
Yöntem-2	5.297	74	5.371	2,25	2.288	96,94
Yöntem-3	6.719	499	7.218	2,47	2.334	97,15

Zamansal analizde, veri ön işleme adımı ham video verisinin derin öğrenme modelleri için giriş verisi olana kadar gerçekleşen tüm süreçleri, özellik çıkarma ve sınıflandırma adımı ise eğitilmiş derin öğrenme modellerinden özelliklerin elde edilmesi ve bu özelliklerin birleştirilerek DVM ile sınıflandırılma aşamalarını kapsamaktadır. Ayrıca toplam tanımlama süresi de tabloda gösterilmiştir. Bununla birlikte gerçek

zamanlı performansı değerlendirmek için kritik bir değer olan Gerçek Zaman Faktörü (Real Time Factor, RTF) (Wang ve diğ., 2021b) kullanılmıştır. RTF, toplam işlem süresinin işaret dili videosunun gerçek süresine bölünmesiyle hesaplanır. RTF değeri ne kadar düşük olursa, kesintisiz iletişim için gereken videoların gerçek zamanlı işlenme performansı o kadar iyi demektir.

Çizelge 4.25 incelendiğinde, tüm yöntemlerde en çok zaman alan aşamanın veri ön işleme adımı olduğu görülmüştür. Veri ön işleme adımı kapsamında en çok zaman alan süreç ise MediaPipe Holistic ile poz verilerinin elde edilmesidir. 82 karelik bir video görüntüsünden poz verilerinin elde edilme süreci 4.237 ms olarak ölçülmüştür. Videodan poz verilerinin elde edilmesi toplam işlem süresinin büyük bir bölümünü oluşturmakta ve yöntemlerin genel verimliliğinde önemli bir rol oynamaktadır. Dolayısıyla bu durum, iyileştirme ihtiyacını vurgulamaktadır. Poz verileri çıkarıldıktan sonra gerçekleştirilen veri ön işleme adımlarındaki diğer süreçler çeşitli uzamsal ve zamansal özelliklerin elde edilmesini, yeniden boyutlandırma ve normalizasyon gibi işlemleri kapsamaktadır ve poz çıkarma işlemine kıyasla daha az maliyetlidir.

Veri ön işleme adımında en iyi zamansal performans Yöntem-1 ile elde edilmiştir. Çünkü Yöntem-1'de, elde edilen poz verilerinden sadece vücut, eller ve yüzlere dair üç görsel poz verisi oluşturulmuş, Yöntem-2'de, ilave olarak, poz görüntü verilerine HTG yöntemi uygulanmış ve parmak pozu tabanlı HTG verisi oluşturulmuştur, bu da ek zamansal maliyet getirmiştir. Yöntem-3'te ise, Yöntem-1'deki verilere ilave olarak her bir parmak pozu görselinin ayrı bir kanalda görselleştirilmesi ve görsel poz verilerine Çerçeve Farkı Yöntemi uygulanması işlemleri gerçekleştirilmiş, bu işlemler de zamansal maliyeti artırmıştır.

Özellik çıkarımı ve sınıflandırma aşaması veri ön işleme aşamasına göre daha az zaman almıştır. Bu aşama, üç adet R3(2+1)D-SLR modeli kullanan Yöntem-1 için özellik çıkarımı ve sınıflandırma süreciyle birlikte 294 ms'de gerçekleşmiş, Yöntem-2'de kullanılan sekiz ResNet-18 modeliyle ise 74 ms'de tamamlanmıştır. Yöntem-2'deki bu performans farkı uzamsal ve zamansal özelliklerin birlikte ele alındığı R3(2+1)D-SLR modelleri yerine daha az karmaşık olan ResNet-18 modellerinin kullanımından kaynaklanmaktadır. Ancak burada da zamansal özelliklerin ele alınması işlemi veri ön işleme aşamasında elde edilen HTG verilerine devredilmiştir, dolayısıyla da bu durum veri ön işleme adımına ek bir maliyet olarak yansımıştır. Yöntem 3'te ise üç adet Çok Kanallı-MobileNetV2, dört adet MobileNetV2 modeli ile bu modellerden elde edilen özelliklere TBA uygulanması ve sınıflandırma süreçleri 499 ms zaman almıştır. Her bir

Çok Kanallı-MobileNetV2 modelinde beş adet MobileNetV2 bulunması, parmak pozu verilerinin ayrıntılı analizi ve elde edilen tanıma doğrulukları bu sürenin kabul edilebilir olduğunu göstermektedir.

Toplam tanıma süresi ve RTF değerlerine bakıldığında Yöntem-1'den Yöntem-3'e doğru bir artış olduğu görülmüştür. Yöntem-1 ile elde edilen 1,86 değeri en iyi RTF değeri olmasına rağmen, gerçek zamanlı çalışma gereksinimleri göz önüne alındığında, ideal olarak RTF değerinin 1'e yakın veya daha düşük olması gerekir. Bu durum, her üç yöntemin gerçek zamanlı performanslarında iyileştirmelere ihtiyacı olduğunu göstermektedir.

Maksimum bellek kullanımı da zamansal maliyete benzer bir tablo sergilemiştir. 82 çerçevelik bir video görüntüsü için gerçekleştirilen tanıma sürecinde, Yöntem-1'de 2.268 MB, Yöntem-2'de 2.288 MB, Yöntem-3'te ise 2.334 MB bellek ihtiyacı olmuştur. İhtiyaç olan bellek miktarı veri ön işleme aşamasında elde edilen özelliklere ve kullanılan modellerin karmaşıklığına bağlı olarak değişkenlik göstermiştir. Elde edilen maksimum bellek kullanım değerleri yüksek performanslı bilgisayarlar için makul kabul edilebilir. Ancak mobil cihazlar gibi sınırlı bellek kapasitesine sahip ortamlarda bu miktarların yönetilmesi zorluk oluşturabilir ve daha etkili bellek yönetim stratejilerinin geliştirilmesini gerektirebilir.

Yöntem-1'den Yöntem-3'e doğru zaman ve bellek maliyetlerinde yaşanan artışa rağmen tanıma doğruluğunda önemli bir iyileşme gözlemlenmiştir. Bu olumlu etki, veri ön işleme aşamasında elde edilen özelliklerin sayısının, niteliğinin ve oluşturulan modellerin karmaşıklığının artmasıyla doğrudan ilişkilidir. Detaylı ve kapsamlı özellik çıkarımı, maliyetlerin yükselmesine rağmen, sonuçların doğruluğunu önemli ölçüde iyileştirmiştir. Bu bağlamda, önerilen yöntemlerin tez kapsamında kullanılan veri kümeleri üzerinde elde ettiği tanıma doğrulukları literatürde yer alan en yüksek skorlarla birlikte Çizelge 4.26'da sunulmuştur.

Çizelge 4.26. Önerilen yöntemlerin farklı veri kümelerindeki performans karşılaştırması

Yöntemler	Doğruluk (%)					
	Bosphorus-Sign22k-genel	Bosphorus-Sign22k	LSA64 (D1)	LSA64 (D2)	LSA64 (D3)	GSL
Gündüz ve Polat (2021)	89,35	-	-	-	-	-
Gökçe ve diğ. (2020)	-	94,94	-	-	-	-
Marais ve diğ. (2022b)	-	-	74,22	-	-	-
Ronchetti ve diğ. (2016a)	-	-	-	91,70	-	-
Bohacek ve Hruz (2022)	-	-	-	-	100	-
Selvaraj ve diğ. (2021)	-	-	-	-	-	95,40
Yöntem-1	94,52	-	96,56	98,53	100	-
Yöntem-2	96,94	94,87	98,43	98,68	100	95,14
Yöntem-3	97,15	95,13	97,96	98,93	99,78	95,37

Bu tez çalışması kapsamında gerçekleştirilen yöntemlerin karşılaştırmalı analizi, çeşitli veri kümeleri üzerinde elde edilen tanıma doğruluk oranları açısından önemli bulgular sunmaktadır. Yöntem-1, Yöntem-2 ve Yöntem-3 literatürde yer alan diğer yöntemlere göre genellikle daha yüksek veya rekabetçi tanıma oranları sağlamıştır, bu da önerilen yöntemlerin etkin özellik çıkarımı ve sınıflandırma stratejilerine sahip olduğunu göstermektedir. Özellikle, Yöntem-3'ün BosphorusSign22k-genel, BosphorusSign22k ve GSL gibi farklı veri kümelerinde sağladığı %95 ve üzeri doğruluk oranları, bu yöntemin geniş uygulama alanlarına adapte olabileceğini göstermektedir.

Bu bölümdeki analizler, önerilen İDT yöntemlerinin performansını zaman ve maliyet açısından kapsamlı bir şekilde değerlendirmiştir. Sonuçlar, her üç yöntemin de yüksek tanıma oranlarına sahip olduğunu göstermekte; ancak, gerçek zamanlı uygulamalar ve sınırlı kaynaklara sahip ortamlar için optimizasyon ihtiyacını da ortaya koymakta ve böylece gelecekteki çalışmalar için bir temel oluşturmaktadır.

5. SONUÇLAR VE ÖNERİLER

5.1. Sonuçlar

İşaret dili, işitme engelli bireylerin toplumla iletişimini sağlayan hayati bir araçtır; ancak, bu dilin doğası gereği içerdiği karmaşık uzamsal ve zamansal yapılar, otomatik tanıma ve tespit süreçlerini zorlaştırmaktadır. Bu tez çalışması kapsamında video tabanlı izole İşaret Dili Tanıma (İDT) alanında üç farklı çalışma gerçekleştirilmiştir. Her bir çalışma ile işaret dilinin daha doğru bir şekilde anlaşılması için yenilikçi yöntemler geliştirilmiş, işaret dili kelimelerinin otomatik olarak sınıflandırılma performansı iyileştirilmiştir.

İlk çalışmada, R3D ve R(2+1)D evrişim bloklarının birleştirilmesiyle oluşturulan R3(2+1)D blok yapısına dayalı olarak geliştirilen R3(2+1)D-SLR ağı önerilmiştir. Bu yenilikçi ağ modeli, işaret dilinin içerdiği karmaşık uzamsal ve zamansal özellikleri etkin bir şekilde öğrenebilme kapasitesiyle dikkat çekmektedir. Mevcut R3D-18 ve R(2+1)D-18 modelleri daha derin öğrenme katmanlarına sahip olmasına rağmen, önerilen model bu modellerle kıyaslandığında gözle görülür bir performans artışı sağlamıştır. Çalışma kapsamında önerilen R3(2+1)D-SLR modeli, işaretçiyi tamamen gösteren tam bir vücut verisinin beraberinde işaret dilinin önemli bileşenlerinden olan el ve yüz verileri için de eğitilerek, bu özelliklerin İDT'deki etkinliği değerlendirilmiş ve eğitilmiş modellerden özellikler çıkarılmıştır. Ardından bu özellikler birleştirilmiş ve DVM ile sınıflandırılmıştır. Bu yaklaşım, İDT için önerilen sistemin genel performansını önemli ölçüde artırarak yüksek doğruluk oranları göstermiştir. Mevcut başarıyla beraber, önerilen İDT sistemi, literatürde çoğunlukla ihmal edilen ve gerçek dünya senaryolarını daha iyi yansıtan farklı arka planlara sahip veriler kullanılarak değerlendirilmiştir. Bu kapsamda, test verilerine çeşitli arka planlar eklenmiş, değişken arka planlara sahip test kümesinde yapılan değerlendirmelerde performans düşüklüğü gözlemlenmiştir. Bu tespitten hareketle, önerilen İDT sisteminin değişken arka plan koşullarına karşı dayanıklılığını artırmak amacıyla, mevcut giriş verileri, karşılık gelen görsel poz verileri ile güncellenmiştir. Böylece önerilen sistem değişken arka planlara karşı sağlamlaşmış, ayrıca arka plan eklenmemiş test görüntüleriyle yapılan değerlendirmelerde de performansının iyileştiği görülmüştür. Çalışmayla ilgili bir başka bulgu da el pozu görüntülerindeki her bir parmağın farklı bir renge sahip olmasının sistemin performansını artırdığı tespit edilmiştir. Özetle; poz verilerine dayalı olarak vücut, eller ve yüz

özelliklerini entegre eden bu kapsamlı metodoloji, çeşitli veri kümelerinde, farklı arka plan koşullarında ve işaretçiden bağımsız değerlendirmelerde mevcut literatürdeki çalışmalardan daha iyi performans göstermiştir.

İkinci çalışmada, İDT sürecinin doğruluğunu artırmak amacıyla videolardan elde edilen poz görüntü verileri ve bu verilerden türetilen Hareket Tarihçesi Görüntüsü (HTG) verilerini kullanan bir yöntem geliştirilmiştir. Bu iki veri türünün entegrasyonu, işaret dilinin hem uzamsal hem de zamansal boyutlarının daha etkin bir şekilde analiz edilmesini sağlamıştır. Özellik çıkarımı sürecinde, RReLU aktivasyon fonksiyonu kullanılarak optimize edilmiş ResNet-18 modeli tercih edilmiştir. Bu model, daha fazla doğrusal olmayanlık sağlayarak öğrenme sürecini güçlendirmiş ve modelin genel doğruluğunu artırmıştır. Vücut, el, yüz pozu görüntüleri ve bunlardan elde edilen HTG verilerinin, eğitilmiş ResNet-18 modellerinden elde edilen özellikleri birleştirilmiş, DVM kullanılarak sınıflandırılmış ve önemli doğruluk artışı sağlanmıştır. Özellikle, önerilen PARMAK-POZU-HTG, işaret dili tanıma çalışmalarında genellikle ihmal edilen parmak hareketlerinin detaylı bir şekilde analiz edilmesine olanak tanımıştır. Araştırmada, poz ve HTG verilerinin kombinasyonunun İDT'nin genel doğruluğuna olan katkısı dikkat çekici bulunmuştur. Ayrıca, parmak pozu tabanlı HTG'nin, İDT sistemlerinden elde edilen başarıyı artırma potansiyeline sahip olduğu ve bu özelliklerin diğer manuel ve manuel olmayan özelliklerle olan etkileşimlerinin, sistemin genel performansını önemli ölçüde iyileştirdiği gözlemlenmiştir. Bununla birlikte, eksik poz verilerinin doğrusal enterpolasyon yoluyla tamamlanmasının, modellerin genel performansını olumlu yönde etkilediği görülmüştür. Bu yöntem, eksik veya gürültülü verilerin model tarafından daha etkin bir şekilde işlenmesini sağlayarak, işaret dili tanıma sürecinin güvenilirliğini ve etkinliğini artırmıştır.

Üçüncü çalışmada, İDT görevi için, özellikle parmak hareketlerine odaklanan çok akışlı bir yaklaşım geliştirilmiştir. Bu kapsamda işaretçinin her bir parmağı için poz verilerinin ayrı bir kanalda görselleştirildiği PARMAK-POZU, her iki elden karşılıklı parmak bölgelerinin kırpılıp birleştirilmesiyle daha odaklanmış bir görsel olan BİRLEŞTİRİLMİŞ-PARMAK-POZU, PARMAK-POZU verisine Çerçeve Farkı Yönteminin uygulanmasıyla elde edilen ÇF-PARMAK-POZU görüntü verileri oluşturulmuştur. Geliştirilen Çok-Kanallı-MobileNetV2 modeli, bu verileri kullanarak, parmak hareketlerinin zengin ve karmaşık özelliklerini etkin bir şekilde öğrenmiştir. Eğitilmiş modellerden çıkan özelliklerin TBA ile indirgenmiş versiyonları birleştirilerek DVM ile sınıflandırılması sonucu yüksek tanıma oranları elde edilmiştir. Bu başarı,

parmak hareketlerinin işaret dilinin anlaşılması ve tespit edilmesinde ne kadar kritik olduğunu göstermiştir. Ayrıca, önerilen İDT sistemi, işaret dilinin sadece parmak verileriyle sınırlı kalmayıp vücut ve yüz verilerini de içerecek şekilde genişletilmiştir. Bu genişletme, poz görüntü verileri ve bu verilere uygulanan Çerçeve Farkı Yöntemiyle oluşturulmuş görüntülerin MobileNetV2 modelinden çıkarılan özelliklerine dayanmaktadır. Tüm özellikler indirgenerek birleştirilmiş, DVM ile sınıflandırılmış ve yüksek sınıflandırma doğrulukları elde edilmiştir.

Tez kapsamında gerçekleştirilmiş olan çalışmaların her biri, İDT alanında önemli yenilikler sunarak işitme engelli bireylerin toplumla iletişimini kolaylaştırmayı amaçlamaktadır. Geliştirilen teknikler, işaret dilinin karmaşık yapısını daha iyi çözümlemiş ve bu dilin otomatik tanınma performansını iyileştirmiştir.

5.2. Öneriler

Bu tez kapsamında yapılan çalışmalar, İDT alanında önemli ilerlemeler sağlamış ve işitme engelli bireylerin toplumla iletişimini ve entegrasyonunu artıracak çıktılar ortaya koymuştur. Bulgular, bu teknoloji alanının daha da geliştirilmesi gerektiğini ve çeşitli iyileştirmelerle daha geniş kullanım alanlarına ulaşabileceğini göstermektedir. Gelecekteki çalışmalarda;

Karmaşık işaret dili yapılarının tanınmasını iyileştirmek amacıyla, ek modalitelerin entegrasyonu üzerine araştırmalar yapılabilir. İşaret dilinin zengin jest ve ifade dizisi, bireyler arasındaki fiziksel farklılıklar gibi faktörlerden etkilenebilir. Bu nedenle farklı el boyutları ve parmak uzunlukları gibi bireysel farklılıkların modelin doğruluğu ve uygulanabilirliği üzerindeki etkileri incelenebilir.

Derin öğrenme ve makine öğrenimi modellerinde hiperparametre optimizasyonunun yanı sıra, İDT sistemlerinin sağlamlığını artıracak veri artırma teknikleri araştırılabilir. Bu teknikler, modelin çeşitli koşullarda daha iyi performans göstermesine olanak tanıyabilir ve gerçek dünya uygulamalarında kullanılabilirliğini artırabilir.

İDT sistemlerinin gerçek zamanlı performansını iyileştirmek ve daha verimli bellek yönetimi stratejileri geliştirmek üzerine yoğunlaşılabilir. Ayrıca, sürekli işaret dili tanıma görevine odaklanarak, özellikle de uzun cümleleri ve tam konuşmaları tanıma kapasitesini geliştirmek üzerine çalışılabilir. Bu çalışmalar, İDT sistemlerinin gerçek zamanlı kullanılabilirliğini artırarak daha doğru iletişim sağlanmasına yardımcı olacaktır.

KAYNAKLAR

Abdi, H. ve Williams, L. J., 2010, Principal component analysis, *Wiley interdisciplinary reviews: computational statistics*, 2 (4), 433-459.

Adaloglou, N., Chatzis, T., Papastratis, I., Stergioulas, A., Papadopoulos, G. T., Zacharopoulou, V., Xydopoulos, G. J., Atzakas, K. ve Daras, P., 2020, A comprehensive study on sign language recognition methods, *arXiv*.

Adaloglou, N., Chatzis, T., Papastratis, I., Stergioulas, A., Papadopoulos, G. T., Zacharopoulou, V., Xydopoulos, G. J., Atzakas, K., Papazachariou, D. ve Daras, P., 2022, A comprehensive study on deep learning-based methods for sign language recognition, *IEEE Transactions on Multimedia*, 24.

Ahad, M. A. R., Tan, J. K., Kim, H. ve Ishikawa, S., 2012, Motion history image: Its variants and applications, *Machine Vision and Applications*, 23 (2), 255-281.

Ahad, M. A. R., 2013, Motion history image, In: Motion history images for action recognition and understanding, Eds, *London*: Springer London, p. 31-76.

Ahmed, M. A., Zaidan, B. B., Zaidan, A. A., Salih, M. M. ve Lakulu, M. M. B., 2018, A review on systems-based sensory gloves for sign language recognition state of the art between 2007 and 2017, *Sensors 2018, Vol. 18, Page 2208*, 18 (7), 2208-2208.

Ahmed, W., Chanda, K. ve Mitra, S., 2017, Vision based hand gesture recognition using dynamic time warping for indian sign language, *Proceedings - 2016 International Conference on Information Science, ICIS 2016*.

Akdag, A. ve Baykan, O. K., 2024a, Enhancing signer-independent recognition of isolated sign language through advanced deep learning techniques and feature fusion, *Electronics*, 13 (7), 1188.

Akdag, A. ve Baykan, O. K., 2024b, Multi-stream isolated sign language recognition based on finger features derived from pose data, *Electronics*, 13 (8), 1591.

Akdağ, A. ve Baykan, Ö. K., 2024, Isolated sign language recognition through integrating pose data and motion history images, *PeerJ Computer Science*, 10, e2054.

Aldhahri, E., Aljuhani, R., Alfaidi, A., Alshehri, B., Alwadei, H., Aljojo, N., Alshutayri, A. ve Almazroi, A., 2023, Arabic sign language recognition using convolutional neural network and mobilenet, *Arabian Journal for Science and Engineering*, 48 (2).

Aloysius, N. ve Geetha, M., 2020, Understanding vision-based continuous sign language recognition, *Multimedia Tools and Applications*, 79 (31-32).

Alsharif, B., Altaher, A. S., Altaher, A., Ilyas, M. ve Alalwany, E., 2023, Deep learning technology to recognize american sign language alphabet, *Sensors*, 23 (18).

Altman, N. ve Krzywinski, M., 2018, The curse(s) of dimensionality this-month. Nature Methods. 15.

Alyami, S., Luqman, H. ve Hammoudeh, M., 2023, Isolated arabic sign language recognition using a transformer-based model and landmark keypoints, *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*

Amari, S.-i. ve Wu, S., 1999, Improving support vector machine classifiers by modifying kernel functions, *Neural Networks*, 12 (6), 783-789.

Anonymous, 2023, International day of sign languages [online], https://www.un.org/en/observances/sign-languages-day, [Ziyaret Tarihi: 15.01.2024].

Aremu, O. O., Hyland-Wood, D. ve McAree, P. R., 2020, A machine learning approach to circumventing the curse of dimensionality in discontinuous time series machine data, *Reliability Engineering and System Safety*, 195.

Baltrušaitis, T., Ahuja, C. ve Morency, L.-P., 2018, Multimodal machine learning: A survey and taxonomy, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41 (2), 423-443.

Barczak, A. L. C., Reyes, N. H., Abastillas, M., Piccio, A. ve Susnjak, T., 2011, A new 2d static hand gesture colour image dataset for asl gestures.

Bishop, C. M. ve Nasrabadi, N. M., 2006, Pattern recognition and machine learning, 4, Springer, p.

Bobick, A. F. ve Davis, J. W., 2001, The recognition of human movement using temporal templates, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23 (3).

Bohacek, M. ve Hruz, M., 2022, Sign pose-based transformer for word-level sign language recognition, *Proceedings - 2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACVW 2022*.

Bottou, L., 2010, Large-scale machine learning with stochastic gradient descent, *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, 177-186.

Camgoz, N. C., Kindiroglu, A. A., Karabüklü, S., Kelepir, M., Sumru Ozsoy, A. ve Akarun, L., 2016, Bosphorussign: A turkish sign language recognition corpus in health and finance domains, *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*.

Carreira, J. ve Zisserman, A., 2017, Quo vadis, action recognition? A new model and the kinetics dataset, *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*.

Chandragiri, P. S. ve Ijjina, E. P., 2021, Recognizing human actions in video using motion history image and deep learning, *2021 12th International Conference on Computing Communication and Networking Technologies, ICCCNT 2021*.

Cortes, C. ve Vapnik, V., 1995, Support-vector networks, *Machine Learning*, 20 (3), 273-297.

Cui, R., Liu, H. ve Zhang, C., 2019, A deep neural framework for continuous sign language recognition by iterative training, *IEEE Transactions on Multimedia*, 21 (7).

Damaneh, M. M., Mohanna, F. ve Jafari, P., 2023, Static hand gesture recognition in sign language based on convolutional neural network with feature extraction method using orb descriptor and gabor filter, *Expert Systems with Applications*, 211, 118559-118559.

Das, S., Biswas, S. K. ve Purkayastha, B., 2022, A deep sign language recognition system for indian sign language, *Neural Computing and Applications*.

Das, S., Imtiaz, M. S., Neom, N. H., Siddique, N. ve Wang, H., 2023, A hybrid approach for bangla sign language recognition using deep transfer learning model with random forest classifier, *Expert Systems with Applications*, 213.

de Castro, G. Z., Guerra, R. R. ve Guimarães, F. G., 2023, Automatic translation of sign language with multi-stream 3d cnn and generation of artificial depth maps, *Expert Systems with Applications*, 215.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G. ve Gelly, S., 2020, An image is worth 16x16 words: Transformers for image recognition at scale, *arXiv preprint arXiv:2010.11929*.

Erten, H. ve Arıcı, N., 2022, İşaret dilinin tarihi serüveni ve türk İşaret dili, *Afyon Kocatepe Üniversitesi Sosyal Bilimler Dergisi*, 24 (1), 1-14.

Ezel, E., 2018, Derin öğrenme yöntemi kullanılarak görüntü-tabanlı türk işaret dili tanıma, *Selçuk Üniversitesi*.

Fang, Y., Xiao, Z., Cai, S. ve Ni, L., 2023, Adversarial multi-task deep learning for signer-independent feature representation, *Applied Intelligence*, 53 (4).

Ghosh, S. K., Rashmi, M., Mohan, B. R. ve Guddeti, R. M. R., 2023, Deep learning-based multi-view 3d-human action recognition using skeleton and depth data, *Multimedia Tools and Applications*, 82 (13).

Goodfellow, I., Bengio, Y. ve Courville, A., 2016, Deep learning, MIT press, p.

Gökçe, Ç., Özdemir, O., Kındıroğlu, A. A. ve Akarun, L., 2020, Score-level multi cue fusion for sign language recognition, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.

Grishchenko, I. ve Bazarevsky, V., 2020, Mediapipe holistic — simultaneous face, hand and pose prediction, on device.

Gupta, R. ve Kumar, A., 2021, Indian sign language recognition using wearable sensors and multi-label classification, *Computers and Electrical Engineering*, 90.

Gündüz, C. ve Polat, H., 2021, Turkish sign language recognition based on multistream data fusion, *Turkish Journal of Electrical Engineering and Computer Sciences*, 29 (2).

Gweth, Y. L., Plahl, C. ve Ney, H., 2012, Enhanced continuous sign language recognition using pca and neural network features, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*.

Hamada, Y., Shimada, N. ve Shirai, Y., 2004, Hand shape estimation under complex backgrounds for sign language recognition, *Proceedings - Sixth IEEE International Conference on Automatic Face and Gesture Recognition*.

Hamza, H. M. ve Wali, A., 2023, Pakistan sign language recognition: Leveraging deep learning models with limited dataset, *Machine Vision and Applications*, 34 (5).

He, K., Zhang, X., Ren, S. ve Sun, J., 2016, Deep residual learning for image recognition, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Hori, N. ve Yamamoto, M., 2024, Real-time isolated sign language recognition, *Frontiers of Artificial Intelligence, Ethics, and Multidisciplinary Applications*, Singapore, 445-458.

Huang, M., Qian, H., Han, Y. ve Xiang, W., 2021, R(2+1)d-based two-stream cnn for human activities recognition in videos, *Chinese Control Conference, CCC*.

Husein, A. M., Calvin, Halim, D., Leo, R. ve William, 2019, Motion detect application with frame difference method on a surveillance camera, *Journal of Physics: Conference Series*.

Ibrahim, N. B., Zayed, H. H. ve Selim, M. M., 2019, Advances, challenges and opportunities in continuous sign language recognition, *Journal of Engineering and Applied Sciences*, 15 (5).

Imran, J. ve Raman, B., 2020, Deep motion templates and extreme learning machine for sign language recognition, *Visual Computer*, 36 (6).

Inik, Ö. ve Ülker, E., 2017, Derin öğrenme ve görüntü analizinde kullanılan derin öğrenme modelleri.

Ioffe, S. ve Szegedy, C., 2015, Batch normalization: Accelerating deep network training by reducing internal covariate shift, *32nd International Conference on Machine Learning, ICML 2015*.

Ioffe, S. ve Christian, S., 2017, Batch normalization: Accelerating deep network training by reducing internal covariate shift sergey, *Journal of Molecular Structure*, 1134.

Irasiak, A., Kozak, J., Piasecki, A. ve Stęclik, T., 2023, Processing real-life recordings of facial expressions of polish sign language using action units, *Entropy*, 25 (1).

Javaid, S. ve Rizvi, S., 2023, Manual and non-manual sign language recognition framework using hybrid deep learning techniques, *Journal of Intelligent and Fuzzy Systems*, 45 (3).

Jebali, M., Dakhli, A. ve Bakari, W., 2024, Deep learning-based sign language recognition system using both manual and non-manual components fusion, *AIMS Mathematics*, 9 (1), 2105-2122.

Ji, S., Xu, W., Yang, M. ve Yu, K., 2013, 3d convolutional neural networks for human action recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35 (1).

Jolliffe, I. T., 2002, Principal component analysis for special types of data, Springer, p.

Jolliffe, I. T. ve Cadima, J., 2016, Principal component analysis: A review and recent developments, *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, 374 (2065), 20150202.

Kerkhof, M., Wu, L., Perin, G. ve Picek, S., 2023, No (good) loss no gain: Systematic evaluation of loss functions in deep learning-based side-channel analysis, *Journal of Cryptographic Engineering*, 13 (3), 311-324.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M. ve Tang, P. T. P., 2016, On large-batch training for deep learning: Generalization gap and sharp minima, *arXiv preprint arXiv:1609.04836*.

Kılıçarslan, S., Adem, K. ve Çelik, M., 2021, An overview of the activation functions used in deep learning algorithms. J new results sci 10 (3): 75–88.

Kindiroglu, A. A., Ozdemir, O. ve Akarun, L., 2019, Temporal accumulative features for sign language recognition, *Proceedings - 2019 International Conference on Computer Vision Workshop, ICCVW 2019*.

Kındıroglu, A. A., Özdemir, O. ve Akarun, L., 2023, Aligning accumulative representations for sign language recognition, *Machine Vision and Applications*, 34 (1).

Kingma, D. P. ve Ba, J., 2014, Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980*.

Konstantinidis, D., Dimitropoulos, K. ve Daras, P., 2018a, A deep learning approach for analyzing video and skeletal features in sign language recognition, *IST 2018 - IEEE International Conference on Imaging Systems and Techniques, Proceedings*.

Konstantinidis, D., Dimitropoulos, K. ve Daras, P., 2018b, Sign language recognition based on hand and body skeletal data, *3DTV-Conference*.

Kour, K. P. ve Mathew, L., 2017, Sign language recognition using image processing, *International Journal of Advanced Research in Computer Science and Software Engineering*, 7 (8).

Köpüklü, O., Kose, N., Gunduz, A. ve Rigoll, G., 2021, Resource efficient 3d convolutional neural networks.

Krizhevsky, A., Sutskever, I. ve Hinton, G. E., 2012, Imagenet classification with deep convolutional neural networks, *Advances in neural information processing systems*.

Kumar, P., Roy, P. P. ve Dogra, D. P., 2018, Independent bayesian classifier combination based sign language recognition using facial expression, *Information Sciences*, 428.

Laines, D., Gonzalez-Mendoza, M., Ochoa-Ruiz, G. ve Bejarano, G., 2023, Isolated sign language recognition based on tree structure skeleton images, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 276-284.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. ve Jackel, L. D., 1989, Backpropagation applied to handwritten zip code recognition, *Neural Computation*, 1 (4), 541-551.

LeCun, Y., Bengio, Y. ve Hinton, G., 2015, Deep learning, *Nature*, 521 (7553), 436-444.

Li, T. H. S., Kao, M. C. ve Kuo, P. H., 2016, Recognition system for home-service-related sign language using entropy-based k-means algorithm and abc-based hmm, *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 46 (1).

Li, Z., Wang, S. H., Fan, R. R., Cao, G., Zhang, Y. D. ve Guo, T., 2019, Teeth category classification via seven-layer deep convolutional neural network with max pooling and global average pooling, *International Journal of Imaging Systems and Technology*, 29 (4), 577-583.

Lin, M., Chen, Q. ve Yan, S., 2013, Network in network, *arXiv preprint arXiv:1312.4400*.

Lin, R., 2022, Analysis on the selection of the appropriate batch size in cnn neural network, *2022 International Conference on Machine Learning and Knowledge Engineering (MLKE)*, 106-109.

Liu, Y., Jiang, X., Yu, X., Ye, H., Ma, C., Wang, W. ve Hu, Y., 2023, A wearable system for sign language recognition enabled by a convolutional neural network, *Nano Energy*, 116, 108767-108767.

Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T. ve Xie, S., 2022, A convnet for the 2020s, *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11966-11976.

Loukili, S. E., Ezzati, A., Alla, S. B. ve Zraibi, B., 2022, Reducing the training time of deep learning models using synchronous sgd and large batch size, *2022 International Conference on Intelligent Systems and Computer Vision (ISCV)*, 1-3.

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J., Chang, W.-T., Hua, W., Georg, M. ve Grundmann, M., 2019, Mediapipe: A framework for building perception pipelines, *CoRR*, abs/1906.08172.

Ma, Y., Xu, T. ve Kim, K., 2022, Two-stream mixed convolutional neural network for american sign language recognition, *Sensors*, 22 (16).

Madani, H. ve Nahvi, M., 2013, Isolated dynamic persian sign language recognition based on camshift algorithm and radon transform, *1st Iranian Conference on Pattern Recognition and Image Analysis, PRIA 2013*.

Maitra, S., Ojha, R. K. ve Ghosh, K., 2018, Impact of convolutional neural network input parameters on classification performance, *2018 4th International Conference for Convergence in Technology (I2CT)*, 1-5.

Marais, M., Brown, D., Connan, J. ve Boby, A., 2022a, An evaluation of hand-based algorithms for sign language recognition, *2022 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, 1-6.

Marais, M., Brown, D., Connan, J., Boby, A. ve Kuhlane, L. L., 2022b, Investigating signer-independent sign language recognition on the lsa64 dataset, *Southern Africa Telecommunication Networks and Applications Conference (SA TNAC)*.

Masood, S., Srivastava, A., Thuwal Harish, C. ve Ahmad, M., 2018, Real-time sign language gesture (word) recognition from video sequences using cnn and rnn, *Intelligent Engineering Informatics*, Singapore, 623-632.

Masters, D. ve Luschi, C., 2018, Revisiting small batch training for deep neural networks. Arxiv 2018, *arXiv preprint arXiv:1804.07612*.

Mohana, R. M. ve Reddy, A. R. M., 2014, Signer-independent slr system using pca and multi-class svm, *Praise Worthy Prize International Review on Computers and Software (IRESOE)*, 9 (12), 1946-1955.

Mukushev, M., Sabyrov, A., Imashev, A., Koishybay, K., Kimmelman, V. ve Sandygulova, A., 2020, Evaluation of manual and non-manual components for sign language recognition, *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*.

Murphy, K. P., 2012, Machine learning: A probabilistic perspective, MIT press, p.

Naeem, H. B., Murtaza, F., Yousaf, M. H. ve Velastin, S. A., 2020, Multiple batches of motion history images (mb-mhis) for multi-view human action recognition, *Arabian Journal for Science and Engineering*, 45 (8).

Noor, N. M., Al Bakri Abdullah, M. M., Yahaya, A. S. ve Ramli, N. A., 2015, Comparison of linear interpolation method and mean method to replace the missing values in environmental data set, *Materials science forum*, 278-281.

Oz, C. ve Leu, M. C., 2011, American sign language word recognition with a sensory glove using artificial neural networks, *Engineering Applications of Artificial Intelligence*, 24 (7).

Özdemir, O., Kindiroglu, A. A., Camgöz, N. C. ve Akarun, L., 2020, Bosphorussign22k sign language recognition dataset, *CoRR*, abs/2004.01283.

Özdemir, O., Baytaş, İ. M. ve Akarun, L., 2023, Multi-cue temporal modeling for skeleton-based sign language recognition, *Frontiers in Neuroscience*, 17.

Perrone, M., Khan, H., Kim, C., Kyrillidis, A., Quinn, J. ve Salapura, V., 2019, Optimal mini-batch size selection for fast gradient descent. Arxiv, *arXiv preprint arXiv:1911.06459*.

Piao, X., Synn, D., Park, J. ve Kim, J.-K., 2023, Enabling large batch size training for dnn models beyond the memory limit while maintaining performance, *IEEE Access*.

Podder, K. K., Ezeddin, M., Chowdhury, M. E. H., Sumon, M. S. I., Tahir, A. M., Ayari, M. A., Dutta, P., Khandakar, A., Mahbub, Z. B. ve Kadir, M. A., 2023, Signer-independent arabic sign language recognition system using deep learning model, *Sensors*, 23 (16).

Powers, D. M., 2020, Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation, *arXiv preprint arXiv:2010.16061*.

Prechelt, L., 1998, Automatic early stopping using cross validation: Quantifying the criteria, *Neural Networks*, 11 (4), 761-767.

Pu, J., Zhou, W. ve Li, H., 2019, Iterative alignment network for continuous sign language recognition, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Rahim, M. A., Shin, J. ve Yun, K. S., 2020, Hand gesture-based sign alphabet recognition and sentence interpretation using a convolutional neural network, *Annals of Emerging Technologies in Computing*, 4 (4).

Raj, R. D. ve Jasuja, A., 2018, British sign language recognition using hog, *2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science, SCEECS 2018*.

Ramachandram, D. ve Taylor, G. W., 2017, Deep multimodal learning: A survey on recent advances and trends, *IEEE signal processing magazine*, 34 (6), 96-108.

Rastgoo, R., Kiani, K. ve Escalera, S., 2021, Sign language recognition: A deep survey. Expert Systems with Applications. 164.

Ringnér, M., 2008, What is principal component analysis? Nature biotechnology. 26.

Rodríguez, J. ve Martínez, F., 2018, Towards on-line sign language recognition using cumulative sd-vlad descriptors, *Communications in Computer and Information Science*.

Ronchetti, F., Quiroga, F., Estrebou, C., Lanzarini, L. ve Rosete, A., 2016a, Sign languague recognition without frame-sequencing constraints: A proof of concept on the argentinian sign language, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.

Ronchetti, F., Quiroga, F. ve Lanzarini, L., 2016b, Lsa64 : An argentinian sign language dataset, *Congreso Argentino de Ciencias de la Computacion (CACIC)*.

Ruder, S., 2016, An overview of gradient descent optimization algorithms, *arXiv preprint arXiv:1609.04747*.

Salehin, I. ve Kang, D.-K., 2023, A review on dropout regularization approaches for deep neural networks within the scholarly domain, *Electronics*, 12 (14), 3106.

Samaan, G. H., Wadie, A. R., Attia, A. K., Asaad, A. M., Kamel, A. E., Slim, S. O., Abdallah, M. S. ve Cho, Y. I., 2022, Mediapipe's landmarks with rnn for dynamic sign language recognition, *Electronics (Switzerland)*, 11 (19).

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. ve Chen, L. C., 2018, Mobilenetv2: Inverted residuals and linear bottlenecks, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Sarhan, N. ve Frintrop, S., 2023, Unraveling a decade: A comprehensive survey on isolated sign language recognition, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3210-3219.

Scabini, L. F. ve Bruno, O. M., 2023, Structure and performance of fully connected neural networks: Emerging complex network properties, *Physica A: Statistical Mechanics and its Applications*, 615, 128585.

Selvaraj, P., C, G. N., Kumar, P. ve Khapra, M. M., 2021, Openhands: Making sign language recognition accessible with pose-based pretrained models across languages, *CoRR*, abs/2110.05877.

Shaha, M. ve Pawar, M., 2018, Transfer learning for image classification, *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 656-660.

Sharma, A., Mittal, A., Singh, S. ve Awatramani, V., 2020, Hand gesture recognition using image processing and feature extraction techniques, *Procedia Computer Science*.

Sincan, O. M. ve Keles, H. Y., 2021, Using motion history images with 3d convolutional networks in isolated sign language recognition, *CoRR*, abs/2110.12396.

Singla, N., 2014, Motion detection based on frame difference method, *International Journal of Information & Computation Technology*, 4 (15).

Smith, L. N., 2017, Cyclical learning rates for training neural networks, *2017 IEEE winter conference on applications of computer vision (WACV)*, 464-472.

Sokolova, M. ve Lapalme, G., 2009, A systematic analysis of performance measures for classification tasks, *Information processing & management*, 45 (4), 427-437.

Sreemathy, R., Turuk, M., Kulkarni, I. ve Khurana, S., 2023, Sign language recognition using artificial intelligence, *Education and Information Technologies*, 28 (5).

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. ve Salakhutdinov, R., 2014, Dropout: A simple way to prevent neural networks from overfitting, *The journal of machine learning research*, 15 (1), 1929-1958.

Sutskever, I., Martens, J., Dahl, G. ve Hinton, G., 2013, On the importance of initialization and momentum in deep learning, *International conference on machine learning*, 1139-1147.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. ve Wojna, Z., 2016, Rethinking the inception architecture for computer vision, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Tan, M. ve Le, Q. V., 2019, Efficientnet: Rethinking model scaling for convolutional neural networks, *36th International Conference on Machine Learning, ICML 2019*.

Taylor, G. W., Fergus, R., LeCun, Y. ve Bregler, C., 2010, Convolutional learning of spatio-temporal features, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.

Terven, J., Cordova-Esparza, D. M., Ramirez-Pedraza, A. ve Chavez-Urbiola, E. A., 2023, Loss functions and metrics in deep learning. A review, *arXiv preprint arXiv:2307.02694*.

Ther, A., Pandit, B. K., Ganguly, A., Chakraborty, A. ve Banerjee, A., 2023, Resource-efficient vlsi architecture of softmax activation function for real-time inference in deep learning applications, *2023 International Symposium on Devices, Circuits and Systems (ISDCS)*, 01-05.

Tian, Y., Su, D., Lauria, S. ve Liu, X., 2022, Recent advances on loss functions in deep learning for computer vision, *Neurocomputing*, 497, 129-158.

Tian, Y., Han, F., Zhu, M., Xu, X. ve Li, Y., 2023, Research on sign language gesture division and gesture extraction in complex background, *International Conference on Computer Vision, Application, and Algorithm (CVAA 2022)*, 103-113.

Tran, D., Bourdev, L., Fergus, R., Torresani, L. ve Paluri, M., 2015, Learning spatiotemporal features with 3d convolutional networks, *2015 IEEE International Conference on Computer Vision (ICCV)*, 4489-4497.

Tran, D., Wang, H., Torresani, L., Ray, J., Lecun, Y. ve Paluri, M., 2018, A closer look at spatiotemporal convolutions for action recognition, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Usmani, I. A., Qadri, M. T., Zia, R., Alrayes, F. S., Saidani, O. ve Dashtipour, K., 2023, Interactive effect of learning rate and batch size to implement transfer learning for brain tumor classification, *Electronics*, 12 (4), 964.

Vapnik, V. N. ve Vapnik, V., 1998, Statistical learning theory.

Wang, F., Du, Y., Wang, G., Zeng, Z. ve Zhao, L., 2021a, (2+1)d-slr: An efficient network for video sign language recognition, *Neural Computing and Applications*.

Wang, T., Fujita, Y., Chang, X. ve Watanabe, S., 2021b, Streaming end-to-end asr based on blockwise non-autoregressive models, *arXiv preprint arXiv:2107.09428*.

Xu, B., Wang, N., Chen, T. ve Li, M., 2015, Empirical evaluation of rectified activations in convolutional network, *CoRR*, abs/1505.00853.

Yang, Q. ve Peng, J. Y., 2014, Chinese sign language recognition method based on depth image information and surf-bow, *Moshi Shibie yu Rengong Zhineng/Pattern Recognition and Artificial Intelligence*, 27 (8).

Yasir, F., Prasad, P. W. C., Alsadoon, A. ve Elchouemi, A., 2016, Sift based approach on bangla sign language recognition, *2015 IEEE 8th International Workshop on Computational Intelligence and Applications, IWCIA 2015 - Proceedings*.

Zahid, H., Rashid, M., Hussain, S., Azim, F., Syed, S. A. ve Saad, A., 2022, Recognition of urdu sign language: A systematic review of the machine learning classification, *PeerJ Computer Science*, 8, e883-e883.

Zhan, C., Duan, X., Xu, S., Song, Z. ve Luo, M., 2007, An improved moving object detection algorithm based on frame difference and edge detection, *Proceedings of the 4th International Conference on Image and Graphics, ICIG 2007*.

Zhang, E., Xue, B., Cao, F., Duan, J., Lin, G. ve Lei, Y., 2019, Fusion of 2d cnn and 3d densenet for dynamic gesture recognition, *Electronics (Switzerland)*, 8 (12).

Zhang, X. ve Li, X., 2019, Dynamic gesture recognition based on memp network, *Future Internet*, 11, 91-91.