MARMARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BEYİN MRI GÖRÜNTÜLERİNDEN DERİN ÖĞRENME VE TRANSFORMER MODELLERİ KULLANARAK İNME TESPİTİ VE ADAPTİF BÜYÜK DİL MODELLERİYLE OTOMATİK RADYOLOJİK RAPOR ÜRETİMİ

MERVE PINAR

DOKTORA TEZİ
Bilgisayar Mühendisliği Anabilim Dalı
Bilgisayar Mühendisliği Doktora Programı

DANIŞMAN
Doç. Dr. Ayşe Berna ALTINEL

EŞ-DANIŞMAN
Dr. Öğr. Üyesi Abdulsamet AKTAŞ

İSTANBUL, 2026

MARMARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BEYİN MRI GÖRÜNTÜLERİNDEN DERİN ÖĞRENME VE TRANSFORMER MODELLERİ KULLANARAK İNME TESPİTİ VE ADAPTİF BÜYÜK DİL MODELLERİYLE OTOMATİK RADYOLOJİK RAPOR ÜRETİMİ

MERVE PINAR

(723622702)

DOKTORA TEZİ
Bilgisayar Mühendisliği Anabilim Dalı
Bilgisayar Mühendisliği Doktora Programı

DANIŞMAN
Doç. Dr. Ayşe Berna ALTINEL

EŞ-DANIŞMAN
Dr. Öğr. Üyesi Abdulsamet AKTAŞ

İSTANBUL, 2026

MARMARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Marmara Üniversitesi Fen Bilimleri Enstitüsü Doktora Öğrencisi Merve PINAR' ın "Beyin MRI Görüntülerinden Derin Öğrenme Ve Transformer Modelleri Kullanarak İnme Tespiti Ve Adaptif Büyük Dil Modelleriyle Otomatik Radyolojik Rapor Üretimi" başlıklı tez çalışması, 18.05.2026 tarihinde savunulmuş ve jüri üyeleri tarafından başarılı bulunmuştur.

Jüri Üyeleri

Doç. Dr. Ayşe Berna ALTINEL (Danışman)
Marmara Üniversitesi ..... (İMZA) .....

Prof. Dr. Ali BULDU (Üye)
Marmara Üniversitesi ..... (İMZA) .....

Dr. Öğr. Üyesi Ali SARIKAŞ (Üye)
Marmara Üniversitesi ..... (İMZA) .....

Dr. Öğr. Üyesi M. Zahid GÜRBÜZ (Üye)
Marmara Üniversitesi ..... (İMZA) .....

Dr. Öğr. Üyesi Gözde ALP (Üye)
Marmara Üniversitesi ..... (İMZA) .....

ONAY

Marmara Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu'nun …….. tarih ve ……..sayılı kararı ile ……. derecesi alması onanmıştır.

Fen Bilimleri Enstitüsü Müdürü
Prof. Dr. Bülent MERTOĞLU

ÖNSÖZ

Marmara Üniversitesi Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı bünyesinde yürütülen bu tez çalışması, gerçekleştirilmesi sürecinde sağlanan teknik ve mali destekler sayesinde tamamlanmıştır. Bu kapsamda, 11558 numaralı projem için verdikleri destekten dolayı Marmara Üniversitesi Bilimsel Araştırma Projeleri Koordinasyon Birimi'ne teşekkürlerimi sunarım.

Doktora eğitimim boyunca, bilimsel bakış açımın gelişmesinde önemli katkıları bulunan, araştırma sürecimin her aşamasında yol gösteren, bilgi ve tecrübelerini benimle paylaşan kıymetli danışmanım Sayın Doç. Dr. Ayşe Berna ALTINEL'e en içten teşekkürlerimi sunarım. Ayrıca, tez çalışmam süresince teknik bilgi birikimi ve tüm desteği ile her aşamada katkı sağlayan, yönlendirmeleri ve yapıcı önerileri ile çalışmanın gelişimine önemli katkılar sunan Dr. Öğr. Üyesi Abdulsamet AKTAŞ'a da teşekkürlerimi sunarım.

Tez çalışmam boyunca Marmara Eğitim Araştırma Hastanesi Nöroloji Bölümü ile kurulan değerli iş birliği kapsamında verilerin sağlanması ve projeye sunduğu kıymetli katkılar için Prof. Dr. İpek MİDİ'ye; görüntülerin elde edilmesi ve etiketlenmesi sürecindeki özverili çalışmaları, titiz yaklaşımları ve her aşamadaki destekleri için Dr. Sedef Duygu ÖZGÜÇ'e içtenlikle teşekkür ederim.

Bu çalışma süresince TÜBİTAK 2211-A Yurt İçi Doktora Burs Programı kapsamında sağlanan burs desteği için Türkiye Bilimsel ve Teknolojik Araştırma Kurumu'na teşekkürlerimi sunarım. Ayrıca TÜBİTAK 1002 Hızlı Destek Programı kapsamında desteklenen proje aracılığıyla sağlanan araştırma altyapısı desteği için TÜBİTAK'a ayrıca teşekkür ederim.

Bu zorlu ve uzun süreçte her zaman yanımda olan, sabırları, anlayışları ve destekleri ile bana güç veren aileme en derin teşekkürlerimi sunarım.

Mayıs, 2026 **Merve PINAR**

İÇİNDEKİLER

	SAYFA
ÖNSÖZ	i
İÇİNDEKİLER	ii
ÖZET	iv
ABSTRACT	vi
YENİLİK BEYANI	vii
SEMBOLLER	vii
KISALTMALAR	viii
ŞEKİL LİSTESİ	x
TABLO LİSTESİ	xii
1. GİRİŞ	1
1.1. Manyetik Rezonans Görüntüleme ve Difüzyon Sekansları	2
1.2. Otomatik İnme Segmentasyonu	3
1.3. Tıbbi Görüntülerde Derin Öğrenme Yaklaşımları	5
1.4. Tıbbi Metin Üretimi ve Büyük Dil Modelleri	6
1.5. İlgili Çalışmalar	8
1.5.1. Görüntü Tabanlı İnme Tespiti Çalışmaları	8
1.5.2. Tıbbi Metin Analizi Çalışmaları	11
1.5.3. Tıbbi Görüntülerden Otomatik Rapor Üretimi Çalışmaları	14
1.6. Araştırmanın Amacı, Kapsamı ve Tezin Organizasyonu	16
2. MATERYAL VE YÖNTEM	17
2.1. Veri Seti ve Etik Süreç	17
2.2. Veri Etiketleme ve Doğrulama Süreci	18
2.3. Önerilen Sistem Mimarisi	19
2.4. Çoklu Girdi Temsilleri ve Kombinasyonları	22
2.5. Segmentasyon Modelleri	25
2.5.1. nnU-Net	26
2.5.2. SwinUNETR	26
2.5.3. 3D U-Net	27
2.6. Eğitim ve Test Süreci	28
2.7. Görüntü Değerlendirme Metrikleri	29
2.8. Otomatik Rapor Üretim Sistemi Genel Mimarisi	30
2.9. Görüntü Tabanlı Özellik Çıkarımı	31
2.10. Anatomik ve Klinik Bilgi Entegrasyonu	33
2.10.1. Segmentasyon Maskesinden Klinik Metriklerin Çıkarılması	33
2.10.2. Yerleşim Tarafı Bilgisinin Belirlenmesi	34
2.10.3. Anatomik Atlaslarla Bölge Eşlemesi	34

2.10.4. Vasküler Sulama Alanı Eşlemesi 35
2.10.5. ADC Değerinin Klinik Yoruma Dönüştürülmesi 36
2.10.6. Anatomik, Vasküler ve Klinik Bilgilerin Yapılandırılması 37
2.10.7. Sayısal Bulgulardan Klinik İfade Üretimi 38
2.10.8. Anatomik ve Klinik Bilginin Rapor Üretimindeki Rolü 39
2.11. Metin Verisinin Hazırlanması 39
2.12. Büyük Dil Modeli Seçimi ve Mimarisi 40
2.13. LLM Eğitimi ve Adaptasyonu 44
2.14. Otomatik Raporların Değerlendirilmesi ve Klinik Uyum Analizi 48
3. BULGULAR ve SONUÇLAR 58
3.1. Segmentasyon Sonuçları 58
3.2. Özellik Çıkarımı ve Anatomik Eşleme Sonuçları 73
3.3. Otomatik Rapor Üretim Sonuçları 78
4. TARTIŞMA

KAYNAKLAR
ÖZGEÇMİŞ

ÖZET

BEYİN MRI GÖRÜNTÜLERİNDEN DERİN ÖĞRENME VE TRANSFORMER MODELLERİ KULLANARAK İNME TESPİTİ VE ADAPTİF BÜYÜK DİL MODELLERİYLE OTOMATİK RADYOLOJİK RAPOR ÜRETİMİ

İnme, dünya genelinde ölüm ve kalıcı nörolojik sakatlığın en önemli nedenlerinden biri olup, erken tanı ve hızlı tedavi süreci hastaların klinik seyri açısından kritik öneme sahiptir. Özellikle akut iskemik inmenin değerlendirilmesinde manyetik rezonans görüntüleme (MRI) ve difüzyon ağırlıklı görüntüleme (DWI) sekansları önemli bir rol oynamaktadır. Ancak beyin MR görüntülerinin manuel olarak değerlendirilmesi ve radyolojik raporların hazırlanması zaman alıcı, uzman deneyimine bağımlı ve gözlemciler arası farklılıklara açık bir süreçtir. Bu nedenle otomatik görüntü analizi ve rapor üretim sistemlerinin geliştirilmesi klinik uygulamalar açısından büyük önem taşımaktadır.

Bu tez çalışmasında, beyin MR görüntülerinden akut iskemik inme lezyonlarının otomatik segmentasyonu, anatomik ve klinik özelliklerinin çıkarılması ve bu bilgiler kullanılarak otomatik radyolojik rapor üretilmesini sağlayan bütünleşik bir yapay zekâ sistemi geliştirilmiştir. Çalışma kapsamında difüzyon ağırlıklı görüntüleme ve görünür difüzyon katsayısı görüntüleri kullanılmış; segmentasyon aşamasında nnU-Net, SwinUNETR ve 3D U-Net mimarileri değerlendirilmiştir. Ayrıca kenar haritaları ve anatomik/vasküler ön bilgi kanalları kullanılarak farklı çoklu girdi kombinasyonlarının segmentasyon performansına etkisi incelenmiştir.

Elde edilen segmentasyon maskeleri üzerinden lezyon hacmi, anatomik lokalizasyon, yerleşim tarafi bilgisi ve vasküler sulama alanı gibi klinik açıdan önemli özellikler çıkarılmıştır. Anatomik eşleme sürecinde Harvard-Oxford atlasları ve vasküler atlaslardan yararlanılmıştır. Yapılandırılmış klinik bulgular daha sonra büyük dil modelleri kullanılarak Türkçe radyolojik raporlara dönüştürülmüştür. Model performansı Dice, IoU, BLEU, ROUGE, METEOR ve BERTScore metrikleri kullanılarak değerlendirilmiştir.

Elde edilen sonuçlar, önerilen yaklaşımın hem otomatik inme segmentasyonu hem de klinik açıdan anlamlı radyolojik rapor üretimi açısından başarılı sonuçlar verdiğini göstermektedir.

ABSTRACT

STROKE DETECTION FROM BRAIN MRI IMAGES USING DEEP LEARNING AND TRANSFORMER MODELS AND AUTOMATIC RADIOLOGICAL REPORT GENERATION WITH ADAPTIVE LARGE LANGUAGE MODELS

Stroke is one of the leading causes of death and permanent neurological disability worldwide, and early diagnosis together with rapid treatment is critically important for the clinical course of patients. In particular, magnetic resonance imaging (MRI) and diffusion-weighted imaging (DWI) sequences play a significant role in the evaluation of acute ischemic stroke. However, the manual assessment of brain MRI images and the preparation of radiological reports are time-consuming processes that are highly dependent on expert experience and prone to inter-observer variability. Therefore, the development of automatic image analysis and report generation systems has become of great importance for clinical applications.

In this thesis, an integrated artificial intelligence system was developed for the automatic segmentation of acute ischemic stroke lesions from brain MRI images, extraction of anatomical and clinical features, and generation of automatic radiological reports using this information. Within the scope of the study, DWI and apparent diffusion coefficient (ADC) images were utilized, and nnU-Net, SwinUNETR, and 3D U-Net architectures were evaluated during the segmentation stage. In addition, the effects of different multi-input combinations on segmentation performance were investigated by incorporating edge maps and anatomical/vascular prior channels.

Clinically significant features such as lesion volume, anatomical localization, lesion laterality, and vascular territory were extracted from the obtained segmentation masks. During the anatomical mapping process, Harvard-Oxford atlases and vascular atlases were utilized. The structured clinical findings were subsequently transformed into Turkish radiological reports using large language models. Model performance was evaluated using Dice, IoU, BLEU, ROUGE, METEOR, and BERTScore metrics.

The obtained results demonstrate that the proposed approach achieves successful performance in both automatic stroke segmentation and the generation of clinically meaningful radiological reports.

YENİLİK BEYANI

Bu tez çalışması, beyin manyetik rezonans görüntülerinden akut iskemik inme lezyonlarının otomatik olarak segmentasyonu, anatomik ve klinik bilgilerinin çıkarılması ve bu bilgilerin büyük dil modelleri kullanılarak otomatik radyolojik raporlara dönüştürülmesini sağlayan uçtan uca bütünleşik bir yapay zekâ sistemi önermektedir. Çalışma, tıbbi görüntü analizi ile doğal dil üretimi alanlarını tek bir genel akış içerisinde birleştirmesi açısından literatürdeki çalışmalardan ayrılmaktadır.

Literatürdeki çalışmaların büyük bir kısmı yalnızca inme segmentasyonu veya yalnızca tıbbi rapor üretimi üzerine odaklanırken, bu tez kapsamında görüntü işleme, anatomik bilgi çıkarımı, klinik yorumlama ve doğal dil tabanlı rapor üretimi süreçleri tek bir sistem içerisinde bütünleşik olarak ele alınmıştır. Böylece yalnızca lezyon segmentasyonu gerçekleştiren bir yapı yerine, görüntüden başlayarak klinik açıdan anlamlı radyolojik rapor üretimine kadar uzanan çok aşamalı bir yapay zekâ iş akışı geliştirilmiştir.

Çalışmanın önemli yeniliklerinden biri, DWI ve ADC görüntülerine ek olarak kenar haritaları (EDGE) ile anatomik/vasküler ön bilgi (PRIOR) kanallarının birlikte kullanılması ve bu çoklu girdi kombinasyonlarının farklı segmentasyon mimarileri üzerindeki etkilerinin sistematik olarak incelenmesidir. Bu kapsamda nnU-Net, SwinUNETR ve 3D U-Net modelleri aynı veri seti ve aynı eğitim protokolü altında karşılaştırılmış; çoklu kanal kullanımının segmentasyon performansına katkısı ayrıntılı biçimde analiz edilmiştir. Böylece hem sinir ağı tabanlı hem de Transformer tabanlı mimarilerin farklı girdi temsilleri altındaki davranışları karşılaştırmalı olarak değerlendirilmiştir.

Tezin bir diğer özgün yönü, segmentasyon maskelerinden yalnızca lezyon alanının belirlenmesiyle sınırlı kalmayıp, lezyon hacmi, anatomik lokalizasyon, yerleşim tarafı ve vasküler sulama alanı gibi klinik açıdan önemli bilgilerin otomatik olarak çıkarılmasıdır. Anatomik eşleme sürecinde Harvard-Oxford atlasları ve vasküler atlaslardan yararlanılarak görüntü tabanlı bulguların klinik bilgiye dönüştürülmesi sağlanmıştır.

Bu tez çalışmasının en önemli yeniliklerinden biri ise, elde edilen yapılandırılmış klinik bilgilerin büyük dil modelleri kullanılarak Türkçe radyolojik raporlara dönüştürülmesidir. Literatürde Türkçe inme radyoloji raporları üzerinde ince ayar yapılmış büyük dil modeli tabanlı kapsamlı bir çalışmanın bulunmaması, bu çalışmanın özgün değerini artırmaktadır. Bu kapsamda Türkçe radyolojik rapor veri seti kullanılarak büyük dil modelleri adapte edilmiş ve inme odaklı otomatik rapor üretimine yönelik alan-özgü bir sistem geliştirilmiştir. Ayrıca düşük maliyetli ve bellek verimli ince ayar yöntemleri kullanılarak büyük modellerin sınırlı donanım kaynaklarıyla eğitilebilmesi hedeflenmiştir.

Önerilen yaklaşımın, radyologların iş yükünün azaltılması, raporlama süreçlerinin hızlandırılması ve yapay zekâ destekli klinik karar destek sistemlerinin geliştirilmesine katkı sağlaması hedeflenmektedir.

Mayıs, 2026 **Merve PINAR**

SEMBOLLER

I_x : Görüntünün x ekseni yönündeki yoğunluk gradyanı

I_y : Görüntünün y ekseni yönündeki yoğunluk gradyanı

I_z : Görüntünün z ekseni yönündeki yoğunluk gradyanı

E : Kenar haritası

E_{norm} : Normalize edilmiş kenar haritası değeri

$\in$: Sıfıra bölme hatasını önlemek için kullanılan küçük sabit değer

X_{EDGE} : Kenar haritası içeren üç kanallı giriş tensörü

T_0 : Başlangıç hizalama dönüşüm parametresi

A : Atlas görüntüsü

I : Hasta DWI görüntüsü

$T*$: Optimize edilmiş en iyi uzaysal dönüşüm

MI : Karşılıklı bilgi (Mutual Information) benzerlik metriği

X_{PRIOR} : Atlas bilgisi içeren üç kanallı giriş tensörü

X_{IMC} : Entegre Çok Kanallı (Integrated Multi-Channel, IMC) giriş tensörü

P : Tahmin edilen lezyon maskesi

G : Gerçek (ground truth) lezyon maskesi

V : Lezyon hacmi

D_{max} : Lezyonun maksimum aksiyal çapı

$N_{sağ}$: Sağ hemisferde bulunan toplam lezyon voxel sayısı

N_{sol} : Sol hemisferde bulunan toplam lezyon voxel sayısı

M : Üç boyutlu segmentasyon maskesi

t : Öteleme vektörü

P_i : i'nci anatomik bölgenin lezyon ile örtüşme yüzdesi

S_k : k'ncı vasküler sulama alanının toplam skoru

W : Büyük dil modelinin ince ayar sonrası güncellenmiş ağırlık matrisi

BP : Kısalık cezası

p_n : Güncellenmiş n-gram kesinlik değeri

c : Üretilen metnin uzunluğu

r : Referans metnin uzunluğu

m : Referans metindeki toplam kelime sayısı

β : Kesinlik ve duyarlılık arasındaki ağırlık dengesini belirleyen parametre

KISALTMALAR

3D : 3 Boyutlu

AAN : American Academy of Neurology

ADC : Görünür Difüzyon Katsayısı

AERMNet : Attention-Enhanced Relational Memory Network

AIS : Akut İskemik İnme

ANOVA : Analysis of Variance

AUC : Area Under the Curve

BERT : Bidirectional Encoder Representations from Transformers

BiLSTM : Bidirectional Long Short-Term Memory

BLEU : Bilingual Evaluation Understudy

BT : Bilgisayarlı Tomografi

CSCA : Channel and Space Compound Attention

DBSCAN : Density-Based Spatial Clustering of Applications with Noise

DeBERTa : Decoding-enhanced BERT with Disentangled Attention

DICOM : Digital Imaging and Communications in Medicine

DKI : Difüzyon Kurtosis Görüntüleme

DSE : Double Squeeze-and-Excitation

DTI : Difüzyon Tensör Görüntüleme

DWI : Difüzyon Ağırlıklı Görüntüleme

FLAIR : Fluid-Attenuated İnversion Recovery

ISLES : Ischemic Stroke Lesion Segmentation

GPT : Generative Pre-trained Transformer

GRU : Gated Recurrent Unit

GUI : Graphical User Interface

KF+MCW : Kessler Foundation + Medical College of Wisconsin

LLaMA2 : Language Model Meta AI 2

LLM : Büyük Dil Modelleri (Large Language Models)

LR : Learning Rate

LSTM : Long Short-Term Memory

MAIF : Multi-Attribute Identification and Fusion

METEOR : Metric for Evaluation of Translation with Explicit ORdering

MIMIC-III : Medical Information Mart for Intensive Care-III

MRI : Manyetik Rezonans Görüntüleme

MVFF : Çok-Vektörlü Öznitelik Füzyonu (Multi-Vector Feature Fusion)

NIfTI : Neuroimaging Informatics Technology Initiative

NIH : National Institutes of Health

NLP : Doğal Dil İşleme

PWI : Perfüzyon Ağırlıklı Görüntüleme

QLoRA : Quantized Low-Rank Adaptation

RAS : Right–Anterior–Superior

R-CNN : Region-Based Convolutional Neural Network

RNN : Recurrent Neural Networks

ROC : Receiver Operating Characteristic

ROUGE-L : Recall-Oriented Understudy for Gisting Evaluation - Longest

SBERT : Sentence-BERT

SelfONN : Self-Organized Operational Neural Networks

SHAP : SHapley Additive exPlanations

SimCSE : Simple Contrastive Sentence Embedding

T5 : Text-To-Text Transfer Transformer

TF-IDF : Term Frequency – Inverse Document Frequency

TF-RF : Term Frequency – Relevance Frequency

TMS : Tıbbi metin sınıflandırma

WHO : Dünya Sağlık Örgütü

YOLO : You Only Look Once

ŞEKİL LİSTESİ

Şekil 1.1. İskemik ve hemorajik inmenin şematik gösterimi. 1
Şekil 1.2. Akut iskemik inmede farklı MRI sekanslarının karşılaştırılması [12]. 2
Şekil 1.3. İnme lezyonunun manuel ve otomatik segmentasyonunun karşılaştırılması [24].4
Şekil 1.4. Makine öğrenmesi ve derin öğrenme yaklaşımlarının karşılaştırılması. 6
Şekil 1.5. Büyük dil modelleri ile tıbbi metin ve rapor üretim süreci. 7
Şekil 2.1. 3D Slicer ortamında gerçekleştirilen manuel segmentasyon süreci. 18
Şekil 2.2. Segmentasyon maskesinin üç boyutlu uzaydaki konumunun gösterimi. 19
Şekil 2.3. Önerilen sistemin genel mimarisi. 21
Şekil 2.4. Modalitelere ait örnek MRI görüntüleri. 25
Şekil 2.5. nnU-Net mimarisi. 26
Şekil 2.6. SwinUNETR mimarisi. 27
Şekil 2.7. 3D U-Net mimarisi. 28
Şekil 2.8. Segmentasyon modeli eğitim ve değerlendirme metodolojisi. 29
Şekil 2.9. MRI tabanlı özellik çıkarım ve yapılandırılmış klinik bilgi oluşturma aşamaları.32
Şekil 2.10. Otomatik radyolojik rapor üretiminin genel aşamaları. 39
Şekil 2.11. QLoRA mimarisi. 44
Şekil 2.12. Eğitim veri seti oluşturma süreci. 45
Şekil 2.13. QLoRA tabanlı ince ayar iş akışı. 48
Şekil 3.1. nnU-Net modeli için farklı giriş konfigürasyonlarına göre Dice, IoU ve vDSC metriklerinin karşılaştırılması. 59
Şekil 3.2. nnU-Net modeli ve DWI+ADC giriş konfigürasyonu kullanılarak elde edilen segmentasyon sonuçlarının örnek görselleri. 60
Şekil 3.3. nnU-Net modeli ve DWI giriş konfigürasyonu kullanılarak elde edilen segmentasyon sonuçlarının örnek görselleri. 61
Şekil 3.4. nnU-Net modeli ile elde edilen segmentasyon sonuçlarının üç düzlem ve üç boyutlu görünüm üzerinden karşılaştırılması. 62
Şekil 3.5. SwinUNETR modeli için farklı giriş konfigürasyonlarına göre Dice, IoU ve vDSC metriklerinin karşılaştırılması. 63
Şekil 3.6. SwinUNETR modeli ve DWI giriş konfigürasyonu kullanılarak elde edilen segmentasyon sonuçlarının örnek görselleri. 64
Şekil 3.7. SwinUNETR modeli FULL giriş konfigürasyonu kullanılarak elde edilen segmentasyon sonuçlarının örnek görselleri. 64
Şekil 3.8. SwinUNETR modeli ile elde edilen segmentasyon sonuçlarının üç düzlem ve üç boyutlu görünüm üzerinden karşılaştırılması. 65
Şekil 3.9. 3D-U-Net modeli için farklı giriş konfigürasyonlarına göre Dice, IoU ve vDSC metriklerinin karşılaştırılması. 67

Şekil 3.10. 3D U-Net modeli DWI giriş konfigürasyonu kullanılarak elde edilen segmentasyon sonuçlarının örnek görselleri. 67
Şekil 3.11. 3D U-Net modeli FULL giriş konfigürasyonu kullanılarak elde edilen segmentasyon sonuçlarının örnek görselleri. 68
Şekil 3.12. 3D U-Net modeli ile elde edilen segmentasyon sonuçlarının üç düzlem ve üç boyutlu görünüm üzerinden karşılaştırılması. 69
Şekil 3.13. 141 hastaya ait lezyon hacim dağılımı. 74
Şekil 3.14. Veri setindeki lezyon hacmi ve maksimum lezyon çapı dağılımlarının kutu grafikleri. 75
Şekil 3.15. Veri setindeki lezyon hacim kategorilerinin dağılımı. 76
Şekil 3.16. HeaRT çerçevesindeki küme destekli özellik zenginleştirme iş akışı. 86

TABLO LİSTESİ

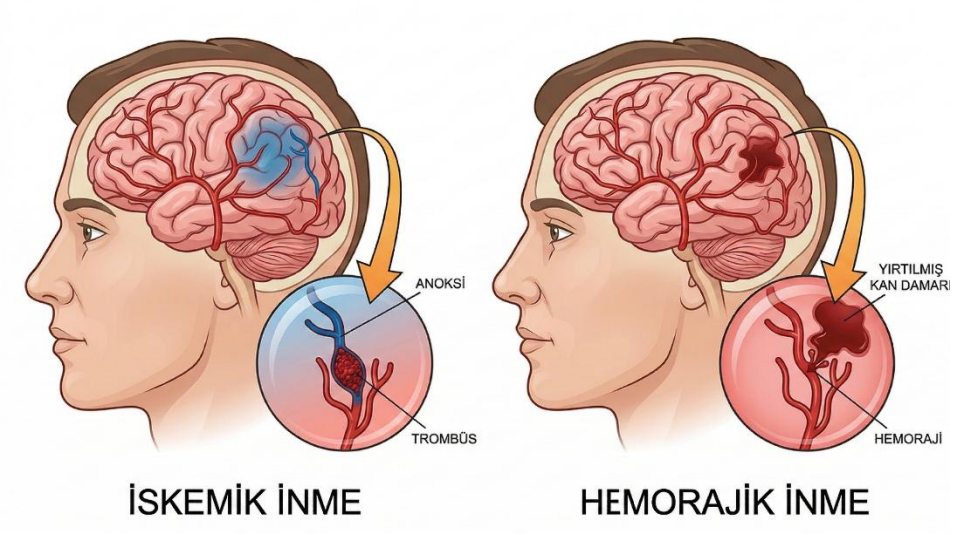
Tablo 1.1. Görüntü tabanlı inme tespiti çalışmalarına yönelik performans karşılaştırması. 10
Tablo 1.2. Tıbbi metin analizi çalışmalarına yönelik performans karşılaştırması. 13
Tablo 1.3. Tıbbi görüntülerden otomatik rapor üretimi çalışmalarına yönelik performans karşılaştırması. 16
Tablo 2.1. Görüntülere ait örnek raporlar. 20
Tablo 2.2. Görüntülerden çıkarılan özellikler. 32
Tablo 2.3. ADC değer aralıklarına göre difüzyon kısıtlanmasının klinik yorumlanması. 37
Tablo 2.4. Görüntü tabanlı analizden elde edilen yapılandırılmış klinik özellikler ve klinik karşılıkları. 37
Tablo 2.5. Yapılandırılmış klinik bilgilerin örnek çıktı formatı. 38
Tablo 2.6. Lezyon hacmine göre nitel rapor ifadesi. 38
Tablo 2.7. Deneysel olarak değerlendirilen açık kaynak modeller. 41
Tablo 2.8. Değerlendirme dışı bırakılan açık kaynak modeller. 42
Tablo 2.9. Ticari (kapalı kaynak) referans modeller. 43
Tablo 2.10. QLoRA hiperparametre detayları. 46
Tablo 2.11. QLoRA eğitim hiperparametre detayları. 47
Tablo 2.12. Uzman değerlendirmesinde kullanılan Likert ölçeği boyutları. 57
Tablo 3.1. nnU-Net için farklı giriş konfigürasyonlarına göre segmentasyon sonuçları. 59
Tablo 3.2. SwinUNETR için farklı giriş konfigürasyonlarına göre segmentasyon sonuçları. 62
Tablo 3.3. 3D U-Net için farklı giriş konfigürasyonlarına göre segmentasyon sonuçları. 66
Tablo 3.4. Cohen's d etki büyüklüğü yorumlama eşikleri. 70
Tablo 3.5. DWI+ADC konfigürasyonunda modeller arası Dice skoru karşılaştırması (n=150, paired t-test). 70
Tablo 3.6. Her model için en iyi konfigürasyon ile diğer konfigürasyonlar arasındaki Dice skoru farkları. 71
Tablo 3.7. DWI+ADC konfigürasyonunda diğer metrikler için modeller arası Cohen's d değerleri. 72
Tablo 3.8. Eğitim ve Doğrulama Seti Karakteristikleri. 73
Tablo 3.9. Lezyon Hacmi İstatistikleri (n=141). 74
Tablo 3.10. Maksimum Lezyon Çapı İstatistikleri (n=141). 75
Tablo 3.11. Niteliksel hacim kategorilerine göre hasta dağılımı. 76
Tablo 3.12. Lezyon yerleşim yeri dağılımı. 77
Tablo 3.13. Vasküler sulama alanına göre lezyon dağılımı. 77
Tablo 3.14. Modellerin BLEU skorları karşılaştırması. 78
Tablo 3.15. Modellerin ROUGE skorları karşılaştırması. 79
Tablo 3.16. Altı modelin genel performans karşılaştırması. 80
Tablo 3.17. Likert sonuçları. 83
Tablo 3.18. Medical Abstracts veri setinin sınıf dağılımı. 84
Tablo 3.19. Biomedical Text Dataset veri setinin sınıf dağılımı. 85
Tablo 3.20. CEFA sürecinde kullanılan kümeleme algoritmaları ve parametre ayarları. 86
Tablo 3.21. Medical Abstracts veri setinde TF-IDF temsili ile sınıflandırma sonuçları. 87

Tablo 3.22. Medical Abstracts veri setinde TF-IDF + SBERT füzyonu ile sınıflandırma sonuçları. 87
Tablo 3.23. Medical Abstracts veri setinde HeaRT tam konfigürasyonu ile sınıflandırma sonuçları. 88
Tablo 3.24. Medical Abstracts veri setinde baseline model (BERT) sonuçları. 88
Tablo 3.25. Medical Abstracts veri setinde HeaRT ve baseline model arasındaki eşleştirilmiş t-testi. 89
Tablo 3.26. Biomedical Text Dataset'te TF-RF temsili ile sınıflandırma sonuçları. 89
Tablo 3.27. Biomedical Text Dataset'te TF-RF + SBERT füzyonu ile sınıflandırma sonuçları. 90
Tablo 3.28. Biomedical Text Dataset'te HeaRT tam konfigürasyonu ile sınıflandırma sonuçları. 90
Tablo 3.29. Biomedical Text Dataset'te baseline model (BERT) sonuçları. 91
Tablo 3.30. Biomedical Text Dataset'te HeaRT ve baseline model arasındaki eşleştirilmiş t-testi. 91

1. GİRİŞ

İnme, beyne giden kan akışının ani ve beklenmedik şekilde kesintiye uğraması sonucu ortaya çıkan, geniş spektrumda nörolojik bozukluklara yol açabilen ciddi bir serebrovasküler hastalıktır [1]. Dünya Sağlık Örgütü (WHO) inmeyi, beynin belirli bir bölgesine giden kan akışının ani ve hızlı bir şekilde kesilmesi sonucunda, o bölgedeki sinir hücrelerinin yeterli oksijen ve besin alamayarak hasar görmesi ya da ölmesiyle sonuçlanan nörolojik bir durum olarak tanımlamıştır [2, 3]. Bu durum, beynin ilgili bölgesinin kontrol ettiği vücut fonksiyonlarında ani kayıplara yol açar. İnme, genellikle kalıcı fiziksel engellere, bilişsel yeti kayıplarına, konuşma ve dil bozukluklarına, duygusal ve psikolojik sorunlara, görme kaybı ya da bozukluklarına, yutma zorluklarına ve hatta ölümle sonuçlanabilecek ciddi komplikasyonlara neden olabilir [4]. Tüm bu etkiler, bireyin yaşam kalitesini önemli ölçüde düşürür ve hem hasta hem de yakın çevresi için uzun süreli bakım ve rehabilitasyon sürecini gerektirebilir. Bu nedenle, inmenin erken evrede tanınması ve hızlı müdahale ile tedavi sürecinin başlatılması, oluşabilecek kalıcı hasarların önlenmesi açısından kritik önem taşımaktadır [5].

İnme, oluş mekanizmasına göre iki ana gruba ayrılmaktadır (Şekil 1.1). Vakaların yaklaşık %80–87'sini oluşturan iskemik inme, beyne giden kan akışının bir pıhtı nedeniyle engellenmesi sonucu ortaya çıkar. Bu pıhtı, beyin damarının içinde oluşabilir (trombüs) ya da vücudun başka bir bölgesinden gelerek damarı tıkayabilir (emboli). Vakaların yaklaşık %13–15'ini oluşturan hemorajik inme ise, beyin damarlarından birinin yırtılması sonucu meydana gelir ve beyin dokusu içine veya çevresine kanama oluşmasına neden olur [6].



Şekil 1.1. İskemik ve hemorajik inmenin şematik gösterimi.

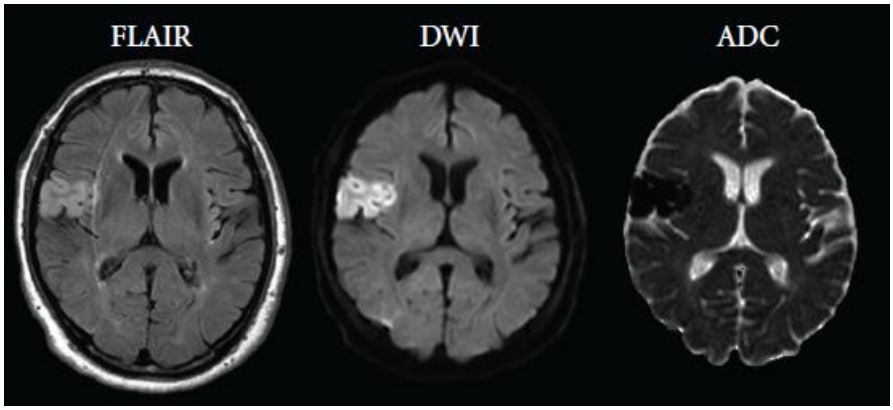
Türkiye'de inme hem ölüm oranları hem de kalıcı sakatlık riski açısından önemli bir halk sağlığı problemidir. Global Burden of Disease verilerine göre, 2019 yılında Türkiye'de yaklaşık 125.345 yeni inme vakası görülmüş olup bu değer yüz binde yaklaşık 154'e karşılık gelmektedir. Aynı yıl ülkede yaklaşık 1.080.380 kişinin inme ile yaşadığı (%1,3 prevalans) tahmin edilmektedir. İnme nedeniyle 48.947 kişi hayatını kaybetmiş, ayrıca inme kaynaklı ölüm ve sakatlıklar nedeniyle toplam yaklaşık 993.082 yaşam yılı kaybı meydana gelmiştir [7]. Yaş dağılımı incelendiğinde, inmelerin %17,4'ünün 50 yaşın altında, %58,5'inin ise 70 yaşın altında görüldüğü; vakaların %54,3'ünün kadınlarda ortaya çıktığı

bildirilmektedir. Bu veriler, inmenin yalnızca ileri yaş grubunu değil, daha genç yaş gruplarını da etkileyen ciddi bir sağlık sorunu olduğunu göstermektedir [7].

İnme sonrası rehabilitasyon süreci, hastaların kaybettikleri motor, bilişsel ve dil fonksiyonlarını yeniden kazanabilmeleri açısından vazgeçilmez olmakla birlikte, genellikle uzun, yorucu ve çok boyutlu bir tedavi dönemini gerektirmektedir. İnme sonucunda ortaya çıkan felç, denge kaybı, konuşma ve yutma güçlükleri gibi kalıcı nörolojik defisitler; aylar hatta bazı olgularda yıllar sürebilen yoğun rehabilitasyon programlarını zorunlu kılmaktadır [8]. Bu süreç yalnızca hastalar için fiziksel ve psikolojik açıdan yıpratıcı olmakla kalmayıp, aile bireyleri üzerinde bakım yükü oluşturmakta ve sağlık sistemi açısından da ciddi ekonomik ve iş gücü kayıplarına neden olmaktadır. Rehabilitasyonun etkinliği erken ve yoğun müdahaleyle artmakla birlikte, ağır hasar gelişmiş vakalarda tam fonksiyonel bağımsızlığa ulaşmak her zaman mümkün olmayabilmektedir. Bu nedenle inmenin erken dönemde doğru ve hızlı biçimde tanınması, uygun tedavinin gecikmeden başlatılması ve kalıcı nörolojik hasarın en aza indirilmesi hem bireysel yaşam kalitesinin korunması hem de uzun süreli rehabilitasyon yükünün azaltılması açısından kritik öneme sahiptir.

1.1. Manyetik Rezonans Görüntüleme ve Difüzyon Sekansları

İnme, dünya genelinde mortalite ve morbiditenin önde gelen nedenlerinden biri olmaya devam etmekte olup, hızlı ve doğru görüntüleme yöntemleri tanı, tedavi planlaması ve klinik seyir değerlendirmesi açısından büyük önem taşımaktadır. Bu bağlamda manyetik rezonans görüntüleme (MRI), akut iskemik inmenin değerlendirilmesinde en önemli görüntüleme yöntemlerinden biri olarak öne çıkmaktadır (Şekil 1.2). Özellikle difüzyon ağırlıklı görüntüleme (Diffusion Weighted Imaging, DWI), serebral iskeminin erken dönemde saptanmasını sağlayan temel sekanslardan biri olarak kabul edilmektedir [9, 10]. DWI sekansı, iskemik olaydan kısa süre sonra ortaya çıkan hücresel düzeydeki su difüzyonu değişimlerini tespit edebilme yeteneği sayesinde, inmenin çok erken evrelerinde bile lezyonların belirlenmesine olanak sağlamaktadır. Bu özelliği sayesinde DWI, akut iskemik inmenin tanısında bilgisayarlı tomografi (BT) gibi diğer görüntüleme yöntemlerine kıyasla daha yüksek duyarlılık ve hassasiyet sağlayan bir yöntem olarak öne çıkmaktadır [11].



Şekil 1.2. Akut iskemik inmede farklı MRI sekanslarının karşılaştırılması [12].

2010 yılında yayımlanan American Academy of Neurology (AAN) kılavuzunda, akut iskemik inmenin en doğru şekilde tanınabilmesi için DWI sekansının uygulanması önerilmiş ve bu öneri DWI'ın BT'ye kıyasla üstün duyarlılık ve hassasiyete sahip olduğunu gösteren çok sayıda çalışma ile desteklenmiştir [13]. Günümüzde DWI, enfarkt çekirdeğinin belirlenmesi, akut iskeminin inme benzeri klinik tablolarından ayırt edilmesi ve tedavi kararlarının yönlendirilmesi açısından altın standart görüntüleme yöntemlerinden biri olarak kabul edilmektedir [14, 15].

DWI görüntülerinin yorumlanmasında görünür difüzyon katsayısı (Apparent Diffusion Coefficient, ADC) haritaları önemli bir tamamlayıcı rol oynamaktadır. ADC haritaları, DWI görüntülerinde gözlenen sinyal değişikliklerinin doğrulanmasına ve gerçek difüzyon kısıtlılığının değerlendirilmesine olanak sağlamaktadır. Akut iskemik bölgelerde hücre içi ödem nedeniyle su moleküllerinin hareketi kısıtlanmakta; bu durum DWI görüntülerinde yüksek sinyal yoğunluğu, ADC haritalarında ise düşük difüzyon değerleri ile karakterize edilmektedir [16]. Bu nedenle DWI ve ADC sekanslarının birlikte değerlendirilmesi, akut iskemik lezyonların güvenilir bir şekilde tanımlanmasını ve diğer patolojik durumların ayırt edilmesini mümkün kılmaktadır.

DWI ve ADC sekanslarının yanı sıra, fluid-attenuated inversion recovery (FLAIR) ve perfüzyon ağırlıklı görüntüleme (Perfusion Weighted Imaging, PWI) gibi diğer MRI sekansları da inme hastalarının değerlendirilmesinde önemli bilgiler sunmaktadır. Ayrıca difüzyon kurtosis görüntüleme (Diffusion Kurtosis Imaging, DKI) ve difüzyon tensör görüntüleme (Diffusion Tensor Imaging, DTI) gibi gelişmiş difüzyon teknikleri, beyin dokusundaki mikroyapısal değişikliklerin daha ayrıntılı biçimde incelenmesine olanak sağlayarak inme hastalarının kapsamlı değerlendirilmesine katkıda bulunmaktadır [17, 18].

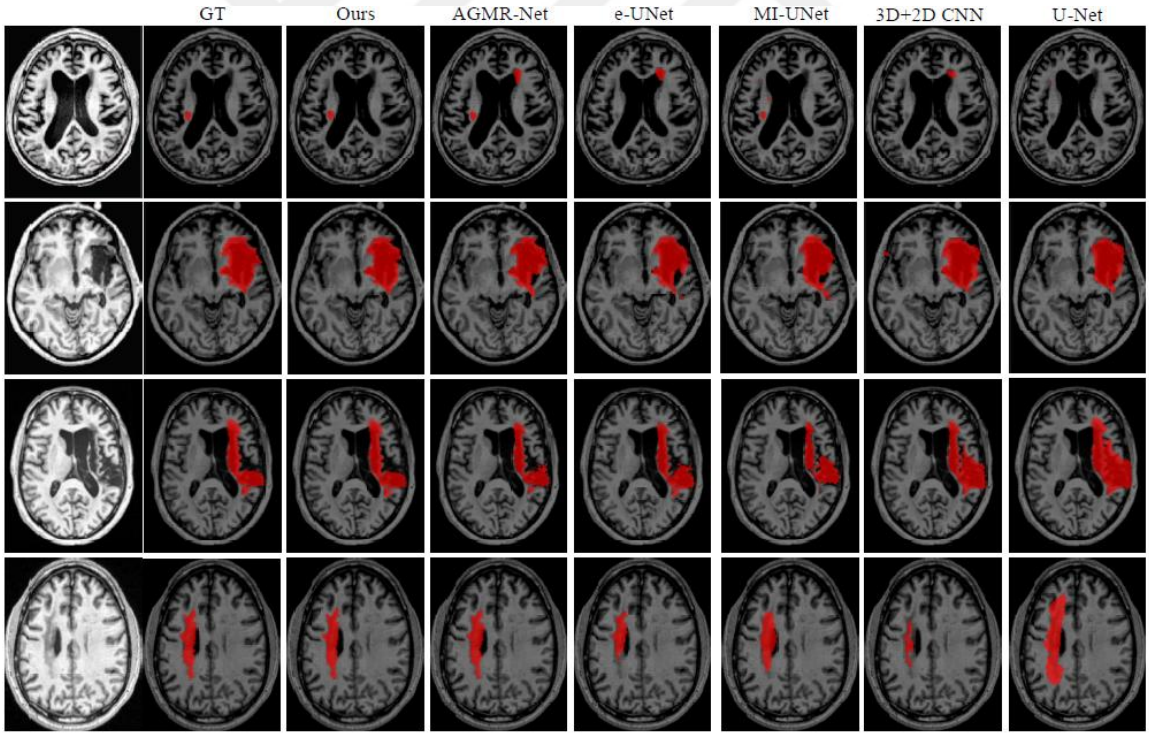
1.2.Otomatik İnme Segmentasyonu

Akut iskemik inmenin değerlendirilmesinde MRI, beyin dokusunda meydana gelen patolojik değişikliklerin ayrıntılı biçimde incelenmesine olanak sağlayan en önemli görüntüleme yöntemlerinden biridir. Özellikle difüzyon ağırlıklı görüntüleme DWI sekansı, iskemik dokuda meydana gelen su difüzyonu değişimlerini erken dönemde tespit edebilmesi sayesinde akut iskemik lezyonların belirlenmesinde kritik bir rol oynamaktadır [19]. Bu görüntülerde ortaya çıkan sinyal değişimleri, iskemik dokunun lokalizasyonu, yayılımı ve şiddeti hakkında önemli bilgiler sunarak inmenin erken teşhisinde ve klinik karar verme süreçlerinde önemli bir yer tutmaktadır. İnmenin zamanında ve doğru şekilde teşhis edilmesi, uygun tedavi stratejilerinin belirlenmesi ve hasta sonuçlarının iyileştirilmesi açısından büyük önem taşımaktadır [20]. Görüntüleme yöntemleri, inme tipinin belirlenmesi, kanamanın dışlanması, enfarkt alanının tespiti ve tedavi planlaması gibi kritik klinik süreçlerde vazgeçilmez bir rol oynamaktadır.

İnme sonrasında ortaya çıkan beyin hasarının, diğer bir ifadeyle enfarkt bölgesinin doğru şekilde belirlenmesi hem klinik değerlendirme hem de tedavi etkinliğinin izlenmesi açısından son derece önemlidir. Özellikle enfarkt hacmi, akut iskemik inme tedavi çalışmalarında önemli bir sonuç ölçütü olarak kabul edilmekte ve hastaların fonksiyonel sonuçları ile güçlü bir ilişki göstermektedir [19]. Bu nedenle enfarkt bölgesinin konumunun ve hacminin doğru bir şekilde belirlenmesi, klinik karar destek süreçleri açısından kritik bir parametre olarak değerlendirilmektedir.

Beyin MR görüntülerinde yer alan inme lezyonlarının belirlenmesi geleneksel olarak uzman nöroloji hekimleri ve radyologlar tarafından manuel olarak gerçekleştirilmektedir. Ancak bu süreç oldukça zaman alıcı ve büyük ölçüde uzman deneyimine bağlıdır. Lezyonların boyutu, şekli ve konumu hastadan hastaya önemli ölçüde değişebildiğinden, lezyon sınırlarının manuel olarak belirlenmesi öznel değerlendirmelere açık olmakta ve gözlemciler arası farklılıkların ortaya çıkmasına neden olabilmektedir [21]. Ayrıca sağlık kuruluşlarında hasta sayısının fazla olması ve uzmanların yoğun çalışma programları gibi faktörler, manuel segmentasyon işleminin klinik uygulamalarda zor ve sürdürülebilir olmayan bir süreç haline gelmesine neden olmaktadır [20]. Enfarkt bölgesinin hacimsel olarak manuel değerlendirilmesi hem zahmetli hem de zaman açısından maliyetli olduğundan, bu yöntemin rutin klinik pratikte uygulanması oldukça güçtür [22].

Bu sınırlılıklar, inme lezyonlarının otomatik olarak tespit edilmesi ve segmentasyonunun yapılmasına yönelik bilgisayar destekli yöntemlerin geliştirilmesini gerekli hale getirmiştir. Otomatik segmentasyon yöntemleri, beyin MR görüntülerinde iskemik lezyonların piksel veya voxel düzeyinde belirlenmesini sağlayarak hem klinik değerlendirme süreçlerini hızlandırmakta hem de lezyon hacmi ve konumu gibi nicel ölçümlerin daha tutarlı ve tekrarlanabilir şekilde elde edilmesine olanak tanımaktadır [23]. Bu tür yöntemler, makine öğrenmesi ve özellikle derin öğrenme alanındaki gelişmelerden yararlanarak lezyon segmentasyonu sürecinin karmaşıklığını azaltmakta ve manuel yöntemlere kıyasla daha hızlı ve nesnel sonuçlar elde edilmesini mümkün kılmaktadır.



Şekil 1.3. İnme lezyonunun manuel ve otomatik segmentasyonunun karşılaştırılması [24].

Otomatik inme segmentasyonu(Şekil 1.3), temel olarak beyin görüntülerinde bulunan anormal dokuların normal beyin dokularından (beyaz madde, gri madde ve beyin omurilik sıvısı gibi) ayrılması sürecini ifade etmektedir [25]. Bununla birlikte bu problem, kullanılan görüntüleme modalitesine, inme tipine, lezyonun boyutuna, konumuna ve morfolojik yapısına bağlı olarak çeşitli zorluklar barındırmaktadır [25, 26]. Lezyonların heterojen

yapısı, küçük lezyonların tespitindeki güçlükler, görüntü artefaktları ve veri setleri arasındaki farklılıklar otomatik segmentasyon yöntemlerinin performansını etkileyen önemli faktörler arasında yer almaktadır. Bu nedenle otomatik inme segmentasyonu problemi, tıbbi görüntü analizi alanında en zorlu ve aynı zamanda en aktif araştırma konularından biri haline gelmiştir. Son yıllarda özellikle derin öğrenme tabanlı yöntemler kullanılarak geliştirilen segmentasyon modelleri, beyin MR görüntülerinde lezyon tespiti ve segmentasyonu görevlerinde önemli başarılar göstermiştir. Bu gelişmeler, klinik karar destek sistemlerinin geliştirilmesi ve inme tanı süreçlerinin daha hızlı ve güvenilir hale getirilmesi açısından büyük bir potansiyel sunmaktadır.

1.3.Tıbbi Görüntülerde Derin Öğrenme Yaklaşımları

Tıbbi görüntülerin otomatik analizi uzun yıllar boyunca eşikleme, kenar belirleme, bölge büyütme ve morfolojik işlemler gibi geleneksel görüntü işleme tekniklerine dayanmaktadır. Bu yöntemler çoğunlukla görüntü yoğunluğu, doku özellikleri ve geometrik yapılar gibi araştırmacı tarafından belirlenen özelliklerin çıkarılmasına dayanmaktadır. Özellikle tıbbi görüntü segmentasyonunda yoğunluk tabanlı eşikleme, aktif kontur modelleri yaygın olarak kullanılmaktadır. Ancak bu yöntemlerin performansı çoğu zaman görüntü kalitesi, gürültü seviyesi ve hasta bazlı anatomik farklılıklardan önemli ölçüde etkilenmektedir. Bunun yanı sıra, manuel olarak belirlenen özelliklerin veri setine özgü olması, farklı görüntüleme protokollerine veya farklı hasta popülasyonlarına genellenebilirliği sınırlayabilmektedir [27].

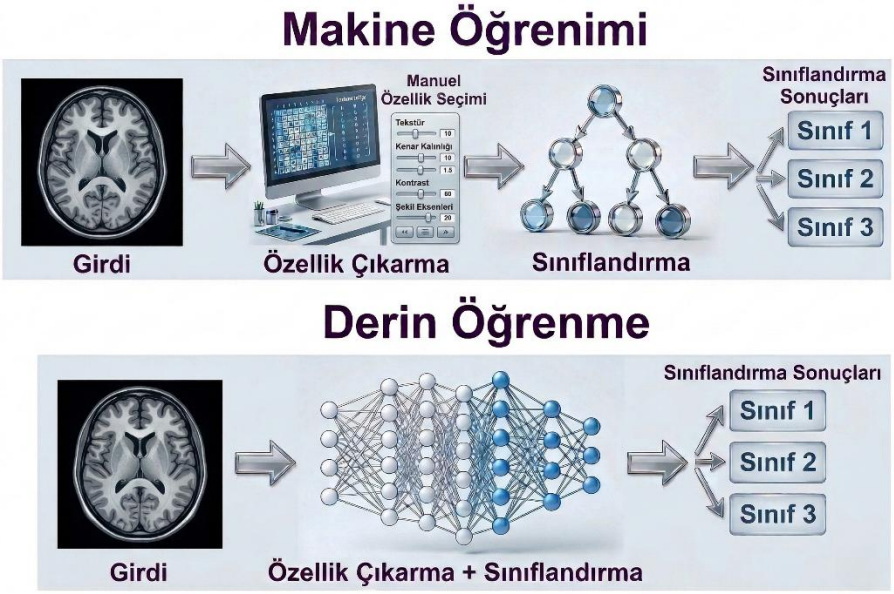
Makine öğrenmesi yöntemlerinin gelişmesiyle birlikte destek vektör makineleri, karar ağaçları ve rastgele ormanlar gibi algoritmalar tıbbi görüntü analizinde kullanılmaya başlanmıştır. Bu yaklaşımlar geleneksel görüntü işleme yöntemlerine kıyasla daha esnek modeller sunmuş olsa da yine de manuel özellik çıkarımına bağımlı olmaları nedeniyle belirli sınırlılıklar taşımaktadır [28]. Özellikle yüksek boyutlu görüntü verilerinde uygun özelliklerin belirlenmesi hem zaman alıcı hem de uzmanlık gerektiren bir süreçtir. Bu durum, daha güçlü ve veri odaklı yöntemlerin geliştirilmesi gerekliliğini ortaya çıkarmıştır.

Son yıllarda hesaplama gücündeki artış ve büyük veri kümelerinin erişilebilir hale gelmesiyle birlikte derin öğrenme yöntemleri, tıbbi görüntü analizi alanında önemli bir dönüşüm sağlamıştır. Derin öğrenme modelleri, çok katmanlı yapıları sayesinde görüntülerden otomatik olarak hiyerarşik özellikler öğrenebilmekte ve manuel özellik çıkarımı gereksinimini büyük ölçüde ortadan kaldırmaktadır. Özellikle konvolüsyonel sinir ağları, görüntü verilerindeki uzamsal ilişkileri öğrenebilme yetenekleri sayesinde sınıflandırma, tespit ve segmentasyon gibi birçok görevde yüksek performans göstermektedir [29].

Derin öğrenme, tıbbi görüntü analizinde otomatik analiz süreçlerinin geliştirilmesine olanak sağlayarak birçok klinik görevde insan uzman performansına yakın hatta bazı durumlarda daha yüksek doğruluk seviyelerine ulaşılmasını mümkün kılmıştır. Günümüzde derin öğrenme tabanlı yöntemler, görüntü segmentasyonu, hastalık tespiti, tanı destek sistemleri ve görüntü rekonstrüksiyonu gibi birçok alanda yaygın olarak kullanılmaktadır. Özellikle tıbbi görüntü segmentasyonu alanında geliştirilen U-Net mimarisi ve türevleri, bu alandaki en yaygın kullanılan mimarilerden biri haline gelmiştir [30, 31].

Ronneberger ve arkadaşları tarafından biyomedikal görüntü segmentasyonu amacıyla önerilen U-Net mimarisi, encoder–decoder yapısı ve bu iki yapı arasındaki geçiş bağlantıları

sayesinde hem yüksek seviyeli semantik bilgiyi hem de düşük seviyeli uzamsal detayları bir araya getirebilmektedir. Bu mimari, özellikle sınırlı veri setleriyle çalışılan biyomedikal uygulamalarda dahi piksel düzeyinde yüksek doğrulukla segmentasyon yapılabilmesini sağlamaktadır [30, 32-34]. U-Net mimarisinin başarısı, zaman içerisinde Attention U-Net, U-Net++ ve Residual U-Net gibi birçok varyantın geliştirilmesine olanak sağlamıştır.



Şekil 1.4. Makine öğrenmesi ve derin öğrenme yaklaşımlarının karşılaştırılması.

Derin öğrenme yöntemlerinin başarısı yalnızca belirli bir görüntüleme tekniği ile sınırlı kalmamış, bilgisayarlı tomografi, manyetik rezonans görüntüleme, ultrason, X-ray gibi farklı tıbbi görüntüleme modalitelerinde de geniş uygulama alanı bulmuştur. Bu durum, derin öğrenme yaklaşımlarının farklı görüntüleme yöntemlerine uyum sağlayabilen esnek ve güçlü modeller olduğunu göstermektedir [35, 36]. Günümüzde bu modeller yalnızca segmentasyon görevleriyle sınırlı kalmayıp, çoklu görev öğrenme yaklaşımları ile tanı, klinik seyir tahmini ve klinik karar destek sistemleri gibi daha kapsamlı uygulamalarda da kullanılmaktadır. Makine öğrenmesi ve derin öğrenme yaklaşımlarında özellik çıkarımı ve sınıflandırma süreçlerinin nasıl gerçekleştirildiği arasındaki temel farklılıklar Şekil 1.4'te gösterilmektedir; makine öğrenmesinde özellikler genellikle manuel olarak belirlenirken, derin öğrenmede bu süreç model tarafından otomatik olarak gerçekleştirilmektedir.

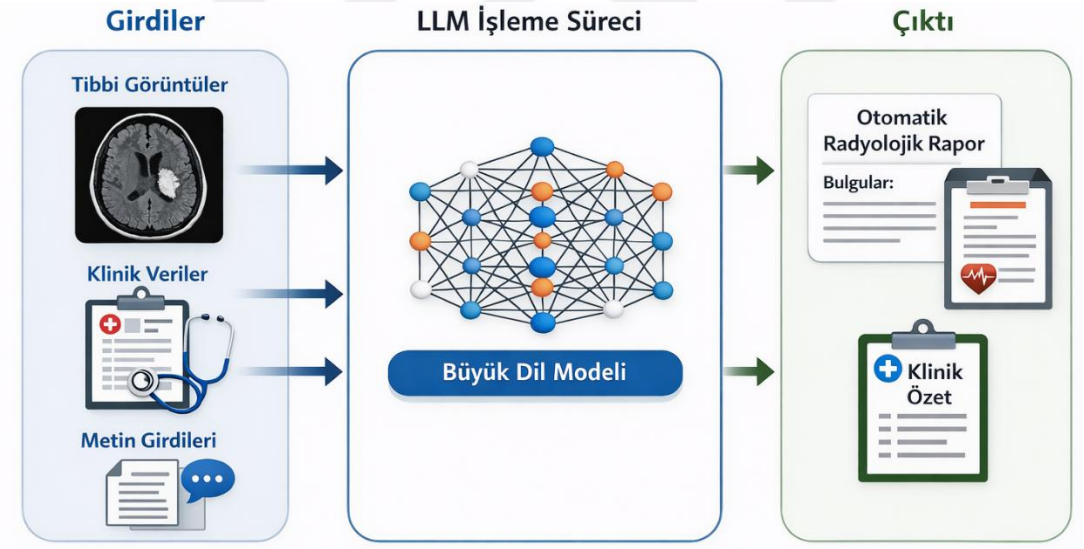
1.4.Tıbbi Metin Üretimi ve Büyük Dil Modelleri

Doğal dil işleme (Natural Language Processing, NLP) alanındaki gelişmeler, tıbbi metinlerin otomatik olarak analiz edilmesi ve üretilmesi konusunda önemli ilerlemeler sağlamıştır. Elektronik sağlık kayıtları, klinik notlar ve radyoloji raporları gibi büyük hacimli tıbbi metin verilerinin hızla artması, bu verilerin otomatik olarak işlenmesini ve anlamlandırılmasını gerekli hale getirmiştir. Geleneksel doğal dil işleme yaklaşımları çoğunlukla kural tabanlı sistemlere veya istatistiksel modellere dayanmakta olup, karmaşık dil yapılarının ve bağlamsal ilişkilerin modellenmesinde sınırlı kalabilmektedir [37, 38].

Derin öğrenme yöntemlerinin gelişmesiyle birlikte doğal dil işleme alanında önemli ilerlemeler kaydedilmiştir. Özellikle tekrarlayan sinir ağları (Recurrent Neural Networks, RNN) ve uzun kısa süreli bellek ağları (Long Short-Term Memory, LSTM) gibi mimariler

metin verilerinin sıralı yapısını modelleyerek metin sınıflandırma, duygu analizi ve metin üretimi gibi görevlerde daha başarılı sonuçlar elde edilmesini sağlamıştır [37]. Bununla birlikte, bu modeller uzun metinlerde bağlamsal ilişkilerin yakalanmasında ve büyük veri kümeleri üzerinde ölçeklenebilirlik açısından bazı sınırlılıklar göstermiştir. Bu sınırlılıkların aşılmasında Transformer mimarisi, doğal dil işleme alanında önemli bir dönüm noktası olmuştur. Vaswani ve arkadaşları tarafından önerilen Transformer modeli, dikkat mekanizmalarına dayalı yapısı sayesinde metin içerisindeki uzun menzilli bağımlılıkları etkili bir şekilde modelleyebilmekte ve paralel hesaplama avantajı sayesinde büyük veri kümeleri üzerinde daha verimli şekilde eğitilebilmektedir [39]. Transformer mimarisinin geliştirilmesiyle birlikte Bidirectional Encoder Representations from Transformers (BERT), Generative Pre-trained Transformer (GPT) ve Text-To-Text Transfer Transformer (T5) gibi modeller ortaya çıkmış ve bu modeller büyük dil modelleri (Large Language Models, LLM) olarak adlandırılan yeni bir araştırma alanının oluşmasına katkı sağlamıştır [40, 41].

Büyük dil modelleri, çok büyük metin veri kümeleri üzerinde ön eğitim süreci ile dilin sözdizimsel ve anlamsal yapısını öğrenebilmekte, ardından belirli görevler için ince ayar yapılarak farklı doğal dil işleme problemlerine uyarlanabilmektedir. Bu modeller metin sınıflandırma, soru-cevap sistemleri, metin özetleme ve metin üretimi gibi çeşitli görevlerde başarılı sonuçlar elde etmektedir [42]. Özellikle metin üretimi görevlerinde LLM mimarileri, verilen bağlamdan anlamlı ve tutarlı metinler oluşturabilme yetenekleri sayesinde önemli bir potansiyel sunmaktadır.



Şekil 1.5. Büyük dil modelleri ile tıbbi metin ve rapor üretim süreci.

Büyük dil modelleri ile tıbbi metin üretiminin kesişimi, klinik doğal dil işleme alanındaki en önemli araştırma konularından biri haline gelmiştir. Sağlık sistemleri klinik iş akışlarını giderek daha fazla dijitalleştirdikçe, otomatik, doğru ve güvenilir tıbbi metin üretimine olan ihtiyaç da artmaktadır. LLM'ler, özellikle Transformer tabanlı mimariler sayesinde klinik dokümantasyon, radyoloji raporu oluşturma, taburculuk özetleri ve hastaya yönelik sağlık bilgilendirme metinleri gibi çeşitli alanlarda akıcı ve tutarlı metinler üretebilme yetenekleri göstermiştir [43-45]. Bu gelişmeler, klinik raporlama süreçlerinin otomatikleştirilmesi ve sağlık profesyonellerinin iş yükünün azaltılması açısından önemli fırsatlar sunmaktadır (Şekil 1.5.). Büyük dil modellerinin yüksek riskli klinik ortamlarda kullanılması bazı

önemli zorlukları da beraberinde getirmektedir. Özellikle model çıktılarının doğruluğu, halüsinasyon problemi, hasta verilerinin gizliliği ve model performansının güvenilir şekilde değerlendirilmesi gibi konular halen aktif araştırma alanları arasında yer almaktadır [44, 46]. Bu nedenle klinik uygulamalarda kullanılacak metin üretim sistemlerinin yalnızca dil üretim kabiliyeti açısından değil, aynı zamanda klinik doğruluk, güvenilirlik ve açıklanabilirlik açısından da değerlendirilmesi gerekmektedir.

1.5. İlgili Çalışmalar

Son yıllarda yapay zekâ ve derin öğrenme yöntemlerinin gelişmesiyle birlikte tıbbi görüntü analizi ve klinik metin işleme alanlarında önemli ilerlemeler kaydedilmiştir. Özellikle manyetik rezonans görüntüleri üzerinden hastalık tespiti ve segmentasyonu, klinik metinlerin otomatik olarak analiz edilmesi ve tıbbi görüntülerden otomatik rapor üretimi gibi konular, sağlık bilişimi alanında yoğun şekilde araştırılan başlıklar arasında yer almaktadır. Bu tez çalışmasının kapsamı göz önünde bulundurularak ilgili çalışmalar üç temel başlık altında değerlendirilmiştir. İlk olarak, manyetik rezonans görüntüleri kullanılarak inme lezyonlarının otomatik olarak tespit edilmesini amaçlayan görüntü tabanlı çalışmalar ele alınmıştır. İkinci olarak, klinik metinlerin doğal dil işleme yöntemleri kullanılarak analiz edilmesine odaklanan tıbbi metin analizi çalışmaları incelenmiştir. Son olarak ise, tıbbi görüntülerden elde edilen bilgilerin doğal dilde raporlara dönüştürülmesini hedefleyen tıbbi görüntülerden otomatik rapor üretimi çalışmaları değerlendirilmiştir. Bu sayede, görüntü analizi, metin analizi ve görüntüden metin üretimi yaklaşımlarını ayrı ayrı inceleyerek bu tez çalışmasının ele aldığı çok modlu yapay zekâ yaklaşımının literatürdeki konumu daha açık şekilde ele alınmıştır.

1.5.1. Görüntü Tabanlı İnme Tespiti Çalışmaları

Tıbbi görüntüler kullanılarak inme lezyonlarının otomatik olarak tespit edilmesi üzerine literatürde çok sayıda çalışma bulunmaktadır. Bu çalışmalar, klasik görüntü işleme yöntemlerinden derin öğrenme ve transformer tabanlı modellere kadar farklı yaklaşımlar önermektedir. Aşağıda görüntü tabanlı inme tespiti alanında literatürde yer alan önemli çalışmalar incelenmiştir.

Chen ve arkadaşları [47], iskemik inme lezyonlarının tespiti için optimize edilmiş You Only Look Once v5 (YOLOv5) tabanlı bir tespit çerçevesi önermiştir. Farklı ölçeklerdeki nesnelerin daha etkin biçimde tanınmasını sağlamak amacıyla, modelin ince ayrıntılara sahip lezyon özelliklerini yakalama yeteneğini geliştiren Multi-Scale Swin Transformer Prediction Head yapısı modele entegre edilmiştir. Ayrıca, Second-Order Channel Attention modülünün eklenmesiyle özellik ayrıştırma kapasitesi artırılmış ve yanlış pozitif sonuçların azaltılması hedeflenmiştir. Önerilen yöntem öncelikle araştırmacılar tarafından toplanan Akut İskemik İnme (AIS) hasta veri seti üzerinde uygulanmış, ardından Ischemic Stroke Lesion Segmentation (ISLES) 2022 veri seti [48] üzerinde doğrulanmıştır. Yapılan deneylerde yöntem %80 ortalama doğruluk değeri elde ederek gerçek dünya iskemik inme tespiti görevlerinde güçlü bir performans sergilemiştir.

Yu ve arkadaşları [49], MRI verileri ile derin öğrenme modellerini kullanarak iskemik inmenin tespiti ve olası tekrarlarının tahmin edilmesini incelemiştir. Çalışmada yapılan literatür incelemesi, özellikle iskemik inme için farklı algoritmaların kullanıldığı performans değerlendirme çalışmalarının bulunduğunu göstermektedir. Tıbbi görüntülerin yorumlanmasının genel olarak "Bilgisayarlı Görü" alanı kapsamında değerlendirildiği ve

sinir ağı tabanlı algoritmaların bu alanda başarılı performans metrikleri sunduğu ifade edilmektedir. Çalışmada ISLES 2015 [50] veri seti kullanılmıştır. Bu veri seti, "inme lezyonları" ve "sağlıklı doku" olmak üzere iki sınıftan oluşan ve açık erişimli bir veri setidir. Deneylerde U-Net, Feature Pyramid Network, DenseNet ve hibrit sinir ağı modelleri gibi farklı mimariler kullanılmıştır. Model performansları Area Under the Curve (AUC) – Receiver Operating Characteristic (ROC) için %75–%80, F1 skoru için %60–%80, kesinlik/duyarlılık (precision/recall) için %70–%85 rapor edilmiştir.

Zhang ve arkadaşları [51], StrokeQD veri setini kullanarak Multi-Attribute Identification and Fusion(MAIF) adlı bir ağ mimarisi önermiştir. Önerilen model, farklı morfolojik özelliklere sahip lezyonların konumunun daha doğru belirlenebilmesi amacıyla deforme olabilir konvolüsyon mekanizmaları ile bölge odaklı havuzlama stratejilerini birlikte kullanmaktadır. Bunun yanı sıra modelde, lezyona özgü anlamsal özellikleri kodlayan özel bir bilgi grafı yapısı entegre edilerek oluşturulan tanısal özetlerdeki belirsizlik ve gereksiz tekrarların azaltılması amaçlanmıştır. Yapılan karşılaştırmalı deneyler, MAIF modelinin özellikle belirsiz ve düzensiz lezyon örüntülerinin tespitinde daha yüksek performans gösterdiğini ortaya koymuştur. Bu sonuçlar, yöntemin iskemik inmenin karmaşık tanısal yorumlanması açısından önemli bir potansiyele sahip olduğunu göstermektedir.

Ahmed ve arkadaşları [52], MRI görüntülerinden inme lezyonlarının segmentasyonu için transformer tabanlı mekanizmalarla güçlendirilmiş bir derin öğrenme çerçevesi önermiştir. Önerilen yöntemde kullanılan deforme olabilir dikkat mekanizmaları, segmentasyon görevlerinde doğruluğu ve modelin sağlamlığını artırmayı amaçlamaktadır. Ayrıca çalışmada, modelin genelleme yeteneğini geliştirmek için gerçek inme lezyonlarının sağlıklı beyin görüntülerine yerleştirilmesine dayanan yenilikçi bir ekolojik veri artırma stratejisi önerilmiştir. Önerilen çerçeve ATLAS veri seti üzerinde %82 Dice katsayısı elde etmiş ve ISLES 2015 veri seti üzerinde de güçlü bir genelleme performansı sergileyerek mevcut yöntemlerden daha iyi sonuçlar üretmiştir. Elde edilen bulgular, önerilen yaklaşımın otomatik inme lezyon analizi süreçlerinde klinik uygulamalara entegre edilebilecek önemli bir potansiyele sahip olduğunu göstermektedir.

Daoudi ve arkadaşları [53], FLAIR, T2 ve DWI görüntülerini içeren çoklu MRI verilerini kullanarak iskemik inme lezyonlarının segmentasyonu için bir yöntem önermiştir. Önerilen yaklaşımda Mask R-CNN mimarisi kullanılarak lezyon tespiti ve segmentasyonu tek bir model içerisinde gerçekleştirilmiştir. Bu bütünleşik yapı, yöntemin klinik açıdan daha uygulanabilir olmasını amaçlamaktadır. Geliştirilen model, yapılan deneylerde %85 ortalama doğruluk elde etmiş ve inme lezyonu segmentasyonu görevlerinde yöntemin etkili ve sağlam bir performans sergilediğini göstermiştir.

Xue ve arkadaşları [54], çalışmalarında ATLAS ve Kessler Foundation + Medical College of Wisconsin (KF+MCW) veri setlerinden elde edilen verileri kullanmıştır. Araştırmacılar, beyin MRI görüntülerinde inme lezyonlarının tespitini iyileştirmek amacıyla modaliteye özgü kodlayıcılar, residual bloklar ve ağırlıklı özellik füzyon modülleri içeren tamamen 3 D çoklu yol sinir ağı mimarisi önermiştir. Zor örneklerde doğruluğu artırmak amacıyla ayrıca Amplified Weighting ve Squeeze-and-Excitation mekanizmaları kullanılarak özellik temsilleri güçlendirilmiştir. Önerilen model, ATLAS veri setinde %54 Dice skoru ve KF+MCW veri setinde %47 Dice skoru elde ederek DeepMedic, AnatomyNet ve 3D U-Net gibi güncel yöntemlere kıyasla daha iyi performans göstermiştir. Elde edilen bu rekabetçi sonuçlar, çoklu yol öğrenme stratejilerinin inme tespiti görevlerinde önemli bir katkı sağlayabileceğini ortaya koymaktadır.

Nouman ve arkadaşları [55], beyin MR görüntülerinde lezyon segmentasyonu için yapay sinir ağı ve transformer tabanlı yapıları birleştiren Neuro-TransUNet adlı yeni bir segmentasyon modeli önermiştir. Önerilen mimari, yapay sinir ağlarının küresel bağlamı modellemedeki sınırlılıklarını ve Transformer modellerinin uzamsal doğruluğu korumadaki zorluklarını gidermeyi amaçlamaktadır. U-Net'in uzamsal özellik çıkarım yeteneğini SwinUNETR'in küresel bağlamsal işleme kapasitesiyle birleştiren hibrit bir yapı kullanmaktadır. Böylece model, uzun menzilli bağımlılıkları yakalayabilmenin yanı sıra görüntülerin uzamsal yapısını da koruyabilmektedir. Önerilen yaklaşımın değerlendirilmesi için ATLAS v2.0 veri seti kullanılmıştır; bu veri seti ATLAS v1.2'nin genişletilmiş bir sürümüdür. Elde edilen sonuçlar, Dice ve Hausdorff mesafesi gibi temel performans ölçütlerinde Neuro-TransUNet modelinin önceki modellere kıyasla daha başarılı sonuçlar elde ettiğini göstermiştir.

Tablo 1.1. Görüntü tabanlı inme tespiti çalışmalarına yönelik performans karşılaştırması

Yıl	Yazar	Yöntem	Veri Seti	Performans Ölçütü	Sonuç (%)
2025	Geetha ve ark. [56]	DeepLabv3+	62 hasta (DWI)	Doğruluk	98
2025	Rahman ve ark. [57]	DenseNet121 + SelfONN + CSCA + DSE	ISLES 2022	Dice	87
2024	Nouman ve ark. [55]	Neuro-TransUNet (UNet + SwinUNETR)	ATLAS 2.0 + ATLAS 1.2	Dice	73
2024	Garcia-Salgado ve ark. [58]	Dikkat Modüllü U-Net + Generalized Dice Focal Loss	ISLES 2015 + ISLES 2022	F1 Skoru	70
2023	Zhang ve ark. [51]	MAIF + VFNet	StrokeQD	F1 Skoru	82
2023	Chen ve ark. [47]	MSA-YOLOv5	AIS veri seti (316 hasta) + ISLES 2022	mAP0.5	80
2020	Xue ve ark. [54]	Çok yollu 2.5D CNN (Multi-path 2.5D U-Net + 3D CNN post-processing)	KF + MCW + ATLAS	Dice	62
2020	Daoudi ve ark. [53]	Mask R-CNN	DWI veri seti, 37 hastadan 540 görüntü	mAP0.5	85

Rahman ve arkadaşları [57], çok kanallı MRI modalitelerini kullanarak iskemik inme lezyonlarının segmentasyonu için yeni bir model önermiştir. Önerilen modelde kodlayıcı olarak DenseNet121, çözücü bölümünde ise Self-Organized Operational Neural Networks (SelfONN) yapısı kullanılmıştır. Özellik çıkarımını güçlendirmek amacıyla modele Channel and Space Compound Attention (CSCA) ve Double Squeeze-and-Excitation (DSE) blokları entegre edilmiştir. Ayrıca model performansını artırmak için Dice Loss ve Jaccard Loss fonksiyonlarının ağırlıklı ortalamasına dayanan yeni bir kayıp fonksiyonu önerilmiştir. Modelin eğitimi ve doğrulama süreçleri ISLES 2022 veri seti kullanılarak DWI, ADC ve geliştirilmiş DWI modaliteleri üzerinde gerçekleştirilmiş ve performans değerlendirmesi Dice metriği ile yapılmıştır. Elde edilen sonuçlar, önerilen çok modlu yaklaşımın DenseNet121_UNet, DenseNet121_UNet++ ve DeepLabV3 gibi mevcut

yöntemlere kıyasla daha yüksek Dice değerleri elde ettiğini ve daha başarılı bir performans sergilediğini göstermiştir.

Geetha ve arkadaşları [56], DWI görüntülerinden inme lezyonlarının otomatik segmentasyonu ve elde edilen bulguların değiştirilmiş Rankin ölçeği ile ilişkilendirilmesine yönelik bir yöntem önermiştir. Çalışmada farklı hastanelerden elde edilen 62 hastaya ait DWI görüntüleri kullanılmış ve her hasta için 60 kesit analiz edilmiştir. Görüntüler üzerinde anizotropik difüzyon filtreleme ve maksimum entropi tabanlı eşikleme ile morfolojik işlemler uygulanarak ön işleme gerçekleştirilmiştir. Ayrıca anlamlı iskemik bulgular içeren kesitlerin belirlenmesi için Çoklu fraktal analizi yöntemi kullanılmıştır. Lezyon segmentasyonu için U-Net ve DeepLabv3+ modelleri karşılaştırmalı olarak değerlendirilmiş ve DeepLabv3+ modeli en iyi performansı göstermiştir. Uzman tarafından oluşturulan referans görüntülerle yapılan değerlendirmede DeepLabv3+ modelinin %98,8 doğruluk elde ettiği rapor edilmiştir. Literatürde görüntü tabanlı inme tespitine yönelik farklı yöntemlerin performans karşılaştırmaları Tablo 1.1'de sunulmaktadır.

1.5.2. Tıbbi Metin Analizi Çalışmaları

Klinik verilerin büyük bir bölümünün metin formatında saklanması, bu verilerin otomatik olarak analiz edilmesini ve anlamlandırılmasını önemli bir araştırma alanı haline getirmiştir. Bu nedenle doğal dil işleme yöntemleri kullanılarak tıbbi metinlerin analiz edilmesine yönelik çok sayıda çalışma literatürde yer almaktadır. Tıbbi Metin Sınıflandırma, yapılandırılmamış metinsel verilerden anlamlı ve eyleme dönüştürülebilir bilgiler elde etmeyi amaçlayan sağlık bilişimi alanındaki temel araştırma konularından biridir. Klinik notlar, elektronik sağlık kayıtları ve biyomedikal literatür gibi çeşitli kaynaklardan elde edilen metinler çoğu zaman kavramsal karmaşıklık, sınırlı bağlamsal bilgi ve dengesiz sınıf dağılımları gibi özellikler göstermektedir. Bu durum, söz konusu alandaki sınıflandırma görevlerini teknik açıdan zorlaştırmaktadır.

Wang ve arkadaşları [59], etiketli hedef alan verisine erişim olmadan çok sınıflı metin sınıflandırması gerçekleştirebilen bilgi entegrasyonlu bir zero-shot çerçevesi önermiştir. Önerilen model, metin ve etiket temsillerini zenginleştirmek amacıyla WordNet'ten elde edilen anlamsal bilgileri ve ConceptNet'ten elde edilen yapısal bilgileri birleştirmektedir. Modelde metinlerin kodlanması için Sentence-BERT kullanılmış; ayrıca adversarial veri artırma, bilgi grafı uzayına projeksiyon ve benzerlik tabanlı çıkarım mekanizmaları aracılığıyla etiket tahmini gerçekleştirilmiştir. Medical Abstracts veri seti de dahil olmak üzere altı farklı veri seti üzerinde yapılan değerlendirmelerde, önerilen model %31,65 F1 skoru elde ederek BERT-NLI gibi temel yöntemlere kıyasla daha yüksek performans göstermiştir.

Schopf ve arkadaşları [60], benzerlik tabanlı ve zero-shot yöntemlerini karşılaştırarak denetimsiz metin sınıflandırma yaklaşımlarını sistematik biçimde değerlendirmiştir. Çalışma kapsamında, yeni oluşturulan Medical Abstracts veri seti de dahil olmak üzere dört farklı veri seti üzerinde deneyler gerçekleştirilmiştir. Araştırmacılar, Simple Contrastive Sentence Embedding (SimCSE) ve Sentence-BERT (SBERT) gibi transformer tabanlı gömme (embedding) yöntemlerini kullanan Lbl2TransformerVec adlı bir yaklaşım önermiş ve bu yöntem çoğu veri setinde en yüksek F1 skorlarını elde etmiştir. Medical Abstracts veri setinde Decoding-enhanced BERT with Disentangled Attention (DeBERTa) tabanlı bir zero-shot model %57,28 ile daha yüksek performans göstermiş olsa da genel olarak benzerlik tabanlı yöntemlerin daha başarılı olduğu görülmüştür. Çalışma, etiketli verinin

bulunmadığı durumlarda semantik gömme temsillerinin denetimsiz sınıflandırmada oldukça etkili olduğunu ortaya koymaktadır.

Makhija ve arkadaşları [61], zaman serileri, görüntüler ve klinik metinler gibi heterojen veri modalitelerinden yorumlanabilir tahmin kontrol listeleri öğrenmek amacıyla olasılıksal mantık programlama tabanlı ProbChecklist adlı bir çerçeve önermiştir. Önerilen çerçeve, modaliteye özgü sinir ağı kodlayıcıları aracılığıyla olasılıksal kavram temsilleri öğrenmekte ve bu temsilleri kontrol listesi modelinde kullanılan ikili özelliklere dönüştürmektedir. Yorumlanabilirliği ve adaleti sağlamak amacıyla, seyrekleştirme ve kavram korelasyonunun azaltılması gibi düzenlileştirme tekniklerinin yanı sıra, hassas gruplar arasında hata farklılıklarını en aza indirmeyi amaçlayan adalet kısıtları uygulanmıştır. Medical Information Mart for Intensive Care-III (MIMIC-III) veri seti üzerinde gerçekleştirilen ölüm riski tahmini görevinde ProbChecklist modeli %77,58 doğruluk, %64,2 kesinlik ve %95,3 özgüllük değerlerine ulaşarak temel kontrol listesi yöntemlerinden daha iyi performans göstermiştir. Ayrıca, Medical Abstracts Corpus veri setinde tümör tespiti görevi için %61,6 dengelenmiş kesinlik ve %62,3 duyarlılık değerleri elde edilmiştir.

Wang ve arkadaşları [62], beş hastalık sınıfına ait 14.438 klinik metin içeren Medical Abstracts veri seti üzerinde KMTLabeler çerçevesini uygulamıştır. Başlangıç aşamasında, her sınıf için 50 örnek olmak üzere toplam 250 etiketli veri ile ince ayar yapılan BioLinkBERT tabanlı sınıflandırıcı yalnızca %21,5 F1 skoru elde etmiş ve bu durum sınırlı gözetimli öğrenmenin kısıtlarını ortaya koymuştur. Bu sorunu gidermek amacıyla, bilgi destekli ve uzman geri bildirimine dayalı bir etiketleme stratejisi uygulanmıştır. Bu süreçte gömme görselleştirmeleri, anahtar kelime ağırlıklandırması ve kural tabanlı pseudo-etiketleme yöntemleri birlikte kullanılmıştır. Uzmanlar, anlamsal kümeleri belirleyerek önemli tıbbi terimlere ağırlık atamış; ayrıca tanısal açıdan önemli örnekleri etiketlemek için sezgisel kurallar kullanılmıştır. Bu yarı otomatik süreç sonucunda yaklaşık 2.900 ek etiketli örnek elde edilmiştir. Genişletilmiş veri seti ile model yeniden eğitildiğinde F1 skoru %65,86'ya yükselmiş ve düşük kaynaklı senaryolarda uzman rehberli etiketlemenin önemli bir katkı sağladığını göstermiştir.

Alammary [63], klinik metin sınıflandırması amacıyla Medical Abstracts veri seti üzerinde on farklı Arapça BERT tabanlı modeli sistematik olarak değerlendirmiştir. En iyi performansı gösteren model %60,25 doğruluk ve %60 F1 skoru elde ederken, diğer modellerin F1 performansları %54,99 ile %59,22 arasında değişmiştir. Elde edilen sonuçlar, genel amaçlı Arapça BERT modellerinin alan-özgü klinik dili ele almada sınırlı bir genelleme yeteneğine sahip olduğunu göstermektedir. Sınıf dengesizliği ve karmaşık semantik yapı gibi zorluklar, modellerin genel etkinliğini azaltmıştır. Çalışma, Arapça tıbbi doğal dil işleme görevlerinde alan-özgü ön eğitim ihtiyacının önemini vurgulamaktadır.

Guan ve arkadaşları [64], zayıf denetimli öğrenme bağlamında etiket fonksiyonlarının otomatik olarak üretilmesini sağlamak amacıyla büyük dil modellerinden yararlanan DataSculpt adlı etkileşimli bir çerçeve geliştirmiştir. Önerilen çerçeve özellikle tıbbi metin sınıflandırma görevlerini içeren MedAbs veri seti üzerinde değerlendirilmiştir. GPT-3.5 ve GPT-4 gibi modeller tarafından üretilen anahtar kelime tabanlı etiket fonksiyonlarının kullanıldığı deneysel sonuçlar, en iyi yapılandırmanın %39,72 F1 skoru elde ettiğini göstermiştir.

Huang ve arkadaşları [65], Kolmogorov karmaşıklığına dayanan bilgi uzaklığını yaklaşık olarak hesaplayabilen ve insan benzeri few-shot öğrenmeyi mümkün kılan GPT tabanlı bir

sıkıştırma yaklaşımı (GPT-AC) önermiştir. Medical Abstracts veri seti üzerinde gerçekleştirilen one-shot sınıflandırma deneylerinde önerilen yöntem, %27,9 doğruluk elde ederek SBERT ve diğer gömme tabanlı yaklaşımlara kıyasla daha yüksek performans göstermiştir. Elde edilen sonuçlar, yöntemin etiketli veriye ihtiyaç duymadan anlamsal benzerliği etkili biçimde yakalayabildiğini ortaya koymaktadır.

Sünnetci [66] tarafından gerçekleştirilen çalışmada, biyomedikal metin verilerinden kolon, akciğer ve tiroid kanserlerinin otomatik sınıflandırılması incelenmiştir. Bu amaçla, ChatGPT kullanılarak anlamsal açıdan çeşitli 3.070 biyomedikal metinden oluşan özgün ve dengeli bir veri seti oluşturulmuştur. Modelin açık sınıf göstergelerini ezberlemesini önlemek amacıyla metinlerde yer alan kanser türlerine ait tüm ifadeler veri setinden çıkarılmıştır. Oluşturulan veri seti %70 eğitim, %15 doğrulama ve %15 test olmak üzere üçe ayrılmıştır. Model geliştirme sürecinden önce metin normalizasyonu ve dizi kodlama gibi ön işleme adımları uygulanmıştır. Çalışma kapsamında LSTM, Gated Recurrent Unit (GRU), Bidirectional Long Short-Term Memory (BiLSTM) ve bunların çok katmanlı ve hibrit varyantları olmak üzere toplam yedi derin öğrenme mimarisi eğitilmiş ve karşılaştırmalı olarak değerlendirilmiştir. Bu modeller arasında tek katmanlı LSTM modeli en yüksek performansı göstermiş; doğrulama veri setinde %91,32 doğruluk ve %91,32 F1 skoru, test veri setinde ise %85,87 doğruluk ve %85,90 F1 skoru elde edilmiştir. Ayrıca, en iyi performansı gösteren model entegre edilerek kullanıcı dostu bir grafiksel kullanıcı arayüzü (GUI) geliştirilmiş ve sistemin pratik ve verimli bir şekilde kullanılabilmesi sağlanmıştır. Literatürde tıbbi metin analizine yönelik farklı yöntemlerin performans karşılaştırmaları Tablo 1.2'de sunulmaktadır.

Tablo 1.2. Tıbbi metin analizi çalışmalarına yönelik performans karşılaştırması.

Yıl	Referans	Yöntem	Veriseti	Doğruluk (%)	Kesinlik (%)	Duyarlılık (%)	F1 Skoru(%)
2025	Sünnetci [66]	LSTM	Biomedical Text Dataset	85,87	86,26	85,88	85,9
		GRU		85	85,4	84,99	84,75
		BiLSTM		83,48	84,24	83,48	83,59
		LSTM + LSTM		84,35	85,61	84,37	84,4
		GRU + GRU		82,39	83,22	82,37	82,34
		BiLSTM + BiLSTM		84,35	84,65	84,34	84,41
2025	Alammary [63]	GigaBERT	Medical Abstracts	60,25			60
2024	Wang ve ark. [62]	Knowledge-assisted labeling					65,86
2024	Makhija ve ark. [61]	ProbChecklist			61,6	62,3	61,9
2024	Schopf ve ark. [60]	Zero-shot Entailment					57,28
2024	Wang ve ark. [59]	BERT-NLI		41,09	37,47		32,52
2023	Guan ve ark. [64]	SC-KATE					39,72
2023	Huang ve ark. [65]	GPT-AC		27,9			

1.5.3. Tıbbi Görüntülerden Otomatik Rapor Üretimi Çalışmaları

Son yıllarda yapay zekâ ve derin öğrenme alanındaki gelişmelerle birlikte, tıbbi görüntülerden otomatik rapor üretimine yönelik çalışmalar hız kazanmıştır. Bu çalışmalar, tıbbi görüntülerden elde edilen görsel bilgilerin doğal dil işleme yöntemleri kullanılarak klinik raporlara dönüştürülmesini amaçlamaktadır. Özellikle görüntü analizi ve metin üretimini birleştiren çok modlu modeller, radyoloji raporlarının otomatik olarak oluşturulması için yaygın olarak kullanılmaktadır. Bu bölümde, tıbbi görüntülerden otomatik rapor üretimine yönelik literatürde yer alan önemli çalışmalar incelenmiştir.

Divya ve arkadaşları [67] tarafından gerçekleştirilen çalışmada, radyolojik görüntülerden otomatik rapor üretimini geliştirmek amacıyla Bellek Rehberli Dönüştürücü (Memory-Guided Transformer) ve Uzamsal-Anlamsal Görsel Özellik Çıkarıcı (Spatial-Semantic Visual Feature Extractor) tabanlı bir mimari önerilmiştir. Önerilen yöntemde görüntülerden hem uzamsal özelliklerin hem de çok ölçekli anlamsal özelliklerin daha etkili biçimde çıkarılabilmesi için ResNet-101 tabanlı bir görsel özellik çıkarıcı kullanılmıştır. Bu yapı içerisinde deforme olabilir evrişim katmanları ile görüntüdeki uzamsal değişimler yakalanırken, genişletilmiş evrişim katmanları aracılığıyla çok ölçekli anlamsal bilgiler elde edilmektedir. Elde edilen bu özellikler birleştirilerek dönüştürücü tabanlı rapor üretim modeline aktarılmaktadır. Model, Indiana University Chest X-ray Collection(IU X-Ray) ve MIMIC-CXR veri kümeleri üzerinde değerlendirilmiş ve performans Bilingual Evaluation Understudy (BLEU), Metric for Evaluation of Translation with Explicit ORdering (METEOR) ve Recall-Oriented Understudy for Gisting Evaluation-Longest (ROUGE-L) gibi doğal dil üretimi ölçütleri ile analiz edilmiştir. Deneysel sonuçlara göre önerilen yöntem IU X-Ray veri setinde BLEU-1 için 0,492, BLEU-4 için 0,180 ve ROUGE-L için 0,379, MIMIC-CXR veri setinde ise BLEU-1 için 0,373, BLEU-4 için 0,112 ve ROUGE-L için 0,297 değerlerine ulaşarak literatürdeki birçok yönteme kıyasla daha başarılı sonuçlar elde etmiştir.

Magalhaes Junior ve arkadaşları [68] tarafından gerçekleştirilen çalışmada, akciğer X-ray görüntülerinden otomatik radyoloji raporu üretimi için XRaySwinGen adı verilen çok modlu bir model önerilmiştir. Önerilen modelde görüntü özelliklerinin çıkarılması için Swin Transformer tabanlı bir görüntü kodlayıcı, metin üretimi için ise GPT-2 tabanlı bir dil çözücü kullanılmıştır. Model, görüntü ve metin arasındaki ilişkiyi öğrenebilmek amacıyla çapraz dikkat katmanları ile birbirine bağlanan uçtan uca bir mimari şeklinde tasarlanmıştır. Eğitim sürecinde IU X-Ray, National Institutes of Health (NIH) Chest X-ray ve özel bir veri seti olmak üzere üç farklı veri kümesi kullanılmıştır. Deneysel sonuçlar, önerilen modelin özellikle önerilen veri setinde BLEU-1 için 0,731, BLEU-4 için 0,661, ROUGE-L için 0,748 ve METEOR için 0,741 değerlerine ulaşarak yüksek başarı elde ettiğini göstermektedir. Ayrıca NIH Chest X-ray veri setinde BLEU-1 için 0,396, ROUGE-L için 0,404 ve METEOR için 0,393 değerleri elde edilmiştir. Bu sonuçlar, Swin Transformer ve GPT-2 tabanlı çok kipli mimarinin X-ray görüntülerinden anlamlı ve tutarlı tıbbi raporlar üretebildiğini göstermektedir.

Zeng ve arkadaşları [69] tarafından gerçekleştirilen çalışmada, tıbbi görüntülerden otomatik rapor üretimi için Attention-Enhanced Relational Memory Network (AERMNet) adlı bir derin öğrenme mimarisi önerilmiştir. Önerilen modelde görüntü özellikleri ResNet-101 tabanlı bir görsel özellik çıkarıcı ile elde edilmekte ve rapor üretim sürecinde çift katmanlı LSTM ile ilişkisel bellek modülü birlikte kullanılmaktadır. Bu yapı, görüntü özellikleri ile oluşturulan metin arasındaki bağlamsal ilişkilerin daha iyi öğrenilmesini amaçlamaktadır.

Model; IU X-Ray, MIMIC-CXR, fetal heart veri seti ve ultrason veri seti üzerinde değerlendirilmiştir. Deneysel sonuçlar, önerilen AERMNet modelinin IU X-Ray veri setinde BLEU-1 için 0,486, METEOR için 0,219 ve ROUGE-L için 0,398 değerlerine ulaştığını göstermektedir. MIMIC-CXR veri setinde ise model BLEU-1 için 0,273, METEOR için 0,157 ve ROUGE-L için 0,232 performansı elde etmiştir. Ayrıca FH veri setinde BLEU-1 için 0,890 ve ROUGE-L için 0,866 gibi yüksek değerler elde edilmiştir. Bu sonuçlar, önerilen bellek destekli mimarinin görüntü ve metin arasındaki ilişkileri daha iyi öğrenerek daha doğru ve bağlamsal olarak tutarlı tıbbi raporlar üretebildiğini göstermektedir.

Wang ve arkadaşları [70] tarafından gerçekleştirilen çalışmada, radyolojik görüntülerden otomatik rapor üretimi amacıyla R2GenGPT adlı büyük dil modeli tabanlı bir yöntem önerilmiştir. Bu yaklaşımda görüntü özellikleri Swin Transformer tabanlı bir görsel kodlayıcı ile çıkarılmakta ve elde edilen görsel özellikler Visual Mapper aracılığıyla büyük dil modelinin kelime gömme uzayına aktarılmaktadır. Metin üretimi aşamasında ise Large Language Model Meta AI 2 (LLaMA2)-7B büyük dil modeli kullanılarak göğüs röntgeni görüntülerinden otomatik radyoloji raporları oluşturulmaktadır. Modelde görüntü ve metin arasındaki uyumu sağlamak amacıyla yüzeysel hizalama, delta hizalama ve derin hizalama olmak üzere üç farklı özellik hizalama stratejisi geliştirilmiştir. Yöntem IU-Xray ve MIMIC-CXR veri kümeleri üzerinde değerlendirilmiştir. Deneysel sonuçlar, özellikle R2GenGPT (Deep) modelinin IU-Xray veri setinde BLEU-1 için 0,488, BLEU-4 için 0,173, ROUGE-L için 0,377 ve METEOR için 0,211 değerlerine ulaştığını göstermektedir. MIMIC-CXR veri setinde ise model BLEU-1 için 0,411, BLEU-4 için 0,134, ROUGE-L için 0,297 ve METEOR için 0,160 performansı elde etmiştir. Ayrıca klinik etkinlik değerlendirmesinde modelin F1 skorunun 0,389 olduğu rapor edilmiştir.

Zhang ve arkadaşları [51] tarafından gerçekleştirilen çalışmada, iskemik inme MR görüntülerinden otomatik radyolojik rapor üretimi amacıyla MAIF tabanlı bir yöntem önerilmiştir. Çalışmada ayrıca StrokeQD adlı geniş ölçekli bir veri seti oluşturulmuştur. Bu veri seti 1181 inme hastasına ait 22.220 adet DWI MR görüntüsü ve 1181 klinik rapor içermektedir. Önerilen yöntemde ilk aşamada VarifocalNet ve deforme olabilir evrişim yapıları kullanılarak lezyonlar tespit edilmektedir. Ardından çok ölçekli özellik çıkarımı ile lezyonların konum, şekil ve sinyal olmak üzere üç temel özelliği belirlenmektedir. Son aşamada ise bu özellikler ön bilgi grafı kullanılarak birleştirilmekte ve mantıksal olarak tutarlı radyolojik raporlar üretilmektedir. Deneysel sonuçlar, önerilen yöntemin lezyon özelliklerini yüksek doğrulukla belirleyebildiğini göstermektedir. Rapor üretim performansında ise model BLEU-1 için 0,5590, BLEU-2 için 0,2925, BLEU-4 için 0,0366, METEOR için 0,4487 ve ROUGE için 0,5664 değerlerine ulaşarak otomatik rapor üretiminde başarılı sonuçlar elde etmiştir.

Yang ve arkadaşları [71] tarafından gerçekleştirilen çalışmada, göğüs röntgeni görüntülerinden otomatik radyolojik rapor üretimi amacıyla çok modlu hizalama ve öğrenilebilir bilgi tabanı içeren bir model önerilmiştir. Önerilen yöntemde radyoloji raporlarında bulunan tıbbi bilgilerin modellenebilmesi için metin temsillerinden otomatik olarak bir bilgi tabanı oluşturulmakta ve bu bilgi tabanı görüntü özellikleri ile kullanılmaktadır. Ayrıca modelde görsel-metin ve görsel-etiket hizalama mekanizmaları kullanılarak görüntüler, rapor metinleri ve hastalık etiketleri arasındaki anlamsal ilişki güçlendirilmektedir. Model IU-Xray ve MIMIC-CXR veri setleri üzerinde değerlendirilmiş ve sonuçlar doğal dil üretimi metrikleri ile klinik etkinlik metrikleri kullanılarak ölçülmüştür. Deneysel sonuçlar önerilen modelin mevcut yöntemlere kıyasla daha iyi

performans elde ettiğini göstermektedir. Özellikle IU-Xray veri setinde BLEU-3 değeri 0,230, BLEU-4 değeri 0,174 ve ROUGE-L değeri 0,399 olarak elde edilmiştir. MIMIC-CXR veri setinde ise BLEU-3 değeri 0,157, BLEU-4 değeri 0,111 ve ROUGE-L değeri 0,274 olarak raporlanmıştır. Klinik değerlendirme metriklerinde ise model %79,5 doğruluk, %42 kesinlik, %33,9 duyarlılık ve %35,2 F1 skor elde ederek önceki yöntemlere göre daha iyi sonuçlar üretmiştir. Tablo 1.3'te, tıbbi görüntülerden otomatik rapor üretimi alanında kullanılan farklı derin öğrenme tabanlı yöntemler, kullanılan veri setleri ve elde edilen performans sonuçları karşılaştırmalı olarak sunulmaktadır.

Tablo 1.3. Tıbbi görüntülerden otomatik rapor üretimi çalışmalarına yönelik performans karşılaştırması.

Referans	Yöntem	Veri Seti	Performans Ölçütü	Sonuç
Divya ve ark. [67]	SSVE + Memory Guided Transformer	IU X-Ray	BLEU-1 METEOR	0.492 0.379
		MIMIC-CXR	BLEU-1 METEOR	0.373 0.297
Magalhaes ve ark. [68]	Swin Transformer + GPT-2 çok modlu model	IU X-ray	BLEU-1 METEOR	0.377 0.322
		NIH Chest X-ray	BLEU-1 METEOR	0.396 0.393
Zeng ve ark. [69]	ResNet-101 + Double-LSTM + Relational Memory	IU X-Ray	BLEU-1 METEOR	0.486 0.219
		MIMIC-CXR	BLEU-1 METEOR	0.273 0.157
Wang ve ark. [70]	Swin Transformer + LLaMA2-7B (R2GenGPT)	IU X-Ray	BLEU-1 METEOR	0.488 0.211
		MIMIC-CXR	BLEU-1 METEOR	0.411 0.160
Zhang ve ark. [51]	MAIF + VarifocalNet	StrokeQD	BLEU-1 METEOR	0.5590 0.5664
Yang ve ark. [71]	Çok Modlu Hizalama + Öğrenilebilir Bilgi Tabanı	IU X-Ray	BLEU-4 ROUGE-L	0.174 0.399
		MIMIC-CXR	BLEU-4 ROUGE-L	0.111 0.274

1.6.Araştırmanın Amacı, Kapsamı ve Tezin Organizasyonu

Literatürde yapılan çalışmalar incelendiğinde, tıbbi görüntülerden otomatik rapor üretimi alanında önemli ilerlemeler kaydedildiği görülmektedir. Ancak mevcut çalışmaların büyük bir kısmı göğüs röntgeni gibi iki boyutlu görüntüler üzerinde yoğunlaşmakta olup, özellikle beyin MR görüntülerinden otomatik radyolojik rapor üretimi üzerine yapılan çalışmalar

sınırlı kalmaktadır. Ayrıca birçok çalışma görüntüden yalnızca genel betimlemeler üretmekte, lezyonun anatomik konumu, hacmi ve klinik yorumunu birlikte ele alan bütüncül yaklaşımlar yeterince incelenmemiştir.

Bu tez çalışmasının temel amacı, beyin manyetik rezonans görüntülerinden elde edilen veriler kullanılarak iskemik inme lezyonlarının otomatik olarak tespit edilmesi, anatomik özelliklerinin analiz edilmesi ve bu bulgulara dayalı olarak klinik açıdan anlamlı radyolojik raporların otomatik biçimde oluşturulmasını sağlayan bütünleşik bir yapay zekâ sisteminin geliştirilmesidir. Beyin MR görüntülerinden elde edilen verilerin otomatik olarak analiz edilmesini sağlayan çok aşamalı bir yapay zekâ tabanlı sistem geliştirilmiştir. Çalışma kapsamında öncelikle DWI ve ADC sekanslarından elde edilen görüntüler üzerinde derin öğrenme tabanlı segmentasyon yöntemleri kullanılarak inme lezyonlarının otomatik olarak tespit edilmesi sağlanmıştır. Segmentasyon sürecinde elde edilen sonuçlar kullanılarak lezyonun konumu, hacmi, anatomik bölgesi ve olası vasküler sulama alanı gibi klinik açıdan önemli özellikler çıkarılmıştır. Elde edilen bu özellikler daha sonra beyin atlasları ile ilişkilendirilmiş ve böylece lezyonun anatomik lokalizasyonunun daha ayrıntılı biçimde belirlenmesi mümkün hale getirilmiştir.

Tez çalışmasının bir diğer önemli amacı, görüntü analizinden elde edilen yapısal ve klinik bilgilerin LLM kullanılarak otomatik radyolojik raporlara dönüştürülmesidir. Bu kapsamda geliştirilen sistemde, görüntülerden elde edilen bulgular ile klinik terminoloji arasında anlamsal bir ilişki kurulması sağlanmış ve radyoloji raporlarında kullanılan dil yapısına uygun metinlerin otomatik olarak üretilmesi gerçekleştirilmiştir. Böylece görüntü tabanlı bulguların yalnızca sayısal analizler şeklinde sunulmasının ötesine geçilerek, klinik uygulamalarda doğrudan kullanılabilecek doğal dil tabanlı radyolojik raporlar oluşturulması mümkün hale getirilmiştir. Bu tez çalışması kapsamında geliştirilen yaklaşım ile tıbbi görüntü analizi ve doğal dil üretimi alanlarını bir araya getiren çok disiplinli bir yapay zekâ çerçevesi oluşturulmuştur. Önerilen sistem sayesinde radyologların iş yükünün azaltılması, raporlama süreçlerinin hızlandırılması ve klinik karar destek sistemlerine katkı sağlanması amaçlanmıştır. Ayrıca geliştirilen yöntemin, beyin MR görüntülerinden elde edilen klinik bilgilerin otomatik olarak yapılandırılması ve raporlanması açısından literatürdeki çalışmalara önemli bir katkı sunması hedeflenmiştir.

2. MATERYAL VE YÖNTEM

2.1. Veri Seti ve Etik Süreç

Bu tez kapsamında kullanılan veri seti, Marmara Üniversitesi Tıp Fakültesi Eğitim ve Araştırma Hastanesi Nöroloji ve Acil Servis birimlerinden elde edilen manyetik rezonans görüntülerinden ve bu görüntülere ait radyoloji raporlarından oluşmaktadır. Veri seti, akut iskemik inme tanısı almış hastalar ile sağlıklı bireylere ait DWI ve ADC kesitlerini içermektedir. Kullanılan verilerin toplanması ve araştırma amaçlı kullanımı için Marmara Üniversitesi Tıp Fakültesi İlaç ve Tıbbi Cihaz Dışı Araştırmalar Etik Kurulu'ndan gerekli izinler alınmıştır. İlgili çalışma, 09.2024.1158 protokol numarası ile etik kurul tarafından onaylanmıştır

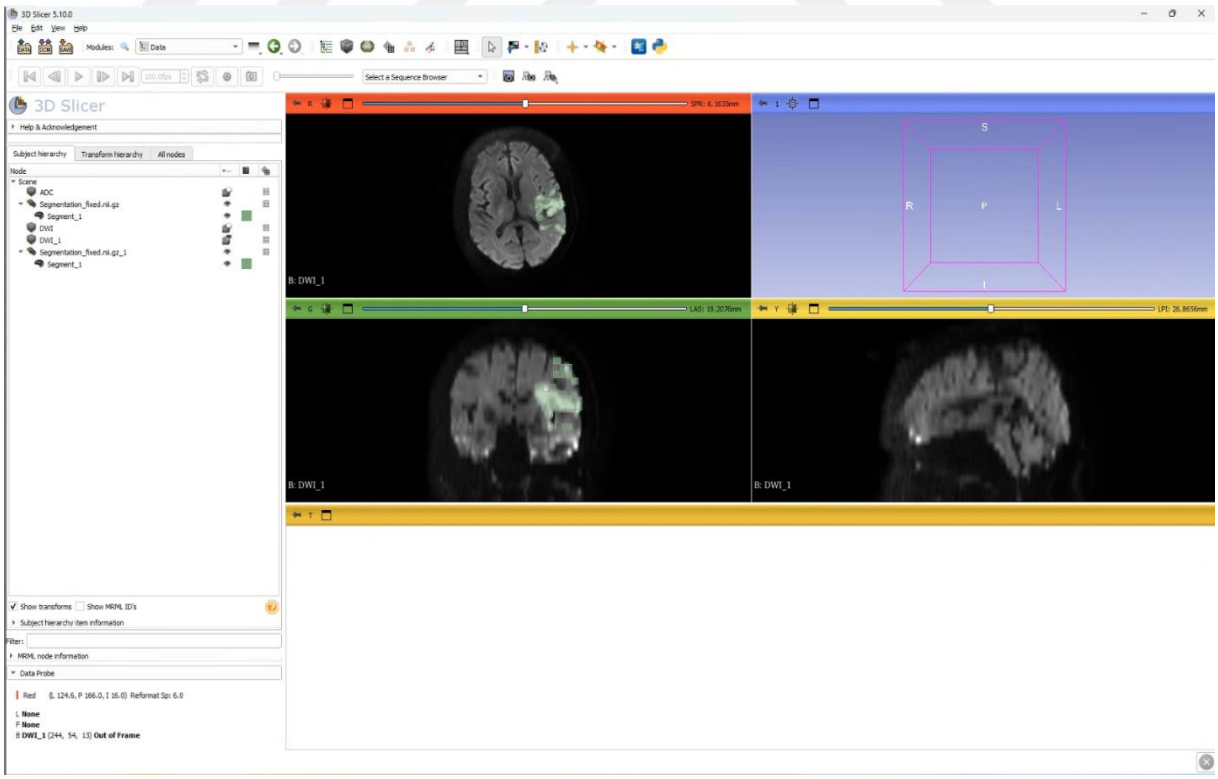
Veriler, etik kurul onayı alındıktan sonra klinik rutin kapsamında elde edilmiş olup, hasta gizliliğini korumak amacıyla tüm veriler anonim hale getirilmiştir. Tez kapsamında 150 inme hastasına ait MR görüntüsü kullanılmıştır. Digital Imaging and Communications in Medicine (DICOM) formatında elde edilen görüntüler, analiz süreçlerinde kullanılmak

üzere uygun veri formatlarına dönüştürülmüş ve her bir görüntüden çoklu kesitler elde edilerek geniş bir veri seti oluşturulmuştur. Ayrıca, bu görüntülere karşılık gelen radyolojik raporlar hastane bilgi sistemlerinden temin edilerek veri setine dahil edilmiştir. Görüntüler ve raporlar, uzman hekimler tarafından kontrol edilerek doğrulanmış ve model geliştirme süreçlerinde kullanılmak üzere hazır hale getirilmiştir. Böylece hem görüntü hem de metin verilerini içeren çok modlu bir veri yapısı oluşturulmuştur.

2.2. Veri Etiketleme ve Doğrulama Süreci

Bu tez çalışması kapsamında kullanılan manyetik rezonans görüntülerinde yer alan iskemik inme lezyonlarının etiketlenmesi, alanında deneyimli uzman hekimler (radyoloji ve nöroloji uzmanları) tarafından manuel olarak yapılmıştır. Etiketleme sürecinde yüksek doğrulukta voxel düzeyinde segmentasyon yapılmasına olanak sağlayan 3D Slicer [72] platformu kullanılmıştır. Segmentasyon için etiketleme işlemi, her bir hastaya ait görüntülerin kesit bazında detaylı olarak incelenmesiyle gerçekleştirilmiştir. Bu sayede lezyon sınırlarının üç boyutlu hacim içerisinde tutarlı ve hassas bir şekilde belirlenmesi sağlanmıştır. Etiketleme sürecinde voxel düzeyinde işaretleme yapılmış, böylece her bir lezyon bölgesi piksel/voxel hassasiyetinde tanımlanmıştır.

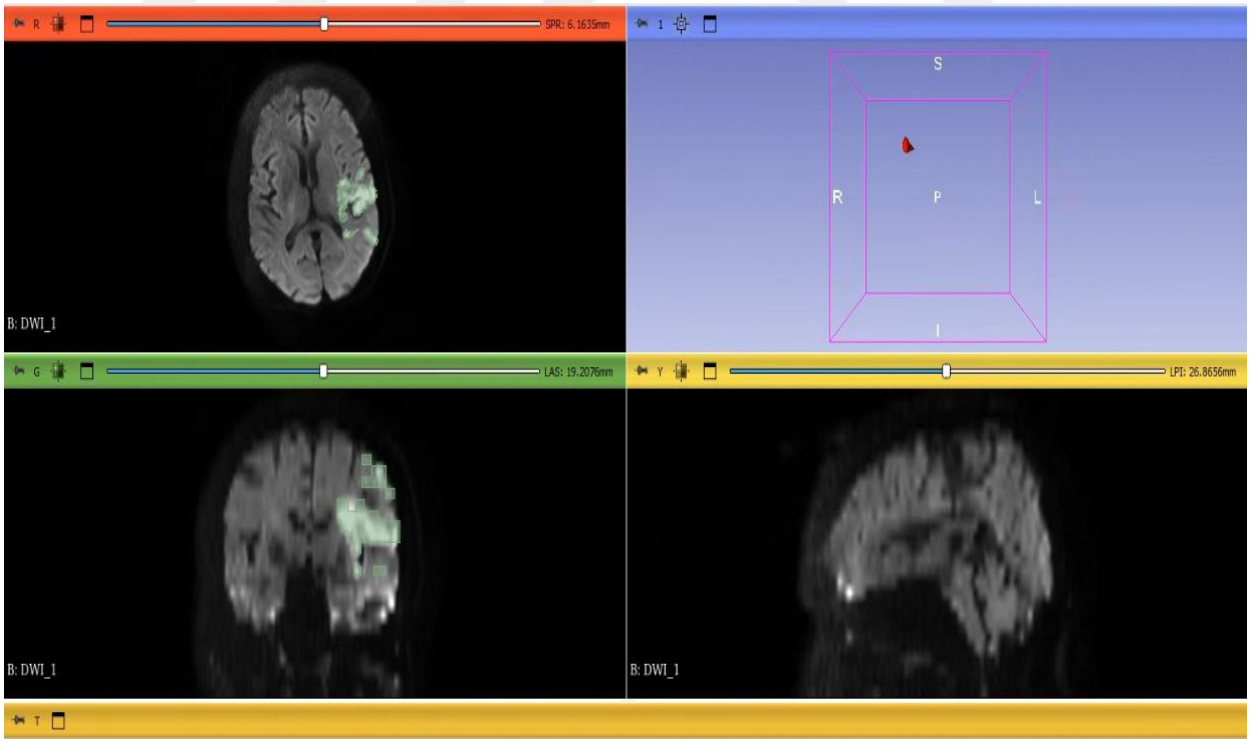
Lezyon segmentasyonu temel olarak DWI görüntüler üzerinden gerçekleştirilmiştir. DWI sekansı, akut iskemik inme lezyonlarını erken dönemde yüksek kontrast ile gösterebilmesi nedeniyle birincil referans olarak kullanılmıştır. Segmentasyon etiketleme süreci, üç boyutlu hacim verisinin farklı düzlemlerde eş zamanlı olarak incelenmesine olanak sağlayan bir ortamda gerçekleştirilmiştir. Bu kapsamda, lezyon bölgeleri voxel düzeyinde manuel olarak etiketlenmiştir. (Şekil 2.1).



Şekil 2.1. 3D Slicer ortamında gerçekleştirilen manuel segmentasyon süreci.

Bu tez kapsamında kullanılan manyetik rezonans görüntülerinin etiketlenmesi süreci, ham görüntü verilerinin hazırlanmasından voxel düzeyinde segmentasyon maskelerinin elde edilmesine kadar çok aşamalı bir işlem hattı şeklinde gerçekleştirilmiştir. İlk aşamada, klinik ortamdan elde edilen görüntüler DICOM formatında temin edilmiş ve analiz süreçlerinde kullanılabilir hale getirilmek üzere uygun veri formatına dönüştürülmüştür. Bu kapsamda, DICOM formatındaki ham görüntü verileri, uzamsal bilgi ve voxel özellikleri korunacak şekilde NIfTI (Neuroimaging Informatics Technology Initiative) formatına dönüştürülmüştür. NIfTI formatına dönüştürülen üç boyutlu görüntü hacimleri, manuel segmentasyon işlemlerinin gerçekleştirilmesi amacıyla 3D Slicer ortamına aktarılmıştır. Segmentasyon süreci bu platform üzerinde gerçekleştirilmiş olup, her bir hasta verisi üç boyutlu hacim olarak yüklenmiştir.

Etiketleme işlemi sırasında, DWI temel referans olarak kullanılmış ve lezyon bölgeleri voxel düzeyinde manuel olarak işaretlenmiştir. Segmentasyon işlemi kesit bazlı (slice-by-slice) ilerletilmiş, ancak her bir kesitte yapılan işaretlemeler komşu kesitler ile değerlendirilerek üç boyutlu hacimsel tutarlılık sağlanmıştır. Böylece yalnızca iki boyutlu segmentasyon değil, aynı zamanda anatomik olarak anlamlı üç boyutlu lezyon maskeleri oluşturulmuştur. Manuel segmentasyon işlemi tamamlandıktan sonra, oluşturulan etiketler 3D Slicer ortamından dışa aktarılmıştır. Segmentasyon çıktıları, her bir hasta için ayrı olacak şekilde NIfTI formatında segmentasyon maskeleri olarak kaydedilmiştir. Bu maskeler, görüntü verileri ile aynı uzamsal boyutlara ve koordinat sistemine sahip olacak şekilde yapılandırılmıştır. Böylece her bir görüntü ile karşılık gelen segmentasyon maskesi arasında birebir eşleşme sağlanmıştır. Elde edilen segmentasyon maskeleri, daha sonraki aşamalarda derin öğrenme tabanlı segmentasyon modellerinin eğitimi için gerçek etiketli (ground truth) veri olarak kullanılmıştır (Şekil 2.2).



Şekil 2.2. Segmentasyon maskesinin üç boyutlu uzaydaki konumunun gösterimi.

Her bir hastaya ait radyolojik rapor verileri, Marmara Üniversitesi Tıp Fakültesi Eğitim ve Araştırma Hastanesi'ne ait hastane bilgi sistemlerinden elde edilmiştir. Görüntü verileri ile eşleştirilen bu raporlar, doğal dil işleme ve otomatik rapor üretimi süreçlerinde kullanılmak üzere yapılandırılmış ve analiz edilebilir hale getirilmiştir. Ham rapor metinleri klinik kullanım amacıyla yazılmış serbest metin formatında olduğundan, doğrudan model eğitiminde kullanılmadan önce bir dizi ön işleme ve standartlaştırma adımından geçirilmiştir.

İlk aşamada, rapor metinleri veri temizleme sürecine tabi tutulmuştur. Bu kapsamda, hasta kimliğini ortaya çıkarabilecek kişisel bilgiler (isim, T.C. kimlik no vb.) veri setinden çıkarılmış ve tüm metinler anonim hale getirilmiştir. Ayrıca yazım hataları, gereksiz tekrarlar, format farklılıkları ve klinik açıdan anlam taşımayan ifadeler temizlenerek daha tutarlı bir metin yapısı elde edilmiştir. Bu süreçte raporların dilsel bütünlüğü korunmuş, ancak model öğrenimini olumsuz etkileyebilecek gürültü bileşenleri ortadan kaldırılmıştır.

İkinci aşamada, rapor içerikleri terminolojik açıdan standardize edilmiştir. Klinik raporlarda aynı bulguların farklı ifadelerle tanımlanabilmesi (örneğin "difüzyon kısıtlılığı", "diff kısıtlılığı", "akut enfarkt bulgusu") modelin öğrenme sürecinde belirsizlik oluşturabileceğinden, bu tür ifadeler mümkün olduğunca ortak bir terminoloji altında birleştirilmiştir. Bu standardizasyon işlemi, modelin semantik olarak daha tutarlı temsiller öğrenmesini sağlamayı amaçlamaktadır.

Elde edilen temizlenmiş ve standardize edilmiş rapor metinleri, görüntü ve segmentasyon maskeleri ile birebir eşleşecek şekilde organize edilmiştir. Her bir hasta için görüntü (DWI/ADC), segmentasyon maskesi ve ilgili radyolojik rapor birlikte ele alınarak çok modlu bir veri yapısı oluşturulmuştur. Bu yapı, görüntü tabanlı bulgular ile metin tabanlı klinik ifadeler arasında doğrudan ilişki kurulmasını mümkün kılmıştır.

Sonuç olarak, radyolojik raporların sistematik bir şekilde temizlenmesi, standardize edilmesi ve görüntü verileri ile hizalanması sayesinde, büyük dil modelleri ile gerçekleştirilecek otomatik rapor üretimi süreci için bir metin veri seti elde edilmiştir. Bu süreç, geliştirilen sistemin yalnızca doğru segmentasyon üretmesini değil, aynı zamanda klinik açıdan anlamlı ve tutarlı radyolojik raporlar oluşturmasını destekleyen temel bileşenlerden biri olmuştur. Bu çalışma kapsamında kullanılan görüntülere karşılık gelen radyolojik raporlara ait örnekler Tablo 2.1'de sunulmuştur.

Tablo 2.1. Görüntülere ait örnek raporlar.

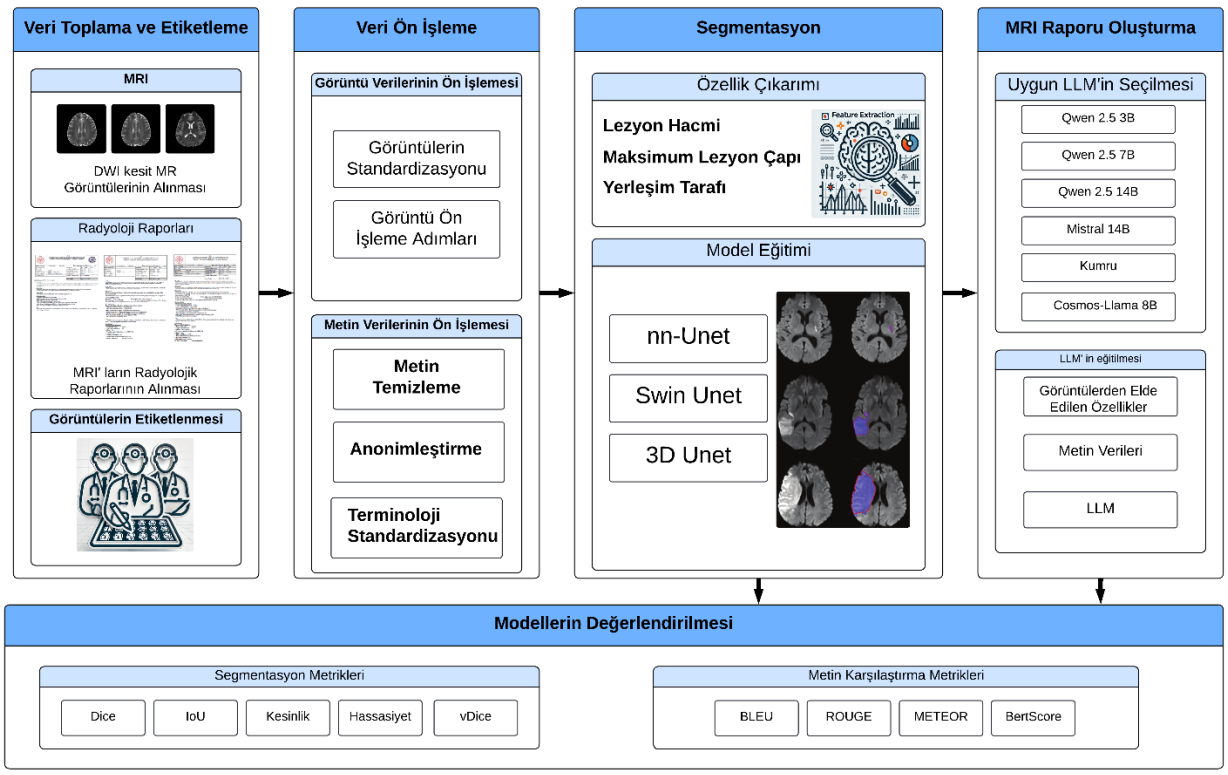
Örnek Raporlar
Bilateral frontal, sol pariyetal ve sol oksipital loblarda dağınık yerleşimli, akut-subakut enfarkt alanları izlenmektedir.
Sağda temporal, frontal ve parietal loblarda akut enfarkt ile uyumlu geniş kısıtlı difüzyon sahası izlenmiştir.
Sağ MCA sulama alanında akut enfarkt ile uyumlu difüzyon kısıtlanması izlenmektedir.
Bilateral periventriküler beyaz cevherde akut iskemiler izlenmektedir.
Sağ frontal lobdan insüler kortekse uzanan akut enfarkt ile uyumlu difüzyon kısıtlanması gösteren alan izlendi.
Sol lateral ventrikül komşuluğundan bazal gangliona uzanarak difüzyonda kısıtlama gösteren alan izlenmiştir (Akut enfarkt-iskemi açısından anlamlı olabilir. Klinik verilerle birlikte değerlendirilmesi önerilir).

2.3. Önerilen Sistem Mimarisi

Bu tez çalışmasında, beyin MR görüntülerinden iskemik inme lezyonlarının otomatik olarak tespit edilmesi, analiz edilmesi ve elde edilen bulguların klinik açıdan anlamlı radyolojik raporlara dönüştürülmesini sağlayan çok aşamalı bütünleşik bir yapay zekâ tabanlı sistem geliştirilmiştir. Önerilen sistem, görüntü işleme, derin öğrenme tabanlı segmentasyon ve doğal dil üretimi bileşenlerini bir araya getiren çok modlu bir işlem hattı yapısı üzerine kurulmuştur.

Sistemin ilk aşamasında, giriş verisi olarak DWI ve ADC haritaları kullanılmaktadır. Bu iki modalite, akut iskemik inme lezyonlarının hem tespiti hem de doğrulanması açısından birbirini tamamlayıcı bilgiler sunmaktadır. DWI görüntüleri lezyon bölgelerini yüksek kontrast ile ortaya koyarken, ADC haritaları bu bölgelerdeki difüzyon kısıtlılığını doğrulamak için kullanılmaktadır. Giriş verileri, model eğitimine uygun hale getirilebilmesi için ön işleme aşamasından geçirilmektedir. Bu aşamada görüntülerin yeniden örneklenmesi , yoğunluk normalizasyonu ve gerekli hizalama işlemleri gerçekleştirilerek veri seti standart bir yapıya dönüştürülmektedir. Böylece farklı hastalardan elde edilen görüntüler arasında oluşabilecek dağılım farklılıklarının azaltılması ve modelin daha kararlı öğrenme gerçekleştirmesi sağlanmaktadır.

Ön işleme aşamasının ardından, segmentasyon modülü devreye girmektedir. Bu aşamada derin öğrenme tabanlı modeller kullanılarak beyin MR görüntüleri üzerinde yer alan iskemik inme lezyonları voxel düzeyinde otomatik olarak tespit edilmektedir. Segmentasyon çıktıları, lezyonun konumu ve sınırlarını belirleyen ikili maskeler şeklinde elde edilmektedir. Bu maskeler, sistemin sonraki aşamalarında kullanılacak temel yapısal bilgiyi oluşturmaktadır.



Şekil 2.3. Önerilen sistemin genel mimarisi.

Segmentasyon sonrasında, lezyon bölgelerine ait çeşitli nicel ve yapısal özellikler çıkarılmaktadır. Bu kapsamda lezyonun hacmi, uzamsal konumu ve anatomik yerleşimi belirlenmektedir. Ayrıca lezyon bölgeleri beyin atlasları ile eşleştirilerek, ilgili anatomik ve vasküler bölgeler (örneğin Middle Cerebral Artery (MCA), Posterior Cerebral Artery (PCA), Anterior Cerebral Artery (ACA) sulama alanları) tespit edilmektedir. Bu özellik çıkarımı aşaması, görüntü tabanlı bulguların klinik açıdan anlamlı hale getirilmesinde önemli bir rol oynamaktadır. Sistemin son aşamasında ise elde edilen tüm yapısal ve klinik bilgiler kullanılarak otomatik radyolojik rapor üretimi gerçekleştirilmektedir. Bu aşamada LLM kullanılarak, segmentasyon ve özellik çıkarımı sonuçları doğal dilde ifade edilmekte ve klinik raporlama formatına uygun metinler oluşturulmaktadır. Böylece sistem yalnızca görüntü analizi yapmakla kalmayıp, aynı zamanda klinik karar destek süreçlerinde doğrudan kullanılabilecek yorumlanabilir çıktılar üretmektedir. Önerilen sistemin genel mimarisi ve işlem adımları Şekil 2.3'te gösterilmektedir. Bu yapı, görüntü tabanlı analiz ile metin tabanlı rapor üretimini bütünleştirerek, uçtan uca çalışan çok modlu bir yapay zekâ çözümü sunmaktadır.

2.4. Çoklu Girdi Temsilleri ve Kombinasyonları

Bu çalışma kapsamında, modele sunulan girdi veri temsillerinin segmentasyon performansı üzerindeki etkisini analiz etmek amacıyla farklı girdi konfigürasyonları tanımlanmış ve sistematik olarak değerlendirilmiştir. Amaç yalnızca çok kanallı giriş kullanımının etkisini incelemek değil, aynı zamanda segmentasyon başarımına anlamlı katkı sağlayan ek bilgi kaynaklarını belirlemektir. Bu doğrultuda, farklı girdi bileşenlerinin tekil ve birleşik etkilerini ortaya koyabilmek için birden fazla deneysel senaryo tasarlanmıştır. Tüm deneyler, aynı veri bölmeleri, aynı ağ mimarileri ve aynı eğitim protokolü kullanılarak gerçekleştirilmiştir. Böylece elde edilen performans farklılıklarının yalnızca kullanılan girdi temsillerinden kaynaklandığı kontrollü bir deneysel çerçeve içerisinde analiz edilmiştir.

İlk deneysel senaryo, "DWI" olarak adlandırılmakta olup, modele yalnızca difüzyon ağırlıklı görüntülerin giriş olarak verildiği tek modaliteli bir yapılandırmayı temsil etmektedir. Bu senaryoda, segmentasyon modeli herhangi bir ek bilgi kaynağı kullanılmadan yalnızca DWI görüntülerinden öğrenme gerçekleştirmektedir. DWI senaryosundan elde edilen sonuçlar, diğer deneysel senaryolarda kullanılan ek veri kaynaklarının (ADC, EDGE ve PRIOR gibi) segmentasyon performansına ne ölçüde katkı sağladığını nicel olarak değerlendirmek için bir referans noktası olarak kullanılmaktadır. Bu sayede, her bir ek bileşenin modele sağladığı katkı açık ve kontrollü bir biçimde analiz edilebilmekte, farklı girdi temsillerinin performans üzerindeki etkileri doğrudan karşılaştırılabilmektedir.

Çoklu modalite senaryosunda(DWI+ADC), DWI görüntülerine ek olarak ADC haritaları ikinci bir giriş kanalı olarak modele dahil edilmiştir. ADC haritaları, difüzyon kısıtlılığına ilişkin nicel bilgiler sunarak DWI görüntülerini tamamlayıcı bir rol üstlenmektedir. Özellikle DWI sinyal özellikleri ile değerlendirildiğinde, lezyon dokusunun normal beyin parankiminden daha doğru bir şekilde ayrıştırılmasına katkı sağlamaktadır. Bu senaryo, tek modaliteli DWI tabanlı temel yapılandırma ile karşılaştırıldığında, çoklu modalite kullanımının segmentasyon performansı üzerindeki etkisini değerlendirmek amacıyla tasarlanmıştır. Bu sayede, DWI ve ADC kombinasyonunun sağladığı tamamlayıcı bilginin modele katkısı kontrollü bir biçimde analiz edilebilmektedir.

Üçüncü deneysel senaryoda(DWI+ADC+EDGE), lezyon sınırlarına ilişkin yapısal bilgiyi açık bir şekilde modele sunmak amacıyla, DWI görüntülerinden türetilen kenar (edge) haritaları ek bir giriş kanalı olarak modele dahil edilmiştir. Bu kapsamda, üç boyutlu DWI hacmi *I(x,y,z)* üzerinde Sobel operatörü uygulanarak üç boyutlu bir kenar temsili elde edilmiştir. Kenar çıkarımı sürecinde, görüntü yoğunluğundaki uzamsal değişimleri yakalayabilmek amacıyla Sobel operatörü [73] üç eksenin (x, y ve z) her biri boyunca bağımsız olarak uygulanmıştır. Bu doğrultuda, hacimsel verinin her bir eksen boyunca türevleri ayrı ayrı hesaplanarak, görüntü içerisindeki kenar ve sınır bilgileri belirgin hale getirilmiştir.

$$I_x = Sobel(I, axis = 0), \qquad I_y = Sobel(I, axis = 1), \qquad I_z = Sobel(I, axis = 2) \quad (1)$$

Denklem 1'de I_x, $I_y$ve I_z, görüntünün üç ortogonal ekseni boyunca yoğunluk değişimlerini temsil eden gradyan bileşenlerini ifade etmektedir. Bu bileşenler, her bir voxel için gradyan büyüklüğünü hesaplamak amacıyla birleştirilmiş ve elde edilen sonuç kullanılarak üç boyutlu kenar haritası aşağıdaki şekilde oluşturulmuştur:

$$E(x, y, z) = \sqrt{I_x(x, y, z)^2 + I_y(x, y, z)^2 + I_z(x, y, z)^2} \qquad (2)$$

Hesaplanan kenar haritaları, farklı hastalar ve kesitler arasında ölçek tutarlılığını sağlamak amacıyla min–maks normalizasyon yöntemi kullanılarak [0,1] aralığına dönüştürülmüştür:

$$E_{norm}(x, y, z) = \frac{E(x, y, z) - \min(E)}{\max(E) - \min(E) + \in} \qquad (3)$$

Denklem 3'te ε, sıfıra bölme hatasını önlemek amacıyla eklenen küçük bir sabiti ifade etmektedir. Normalize edilen kenar hacimleri, DWI ve ADC görüntüleri ile uzamsal olarak hizalanmış ve giriş tensörüne üçüncü bir kanal olarak dahil edilmiştir. Buna göre, model girdisi aşağıdaki şekilde tanımlanmaktadır:

$$X_{EDGE} = [DWI, ADC, EDGE]$$

Bu yaklaşımın temel amacı, modelin yalnızca yoğunluk tabanlı özellikleri değil, aynı zamanda yerel yoğunluk değişimlerinden türetilen sınır odaklı yapısal bilgileri de öğrenmesini sağlamaktır. Akut iskemik inme lezyonlarının düşük kontrastlı yapısı ve belirsiz sınırları göz önünde bulundurulduğunda, kenar temsili lezyon ile sağlıklı doku arasındaki geçiş bölgelerinin daha belirgin hale getirilmesine katkı sağlamaktadır. Ayrıca, kenar bilgisinin doğrudan bir giriş kanalı olarak modele sunulması, modeli sınır odaklı öğrenmeye yönlendiren bir tümevarımsal önyargı kazandırmaktadır. Bu durum, özellikle sınırlı veri koşullarında, daha kararlı ve sınır hassasiyeti yüksek segmentasyon performansının elde edilmesini kolaylaştırmaktadır.

Dördüncü deneysel senaryoda(DWI+ADC+PRIOR), yoğunluk tabanlı özelliklerin ötesinde uzamsal bağlam sağlamak amacıyla atlas tabanlı anatomik ve vasküler ön bilgi (prior) haritaları ek bir giriş kanalı olarak modele dahil edilmiştir. Bu ön bilgi haritaları, referans

uzayda tanımlı standart bir vasküler sulama atlasının her bir hastaya ait DWI hacmine kayıtlanması ile elde edilmiştir.

Burada *A(x,y,z)* atlas hacmini, *I(x,y,z)* ise hastaya özgü DWI görüntüsünü ifade etmektedir. Atlasın hasta verisi ile hizalanabilmesi amacıyla, SimpleITK [74] çerçevesi kullanılarak üç boyutlu düzlem kayıt işlemi gerçekleştirilmiştir. İlk aşamada, görüntü geometrisine dayalı olarak merkezlenmiş bir düzlem dönüşüm tahmin edilerek aşağıdaki şekilde tanımlanmıştır:

$$T_0 = CenteredTransformInitializer(I, A) \quad (4)$$

Bu başlangıç adımını takiben, benzerlik ölçütü olarak Karşılıklı Bilgi (Mutual Information) kullanılarak yoğunluk tabanlı bir optimizasyon süreci gerçekleştirilmiştir:

$$T^* = argmax_T \, MI(I, T(A)) \quad (5)$$

Denklem 5'te T düzlem dönüşümü, MI ise sabit görüntü (DWI) ile dönüştürülmüş atlas arasındaki karşılıklı bilgiyi ifade etmektedir. Optimizasyon süreci, sabit iterasyon sayısı ile gradyan inişi yöntemi kullanılarak gerçekleştirilmiştir. Elde edilen en uygun dönüşüm T^*, doğrusal enterpolasyon kullanılarak atlasa uygulanmış ve böylece hastaya özgü ön bilgi haritası elde edilmiştir.

$$A'(x, y, z) = T^*(A(x, y, z)) \quad (6)$$

Elde edilen atlas hizalanmış ön bilgi hacmi, daha sonra min–maks normalizasyon yöntemi kullanılarak [0,1] aralığına dönüştürülmüştür:

$$A_{norm}(x, y, z) = \frac{A'(x, y, z) - \min(A')}{\max(A') - \min(A') + \in} \quad (7)$$

Denklem 7'de ε, sıfıra bölme hatasını önlemek amacıyla eklenen küçük bir sabiti ifade etmektedir. Normalize edilen ön bilgi haritaları, karşılık gelen DWI ve ADC görüntülerinin uzamsal çözünürlüğü, voxel aralığı ve kesit sıralaması ile uyumlu olacak şekilde yeniden örneklenmiştir. Son olarak, bu ön bilgi kanalı ek bir modalite olarak giriş tensörüne dahil edilmiş ve aşağıdaki çok kanallı giriş temsili elde edilmiştir:

$$X_{PRIOR} = [DWI, ADC, PRIOR]$$

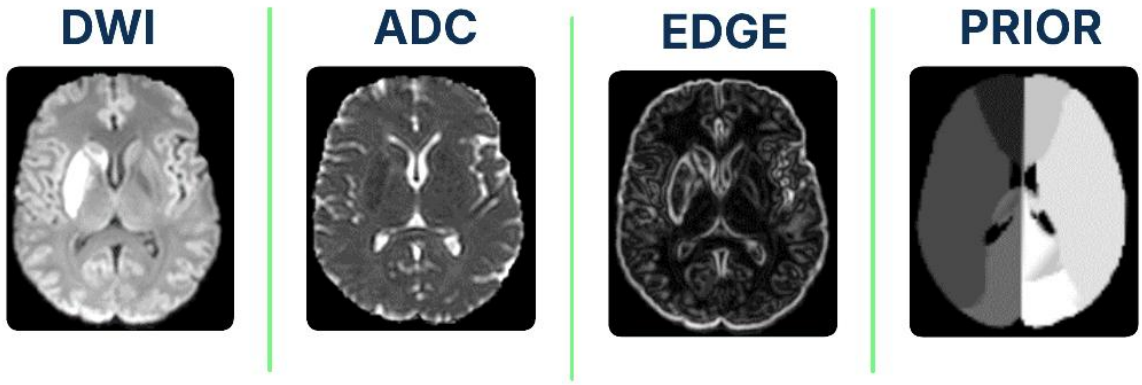
Burada kullanılan ön bilgi haritaları, voxel düzeyinde hassas bir anatomik eşleşme sağlamak amacıyla tasarlanmamıştır. Bunun yerine, lezyonların genel anatomik ve vasküler dağılımını kodlayarak kaba düzeyde uzamsal rehberlik sunmayı hedeflemektedir. Bu kapsamda, 1298 akut iskemik inme hastasına ait lezyon dağılımlarından türetilmiş bir arteriyel sulama atlası kullanılarak, modelin lezyonların daha olası olduğu bölgeler hakkında ön bilgi edinmesi sağlanmıştır [75]. Bu yaklaşımda, ön bilgi haritaları katı anatomik kısıtlamalar dayatmak yerine, segmentasyon sürecinde modeli anatomik ve vasküler açıdan daha olası bölgelere yönlendiren yumuşak bir uzamsal önyargı olarak işlev görmektedir. Böylece model, yalnızca yoğunluk tabanlı ipuçlarına bağlı kalmaksızın, daha anlamlı ve klinik olarak tutarlı segmentasyon çıktıları üretebilmektedir.

Son deneysel senaryo, Entegre Çok Kanallı (Integrated Multi-Channel, IMC) senaryosu olarak adlandırılmakta olup, yoğunluk tabanlı görüntüleme modalitelerinin (DWI ve ADC), yapısal sınır temsillerinin (kenar haritaları-EDGE) ve atlas tabanlı uzamsal ön bilgi haritalarının birlikte kullanıldığı tam yapılandırmayı temsil etmektedir. Bu senaryoda, farklı bilgi kaynakları çok kanallı bir giriş yapısı içerisinde modele eş zamanlı olarak sunulmaktadır.

Bu kapsamda, yoğunluk bilgisi, sınır bilgisi ve uzamsal bağlamın bir araya getirilmesiyle modelin daha zengin ve tamamlayıcı özellikler öğrenmesi hedeflenmektedir. Böylece hem lezyonun sınırlarının daha doğru belirlenmesi hem de anatomik olarak anlamlı bölgelerde segmentasyon yapılması amaçlanmaktadır. Buna göre, model girdisi aşağıdaki şekilde tanımlanmaktadır:

$$X_{IMC} = [DWI, ADC, EDGE, PRIOR]$$

Bu senaryo, yoğunluk, yapısal ve anlamsal/uzamsal ipuçları gibi heterojen bilgi kaynaklarının birlikte kullanımının modelin temsil öğrenme kapasitesi üzerindeki etkisini sistematik olarak değerlendirmek amacıyla tasarlanmıştır. Bu kapsamda, kenar tabanlı sınır bilgisi ile ön bilgi tabanlı uzamsal rehberliğin, düşük kontrastlı ve sınırları belirsiz lezyonların tespit edilmesi ve doğru şekilde sınırlandırılmasına sağladığı tamamlayıcı katkılar incelenmektedir. Ayrıca bu yapılandırma, çoklu modalite entegrasyonunun segmentasyon doğruluğunu artırıp artırmadığını, lezyon sınırlarının daha hassas biçimde belirlenmesini sağlayıp sağlamadığını ve modelin genelleme performansına olan etkisini değerlendirmek için temel deneysel senaryo olarak kullanılmaktadır. Şekil 2.4'te, çalışmada kullanılan DWI, ADC, EDGE ve PRIOR giriş modalitelerine ait örnek MRI görüntüleri sunulmakta olup, her bir modalitenin farklı uzamsal ve anatomik bilgileri modele sağladığı görülmektedir.



Şekil 2.4. Modalitelere ait örnek MRI görüntüleri.

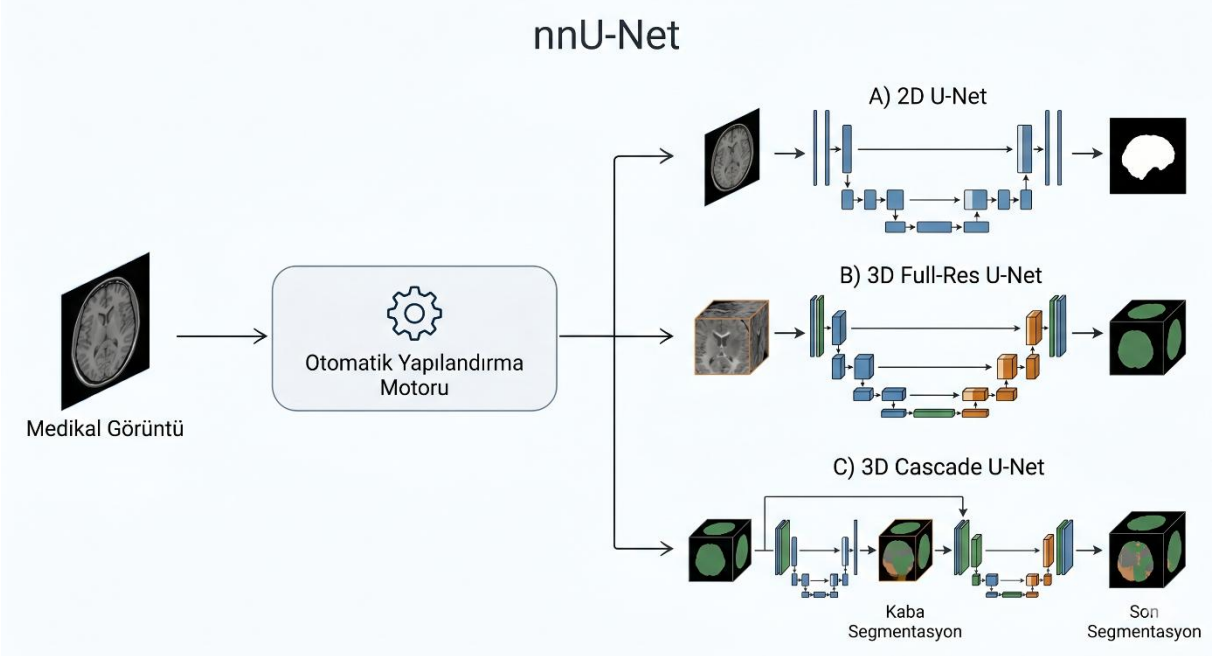
2.5. Segmentasyon Modelleri

2.5.1. nnU-Net

nnU-Net, otomatik yapılandırma yeteneklerine sahip bir derin öğrenme çerçevesi olup, tıbbi görüntü segmentasyonu alanında önemli ilerlemeler sağlamıştır [76]. U-Net mimarisine dayalı eğitim sürecini otomatikleştirerek, manuel hiperparametre ayarlamasına ihtiyaç duymadan yüksek performanslı segmentasyon modellerinin geliştirilmesine olanak

tanımaktadır(Şekil 2.5). Farklı tıbbi görüntüleme modaliteleri ve anatomik yapılarda tutarlı ve güçlü performans sergilemesi sayesinde, nnU-Net günümüzde tıbbi görüntü segmentasyonu çalışmalarında standart bir referans çerçeve haline gelmiştir [77].

nnU-Net, U-Net tabanlı encoder–decoder mimarisi üzerine inşa edilmiştir. Bu yapı, hiyerarşik özellik çıkarımını etkin bir şekilde gerçekleştirirken, geçiş bağlantıları sayesinde ince uzamsal detayların korunmasını sağlayarak voxel düzeyinde hassas segmentasyon yapılmasına olanak tanımaktadır [78]. Ayrıca, çok ölçekli özellik işleme mekanizması sayesinde farklı çözünürlük seviyelerinden elde edilen bilgiler birleştirilerek, modelin farklı boyutlardaki lezyonları daha etkili bir şekilde öğrenmesi ve segmentasyon performansının artırılması sağlanmaktadır [77].

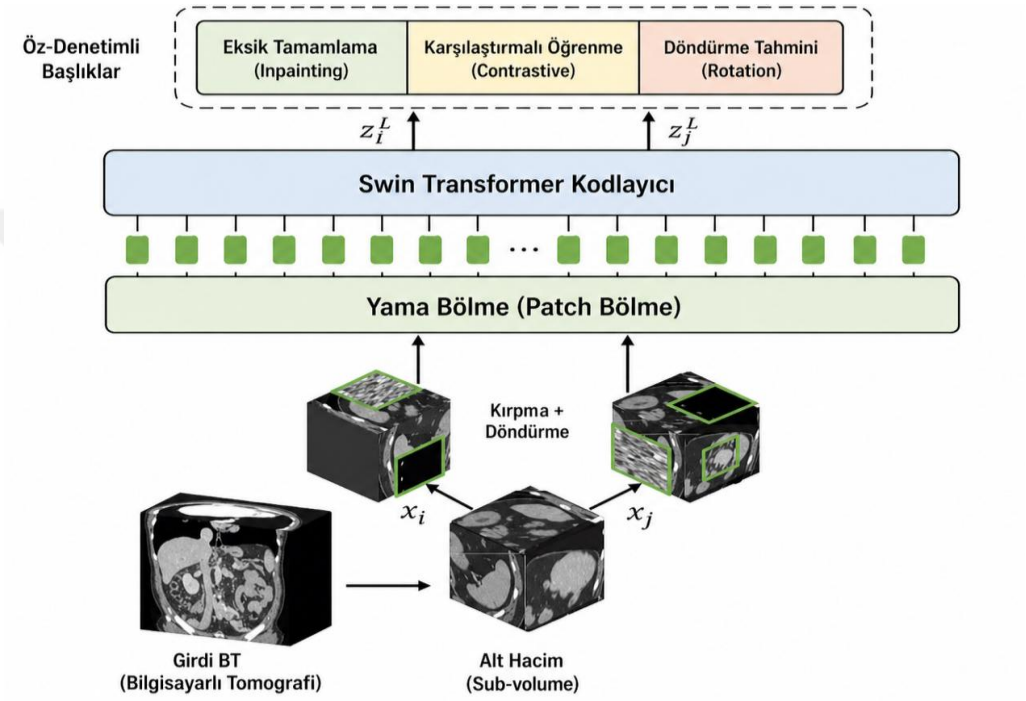


Şekil 2.5. nnU-Net mimarisi.

2.5.2. SwinUNETR

Hacimsel tıbbi görüntü segmentasyonu için Transformer tabanlı SwinUNETR mimarisi ile encoder–decoder yapısını benimsemekte olup, hiyerarşik özellik temsilleri 3D Swin Transformer blokları aracılığıyla çıkarılmaktadır [79]. Giriş olarak verilen ve üç boyutlu görüntünün uzamsal boyutlarını temsil eden H×W×D boyutundaki hacim, örtüşmeyen parçalara ayrılarak her bir parça token temsiline dönüştürülmektedir. Bu token'lar, pencere tabanlı çok başlı öz-dikkat mekanizması kullanılarak işlenmektedir. Pencereler arası bilgi etkileşimini sağlamak amacıyla, ardışık katmanlar boyunca kaydırılmış pencere dikkat mekanizması uygulanmaktadır. Bu yapı sayesinde model, uzun menzilli uzamsal bağımlılıkları etkin bir şekilde öğrenebilmekte ve hacimsel görüntülerdeki karmaşık yapıları daha başarılı bir şekilde temsil edebilmektedir [80]. Ağ yapısı, DWI, ADC, EDGE ve PRIOR temsillerinden oluşan çok kanallı bir giriş konfigürasyonu aracılığıyla çok modlu bilgi entegrasyonunu destekleyecek şekilde tasarlanmıştır. Encoder aşamasında çıkarılan çok ölçekli özellik haritaları, decoder aşamasında yukarı örnekleme işlemleri ve skip bağlantıları aracılığıyla kademeli olarak birleştirilmekte, böylece voxel düzeyinde hassas segmentasyon çıktıları elde edilmektedir.

SwinUNETR mimarisi (Şekil 2.6.), Transformer tabanlı küresel bağlam modelleme ile konvolüsyon tabanlı yerel özellik öğrenimini birlikte kullanan hibrit bir tasarım benimsemektedir. Swin Transformer blokları uzun menzilli uzamsal bağımlılıkları yakalarken, patch embedding ve downsampling aşamalarına entegre edilen konvolüsyon işlemleri ince detaylı yerel doku özelliklerinin korunmasını sağlamaktadır [81]. Bu yapı, CNN tabanlı yöntemlerin sınırlı bağlamsal modelleme kapasitesi ile saf Transformer mimarilerinin düşük seviyeli özellik temsili konusundaki yetersizlikleri arasında dengeli bir çözüm sunmaktadır. Model, 64×64×64 boyutlarında sabit hacimli ilgi bölgeleri (region of interest, ROI) üzerinde eğitilmiş olup, parametreleri Dice tabanlı bir kayıp fonksiyonu kullanılarak optimize edilmiştir.



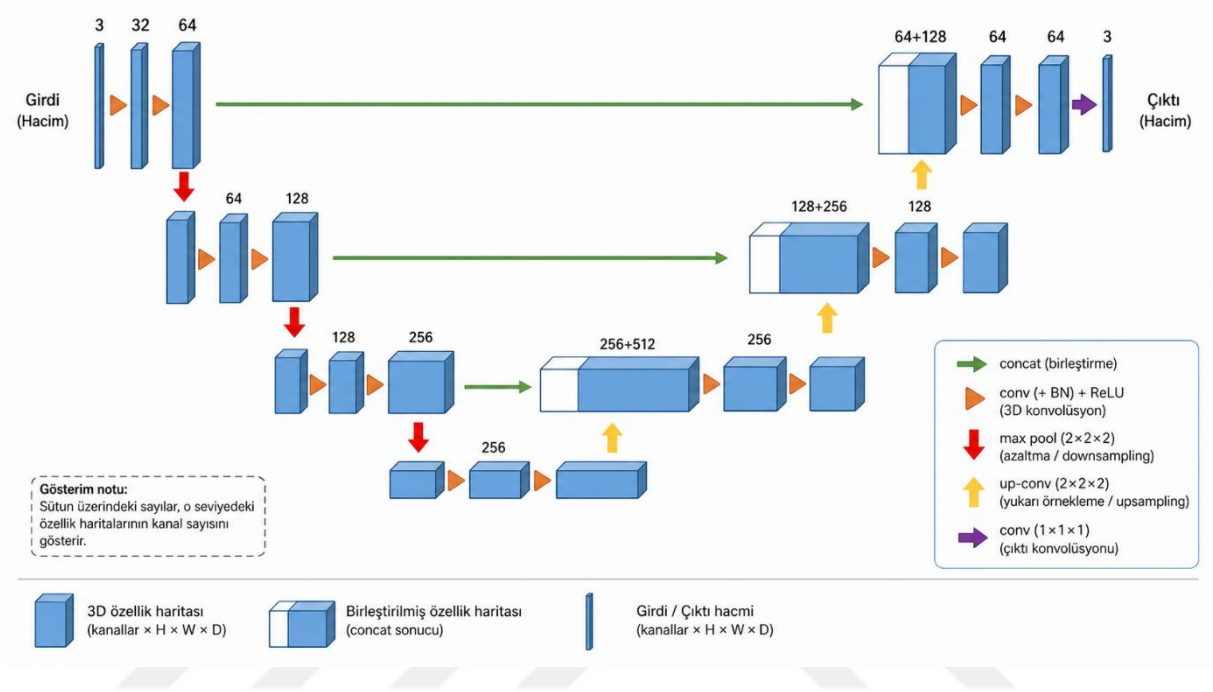
Şekil 2.6. SwinUNETR mimarisi.

2.5.3. 3D U-Net

3D U-Net segmentasyonu, U-Net mimarisinin hacimsel tıbbi görüntü verilerine genişletilmiş bir versiyonu olup, üç boyutlu bir görüntü hacmi içerisindeki yapıların voxel düzeyinde segmentasyonunu gerçekleştirmeyi mümkün kılmaktadır. 3D U-Net, orijinal 2D U-Net mimarisinin hacimsel segmentasyon görevleri için uyarlanmış bir türevidir. Bu mimari, tekil 2D kesitler yerine doğrudan hacimsel veriler üzerinde çalışabilmek için 3B konvolüsyon, 3D havuzlama ve 3D yukarı örnekleme işlemlerini kullanmaktadır [82]. Bu yaklaşım, tüm hacim verisinin doğrudan işlenmesini sağlayarak kesitler arası tutarlılığı artırmakta ve özellikle organ ve patoloji segmentasyonu gibi hacimsel veri içeren uygulamalarda yaygın olarak kullanılmaktadır [83].

Segmentasyon modeli(Şekil 2.7.), Özgün Çiçek ve arkadaşları tarafından önerilen 3D U-Net mimarisine dayanmaktadır [82]. Bu mimari, encoder–decoder yapısını takip etmekte olup, encoder kısmında 3D konvolüsyon katmanları aracılığıyla hiyerarşik özellik temsilleri kademeli olarak çıkarılmakta, decoder kısmında ise yukarı örnekleme işlemleri ve düşük seviyeli ile yüksek seviyeli özellikleri birleştiren skip bağlantıları sayesinde uzamsal

çözünürlük yeniden elde edilmektedir. Uygulanan yapılandırmada ağ, DWI, ADC, kenar tabanlı temsiller ve anatomik ön bilgi haritalarından oluşan dört kanallı bir giriş verisini işlemekte ve lezyon ile arka plan sınıfları için voxel düzeyinde tahminler üretmektedir. Çıkarım aşamasında, yüksek çözünürlüklü hacimsel verilerin etkin bir şekilde işlenebilmesi amacıyla sliding window stratejisi kullanılmaktadır. Model performansı ise, tıbbi görüntü segmentasyonu literatüründe yaygın olarak kullanılan standart bir metrik olan Dice Benzerlik Katsayısı ile nicel olarak değerlendirilmektedir.



Şekil 2.7. 3D U-Net mimarisi.

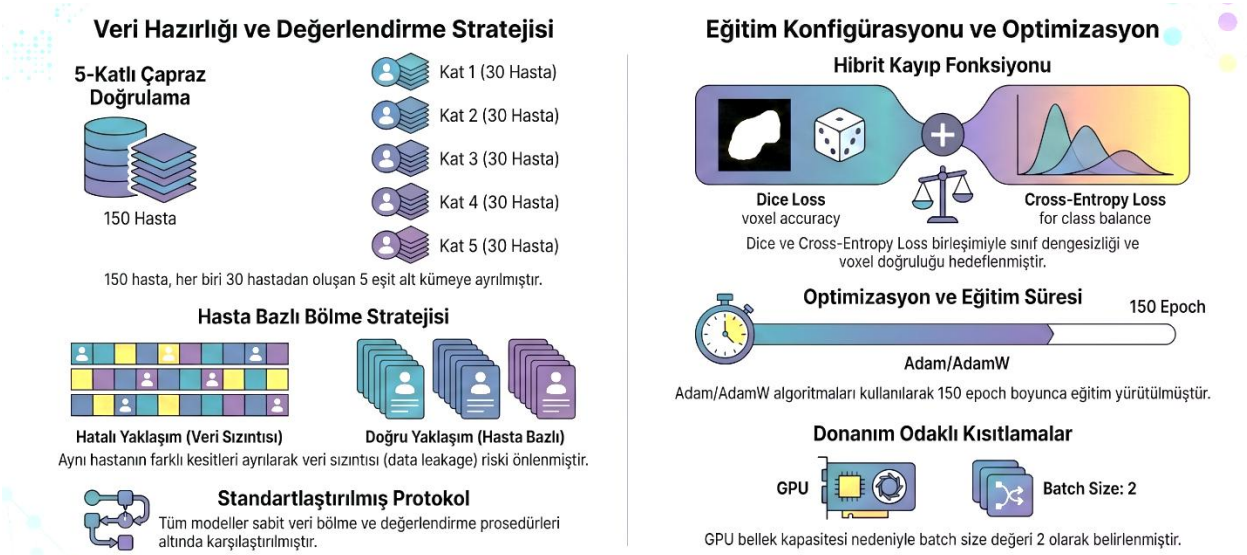
2.6. Eğitim ve Test Süreci

Bu tez çalışmasında geliştirilen segmentasyon modellerinin performansını güvenilir ve karşılaştırılabilir biçimde değerlendirebilmek amacıyla tüm deneysel süreçler standartlaştırılmış bir eğitim ve test protokolü altında yürütülmüştür. Farklı model mimarileri ve giriş konfigürasyonları arasında adil bir karşılaştırma sağlanabilmesi için veri bölme stratejisi, eğitim parametreleri ve değerlendirme prosedürleri tüm deneylerde sabit tutulmuştur.

Model performansının istatistiksel olarak güvenilir biçimde değerlendirilebilmesi amacıyla, hasta bazlı 5-katlı çapraz doğrulama yöntemi kullanılmıştır. Veri seti, her biri 30 hastadan oluşan ve toplamda 150 hastayı kapsayan beş eşit alt kümeye ayrılmıştır. Her iterasyonda dört alt küme eğitim amacıyla kullanılırken, kalan bir alt küme test seti olarak değerlendirilmiştir. Bu süreç beş kez tekrarlanarak her veri örneğinin bir kez test setinde yer alması sağlanmıştır. Hasta bazlı bölme stratejisi tercih edilerek aynı hastaya ait farklı kesitlerin hem eğitim hem de test setinde yer almasının önüne geçilmiş ve veri sızıntısı riski ortadan kaldırılmıştır. Böylece modelin gerçek genelleme performansı daha doğru biçimde ölçülmüştür.

Eğitim sürecinde modelin kararlı ve etkin bir şekilde yakınsamasını sağlamak amacıyla öğrenme oranı(Learning Rate-LR) düşük bir başlangıç değeri ile başlatılmış ve eğitim

süreci boyunca adaptif olarak güncellenmiştir. Optimizasyon algoritması olarak, parametre güncellemelerinde yüksek performans ve hızlı yakınsama sağlayan Adam ve AdamW yöntemleri tercih edilmiştir. Modelin segmentasyon başarımını artırmak ve hem sınıf dengesizliğini hem de voxel düzeyindeki doğruluğu dikkate almak amacıyla kayıp fonksiyonu olarak Dice Loss ve Cross-Entropy Loss'un birleşimi kullanılmıştır(Şekil 2.8). Eğitim süreci toplam 150 epoch boyunca yürütülmüş olup, bu değer modelin yakınsama davranışı göz önünde bulundurularak belirlenmiştir. Üç boyutlu volumetrik verilerin yüksek bellek gereksinimi nedeniyle batch size değeri 2 olarak seçilmiş ve bu tercih doğrudan GPU/ram kapasitesi ile sınırlı hesaplama kaynaklarına bağlı olarak yapılmıştır.



Şekil 2.8. Segmentasyon modeli eğitim ve değerlendirme metodolojisi.

2.7. Görüntü Değerlendirme Metrikleri

Model performansı hem segmentasyon doğruluğunu hem de klinik tespit güvenilirliğini kapsamlı bir şekilde değerlendirebilmek amacıyla voxel düzeyinde ve lezyon düzeyinde çeşitli metrikler kullanılarak nicel olarak analiz edilmiştir. Tüm metrikler, beş katlı çapraz doğrulama şeması kapsamında her bir fold için ayrı ayrı hesaplanmış ve sonuçlar fold'lar boyunca ortalama ± standart sapma şeklinde raporlanmıştır.

Tahmin edilen lezyon maskesi ile gerçek maskeler arasındaki uzamsal örtüşmeyi ölçmek amacıyla birincil değerlendirme metriği olarak Dice benzerlik katsayısı kullanılmıştır [84]. Dice metriği, lezyon ve arka plan voxel'ları arasındaki ciddi sınıf dengesizliği nedeniyle iskemik inme segmentasyonu problemleri için özellikle uygun bir ölçüttür. Burada P tahmin edilen lezyon maskesini, G ise gerçek lezyon maskesini temsil etmektedir.

$$Dice = \frac{2\,|\,P \cap G\,|}{|\,P\,| \cup |\,G\,|} \quad (8)$$

Tahmin edilen ve gerçek maskeler arasındaki kesişimin birleşime oranını hesaplayarak voxel düzeyinde örtüşmeyi nicel olarak ifade eder; bu yönüyle Dice katsayısına kıyasla daha katı bir uzamsal uyum ölçüsü sunmaktadır [85].

$$IoU = \frac{|P \cap G|}{|P \cup G|} \quad (9)$$

Voxel düzeyinde sınıflandırma performansını değerlendirmek amacıyla kesinlik ve duyarlılık metrikleri hesaplanmıştır. Kesinlik, lezyon olarak tahmin edilen tüm voxel'ler içerisinde doğru şekilde tahmin edilen lezyon voxel'lerinin oranını ifade ederken, duyarlılık ise modelin gerçek lezyon voxel'larını doğru şekilde tespit edebilme yeteneğini ölçmektedir.

Volumetrik Dice metriği, tüm kesitler boyunca birleştirilen tahmin edilen ve gerçek lezyon hacimleri arasındaki örtüşmeyi hesaplayarak, üç boyutlu lezyon hacmi düzeyinde segmentasyon uyumunu değerlendirmektedir:

$$vDSC = \frac{2\ x\ \sum_{i=1}^{N} |P_i \cap G_i|}{\sum_{i=1}^{N} |P_i| + \sum_{i=1}^{N} |G_i|} \quad (10)$$

Burada P_i ve G_i, sırasıyla $i.$ kesite ait tahmin edilen ve gerçek lezyon maskelerini, N ise bir hacimsel taramaya ait toplam kesit sayısını ifade etmektedir. Kesit bazlı Dice metriğinden farklı olarak, volumetrik Dice tüm kesitler boyunca biriken üç boyutlu lezyon hacmini dikkate almakta olup, bu yönüyle hacimsel tutarlılığın değerlendirilmesi ve klinik açıdan anlamlı lezyon yükünün belirlenmesi için daha uygun bir ölçüt sunmaktadır.

2.8. Otomatik Rapor Üretim Sistemi Genel Mimarisi

Bu tez kapsamında geliştirilen otomatik radyolojik rapor üretim sistemi, beyin MRI görüntülerinden elde edilen bulguların klinik açıdan anlamlı ve yapılandırılmış metinlere dönüştürülmesini sağlayan çok aşamalı ve çok modlu bir yapay zekâ çerçevesi olarak tasarlanmıştır. Önerilen sistem, görüntü işleme, özellik çıkarımı, klinik bilgi entegrasyonu ve doğal dil üretimi bileşenlerini bütünleşik bir mimari altında bir araya getirmektedir.

Sistem genel olarak dört temel aşamadan oluşmaktadır: (i) görüntü tabanlı analiz ve segmentasyon, (ii) klinik olarak anlamlı özelliklerin çıkarımı, (iii) yapılandırılmış bilginin oluşturulması ve (iv) büyük dil modeli aracılığıyla doğal dilde rapor üretimi. İlk aşamada, DWI ve yardımcı sekanslar kullanılarak derin öğrenme tabanlı segmentasyon modelleri aracılığıyla inme lezyonları voxel düzeyinde tespit edilmektedir. Bu aşamanın çıktısı, lezyonun uzamsal konumunu ve hacimsel dağılımını içeren üç boyutlu segmentasyon maskeleridir.

İkinci aşamada, elde edilen segmentasyon çıktıları üzerinden klinik açıdan anlamlı özellikler çıkarılmaktadır. Bu kapsamda lezyonun hacmi, konumu, anatomik bölgesi ve olası vasküler sulama alanı gibi parametreler hesaplanmaktadır. Bu özellikler, yalnızca sayısal değerler olarak değil, aynı zamanda klinik yorumlamaya uygun yapılandırılmış bilgi temsilleri şeklinde düzenlenmektedir. Üçüncü aşamada, görüntü tabanlı özellikler anatomik atlaslar ve klinik bilgi kaynakları ile entegre edilerek semantik açıdan zenginleştirilmiş bir veri yapısı oluşturulmaktadır. Bu entegrasyon sayesinde model, yalnızca görüntüdeki sinyal değişimlerini değil, aynı zamanda bu değişimlerin klinik anlamını da öğrenebilmektedir. Böylece lezyonların yalnızca tespiti değil, klinik bağlam içerisinde yorumlanması da mümkün hale gelmektedir.

Son aşamada ise, elde edilen yapılandırılmış klinik bilgiler LLM tabanlı bir metin üretim sistemi tarafından işlenerek doğal dilde radyolojik raporlara dönüştürülmektedir. Bu süreçte model, klinik terminolojiye uygun, tutarlı ve bağlamsal olarak anlamlı cümleler üretecek şekilde eğitilmiştir. Üretilen raporlar, geleneksel radyoloji raporlarının yapısına uygun olarak bulguların tanımlanması, lokalizasyon bilgisi ve klinik yorum içerecek şekilde oluşturulmaktadır.

Önerilen bu mimari, görüntü tabanlı analiz ile doğal dil üretimi süreçlerini tek bir çatı altında birleştirerek, uçtan uca çalışan bir klinik karar destek sistemi sunmaktadır. Sistem, radyologların iş yükünü azaltmayı, raporlama sürecini hızlandırmayı ve raporların tutarlılığını artırmayı hedeflemektedir. Ayrıca, çok modlu veri entegrasyonu sayesinde yalnızca görüntüye dayalı değil, klinik açıdan anlamlı ve yorumlanabilir çıktılar üretilmesini sağlamaktadır.

2.9. Görüntü Tabanlı Özellik Çıkarımı

Otomatik rapor üretim sisteminin temel bileşenlerinden biri, manyetik rezonans görüntülerinden elde edilen verilerin klinik açıdan anlamlı özelliklere dönüştürülmesini sağlayan görüntü tabanlı özellik çıkarım sürecidir. Bu süreçte amaç, ham görüntü verilerinden yalnızca görsel temsiller elde etmek değil, aynı zamanda bu temsilleri klinik karar verme süreçlerinde kullanılabilecek yapılandırılmış bilgiye dönüştürmektir.

Özellik çıkarım süreci, derin öğrenme tabanlı segmentasyon modelleri tarafından elde edilen lezyon maskeleri temel alınarak gerçekleştirilmiştir. Önceki aşamada elde edilen üç boyutlu segmentasyon çıktıları, her bir voxel'in lezyon ya da sağlıklı doku olarak sınıflandırıldığı ikili maskeler şeklinde temsil edilmektedir. Bu maskeler kullanılarak lezyon bölgelerine ait uzamsal ve hacimsel özellikler hesaplanmaktadır.

İlk olarak, lezyon hacmi hesaplanmaktadır. Hacim hesaplama işlemi, segmentasyon maskesinde lezyon olarak işaretlenen voxel sayısının, her bir voxel'in fiziksel boyutları ile çarpılmasıyla gerçekleştirilmektedir. Bu işlem aşağıdaki şekilde ifade edilmektedir:

$$V = N_{lezyon}\ x\ s_x\ x\ s_y\ x\ s_z \qquad (11)$$

Burada N_{lesion}, lezyon olarak etiketlenen toplam voxel sayısını; s_x, s_y, s_z ise sırasıyla x, y ve z eksenlerindeki voxel boyutlarını temsil etmektedir. Bu hesaplama sonucunda lezyon hacmi milimetreküp (mm^3) veya santimetreküp (cc) cinsinden elde edilmektedir. Bu metrik, klinik değerlendirme açısından önemli bir parametre olup, inmenin şiddeti ve yayılımı hakkında doğrudan bilgi sunmaktadır.

İkinci olarak, lezyonun uzamsal konumu belirlenmektedir. Bu kapsamda, segmentasyon maskesi üzerinde lezyon bölgelerinin koordinat bilgileri analiz edilerek lezyonun görüntü hacmi içerisindeki dağılımı çıkarılmaktadır. Lezyonun merkez noktası, sınır koordinatları ve uzamsal yayılımı gibi bilgiler hesaplanarak, lezyonun beynin hangi bölgesinde yer aldığına dair nicel bir temsil elde edilmektedir. Buna ek olarak, lezyon sınırlarının daha belirgin şekilde modellenebilmesi amacıyla kenar tabanlı özellikler de çıkarılmaktadır. Bu kapsamda, DWI görüntüleri üzerinde Sobel operatörü kullanılarak üç boyutlu gradyan hesaplaması yapılmakta ve görüntüdeki yoğunluk değişimlerine karşılık gelen kenar haritaları elde edilmektedir. Elde edilen bu kenar temsilleri, özellikle düşük kontrastlı veya

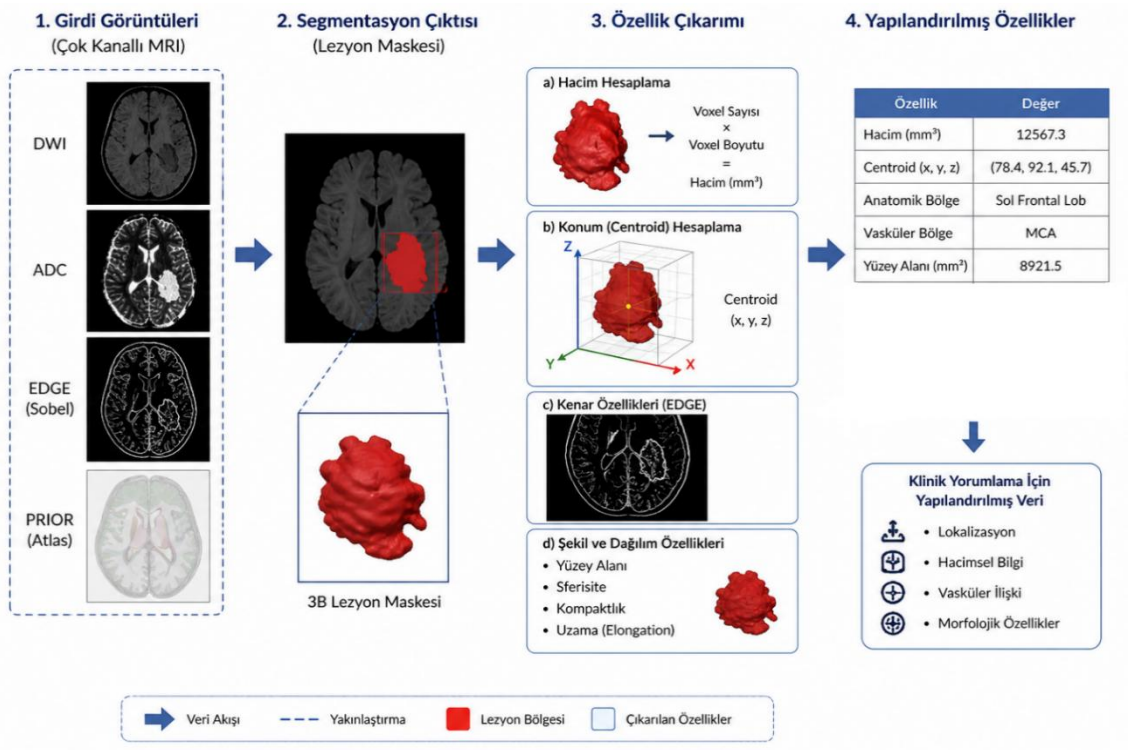
sınırları belirsiz lezyon bölgelerinde, lezyonun morfolojik yapısının daha iyi anlaşılmasına katkı sağlamaktadır.

Özellik çıkarımı sürecinde ayrıca çok kanallı giriş yapısının sağladığı avantajlardan da yararlanılmaktadır. DWI ve ADC sekanslarının birlikte kullanılması sayesinde, lezyon bölgelerinde gözlenen sinyal değişimlerinin doğrulanması ve yanlış pozitif tespitlerin azaltılması hedeflenmektedir. Bu sayede model, yalnızca tek bir görüntü kaynağına bağlı kalmadan daha güvenilir ve sağlam özellik temsilleri öğrenebilmektedir. Tablo 2.2'de, MRI görüntülerinden çıkarılan temel özellikler, bu özelliklerin açıklamaları ve klinik açıdan taşıdığı önem birlikte sunulmaktadır.

Tablo 2.2. Görüntülerden çıkarılan özellikler.

Özellik	Açıklama	Klinik Önemi
Lezyon Hacmi	Toplam lezyon voxel hacmi	İnme şiddeti
Konum	Lezyonun merkez koordinatı	Lokalizasyon
Anatomik Bölge	Atlas ile eşleşen bölge	Klinik yorum
Edge Özellikleri	Sobel tabanlı kenar bilgisi	Sınır doğruluğu

Elde edilen tüm bu özellikler, doğrudan ham veri olarak kullanılmak yerine, klinik açıdan anlamlı ve yorumlanabilir bir yapı içerisinde yeniden düzenlenmektedir. Bu kapsamda lezyon hacmi, konum bilgisi ve morfolojik özellikler yapılandırılmış veri formatına dönüştürülerek bir sonraki aşamada kullanılacak olan klinik bilgi entegrasyonu ve rapor üretim süreçlerine aktarılmaktadır.



Şekil 2.9. MRI tabanlı özellik çıkarım ve yapılandırılmış klinik bilgi oluşturma aşamaları.

Bu sürece ilişkin genel iş akışı, ham MRI görüntülerinden segmentasyon maskelerinin elde edilmesi, ardından klinik açıdan anlamlı özelliklerin çıkarılması ve yapılandırılmış veri formatına dönüştürülmesi adımlarını içerecek şekilde Şekil 2.9'da sunulmuştur.

2.10. Anatomik ve Klinik Bilgi Entegrasyonu

Bu tez kapsamında geliştirilen otomatik radyolojik rapor üretim sisteminde, segmentasyon modelleri tarafından elde edilen lezyon maskelerinin yalnızca sayısal çıktılar olarak değerlendirilmesinin yeterli olmadığı görülmüştür. Maskelerden elde edilen uzamsal ve hacimsel bilgiler anatomik atlaslar, vasküler sulama alanı haritaları ve klinik yorumlama kuralları ile bütünleştirilerek klinik açıdan anlamlı yapılandırılmış bilgilere dönüştürülmüştür. Bu aşama, görüntü tabanlı yapay zekâ çıktıları ile radyolojik rapor üretimi arasındaki en önemli köprü bileşenlerinden biridir.

Önerilen sistemde anatomik ve klinik bilgi entegrasyonu dört temel aşamadan oluşmaktadır: ilk olarak lezyon maskesinden temel ölçümsel bilgiler çıkarılmış, ikinci olarak lezyonun hemisferik (beynin sağ veya sol yarısı) yerleşimi belirlenmiş, üçüncü olarak anatomik atlaslar aracılığıyla kortikal ve subkortikal bölge eşlemesi yapılmış, son olarak ise vasküler sulama alanı atlası kullanılarak lezyonun olası arteriyel dağılım alanı tespit edilmiştir. Elde edilen tüm bu bilgiler daha sonra büyük dil modeline aktarılmak üzere yapılandırılmış bulgu formatına dönüştürülmüştür.

2.10.1. Segmentasyon Maskesinden Klinik Metriklerin Çıkarılması

Anatomik ve klinik entegrasyon sürecinin ilk aşamasında, derin öğrenme tabanlı segmentasyon modeli tarafından üretilen lezyon maskesi kullanılarak temel klinik metrikler hesaplanmıştır. Segmentasyon maskesi, her bir voxel'in lezyon veya arka plan olarak sınıflandırıldığı ikili bir hacimsel veri yapısıdır. Bu maske üzerinden lezyon sayısı, toplam lezyon hacmi, maksimum lezyon çapı, yerleşim tarafı ve ADC değeri gibi klinik açıdan anlamlı parametreler çıkarılmıştır.

Lezyon hacmi, segmentasyon maskesi içerisinde lezyon olarak etiketlenen toplam voxel sayısının, her bir voxel'in fiziksel boyutları ile çarpılmasıyla hesaplanmıştır. Bu işlem aşağıdaki denklem (12) ile ifade edilmektedir:

$$V_{lezyon} = N_{lezyon} \times s_x \times s_y \times s_z \qquad (12)$$

Burada V_{lesion} toplam lezyon hacmini, N_{lesion} olarak işaretlenen toplam voxel sayısını, s_x, s_y ve s_z ise sırasıyla x, y ve z eksenlerindeki voxel boyutlarını ifade etmektedir. Hacim değeri milimetreküp cinsinden elde edildikten sonra klinik raporlama açısından daha anlaşılır olması amacıyla mililitreye dönüştürülmüştür:

$$V_{ml} = \frac{V_{mm^3}}{1000} \qquad (13)$$

Bu dönüşüm sayesinde lezyon yükü klinik açıdan daha yorumlanabilir hale getirilmiştir. Örneğin, küçük hacimli odaklar "milimetrik difüzyon kısıtlanması", daha yüksek hacimli lezyonlar ise "geniş difüzyon kısıtlanma alanı" olarak rapor üretim sürecine aktarılmıştır.

Lezyonun maksimum çapı ise segmentasyon maskesi üzerinde en büyük lezyon bileşeni dikkate alınarak hesaplanmıştır. Bu işlemde bağlı bileşen analizi ile lezyon odakları ayrılmış, en büyük bileşen seçilmiş ve bu bileşenin aksiyal düzlemdeki maksimum yayılımı ölçülmüştür. Maksimum çap aşağıdaki şekilde ifade edilebilir:

$$D_{max} = \max \sqrt{(x_{max} - x_{min})^2 s_x^2 + (y_{max} - y_{min})^2 s_y^2} \tag{14}$$

Burada x_{max}, x_{min}, y_{max} ve y_{min}, lezyonun ilgili kesitteki sınır koordinatlarını; s_x ve s_y ise voxel boyutlarını göstermektedir. Bu değer, özellikle lezyonun yayılımını raporlamak ve büyük/küçük lezyon ayrımını desteklemek amacıyla kullanılmıştır.

2.10.2. Yerleşim Tarafı Bilgisinin Belirlenmesi

Lezyonun sağ, sol veya bilateral hemisfer yerleşimli olup olmadığının belirlenmesi, radyolojik rapor üretimi açısından temel bir adımdır. Bu tezde yerleşim tarafı hesabı, segmentasyon maskesinin standart RAS (Right–Anterior–Superior) koordinat sistemine hizalanmış görüntü uzayına dönüştürülmesinden sonra x ekseni boyunca lezyon voxel dağılımının analiz edilmesiyle gerçekleştirilmiştir.

RAS koordinat sisteminde x ekseninin pozitif yönü anatomik sağ hemisferi, negatif yönü ise anatomik sol hemisferi temsil etmektedir. Bu nedenle maske hacmi orta hattan ikiye ayrılmış ve sağ ve sol hemisferdeki lezyon voxel sayıları ayrı ayrı hesaplanmıştır:

$$\begin{aligned} N_{sol} &= \sum M(x,y,z), x < x_{mid} \\ N_{sağ} &= \sum M(x,y,z), x \geq x_{mid} \end{aligned} \tag{15}$$

Burada $M(x,y,z)$ segmentasyon maskesini, x_{mid} ise hacmin orta hattını temsil etmektedir. Yerleşim tarafı kararı aşağıdaki kurala göre verilmiştir:

$$Yerleşim\ Tarafı = \begin{cases} Sol, & N_{sol} > 1.5 \times N_{sağ} \\ Sağ, & N_{sağ} > 1.5 \times N_{sol} \\ Bilateral, & \text{diğer durumlarda} \end{cases} \tag{16}$$

Bu yaklaşım, küçük hacimli gürültülerin veya sınıra yakın voxel dağılımlarının yanlış yerleşim tarafı üretmesini azaltmak amacıyla oran tabanlı bir karar mekanizması kullanmaktadır. Elde edilen yerleşim tarafı bilgisi hem anatomik bölge yorumlamasında hem de vasküler atlas eşlemesinde kontrol değişkeni olarak kullanılmıştır.

2.10.3. Anatomik Atlaslarla Bölge Eşlemesi

Lezyonun yalnızca görüntü koordinatları ile ifade edilmesi klinik açıdan yeterli değildir. Radyolojik raporlarda lezyonun frontal lob, parietal lob, temporal lob, oksipital lob, talamus, putamen veya serebral korteks gibi anatomik yapılarla ilişkilendirilmesi beklenmektedir. Bu nedenle bu tezde segmentasyon maskeleri anatomik atlaslarla eşleştirilmiştir.

Bu amaçla Harvard-Oxford kortikal ve subkortikal atlasları [86-88] kullanılmıştır. Kortikal atlas, serebral korteks içerisindeki girüs ve lob düzeyindeki anatomik yapıları temsil ederken; subkortikal atlas talamus, kaudat nükleus, putamen, globus pallidus, hipokampus, amigdala ve beyin sapı gibi derin gri cevher ve subkortikal yapıları içermektedir.

Hasta uzayındaki DWI görüntüsü ile atlas uzayı arasında doğrudan birebir örtüşme bulunmadığından, atlas görüntüleri hasta DWI uzayına hizalanmıştır. Bu işlem düzlem hizalama kullanılarak gerçekleştirilmiştir. Genel hizalama dönüşüm aşağıdaki şekilde ifade edilmektedir:

$$\mathrm{x}' = \mathrm{Ax} + \mathrm{t} \qquad (17)$$

Burada x atlas uzayındaki voxel koordinatını, x' hasta görüntü uzayındaki karşılık gelen koordinatı, A üç boyutlu hizalama dönüşüm matrisini ve t öteleme vektörünü ifade etmektedir. Hizalama dönüşüm matrisi daha açık biçimde aşağıdaki gibi gösterilebilir:

$$\begin{bmatrix} x' \\ y' \\ z' \\ 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & t_x \\ a_{21} & a_{22} & a_{23} & t_y \\ a_{31} & a_{32} & a_{33} & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \qquad (18)$$

Bu hizalama sonucunda atlas etiketleri hasta görüntüsüyle aynı uzamsal koordinat sistemine taşınmıştır. Atlas verileri etiket haritası niteliğinde olduğundan, yeniden örnekleme sırasında en yakın komşu yaklaşımıyla yeniden örnekleme kullanılmıştır. Böylece atlas etiketlerinin kesirli değerlere dönüşmesi engellenmiş ve anatomik etiket bütünlüğü korunmuştur.

Atlas ve lezyon maskesi aynı uzaya getirildikten sonra, lezyon maskesi ile atlas etiketleri voxel düzeyinde çakıştırılmıştır. Lezyon maskesi içinde kalan atlas etiketleri sayılarak en fazla örtüşen anatomik bölgeler belirlenmiştir. Bir anatomik bölgeye ait örtüşme oranı aşağıdaki şekilde hesaplanmıştır:

$$P_i = \frac{N_i}{N_{toplam}} \times 100 \qquad (19)$$

Burada P_i, i'nci anatomik bölgenin lezyon ile örtüşme yüzdesini; N_i, ilgili atlas etiketine düşen lezyon voxel sayısını; N_{toplam} ise atlas üzerinde anatomik karşılığı bulunan toplam lezyon voxel sayısını ifade etmektedir. Bu işlem sonucunda her hasta için en fazla örtüşen kortikal ve subkortikal bölgeler belirlenmiştir. Örneğin, bir lezyonun büyük bölümü "superior frontal girus" ve "sol serebral korteks" ile örtüşüyorsa sistem bunu kortikal ve subkortikal anatomik bilgi olarak kaydetmektedir.

2.10.4. Vasküler Sulama Alanı Eşlemesi

İskemik inme değerlendirmesinde lezyonun yalnızca anatomik lokalizasyonu değil, aynı zamanda hangi arteriyel sulama alanında yer aldığı da klinik açıdan önemlidir. Bu nedenle bu tezde anatomik atlaslara ek olarak vasküler sulama alanı atlası kullanılmıştır [75].

Vasküler atlas, beyin dokusunu anterior ACA, MCA, PCA ve vertebrobaziler sistem gibi arteriyel dağılım bölgelerine ayırmaktadır. Segmentasyon maskesi bu atlas ile çakıştırılarak lezyonun hangi vasküler sulama alanına karşılık geldiği belirlenmiştir. Vasküler alan skoru aşağıdaki şekilde hesaplanmıştır:

$$S_k = \sum_{i \in T_k} N_i \qquad (20)$$

Burada S_k, k'ncı vasküler sulama alanının toplam skorunu; T_k, ilgili vasküler alana ait atlas etiketlerini; N_i ise i'nci vasküler atlas etiketine karşılık gelen lezyon voxel sayısını ifade etmektedir. En yüksek skora sahip vasküler alan, baskın vasküler sulama alanı olarak seçilmiştir:

$$T_{en\,iyi} = \arg\max_k S_k \qquad (21)$$

Bu işlem sonucunda her hasta için vasküler sulama alanı "MCA", "ACA", "PCA" veya "vertebrobaziler" gibi üst düzey bir kategori olarak belirlenmiştir. Ayrıca daha ayrıntılı vasküler bilgi üretmek amacıyla atlas etiketleri "sol MCA frontal dal sulama alanı", "sağ MCA parietal dal sulama alanı", "sol PCA oksipital dal sulama alanı" gibi detaylı klinik ifadelere dönüştürülmüştür. Vasküler atlas eşlemesinde yerleşim tarafı tutarlılığı ayrıca kontrol edilmiştir. Atlas hizalama sürecinde küçük uzaysal kaymalar meydana gelebileceğinden, atlas tarafından verilen vasküler taraf bilgisi segmentasyon maskesinden hesaplanan yerleşim tarafı ile karşılaştırılmıştır. Eğer atlas tahmini ile maske tabanlı yerleşim tarafı çelişirse, aynı vasküler bölgenin doğru hemisfer tarafındaki karşılığı seçilmiştir. Bu kontrol, özellikle otomatik rapor üretimi aşamasında yanlış hemisfer bildiriminin önlenmesi açısından önemlidir.

2.10.5. ADC Değerinin Klinik Yoruma Dönüştürülmesi

Bu tezde DWI görüntülerine ek olarak ADC haritaları da kullanılmıştır. ADC değeri, akut iskemik lezyonların değerlendirilmesinde önemli bir klinik göstergedir. Akut iskemide hücre içi ödem nedeniyle su moleküllerinin hareketi kısıtlanır; bu durum DWI görüntülerinde hiperintensite, ADC haritalarında ise düşük değer olarak izlenir.

Segmentasyon maskesi ADC haritası üzerine uygulanarak yalnızca lezyon bölgesindeki ADC değerleri alınmıştır. Aykırı değerlerin etkisini azaltmak amacıyla ortalama yerine medyan ADC değeri kullanılmıştır:

$$ADC_{lezyon} = median(ADC(x,y,z)), (x,y,z) \in M \qquad (22)$$

Burada M, lezyon maskesi içerisinde kalan voxel kümesini temsil etmektedir. ADC değerleri farklı cihaz veya dışa aktarım biçimlerine bağlı olarak farklı ölçeklerde olabileceğinden, değerler otomatik ölçek düzeltme işleminden geçirilmiştir. Böylece ADC değerleri klinik olarak yorumlanabilir aralığa dönüştürülmüştür. Lezyon bölgesinden elde edilen ADC değerleri, klinik anlamlılık açısından belirli aralıklara ayrılarak yorumlanmış olup, bu sınıflandırma Tablo 2.3'te detaylı olarak verilmiştir.

Tablo 2.3. ADC değer aralıklarına göre difüzyon kısıtlanmasının klinik yorumlanması.

ADC aralığı ($\times 10^{-6}$ mm^2/s)	Klinik yorum
< 400	Çok düşük ADC; belirgin akut difüzyon kısıtlanması
400 -620	Düşük ADC; akut iskemik difüzyon kısıtlanması ile uyumlu
620–900	Orta-düşük ADC; subakut iskemi ile uyumlu olabilir
900–1200	Normal sınıra yakın ADC; subakut/iyileşme dönemi veya ödem ile ilişkili olabilir
> 1200	Yüksek ADC; kronik değişiklik veya sıvı komponenti içerebilir

Bu yorumlama sonucunda ADC değeri doğrudan sayısal bir parametre olarak değil, klinik anlam taşıyan bir ifade olarak rapor üretim sistemine aktarılmıştır. Örneğin, "ADC değeri düşüktür" ifadesi sistem tarafından "akut iskemik difüzyon kısıtlanması ile uyumlu" şeklinde yapılandırılmış bilgiye dönüştürülmüştür.

2.10.6. Anatomik, Vasküler ve Klinik Bilgilerin Yapılandırılması

Anatomik ve klinik entegrasyon aşamasında elde edilen tüm bilgiler, büyük dil modelinin anlayabileceği yapılandırılmış bir giriş formatına dönüştürülmüştür. Bu yapılandırılmış format, ham görüntü verisi yerine klinik açıdan anlamlı öznitelikler içermektedir. Böylece LLM'e doğrudan görüntü veya maske verilmemekte; bunun yerine görüntüden çıkarılmış, klinik olarak yorumlanabilir bulgular aktarılmaktadır.

Tablo 2.4. Görüntü tabanlı analizden elde edilen yapılandırılmış klinik özellikler ve klinik karşılıkları.

Değişken	Açıklama	Klinik karşılık
lesion_count	Lezyon odaklarının sayısı	Tek/multifokal enfarkt değerlendirmesi
total_volume_ml	Toplam lezyon hacmi	Lezyon yükü ve yaygınlık
max_diameter_mm	En büyük lezyon çapı	Lezyon boyutu
laterality	Sağ, sol veya bilateral yerleşim	Hemisferik lokalizasyon
cortical_region	Kortikal atlas eşleşmesi	Kortikal anatomik bölge
subcortical_region	Subkortikal atlas eşleşmesi	Derin gri cevher/subkortikal tutulum
anatomical_region	Kortikal ve subkortikal bilgilerin birleşimi	Genel anatomik lokalizasyon
vascular_territory	ACA/MCA/PCA/vertebrobaziler alan	Arteriyel sulama alanı
adc_mean	Düzeltilmiş medyan ADC değeri	Difüzyon kısıtlanmasının nicel göstergesi
adc_interpretation	ADC'nin klinik yorumu	Akut/subakut/kronik süreç yorumu

Görüntü tabanlı analiz ve anatomik eşleme süreçleri sonucunda elde edilen yapılandırılmış klinik değişkenler ve bu değişkenlerin klinik karşılıkları Tablo 2.4'te sunulmuştur. Bu yapılandırılmış bilgiler daha sonra doğal dil üretimi için kullanılacak olan kullanıcı girdisine dönüştürülmüştür. Örnek bir yapılandırılmış bulgu formatı Tablo 2.5'te verilmiştir:

Tablo 2.5. Yapılandırılmış klinik bilgilerin örnek çıktı formatı.

Yapılandırılmış bilgi	Örnek çıktı
Lezyon sayısı	1
Lezyon büyüklüğü	Küçük-orta boyutlu difüzyon kısıtlanması
Yerleşim tarafı	Sol hemisfer
Kortikal bölge	Superior frontal girus
Subkortikal bölge	Sol serebral beyaz cevher
Vasküler sulama alanı	MCA
ADC yorumu	Düşük ADC; akut iskemik difüzyon kısıtlanması ile uyumlu

Bu yapı, otomatik rapor üretim sisteminin halüsinasyon riskini azaltmak ve yalnızca verilen klinik bulgular üzerinden rapor oluşturmasını sağlamak amacıyla kullanılmıştır. Böylece modelin rapor üretimi sırasında görüntüde bulunmayan bir lokalizasyon, ek lezyon veya vasküler alan uydurmasının önüne geçilmesi hedeflenmiştir.

2.10.7. Sayısal Bulgulardan Klinik İfade Üretimi

Radyoloji raporlarında sayısal değerlerin doğrudan verilmesi çoğu zaman yeterli değildir. Bu nedenle bu tezde bazı sayısal parametreler, raporlamada kullanılan nitel ifadelere dönüştürülmüştür. Özellikle lezyon hacmi ve lezyon sayısı birlikte değerlendirilerek lezyon büyüklüğüne ilişkin klinik ifadeler oluşturulmuştur.

Lezyon hacmine dayalı nitel sınıflandırma Tablo 2.6'da verilmiştir.

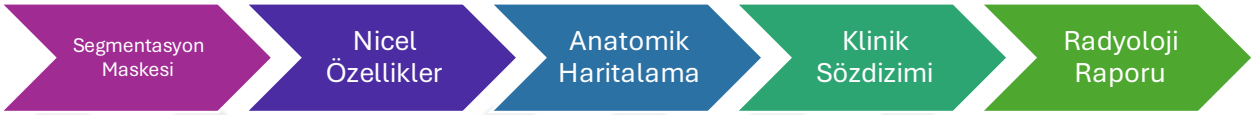
Tablo 2.6. Lezyon hacmine göre nitel rapor ifadesi.

Koşul	Üretilen klinik ifade
Çok sayıda küçük odak ve toplam hacim < 30 mL	Milimetrik-noktasal difüzyon kısıtlanma odakları
≥3 lezyon ve toplam hacim < 10 mL	Milimetrik boyutlu çok odaklı difüzyon kısıtlanması
Hacim < 2 mL	Noktasal/milimetrik difüzyon kısıtlanması
2–5 mL	Milimetrik boyutlu difüzyon kısıtlanması
5–15 mL	Küçük-orta boyutlu difüzyon kısıtlanması
15–40 mL	Orta büyüklükte difüzyon kısıtlanma alanı
40–100 mL	Geniş difüzyon kısıtlanma alanı
>100 mL	Çok geniş, yaygın difüzyon kısıtlanma alanı

Bu dönüşüm sayesinde model çıktısı yalnızca "12.4 mL lezyon" gibi nicel bir ifade ile sınırlı kalmamakta, bunun yerine radyoloji pratiğinde daha doğal kullanılabilecek "küçük-orta boyutlu difüzyon kısıtlanması" gibi klinik ifadeler üretebilmektedir.

2.10.8. Anatomik ve Klinik Bilginin Rapor Üretimindeki Rolü

Bu tezde anatomik ve klinik bilgi entegrasyonu, otomatik rapor üretim sisteminin doğruluğunu ve klinik tutarlılığını artıran temel aşamalardan biridir. Segmentasyon modeli yalnızca lezyonun nerede olduğunu voxel düzeyinde belirlerken, anatomik ve vasküler entegrasyon aşaması bu voxel bilgilerinin klinik dilde yorumlanmasını sağlamaktadır. Bu dönüşüm aşağıdaki şekilde özetlenebilir:



Şekil 2.10. Otomatik radyolojik rapor üretiminin genel aşamaları.

Önerilen sistemde, segmentasyon çıktılarının nicel özelliklere dönüştürülmesi, anatomik ve klinik bilgi ile zenginleştirilmesi ve nihayetinde doğal dilde radyolojik rapora aktarılması süreci Şekil 2.10'da şematik olarak gösterilmektedir. Bu süreçte segmentasyon maskesi üzerinden hacim, çap ve yerleşim tarafı gibi nicel özellikler çıkarılmakta; anatomik atlaslar aracılığıyla lezyonun kortikal ve subkortikal karşılığı belirlenmekte; vasküler atlas ile arteriyel sulama alanı saptanmakta; ADC analizi ile difüzyon kısıtlanmasının klinik yorumu yapılmakta ve tüm bu bilgiler yapılandırılmış bir LLM girdisine dönüştürülmektedir.

Bu yaklaşım, klasik görüntü tabanlı segmentasyon sistemlerinden farklı olarak, model çıktısının yalnızca maske veya olasılık haritası olarak kalmasını engellemekte ve bu çıktıları klinik karar destek sürecinde kullanılabilecek açıklanabilir bilgilere dönüştürmektedir. Böylece geliştirilen sistem, yalnızca "lezyon var/yok" veya "maske üretimi" yapan bir araç olmaktan çıkarak, lezyonun nerede olduğu, ne kadar büyük olduğu, hangi vasküler alanda bulunduğu ve difüzyon özelliklerinin neyi düşündürdüğü gibi radyolojik açıdan kritik sorulara yanıt verebilen bütüncül bir yapıya kavuşmaktadır.

2.11. Metin Verisinin Hazırlanması

Bu tez kapsamında geliştirilen otomatik radyolojik rapor üretim sisteminde, LLM'in etkin şekilde eğitilebilmesi ve doğru çıktılar üretebilmesi amacıyla metin verileri kapsamlı bir ön işleme sürecinden geçirilmiştir. Bu süreçte, ham radyoloji raporları klinik anlam bütünlüğü korunacak şekilde temizlenmiş, standartlaştırılmış ve yapılandırılmış veri formatına dönüştürülmüştür.

İlk aşamada, veri setinde yer alan radyolojik raporlar incelenmiş ve modelin öğrenmesini olumsuz etkileyebilecek gereksiz ifadeler, özel karakterler ve format farklılıkları giderilmiştir. Bu kapsamda metinler küçük harfe dönüştürülmüş, noktalama işaretleri standartlaştırılmış ve tekrar eden boşluklar temizlenmiştir. Bunun yanı sıra, rapor içerisinde

klinik anlam taşımayan başlık, tarih veya teknik cihaz bilgileri gibi bölümler veri setinden çıkarılmıştır.

İkinci aşamada, metinlerdeki terminolojik tutarlılığın sağlanmasına yönelik normalizasyon işlemleri gerçekleştirilmiştir. Aynı klinik durumu ifade eden farklı yazım biçimleri (örneğin "difüzyon kısıtlanması", "difüzyon kısıtlılığı") ortak bir ifade altında birleştirilmiştir. Bu işlem, modelin farklı ifadeleri ayrı kavramlar olarak öğrenmesini engelleyerek daha tutarlı bir temsil öğrenmesini sağlamaktadır. Üçüncü aşamada, metinler doğal dil işleme süreçlerine uygun hale getirilmiştir. Bu kapsamda raporlar cümlelere ayrılmış ve modelin giriş formatına uygun şekilde düzenlenmiştir. Buna ek olarak, modelin yalnızca serbest metin üzerinden değil, aynı zamanda görüntü tabanlı analizden elde edilen yapılandırılmış klinik bilgiler ile öğrenebilmesi amacıyla metin verileri bu bilgiler ile entegre edilmiştir. Bu doğrultuda, her bir hasta için elde edilen yapılandırılmış klinik özellikler (örneğin lezyon hacmi, anatomik bölge, vasküler sulama alanı ve ADC yorumu) metin girdisine dahil edilmiştir. Bu yaklaşım, modelin yalnızca dilsel kalıpları değil, aynı zamanda klinik bağlamı da öğrenmesini sağlamaktadır.

Metin verilerinin hazırlanması sürecinin son aşamasında, model eğitimi için uygun veri çiftleri oluşturulmuştur. Bu kapsamda her bir örnek için giriş ve hedef metinler tanımlanmıştır. Giriş verisi yapılandırılmış klinik bilgileri içerirken, hedef veri karşılık gelen radyolojik rapor metnini temsil etmektedir. Bu eşleştirme sayesinde model, verilen klinik bulgulara karşılık gelen doğal dilde rapor üretimini öğrenmektedir.

2.12. Büyük Dil Modeli Seçimi ve Mimarisi

Bu tez çalışmasının ikinci temel bileşeni, segmentasyon modellerinin çıktılarından elde edilen yapısal bulgulara dayalı olarak Türkçe radyoloji raporu üreten bir büyük dil modeli akışının geliştirilmesidir. Segmentasyon aşaması tamamlandıktan sonra lezyon maskesinden çıkarılan nicel özellikler (lezyon hacmi, yerleşim tarafı, anatomik lob ve derin yapı lokalizasyonu, vasküler alan tahmini) yapılandırılmış bir metin girdisine dönüştürülmekte ve bu girdi bir büyük dil modeline sunularak standart formatta Türkçe radyoloji raporu üretilmektedir.

İnme radyolojisi alanında Türkçe rapor üreten bir dil modeli geliştirmek için temel model seçiminde dört ana kriter belirlenmiştir. Birincisi, modelin ön eğitim verisi içinde yeterli düzeyde Türkçe metin bulunması gerekliliğidir; Türk dilinin eklemeli morfolojik yapısını, ünlü uyumunu ve radyolojik terminolojiyi doğru şekilde üretebilmesi açısından bu kriter kritik önem taşımaktadır. İkincisi, klinik ortamlarda gerçek zamanlı veya yarı gerçek zamanlı rapor üretimi hedeflendiğinden modelin mevcut donanım kısıtları altında ince ayar yapılabilmesi ve çıkarım işleminin makul sürede tamamlanabilmesidir. Üçüncüsü, modelin yapılandırılmış girdi alarak belirli bir formatta (Bulgular ve Sonuç bölümlerinden oluşan radyoloji raporu) çıktı üretebilme kapasitesine sahip olmasıdır. Dördüncüsü ise akademik araştırma ve klinik uygulama amacıyla modelin serbestçe kullanılabilir, değiştirilebilir ve dağıtılabilir bir açık kaynak lisans altında sunulmasıdır.

Tablo 2.7, 2.8 ve 2.9'da Türkçe radyoloji raporu üretimi görevi için değerlendirilen açık kaynak ve ticari büyük dil modelleri sunulmaktadır. Modeller; parametre boyutu, dil desteği, tıbbi alan optimizasyonu, lisans durumu ve bu çalışma kapsamındaki kullanılabilirlik kriterleri açısından karşılaştırılmıştır.

Deneysel değerlendirmeye alınan altı modelden üçü Qwen2.5 ailesine aittir. Qwen2.5, Alibaba Group bünyesindeki Qwen ekibi tarafından Eylül 2024'te yayımlanan açık ağırlıklı büyük dil modeli serisidir [89]. Serinin tamamı 18 trilyon token içeren, Türkçe dahil 29'dan fazla dili kapsayan büyük ölçekli bir veri kümesi üzerinde ön eğitimden geçirilmiştir. Önceki sürüm olan Qwen2'ye kıyasla ön eğitim verisinin 7 trilyon tokendan 18 trilyon tokena genişletilmesi, yapılandırılmış veri anlama ve talimat takip yeteneklerinde belirgin iyileşmeler sağlamıştır. Post-training aşamasında 1 milyonun üzerinde örnek ile denetimli ince ayar (SFT), doğrudan tercih optimizasyonu (DPO) ve grup göreli politika optimizasyonu (GRPO) uygulanmıştır. Üç Qwen2.5 modeli de aynı decoder-only Transformer mimarisini, aynı 151.646 tokenlık bayt düzeyinde BPE tokenizer'ını ve aynı ChatML diyalog şablonunu paylaşmaktadır. Mimari açıdan geleneksel çok başlıklı dikkat yerine gruplandırılmış sorgu dikkati (Grouped-Query Attention, GQA) kullanılmaktadır; bu mekanizmada birden fazla sorgu başlığı aynı anahtar-değer çiftini paylaşarak çıkarım sırasında bellek tüketimini önemli ölçüde azaltmaktadır. Konum bilgisi döner konum gömmeleri (RoPE) ile kodlanmakta, ileri beslemeli ağ katmanlarında SwiGLU aktivasyon fonksiyonu kullanılmakta ve katman normalizasyonu olarak RMSNorm tercih edilmektedir. Üç model arasındaki fark yalnızca ölçektedir: 3B modeli 36 katman ve 2.048 gizli boyut, 7B modeli 28 katman ve 3.584 gizli boyut, 14B modeli ise 48 katman ve 5.120 gizli boyut içermektedir. Bu ölçek farklılığı, aynı mimari aile içinde parametre sayısının rapor üretim kalitesine etkisinin izole edilerek analiz edilmesini mümkün kılmaktadır.

Tablo 2.7. Deneysel olarak değerlendirilen açık kaynak modeller.

Model	Geliştirici	Parametre	Eğitim Tipi	Dil Desteği	VRAM	Araştırma Sorusu
Qwen2.5-3B-Instruct [89]	Alibaba (Qwen)	3,09B	Çok dilli ön eğitim + SFT/DPO/GRPO	29+ dil, Türkçe dahil	~3 GB	Küçük ölçekli çok dilli baseline
Qwen2.5-7B-Instruct [89]	Alibaba (Qwen)	7,61B	Çok dilli ön eğitim + SFT/DPO/GRPO	29+ dil, Türkçe dahil	~5 GB	Küçük ölçekli çok dilli baseline
Qwen2.5-14B-Instruct [89]	Alibaba (Qwen)	14,7B	Çok dilli ön eğitim + SFT/DPO/GRPO	29+ dil, Türkçe dahil	~10 GB	Model boyutu etkisi (3B vs 14B)
Mistral-7B-Instruct-v0.3 [90]	Mistral AI	7,3B	Çok dilli ön eğitim + SFT	Çok dilli (İngilizce odaklı)	~5 GB	Farklı mimari aile karşılaştırması
Kumru-2B-Instruct [91]	VNGRS	2B	Türkçe sıfırdan ön eğitim + 1M örnek SFT	Yalnızca Türkçe	~1,5 GB	Türkçe-özgü model etkisi
Cosmos Turkish-Llama-8B-DPO [92]	YTÜ COSMOS	8B	LLaMA-3'ün Türkçe ince ayarı + DPO	Türkçe ince ayarlı	~5 GB	Türkçe ince ayarlı transfer öğrenme

Diğer üç model, Qwen2.5'ten farklı tasarım felsefeleri taşımaktadır. Mistral-7B-Instruct-v0.3 [90], Qwen modellerinin aksine tam bağlam dikkat yerine kayan pencere dikkati (Sliding Window Attention, SWA) mekanizması kullanmaktadır. SWA'da her token yalnızca belirli bir pencere boyutu (4.096 token) içindeki komşu tokenlere dikkat edebilmekte, bu sayede hesaplama maliyeti bağlam uzunluğu ile doğrusal olarak ölçeklenmektedir. Ayrıca Mistral'in tokenizer'ı 32.768 tokenlık bir kelime hazinesine sahip olup Qwen'in 151.646 tokenlık kelime hazinesine kıyasla önemli ölçüde küçüktür; bu fark özellikle Türkçe gibi eklemeli dillerde token başına düşen karakter sayısını ve dolayısıyla modelin dili temsil etme verimliliğini doğrudan etkilemektedir. Kumru-2B-Instruct [91] ise VNGRS tarafından 500 GB temizlenmiş ve tekilleştirilmiş Türkçe derlem üzerinde sıfırdan eğitilmiş bir modeldir. Kumru'nun en belirgin avantajı Türkçe'ye özgü olarak tasarlanmış 50.176 tokenlık BPE tokenizer'ıdır: diğer çok dilli modeller aynı Türkçe metni kodlamak için Kumru'ya kıyasla %38 ila %98 daha fazla token harcamakta, bu da hem bağlam penceresinin daha verimli kullanılması hem de çıkarım hızının artması anlamına gelmektedir. Cosmos Turkish-Llama-8B-DPO [92] ise transfer öğrenme yaklaşımını temsil etmektedir: Meta'nın LLaMA-3 8B temel modeli üzerine Yıldız Teknik Üniversitesi COSMOS grubu tarafından 30 GB Türkçe veri ile ince ayar yapılmış ve ardından DPO uygulanarak diyalog kalitesi iyileştirilmiştir. Bu model, LLaMA-3'ün güçlü İngilizce temsil kapasitesini Türkçe alan bilgisiyle birleştirmeyi hedeflemektedir.

Bu altı modelin tokenizer yapıları ve Türkçe metin temsil verimlilikleri arasındaki farklar, ince ayar ve çıkarım performansını doğrudan etkileyen önemli bir değişkendir. Tablo 2.10'da modellerin tokenizer özellikleri karşılaştırılmaktadır. Qwen2.5 modelleri 151.646 tokenlık geniş kelime hazinesiyle çok dilli kapsamı artırırken, Kumru'nun Türkçe'ye özgü 50.176 tokenlık kompakt tokenizer'ı aynı Türkçe metni çok daha az tokenle kodlayabilmektedir. Mistral'in 32.768 tokenlık kelime hazinesi en küçük olanıdır ve eklemeli dillerde daha fazla parçalanmaya neden olmaktadır. Cosmos ise LLaMA-3'ün 128.256 tokenlık tokenizer'ını miras almıştır. Bu farklılıklar, aynı radyoloji rapor metninin farklı modellerde farklı uzunluklarda token dizilerine dönüşmesi anlamına gelmekte olup modellerin bağlam penceresi kullanım verimliliğini ve eğitim hızını doğrudan etkilemektedir.

Tablo 2.8. Değerlendirme dışı bırakılan açık kaynak modeller.

Model	Geliştirici	Parametre	Dil Desteği	Tıbbi Optimizasyon	Dışlanma Gerekçesi
MedGemma-4B [93]	Google	4B	Çok dilli, Türkçe sınırlı	Tıbbi veri ile ön eğitim (radyoloji, histopatoloji, oftalmoloji, dermatoloji)	Multimodal (görüntü+metin) odaklı; bu çalışmada LLM'e görüntü değil yapılandırılmış metin verilmektedir
MedGemma-27B [93]	Google	27B	Çok dilli, Türkçe sınırlı	Tıbbi metin optimizasyonlu	Donanım kısıtı (27B); Türkçe radyoloji eğitim verisi bulunmamaktadır

Ticari modeller (Tablo 2.9) arasında GPT-4o, radyoloji rapor özetleme ve yapılandırma görevlerinde en yüksek performansı sergilemekte olup Eurorad vaka çalışmasında %79,6 tanısal doğruluk elde etmiştir. Ancak ticari modeller bu çalışmada üç temel nedenle kullanılmamıştır: birincisi, kapalı kaynak yapıları nedeniyle model ağırlıklarına erişim mümkün olmayıp alan-spesifik ince ayar yapılamamaktadır; ikincisi, klinik hasta verilerinin API üzerinden üçüncü taraf sunuculara gönderilmesi veri gizliliği ve etik kurul düzenlemeleri açısından sorun oluşturmaktadır; üçüncüsü, model versiyonlarının zaman içinde değişebilmesi deneysel tekrarlanabilirliği tehlikeye atmaktadır.

Tablo 2.9. Ticari (kapalı kaynak) referans modeller.

Model	Geliştirici	Dil Desteği	Tıbbi Performans	Erişim	Kullanılmama Gerekçesi
GPT-4o [94]	OpenAI	Güçlü çok dilli, Türkçe dahil	Radyoloji rapor yapılandırmada SOTA; Eurorad'da %79,6 doğruluk	Ücretli API	Kapalı kaynak; ince ayar yapılamaz, veri gizliliği; tekrarlanabilirlik sorunu
GPT-4o-mini [94]	OpenAI	Çok dilli	GPT-4o'ya yakın performans	Ücretli API	Kapalı kaynak; ince ayar yapılamaz, veri gizliliği; tekrarlanabilirlik sorunu
Claude 3.5 Sonnet [95]	Anthropic	Çok dilli, Türkçe dahil	Radyoloji görevlerinde rekabetçi	Ücretli API	Kapalı kaynak; ince ayar yapılamaz, veri gizliliği
Gemini 1.5 Pro [96]	Google	Güçlü çok dilli	Tıbbi çıkarım kapasitesi yüksek	Ücretli API	Kapalı kaynak; model davranışı değişkenlik gösterebilir

Açık kaynak modeller arasında MedGemma, Google tarafından tıbbi veri ile önceden eğitilmiş özelleştirilmiş bir model ailesi olarak dikkat çekmektedir. MedGemma-4B varyantı tıbbi görüntü yorumlama, radyoloji ve histopatoloji alanlarında güçlü performans sunmakla birlikte bu çalışmada LLM'e ham görüntü değil segmentasyon çıktısından elde edilen yapılandırılmış metin verilmekte olup multimodal mimarisi bu kullanım senaryosu için gereksiz karmaşıklık getirmektedir. Ayrıca MedGemma'nın Türkçe radyoloji terminolojisine yönelik eğitim verisi bulunmamaktadır.

Bütün bunlar değerlendirildiğinde, altı model deneysel değerlendirmeye alınmıştır. Qwen2.5 ailesinden 3B, 7B ve 14B Instruct modelleri, aynı mimari aile içinde model boyutu etkisinin izole edilerek incelenmesine olanak tanımaktadır. Mistral-7B-Instruct-v0.3, sliding window attention (SWA) mekanizması ve farklı tokenizer yapısıyla alternatif birçok dilli mimari perspektifi sunmaktadır. Kumru-2B-Instruct, VNGRS tarafından 500 GB Türkçe veri üzerinde sıfırdan eğitilmiş olup Türkçe'ye özgü BPE tokenizer'ı sayesinde diğer modellere kıyasla %38–98 daha az token kullanma avantajına sahiptir; Cetvel değerlendirme ölçütünde LLaMA-3.3-70B ve Gemma-3-27B gibi çok daha büyük modelleri geride bıraktığı rapor edilmiştir (VNGRS, 2024). Cosmos Turkish-Llama-8B-DPO ise Yıldız Teknik Üniversitesi COSMOS araştırma grubu tarafından Meta'nın LLaMA-3 8B modeli üzerine 30 GB Türkçe veri ile ince ayar yapılarak ve ardından DPO (Direct Preference Optimization) uygulanarak geliştirilmiştir. Bu altılı karşılaştırma, tezin deneysel bölümünde dört farklı araştırma sorusunun cevaplanmasına imkân tanımaktadır: model boyutu etkisi (Qwen 3B, 7B ve 14B), mimari aile etkisi (Qwen vs Mistral), Türkçe-özgü ön eğitimin etkisi (Kumru vs çok dilli modeller) ve Türkçe ince ayarlı transfer öğrenmenin etkisi (Cosmos vs genel çok dilli modeller).

2.13. LLM Eğitimi ve Adaptasyonu

Bu bölümde, Bölüm 2.12'de tanıtılan altı büyük dil modelinin Türkçe radyoloji rapor üretimi görevine uyarlanması için uygulanan ince ayar stratejisi, eğitim veri setinin yapısı, kullanılan parametre-verimli eğitim teknikleri ve eğitim hiperparametreleri ayrıntılı olarak açıklanmaktadır.

Büyük dil modellerinin alan-spesifik görevlere uyarlanmasında iki temel yaklaşım mevcuttur: tam ince ayar (full fine-tuning) ve parametre-verimli ince ayar (parameter-efficient fine-tuning, PEFT). Tam ince ayarda modelin tüm parametreleri güncellenmekte olup bu yaklaşım yüksek GPU belleği ve hesaplama kaynağı gerektirmektedir; örneğin 7B parametreli bir modelin tam ince ayarı yaklaşık 56 GB, 14B parametreli bir modelin tam ince ayarı ise yaklaşık 120 GB VRAM gerektirmektedir. Mevcut donanım kısıtları ve sınırlı eğitim verisi boyutu (141 örnek) göz önünde bulundurulduğunda, tam ince ayar yaklaşımı hem hesaplama hem de aşırı öğrenme riski açısından uygun değildir. Bu nedenle, çalışmada tüm modellere parametre-verimli ince ayar tekniklerinden QLoRA (Quantized Low-Rank Adaptation) yöntemi uygulanmıştır. Şekil 2.11'de, büyük dil modellerinin düşük maliyetli ve parametre-verimli biçimde ince ayarlanmasını sağlayan QLoRA mimarisinin temel bileşenleri ve çalışma yapısı gösterilmektedir.



Şekil 2.11. QLoRA mimarisi.

QLoRA, Dettmers ve arkadaşları [97] tarafından önerilen ve LoRA (Low-Rank Adaptation) ile nicemleme tekniklerini birleştiren bir parametre-verimli ince ayar yöntemidir. Bu yöntemde, modelin orijinal ağırlık matrisleri dondurulmuş olarak tutulmakta ve eğitim sırasında yalnızca düşük rütbeli(rank) adaptör matrisleri güncellenmektedir. Böylece eğitilmesi gereken parametre sayısı, toplam model parametrelerinin küçük bir yüzdesiyle sınırlı kalmaktadır. LoRA yönteminin temel prensibi, önceden eğitilmiş bir ağırlık matrisi $W_0 \in \mathbb{R}^{(d\times k)}$ için güncellemenin düşük rütbeli bir ayrıştırma ile yaklaştırılmasıdır. Buna göre, ince ayar sonrası elde edilen ağırlık matrisi aşağıdaki şekilde ifade edilmektedir:

$$W = W_0 + \Delta W = W_0 + B \times A\,(X) \qquad (23)$$

Burada B ∈ ℝ^(d×r) ve A ∈ ℝ^(r×k) düşük rütbeli adaptör matrisleridir; r ise rütbe hiperparametresidir ve r < min(d, k) koşulunu sağlamaktadır. Eğitim sürecinde yalnızca A ve B matrisleri güncellenmekte, orijinal ağırlık matrisi W_0 sabit tutulmaktadır. Bu yaklaşım sayesinde eğitilmesi gereken parametre sayısı d×k yerine r×(d+k) seviyesine düşürülmektedir.

QLoRA, bu LoRA yapısının üzerine iki ek yenilik getirmektedir. Birincisi, dondurulmuş model ağırlıkları 4-bit NormalFloat (NF4) nicemleme ile sıkıştırılarak bellekte kapladığı alan önemli ölçüde azaltılmaktadır. NF4, normal dağılıma sahip ağırlıklar için bilgi-kuramsal açıdan optimal bir veri tipi olup, standart 4-bit tamsayı nicemlemeye kıyasla daha düşük nicemleme hatası üretmektedir. İkincisi, çift nicemleme (double quantization) tekniği ile nicemleme sabitlerinin kendileri de nicelenerek ek bellek tasarrufu sağlanmaktadır. Böylece, örneğin 14B parametreli bir modelin ince ayarı yaklaşık 10 GB VRAM ile gerçekleştirilebilmekte; bu da mevcut donanım kısıtları altında büyük modellerin eğitimini mümkün kılmaktadır. QLoRA yönteminde ileri besleme sırasında dondurulmuş ağırlıklar 4-bit depolamadan BFloat16 hassasiyetine geri dönüştürülmekte, adaptör ağırlıkları ile toplanmakta ve çıkış hesaplanmaktadır. Geri yayılım sırasında ise gradyanlar yalnızca adaptör matrislerine aktarılmakta, dondurulmuş ağırlıklar güncellenmemektedir. Bu mekanizma sayesinde model, orijinal bilgi birikimini korurken görev-spesifik adaptasyonu yalnızca küçük bir parametre alt kümesi üzerinden gerçekleştirmektedir. Şekil 2.12'de, segmentasyon maskeleri, anatomik atlas bilgileri ve radyoloji raporlarının birleştirilmesiyle oluşturulan eğitim veri seti hazırlama süreci ayrıntılı olarak gösterilmektedir.



Şekil 2.12. Eğitim veri seti oluşturma süreci.

Başlangıçta 150 hastadan oluşan veri setinden segmentasyon–rapor tutarsızlığı gösteren 9 örnek çıkarılarak 141 hastalık nihai veri seti elde edilmiştir. Bu veri seti, sabit bir rastgele tohum (seed=42) kullanılarak 120 eğitim ve 21 doğrulama örneğine ayrılmıştır. Her bir eğitim örneği, giriş (input) ve hedef (output) olmak üzere iki bileşenden oluşmaktadır. Giriş bileşeni, segmentasyon ve anatomik eşleme aşamalarından elde edilen yapılandırılmış klinik bilgileri içeren bir metin şablonudur. Bu şablon; lezyon sayısı, lezyon büyüklüğü kategorisi, yerleşim tarafı, kortikal bölge, subkortikal bölge, vasküler sulama alanı ve ADC yorumu gibi alanları kapsamaktadır. Hedef bileşeni ise ilgili hastaya ait radyolojik rapor metnidir. Örnek bir eğitim çifti aşağıda gösterilmektedir:

Giriş: "Lezyon sayısı: 1, Lezyon büyüklüğü: Küçük-orta boyutlu difüzyon kısıtlanması, Yerleşim Tarafı: Sol hemisfer, Kortikal bölge: Superior frontal girus, Subkortikal bölge: Sol serebral beyaz cevher, Vasküler sulama alanı: MCA, ADC yorumu: Düşük ADC; akut iskemik difüzyon kısıtlanması ile uyumlu"

Hedef: "Sol frontal lobda MCA sulama alanında akut enfarkt ile uyumlu difüzyon kısıtlanması izlenmektedir."

Eğitim sırasında her bir örnek, modelin diyalog şablonuna uygun biçimde yapılandırılmıştır. Qwen modelleri için ChatML formatı, Mistral için Mistral Instruct şablonu, Kumru için uygun diyalog formatı ve Cosmos için LLaMA-3 şablonu kullanılmıştır. Tüm modellere aynı sistem istemi (system prompt) verilmiş olup bu istem, modelin görevini tanımlayan ve çıktı formatını belirleyen yönergeler içermektedir. Sistem isteminde modelin yalnızca verilen yapılandırılmış bilgiler üzerinden rapor üretmesi, ek bilgi uydurmaması (halüsinasyon yapmaması) ve raporun "Bulgular" ile "Sonuç" bölümlerinden oluşması gerektiği belirtilmiştir. Tüm modellerde aynı QLoRA hiperparametreleri kullanılarak adil bir karşılaştırma ortamı sağlanmıştır. Kullanılan hiperparametreler Tablo 2.10' da özetlenmektedir.

Tablo 2.10. QLoRA hiperparametre detayları.

Hiperparametre	Değer	Açıklama
LoRA rütbesi (r)	16	Adaptör matrislerinin boyutu
LoRA alfa (α)	32	Ölçekleme katsayısı (α/r = 2)
LoRA dropout	0,05	Adaptör katmanlarına uygulanan dropout oranı
Hedef modüller	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj	Adaptör eklenen Transformer katmanları
Nicemleme tipi	NF4 (4-bit NormalFloat)	Dondurulmuş ağırlıkların nicemleme formatı
Hesaplama hassasiyeti	BFloat16	İleri ve geri beslemedeki kayan nokta hassasiyeti
Çift nicemleme	Evet	Nicemleme sabitlerinin de nicelenmesi

LoRA rütbesi (r=16), adaptör matrislerinin ifade gücü ile hesaplama maliyeti arasında denge kurmak amacıyla seçilmiştir. Düşük rütbe değerleri (r=4 veya r=8), modelin görev-

spesifik bilgiyi yeterince temsil edememe riskini taşırken; çok yüksek değerler (r=64 veya r=128), sınırlı eğitim verisinde aşırı öğrenmeye yol açabilmektedir. LoRA alfa değeri (α=32), adaptör çıkışlarının ölçeklenme oranını belirlemekte olup $\alpha/r = 2$ oranı adaptör güncellemelerinin model davranışı üzerinde yeterli etkiyi göstermesini sağlamaktadır. Hedef modüller olarak Transformer bloklarındaki dikkat mekanizması katmanlarının tamamı (sorgu, anahtar, değer ve çıkış projeksiyonları) ile ileri beslemeli ağ katmanları (gate, up ve down projeksiyonları) seçilmiştir. Bu geniş kapsam, modelin hem dikkat kalıplarını hem de özellik dönüşümlerini görev-spesifik olarak uyarlayabilmesini sağlamaktadır.

Model eğitiminde kullanılan genel eğitim hiperparametreleri Tablo 2.11'de sunulmaktadır. Bu hiperparametreler tüm modellerde sabit tutularak, performans farklılıklarının yalnızca model mimarisi ve ön eğitim bilgisinden kaynaklanması sağlanmıştır.

Tablo 2.11. QLoRA eğitim hiperparametre detayları.

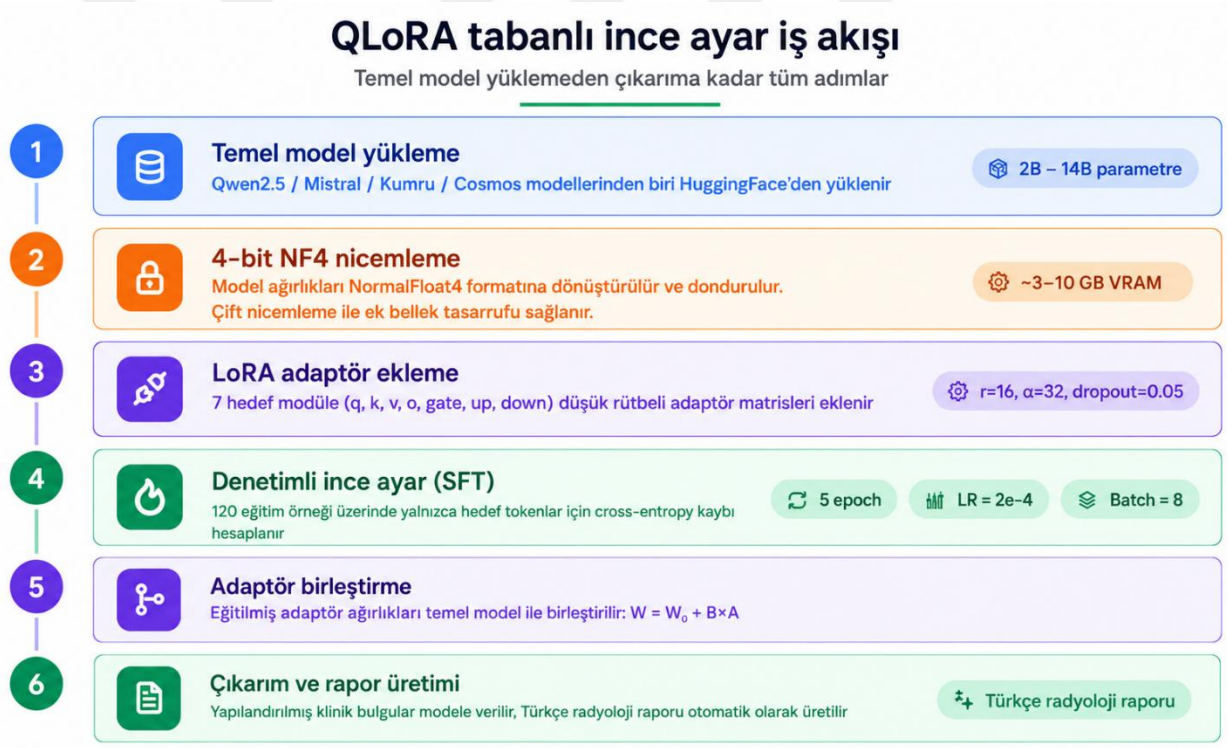
Hiperparametre	Değer	Açıklama
Epoch sayısı	5	Toplam eğitim döngüsü
Batch boyutu	8	GPU' da eş zamanlı işlenen örnek sayısı
Öğrenme oranı	2×10^{-4}	AdamW optimizasyonu için başlangıç öğrenme oranı
Öğrenme oranı zamanlayıcısı	Kosinüs	Eğitim boyunca öğrenme oranının kademeli azaltılması
Maksimum dizi uzunluğu	512	Giriş ve çıkış tokenlarının toplam maksimum uzunluğu
Optimizasyon algoritması	AdamW (paged, 8-bit)	Bellek verimli adaptif optimizasyon

Epoch sayısı 5 olarak belirlenmiştir. Bu değer, 120 eğitim örneğinden oluşan küçük bir veri seti üzerinde modelin yeterli düzeyde öğrenme gerçekleştirmesini sağlarken, aşırı öğrenme riskini kontrol altında tutmak amacıyla sınırlı tutulmuştur. Eğitim süreci boyunca doğrulama kaybı (validation loss) izlenerek modelin genelleme performansı takip edilmiştir. Efektif batch boyutu, GPU başına 2 örnek ve 4 gradyan birikim adımı ile toplam 8 olarak hesaplanmıştır. Gradyan birikim yöntemi, sınırlı GPU belleği nedeniyle büyük batch boyutlarının doğrudan kullanılamaması durumunda, birden fazla mini-batch üzerinden gradyanların biriktirilmesi ve ardından tek bir güncelleme adımı yapılması prensibine dayanmaktadır. Bu yaklaşım, daha kararlı gradyan tahminleri sağlayarak eğitim sürecinin yakınsamasını iyileştirmektedir.

Öğrenme oranı olarak 2×10^{-4} değeri tercih edilmiştir. Bu değer, QLoRA tabanlı ince ayar çalışmalarında yaygın olarak kullanılan bir aralıkta yer almakta olup, adaptör parametrelerinin yeterli hızda güncellenmesini sağlarken eğitim kararsızlığını önleyecek düzeyde düşük tutulmuştur. Kosinüs zamanlayıcı ile eğitimin ilk aşamalarında hızlı öğrenme gerçekleştirilirken, son aşamalarda öğrenme oranı kademeli olarak sıfıra yaklaştırılarak ince ayarın daha hassas bir şekilde tamamlanması hedeflenmiştir. Paged AdamW (8-bit) optimizasyon algoritması, standart AdamW'nin bellek verimli bir

varyantıdır. Bu yaklaşımda, optimizasyon durumları (birinci ve ikinci moment tahminleri) 8-bit hassasiyette saklanmakta ve gerekli olduğunda CPU belleğine taşınarak GPU bellek kullanımı azaltılmaktadır. Bu özellik, özellikle 14B gibi büyük modellerin 12 GB VRAM ile eğitilmesini mümkün kılan kritik bir bileşendir.

Altı modelin her biri, aynı eğitim ve doğrulama bölmesi (120/21), aynı QLoRA konfigürasyonu ve aynı eğitim hiperparametreleri kullanılarak ince ayardan geçirilmiştir. Eğitim sürecinde Hugging Face Transformers kütüphanesi, PEFT (Parameter-Efficient Fine-Tuning) kütüphanesi ve BitsAndBytes nicemleme kütüphanesi birlikte kullanılmıştır. Denetimli ince ayar (Supervised Fine-Tuning, SFT) yöntemi uygulanmış olup eğitim sırasında kayıp fonksiyonu olarak standart dil modelleme kaybı (cross-entropy loss) kullanılmıştır. Bu kayıp fonksiyonunda yalnızca hedef (rapor) tokenlarına ait kayıp hesaplanmış, giriş (yapılandırılmış klinik bilgi) tokenları kayıp hesaplamasından maskelenmiştir. Bu yaklaşım, modelin yalnızca rapor üretim becerisini öğrenmesini sağlayarak, giriş metnini ezberleme eğilimini azaltmaktadır. Şekil 2.13'te, temel modelin yüklenmesinden başlayarak QLoRA tabanlı ince ayar, adaptör eğitimi ve otomatik radyoloji raporu üretimine kadar gerçekleştirilen tüm işlem adımları gösterilmektedir.



Şekil 2.13. QLoRA tabanlı ince ayar iş akışı.

Eğitim tamamlandıktan sonra, her bir model için elde edilen en iyi adaptör ağırlıkları kaydedilmiştir. Çıkarım (inference) aşamasında, dondurulmuş temel model ağırlıkları ile eğitilmiş adaptör matrisleri birleştirilerek nihai model oluşturulmuştur. Çıkarım sırasında modele yapılandırılmış klinik bilgiler giriş olarak verilmiş ve model tarafından üretilen radyolojik raporlar Bölüm 2.14'te açıklanan değerlendirme metrikleri ile analiz edilmiştir.

2.14. Otomatik Raporların Değerlendirilmesi ve Klinik Uyum Analizi

Bölüm 2.13'te açıklanan ince ayar sürecinden geçirilen altı büyük dil modelinin ürettiği Türkçe radyoloji raporlarının değerlendirilmesi için kullanılan metrikler, klinik uyum analiz yöntemleri ve uzman değerlendirme protokolü ayrıntılı olarak açıklanmaktadır. Değerlendirme süreci üç temel katmandan oluşmaktadır: (i) otomatik doğal dil işleme metrikleri, (ii) klinik içerik doğruluk metrikleri ve (iii) uzman radyolog değerlendirmesi. Bu çok katmanlı değerlendirme yaklaşımı, üretilen raporların yalnızca dilsel kalitesinin değil, aynı zamanda klinik doğruluğunun ve kullanılabilirliğinin de kapsamlı bir şekilde analiz edilmesini amaçlamaktadır.

Otomatik rapor üretim sistemlerinin performansının nicel olarak değerlendirilebilmesi amacıyla, doğal dil üretimi literatüründe yaygın olarak kullanılan standart metrikler uygulanmıştır. Bu metrikler, model tarafından üretilen raporları uzman radyolog tarafından yazılmış referans raporlarla karşılaştırarak dilsel benzerliği ölçmektedir. Her bir metrik, üretilen metnin farklı bir boyutunu değerlendirmektedir.

BLEU (Bilingual Evaluation Understudy), doğal dil üretimi görevlerinde üretilen metinlerin referans metinlerle benzerliğini ölçmek amacıyla kullanılan en yaygın otomatik değerlendirme metriklerinden biridir. İlk olarak Papineni ve arkadaşları [98] tarafından makine çevirisi sistemlerini değerlendirmek amacıyla önerilmiş olsa da günümüzde otomatik özetleme, görüntü açıklama üretimi, konuşma üretimi ve radyolojik rapor oluşturma gibi birçok doğal dil üretimi probleminde yaygın biçimde kullanılmaktadır. Bu metrik, model tarafından üretilen metindeki n-gram yapılarının referans metinlerde bulunma oranını ölçerek dilsel benzerliği nicel olarak değerlendirmektedir.

BLEU metriği, üretilen metin ile referans metin arasındaki unigram, bigram, trigram ve four-gram düzeyindeki örtüşmeleri dikkate almaktadır. Bu yaklaşım sayesinde yalnızca tek tek kelimelerin eşleşmesi değil, aynı zamanda kelime sıralaması ve yerel bağlam tutarlılığı da değerlendirmeye dahil edilmektedir. BLEU skorunun yüksek olması, model tarafından üretilen metnin referans rapora hem sözcüksel hem de yapısal açıdan daha yakın olduğunu göstermektedir. BLEU metriğinin genel denklemi aşağıdaki (24) gibidir:

$$BLEU = BP\ x\ \exp\left(\sum_{n=1}^{N} w_n log p_n\right) \tag{24}$$

Denklem 24' te BP, kısalık cezası (brevity penalty), w_n, her n-gram düzeyi için ağırlık katsayısı, p_n: güncellenmiş n-gram kesinlik değeri, N kullanılan maksimum n-gram seviyesi olarak tanımlanmaktadır. BLEU hesaplamasında kullanılan güncellenmiş n-gram kesinlik değeri aşağıdaki şekilde ifade edilmektedir:

$$p_n = \frac{\sum_{\text{ngram}_n} \text{Count}_{clip}\ (\text{ngram}_n)}{\sum_{\text{ngram}_n} \text{Count}\ (\text{ngram}_n)} \tag{25}$$

Bu denklemde, Count_{clip} değeri aday metindeki n-gram tekrarlarının referans metindeki maksimum tekrar sayısı ile sınırlandırılmış halini ifade etmektedir. Böylece modelin aynı kelime veya ifadeyi aşırı tekrar ederek yapay biçimde yüksek skor elde etmesi engellenmektedir. BLEU metriğinde ayrıca kısa metin üretimini cezalandırmak amacıyla kısalık cezası terimi kullanılmaktadır. Bu ceza mekanizması aşağıdaki şekilde tanımlanmaktadır:

$$BP = \begin{cases} 1, & c > r \\ e^{(1-r/c)}, & c \leq r \end{cases}$$

(26)

Burada c, üretilen metnin uzunluğu, r, referans metnin uzunluğu olarak tanımlanmaktadır. Eğer model referans metne kıyasla çok kısa bir çıktı üretirse, BLEU skoru ceza uygulanarak düşürülmektedir.

Ancak BLEU metriğinin bazı sınırlılıkları da bulunmaktadır. BLEU yalnızca sözcüksel örtüşmeye dayalı bir değerlendirme yaptığı için, aynı klinik anlamı taşıyan ancak farklı sözcüklerle ifade edilen cümleleri tam olarak değerlendiremeyebilmektedir. Özellikle tıbbi raporlama gibi semantik doğruluğun kritik olduğu alanlarda, BLEU skorunun tek başına yeterli olmadığı ve ROUGE, METEOR [99], BERTScore [100] veya klinik uzman değerlendirmeleri gibi ek ölçütlerle birlikte kullanılması gerektiği belirtilmektedir.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation), doğal dil üretimi ve özellikle otomatik özetleme sistemlerinin değerlendirilmesinde yaygın olarak kullanılan otomatik bir metin değerlendirme metriğidir. Lin tarafından önerilen bu metrik, model tarafından üretilen metin ile referans metin arasındaki örtüşmeyi ölçerek metinsel benzerliği nicel olarak değerlendirmektedir [101]. Günümüzde ROUGE metriği yalnızca özetleme problemlerinde değil, aynı zamanda görüntü açıklama üretimi, sohbet sistemleri ve otomatik radyolojik rapor oluşturma gibi doğal dil üretim görevlerinde de yaygın olarak kullanılmaktadır.

ROUGE metriği, BLEU'dan farklı olarak kesinlik odaklı değil, duyarlılık odaklı bir değerlendirme yaklaşımı benimsemektedir. Başka bir ifadeyle ROUGE, referans metindeki bilginin ne kadarının model tarafından üretilen metinde yakalanabildiğini ölçmektedir. Bu özellik özellikle tıbbi raporlama gibi kritik bilgi kaybının önemli olduğu uygulamalarda büyük önem taşımaktadır. ROUGE ailesi farklı varyantlardan oluşmaktadır. Bu çalışmada en yaygın kullanılan ROUGE-1, ROUGE-2 ve ROUGE-L skorları değerlendirme amacıyla kullanılmıştır. ROUGE-N metriği aşağıdaki şekilde tanımlanmaktadır:

$$\text{ROUGE-N} = \frac{\sum_{\text{ngram}_n \in Ref} \text{Count}_{match}\,(\text{ngram}_n)}{\sum_{\text{ngram}_n \in Ref} \text{Count}\,(\text{ngram}_n)}$$

(27)

Denklem 27'de Count_{match}, üretilen metin ile referans metin arasında eşleşen n-gram sayısını, Count, referans metindeki toplam n-gram sayısını ifade etmektedir. ROUGE-1 unigram düzeyindeki örtüşmeyi ölçerken, ROUGE-2 bigram düzeyindeki kelime dizisi benzerliğini değerlendirmektedir. Daha yüksek ROUGE skorları, modelin referans

rapordaki klinik ifadeleri ve terminolojik yapıları daha başarılı biçimde yakaladığını göstermektedir.

ROUGE-L metriği ise en uzun ortak alt diziye dayanmaktadır. Bu yaklaşım yalnızca kelime eşleşmelerini değil, aynı zamanda cümle yapısı ve sıralama tutarlılığını da değerlendirmektedir. ROUGE-L hesaplamasında kullanılan duyarlılık değeri aşağıdaki şekilde ifade edilmektedir:

$$R_{LCS} = \frac{LCS(X,Y)}{m} \tag{28}$$

Burada $LCS(X,Y)$ üretilen metin ile referans metin arasındaki en uzun ortak alt diziyi, m referans metindeki toplam kelime sayısını ifade etmektedir.

$$P_{LCS} = \frac{LCS(X,Y)}{n} \tag{29}$$

Kesinlik değeri Denklem 29'da tanımlanmaktadır. Burada n, üretilen metindeki toplam kelime sayısını göstermektedir. ROUGE-L için F-skor değeri ise aşağıdaki şekilde hesaplanmaktadır:

$$F_{LCS} = \frac{(1+\beta^2)R_{LCS}P_{LCS}}{R_{LCS} + \beta^2 P_{LCS}} \tag{30}$$

Bu tez kapsamında ROUGE metriği, otomatik olarak üretilen radyolojik raporların uzman radyologlar tarafından yazılmış referans raporlarla benzerliğini değerlendirmek amacıyla kullanılmıştır. Özellikle ROUGE-1 ve ROUGE-2 skorları terminolojik örtüşmeyi ölçerken, ROUGE-L metriği raporların yapısal ve anlamsal tutarlılığı hakkında ek bilgi sağlamaktadır. Böylece yalnızca kelime düzeyindeki benzerlik değil, aynı zamanda klinik ifade akışı ve rapor yapısının korunma düzeyi de değerlendirilebilmektedir.

ROUGE metriği tıbbi doğal dil üretimi çalışmalarında önemli avantajlar sağlamasına rağmen bazı sınırlılıklara da sahiptir. Özellikle eş anlamlı ifadeler veya farklı sözdizimsel yapılar aynı klinik anlamı taşımasına rağmen düşük ROUGE skorları üretebilmektedir. Bu nedenle ROUGE skorlarının BLEU [98], METEOR [99], BERTScore [100] ve uzman klinik değerlendirmeleri ile yorumlanması önerilmektedir.

METEOR (Metric for Evaluation of Translation with Explicit ORdering), doğal dil üretimi sistemlerinin değerlendirilmesinde kullanılan otomatik bir metin benzerlik metriğidir. İlk olarak Banerjee ve Lavie [99] tarafından makine çevirisi sistemlerini değerlendirmek amacıyla geliştirilmiş olup, günümüzde otomatik özetleme, görüntü açıklama üretimi, diyalog sistemleri ve otomatik radyolojik rapor oluşturma gibi birçok doğal dil üretimi görevinde yaygın biçimde kullanılmaktadır. METEOR metriği, BLEU ve ROUGE gibi klasik n-gram tabanlı metriklere kıyasla daha esnek bir değerlendirme yaklaşımı

sunmaktadır. Özellikle yalnızca birebir kelime eşleşmelerini değil; kök benzerliklerini (stemming), eş anlamlı kelimeleri (synonym matching) ve kelime sıralamasını da dikkate alabilmesi nedeniyle semantik benzerliği daha başarılı biçimde değerlendirebilmektedir.

METEOR metriği temel olarak kesinlik ve duyarlılık değerlerini birlikte kullanmaktadır. BLEU metriğinin ağırlıklı olarak kesinlik odaklı olmasına karşın, METEOR her iki ölçütü dengeli biçimde değerlendirerek üretilen metnin referans metni ne ölçüde kapsadığını ve ne kadar doğru ifade ettiğini birlikte analiz etmektedir. Bu özellik özellikle tıbbi raporlama gibi anlam bütünlüğünün kritik olduğu alanlarda önemli avantaj sağlamaktadır. METEOR metriğinde kesinlik değeri aşağıdaki şekilde hesaplanmaktadır:

$$P = \frac{m}{w_t} \tag{31}$$

Denklem 31'de m eşleşen unigram sayısını, w_t üretilen metindeki toplam unigram sayısını ifade etmektedir. Duyarlılık değeri ise aşağıdaki şekilde tanımlanmaktadır:

$$R = \frac{m}{w_r} \tag{32}$$

Denklem 32'de w_r referans metindeki toplam unigram sayısını göstermektedir. METEOR metriği kesinlik ve duyarlılık değerlerini harmonik ortalama ile birleştirmektedir. Kullanılan ağırlıklı F-mean denklemi aşağıdaki gibidir:

$$F_{mean} = \frac{10PR}{R + 9P} \tag{33}$$

Denklem 33'te duyarlılık değerine kesinliğe göre daha yüksek ağırlık verilmektedir. Bunun nedeni, özellikle doğal dil üretimi görevlerinde referans metindeki önemli bilgilerin eksiksiz biçimde yakalanmasının kritik kabul edilmesidir. METEOR metriğinde ayrıca kelime sıralamasını değerlendirmek amacıyla bir ceza terimi kullanılmaktadır. Ceza değeri aşağıdaki şekilde hesaplanmaktadır:

$$Penalty = 0.5\left(\frac{c}{m}\right)^3 \tag{34}$$

Denklem 34'te c eşleşen kelime bloklarının sayısını, m toplam eşleşen unigram sayısını ifade etmektedir. Son METEOR skoru ise aşağıdaki şekilde hesaplanmaktadır:

$$METEOR = (1 - Penalty) \times F_{mean} \tag{35}$$

Bu yapı sayesinde METEOR yalnızca kelime örtüşmesini değil, aynı zamanda kelime sıralaması ve anlamsal tutarlılığı da değerlendirmeye dahil etmektedir. Özellikle eş anlamlı ifadelerin bulunduğu durumlarda BLEU ve ROUGE metriklerine göre daha güçlü bir değerlendirme sağlayabilmektedir.

Bu tez kapsamında METEOR metriği, otomatik olarak üretilen radyolojik raporların uzman radyologlar tarafından hazırlanan referans raporlarla anlamsal benzerliğini değerlendirmek amacıyla kullanılmıştır. METEOR'un stemming ve synonym matching mekanizmaları sayesinde, aynı klinik anlamı taşıyan ancak farklı kelimelerle ifade edilen cümleler daha doğru biçimde değerlendirilebilmektedir. Bu özellik özellikle tıbbi terminolojide farklı ifade biçimlerinin sık görülmesi nedeniyle önem taşımaktadır.

Bununla birlikte METEOR metriğinin de bazı sınırlılıkları bulunmaktadır. Her ne kadar anlamsal benzerliği klasik n-gram tabanlı metriklere göre daha başarılı değerlendirse de derin bağlamsal anlam ilişkilerini tam olarak yakalayamamaktadır. Bu nedenle METEOR skorlarının BLEU, ROUGE, BERTScore ve uzman klinik değerlendirmeleri ile yorumlanması önerilmektedir.

BERTScore, doğal dil üretimi sistemlerinin değerlendirilmesinde kullanılan, bağlamsal anlamsal benzerliği ölçmeye yönelik gelişmiş bir otomatik değerlendirme metriğidir. Zhang ve arkadaşları tarafından önerilen bu yöntem, klasik n-gram tabanlı değerlendirme metriklerinden farklı olarak Transformer tabanlı dil modellerinin ürettiği bağlamsal kelime gömme (contextual embedding) temsillerini kullanmaktadır. Böylece yalnızca birebir kelime eşleşmeleri değil, aynı zamanda anlamsal yakınlık da değerlendirilebilmektedir.

BLEU, ROUGE ve METEOR gibi geleneksel metrikler çoğunlukla yüzeysel kelime örtüşmelerine dayandığından, aynı anlamı taşıyan fakat farklı sözcüklerle ifade edilen cümlelerde düşük skor üretebilmektedir. Özellikle tıbbi raporlama gibi semantik doğruluğun kritik olduğu alanlarda bu durum önemli bir sınırlılık oluşturmaktadır. BERTScore ise önceden eğitilmiş BERT tabanlı dil modellerinden elde edilen bağlamsal embedding'leri kullanarak, kelimeler arasındaki anlamsal benzerliği ölçmektedir. Bu nedenle farklı sözcüklerle ifade edilen ancak aynı klinik anlamı taşıyan cümleleri daha doğru biçimde değerlendirebilmektedir.

BERTScore metriğinde öncelikle hem referans metindeki hem de üretilen metindeki her kelime, BERT tabanlı bir model aracılığıyla yüksek boyutlu vektör temsillerine dönüştürülmektedir. Daha sonra aday metindeki her kelime için referans metindeki en benzer kelime belirlenmekte ve kosinüs benzerliği hesaplanmaktadır.

İki embedding vektörü arasındaki kosinüs benzerliği aşağıdaki şekilde tanımlanmaktadır:

$$\mathrm{sim}(x_i, y_j) = \frac{x_i^T y_j}{\| x_i \| \| y_j \|} \tag{36}$$

Denklem 36'da x_i aday metindeki i'inci kelimenin embedding vektörünü, y_j referans metindeki j'inci kelimenin embedding vektörünü ifade etmektedir. BERTScore kesinlik değeri aşağıdaki şekilde hesaplanmaktadır:

$$P_{BERT} = \frac{1}{|\hat{x}|}\sum_{x_i \in \hat{x}} \max_{y_j \in y} \mathrm{sim}(x_i, y_j)$$

(37)

Denklem 37'de $\hat{x}$ aday metindeki kelimeleri, y referans metindeki kelimeleri göstermektedir.

Duyarlılık değeri ise denklem 38'deki şekilde tanımlanmaktadır:

$$R_{BERT} = \frac{1}{|y|}\sum_{y_j \in y} \max_{x_i \in \hat{x}} \mathrm{sim}(x_i, y_j)$$

(38)

Son olarak BERTScore F1 değeri kesinlik ve duyarlılık skorlarının harmonik ortalaması alınarak hesaplanmaktadır:

$$F1_{BERT} = 2 \cdot \frac{P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}}$$

(39)

BERTScore'un en önemli avantajlarından biri, bağlamsal dil temsillerini kullanması sayesinde anlamsal benzerliği klasik n-gram tabanlı metriklerden çok daha başarılı değerlendirebilmesidir. Özellikle tıbbi raporlama sistemlerinde, aynı klinik bulgunun farklı ifadelerle tanımlanabilmesi nedeniyle bu özellik büyük önem taşımaktadır. Örneğin "acute infarction" ve "recent ischemic lesion" gibi klinik olarak benzer anlam taşıyan ifadeler, klasik metriklerde düşük eşleşme üretebilirken, BERTScore bu anlamsal yakınlığı daha başarılı biçimde yakalayabilmektedir.

Bu tez kapsamında BERTScore metriği, otomatik olarak üretilen radyolojik raporların uzman radyologlar tarafından hazırlanan referans raporlarla anlamsal uyumunu değerlendirmek amacıyla kullanılmıştır. Özellikle klinik terminoloji, anatomik tanımlar ve patolojik ifadelerin semantik doğruluğunu ölçmek açısından BERTScore önemli avantaj sağlamaktadır. Böylece yalnızca sözcüksel benzerlik değil, aynı zamanda klinik anlam tutarlılığı da değerlendirme sürecine dahil edilmektedir.

Bununla birlikte BERTScore metriğinin bazı sınırlılıkları da bulunmaktadır. Hesaplama maliyeti klasik metriklere göre daha yüksektir ve kullanılan önceden eğitilmiş dil modelinin kalitesi sonuçları doğrudan etkileyebilmektedir. Ayrıca genel amaçlı BERT modelleri bazı tıbbi terminolojileri yeterince temsil edemeyebilir. Bu nedenle medikal uygulamalarda BioBERT, ClinicalBERT veya PubMedBERT gibi alan-özgü modellerin kullanılması önerilmektedir. Buna rağmen BERTScore, günümüzde doğal dil üretimi sistemlerinin anlamsal değerlendirilmesinde en güçlü otomatik metriklerden biri olarak kabul edilmektedir.

Otomatik NLP metrikleri, üretilen raporların dilsel kalitesini ölçmekte ancak klinik doğruluğu doğrudan değerlendirememektedir. Bu nedenle, üretilen raporların klinik açıdan

anlamlılığını ve doğruluğunu nicel olarak ölçmek amacıyla alana özgü klinik içerik doğruluk metrikleri tanımlanmıştır. Bu metrikler, radyoloji raporlarında yer alması gereken kritik klinik bilgilerin doğru şekilde üretilip üretilmediğini kontrol etmektedir.

Yerleşim tarafı bilgisi, radyoloji raporlarında en kritik klinik parametrelerden biridir. Yanlış hemisfer bildirimi doğrudan klinik hatalara yol açabileceğinden, üretilen raporlarda yerleşim tarafı bilgisinin doğruluğu ayrı bir metrik olarak değerlendirilmiştir. Bu metrik, üretilen rapordaki yerleşim tarafı ifadesinin (sağ, sol veya bilateral) referans rapordaki ifade ile uyumlu olup olmadığını kontrol etmektedir. Yerleşim tarafı doğruluğu aşağıdaki şekilde hesaplanmıştır:

$$Yerleşim\ Tarafı\ Doğruluğu = \frac{Doğru\ lateralite\ içeren\ rapor\ sayısı}{Toplam\ rapor\ sayısı} \times 100 \quad (40)$$

Üretilen raporlarda belirtilen anatomik bölgelerin (frontal lob, temporal lob, parietal lob, oksipital lob, bazal ganglion, talamus, insüler korteks vb.) referans raporlarla uyumu değerlendirilmiştir. Bu değerlendirmede, referans raporda bahsedilen anatomik yapıların üretilen raporda da doğru şekilde yer alıp almadığı kontrol edilmiştir. Anatomik bölge doğruluğu hem kesinlik (üretilen rapordaki anatomik terimlerin doğruluk oranı) hem de duyarlılık (referans rapordaki anatomik terimlerin üretilen raporda yakalanma oranı) boyutlarında hesaplanmıştır.

İskemik inme raporlarında lezyonun ilişkili olduğu vasküler sulama alanının (MCA, ACA, PCA, vertebrobaziler) doğru şekilde bildirilmesi klinik açıdan önemlidir. Bu metrik, üretilen rapordaki vasküler alan bilgisinin segmentasyon ve atlas eşleme aşamasından elde edilen referans bilgi ile uyumunu değerlendirmektedir.

Radyoloji raporlarında kullanılan klinik terminolojinin standart ve tutarlı olması beklenmektedir. Bu kapsamda, üretilen raporlarda "difüzyon kısıtlanması", "akut enfarkt", "iskemik değişiklik" gibi anahtar klinik terimlerin doğru ve uygun bağlamda kullanılıp kullanılmadığı değerlendirilmiştir. Terminoloji tutarlılığı, önceden tanımlanmış bir tıbbi terim sözlüğü ile otomatik eşleştirme yoluyla ölçülmüştür.

Üretilen raporların yapısal olarak standart radyoloji rapor formatına uygunluğu değerlendirilmiştir. Bu kapsamda, raporların "Bulgular" ve "Sonuç" bölümlerini içerip içermediği, uygun başlık yapısına sahip olup olmadığı ve klinik raporlama kurallarına uygun bir şekilde yapılandırılıp yapılandırılmadığı kontrol edilmiştir. Format doğruluğu ikili (uygun/uygun değil) bir değerlendirme olarak hesaplanmıştır.

Model tarafından üretilen raporların genel üretim kalitesini değerlendirmek ve olası sorunları tespit etmek amacıyla ek kalite kontrol metrikleri tanımlanmıştır.

Dil modellerinin metin üretimi sırasında karşılaşılan yaygın problemlerden biri, aynı ifadelerin tekrar tekrar üretilmesidir. Bu sorunu nicel olarak ölçmek amacıyla, üretilen metindeki tekrarlayan n-gram oranı hesaplanmıştır. Tekrar oranı, metindeki benzersiz n-gram sayısının toplam n-gram sayısına bölünmesiyle elde edilen değerin 1'den çıkarılmasıyla hesaplanmaktadır:

$$Tekrar\ oranı\ =\ 1\ -\ (\frac{Benzersiz\ n-gram\ sayısı}{Toplam\ n-gram\ sayısı})$$

(41)

Yüksek tekrar oranı, modelin metin üretimi sırasında döngüsel kalıplara girdiğini ve bu durumun rapor kalitesini olumsuz etkilediğini göstermektedir.

Üretilen raporların uzunluğunun referans raporlarla karşılaştırılması, modelin bilgi yoğunluğu ve özetleme kapasitesi hakkında bilgi sunmaktadır. Uzunluk oranı aşağıdaki şekilde hesaplanmıştır:

$$Uzunluk\ oranı\ =\ \frac{Üretilen\ rapor\ kelime\ sayısı}{Referans\ rapor\ kelime\ sayısı}$$

(42)

Bu oran 1,0'a yakın değerler aldığında üretilen raporun referans raporla benzer uzunlukta olduğunu, 1,0'dan önemli ölçüde düşük olduğunda modelin eksik bilgi ürettiğini, 1,0'dan yüksek olduğunda ise modelin gereksiz detay veya tekrar içeren uzun metinler ürettiğini göstermektedir.

Büyük dil modellerinin en önemli risklerinden biri, girdi verilerinde bulunmayan bilgileri uyduran (halüsinasyon yapan) çıktılar üretmesidir. Klinik raporlamada halüsinasyon, var olmayan lezyonların bildirilmesi, yanlış anatomik bölge atanması veya uygun olmayan klinik yorumların eklenmesi şeklinde ortaya çıkabilir. Bu çalışmada halüsinasyon tespiti, üretilen rapordaki anatomik ve klinik ifadelerin segmentasyon ve atlas eşleme aşamalarından elde edilen yapılandırılmış bilgilerle karşılaştırılması yoluyla gerçekleştirilmiştir. Yapılandırılmış girdide bulunmayan ancak üretilen raporda yer alan klinik bilgiler halüsinasyon olarak işaretlenmiştir. Otomatik metriklerin ölçemediği klinik yargı, terminolojik uygunluk ve raporun genel kullanılabilirliği gibi boyutların değerlendirilebilmesi amacıyla uzman radyolog değerlendirmesi gerçekleştirilmiştir. Bu değerlendirme, otomatik rapor üretim sistemlerinin klinik geçerliliğinin belirlenmesinde altın standart olarak kabul edilmektedir.

Değerlendirme sürecinde, doğrulama setindeki hastalar için altı farklı model tarafından üretilen raporlar uzman radyologlara sunulmuştur. Değerlendirme, kör bir protokol ile gerçekleştirilmiş olup değerlendiricilere raporun hangi model tarafından üretildiği bilgisi verilmemiştir. Her bir rapor, beş farklı boyutta 5 puanlık Likert ölçeği kullanılarak değerlendirilmiştir. Bu boyutlar Tablo 2.12'de açıklanmaktadır.

Her bir değerlendirme boyutu için puanlama kriterleri önceden tanımlanmış olup değerlendiricilere detaylı bir puanlama kılavuzu sağlanmıştır. Bu yaklaşım, değerlendiriciler arasındaki tutarlılığı artırmayı ve öznel yargı farklılıklarını en aza indirmeyi amaçlamaktadır. Bölüm 2.12'de tanıtılan altı modelin performanslarının sistematik olarak karşılaştırılması amacıyla tüm metrikler aynı doğrulama seti (21 hasta) üzerinde hesaplanmıştır. Modeller arası karşılaştırma dört farklı araştırma sorusu çerçevesinde yapılandırılmıştır.

Tablo 2.12. Uzman değerlendirmesinde kullanılan Likert ölçeği boyutları.

Değerlendirme boyutu	**Açıklama**	**1 (en düşük)**	**5 (en yüksek)**
Klinik doğruluk	Rapordaki bulguların segmentasyon sonuçlarıyla uyumu	Tamamen yanlış bulgular	Tüm bulgular doğru ve eksiksiz
Anatomik doğruluk	Belirtilen anatomik bölge ve yerleşim tarafı bilgisinin doğruluğu	Yanlış taraf veya bölge	Doğru ve detaylı lokalizasyon
Terminoloji uygunluğu	Kullanılan tıbbi terminolojinin radyoloji pratiğine uygunluğu	Uygunsuz veya hatalı terimler	Standart radyoloji terminolojisi
Dilsel akıcılık	Raporun dilbilgisel doğruluğu ve okunabilirliği	Anlaşılmaz veya bozuk cümleler	Akıcı ve profesyonel dil
Klinik kullanılabilirlik	Raporun klinik karar destek amacıyla kullanılabilirliği	Hiç kullanılamaz	Doğrudan kullanılabilir

Birinci karşılaştırma, Qwen2.5 ailesindeki 3B, 7B ve 14B modelleri arasında gerçekleştirilmiş olup model boyutunun rapor üretim kalitesine etkisini analiz etmeyi amaçlamaktadır. Bu karşılaştırma, aynı mimari aile içinde yalnızca parametre sayısının değiştiği kontrollü bir deney ortamı sunmaktadır.

İkinci karşılaştırma, aynı parametre ölçeğindeki Qwen2.5-7B ve Mistral-7B modelleri arasında gerçekleştirilmiş olup farklı mimari tasarımların (tam bağlam dikkat mekanizması ile kayan pencere dikkat mekanizması) ve tokenizer yapılarının Türkçe radyoloji rapor üretimine etkisini incelemeyi hedeflemektedir.

Üçüncü karşılaştırma, Türkçe'ye özgü olarak sıfırdan eğitilmiş Kumru-2B modeli ile çok dilli modeller arasında yapılmış olup dile özgü ön eğitimin alan-spesifik ince ayar performansına etkisini araştırmaktadır.

Dördüncü karşılaştırma, LLaMA-3 üzerine Türkçe ince ayar yapılmış Cosmos-8B modeli ile genel çok dilli modeller arasında gerçekleştirilmiş olup transfer öğrenme yaklaşımının Türkçe tıbbi metin üretimindeki etkinliğini değerlendirmektedir.

Her bir karşılaştırma için otomatik NLP metrikleri, klinik doğruluk metrikleri ve uzman değerlendirme sonuçları birlikte analiz edilerek, modellerin güçlü ve zayıf yönleri çok boyutlu olarak ortaya konmuştur.

Değerlendirme süreci, model çıkarımından başlayarak otomatik ve uzman değerlendirme aşamalarını kapsayan çok adımlı bir iş akışı şeklinde gerçekleştirilmiştir. İlk olarak, doğrulama setindeki 21 hastanın yapılandırılmış klinik bilgileri altı modele giriş olarak verilmiş ve her model tarafından ayrı ayrı Türkçe radyoloji raporları üretilmiştir. Üretilen

raporlar, aynı hastaya ait referans raporlarla eşleştirilerek otomatik NLP metrikleri hesaplanmıştır. Eş zamanlı olarak, klinik içerik doğruluk metrikleri hesaplanarak yerleşim tarafı, anatomik bölge ve vasküler alan doğruluğu değerlendirilmiştir. Son olarak, üretilen raporlar uzman radyologlara sunularak kör değerlendirme gerçekleştirilmiş ve Likert ölçeği puanları elde edilmiştir.

Bu çok katmanlı değerlendirme yaklaşımı, üretilen raporların yalnızca kelime düzeyinde benzerliğini değil, aynı zamanda anlamsal uyumunu, klinik doğruluğunu ve pratik kullanılabilirliğini de kapsamlı bir şekilde değerlendirmektedir. Böylece her bir modelin güçlü ve zayıf yönleri farklı boyutlarda ortaya konmakta ve en uygun modelin belirlenmesi için çok boyutlu bir karar çerçevesi oluşturulmaktadır.

3. BULGULAR ve SONUÇLAR

Bu bölümde, tez kapsamında geliştirilen çok aşamalı otomatik radyolojik rapor üretim sisteminin deneysel sonuçları sunulmaktadır. Önerilen sistem; derin öğrenme tabanlı segmentasyon, görüntü tabanlı özellik çıkarımı, anatomik ve vasküler bilgi entegrasyonu ile LLM tabanlı rapor üretimi bileşenlerinden oluşmaktadır. Elde edilen bulgular, sistemin her bir bileşeninin performansını ayrı ayrı değerlendirecek şekilde yapılandırılmıştır.

İlk olarak, segmentasyon modellerinin lezyon tespitindeki başarımı Dice ve IoU gibi nicel metrikler üzerinden analiz edilmiştir. Ardından, segmentasyon çıktılarından elde edilen özelliklerin doğruluğu ve anatomik atlaslar ile gerçekleştirilen eşleme sürecinin etkinliği incelenmiştir. Son olarak, yapılandırılmış klinik bilgiler kullanılarak üretilen otomatik radyolojik raporlar değerlendirilmiş ve modelin klinik açıdan anlamlı ve tutarlı metinler üretme kapasitesi analiz edilmiştir. Buna ek olarak, elde edilen performans farklarının istatistiksel olarak anlamlı olup olmadığını değerlendirmek amacıyla uygun istatistiksel testler uygulanmıştır. Bu kapsamda, farklı model ve giriş konfigürasyonları arasındaki performans karşılaştırmaları istatistiksel anlamlılık analizleri ile desteklenmiş ve elde edilen bulguların güvenilirliği artırılmıştır.

Elde edilen sonuçlar, önerilen yaklaşımın yalnızca görüntü tabanlı analiz ile sınırlı kalmayıp, bu analizleri klinik bağlam içerisinde anlamlandırarak bütüncül ve istatistiksel olarak desteklenmiş bir karar destek sistemi sunduğunu göstermektedir.

3.1. Segmentasyon Sonuçları

Bu bölümde, kullanılan segmentasyon modellerinin performansı nicel metrikler ve istatistiksel analizler ile değerlendirilmiştir. Modellerin başarımı, Dice benzerlik katsayısı, IoU, kesinlik ve duyarlılık metrikleri kullanılarak ölçülmüştür. Tüm metrikler, her bir hasta için ayrı ayrı hesaplanmış ve ortalama değerler üzerinden raporlanmıştır.

Farklı modeller arasında elde edilen performans farklarının istatistiksel olarak anlamlı olup olmadığını değerlendirmek amacıyla eşleştirilmiş t-testi uygulanmıştır. Bu testte, her bir modelin aynı hasta örnekleri üzerindeki performans değerleri karşılaştırılmıştır. İstatistiksel anlamlılık seviyesi $p<0,05$ olarak kabul edilmiştir. Bu sayede modeller arasındaki performans farklarının büyüklüğü yorumlanabilir hale getirilmiştir.

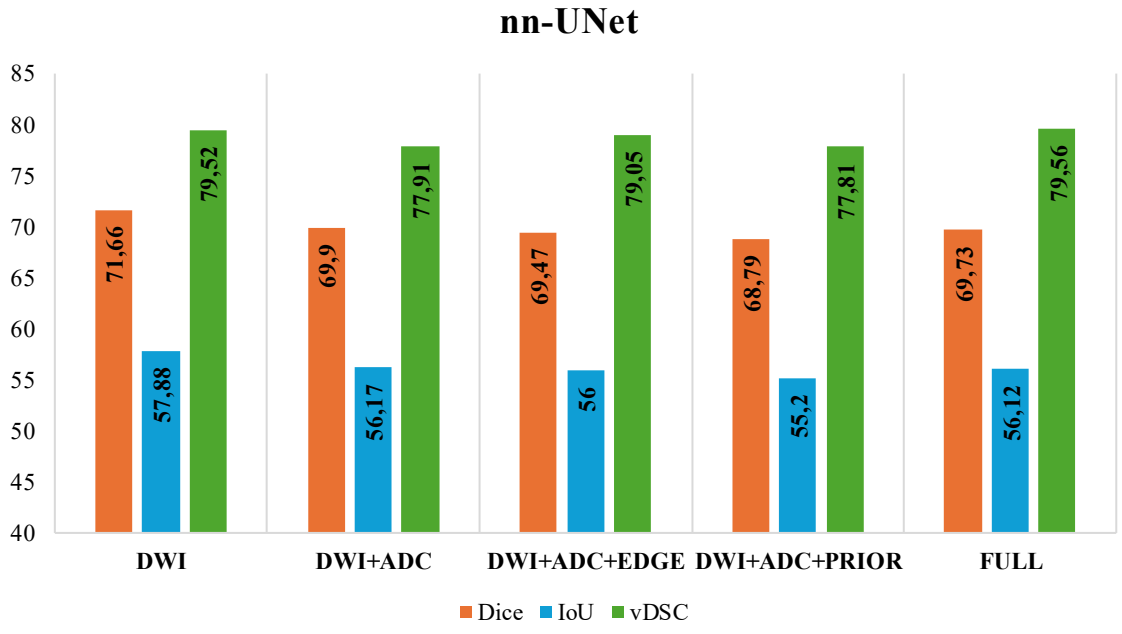
Tablo 3.1 incelendiğinde, nnU-Net modelinin farklı giriş konfigürasyonları altında tutarlı bir performans sergilediği görülmektedir. Özellikle yalnızca DWI verisinin kullanıldığı

temel konfigürasyon yüksek Dice (%71,66) ve IoU (%57,88) değerlerine ulaşmıştır. Bu durum, DWI görüntülerinin akut iskemik lezyonların tespitinde güçlü ayırt edici özellikler sunduğunu ve nnU-Net mimarisinin bu bilgiyi etkin bir şekilde öğrenebildiğini göstermektedir.

Tablo 3.1. nnU-Net için farklı giriş konfigürasyonlarına göre segmentasyon sonuçları.

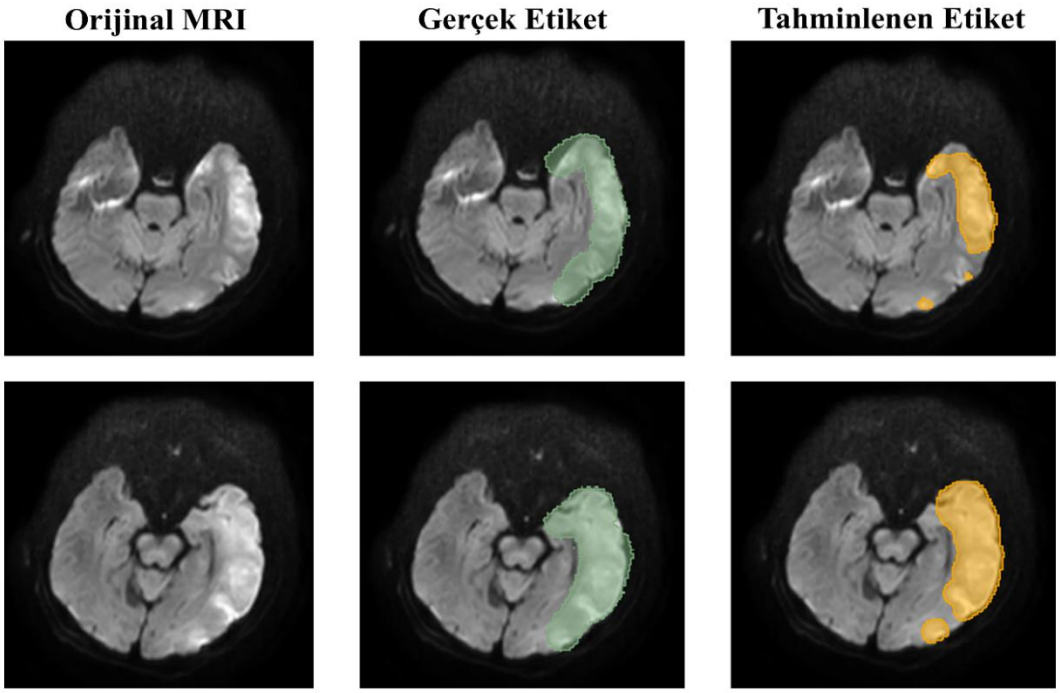
Giriş Konfigürasyonu	Dice (%)	IoU (%)	Kesinlik (%)	Duyarlılık (%)	vDSC (%)
DWI	**71,66 ± 0,16**	**57,88 ± 0,16**	**75,90 ± 0,19**	**72,30 ± 0,20**	79,52
DWI+ADC	69,90 ± 0,18	56,17 ± 0.17	73,31 ± 0,21	71,75 ± 0,20	77,91
DWI+ADC+EDGE	69,47 ± 0,19	56,00 ± 0,18	73,85 ± 0,21	69,69 ± 0,23	79,05
DWI+ADC+PRIOR	68,79 ± 0,19	55,20 ± 0,18	74,10 ± 0,21	70,33 ± 0,22	77,81
FULL	69,73 ± 0,18	56,12 ± 0,18	74,73 ± 0,21	69,99 ± 0,22	**79,56**

DWI ve ADC sekanslarının birlikte kullanıldığı konfigürasyonda (DWI+ADC), performansın bir miktar düştüğü gözlemlenmiştir. Bu sonuç, teorik olarak tamamlayıcı bilgi sağlaması beklenen ADC verisinin her durumda performans artışı sağlamadığını göstermektedir. Bunun olası nedeni, farklı görüntü sekanslarının erken aşamada birlikte kullanılması sonucunda modelin öğrenme sürecinde temsil karmaşıklığının artması ve bu durumun model optimizasyonunu zorlaştırması olabilir. Tablo 3.1'de sayısal olarak sunulan performans sonuçları, Şekil 3.1'de grafiksel olarak da gösterilerek farklı giriş konfigürasyonlarının etkisi daha açık biçimde ortaya konulmuştur.



Şekil 3.1. nnU-Net modeli için farklı giriş konfigürasyonlarına göre Dice, IoU ve vDSC metriklerinin karşılaştırılması.

EDGE bilgisinin eklendiği durumda (DWI+ADC+EDGE), Dice ve IoU değerlerinde sınırlı bir değişim gözlemlenmiştir. Bununla birlikte, vDSC değerinin nispeten yüksek kalması, kenar tabanlı bilgilerin özellikle lezyon sınırlarının belirlenmesinde faydalı olabileceğini göstermektedir. Ancak bu katkı, genel segmentasyon performansına doğrudan ve güçlü bir artış olarak yansımamıştır. nnU-Net modeli altında DWI+ADC verisinin kullanıldığı konfigürasyonda elde edilen segmentasyon sonuçlarının karşılaştırmaları, orijinal görüntü, gerçek etiket ve model tahminlerini içerecek şekilde Şekil 3.2'de gösterilmektedir.

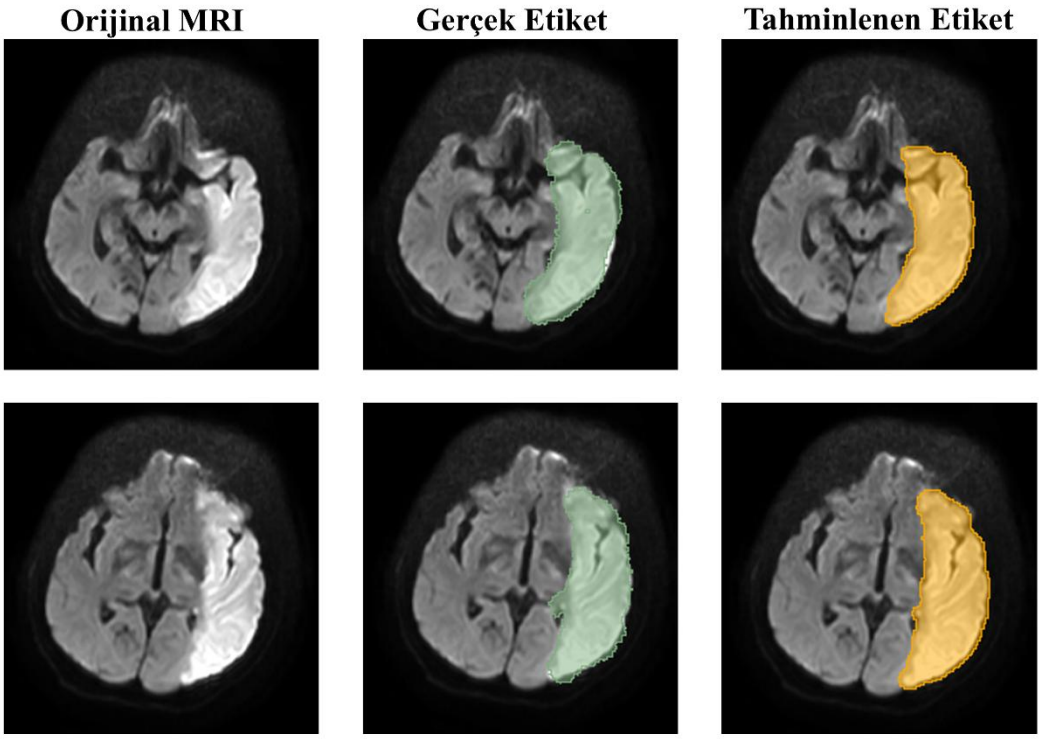


Şekil 3.2. nnU-Net modeli ve DWI+ADC giriş konfigürasyonu kullanılarak elde edilen segmentasyon sonuçlarının örnek görselleri.

Anatomik ön bilgi (PRIOR) eklendiğinde (DWI+ADC+PRIOR), performansın tekrar bir miktar düştüğü görülmektedir. Bu durum, atlas tabanlı ön bilgilerin her zaman model performansını artırmadığını ve özellikle bireysel hasta anatomisi ile atlas arasındaki uyumsuzlukların model üzerinde olumsuz etki oluşturabileceğini düşündürmektedir. Ayrıca, bu tür ön bilgilerin kaba uzamsal rehberlik sağlamasına rağmen, voxel düzeyinde kesin sınırlar sunmaması performans üzerinde sınırlayıcı bir faktör olabilir. Tüm bileşenlerin birlikte kullanıldığı tam konfigürasyonda (FULL), performansın en yüksek seviyeye ulaşmadığı, aksine bazı metriklerde düşüş yaşandığı gözlemlenmiştir. Bu bulgu, çoklu giriş bileşenlerinin basit bir şekilde bir araya getirilmesinin her zaman performans artışı sağlamadığını açıkça ortaya koymaktadır. Özellikle erken füzyon yaklaşımında, farklı bilgi kaynaklarının temsil uzayında çakışma oluşturabileceği ve bu durumun modelin öğrenme kapasitesini olumsuz etkileyebileceği değerlendirilmektedir.

Şekil 3.3'te, nnU-Net modeli ve yalnızca DWI giriş konfigürasyonu kullanılarak elde edilen segmentasyon sonuçlarına ait örnek bir vaka sunulmaktadır. Görsellerde sırasıyla orijinal MRI kesitleri, uzman tarafından oluşturulmuş gerçek etiketler ve model tarafından tahmin edilen segmentasyon çıktıları karşılaştırmalı olarak gösterilmektedir. Görseller incelendiğinde, modelin lezyon bölgesini genel hatlarıyla doğru şekilde tespit ettiği görülmektedir. Özellikle lezyonun hemisferik yerleşimi ve uzamsal yayılımı açısından

tahmin edilen maske ile gerçek etiket arasında yüksek düzeyde örtüşme bulunmaktadır. Bu durum, modelin lezyonun konumunu ve makroskopik morfolojisini başarılı bir şekilde öğrenebildiğini göstermektedir.

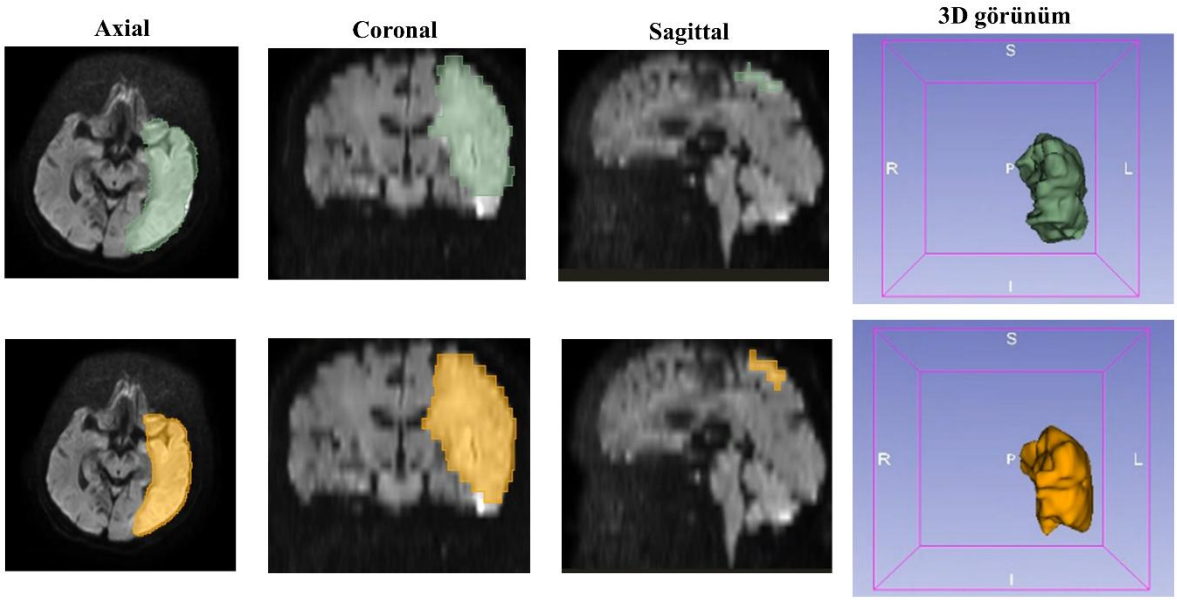


Şekil 3.3. nnU-Net modeli ve DWI giriş konfigürasyonu kullanılarak elde edilen segmentasyon sonuçlarının örnek görselleri.

Nicel sonuçlar da bu gözlemi desteklemektedir. İlgili vaka için elde edilen Dice katsayısı %89,00, IoU değeri %80,17, kesinlik değeri %90,30 ve duyarlılık değeri %87,73 olarak hesaplanmıştır. Dice ve IoU değerlerinin yüksek olması, model tahmininin gerçek etiket ile güçlü bir uzamsal örtüşme sağladığını göstermektedir. Kesinlik değerinin %90,30 olması, modelin yanlış pozitif üretme eğiliminin düşük olduğunu, yani lezyon olmayan bölgeleri yanlışlıkla lezyon olarak sınıflandırmadığını göstermektedir. Buna karşılık duyarlılık değerinin %87,73 olması, modelin gerçek lezyon bölgelerinin büyük çoğunluğunu başarıyla tespit ettiğini, ancak bazı küçük bölgelerde eksik segmentasyon oluşabileceğini göstermektedir.

Görsel analiz ile değerlendirildiğinde, modelin özellikle lezyonun ana gövdesini doğru şekilde segment ettiği, ancak sınır bölgelerinde kısmi sapmaların oluşabildiği gözlemlenmektedir. Bu sapmalar, özellikle düşük kontrastlı geçiş bölgelerinde ve lezyon sınırlarının belirsiz olduğu alanlarda ortaya çıkmaktadır. Bunun temel nedeni, DWI görüntülerinde bazı bölgelerde sinyal değişiminin kademeli olması ve lezyon sınırlarının keskin olmamasıdır. Bu durum, modelin voxel düzeyinde kesin ayrım yapmasını zorlaştırmaktadır. Bu örnek vaka üzerinden elde edilen sonuçlar, nnU-Net modelinin yalnızca DWI verisi kullanılarak bile yüksek performanslı segmentasyon gerçekleştirebildiğini göstermektedir. Bu durum, DWI sekansının akut iskemik lezyonların tespitinde güçlü bir bilgi kaynağı olduğunu ve modelin bu bilgiyi etkin şekilde kullanabildiğini ortaya koymaktadır. Aynı zamanda, görsel ve nicel sonuçların birlikte

değerlendirilmesi, model performansının yalnızca metrikler üzerinden değil, klinik açıdan anlamlı görsel karşılaştırmalar ile de desteklenmesi gerektiğini göstermektedir.



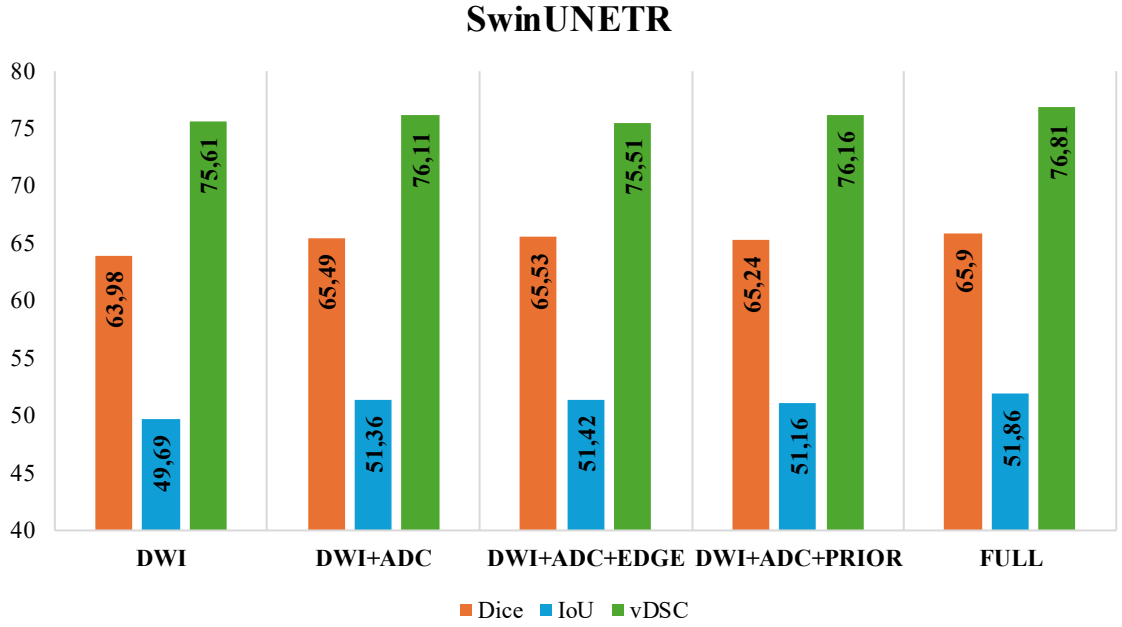
Şekil 3.4. nnU-Net modeli ile elde edilen segmentasyon sonuçlarının üç düzlem ve üç boyutlu görünüm üzerinden karşılaştırılması.

Şekil 3.4 incelendiğinde, modelin lezyonun üç boyutlu uzamsal yapısını genel olarak doğru şekilde yakalayabildiği görülmektedir. Özellikle aksiyal düzlemde elde edilen segmentasyon sonuçlarının, koronal ve sagittal düzlemlerle tutarlı olduğu gözlemlenmektedir. Üç boyutlu hacimsel gösterim, lezyonun morfolojik yapısını daha bütüncül bir şekilde ortaya koymakta ve modelin lezyonun uzamsal yayılımını doğru şekilde tahmin edebildiğini göstermektedir. Bununla birlikte, bazı bölgelerde sınırların hafif genişlediği veya küçük alanların eksik segment edildiği gözlemlenmektedir. Bu durum, modelin özellikle ince ve düşük kontrastlı yapıların bulunduğu bölgelerde hata yapabildiğini göstermektedir.

Tablo 3.2. SwinUNETR için farklı giriş konfigürasyonlarına göre segmentasyon sonuçları.

Giriş Konfigürasyonu	**Dice (%)**	**IoU (%)**	**Kesinlik (%)**	**Duyarlılık (%)**	**vDSC (%)**
DWI	63,98 ± 0,19	49,69 ± 0,18	70,99 ± 0,21	64,41 ± 0,22	75,61
DWI+ADC	65,49 ± 0,19	51,36 ± 0,18	**73,24 ± 0,21**	64,35 ± 0,22	76,11
DWI+ADC+EDGE	65,53 ± 0,19	51,42 ± 0,18	72,12 ± 0,21	64,93 ± 0,22	75,51
DWI+ADC+PRIOR	65,24 ± 0,20	51,16 ± 0,18	72,17 ± 0,21	64,46 ± 0,24	76,16
FULL	**65,90 ± 0,19**	**51,86 ± 0,18**	72,36 ± 0,21	**65,54 ± 0,23**	**76,81**

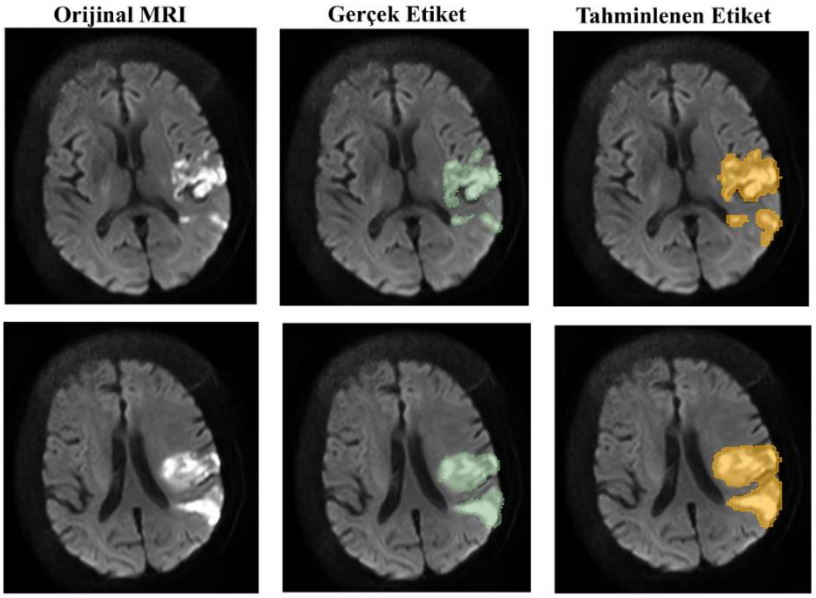
Tablo 3.2 incelendiğinde, SwinUNETR modelinin farklı giriş konfigürasyonları altında daha dengeli ancak nnU-Net modeline kıyasla genel olarak daha düşük performans sergilediği görülmektedir. Yalnızca DWI verisinin kullanıldığı temel konfigürasyonda Dice değeri %63,98 olarak elde edilirken, çok kanallı girişlerin eklenmesiyle performansta sınırlı düzeyde iyileşmeler gözlemlenmiştir. DWI ve ADC sekanslarının birlikte kullanıldığı konfigürasyonda (DWI+ADC), Dice değerinin %65,49'a yükseldiği görülmektedir. Bu durum, ADC verisinin SwinUNETR modeli için tamamlayıcı bir bilgi kaynağı sunduğunu ve modelin çoklu görüntü sekanslarından fayda sağlayabildiğini göstermektedir. Ancak bu artışın sınırlı kalması, modelin multimodal veriyi etkin şekilde birleştirme kapasitesinin sınırlı olabileceğine işaret etmektedir. Şekil 3.5'te, SwinUNETR modeli için farklı giriş konfigürasyonları kullanılarak elde edilen Dice, IoU ve vDSC performans sonuçları karşılaştırmalı olarak sunulmaktadır.



Şekil 3.5. SwinUNETR modeli için farklı giriş konfigürasyonlarına göre Dice, IoU ve vDSC metriklerinin karşılaştırılması.

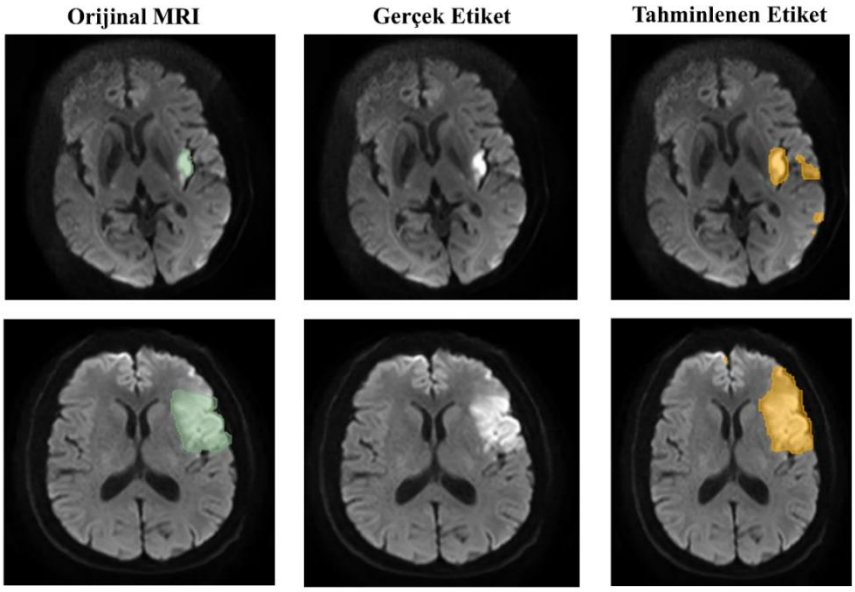
EDGE bilgisinin eklendiği durumda (DWI+ADC+EDGE), Dice ve IoU değerlerinde çok küçük bir artış gözlemlenmiştir. Bu bulgu, SwinUNETR modelinin kenar tabanlı yüksek frekanslı bilgileri kısmen kullanabildiğini ortaya koymaktadır. Özellikle IoU değerindeki artış, modelin lezyon sınırlarını daha iyi yakalayabildiğini göstermektedir. Bununla birlikte, bu iyileşmenin sınırlı düzeyde kalması, Transformer tabanlı modellerin yerel detaylara duyarlılığının konvolüsyon tabanlı modellere kıyasla daha düşük olabileceğini göstermektedir. Anatomik ön bilgi (PRIOR) eklendiğinde (DWI+ADC+PRIOR), performansın EDGE konfigürasyonuna benzer seviyelerde kaldığı görülmektedir. Bu durum, SwinUNETR modelinin atlas tabanlı ön bilgileri doğrudan kullanmakta zorlandığını ve bu tür bilgilerin model tarafından yeterince etkin şekilde entegre edilemediğini düşündürmektedir. Özellikle kaba anatomik rehberlik sağlayan prior haritalarının, Transformer mimarilerinde doğrudan fayda sağlamayabileceği değerlendirilmektedir. Tüm bileşenlerin birlikte kullanıldığı FULL konfigürasyonda ise modelin en yüksek Dice (%65,90) ve IoU (%51,86) değerlerine ulaştığı görülmektedir. Bu sonuç, SwinUNETR modelinin çoklu bilgi kaynaklarını birlikte kullanabildiğini ancak bu katkının sınırlı

kaldığını göstermektedir. nnU-Net modelinin aksine, bu modelde çoklu girişlerin bir araya getirilmesi performansı düşürmemiş, ancak belirgin bir sıçrama da yaratmamıştır. SwinUNETR modeli ve DWI giriş konfigürasyonu kullanılarak elde edilen segmentasyon sonuçlarının, orijinal MRI görüntüleri ile gerçek etiket ve model tahminleri üzerinden karşılaştırması Şekil 3.6'da sunulmuştur.



Şekil 3.6. SwinUNETR modeli ve DWI giriş konfigürasyonu kullanılarak elde edilen segmentasyon sonuçlarının örnek görselleri.

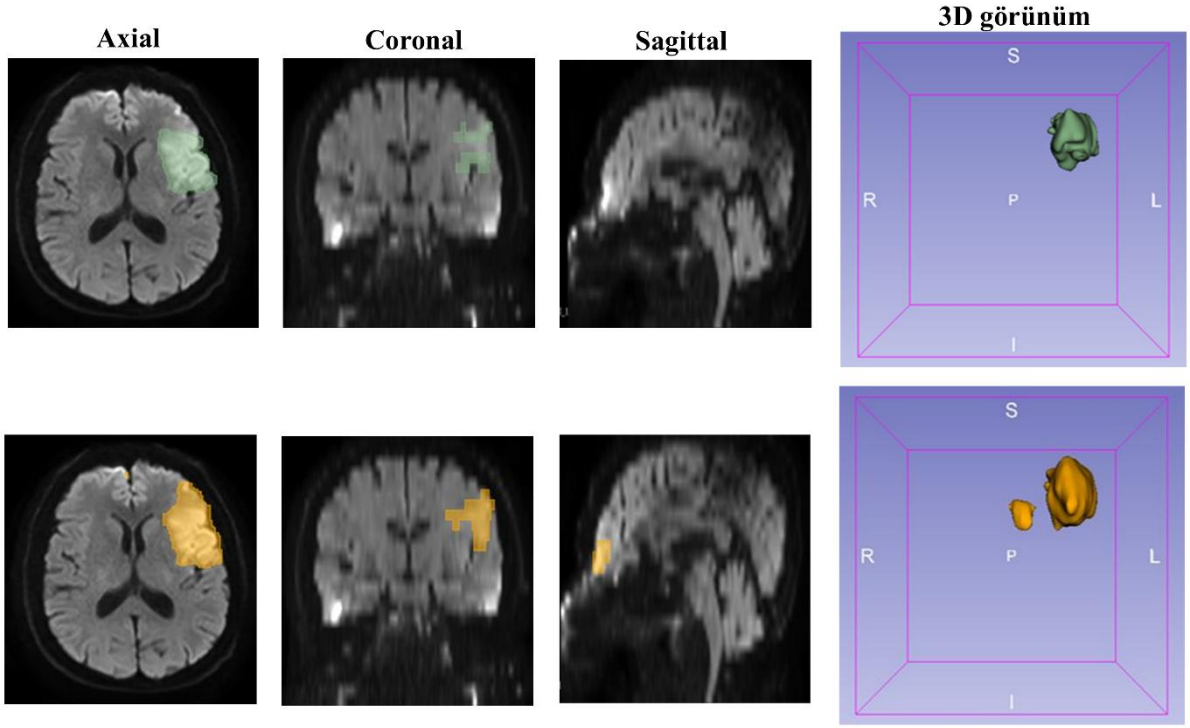
Şekil 3.7'de, SwinUNETR modeli ve FULL giriş konfigürasyonu kullanılarak elde edilen segmentasyon sonuçlarına ait örnek bir vaka sunulmaktadır. Görsellerde orijinal MRI kesitleri, uzman tarafından oluşturulmuş gerçek etiketler ve model tarafından tahmin edilen segmentasyon çıktıları karşılaştırmalı olarak gösterilmektedir.



Şekil 3.7. SwinUNETR modeli FULL giriş konfigürasyonu kullanılarak elde edilen segmentasyon sonuçlarının örnek görselleri.

İlgili vaka için elde edilen nicel performans metrikleri incelendiğinde, Dice değeri %74,02, IoU değeri %58,76, kesinlik değeri %59,57 ve duyarlılık değeri %97,74 olarak hesaplanmıştır. Bu metrikler birlikte değerlendirildiğinde, modelin lezyon bölgelerini büyük ölçüde yakalayabildiği ancak tahminlerin uzamsal doğruluğu açısından bazı sınırlılıklar içerdiği görülmektedir.

Özellikle duyarlılık değerinin oldukça yüksek olması (%97,74), modelin gerçek lezyon bölgelerinin neredeyse tamamını tespit edebildiğini göstermektedir. Bu durum, modelin lezyon kaçırma (false negative) eğiliminin oldukça düşük olduğunu ortaya koymaktadır. Ancak kesinlik değerinin görece düşük olması (%59,57), modelin lezyon olmayan bölgeleri de lezyon olarak sınıflandırma eğiliminde olduğunu, yani belirgin düzeyde yanlış pozitif (false positive) üretildiğini göstermektedir.



Şekil 3.8. SwinUNETR modeli ile elde edilen segmentasyon sonuçlarının üç düzlem ve üç boyutlu görünüm üzerinden karşılaştırılması.

Şekil 3.8'de, SwinUNETR modeli kullanılarak elde edilen segmentasyon sonuçları, aksiyal, koronal ve sagittal düzlemler ile üç boyutlu hacimsel görünüm üzerinden karşılaştırmalı olarak sunulmaktadır. Üst satırda uzman tarafından oluşturulan gerçek etiketler, alt satırda ise model tarafından tahmin edilen segmentasyon çıktıları gösterilmektedir. Görseller incelendiğinde, modelin lezyonun genel konumunu ve uzamsal dağılımını doğru şekilde yakalayabildiği görülmektedir. Özellikle aksiyal düzlemde lezyonun ana kütlesinin büyük ölçüde doğru segment edildiği dikkat çekmektedir. Bununla birlikte, koronal ve sagittal düzlemlerde bazı bölgelerde parçalı yapılar ve düzensiz sınırlar oluştuğu gözlemlenmektedir. Üç boyutlu görünüm bu durumu daha açık ortaya koymakta olup, modelin lezyon hacmini genel olarak yakaladığı ancak bazı ek küçük odaklar üreterek aşırı segmentasyon eğilimi gösterdiği anlaşılmaktadır. Bu durum, modelin yüksek duyarlılığa sahip olmasına karşın sınır doğruluğu ve özgüllük açısından sınırlılıklar taşıdığını göstermektedir. Özellikle düşük kontrastlı bölgelerde ve lezyon çevresindeki benzer

yoğunluklu dokularda yanlış pozitif segmentasyonların ortaya çıkması, Transformer tabanlı mimarilerin geniş bağlamsal alanı kullanırken lokal detayları yeterince ayrıştıramamasından kaynaklanabilir. Genel olarak değerlendirildiğinde, SwinUNETR modelinin lezyonun makroskopik yapısını başarılı bir şekilde temsil edebildiği, ancak ince sınır detayları ve küçük lezyon odaklarında iyileştirmeye ihtiyaç duyduğu görülmektedir.

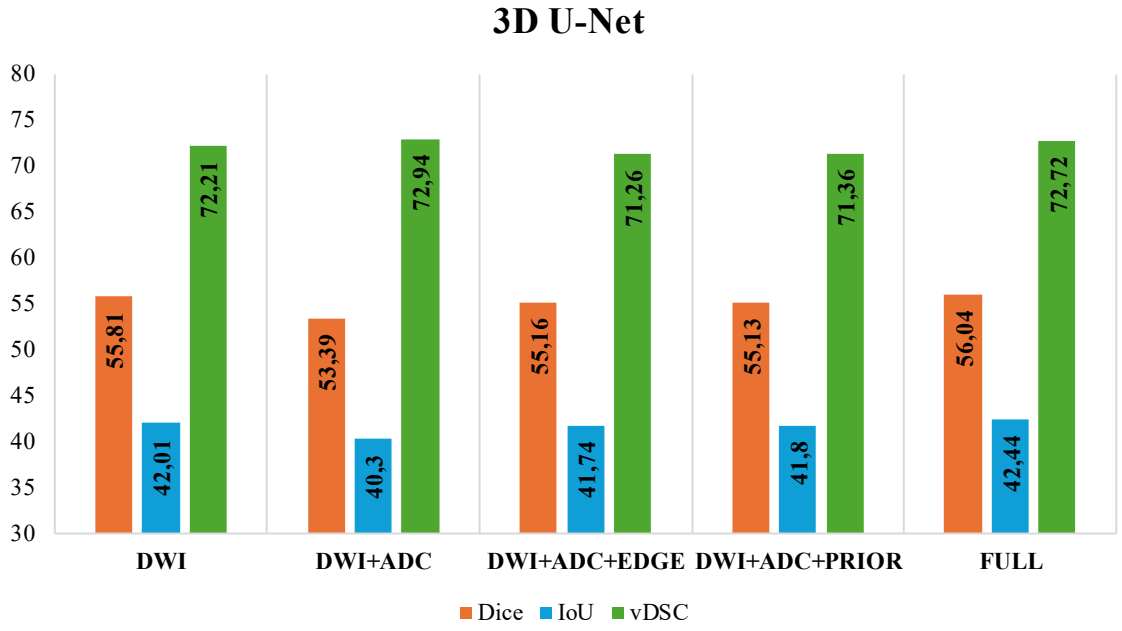
Tablo 3.3. 3D U-Net için farklı giriş konfigürasyonlarına göre segmentasyon sonuçları.

Giriş Konfigürasyonu	Dice (%)	IoU (%)	Kesinlik (%)	Duyarlılık (%)	vDSC (%)
DWI	55,81 ± 0,23	42,01 ± 0,20	**67,10 ± 0,27**	53,87 ± 0,26	72,21
DWI+ADC	53,39 ± 0,25	40,30 ± 0,22	64,45 ± 0,27	52,08 ± 0,28	**72,94**
DWI+ADC+EDGE	55,16 ± 0,24	41,74 ± 0,21	64,29 ± 0,28	53,83 ± 0,27	71,26
DWI+ADC+PRIOR	55,13 ± 0,25	41,80 ± 0,21	63,34 ± 0,27	54,36 ± 0,27	71,36
FULL	**56,04 ± 0,24**	**42,44 ± 0,21**	65,15 ± 0,26	**55,09 ± 0,27**	72,72

Tablo 3.3 incelendiğinde, 3D U-Net modelinin farklı giriş konfigürasyonları altında genel olarak daha düşük ve dalgalı bir performans sergilediği görülmektedir. Özellikle Dice ve IoU metrikleri açısından elde edilen değerler, nnU-Net ve SwinUNETR modellerine kıyasla belirgin şekilde daha düşüktür. Bu durum, 3D U-Net mimarisinin kullanılan veri seti ve problem yapısı için daha sınırlı bir temsil kapasitesine sahip olduğunu göstermektedir.

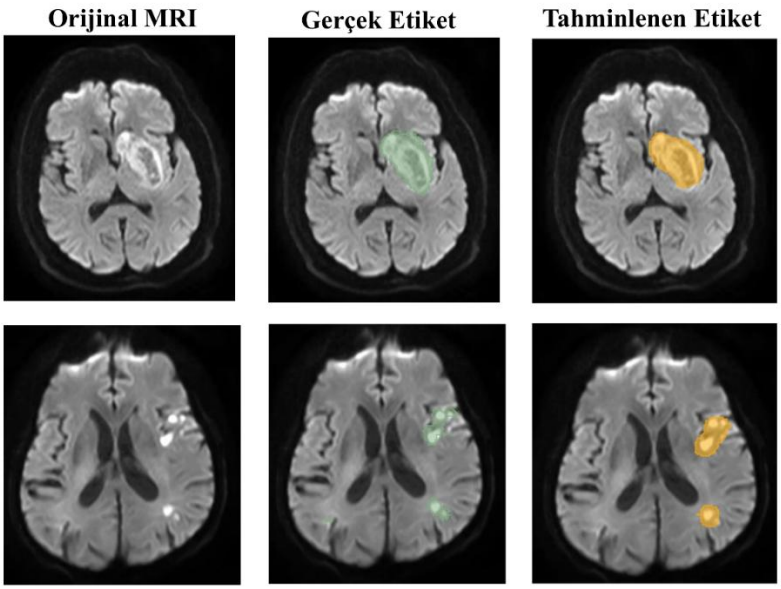
Yalnızca DWI verisinin kullanıldığı temel konfigürasyonda modelin Dice değeri %55,81 ve IoU değeri %42,01 olarak elde edilmiştir. Bu sonuç, modelin lezyon bölgelerini genel hatlarıyla tespit edebildiğini ancak uzamsal örtüşme açısından sınırlı kaldığını göstermektedir. DWI+ADC konfigürasyonunda performansın beklenenin aksine düşmesi (Dice: %53,39) dikkat çekicidir. 3D U-Net gibi daha klasik konvolüsyon tabanlı mimariler, farklı dağılımlara sahip giriş kanallarını etkin şekilde birleştirmekte zorlanabilmekte ve bu da performans kaybına yol açabilmektedir. 3D U-Net modeli için farklı giriş konfigürasyonlarına bağlı olarak elde edilen Dice, IoU ve vDSC metriklerinin karşılaştırmalı analizi Şekil 3.9'da sunulmuştur.

EDGE bilgisinin eklendiği durumda (DWI+ADC+EDGE), performansta sınırlı bir iyileşme gözlemlenmiştir (Dice: %55,16). Bu durum, kenar tabanlı bilgilerin özellikle lezyon sınırlarının belirlenmesinde model için kısmi fayda sağladığını göstermektedir. Ancak bu katkının güçlü olmaması, modelin yüksek frekanslı sınır bilgilerini derin özellik temsillerine etkin şekilde entegre edemediğini ortaya koymaktadır. Anatomik ön bilgi eklendiğinde (DWI+ADC+PRIOR), performansın EDGE konfigürasyonuna benzer seviyelerde kaldığı görülmektedir. Bu durum, atlas tabanlı ön bilgilerin 3D U-Net modeli tarafından etkin şekilde kullanılamadığını düşündürmektedir. Özellikle bu tür ön bilgilerin hasta-özel anatomik farklılıkları tam olarak yansıtamaması ve yalnızca kaba bir uzamsal rehberlik sunması, modelin voxel düzeyinde doğru öğrenme gerçekleştirmesini sınırlayabilir.



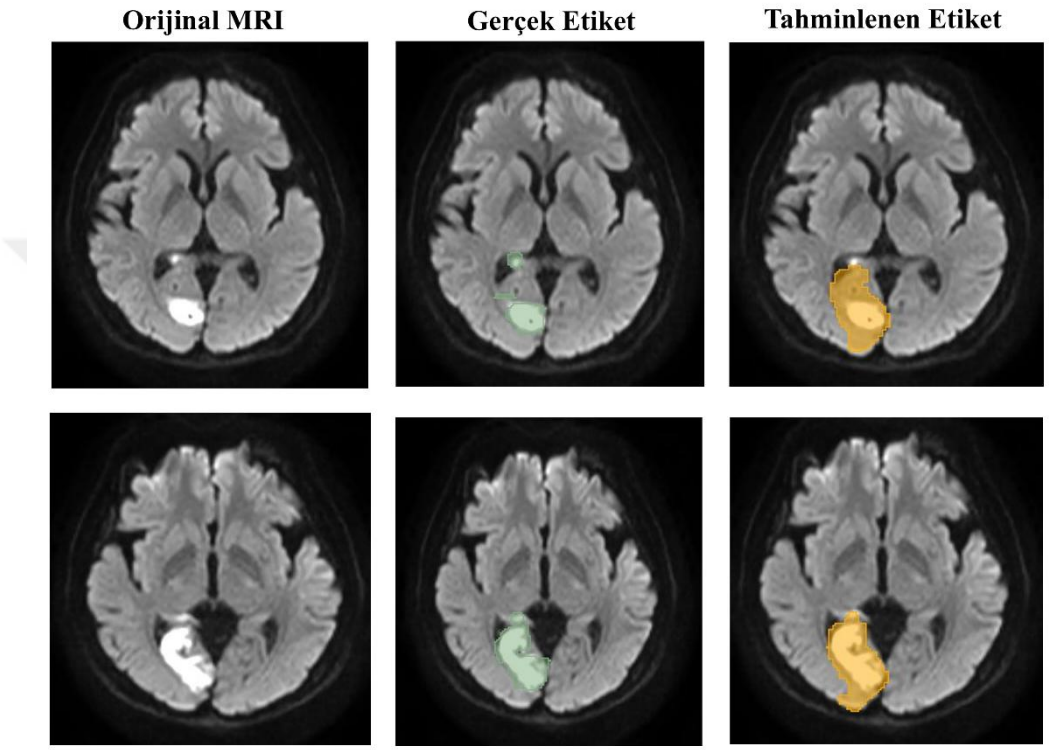
Şekil 3.9. 3D-U-Net modeli için farklı giriş konfigürasyonlarına göre Dice, IoU ve vDSC metriklerinin karşılaştırılması.

Tüm bileşenlerin birlikte kullanıldığı FULL konfigürasyonda modelin en yüksek Dice değerine (%56,04) ulaşması, çoklu girişlerin birlikte kullanılmasının sınırlı da olsa katkı sağladığını göstermektedir. Ancak bu artışın oldukça düşük olması, modelin giriş bileşenleri arasındaki karmaşık ilişkileri öğrenmekte zorlandığını ve bu bilgileri etkin şekilde birleştiremediğini ortaya koymaktadır. 3D U-Net modeli kullanılarak elde edilen segmentasyon sonuçlarının, orijinal MRI görüntüleri ile gerçek etiket ve model tahminleri üzerinden karşılaştırmalı nitel analizi Şekil 3.10'da sunulmuştur.



Şekil 3.10. 3D U-Net modeli DWI giriş konfigürasyonu kullanılarak elde edilen segmentasyon sonuçlarının örnek görselleri.

Şekil 3.11'de 3D U-Net modeli ve FULL giriş konfigürasyonu kullanılarak elde edilen segmentasyon sonuçlarına ait örnek bir vaka sunulmaktadır. Görseller incelendiğinde, modelin lezyonun genel konumunu doğru şekilde belirleyebildiği, ancak lezyon sınırlarını kesin olarak yakalamakta zorlandığı görülmektedir. Nicel metrikler de bu durumu desteklemektedir; ilgili vaka için Dice değeri %65,08 ve IoU değeri %48,24 olarak hesaplanmıştır. Bu değerler, modelin gerçek etiket ile belirli bir düzeyde örtüşme sağladığını ancak uzamsal doğruluğun sınırlı kaldığını göstermektedir. Özellikle kesinlik değerinin %52,60 olması, modelin yanlış pozitif üretme eğiliminin yüksek olduğunu, yani lezyon olmayan bölgeleri de lezyon olarak sınıflandırabildiğini ortaya koymaktadır.

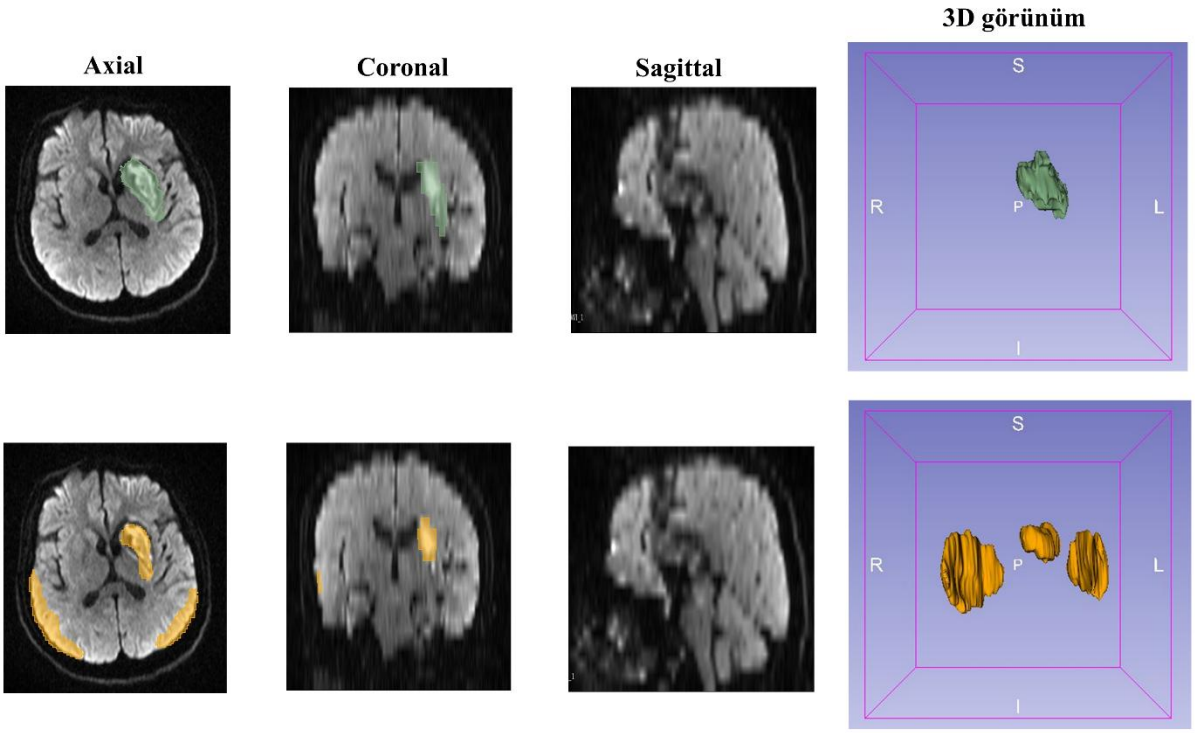


Şekil 3.11. 3D U-Net modeli FULL giriş konfigürasyonu kullanılarak elde edilen segmentasyon sonuçlarının örnek görselleri.

Buna karşılık duyarlılık değerinin %85,33 gibi görece yüksek bir seviyede olması, modelin gerçek lezyon bölgelerinin büyük bir kısmını tespit edebildiğini göstermektedir. Bu durum, modelin lezyon kaçırma olasılığının düşük olduğunu ancak sınır belirleme konusunda aşırı genişleme eğilimi gösterdiğini ortaya koymaktadır. Görsel analizde de model tahminlerinin gerçek etikete kıyasla daha geniş alanlara yayıldığı ve özellikle lezyon çevresindeki düşük kontrastlı bölgelerde hatalı segmentasyonların oluştuğu dikkat çekmektedir. Bu bulgular, 3D U-Net modelinin yüksek duyarlılık sağlayabilmesine rağmen, sınırlı temsil kapasitesi ve multimodal girişleri etkin şekilde bütünleştirememe nedeniyle özgüllük ve sınır doğruluğu açısından zayıf kaldığını göstermektedir.

Şekil 3.12'de, 3D U-Net modelinin segmentasyon sonuçları aksiyal, koronal ve sagittal düzlemler ile üç boyutlu hacimsel görünüm üzerinden karşılaştırmalı olarak sunulmaktadır. Üst satırda gerçek etiketler, alt satırda ise model tarafından üretilen tahminler yer almaktadır. Görseller incelendiğinde, modelin lezyonun genel konumunu doğru şekilde yakalayabildiği, ancak özellikle aksiyal düzlemde belirgin olmak üzere tahmin edilen maskenin gerçek etikete kıyasla daha geniş alanlara yayıldığı görülmektedir. Bu durum,

modelin lezyon sınırlarını tam olarak belirleyemediğini ve çevre dokuları da kapsayacak şekilde aşırı segmentasyon eğilimi gösterdiğini ortaya koymaktadır.



Şekil 3.12. 3D U-Net modeli ile elde edilen segmentasyon sonuçlarının üç düzlem ve üç boyutlu görünüm üzerinden karşılaştırılması.

Koronal ve sagittal düzlemlerde ise lezyonun uzamsal sürekliliğinin tam olarak korunamadığı ve bazı bölgelerde parçalı veya eksik temsil edildiği dikkat çekmektedir. Üç boyutlu görünüm bu bulguları daha net ortaya koymakta olup, modelin tek bir kompakt lezyon yapısı yerine birden fazla ayrık ve genişlemiş segmentasyon bölgeleri oluşturduğu gözlemlenmektedir. Bu sonuçlar, modelin yüksek duyarlılığa sahip olmasına rağmen uzamsal tutarlılık ve sınır doğruluğu açısından sınırlılıklar taşıdığını göstermektedir.

Üç segmentasyon mimarisinin (nnU-Net, SwinUNETR, 3D U-Net) ve beş giriş konfigürasyonunun (DWI, DWI+ADC, DWI+ADC+EDGE, DWI+ADC+PRIOR, FULL) performanslarını karşılaştırmak amacıyla eşleştirilmiş t-testi (paired t-test) ve Cohen's d etki büyüklüğü kullanılmıştır. Tüm modeller aynı 150 hastalık veri seti üzerinde 5-katlı çapraz doğrulama ile değerlendirildiğinden, her hasta için üç modelden elde edilen Dice skorları birebir eşleştirilerek istatistiksel testler uygulanmıştır.

Eşleştirilmiş t-testi [102], aynı denekler üzerinde ölçülen iki koşul arasındaki ortalama farkın istatistiksel olarak anlamlı olup olmadığını test eder. Bu test, bireyler arası varyansı ortadan kaldırarak daha hassas bir karşılaştırma sağlar. Bununla birlikte, p-değeri tek başına farkın pratik anlamlılığını yansıtmaz; büyük örneklem hacimlerinde (n=150) küçük farklar bile istatistiksel olarak anlamlı çıkabilmektedir. Bu nedenle, etki büyüklüğü ölçüsü olarak Cohen's d hesaplanmış ve her karşılaştırma için %95 güven aralığı (GA) raporlanmıştır. Cohen's d, gözlemlenen farkın standart sapma birimleri cinsinden büyüklüğünü ifade eder ve örneklem büyüklüğünden bağımsız bir yorumlama çerçevesi sunar.

Her hasta *i* için A ve B modellerinin Dice skorları arasındaki fark $d_i = x_{A,i} - x_{B,i}$ olarak tanımlanmıştır. t-istatistiği [103] aşağıdaki denklem ile hesaplanmaktadır:

$$t = \bar{d} / (s_d / \sqrt{n}) \quad (43)$$

Burada $\bar{d}$ farkların ortalamasını, s_d farkların standart sapmasını ve n eşleştirilmiş gözlem sayısını (n=150) göstermektedir. Serbestlik derecesi $df = n - 1 = 149$ olarak belirlenmiştir.

Eşleştirilmiş örnekler için Cohen's d [104], fark skorlarının ortalamasının fark skorlarının standart sapmasına bölünmesiyle elde edilir:

$$d = \bar{d} / s_d \quad (44)$$

Cohen tarafından önerilen yorumlama eşikleri Tablo 3.4' te verilmiştir.

Tablo 3.4. Cohen's d etki büyüklüğü yorumlama eşikleri.

|d| Aralığı	**Etki Büyüklüğü**	**Yorumlama**
< 0.20	İhmal edilebilir	Pratik fark yok
0.20 – 0.49	Küçük	Zayıf fark
0.50 – 0.79	Orta	Belirgin fark
≥ 0.80	Büyük	Güçlü fark

Tablo 2'de, DWI+ADC giriş konfigürasyonunda üç model arasındaki Dice skoru karşılaştırması ve ilgili istatistiksel sonuçlar sunulmaktadır. Bu konfigürasyon, nnU-Net için en yüksek ortalama Dice skorunu (71,66) veren konfigürasyon olması nedeniyle referans olarak seçilmiştir.

Tablo 3.5. DWI+ADC konfigürasyonunda modeller arası Dice skoru karşılaştırması (n=150, paired t-test).

Karşılaştırma	**Model A (Ort ± SS)**	**Model B (Ort ± SS)**	**Δ**	**Cohen's d**	**%95 GA**	**Etki**	**t**	**p-değeri**
nnU-Net vs SwinUNETR	71,66 ± 0,16	65,49 ± 0,19	+0.0617	+0.60	[+0.37, +0.82]	Orta	+7.31	1.52×10^{-11}
nnU-Net vs 3D U-Net	71,66 ± 0,16	53,39 ± 0,25	+0.1827	+0.98	[+0.74, +1.21]	Büyük	+11.96	1.60×10^{-23}
SwinUNETR vs 3D U-Net	65,49 ± 0,19	53,39 ± 0,25	+0.1210	+0.76	[+0.52, +0.99]	Orta	+9.27	1.97×10^{-16}

Tablo 3.5'teki sonuçlara göre, nnU-Net tüm karşılaştırmalarda en yüksek Dice skorunu elde etmiştir. nnU-Net ile SwinUNETR arasındaki Dice farkı 6,17 olup, Cohen's d = 0.60 ile orta düzeyde bir etki büyüklüğüne karşılık gelmektedir. %95 güven aralığının [+0.37, +0.82] sıfırı içermemesi, bu farkın istatistiksel olarak anlamlı olduğunu doğrulamaktadır.

nnU-Net ile 3D U-Net arasındaki fark daha belirgindir: ΔDice = 18,27, Cohen's d = 0.98. Bu değer Cohen'in büyük etki eşiğini (≥0.80) aştığından, nnU-Net'in 3D U-Net'e göre klinik açıdan anlamlı bir üstünlük sağladığı söylenebilir. SwinUNETR ile 3D U-Net arasındaki fark ise ΔDice = 12,10, d = 0.76 ile orta düzeyde bir etki göstermektedir.

Her model için en iyi performans gösteren giriş konfigürasyonu referans alınarak, diğer konfigürasyonlarla eşleştirilmiş karşılaştırmalar yapılmıştır. Sonuçlar Tablo 3.6'da özetlenmektedir.

Tablo 3.6. Her model için en iyi konfigürasyon ile diğer konfigürasyonlar arasındaki Dice skoru farkları.

Model	Karşılaştırma	Δ Dice	Cohen's d	%95 GA	Etki
nnU-Net	DWI+ADC vs DWI	+0.0176	+0.17	[−0.00, +0.34]	İhmal ed.
nnU-Net	DWI+ADC vs DWI+ADC+Edge	+0.0219	+0.22	[+0.06, +0.38]	Küçük
nnU-Net	DWI+ADC vs DWI+ADC+Prior	+0.0287	+0.33	[+0.17, +0.49]	Küçük
nnU-Net	DWI+ADC vs Full	+0.0193	+0.25	[+0.09, +0.41]	Küçük
SwinUNETR	Full vs DWI	+0.0192	+0.14	[−0.02, +0.30]	İhmal ed.
SwinUNETR	Full vs DWI+ADC	+0.0041	+0.05	[−0.11, +0.21]	İhmal ed.
SwinUNETR	Full vs DWI+ADC+Edge	+0.0037	+0.04	[−0.12, +0.20]	İhmal ed.
SwinUNETR	Full vs DWI+ADC+Prior	+0.0066	+0.10	[−0.06, +0.26]	İhmal ed.
3D U-Net	Full vs DWI	+0.0023	+0.03	[−0.13, +0.19]	İhmal ed.
3D U-Net	Full vs DWI+ADC	+0.0264	+0.25	[+0.09, +0.41]	Küçük
3D U-Net	Full vs DWI+ADC+Edge	+0.0087	+0.13	[−0.03, +0.29]	İhmal ed.
3D U-Net	Full vs DWI+ADC+Prior	+0.0090	+0.10	[−0.06, +0.26]	İhmal ed.

Tablo 3.6'daki sonuçlar önemli bir bulguyu ortaya koymaktadır: giriş konfigürasyonları arasındaki farklar, modeller arası farklarla kıyaslandığında çok daha küçüktür. nnU-Net için en büyük etki büyüklüğü d = 0,33 (küçük) iken, modeller arası en küçük etki büyüklüğü d = 0,60 (orta) olarak gözlenmiştir.

SwinUNETR ve 3D U-Net modelleri için giriş konfigürasyonları arasındaki farklar istatistiksel olarak anlamsızdır (tüm güven aralıkları sıfırı içermektedir). Bu durum, bu iki mimarinin ek kanal bilgisinden (Edge, Prior) yeterince faydalanmadığını göstermektedir. Buna karşın, nnU-Net'te DWI+ADC konfigürasyonu DWI+ADC+PRIOR ve DWI+ADC+EDGE konfigürasyonlarına göre istatistiksel olarak anlamlı biçimde üstün performans sergilemiştir; ancak etki büyüklükleri küçük düzeyde kalmıştır (d = 0,22–0,33).

Bu bulgular, iskemik inme lezyon segmentasyonunda mimari seçiminin giriş konfigürasyonundan daha belirleyici bir faktör olduğunu göstermektedir. Edge haritası ve atlas prior gibi ek bilgi kanalları, beklenilenin aksine, performansı tutarlı biçimde artırmamıştır.

DWI+ADC konfigürasyonunda IoU, kesinlik ve duyarlılık metrikleri için modeller arası Cohen's d değerleri Tablo 3.7'de sunulmaktadır.

Tablo 3.7. DWI+ADC konfigürasyonunda diğer metrikler için modeller arası Cohen's d değerleri.

Metrik	Karşılaştırma	Δ	Cohen's d	Etki
IoU	nnU-Net vs SwinUNETR	+0,0652	+0,64	Orta
IoU	nnU-Net vs 3D U-Net	+0,1758	+1,15	Büyük
IoU	SwinUNETR vs 3D U-Net	+0,1106	+0,86	Büyük
Kesinlik	nnU-Net vs SwinUNETR	+0,0266	+0,21	Küçük
Kesinlik	nnU-Net vs 3D U-Net	+0,1144	+0,47	Küçük
Kesinlik	SwinUNETR vs 3D U-Net	+0,0879	+0,37	Küçük
Duyarlılık	nnU-Net vs SwinUNETR	+0,0795	+0,52	Orta
Duyarlılık	nnU-Net vs 3D U-Net	+0,2022	+0,87	Büyük
Duyarlılık	SwinUNETR vs 3D U-Net	+0,1227	+0,73	Orta

IoU metriğinde nnU-Net–3D U-Net farkı d = 1,15 ile en büyük etki büyüklüğünü göstermiştir. Duyarlılık metriğinde de benzer bir eğilim gözlenmiş olup (d = 0,87, büyük etki), nnU-Net'in lezyon bölgelerini daha eksiksiz tespit ettiği anlaşılmaktadır. Kesinlik metriğinde ise farklar nispeten daha küçüktür (d = 0,21–0,47, küçük etki), bu da modeller arasındaki temel farkın kesinlikten ziyade duyarlılık boyutunda olduğunu ortaya koymaktadır.

3.2. Özellik Çıkarımı ve Anatomik Eşleme Sonuçları

Bu tez çalışmasında radyoloji arşivinden retrospektif olarak elde edilen toplam 150 akut iskemik inme hastasına ait veriler kullanılmıştır. Her hasta için DWI, ADC haritaları, nnU-NET modeli ile üretilmiş ve uzman radyolog tarafından doğrulanmış üç boyutlu lezyon segmentasyon maskeleri ile serbest metin formatındaki radyoloji raporları toplanmıştır. Bu çok bileşenli veri yapısı hem görüntü tabanlı analiz hem de LLM tabanlı otomatik rapor üretimi süreçleri için kapsamlı bir veri altyapısı sağlamaktadır.

LLM' in ince ayar sürecinde kullanılacak veri seti oluşturulurken, modelin öğrenme sürecini olumsuz etkileyebilecek tutarsız örneklerin elenmesi amacıyla belirli dışlama kriterleri uygulanmıştır. Bu kapsamda, özellikle giriş (segmentasyon çıktıları) ile hedef (radyoloji raporu) arasındaki anlamsal uyumsuzluklar detaylı olarak incelenmiş ve model performansını düşürebilecek örnekler veri setinden çıkarılmıştır. Bu nedenle en kritik dışlama kriterlerinden biri segmentasyon–rapor tutarsızlığı olarak belirlenmiştir. Bazı vakalarda, segmentasyon modeli anlamlı bir lezyon tespit etmiş olmasına rağmen, ilgili

radyoloji raporunda akut iskemik bulgu bulunmadığı ifade edilmiştir. Bu durum, modelin öğrenme sürecinde çelişkili örneklerle karşılaşmasına neden olarak hatalı genellemeler yapmasına yol açabilmektedir.

Bu tür tutarsızlıkların iki temel nedenden kaynaklandığı gözlemlenmiştir. İlk olarak, görüntüleme sırasında oluşan artefaktlar (örneğin hasta hareketi veya manyetik alan inhomojenitesi) nedeniyle radyolog tarafından incelemenin "değerlendirilemedi" şeklinde raporlanması söz konusu olabilmektedir. İkinci olarak ise, segmentasyon çıktıları ile radyoloji raporlarının farklı zamanlara veya farklı görüntüleme serilerine ait olması, yani veri eşleşmesinde zamansal veya seri bazlı uyumsuzluklar bulunması bu tür çelişkilere neden olabilmektedir. Bu tip örneklerde, giriş verisi olan segmentasyon çıktısı ile hedef çıktı olan rapor metni arasında doğrudan bir uyumsuzluk oluşmaktadır. Bu durum, modelin pozitif bir segmentasyon girdisine karşılık negatif içerikli bir rapor üretmeyi öğrenmesine neden olarak, özellikle klinik karar destek sistemi olarak tasarlanan yapıda istenmeyen sonuçlara yol açabilmektedir. Bu nedenle, segmentasyonun pozitif olduğu durumlarda raporun da klinik olarak tutarlı ve pozitif içerik üretmesi hedeflenmiş, bu kriteri sağlamayan örnekler veri setinden çıkarılmıştır.

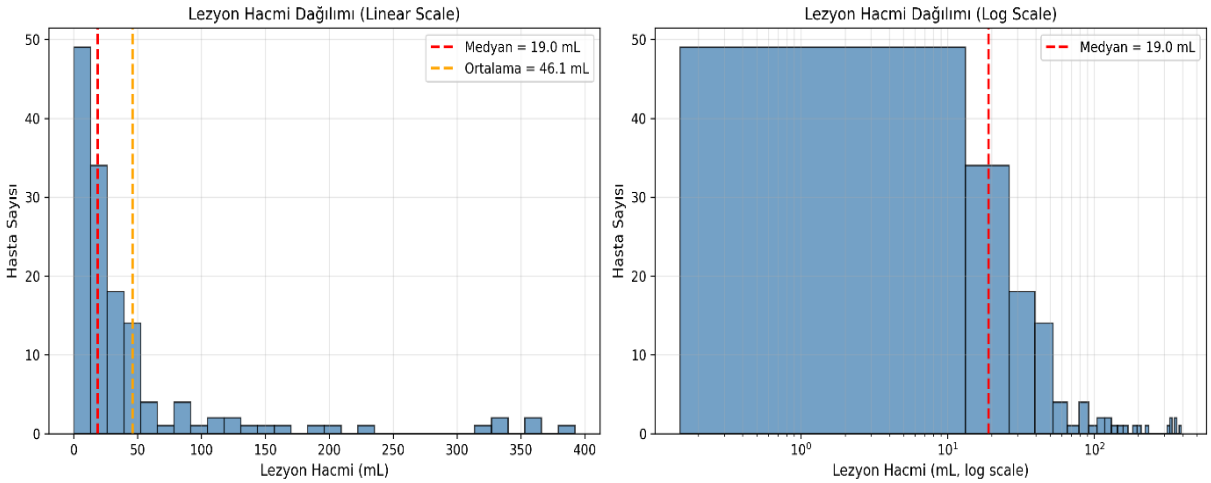
Negatif içerikli raporların otomatik olarak tespit edilebilmesi amacıyla, Türkçe diline özgü ifadeleri kapsayan düzenli ifade (regex) tabanlı bir filtre geliştirilmiştir. Bu filtre kapsamında, "akut iskemi/enfarkt izlenmedi", "difüzyon kısıtlanması saptanmadı", "iskemiyi düşündürecek görünüm izlenmedi" gibi klinik olarak negatif anlam taşıyan ifade kalıpları belirlenmiş ve bu tür metinleri içeren örnekler veri setinden sistematik olarak elenmiştir.

Tablo 3.8. Eğitim ve Doğrulama Seti Karakteristikleri

Özellik	Eğitim (n=120)	Doğrulama (n=21)
Toplam hasta	120	21
Sol hemisfer	~%50	~%50
Sağ hemisfer	~%44	~%44
Bilateral	~%6	~%6
Anatomik etiket içeren (Harvard-Oxford)	%84	%84
Vasküler territory etiketli	%90	%90
ADC değeri içeren	%67	%67

Uygulanan dışlama kriterleri sonucunda (Tablo 3.8), başlangıçta yer alan toplam 150 hastadan 9'u veri tutarsızlıkları nedeniyle LLM fine tune kapsamı dışında bırakılmış ve analizler 141 hasta üzerinden gerçekleştirilmiştir. Eksik dosya ve yetersiz rapor uzunluğu kriterlerine uyan herhangi bir örnek bulunmazken, tüm dışlamalar segmentasyon–rapor tutarsızlığına dayalı negatif rapor filtrelemesi sonucunda yapılmıştır. Elde edilen nihai veri seti 141 hastadan oluşmakta olup, bu veri seti sabit bir rastgele tohum (seed=42) kullanılarak eğitim (n=120) ve doğrulama (n=21) olmak üzere iki alt kümeye ayrılmıştır. Bu yapı sayesinde modelin eğitimi ve performans değerlendirmesi kontrollü ve tekrarlanabilir bir şekilde gerçekleştirilmiştir.

Veri kümesindeki 141 hastanın lezyon hacimleri çok geniş bir aralıkta dağılım göstermektedir (Tablo 3.9, Şekil 3.13). Ortalama lezyon hacmi 46,10 ± 76,09 mL olarak hesaplanmış; ancak yüksek standart sapma değeri, hacim dağılımının belirgin biçimde sağa çarpık olduğunu ve aritmetik ortalamanın merkezi eğilimi yeterince yansıtmadığını göstermektedir. Bu nedenle medyan değeri (19,02 mL) ve çeyrekler arası genişlik (IQR: 7,60 – 42,49 mL) merkezi eğilim ölçütü olarak daha uygun bulunmaktadır.



Şekil 3.13. 141 hastaya ait lezyon hacim dağılımı.

Veri kümesindeki en küçük lezyon 0,15 mL, en büyük lezyon ise 392,27 mL hacme sahiptir; bu yelpaze, küçük lakuner enfarktlardan kapsamlı territorial enfarktlara kadar uzanan tüm klinik şiddet spektrumunu içermektedir.

Tablo 3.9. Lezyon Hacmi İstatistikleri (n=141)

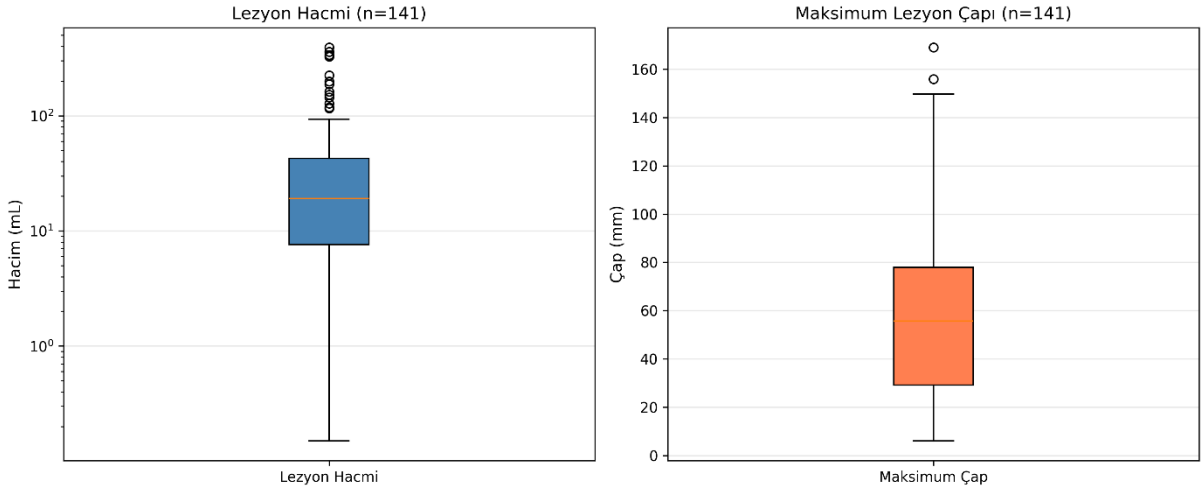
İstatistik	Değer (mL)
Ortalama ± Standart Sapma	46,10 ± 76,09
Medyan	19,02
Q1 (25. yüzdelik)	7,60
Q3 (75. yüzdelik)	42,49
Çeyrekler arası genişlik (IQR)	34,89
Minimum	0,15
Maksimum	392,27

Lezyonların aksiyel düzlemdeki maksimum çapı, segmentasyon maskesi içindeki bağlı bileşenlerin her aksiyel kesitinde ölçülen en uzun mesafenin millimetre cinsinden hesaplanması yoluyla belirlenmiştir. Veri kümesindeki 141 hastanın maksimum lezyon çapları için ortalama 59,45 ± 35,71 mm, medyan 55,60 mm (IQR: 29,20 – 77,90 mm), minimum 6,00 mm ve maksimum 169,10 mm olarak ölçülmüştür (Tablo 3.10).

Tablo 3.10. Maksimum Lezyon Çapı İstatistikleri (n=141)

İstatistik	Değer (mm)
Ortalama ± Standart Sapma	59,45 ± 35,71
Medyan	55,60
Q1 (25. yüzdelik)	29,20
Q3 (75. yüzdelik)	77,90
Minimum	6,00
Maksimum	169,10

Akut iskemik inme bazı hastalarda tek bir lezyon olarak görülürken, multipl emboli veya watershed enfarktı gibi klinik tablolarda birden fazla ayrık lezyon halinde de izlenebilir. Veri kümesindeki bu çeşitliliği değerlendirmek üzere her segmentasyon maskesi, birbirine komşu vokselleri tek bir lezyon olarak gruplayan bir bağlı bileşen analizinden geçirilmiştir. Üç boyutlu uzayda komşuluk ilişkisi, ortak yüz, kenar veya köşe paylaşımını kapsayan 26-bağlantı kuralı ile tanımlanmıştır.



Şekil 3.14. Veri setindeki lezyon hacmi ve maksimum lezyon çapı dağılımlarının kutu grafikleri.

Veri kümesindeki 141 hastanın 45'inde (%31,9) tek lezyon, 96'sında (%68,1) çoklu lezyon mevcuttur. Çoklu lezyonlu vakalarda ortalama lezyon sayısı 5,6, medyan değer 3, maksimum değer ise 46 olarak hesaplanmıştır. Çalışma popülasyonunun büyük çoğunluğunun (yaklaşık üçte ikisinin) multifokal lezyon içerdiği bu bulgu, akut iskemik inmenin gerçek dünyadaki klinik bulgular ile uyumludur ve modelimizin tek lezyon - çoklu lezyon ayrımını öğrenmesi için yeterli çeşitlilikte örnek sağladığını göstermektedir. Veri setinde yer alan lezyon hacmi ve maksimum lezyon çapına ilişkin dağılımlar, aykırı değerler ve merkezi eğilim ölçüleri ile Şekil 3.13'te sunulmuştur.

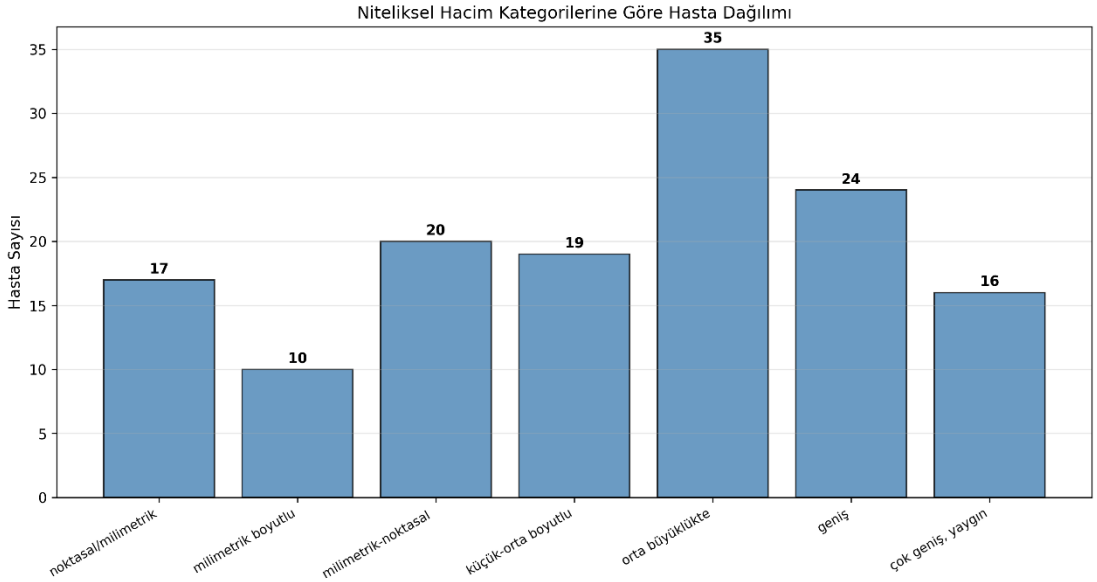
Klinik radyoloji raporlarında lezyon büyüklüğü tipik olarak mililitre cinsinden değil, niteliksel ifadeler (örn. "milimetrik", "küçük-orta boyutlu", "geniş", "yaygın") ile belirtilmektedir. Bu nedenle veri kümesindeki tüm hastalar, Bölüm 2'de Tablo 2.6'da tanımlanan ve veri-tabanlı eşiklerle kalibre edilen niteliksel sınıflandırma şemasına göre kategorize edilmiştir.

Tablo 3.11, 141 hastanın yedi farklı kategoriye dağılımını göstermektedir.

Tablo 3.11. Niteliksel hacim kategorilerine göre hasta dağılımı.

Kategori	**Hacim Aralığı**	**Hasta Sayısı**	**Yüzde**
Noktasal / milimetrik	< 2 mL	17	%12,1
Milimetrik boyutlu	2 – 5 mL	10	%7,1
Milimetrik-noktasal odaklar*	< 30 mL	20	%14,2
Küçük-orta boyutlu	5 – 15 mL	19	%13,5
Orta büyüklükte	15 – 40 mL	35	%24,8
Geniş	40 – 100 mL	24	%17,0
Çok geniş, yaygın	≥ 100 mL	16	%11,3
TOPLAM	–	**141**	**%100,0**

"Milimetrik-noktasal odaklar" kategorisi, lezyon sayısı ≥ 5 ve toplam hacim < 30 mL olan multifokal vakalar için ayrılmış özel bir kategoridir; radyologların bu klinik tabloyu "multipl milimetrik odak" olarak ifade etme eğilimini yakalamak için tanımlanmıştır.



Şekil 3.15. Veri setindeki lezyon hacim kategorilerinin dağılımı.

Veri kümesinde en yoğun temsil edilen kategori "orta büyüklükte" lezyonlar (%24,8)'dir; bunu sırasıyla "geniş" (%17,0), "milimetrik-noktasal odaklar" (%14,2), "küçük-orta boyutlu" (%13,5), "noktasal/milimetrik" (%12,1), "çok geniş, yaygın" (%11,3) ve "milimetrik boyutlu" (%7,1) kategorileri izlemektedir. Bu görece dengeli dağılım, modelin

tüm boyut kategorilerinde yeterli eğitim örneği görmesini sağlamakta ve genelleme yeteneğini desteklemektedir.

Lezyon yerleşim tarafı bilgisi, segmentasyon maskesinin RAS formatına dönüştürülmüş hacim üzerinden hesaplanmıştır. Hacmin orta-sagital düzlemine göre sağ ve sol hemisferdeki voksel sayıları karşılaştırılmış; bir hemisferdeki voksel sayısı diğerinin 1,5 katından fazla olduğunda lezyon o hemisfere atanmış, aksi halde bilateral olarak işaretlenmiştir.

Tablo 3.12. Lezyon yerleşim yeri dağılımı.

Yerleşim Tarafı	Hasta Sayısı	Yüzde
Sol hemisfer	71	%50,4
Sağ hemisfer	62	%44,0
Bilateral	8	%5,7

Sol-sağ hemisfer dağılımı görece dengelidir (%50,4 vs %44,0). Sol hemisfer baskınlığı (%6,4 fark) literatürde rapor edilen akut iskemik inme epidemiyolojisi ile uyumludur. Bilateral lezyonlu hasta oranı (%5,7) ise embolik kaynaklı multifokal inme olgularını yansıtmaktadır (Tablo 3.12).

Lezyonların hangi serebral arter sulama alanına ait olduğunu belirlemek üzere, MNI tabanlı bir vasküler territory atlası kullanılmıştır. Atlas, beyni dört ana vasküler bölgeye ayırmaktadır: Anterior Serebral Arter (ACA), Orta Serebral Arter (MCA), Posterior Serebral Arter (PCA) ve Vertebrobaziler sistem.

Tablo 3.13. Vasküler sulama alanına göre lezyon dağılımı

Yerleşim Tarafı	Hasta Sayısı	Yüzde
Orta Serebral Arter (MCA)	76	%53,9
Anterior Serebral Arter (ACA)	42	%29,8
Posterior Serebral Arter (PCA)	9	%6,4
Belirtilmemiş	14	%9,9

"Belirtilmemiş" kategorisi, lezyon vokselleri atlasta tanımlanmış vasküler bölgelerin dışına düştüğünde (genellikle çok küçük veya periferik lezyonlar) atanmaktadır. Beklenen biçimde MCA sulama alanı en sık etkilenen vasküler bölgedir (%53,9), bunu ACA (%29,8) ve PCA (%6,4) izlemektedir. Bu dağılım, akut iskemik inme literatüründe rapor edilen oranlarla uyumludur: MCA tüm iskemik inmelerin yaklaşık yarısını kapsayan en sık etkilenen damarlardır; ACA ise daha düşük oranda görülür. Veri kümesindeki ACA oranının görece yüksek olmasının nedeni, atlasın watershed bölgelerinde bazı kortikal alanları sınır olarak ACA'ya atamasıdır.

3.3. Otomatik Rapor Üretim Sonuçları

Bu bölümde, Bölüm 2.13'te açıklanan QLoRA tabanlı ince ayar sürecinden geçirilen altı büyük dil modelinin Türkçe radyoloji rapor üretim performansı sunulmaktadır.

Değerlendirme, doğrulama setindeki 21 hasta üzerinde gerçekleştirilmiş olup sonuçlar Bölüm 2.14'te tanımlanan metrikler kullanılarak raporlanmıştır.

3.3.1. BLEU Metrikleri

Altı modelin BLEU-1, BLEU-2, BLEU-3 ve BLEU-4 düzeylerindeki performansları Tablo 3.14'te sunulmaktadır. BLEU-1 tekil kelime eşleşmesini, BLEU-4 ise dört kelimelik ardışık dizilerin doğruluğunu ölçmektedir.

Tablo 3.14. Modellerin BLEU skorları karşılaştırması.

Model	Parametre	Dil	Aile	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Qwen 2.5 3B	3B	Multilingual	Qwen	0,4659	0,3213	0,2527	0,2149
Qwen 2.5 7B	7B	Multilingual	Qwen	0,4471	0,3055	0,2463	0,211
Qwen 2.5 14B	14B	Multilingual	Qwen	0,2903	0,1389	0,0997	0,0846
Mistral 7B	7B	Multilingual	Mistral	**0,5505**	**0,4631**	**0,4171**	**0,3808**
Kumru 2B	2B	Turkish-native	Kumru	0,0984	0,0585	0,0434	0,0322
Cosmos 8B	8B	Turkish fine-tuned	LLaMA	0,0259	0,0007	0	0

BLEU metrikleri açısından Mistral 7B modeli tüm n-gram düzeylerinde en yüksek performansı göstermiştir. Mistral 7B'nin BLEU-1 skoru 0,5505, BLEU-4 skoru ise 0,3808 olarak elde edilmiştir. Bu sonuç, Mistral 7B'nin ürettiği raporların kelime ve ifade düzeyinde referans raporlarla en yüksek örtüşmeyi sağladığını göstermektedir. Qwen ailesinde 3B ve 7B modelleri birbirine yakın BLEU-4 değerleri (sırasıyla 0,2149 ve 0,2110) üretirken, 14B modeli 0,0846 ile belirgin şekilde düşük kalmıştır.

Türkçe-native modellerin BLEU performansı oldukça düşük düzeydedir. Kumru 2B'nin BLEU-4 skoru 0,0322, Cosmos 8B'nin ise 0,0000 olarak hesaplanmıştır. Cosmos 8B modelinin BLEU-3 ve BLEU-4 değerlerinin sıfır olması, modelin ürettiği raporların referans raporlarla üç ve dört kelimelik ardışık dizi düzeyinde hiçbir eşleşme içermediğini göstermektedir.

BLEU-1'den BLEU-4'e doğru tüm modellerde beklenen düşüş gözlenmektedir. Bu düşüş, daha uzun n-gram dizilerinin eşleşmesinin doğal olarak daha zor olmasından kaynaklanmaktadır. Ancak Mistral 7B'de bu düşüş oranı (BLEU-1: 0,5505 → BLEU-4: 0,3808, %31 düşüş) diğer modellere kıyasla daha düşüktür. Qwen 2.5 3B'de ise bu düşüş %54 (0,4659 → 0,2149) seviyesindedir. Bu fark, Mistral 7B'nin yalnızca kelime düzeyinde değil, ifade düzeyinde de referans raporlarla daha tutarlı metinler ürettiğini göstermektedir.

3.3.2. ROUGE Metrikleri

Altı modelin ROUGE-1, ROUGE-2 ve ROUGE-L skorları Tablo 3.15'te sunulmaktadır. ROUGE-1 tekil kelime düzeyinde duyarlılığı, ROUGE-2 bigram düzeyinde duyarlılığı, ROUGE-L ise en uzun ortak alt dizi yöntemini kullanarak yapısal benzerliği ölçmektedir.

ROUGE metrikleri açısından Qwen 2.5 3B modeli tüm düzeylerde en yüksek performansı sergilemiştir. Qwen 2.5 3B'nin ROUGE-1 skoru 0,5300, ROUGE-L skoru ise 0,5129 olarak

elde edilmiştir. Bu sonuç, BLEU metriklerindeki sıralamadan farklı bir tablo ortaya koymaktadır: BLEU'da Mistral 7B liderken, ROUGE'da Qwen 2.5 3B liderdir.

Tablo 3.15. Modellerin ROUGE skorları karşılaştırması.

Model	Parametre	ROUGE-1	ROUGE -2	ROUGE -L
Qwen 2.5 3B	3B	**0,53**	**0,3945**	**0,5129**
Qwen 2.5 7B	7B	0,4467	0,2945	0,4313
Qwen 2.5 14B	14B	0,3391	0,1979	0,3295
Mistral 7B	7B	0,3712	0,2789	0,3403
Kumru 2B	2B	0,1969	0,1267	0,1569
Cosmos 8B	8B	0,1146	0,0138	0,076

Bu farklılık, iki metrik ailesinin ölçtüğü boyutlardan kaynaklanmaktadır. BLEU kesinlik odaklı bir metriktir ve üretilen metindeki n-gramların referansta ne oranda bulunduğunu ölçmektedir. ROUGE ise duyarlılık odaklıdır ve referans metindeki bilgilerin üretilen raporda ne kadarının yakalandığını değerlendirmektedir. Buna göre, Mistral 7B'nin ürettiği raporlar kelime seçimi açısından daha isabetli ancak kapsam olarak daha dar; Qwen 2.5 3B'nin raporları ise referans rapordaki klinik bilgileri daha kapsamlı biçimde yakalayan ancak ek ifadeler de içeren metinlerdir.

Qwen ailesinde ROUGE metrikleri de BLEU ile benzer şekilde 3B > 7B > 14B sıralamasını göstermektedir. Qwen 2.5 14B'nin ROUGE-L skoru (0,3295), Mistral 7B'nin skoruyla (0,3403) benzer düzeyde olup her ikisi de Qwen 2.5 3B'nin (0,5129) önemli ölçüde gerisindedir.

3.3.3. Genel Performans Karşılaştırması

Altı modelin tüm otomatik NLP metrikleri, klinik doğruluk metrikleri ve üretim kalitesi metrikleri Tablo 3.16'da özet olarak sunulmaktadır.

Genel performans tablosu incelendiğinde, tek bir modelin tüm metriklerde üstün olmadığı görülmektedir. Qwen 2.5 3B modeli ROUGE-L (0,5129), METEOR (0,4526) ve BERTScore (0,7630) metriklerinde en yüksek değerleri elde ederek semantik benzerlik ve kapsam açısından en güçlü model olmuştur. Mistral 7B ise BLEU-4 (0,3808) metriğinde açık farkla lider olup kelime düzeyinde kesinlik açısından en başarılı modeldir. Qwen 2.5 7B, yerleşim tarafı doğruluğunda (0,7143) en yüksek skoru elde etmiştir.

Qwen 2.5 ailesindeki üç farklı boyuttaki modelin (3B, 7B, 14B) karşılaştırması, aynı mimari aile içinde yalnızca parametre sayısının değiştiği kontrollü bir deney ortamı sunmaktadır. Sonuçlar, model boyutu arttıkça performansın iyileşmediğini; aksine, genel bir düşüş

eğilimi gösterdiğini ortaya koymaktadır. BLEU-4 değerleri 3B'de 0,2149, 7B'de 0,2110 ve 14B'de 0,0846 olarak elde edilmiştir. BERTScore değerleri de benzer bir örüntü izleyerek 3B'de 0,7630, 7B'de 0,7422 ve 14B'de 0,6954 olarak hesaplanmıştır.

Tablo 3.16. Altı modelin genel performans karşılaştırması.

Model	BLEU-4	ROUGE-L	METEOR	BERTScore	Yerleşim tarafı	Format_B	Format_S	Tekrar	Uzunluk
Qwen 2.5 3B	0,2149	**0,5129**	**0,4526**	**0,763**	0,5238	**1**	**1**	0	0,8951
Qwen 2.5 7B	0,211	0,4313	0,429	0,7422	**0,7143**	0,9524	**1**	0	**0,9033**
Qwen 2.5 14B	0,0846	0,3295	0,2357	0,6954	0,6667	0,2857	**1**	0	0,5491
Mistral 7B	**0,3808**	0,3403	0,39	0,6538	0,5714	**1**	0,2381	0	0,7752
Kumru 2B	0,0322	0,1569	0,2121	0,5864	0,381	**1**	**1**	0	4,7808
Cosmos 8B	0	0,076	0,0469	0,5301	0,0952	0,619	0,9048	0	3,2941

Bu sonuç, küçük veri setleriyle (120 eğitim örneği) gerçekleştirilen QLoRA ince ayarında büyük modellerin daha küçük modellere göre dezavantajlı olduğunu göstermektedir. Bu durumun olası nedenleri şunlardır: birincisi, büyük modellerin sahip olduğu geniş parametre uzayı sınırlı sayıda eğitim örneğiyle yeterince yönlendirilememektedir. İkincisi, düşük rütbeli adaptör matrislerinin (r=16) toplam parametre sayısına oranı büyük modellerde daha düşük kalmakta ve bu durum modelin görev-spesifik adaptasyon kapasitesini sınırlamaktadır. Üçüncüsü, 14B modelin 4-bit nicemleme ile 12 GB VRAM'e sığdırılması, hesaplama hassasiyetinde ek kayıplara yol açmış olabilir. Öte yandan, yerleşim tarafı doğruluğu açısından 7B model (0,7143), 3B modelden (0,5238) belirgin şekilde üstündür. Bu bulgu, daha büyük modellerin ön eğitim aşamasında edindiği daha kapsamlı dil anlayışının belirli klinik görevlerde avantaj sağlayabildiğini düşündürmektedir.

Aynı parametre ölçeğindeki iki çok dilli modelin karşılaştırması, mimari tasarım farklılıklarının rapor üretim performansına etkisini ortaya koymaktadır.

Mistral 7B, BLEU metriklerinin tüm düzeylerinde Qwen 2.5 7B'yi önemli farkla geçmiştir (BLEU-4: 0,3808'e karşı 0,2110). Ancak Qwen 2.5 7B, ROUGE-L (0,4313'e karşı 0,3403), METEOR (0,4290'a karşı 0,3900) ve BERTScore (0,7422'ye karşı 0,6538) metriklerinde üstün performans sergilemiştir. Bu zıt yönlü sonuç, iki modelin farklı üretim stratejileri benimsediğine işaret etmektedir.

Mistral 7B'nin yüksek BLEU ancak düşük ROUGE performansı, modelin kısa ve referans metinle kelime düzeyinde yüksek örtüşme gösteren raporlar ürettiğini göstermektedir. Nitekim uzunluk oranı verileri bu yorumu desteklemektedir: Mistral 7B'nin uzunluk oranı 0,7752 iken Qwen 2.5 7B'nin uzunluk oranı 0,9033'tür. Qwen 2.5 7B, referans raporlara daha yakın uzunlukta metinler üreterek daha yüksek kapsam sağlamıştır.

Bir diğer önemli fark, format doğruluğunda gözlenmektedir. Mistral 7B'nin "Sonuç" bölümü format doğruluğu yalnızca 0,2381 iken Qwen 2.5 7B'de bu değer 1,0000'dir. Bu

bulgu, Mistral 7B'nin "Bulgular" bölümünü tutarlı şekilde üretirken "Sonuç" bölümünü çoğu raporda atlama eğiliminde olduğunu göstermektedir.

Segmentasyon Bulguları (Model Girdisi):

Lezyon sayısı	2
Lezyon büyüklüğü	noktasal/milimetrik difüzyon kısıtlanması
Yerleşim tarafı	sağ hemisfer
Kortikal bölge	parietal operkülum korteksi
Subkortikal bölge	sağ serebral korteks, sağ serebral beyaz cevher, sağ hipokampus
Vasküler sulama alanı	MCA

Ground Truth (Uzman Raporu)	**Qwen 2.5 3B Tahmini**
BULGULAR: Kranial difüzyon MR incelemesinde; Pons sag yariminda akut enfarkt ile uyumlu difüzyon kısıtlılığı gösteren odaklar izlendi. SONUÇ/ÖNERİLER: Bulgular klinik, fizik muayene ve laboratuvar verileri ışığında değerlendirilmelidir.	BULGULAR: Kranial difüzyonal MR incelemesinde; Sagda bazal ganglionlar düzeyinde akut enfarkt ile uyumlu difüzyona kısıtlaması gösteren alan izlenmiştir. SONUÇ / ÖNERİLER; Bulgular klinik, fizik muayene ve laboratuvar verileri ışığında değerlendirilmelidir.

Türkçe'ye özgü olarak eğitilmiş Kumru 2B ve Türkçe ince ayar yapılmış Cosmos 8B modellerinin performansı, çok dilli modellerle karşılaştırıldığında belirgin şekilde düşük kalmıştır.

Kumru 2B modeli, BLEU-4 (0,0322), ROUGE-L (0,1569), METEOR (0,2121) ve BERTScore (0,5864) metriklerinin tümünde en düşük ikinci performansı göstermiştir. Cosmos 8B ise tüm metriklerde en düşük performansı sergilemiş olup BLEU-3 ve BLEU-4 değerleri 0,0000 olarak hesaplanmıştır. Cosmos 8B'nin BERTScore değeri 0,5301 ile tüm modeller arasında en düşük semantik benzerliği yansıtmaktadır.

Bu düşük performansın en dikkat çekici yansıması uzunluk oranlarında görülmektedir. Kumru 2B'nin uzunluk oranı 4,7808, Cosmos 8B'nin ise 3,2941 olarak hesaplanmıştır. Bu değerler, her iki modelin referans raporların sırasıyla yaklaşık 5 ve 3 katı uzunluğunda metinler ürettiğini göstermektedir. Bu aşırı uzun çıktılar, modellerin radyoloji rapor formatını yeterince öğrenemediğine ve tekrarlayan veya ilgisiz içerik üreterek raporları gereksiz yere uzattığına işaret etmektedir. Tekrar oranının tüm modellerde 0,0000 olarak hesaplanması, tekrar engelleme parametrelerinin (no_repeat_ngram_size) etkili çalıştığını göstermekle birlikte, uzunluk problemi farklı ifadelerle benzer içeriğin tekrarlanmasından kaynaklanıyor olabilir.

Yerleşim tarafı doğruluğu açısından da Türkçe-native modeller düşük performans sergilemiştir. Kumru 2B'nin yerleşim tarafı doğruluğu 0,3810, Cosmos 8B'nin ise yalnızca

0,0952 olarak hesaplanmıştır. Cosmos 8B'nin %10'un altındaki yerleşim tarafı doğruluğu, modelin hastanın lezyon tarafını doğru bildirme konusunda neredeyse tamamen başarısız olduğunu göstermektedir.

Bu sonuçlar, Türkçe genel dil bilgisinin tıbbi radyoloji rapor üretimi için yeterli olmadığını ortaya koymaktadır. Çok dilli modellerin daha geniş ve çeşitli ön eğitim verilerinden edindiği tıbbi bilgi birikimi ve yapılandırılmış metin üretim kapasitesi, alana özgü ince ayarda belirleyici bir avantaj sağlamaktadır.

Klinik doğruluk metrikleri, modellerin rapor üretiminin klinik açıdan anlamlılığını ortaya koymaktadır. Yerleşim tarafı doğruluğunda en yüksek performansı Qwen 2.5 7B (0,7143) göstermiştir. Bu model, raporların yaklaşık %71'inde hastanın lezyon tarafını doğru olarak bildirmiştir. Qwen 2.5 14B (0,6667), Mistral 7B (0,5714) ve Qwen 2.5 3B (0,5238) orta düzey yerleşim tarafı doğruluğu sağlarken, Kumru 2B (0,3810) ve Cosmos 8B (0,0952) düşük performans göstermiştir.

Rapor format uyumu iki boyutta değerlendirilmiştir: "Bulgular" bölümü (Format_B) ve "Sonuç" bölümü (Format_S) varlığı. Qwen 2.5 3B, her iki format boyutunda da tam uyum (1,0000) sağlayan tek modeldir. Mistral 7B, "Bulgular" bölümünde %100 uyum gösterirken "Sonuç" bölümünde yalnızca %24 uyum sağlamıştır. Bu durum, Mistral 7B'nin raporlarını çoğunlukla yalnızca bulgularla sınırlı tutarak sonuç ve öneri kısmını atladığını göstermektedir. Qwen 2.5 14B ise tersine, "Sonuç" bölümünde %100 uyum gösterirken "Bulgular" bölümünde yalnızca %29 uyum sağlamıştır.

Tüm modellerde tekrar oranı 0,0000 olarak hesaplanmıştır. Bu sonuç, çıkarım aşamasında uygulanan tekrar engelleme stratejilerinin (no_repeat_ngram_size parametresi) etkin bir şekilde çalıştığını doğrulamaktadır.

Uzunluk oranı açısından referans raporlara en yakın değerleri Qwen 2.5 7B (0,9033) ve Qwen 2.5 3B (0,8951) modelleri vermiştir. Bu iki model, referans raporların yaklaşık %90'ı uzunluğunda metinler üreterek hem yeterli bilgi kapsamı hem de gereksiz uzunluktan kaçınma arasında iyi bir denge kurmuştur. Mistral 7B'nin uzunluk oranı (0,7752) biraz düşük kalırken, Qwen 2.5 14B'nin oranı (0,5491) referans raporların yarısı kadar kısa raporlar üretildiğine işaret etmektedir.

Altı modelin çok boyutlu karşılaştırması, tek bir modelin tüm değerlendirme boyutlarında kesin üstünlük sağlamadığını ortaya koymaktadır. Ancak genel performans profilleri incelendiğinde modeller üç gruba ayrılabilir.

Birinci grupta Qwen 2.5 3B ve Qwen 2.5 7B yer almaktadır. Bu iki model, semantik benzerlik (BERTScore), kapsam (ROUGE-L), format uyumu ve uzunluk oranı metriklerinde dengeli ve yüksek performans sergilemiştir. Qwen 2.5 3B en yüksek BERTScore (0,7630) ve ROUGE-L (0,5129) değerlerini elde ederken, Qwen 2.5 7B en yüksek yerleşim tarafı doğruluğunu (0,7143) sağlamıştır.

İkinci grupta Mistral 7B yer almaktadır. Bu model, BLEU metriklerinde açık farkla lider olmakla birlikte, düşük ROUGE skoru, düşük format uyumu ve kısa rapor uzunluğu gibi sınırlılıklara sahiptir. Mistral 7B, referans raporlarla kelime düzeyinde yüksek örtüşme gösteren ancak kapsamı sınırlı raporlar üretmektedir.

Üçüncü grupta Qwen 2.5 14B, Kumru 2B ve Cosmos 8B yer almaktadır. Bu modeller, tüm metriklerde düşük veya çok düşük performans göstermiştir. Qwen 2.5 14B'nin düşük performansı sınırlı eğitim verisi ile büyük model boyutu arasındaki uyumsuzluktan, Kumru 2B ve Cosmos 8B'nin düşük performansı ise Türkçe genel dil modellerinin tıbbi rapor üretimi görevine yeterince adapte olamamasından kaynaklanmaktadır.

Oluşturulan otomatik radyolojik raporların klinik uygunluğunu değerlendirmek amacıyla uzman hekim değerlendirmesi gerçekleştirilmiştir. Bu kapsamda raporlar; klinik doğruluk, anatomik uyum, dil akıcılığı, tıbbi tutarlılık ve genel yeterlilik kriterleri açısından 5'li Likert ölçeği kullanılarak değerlendirilmiştir. Uzman değerlendirmeleri sonucunda üretilen raporların büyük çoğunluğunun klinik açıdan yeterli ve doğru bulunduğu gözlemlenmiştir. Özellikle anatomik lokalizasyon ve vasküler bölge bilgilerinin doğru şekilde raporlara yansıtıldığı belirtilmiştir. Ortalama Likert skorlarının tüm değerlendirme boyutlarında yüksek düzeyde olması, önerilen sistemin klinik kullanım açısından umut verici sonuçlar sunduğunu göstermektedir.

Uzman hekim değerlendirmelerinde otomatik üretilen raporlar genel yeterlilik açısından ortalama 4.6/5, klinik doğruluk açısından 4.5/5 ve anatomik uyum açısından 4.7/5 Likert skoru elde etmiştir. Değerlendirmeler, sistem tarafından üretilen raporların büyük ölçüde klinik olarak tutarlı ve kullanılabilir olduğunu göstermiştir.

Tablo 3.17. Likert sonuçları.

Değerlendirme Kriteri	Ortalama Likert Skoru
Klinik doğruluk	4,5
Anatomik uyum	4,7
Dil akıcılığı	4,4
Tıbbi tutarlılık	4,6
Genel yeterlilik	4,6

Uzman hekimlerin verdikleri Likert skorları(Tablo 3.17), yalnızca metin benzerlik metriklerinin değil, aynı zamanda klinik anlamlılığın da sağlandığını göstermektedir. Bu durum, önerilen sistemin yalnızca dilsel benzerlik üreten bir yapı olmadığını, aynı zamanda klinik olarak anlamlı ve kullanılabilir radyolojik raporlar oluşturabildiğini ortaya koymaktadır.

3.4. Tıbbi Metin Sınıflandırması Üzerine Temel Çalışma: HeaRT Çerçevesi

Bu tez çalışmasının temel odağı beyin MR görüntülerinden inme lezyonlarının segmentasyonu ve yapılandırılmış klinik bilgilerin büyük dil modelleri aracılığıyla Türkçe radyolojik raporlara dönüştürülmesidir. Ancak otomatik rapor üretim sisteminin başarılı biçimde geliştirilmesi, tıbbi metinlerin yapısal ve anlamsal özelliklerinin derinlemesine anlaşılmasını gerektirmektedir. Bu gereksinimden yola çıkarak, tez kapsamındaki radyolojik rapor üretim çalışmalarına zemin hazırlamak amacıyla, tıbbi metinlerin analizi ve sınıflandırılması üzerine kapsamlı bir temel çalışma gerçekleştirilmiştir [105].

HeaRT çerçevesinin geliştirilmesi sürecinde elde edilen deneyimler, tezin LLM tabanlı rapor üretim bileşenine birkaç kritik açıdan katkı sağlamıştır. Birincisi, tıbbi terminolojinin karmaşıklığı, eşanlamlılık ve çok anlamlılık gibi doğal dil işleme zorluklarının biyomedikal bağlamda nasıl ele alınabileceğine ilişkin önemli bir bilgi birikimi oluşturulmuştur. İkincisi,

SBERT tabanlı bağlamsal gömme temsillerinin tıbbi metinlerde semantik ilişkileri yakalama kapasitesi sistematik olarak değerlendirilmiş ve bu deneyim, radyolojik raporların otomatik değerlendirmesinde BERTScore metriğinin seçilmesine ve yorumlanmasına rehberlik etmiştir. Üçüncüsü, sınıf dengesizliği ve bağlamsal yetersizlik gibi tıbbi metin işlemede sıklıkla karşılaşılan sorunlara yönelik geliştirilen çözüm stratejileri, LLM ince ayar sürecindeki veri hazırlama ve değerlendirme aşamalarında doğrudan uygulanabilir metodolojik bir temel oluşturmuştur. Bu alt bölümde HeaRT çerçevesinin yöntemi, kullanılan veri setleri ve elde edilen deneysel sonuçlar ayrıntılı olarak sunulmaktadır.

Tıbbi metin sınıflandırma (TMS), sağlık bilişimi alanındaki temel araştırma konularından biri olup, yapılandırılmamış klinik verilerden anlamlı ve eyleme dönüştürülebilir bilgilerin çıkarılmasını amaçlamaktadır. Klinik notlar, elektronik sağlık kayıtları ve biyomedikal literatür gibi kaynaklardan elde edilen metinler; kavramsal karmaşıklık, sınırlı bağlamsal bilgi ve dengesiz sınıf dağılımları gibi özellikler göstermekte ve bu durum sınıflandırma görevlerini teknik açıdan zorlaştırmaktadır. HeaRT, seyrek etiketli ve anlamsal açıdan karmaşık tıbbi metinler için özel olarak tasarlanmış, açıklanabilir ve çok katmanlı bir sınıflandırma mimarisidir. Bildiğimiz kadarıyla bu çalışma, Term Frequency (TF) tabanlı sözcüksel ağırlıklandırma, SBERT bağlamsal gömme temsilleri ve kümeleme tabanlı yapısal öznitelikleri birleşik bir Çok-Vektörlü Öznitelik Füzyonu (Multi-Vector Feature Fusion, MVFF) mimarisi çerçevesinde sistematik olarak bütünleştiren ilk çalışmadır. HeaRT çerçevesinin temel katkıları şu şekilde özetlenebilir: (i) Term Frequency – Inverse Document Frequency (TF-IDF) ve Term Frequency – Relevance Frequency (TF-RF) yöntemlerini alan-özgü SBERT gömme temsilleri ile bütünleştiren bir hibrit metin temsil stratejisinin geliştirilmesi; (ii) Analysis of Variance (ANOVA) ve SHapley Additive exPlanations (SHAP) yöntemlerini birleştiren, istatistiksel anlamlılığı açıklanabilir yapay zekâ teknikleri ile entegre eden yenilikçi bir hibrit öznitelik seçim mekanizmasının önerilmesi; (iii) K-means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) ve agglomeratif kümeleme algoritmalarının çıktılarını bütünleştirerek verinin hem küresel hem de yerel yapısal özelliklerini yakalayan kapsamlı bir öznitelik çıkarım sürecinin tasarlanması; (iv) sözcüksel, anlamsal ve yapısal bilgiyi tek bir öznitelik uzayında birleştiren çok-vektörlü bir öznitelik füzyon yaklaşımının geliştirilmesi.

HeaRT çerçevesi, birbirini tamamlayan iki farklı referans biyomedikal metin veri seti üzerinde değerlendirilmiştir. Medical Abstracts veri seti [60], Kaggle platformundan elde edilen, beş farklı klinik durumu tanımlayan toplam 14.438 tıbbi özetten oluşmaktadır. Özetlerin ortalama uzunluğu 50 ila 550 kelime arasında değişmektedir. Veri setindeki sınıf dağılımı dengesiz bir yapı sergilemekte olup, sınıflar ve örnek sayıları Tablo 3.18'de sunulmaktadır.

Tablo 3.18. Medical Abstracts veri setinin sınıf dağılımı.

Sınıf Adı	Örnek Sayısı
Neoplazmlar	3.163
Sindirim Sistemi Hastalıkları	1.494
Sinir Sistemi Hastalıkları	1.925
Kardiyovasküler Hastalıklar	3.051
Genel Patolojik Durumlar	4.805
Toplam	14.438

Biomedical Text Dataset, ChatGPT kullanılarak üretilmiş, anlamsal açıdan çeşitli 3.060 biyomedikal metinden oluşmaktadır. Veri seti, kolon, akciğer ve tiroid kanseri olmak üzere üç sınıfa aittir ve dengeli bir dağılım göstermektedir (her sınıfta 1.020 örnek). Etiket sızıntısını önlemek amacıyla metinlerden "kolon", "akciğer" ve "tiroid" kelimeleri çıkarılmıştır. Belgelerin ortalama uzunluğu 40 ila 100 kelime arasındadır (Tablo 3.19).

Tablo 3.19. Biomedical Text Dataset veri setinin sınıf dağılımı.

Sınıf Adı	Örnek Sayısı
Kolon Kanseri	1.020
Akciğer Kanseri	1.020
Tiroid Kanseri	1.020
Toplam	3.060

Ham tıbbi metinler sistematik bir ön işleme sürecinden geçirilmiştir. Bu süreç sırasıyla şu adımları kapsamaktadır: (1) küçük harfe dönüştürme, (2) noktalama işaretleri ve özel karakterlerin kaldırılması, (3) NLTK İngilizce durak kelime listesi kullanılarak durak kelimelerin çıkarılması, (4) kelime düzeyinde tokenizasyon. SBERT'in Transformer tabanlı mimarisi bağlam duyarlı temsiller ürettiğinden, kökleme (stemming) veya lemmatizasyon gibi geleneksel ön işleme adımları uygulanmamıştır. Ön işleme sonrasında tokenize edilmiş kelime listeleri, gömme modelleri tarafından işlenebilmesi için boşluklarla ayrılmış cümle benzeri dizilere yeniden birleştirilmiştir.

HeaRT çerçevesinde metin temsili iki tamamlayıcı bileşenden oluşmaktadır. Birinci bileşen, TF tabanlı vektörizasyon yöntemleridir. TF-IDF [106], bir terimin belirli bir belgedeki yerel frekansını korpus genelindeki nadir olma durumu ile birleştirerek temsil gücünü artırmaktadır. TF-RF ise sınıf koşullu ilgilik bilgisini açıkça dahil ederek bir sınıfta sık geçen ancak diğer sınıflarda nadir bulunan terimleri ön plana çıkarmaktadır. Her iki yöntem de farklı veri seti özelliklerine göre tamamlayıcı avantajlar sunmaktadır: heterojen ve uzun metinlerden oluşan Medical Abstracts veri setinde TF-IDF daha etkili olurken, dengeli ve sınıf-özgü terminolojinin belirgin olduğu Biomedical Text Dataset'te TF-RF [107] üstün performans göstermiştir.

İkinci bileşen, SBERT [108] tabanlı derin anlamsal gömme temsilleridir. Bu çalışmada all-MiniLM-L6-v2 ön eğitimli SBERT modeli kullanılarak 768 boyutlu anlamsal gömme vektörleri elde edilmiştir. MVFF stratejisi kapsamında, TF-RF veya TF-IDF sözcüksel vektörleri ile SBERT gömme temsilleri vektör birleştirme yoluyla tek bir hibrit öznitelik matrisine dönüştürülmüştür. Yüksek boyutlu ve seyrek öznitelik uzaylarının genelleme kapasitesini olumsuz etkilemesi sorununu çözmek amacıyla ANOVA [109] ve SHAP [110] yöntemlerini birleştiren hibrit bir öznitelik seçim stratejisi önerilmiştir. Birinci aşamada tek yönlü ANOVA kullanılarak sözcüksel özniteliklerin hedef sınıf değişkenleri ile istatistiksel ilişkisi değerlendirilmiştir. İkinci aşamada SHAP kullanılarak her bir özniteliğin model çıktısına olan marjinal katkısı hesaplanmıştır. Hibrit öznitelik önem skoru, iki yöntemin normalize edilmiş skorlarının eşit ağırlıklı birleşimi ($0{,}5 \times \text{ANOVA} + 0{,}5 \times \text{SHAP}$) olarak hesaplanmıştır.

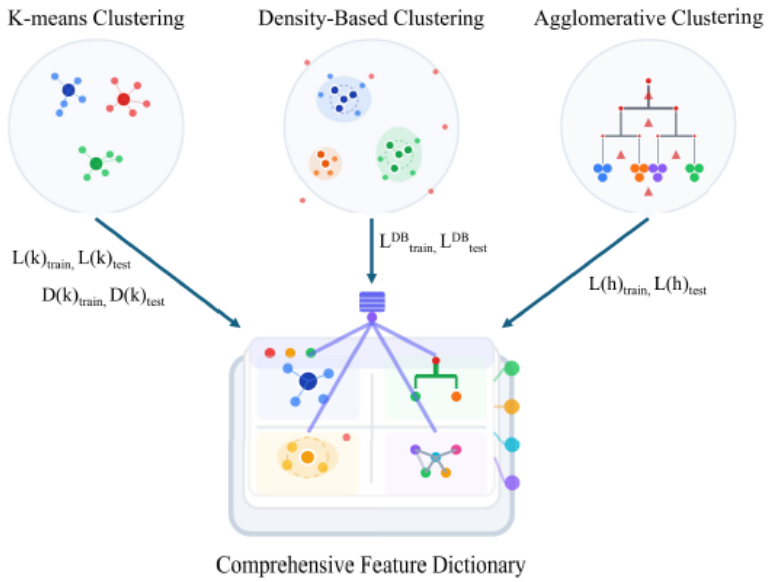
HeaRT çerçevesinin en ayırt edici bileşeni olan Cluster-Enhanced Feature (CEFA) yaklaşımı, denetimsiz kümeleme algoritmalarından elde edilen yapısal göstergeleri

öznitelik uzayına entegre etmektedir. CEFA süreci üç aşamadan oluşmaktadır: (I) K-means kümeleme ($k \in \{3, 5, 8, 10, 15\}$), küme etiketleri ve Öklid centroid uzaklıklarının çıkarılması; (II) DBSCAN ($\varepsilon = 0{,}5$, min_samples = 5) ile yoğunluk tabanlı kümeleme ve gürültü/aykırı değer tespiti; (III) Agglomeratif hiyerarşik kümeleme ($h \in \{5, 10\}$) ile iç içe geçmiş yapıların yakalanması. Üç algoritmadan elde edilen tüm yapısal çıktılar kapsamlı bir öznitelik sözlüğünde birleştirilerek sınıflandırma için nihai öznitelik vektörünü oluşturmuştur (Şekil 3.16). Kullanılan parametreler Tablo 3.20'de özetlenmektedir.

Tablo 3.20. CEFA sürecinde kullanılan kümeleme algoritmaları ve parametre ayarları.

Algoritma	Parametre(ler)	Değer(ler)
K-means	n_clusters (k)	3, 5, 8, 10, 15
	n_init	10
	random_state	42
	Çıktı	Küme etiketleri ve centroid uzaklıkları
DBSCAN	ε	0,5
	min_samples	5
Agglomeratif	n_clusters	5, 10
	Bağlantı yöntemi	Ortalama (average)
	Uzaklık metriği	Kosinüs

Sınıflandırma aşamasında on farklı algoritma değerlendirilmiştir: AdaBoost [111], ExtraTrees [112], Gradient Boosting [112], LightGBM [113], Passive Aggressive [114], Random Forest [115], RidgeClassifier [116], XGBoost [117], MLP (128-64 nöronlu iki gizli katman, ReLU aktivasyonu, 500 iterasyon) ve SGDClassifier [118]. Her veri seti tabakalı örnekleme ile %80 eğitim ve %20 test olarak bölünmüş; tüm deneyler beş farklı rastgele tohum (0, 42, 64, 123, 1234) üzerinde yürütülmüştür.



Şekil 3.16. HeaRT çerçevesindeki küme destekli özellik zenginleştirme iş akışı.

Performans metrikleri ortalama ± standart sapma olarak raporlanmıştır. Tüm deneyler Intel i7 13700K işlemci, NVIDIA RTX 4070 Ti GPU (12 GB), 64 GB DDR4 RAM ortamında Python 3.10 ve PyTorch 2.7.0 ile gerçekleştirilmiştir. Medical Abstracts veri seti üzerinde üç farklı konfigürasyon sistematik olarak karşılaştırılmıştır. Tablo 3.21'de yalnızca TF-IDF vektörizasyonu sonuçları sunulmaktadır. Sonuçlara göre en yüksek ortalama başarıyı doğrulukta %53,42 ve F1-skorunda %53,20 ile Gradient Boosting modeli elde ederken, LightGBM modeli de benzer şekilde yüksek performans göstermiştir.

Tablo 3.21. Medical Abstracts veri setinde TF-IDF temsili ile sınıflandırma sonuçları.

Model	Tohum 0 Doğruluk(%) / F1 skor(%)	Tohum 42 Doğruluk(%) / F1 skor(%)	Tohum 64 Doğruluk(%) / F1 skor(%)	Tohum 123 Doğruluk(%) / F1 skor(%)	Tohum 1234 Doğruluk(%) / F1 skor(%)	Ortalama Doğruluk(%) / F1 skor(%)
XGBoost	50,38 / 50,39	50,90 / 50,82	50,21 / 50,29	51,28 / 51,04	51,97 / 51,92	50,95 / 50,89
ExtraTrees	47,23 / 46,73	47,75 / 47,18	46,92 / 46,32	47,75 / 46,81	47,58 / 46,79	47,44 / 46,77
LightGBM	52,80 / 52,64	54,19 / 54,02	53,32 / 53,20	53,36 / 52,94	53,43 / 53,21	53,42 / 53,20
Passive Agg.	43,84 / 43,54	43,18 / 42,88	43,52 / 43,60	45,64 / 45,24	45,84 / 46,10	44,40 / 44,27
Random Forest	47,09 / 46,63	47,78 / 47,18	46,50 / 45,95	47,61 / 46,74	47,61 / 46,95	47,32 / 46,69
Grad. Boosting	56,72 / 56,47	57,06 / 56,76	56,34 / 56,09	55,99 / 55,33	56,44 / 56,05	56,51 / 56,14
Log. Regression	41,20 / 41,44	40,96 / 41,19	40,89 / 41,13	42,66 / 42,57	42,35 / 42,55	41,61 / 41,78
MLP	45,15 / 45,22	46,92 / 47,01	44,74 / 44,78	46,12 / 45,64	47,78 / 47,76	46,14 / 46,08
SGD Classifier	43,77 / 43,81	45,57 / 44,93	45,39 / 44,28	45,74 / 45,57	46,09 / 44,47	45,31 / 44,61
Ridge Classifier	41,17 / 41,17	42,28 / 42,38	42,11 / 42,11	43,04 / 42,78	43,07 / 43,12	42,33 / 42,31

TF-IDF ve SBERT füzyonu ile elde edilen sonuçlar Tablo 3.22'de sunulmaktadır. AdaBoost modeli %58,78 ortalama F1 skoru ile en yüksek performansı göstermiştir.

Tablo 3.22. Medical Abstracts veri setinde TF-IDF + SBERT füzyonu ile sınıflandırma sonuçları.

Model	Tohum 0 Doğruluk(%) / F1 skor(%)	Tohum 42 Doğruluk(%) / F1 skor(%)	Tohum 64 Doğruluk(%) / F1 skor(%)	Tohum 123 Doğruluk(%) / F1 skor(%)	Tohum 1234 Doğruluk(%) / F1 skor(%)	Ortalama Doğruluk(%) / F1 skor(%)
AdaBoost	57,13 / 56,86	59,21 / 58,82	60,15 / 60,21	59,11 / 58,85	59,28 / 59,15	58,98 / 58,78
Grad. Boosting	54,71 / 54,02	57,41 / 56,79	58,03 / 56,89	56,30 / 55,26	52,22 / 51,07	55,73 / 54,80
RidgeClassifier	55,06 / 54,20	57,62 / 56,36	57,10 / 55,76	56,34 / 55,03	47,85 / 44,56	54,79 / 53,18
LightGBM	49,17 / 48,50	51,11 / 50,36	50,90 / 50,12	49,93 / 49,30	50,24 / 49,02	50,27 / 49,46
MLP	47,26 / 46,72	50,42 / 49,98	47,30 / 46,29	48,75 / 48,30	46,88 / 45,98	48,12 / 47,45
ExtraTrees	47,71 / 45,08	48,75 / 45,72	47,71 / 44,77	47,61 / 45,12	55,85 / 54,60	49,53 / 47,06
PassiveAgg.	47,06 / 45,67	49,07 / 46,98	48,96 / 47,96	48,34 / 47,18	46,54 / 45,94	47,99 / 46,75
XGBoost	48,96 / 48,11	50,83 / 49,74	50,42 / 49,31	48,72 / 47,78	37,08 / 35,96	47,20 / 46,18
SGDClassifier	47,26 / 46,75	47,06 / 46,80	46,68 / 46,67	46,54 / 46,74	41,03 / 40,80	45,71 / 45,55
RandomForest	47,16 / 45,12	48,61 / 46,17	48,13 / 45,73	47,40 / 45,37	48,23 / 44,75	47,91 / 45,43

ANOVA-SHAP hibrit öznitelik seçimi ve CEFA yapısal zenginleştirmesinin entegrasyonu sonrasında elde edilen HeaRT tam konfigürasyon sonuçları Tablo 3.23'te sunulmaktadır. AdaBoost, %60,74 ortalama F1 skoru ve %60,94 doğruluk ile en yüksek performansı elde etmiştir. Elde edilen bulgular, TF-IDF tabanlı klasik temsil yöntemlerine kıyasla HeaRT mimarisinin biyomedikal metinlerin anlamsal ve yapısal özelliklerini daha etkili temsil ederek sınıflandırma performansını artırdığını göstermektedir.

Tablo 3.23. Medical Abstracts veri setinde HeaRT tam konfigürasyonu ile sınıflandırma sonuçları.

Model	Tohum 0 Doğruluk(%) / F1 skor(%)	Tohum 42 Doğruluk(%) / F1 skor(%)	Tohum 64 Doğruluk(%) / F1 skor(%)	Tohum 123 Doğruluk(%) / F1 skor(%)	Tohum 1234 Doğruluk(%) / F1 skor(%)	Ortalama Doğruluk(%) / F1 skor(%)
AdaBoost	60,04 / 60,35	61,25 / 60,50	61,18 / 60,96	60,49 / 60,13	61,74 / 61,78	60,94 / 60,74
Grad. Boosting	55,99 / 55,45	57,96 / 57,13	58,14 / 57,36	56,82 / 55,85	56,27 / 55,14	57,04 / 56,19
RidgeClassifier	55,44 / 54,55	57,65 / 56,51	56,89 / 55,59	56,02 / 54,84	56,27 / 54,90	56,45 / 55,28
LightGBM	49,48 / 48,84	51,39 / 50,59	51,49 / 50,70	49,52 / 48,48	51,21 / 50,30	50,62 / 49,78
XGBoost	48,65 / 47,95	50,55 / 49,63	50,83 / 49,95	49,69 / 48,87	50,42 / 49,09	50,03 / 49,10
SGDClassifier	54,47 / 49,22	51,42 / 51,58	55,78 / 50,22	50,17 / 47,67	49,03 / 44,02	52,18 / 48,54
PassiveAgg.	57,38 / 53,43	56,20 / 55,81	52,15 / 46,56	41,66 / 31,27	54,88 / 48,07	52,45 / 47,03
MLP	46,43 / 45,83	47,71 / 46,34	47,65 / 46,67	46,78 / 46,25	47,92 / 46,86	47,30 / 46,39
ExtraTrees	47,82 / 45,91	48,27 / 46,05	48,93 / 47,03	47,40 / 45,69	48,72 / 46,43	48,23 / 46,22
RandomForest	47,58 / 45,96	48,20 / 45,95	48,58 / 47,05	47,23 / 45,85	48,37 / 46,31	47,99 / 46,22

Karşılaştırma amacıyla, Tablo 3.24'te yalnızca BERT tabanlı cümle gömme temsilleri kullanılarak elde edilen baseline sonuçları sunulmaktadır.

Tablo 3.24. Medical Abstracts veri setinde baseline model (BERT) sonuçları.

Model	Tohum 0 Doğruluk(%) / F1 skor(%)	Tohum 42 Doğruluk(%) / F1 skor(%)	Tohum 64 Doğruluk(%) / F1 skor(%)	Tohum 123 Doğruluk(%) / F1 skor(%)	Tohum 1234 Doğruluk(%) / F1 skor(%)	Ortalama Doğruluk(%) / F1 skor(%)
RidgeClassifier	53,19 / 49,73	54,64 / 50,33	54,05 / 49,54	52,98 / 48,84	55,02 / 50,96	53,98 / 49,88
SGDClassifier	46,16 / 42,57	45,15 / 43,55	48,61 / 45,01	50,17 / 45,32	44,46 / 39,49	46,91 / 43,19
AdaBoost	45,88 / 42,19	46,19 / 40,85	44,43 / 38,88	45,81 / 40,93	45,53 / 40,62	45,57 / 40,69
PassiveAgg.	45,53 / 36,99	50,87 / 47,78	46,33 / 39,27	42,52 / 39,50	45,15 / 39,62	46,08 / 40,63
Grad. Boosting	43,01 / 38,92	44,04 / 39,64	43,84 / 38,60	43,63 / 38,81	43,77 / 39,44	43,66 / 39,08
MLP	37,47 / 36,63	38,19 / 36,88	36,88 / 34,42	36,91 / 35,58	38,95 / 37,65	37,68 / 36,23
LightGBM	37,74 / 33,53	39,13 / 35,03	38,19 / 33,84	38,40 / 34,29	39,16 / 34,93	38,52 / 34,32
XGBoost	36,70 / 32,83	37,36 / 32,60	36,98 / 32,47	36,46 / 32,25	38,16 / 33,75	37,13 / 32,78
ExtraTrees	34,28 / 26,83	34,52 / 26,87	33,86 / 26,53	33,34 / 25,97	34,49 / 27,13	34,10 / 26,67
RandomForest	34,28 / 27,02	34,07 / 26,67	33,17 / 25,85	33,34 / 26,34	34,21 / 27,28	33,82 / 26,63

Tablo 3.25'te HeaRT ve temel çizgi modelleri arasındaki eşleştirilmiş t-testi sonuçları sunulmaktadır. HeaRT, PassiveAggressive dışındaki tüm modellerde istatistiksel olarak anlamlı iyileştirmeler sağlamıştır ($p < 0,05$).

Tablo 3.25. Medical Abstracts veri setinde HeaRT ve baseline model arasındaki eşleştirilmiş t-testi.

Model	Ort HeaRT	SS HeaRT	Ort Temel	SS Temel	p-değeri	Anlamlı	Kazanım
AdaBoost	60,744	0,65	40,69	1,18	0,000009	Evet	HeaRT
ExtraTrees	46,222	0,53	26,67	0,44	0	Evet	HeaRT
Grad. Boosting	56,186	1,00	39,08	0,44	0,000005	Evet	HeaRT
LightGBM	49,782	1,04	34,32	0,66	0,000003	Evet	HeaRT
MLP	46,39	0,40	36,23	1,25	0,000066	Evet	HeaRT
PassiveAgg.	47,028	9,59	40,63	4,14	0,18634	Hayır	–
RandomForest	46,224	0,49	26,63	0,56	0,000001	Evet	HeaRT
RidgeClassifier	55,278	0,79	49,88	0,81	0,000254	Evet	HeaRT
SGDClassifier	48,542	2,90	43,19	2,35	0,00514	Evet	HeaRT
XGBoost	49,098	0,77	32,78	0,58	0,000004	Evet	HeaRT

Biomedical Text Dataset üzerinde sözcüksel bileşen olarak TF-RF kullanılmıştır. Bu seçimin nedeni, veri setinin dengeli üç sınıftan oluşması ve sınıf-özgü terminolojinin belirgin olmasıdır. TF-RF temsili ile elde edilen sonuçlar Tablo 3.26'da sunulmaktadır.

Tablo 3.26. Biomedical Text Dataset'te TF-RF temsili ile sınıflandırma sonuçları.

Model	Tohum 0 Doğruluk(%) / F1 skor(%)	Tohum 42 Doğruluk(%) / F1 skor(%)	Tohum 64 Doğruluk(%) / F1 skor(%)	Tohum 123 Doğruluk(%) / F1 skor(%)	Tohum 1234 Doğruluk(%) / F1 skor(%)	Ortalama Doğruluk(%) / F1 skor(%)
XGBoost	90,39 / 90,43	92,18 / 92,18	92,02 / 92,01	90,72 / 90,71	90,07 / 90,05	91,07 / 91,08
ExtraTrees	92,35 / 92,37	92,67 / 92,70	91,21 / 91,23	92,35 / 92,35	90,88 / 90,90	91,89 / 91,91
LightGBM	91,86 / 91,87	91,86 / 91,85	91,86 / 91,85	92,18 / 92,18	90,55 / 90,52	91,66 / 91,65
Passive Agg.	86,32 / 86,32	87,30 / 87,28	88,60 / 88,60	88,11 / 88,13	88,44 / 88,41	87,75 / 87,75
Random Forest	90,88 / 90,94	91,37 / 91,38	90,72 / 90,75	91,04 / 91,09	89,58 / 89,62	90,72 / 90,75
Grad. Boosting	88,11 / 88,19	91,04 / 91,08	89,58 / 89,63	89,58 / 89,59	89,41 / 89,46	89,54 / 89,59
Log. Regression	86,81 / 86,82	88,60 / 88,63	88,11 / 88,11	88,60 / 88,62	88,60 / 88,61	88,14 / 88,16
MLP	89,25 / 89,25	88,27 / 88,27	89,09 / 89,08	88,11 / 88,12	89,25 / 89,26	88,79 / 88,79
SGD Classifier	85,18 / 85,23	86,48 / 86,49	86,48 / 86,52	88,60 / 88,60	86,81 / 86,80	86,71 / 86,73
Ridge Classifier	77,52 / 77,43	77,69 / 77,51	79,32 / 79,26	77,20 / 77,12	77,52 / 77,52	77,85 / 77,75

TF-RF ve SBERT füzyonu sonuçları Tablo 3.27'de sunulmaktadır. XGBoost (%93,32) ve ExtraTrees (%93,27) en yüksek performansı sağlamıştır.

Tablo 3.27. Biomedical Text Dataset'te TF-RF + SBERT füzyonu ile sınıflandırma sonuçları.

Model	Tohum 0 Doğruluk(%) / F1 skor(%)	Tohum 42 Doğruluk(%) / F1 skor(%)	Tohum 64 Doğruluk(%) / F1 skor(%)	Tohum 123 Doğruluk(%) / F1 skor(%)	Tohum 1234 Doğruluk(%) / F1 skor(%)	Ortalama Doğruluk(%) / F1 skor(%)
XGBoost	93,49 / 93,55	93,65 / 93,67	92,54 / 92,51	93,97 / 93,97	92,83 / 92,86	93,29 / 93,32
ExtraTrees	93,16 / 93,19	93,65 / 93,66	92,83 / 92,84	93,81 / 93,80	92,83 / 92,87	93,26 / 93,27
LightGBM	93,49 / 93,54	94,46 / 94,46	92,02 / 92,06	94,30 / 94,26	91,86 / 91,91	93,23 / 93,25
Passive Agg.	92,51 / 92,54	93,49 / 93,49	93,00 / 92,96	93,32 / 93,31	91,86 / 91,86	92,84 / 92,83
Random Forest	92,18 / 92,24	93,81 / 93,83	92,51 / 92,51	93,81 / 93,81	91,69 / 91,73	92,80 / 92,82
Grad. Boosting	92,35 / 92,42	93,65 / 93,70	92,02 / 92,01	92,83 / 92,78	91,69 / 91,73	92,51 / 92,53
Log. Regression	91,86 / 91,88	93,00 / 93,01	92,02 / 92,02	93,16 / 93,18	91,37 / 91,40	92,28 / 92,29
MLP	92,02 / 92,05	93,65 / 93,71	92,02 / 92,02	92,35 / 92,34	89,74 / 89,74	91,96 / 91,97
SGD Classifier	91,04 / 91,09	91,04 / 91,05	91,86 / 91,79	91,37 / 91,39	91,37 / 91,39	91,34 / 91,34
Ridge Classifier	87,95 / 87,99	87,13 / 87,13	86,97 / 86,95	87,95 / 87,90	87,95 / 87,98	87,59 / 87,59

HeaRT tam konfigürasyonu sonuçları Tablo 3.28'de sunulmaktadır. LightGBM, %94,01 doğruluk ve %94,02 F1 skoru ile en yüksek performansı elde etmiştir.

Tablo 3.28. Biomedical Text Dataset'te HeaRT tam konfigürasyonu ile sınıflandırma sonuçları.

Model	Tohum 0 Doğruluk(%) / F1 skor(%)	Tohum 42 Doğruluk(%) / F1 skor(%)	Tohum 64 Doğruluk(%) / F1 skor(%)	Tohum 123 Doğruluk(%) / F1 skor(%)	Tohum 1234 Doğruluk(%) / F1 skor(%)	Ortalama Doğruluk(%) / F1 skor(%)
LightGBM	93,81 / 93,84	95,28 / 95,27	93,32 / 93,34	95,28 / 95,23	92,35 / 92,39	94,01 / 94,02
ExtraTrees	92,67 / 92,69	95,44 / 95,44	93,65 / 93,66	93,81 / 93,81	92,83 / 92,87	93,68 / 93,69
XGBoost	93,49 / 93,53	93,97 / 93,96	92,51 / 92,55	94,79 / 94,78	92,51 / 92,53	93,45 / 93,47
RandomForest	93,00 / 93,03	93,97 / 93,99	93,49 / 93,52	93,16 / 93,14	91,53 / 91,58	93,03 / 93,05
Grad. Boosting	92,51 / 92,59	93,00 / 93,01	91,37 / 91,37	93,81 / 93,78	92,18 / 92,19	92,57 / 92,59
Log. Regression	92,02 / 92,03	93,16 / 93,17	92,18 / 92,16	92,67 / 92,68	92,02 / 92,04	92,41 / 92,42
PassiveAgg.	92,51 / 92,52	93,81 / 93,84	91,86 / 91,89	91,21 / 91,28	92,18 / 92,23	92,31 / 92,35
MLP	91,86 / 91,90	92,83 / 92,85	92,51 / 92,50	92,35 / 92,31	91,53 / 91,53	92,22 / 92,22
SGDClassifier	90,39 / 90,42	93,81 / 93,83	91,53 / 91,53	91,53 / 91,54	89,58 / 89,61	91,37 / 91,39
RidgeClassifier	87,13 / 87,18	87,46 / 87,44	88,11 / 88,09	87,62 / 87,58	88,76 / 88,78	87,82 / 87,81

Tablo 3.29'da, Biomedical Text Dataset üzerinde baseline olarak kullanılan BERT tabanlı temsil yöntemi ile farklı makine öğrenmesi modellerinin sınıflandırma performansları karşılaştırılmıştır. Sonuçlara göre en yüksek başarı LightGBM modeli ile elde edilmiş olup model ortalama %90,23 doğruluk ve %90,23 F1-skoru değerine ulaşmıştır.

Tablo 3.29. Biomedical Text Dataset'te baseline model (BERT) sonuçları.

Model	Tohum 0 Doğruluk(%) / F1 skor(%)	Tohum 42 Doğruluk(%) / F1 skor(%)	Tohum 64 Doğruluk(%) / F1 skor(%)	Tohum 123 Doğruluk(%) / F1 skor(%)	Tohum 1234 Doğruluk(%) / F1 skor(%)	Ortalama Doğruluk(%) / F1 skor(%)
LightGBM	89,90 / 89,93	92,35 / 92,37	90,55 / 90,50	87,79 / 87,80	90,55 / 90,54	90,23 / 90,23
XGBoost	85,50 / 85,50	90,39 / 90,39	88,60 / 88,57	86,16 / 86,15	88,27 / 88,19	87,78 / 87,76
Grad. Boosting	85,83 / 85,84	87,95 / 88,03	87,95 / 87,95	85,18 / 85,27	86,16 / 86,19	86,61 / 86,65
RandomForest	83,88 / 83,82	85,67 / 85,73	84,20 / 84,24	87,46 / 87,46	85,99 / 86,02	85,44 / 85,45
ExtraTrees	83,22 / 83,27	86,32 / 86,44	84,20 / 84,23	85,18 / 85,25	86,16 / 86,19	85,02 / 85,08
RidgeClassifier	51,95 / 50,90	50,83 / 48,37	47,07 / 53,09	51,79 / 51,77	51,79 / 50,49	50,69 / 50,92
Log. Regression	43,97 / 42,99	41,53 / 40,40	43,84 / 48,21	47,64 / 42,83	40,18 / 40,18	43,43 / 42,92
MLP	39,58 / 39,47	42,02 / 41,75	42,51 / 42,51	43,65 / 43,51	44,14 / 43,83	42,38 / 42,21
SGDClassifier	38,44 / 37,46	36,32 / 28,84	36,81 / 34,65	35,18 / 31,06	40,23 / 38,21	37,39 / 34,04
PassiveAgg.	39,41 / 37,11	38,76 / 33,76	39,25 / 37,39	37,13 / 29,98	37,13 / 30,95	38,34 / 33,84

Tablo 3.30'da HeaRT ve temel çizgi modelleri arasındaki eşleştirilmiş t-testi sonuçları sunulmaktadır. HeaRT, tüm modellerde istatistiksel olarak anlamlı iyileştirmeler sağlamıştır ($p < 0,05$).

Tablo 3.30. Biomedical Text Dataset'te HeaRT ve baseline model arasındaki eşleştirilmiş t-testi.

Model	Ort HeaRT	SS HeaRT	Ort Temel	SS Temel	p-değeri	Anlamlı	Kazanım
LightGBM	94,01	1,24	90,23	1,64	0,0173	Evet	HeaRT
ExtraTrees	93,69	1,09	85,08	1,33	0,000072	Evet	HeaRT
XGBoost	93,47	0,96	87,76	1,97	0,0062	Evet	HeaRT
RandomForest	93,05	0,90	85,45	1,46	0,0008	Evet	HeaRT
Grad. Boosting	92,59	0,90	86,66	1,26	0,0022	Evet	HeaRT
Log. Regression	92,42	0,50	43,01	3,04	0,000004	Evet	HeaRT
PassiveAgg.	92,35	0,95	33,84	3,41	0,000002	Evet	HeaRT
MLP	92,22	0,52	42,21	1,74	<0,000001	Evet	HeaRT
SGDClassifier	91,39	1,59	34,04	4,04	0,000021	Evet	HeaRT
RidgeClassifier	87,81	0,63	50,21	1,82	0,000002	Evet	HeaRT

HeaRT çerçevesinin iki farklı biyomedikal veri seti üzerinde elde ettiği sonuçlar, önerilen çok katmanlı yaklaşımın tutarlı ve istatistiksel olarak anlamlı iyileştirmeler sağladığını ortaya koymaktadır. Sözcüksel ve anlamsal temsillerin füzyonu, yalnızca sözcüksel veya yalnızca anlamsal yöntemlere kıyasla tutarlı performans artışı sağlamıştır. TF-IDF'nin Medical Abstracts gibi heterojen ve geniş veri setlerinde, TF-RF'nin ise Biomedical Text Dataset gibi dengeli ve sınıf-özgü terminolojisi belirgin veri setlerinde SBERT ile füzyon halinde daha etkili olduğu belirlenmiştir.

ANOVA-SHAP hibrit öznitelik seçim stratejisi, yüksek boyutlu öznitelik uzaylarında gereksiz ve gürültülü özniteliklerin filtrelenmesini sağlayarak özellikle topluluk tabanlı sınıflandırıcılarda genelleme performansını artırmıştır. CEFA süreci, kümeleme algoritmalarından elde edilen yapısal göstergelerin öznitelik uzayına entegrasyonu yoluyla, anlamsal olarak örtüşen sınıflar arasındaki ayrımcılığı iyileştirmiştir. Bu katkı özellikle Medical Abstracts veri setinde Genel Patolojik Durumlar ile Sinir Sistemi Hastalıkları arasındaki karışıklığın azaltılmasında ve Biomedical Text Dataset'te Kolon Kanseri ile Akciğer Kanseri arasındaki örtüşmenin giderilmesinde gözlenmiştir.

Ağaç tabanlı topluluk yöntemleri (AdaBoost, LightGBM, ExtraTrees, XGBoost), doğrusal modellere kıyasla çok-vektörlü öznitelik uzayındaki karmaşık doğrusal olmayan ilişkileri daha etkili biçimde modellemiştir. Literatürdeki en ileri yöntemlerle karşılaştırıldığında, HeaRT Medical Abstracts veri setinde %60,74 F1 skoru ile tam otomatik yöntemler arasında en yüksek performansı sağlamıştır. Biomedical Text Dataset'te ise HeaRT %94,02 F1 skoru ile daha önce raporlanmış en iyi sonucu (%85,90 — LSTM) yaklaşık 8 puanla geçerek yeni bir referans değer oluşturmuştur.

Bu temel çalışmadan elde edilen deneyimler, tezin asıl odağı olan LLM tabanlı radyolojik rapor üretim sisteminin geliştirilmesine doğrudan katkı sağlamıştır. Tıbbi metin yapılarının anlaşılması, bağlamsal gömme temsillerinin tıbbi alanda değerlendirilmesi ve sınıf dengesizliği ile başa çıkma stratejilerinin geliştirilmesi; radyolojik raporların hem üretilmesinde hem de otomatik değerlendirilmesinde kritik bir metodolojik temel oluşturmuştur.

4. TARTIŞMA

Bu tez çalışmasında, akut iskemik inme lezyonlarının beyin MR görüntülerinden otomatik olarak segmentasyonu, segmentasyon maskelerinden anatomik ve klinik bilgilerin çıkarılması ve elde edilen yapılandırılmış bulguların büyük dil modelleri kullanılarak otomatik radyolojik raporlara dönüştürülmesi amacıyla bütünleşik bir yapay zekâ sistemi geliştirilmiştir. Çalışma kapsamında hem konvolüsyon tabanlı hem de Transformer tabanlı segmentasyon mimarileri değerlendirilmiş; ayrıca farklı çoklu girdi kombinasyonlarının segmentasyon performansı üzerindeki etkileri sistematik olarak incelenmiştir. Aşağıda elde edilen bulgular, literatürdeki ilgili çalışmalarla karşılaştırmalı olarak tartışılmakta; çalışmanın güçlü yönleri, sınırlılıkları ve gelecek çalışma önerileri ele alınmaktadır.

Segmentasyon sonuçları incelendiğinde, nnU-Net mimarisinin genel olarak en kararlı ve en yüksek performansı sağlayan model olduğu gözlemlenmiştir. nnU-Net, DWI konfigürasyonunda %71,66 Dice skoru ile hem SwinUNETR'ı (%63,98) hem de 3D U-Net'i (%55,81) belirgin biçimde geride bırakmıştır. İstatistiksel analizler bu farkın klinik açıdan anlamlı olduğunu doğrulamıştır: nnU-Net ile 3D U-Net arasındaki Cohen's d değeri 0,98 (büyük etki), nnU-Net ile SwinUNETR arasındaki değer ise 0,60 (orta etki) olarak

hesaplanmıştır. nnU-Net'in üstün performansı, bu mimarinin veri setine özgü yapılandırmaları (yama boyutu, batch boyutu, ağ topolojisi, ön işleme stratejisi) otomatik olarak optimize edebilmesi ile doğrudan ilişkilidir. Isensee ve arkadaşları [76] tarafından gösterildiği üzere, nnU-Net'in kural tabanlı yapılandırma motoru, manuel hiperparametre ayarlamasına gerek kalmadan farklı veri setlerinde tutarlı performans sağlamaktadır. Özellikle 150 hastalık görece küçük bir veri setinde, bu otomatik optimizasyon mekanizması güçlü bir tümevarımsal önyargı sunarak modelin hızlı yakınsamasına katkıda bulunmuştur.

Transformer tabanlı SwinUNETR modelinin nnU-Net'e kıyasla daha düşük performans göstermesi ise bu mimarinin yüksek veri gereksinimi ile açıklanabilir. Tang ve arkadaşları [79] SwinUNETR'ı 5.050 BT taramasından oluşan büyük ölçekli bir veri seti üzerinde ön eğitimden geçirerek başarılı sonuçlar elde etmişlerdir; buna karşın bu tezde kullanılan 150 hastalık veri seti, Transformer bloklarındaki çok başlı öz-dikkat mekanizmalarının etkili biçimde eğitilebilmesi için sınırlı kalmaktadır. Bununla birlikte SwinUNETR'ın FULL konfigürasyonunda (%65,90) çoklu girdi bileşenlerinden sınırlı da olsa fayda sağlaması, bu mimarinin daha büyük veri setlerinde potansiyelini ortaya koyabileceğini düşündürmektedir. 3D U-Net modelinin tüm konfigürasyonlarda en düşük performansı göstermesi (Dice: %53,39–%56,04), bu mimarinin nnU-Net'in sunduğu otomatik yapılandırma ve gelişmiş eğitim stratejilerinden yoksun olması ile ilişkilendirilmektedir.

Bu çalışmada elde edilen segmentasyon sonuçları, literatürdeki ilgili çalışmalarla karşılaştırıldığında rekabetçi bir konumda yer almaktadır. ISLES 2022 yarışmasının resmi değerlendirme sonuçlarına göre, yarışma katılımcılarının DWI ve ADC verileri kullanarak elde ettiği ortalama Dice skorları 0,50–0,60 aralığında yoğunlaşmaktadır [48]. Bu çalışmada nnU-Net ile elde edilen %71,66'lık Dice skoru, ISLES 2022 ortalamalarının üzerinde yer almaktadır. Ancak bu karşılaştırmanın dikkatli yorumlanması gerekmektedir; zira ISLES 2022 çok merkezli ve heterojen bir veri seti kullanırken, bu çalışmada tek merkezli bir klinik veri seti kullanılmıştır ve tek merkezli veri setleri, görüntüleme protokollerinin homojenliği nedeniyle genellikle daha yüksek performans değerlerine yol açabilmektedir.

Rahman ve arkadaşları [57] ISLES 2022 veri seti üzerinde DenseNet121 tabanlı çok kanallı bir model ile %87 Dice skoru raporlamışlardır; ancak bu yüksek değer, DWI, ADC ve geliştirilmiş DWI gibi üç modaliteyi birlikte kullanan özelleştirilmiş bir mimarinin sonucudur. Nouman ve arkadaşları [55] ATLAS 2.0 veri seti üzerinde Neuro-TransUNet hibrit modeli ile %73 Dice skoru elde etmişlerdir ve bu değer bu çalışmadaki nnU-Net performansı ile karşılaştırılabilir düzeydedir. Xue ve arkadaşları [54] çok yollu 2.5D CNN yaklaşımı ile ATLAS veri setinde %54, KF+MCW veri setinde %47 Dice skorları raporlamışlardır; bu çalışmada 3D U-Net ile elde edilen sonuçlar bile (%55,81) bu değerlerin üzerindedir. Ahmed ve arkadaşları [52] Transformer tabanlı deforme olabilir dikkat mekanizmalı bir model ile ATLAS veri setinde %82 Dice skoru elde etmişlerdir; ancak bu çalışmada ekolojik veri artırma stratejisi kullanılarak eğitim seti önemli ölçüde genişletilmiştir. Tüm bu karşılaştırmalar değerlendirildiğinde, bu tez çalışmasının diğer çalışmalardan ayrılan temel yönü, segmentasyonun nihai amaç olmayıp daha geniş bir uçtan uca sistemin ilk aşaması olarak konumlandırılmasıdır. Literatürdeki çalışmaların büyük çoğunluğu yalnızca segmentasyon performansını optimize etmeye odaklanırken, bu tezde segmentasyon çıktıları anatomik bilgi çıkarımı ve otomatik rapor üretimi ile bütünleştirilmiştir.

Çalışmada değerlendirilen çoklu girdi kombinasyonlarına ilişkin istatistiksel analiz sonuçları (Tablo 3.6), giriş konfigürasyonları arasındaki farkların modeller arası farklarla kıyaslandığında çok daha küçük olduğunu açıkça göstermiştir. nnU-Net için konfigürasyonlar arası en büyük etki büyüklüğü d = 0,33 (küçük) iken, modeller arası en küçük etki büyüklüğü d = 0,60 (orta) olarak gözlenmiştir. Bu bulgu, iskemik inme lezyon segmentasyonunda mimari seçiminin giriş konfigürasyonundan daha belirleyici bir faktör olduğunu göstermektedir.

DWI ve ADC sekanslarının birlikte kullanılması SwinUNETR modelinde Dice değerini %63,98'den %65,49'a yükseltmiş olsa da nnU-Net'te DWI+ADC konfigürasyonu yalnızca DWI konfigürasyonuna kıyasla hafif bir performans düşüşü göstermiştir (Dice: %71,66 → %69,90). Bu durum, erken füzyon stratejisinde farklı dağılımlara sahip giriş kanallarının birleştirilmesinin her zaman doğrudan fayda sağlamadığına işaret etmektedir. EDGE kanalının eklenmesiyle gözlemlenen sınırlı performans artışları, Sobel tabanlı kenar çıkarımının gürültüye duyarlılığı, kenar haritalarının semantik bağlam sağlayamaması ve düşük kontrastlı lezyonlarda yanlış sınır vurguları oluşturabilmesi ile ilişkili olabilir. Anatomik ön bilgi kanalının sınırlı katkısı ise atlas tabanlı ön bilgilerin hasta-özgü anatomik varyasyonları tam olarak yansıtamaması ve yalnızca kaba uzamsal rehberlik sunmasından kaynaklanmış olabileceği düşünülmektedir.

EDGE ve PRIOR kanallarının aynı anda kullanıldığı FULL konfigürasyonun bazı modellerde beklenen performans artışını sağlayamaması, erken füzyon stratejisinin sınırlılıklarını ortaya koymaktadır. Lokal yüksek frekanslı sınır bilgisi ile küresel düşük frekanslı anatomik bağlam gibi farklı temsil biçimlerinin giriş seviyesinde birleştirilmesi, özellikle sınırlı veri koşullarında temsil çakışmalarına ve öğrenme kararsızlığına neden olabilir. Literatürde farklı füzyon stratejilerinin bu soruna yönelik çözümler sunduğu bilinmektedir. Zhou ve arkadaşları [119] tarafından yapılan kapsamlı bir derlemede, tıbbi görüntü segmentasyonunda erken füzyon, geç füzyon ve hibrit füzyon stratejilerinin karşılaştırılması sonucunda, özellikle heterojen bilgi kaynaklarının birleştirilmesinde dikkat mekanizması tabanlı geç füzyon yaklaşımlarının daha yüksek performans sağladığı gösterilmiştir. Benzer şekilde, Zhu ve arkadaşları [120] çok modlu beyin tümörü segmentasyonunda çapraz dikkat füzyon mekanizmasının erken kanal birleştirmeye kıyasla %3-5 Dice artışı sağladığını raporlamışlardır. Bu çalışmada kullanılan erken füzyon stratejisinin sınırlı başarısı, gelecekte dikkat tabanlı geç füzyon, çapraz dikkat füzyon veya çok akışlı füzyon stratejilerinin uygulanmasıyla daha etkili sonuçlar elde edilebileceğini güçlü biçimde düşündürmektedir. Özellikle her bir bilgi kaynağının (DWI, ADC, EDGE, PRIOR) ayrı kodlayıcı dallarında işlendikten sonra dikkat mekanizmaları aracılığıyla birleştirilmesi, farklı modalitelerin öğrenme sürecine göreve uygun ağırlıklarla katkıda bulunmasını sağlayabilir.

Otomatik rapor üretim sonuçları, altı büyük dil modelinin farklı değerlendirme boyutlarında birbirinden belirgin şekilde ayrışan performans profilleri sergilediğini ortaya koymuştur. Mistral 7B modelinin BLEU metriklerinde açık farkla lider olması (BLEU-4: 0,3808) ile Qwen 2.5 3B modelinin ROUGE-L (0,5129), METEOR (0,4526) ve BERTScore (0,7630) metriklerinde en yüksek değerleri elde etmesi, bu iki modelin farklı güçlü yönlere sahip olduğunu göstermektedir. Bu farklılaşmanın kökeninde iki metrik ailesinin ölçtüğü boyutlar yatmaktadır. BLEU, üretilen metindeki n-gramların referansta ne oranda bulunduğunu ölçen kesinlik odaklı bir metriktir; dolayısıyla Mistral 7B'nin yüksek BLEU performansı, bu modelin referans raporlardaki kelime ve ifade kalıplarını daha isabetli biçimde ürettiğini göstermektedir. Buna karşın Mistral 7B'nin düşük ROUGE-L skoru (0,3403) ve düşük

uzunluk oranı (0,7752), modelin referans raporlardaki klinik bilgilerin önemli bir kısmını atladığını ortaya koymaktadır. Özellikle "Sonuç" bölümünü yalnızca %24 oranında üreten Mistral 7B, klinik açıdan eksik raporlar oluşturma eğilimindedir.

Qwen 2.5 3B'nin ROUGE ve BERTScore metriklerindeki üstünlüğü ise bu modelin referans rapordaki klinik bilgileri daha kapsamlı biçimde yakaladığını göstermektedir. ROUGE duyarlılık odaklı bir metrik olduğundan, yüksek ROUGE-L değeri modelin referans rapordaki kritik klinik terimleri ve ifadeleri büyük ölçüde koruduğuna işaret etmektedir. BERTScore'un bağlamsal gömme temsillerini kullanması nedeniyle, Qwen 2.5 3B'nin yüksek BERTScore değeri aynı klinik anlamı farklı sözcüklerle ifade etme becerisini de yansıtmaktadır. Bu bulgu, Türkçe radyoloji terminolojisinde aynı bulgunun farklı ifade biçimleriyle tanımlanabilmesi (örneğin "difüzyon kısıtlanması" ile "kısıtlı difüzyon alanı") nedeniyle klinik açıdan özellikle önemlidir. Klinik uygulamalar açısından değerlendirildiğinde, bir raporun eksik bilgi içermesi klinik karar süreçlerini doğrudan olumsuz etkileyebileceğinden, yüksek ROUGE ve BERTScore değerlerinin yüksek BLEU değerlerine kıyasla daha kritik olduğu değerlendirilmektedir.

Qwen 2.5 ailesindeki üç modelin (3B, 7B, 14B) karşılaştırması, aynı mimari aile içinde model boyutu arttıkça performansın iyileşmediğini, aksine belirgin bir düşüş eğilimi gösterdiğini ortaya koymuştur. BLEU-4 değerleri 3B'de 0,2149, 7B'de 0,2110 ve 14B'de 0,0846; BERTScore değerleri ise sırasıyla 0,7630, 0,7422 ve 0,6954 olarak elde edilmiştir. Bu ters orantılı ilişkinin en olası açıklaması, eğitim veri seti büyüklüğü ile model kapasitesi arasındaki uyumsuzluktur. 120 eğitim örneği, 14,7 milyar parametreli bir modelin etkin biçimde adapte edilebilmesi için yeterli değildir. QLoRA ile yalnızca düşük rütbeli adaptör matrisleri güncellenmesine rağmen, 14B modelde eğitilmesi gereken parametre sayısı ($r \times (d+k)$ formülünde $d = 5.120$) hâlâ diğer modellere kıyasla çok daha yüksektir ve bu durum sınırlı eğitim verisinde aşırı öğrenme riskini artırmaktadır. Ayrıca 14B modelin daha düşük uzunluk oranı (0,5491) ve düşük "Bulgular" format uyumu (%29), modelin eğitim sırasında kısa ve yapısal olarak eksik raporlar üretmeyi öğrendiğini düşündürmektedir. Buna karşın Qwen 2.5 7B yerleşim tarafı doğruluğunda en yüksek skoru (%71,43) elde etmiştir. Yerleşim tarafı bilgisi radyoloji raporlarında yanlış bildirildiğinde doğrudan klinik hatalara yol açabileceğinden, bu modelin klinik kullanılabilirlik açısından güçlü bir aday olduğu değerlendirilmektedir.

Türkçe'ye özgü olarak eğitilmiş Kumru 2B ve Türkçe ince ayar yapılmış Cosmos 8B modellerinin çok dilli modellere kıyasla belirgin şekilde düşük performans göstermesi, dikkat çekici ve öğretici bir bulgudur. Kumru 2B'nin BERTScore değeri 0,5864, Cosmos 8B'nin ise 0,5301 ile tüm modeller arasında en düşük semantik benzerliği yansıtmaktadır. Cosmos 8B'nin BLEU-3 ve BLEU-4 değerlerinin sıfır olması, modelin referans raporlarla üç ve dört kelimelik ardışık dizi düzeyinde hiçbir eşleşme üretmediğini göstermektedir. Bu düşük performansın birkaç temel nedeni bulunmaktadır. Kumru 2B'nin 500 GB'lık Türkçe genel derlem üzerinde eğitilmiş olması, modelin günlük dil yapısını güçlü biçimde öğrenmesini sağlamakla birlikte, tıbbi terminoloji, radyoloji rapor formatı ve yapılandırılmış klinik çıktı üretimi gibi alan-özgü yetenekler açısından yeterli ön bilgi sağlayamamıştır. Cosmos 8B'nin LLaMA-3 üzerine yapılmış Türkçe ince ayarı ise ağırlıklı olarak genel diyalog kalitesini iyileştirmeye yönelik olup, yapılandırılmış girdi-çıktı eşlemesi ve alan-spesifik metin üretimi konusunda yeterli adaptasyon sağlayamamıştır. Her iki modelin de aşırı uzun raporlar üretmesi (Kumru: 4,78×, Cosmos: 3,29× referans uzunluğu), radyoloji rapor formatının yeterince öğrenilemediğinin açık bir göstergesidir. Cosmos 8B'nin yerleşim tarafı doğruluğunun yalnızca %9,52 olması ise modelin

yapılandırılmış girdideki taraf bilgisini rapor çıktısına yansıtma konusunda neredeyse tamamen başarısız olduğunu göstermektedir. Bu sonuçlar, Türkçe genel dil yetkinliğinin tıbbi rapor üretimi için gerekli ama yeterli olmayan bir koşul olduğunu açıkça ortaya koymaktadır. Çok dilli modellerin daha geniş ve çeşitli ön eğitim verilerinden edindiği yapılandırılmış metin üretim kapasitesi, talimat takip yeteneği ve çok dilli tıbbi terminoloji bilgisi, alana özgü ince ayarda belirleyici bir avantaj sağlamaktadır.

Bu çalışmada elde edilen rapor üretim sonuçları, literatürdeki ilgili çalışmalarla karşılaştırıldığında, farklı veri setleri ve diller üzerinde çalışılmasına rağmen rekabetçi bir konumda yer almaktadır. Zhang ve arkadaşları [51] StrokeQD veri seti üzerinde MAIF modeli ile BLEU-1: 0,5590, METEOR: 0,4487 ve ROUGE: 0,5664 değerleri raporlamışlardır. Bu çalışmada Mistral 7B ile elde edilen BLEU-1 (0,5505) ve Qwen 2.5 3B ile elde edilen ROUGE-L (0,5129) değerleri Zhang ve arkadaşlarının sonuçlarına yakın düzeydedir. Ancak Zhang ve arkadaşlarının çalışması 1.181 hastadan oluşan çok daha büyük bir veri seti ve uçtan uca eğitilmiş özel bir mimari kullanırken, bu çalışmada yalnızca 120 eğitim örneği ve parametre-verimli ince ayar ile karşılaştırılabilir sonuçlara ulaşılması dikkat çekicidir.

Göğüs röntgeni tabanlı rapor üretim çalışmaları ile karşılaştırıldığında, Wang ve arkadaşları [70] R2GenGPT modeli ile IU-Xray veri setinde BLEU-4: 0,173 ve ROUGE-L: 0,377 değerleri elde etmişlerdir; bu çalışmada Mistral 7B'nin BLEU-4 değeri (0,3808) bu sonucun üzerindedir. Divya ve arkadaşları [67] Memory Guided Transformer modeli ile IU X-Ray veri setinde BLEU-1: 0,492 ve ROUGE-L: 0,379 değerlerine ulaşmışlardır; bu çalışmadaki Mistral 7B'nin BLEU-1 değeri (0,5505) ve Qwen 2.5 3B'nin ROUGE-L değeri (0,5129) her iki metrikte de bu sonuçların üzerindedir. Zeng ve arkadaşları [69] AERMNet modeli ile IU X-Ray veri setinde BLEU-1: 0,486 ve METEOR: 0,219 elde etmişlerdir; bu çalışmadaki METEOR değerleri (Qwen 2.5 3B: 0,4526; Qwen 2.5 7B: 0,4290) bu sonuçların yaklaşık iki katı düzeyindedir. Bununla birlikte bu karşılaştırmalar farklı modaliteler, farklı diller ve farklı rapor yapıları içerdiğinden dikkatli yorumlanmalıdır.

Bu tez çalışmasının rapor üretimi açısından en önemli farkı, modele doğrudan ham görüntü verilmek yerine, segmentasyon ve anatomik eşleme aşamalarından elde edilen yapılandırılmış klinik bilgilerin girdi olarak kullanılmasıdır. Bu yaklaşım, modelin halüsinasyon riskini azaltmakta ve üretilen raporların klinik tutarlılığını artırmaktadır. Literatürdeki çoğu çalışma görüntüden doğrudan metin üretimi paradigmasını benimserken, bu çalışmada benimsenen yapılandırılmış bilgiden metin üretimi yaklaşımı, rapor içeriğinin kontrol edilebilirliğini ve açıklanabilirliğini önemli ölçüde artırmaktadır.

Bu tez çalışmasının önemli katkılarından biri de segmentasyon maskelerinden yalnızca lezyon alanının belirlenmesiyle yetinmeyip, klinik açıdan anlamlı yapılandırılmış bilgilerin otomatik olarak çıkarılmasıdır. Lezyon hacmi, anatomik lokalizasyon, hemisfer bilgisi ve vasküler sulama alanı gibi özelliklerin otomatik olarak belirlenmesi, segmentasyon çıktılarının doğrudan klinik yorumlama süreçlerine entegre edilmesine olanak sağlamıştır. Harvard-Oxford atlasları ve vasküler atlasların kullanılması sayesinde görüntü tabanlı bulgular semantik klinik bilgiye dönüştürülebilmiş ve bu yapılandırılmış bilgiler büyük dil modelleri için anlamlı girdiler oluşturmuştur.

Bu yapılandırılmış bilgi yaklaşımı, klasik görüntüden metin üretimine dayalı rapor üretim sistemlerine kıyasla iki temel avantaj sunmaktadır: birincisi modelin halüsinasyon yapma riskini azaltmaktadır, çünkü model yalnızca kendisine verilen yapılandırılmış bulgular

üzerinden rapor üretmektedir; ikincisi rapor içeriğinin izlenebilirliğini ve açıklanabilirliğini artırmaktadır, çünkü üretilen rapordaki her bulgu görüntü analizinden elde edilmiş somut bir veri noktasına dayandırılabilmektedir. Bununla birlikte, mevcut anatomik eşleme sürecinin bazı sınırlılıkları da gözlemlenmiştir. Düzlem hizalama yönteminin kullanılması, atlas ile hasta anatomisi arasında kaba bir hizalama sağlamakla birlikte, özellikle geniş lezyonların beyin anatomisini deforme ettiği durumlarda doğruluğu sınırlı kalabilmektedir. Deformable registration yöntemlerinin kullanılması bu sorunu kısmen giderebilir; ancak deformable yöntemlerin lezyonlu bölgelerde güvenilirliği de tartışmalıdır ve lezyon maskeleme stratejileri ile uygulanması gerekebilir.

Bu çalışmanın sonuçlarının yorumlanmasında dikkate alınması gereken en kritik sınırlılıklardan biri veri seti büyüklüğüdür. 150 hastalık segmentasyon veri seti ve 120 örneklik LLM eğitim veri seti, özellikle büyük parametre boyutlu modellerin tam potansiyelini ortaya çıkarmak için yetersiz kalmış olabilir; 14B modelin düşük performansı bu hipotezi desteklemektedir. Literatürde başarılı tıbbi görüntü segmentasyonu ve rapor üretimi çalışmalarının çoğunlukla çok daha büyük veri setleri üzerinde gerçekleştirildiği görülmektedir: ISLES 2022 veri seti 400 hastadan, ATLAS v2.0 1.271 hastadan, Zhang ve arkadaşlarının [51] StrokeQD veri seti ise 1.181 hastadan oluşmaktadır. Göğüs röntgeni tabanlı rapor üretim çalışmalarında kullanılan MIMIC-CXR veri seti yaklaşık 227.835 görüntü içermektedir. Bu çalışmadaki 150 hastalık veri seti, bu büyüklüklerin çok altında kalmaktadır.

Veri seti büyüklüğünün sınırlı olmasının çalışmanın farklı bileşenlerine etkisi de farklı düzeylerdedir. Segmentasyon aşamasında nnU-Net'in otomatik yapılandırma mekanizması sayesinde 150 hastayla bile rekabetçi sonuçlar elde edilmiştir; bu durum, güçlü tümevarımsal önyargıya sahip mimarilerin sınırlı veri koşullarında avantajlı olduğunu göstermektedir. Buna karşın, LLM ince ayar aşamasında 120 eğitim örneği, özellikle büyük modellerin (14B) adaptasyon kapasitesini ciddi biçimde sınırlamıştır. Bu gözlem, farklı sistem bileşenlerinin veri ihtiyacının birbirinden bağımsız olarak değerlendirilmesi gerektiğini ortaya koymaktadır.

Veri setinin çeşitliliği açısından da önemli sınırlılıklar bulunmaktadır. Tüm veriler tek bir merkezden (Marmara Üniversitesi Eğitim ve Araştırma Hastanesi) toplanmış olup belirli bir MRI cihazı, görüntüleme protokolü ve hasta popülasyonunu yansıtmaktadır. Bu durum, modelin farklı MRI cihazları (farklı üreticiler, farklı manyetik alan güçleri), farklı görüntüleme protokolleri (farklı b-değerleri, farklı kesit kalınlıkları) ve farklı hasta demografileri üzerindeki genelleme performansının değerlendirilmemiş olması anlamına gelmektedir. Tek merkezli çalışmaların görüntüleme protokol homojenliğinden kaynaklanan yapay performans artışı riski taşıdığı bilinmektedir; bu nedenle çok merkezli doğrulama çalışmaları olmadan elde edilen sonuçların genellenebilirliği konusunda dikkatli olunması gerekmektedir.

Ayrıca, veri setindeki radyoloji raporlarının yalnızca tek bir hastaneden ve sınırlı sayıda radyologdan elde edilmiş olması, rapor terminolojisi ve yapısının belirli bir klinik pratiğin yansıması olduğu anlamına gelmektedir. Farklı hastanelerde radyolojik raporlama stilleri, terminoloji tercihleri ve rapor uzunluğu önemli ölçüde farklılık gösterebilmektedir. Bu nedenle, geliştirilen rapor üretim modelinin farklı raporlama geleneklerine sahip kliniklerde performansının ayrıca değerlendirilmesi gerekmektedir.

Segmentasyon modellerinde yalnızca DWI ve ADC sekansları kullanılmıştır. FLAIR, PWI, T2 veya Susceptibility Weighted Imaging gibi ek sekansların dahil edilmesi, çalışmanın klinik kapsamını önemli ölçüde genişletebilirdi. FLAIR sekansının eklenmesi özellikle iki kritik avantaj sağlayabilir: birincisi, DWI-FLAIR uyumsuzluğu inme zamanlamasının belirlenmesinde klinik açıdan değerli bir göstergedir ve otomatik tespit edilmesi akut-subakut lezyon ayrımını kolaylaştırabilir; ikincisi, FLAIR görüntüleri kronik vasküler değişikliklerin ve lökoareozis gibi eşlik eden bulguların tanımlanmasına olanak sağlamaktadır. PWI sekansının eklenmesi ise penumbra (kurtarılabilir beyin dokusu) değerlendirmesini mümkün kılarak, klinik karar destek sisteminin terapötik karar süreçlerine doğrudan katkı sağlamasına imkân tanıyabilir. Mevcut çalışmada PWI verisinin bulunmaması, sistemin yalnızca mevcut enfarkt alanını değerlendirmesini mümkün kılmakta, ancak risk altındaki dokunun tahmin edilmesini engellenmektedir.

Bunun ötesinde, mevcut sistemin yalnızca iskemik inme lezyonlarına odaklanması da bir sınırlılık olarak değerlendirilebilir. Klinik pratikte beyin MR görüntülemesinde iskemik inme ile hemorajik dönüşüm, mikrokanamalar, kitle lezyonları, vasküler malformasyonlar veya enflamatuar değişiklikler gibi eşlik eden patolojiler sıklıkla karşılaşılmaktadır. Mevcut sistemin bu tür ek bulguları tanımlayamaması, üretilen raporların kapsamını kısıtlamaktadır.

Rapor üretimi açısından da klinik bilgi entegrasyonunun genişletilebileceği alanlar bulunmaktadır. Mevcut sistem segmentasyon maskesinden hacim, konum, anatomik bölge ve vasküler alan gibi temel özellikleri çıkarmaktadır; ancak klinik karar verme sürecinde hasta yaşı, klinik semptomlar, onset zamanı, NIHSS skoru ve tedavi öncesi/sonrası karşılaştırma gibi ek klinik bilgilerin de rapor üretim sürecine entegre edilmesi, üretilen raporların klinik değerini önemli ölçüde artırabilir. Elektronik sağlık kayıtlarından elde edilen bu tür bilgilerin yapılandırılmış girdi şablonuna dahil edilmesi, modelin daha kişiselleştirilmiş ve klinik bağlama duyarlı raporlar üretmesini sağlayabilir.

Bu çalışmada kullanılan değerlendirme metodolojisi, otomatik NLP metrikleri, klinik içerik doğruluk metrikleri ve uzman değerlendirmesi olmak üzere üç katmandan oluşmaktadır. Ancak her bir katmanın belirli sınırlılıkları bulunmaktadır. BLEU, ROUGE ve METEOR gibi otomatik metrikler yüzeysel kelime örtüşmesine dayandığından, aynı klinik anlamı farklı sözcüklerle ifade eden raporları yeterince değerlendiremeyebilmektedir. BERTScore bu sorunu kısmen çözmekle birlikte, kullanılan genel amaçlı BERT modelinin Türkçe tıbbi terminolojiyi yeterince temsil edememesi nedeniyle değerlendirme doğruluğu sınırlı kalabilmektedir. Türkçe tıbbi dil modelleri (örneğin BioTurkBERT veya benzeri alan-özgü modeller) kullanılarak hesaplanan BERTScore değerlerinin, klinik uyumu daha doğru yansıtması beklenmektedir.

Uzman değerlendirmesi çalışmanın güçlü yönlerinden biri olmakla birlikte, değerlendiricilerin sayısının ve değerlendirilen örnek sayısının sınırlı olması, sonuçların genellenebilirliğini kısıtlamaktadır. Değerlendiriciler arası güvenilirlik ölçütlerinin raporlanması ve daha büyük bir değerlendirme panelinin oluşturulması, gelecek çalışmalarda uzman değerlendirmesinin gücünü artırabilir. Ayrıca, klinik doğruluk metriklerinin otomatik hesaplanmasında kullanılan kural tabanlı yaklaşımların, karmaşık klinik ifadeleri tam olarak yakalayamaması da bir sınırlılık olarak değerlendirilebilir.

Segmentasyon performansının değerlendirilmesinde kullanılan 5-katlı çapraz doğrulama yöntemi, 150 hastalık veri setinin boyutu göz önünde bulundurulduğunda uygun bir strateji

olarak değerlendirilmektedir. Ancak her katlama yalnızca 30 test hastası içerdiğinden, özellikle küçük ve nadir lokalizasyonlu lezyonların değerlendirmesinde istatistiksel güvenilirlik sınırlı kalabilmektedir. Ayrıca, Dice ve IoU gibi örtüşme tabanlı metriklerin küçük lezyonlarda aşırı duyarlı olması (birkaç voxel hatası bile skorları ciddi biçimde düşürebilir) ve büyük lezyonlarda daha toleranslı davranması, performans değerlendirmesinin lezyon boyutuna göre tabakalandırılması gerekliliğini ortaya koymaktadır.

Geliştirilen sistemin klinik pratikte uygulanabilirliği değerlendirildiğinde, çeşitli pratik ve etik boyutların dikkate alınması gerekmektedir. Öncelikle, otomatik üretilen radyolojik raporların klinik kullanımı, mevcut sağlık mevzuatı çerçevesinde radyolog onayını gerektirmektedir. Bu nedenle, geliştirilen sistem bir radyolog yerine geçmeyi değil, radyologa taslak rapor sunarak raporlama sürecini hızlandırmayı hedeflemektedir. Bu bağlamda, sistemin klinik iş akışına entegrasyonu sırasında "insan-döngüsünde" (human-in-the-loop) bir yaklaşımın benimsenmesi zorunludur.

Halüsinasyon problemi, büyük dil modellerinin klinik ortamlarda kullanılmasının önündeki en önemli engellerden biridir. Bu çalışmada yapılandırılmış bilgiden metin üretimi yaklaşımının benimsenmesi, halüsinasyon riskini önemli ölçüde azaltmaktadır; ancak modelin verilen bilgilerin ötesinde çıkarımlar yapması veya uygunsuz klinik ifadeler üretmesi riski tamamen ortadan kalkmamıştır. Halüsinasyon tespiti ve düzeltmesi için otomatik doğrulama mekanizmalarının geliştirilmesi, örneğin üretilen rapordaki anatomik ve klinik ifadelerin girdi yapılandırılmış verilerle karşılaştırılması yoluyla tutarlılık kontrolü yapılması, sistemin klinik güvenilirliğini artıracaktır.

Veri gizliliği ve etik boyutlar açısından, bu çalışmada açık kaynak modellerin yerel donanım üzerinde çalıştırılması, hasta verilerinin üçüncü taraf sunuculara gönderilmesi gerekliliğini ortadan kaldırmaktadır. Bu durum, ticari API tabanlı çözümlere kıyasla önemli bir avantaj sunmaktadır. Ancak klinik uygulamada modelin düzenli olarak güncellenmesi, yeni veri setleriyle yeniden eğitilmesi ve performansının sürekli izlenmesi gerekecektir. Bu süreçlerin standartlaştırılması ve uygun denetim mekanizmalarının oluşturulması, sistemin klinik kabul görmesi açısından kritik öneme sahiptir.

Ayrıca, otomatik rapor üretim sistemlerinin potansiyel yanlılık (bias) riskleri de göz ardı edilmemelidir. Eğitim veri setinin belirli bir hasta popülasyonundan (yaş, cinsiyet, etnisite, inme tipi dağılımı) toplanmış olması, modelin bu popülasyonun dışındaki hasta gruplarında farklı performans göstermesine yol açabilir. Bu nedenle, sistemin klinik doğrulaması sırasında çeşitli demografik gruplar üzerinde performansının değerlendirilmesi ve olası yanlılıkların belirlenmesi önemlidir.

4.1. Gelecek Çalışmalar

Elde edilen sonuçlar ve tespit edilen sınırlılıklar doğrultusunda gelecek çalışmalar için çeşitli öneriler sunulmaktadır. Farklı hastaneler ve MRI cihazlarından elde edilen verilerle eğitim setinin genişletilmesi modelin genelleme yeteneğini artıracaktır. Özellikle 500–1.000 hastalık çok merkezli bir eğitim seti ile büyük modellerin performansının yeniden değerlendirilmesi önerilmektedir. Çok merkezli veri toplama sürecinde farklı MRI üreticileri (Siemens, GE, Philips), farklı alan güçleri (1,5T ve 3T) ve farklı görüntüleme protokollerinin temsil edilmesi, modelin gerçek klinik koşullardaki performansının daha güvenilir biçimde değerlendirilmesini sağlayacaktır. Buna ek olarak, ISLES 2022 ve

ATLAS v2.0 gibi halka açık veri setlerinin eğitim sürecine dahil edilmesi, veri çeşitliliğini artırarak modelin genelleme kapasitesini güçlendirebilir.

FLAIR ve PWI sekanslarının segmentasyon sürecine dahil edilmesi, akut-subakut ayrımı ve penumbra değerlendirmesi açısından önemli katkılar sağlayabilir. DWI-FLAIR uyumsuzluğu, inme zamanlamasının belirlenmesinde klinik açıdan değerli bir gösterge olarak sisteme entegre edilebilir. Bunun ötesinde, SWI sekansının eklenmesi hemorajik dönüşümün otomatik tespitini mümkün kılarak raporun klinik kapsamını genişletebilir. DTI verisinin entegrasyonu ise beyaz cevher yollarının lezyon tarafından etkilenme derecesinin değerlendirilmesine olanak sağlayarak rehabilitasyon planlaması açısından önemli bilgiler sunabilir.

Bu çalışmada erken füzyon stratejisinin sınırlılıkları açıkça ortaya konmuştur. Gelecek çalışmalarda, her bir bilgi kaynağının ayrı kodlayıcı dallarında işlendiği ve dikkat mekanizmaları aracılığıyla birleştirildiği çok akışlı mimari yaklaşımları değerlendirilebilir. Çapraz dikkat mekanizmaları, farklı modaliteler arasındaki komplementer bilgiyi daha etkin biçimde kullanarak segmentasyon performansını artırabilir. Ayrıca, öğrenilebilir füzyon ağırlıkları kullanılarak her bir bilgi kaynağının modele optimal katkısının otomatik olarak belirlenmesi sağlanabilir. Dinamik füzyon stratejileri, farklı hasta vakaları için farklı modalite kombinasyonlarının otomatik olarak seçilmesine olanak tanıyarak, hasta-spesifik segmentasyon performansını iyileştirebilir.

Reinforcement Learning from Human Feedback veya Direct Preference Optimization gibi ileri eğitim stratejilerinin uygulanması, rapor kalitesini ve klinik tutarlılığı artırabilir. Özellikle DPO, radyolog tercihleri doğrultusunda modelin çıktılarını optimize etmek için etkili bir yöntem olabilir. Ayrıca, sentetik veri artırma yöntemleri kullanılarak eğitim veri setinin genişletilmesi düşünülmelidir; mevcut raporların yapılandırılmış bilgilerinde küçük varyasyonlar oluşturularak (farklı lezyon hacimleri, farklı anatomik bölgeler, farklı vasküler alanlar) eğitim örneği sayısı artırılabilir. Halüsinasyon tespiti ve düzeltmesi için otomatik doğrulama mekanizmalarının geliştirilmesi, örneğin üretilen rapordaki anatomik terimlerin ve vasküler alan bilgisinin yapılandırılmış girdiyle karşılaştırılarak tutarlılığının kontrol edilmesi, sistemin klinik güvenilirliğini önemli ölçüde artıracaktır.

Genel amaçlı NLP metrikleri (BLEU, ROUGE) tıbbi raporların değerlendirilmesinde tam olarak yeterli olmayabilir. RadGraph tabanlı klinik etkinlik metrikleri veya Türkçe tıbbi dil modelleri kullanılarak hesaplanan BERTScore gibi alana özgü değerlendirme ölçütlerinin geliştirilmesi, rapor kalitesinin daha doğru biçimde ölçülmesini sağlayabilir. Ayrıca, F1-RadGraph veya CheXbert gibi radyoloji odaklı metriklerin Türkçe'ye uyarlanması, gelecek çalışmaların değerlendirme kalitesini artırabilir.

Geliştirilen sistemin yalnızca iskemik inme ile sınırlı kalmayıp hemorajik inme, beyin tümörleri, demiyelinizan lezyonlar ve diğer intrakranial patolojilere genişletilmesi, daha kapsamlı bir nöroradyoloji karar destek sistemi oluşturulmasına katkı sağlayabilir. Bu genişleme, çoklu sınıf segmentasyonu ve patoloji-özgü rapor şablonları gerektirecek olup, sistemin mimarisinde önemli adaptasyonları beraberinde getirecektir.

Klinik ortamda sistemin kullanılabilirliği, çıkarım hızına doğrudan bağlıdır. Model sıkıştırma (pruning, distillation), niceleme (quantization) ve ONNX/TensorRT optimizasyonları ile çıkarım süresinin klinik iş akışına uygun seviyelere düşürülmesi önerilmektedir. Özellikle segmentasyon ve rapor üretimi aşamalarının toplam çıkarım

süresinin birkaç dakikayı geçmemesi, sistemin klinik kabul görmesi açısından önemli bir gereklilik olarak değerlendirilmektedir.

Sonuç olarak bu tez çalışması, akut iskemik inmenin otomatik segmentasyonu ile büyük dil modelleri tabanlı radyolojik rapor üretimini aynı sistem içerisinde bütünleştiren kapsamlı bir yapay zekâ yaklaşımı sunmaktadır. Segmentasyon aşamasında nnU-Net modeli DWI konfigürasyonunda %71,66 Dice skoru ile en yüksek performansı sağlamış ve model mimarisi seçiminin giriş konfigürasyonundan daha belirleyici bir faktör olduğu istatistiksel olarak gösterilmiştir. Otomatik rapor üretiminde ise çok dilli küçük-orta boyutlu modellerin (Qwen 2.5 3B: BERTScore 0,7630; Mistral 7B: BLEU-4 0,3808) 120 eğitim örneği ve 12 GB VRAM kısıtı altında Türkçe radyoloji rapor üretiminde en etkili seçenekler olduğu belirlenmiştir. Çok dilli modellerin Türkçe-özgü modellere kıyasla alan-spesifik ince ayarda belirgin üstünlük sağladığı ve model boyutu artışının sınırlı eğitim verisi koşullarında performansı olumsuz etkileyebildiği ortaya konmuştur.

Çalışmanın iki temel kısıtı olan tek merkezli ve sınırlı büyüklükteki veri seti ile yalnızca DWI ve ADC sekanslarının kullanılması, sonuçların genellenebilirliğini sınırlamaktadır. Bununla birlikte, yapılandırılmış bilgiden metin üretimi paradigmasının halüsinasyon riskini azaltması, erken füzyon stratejisinin sınırlılıklarının deneysel olarak ortaya konması ve sınırlı kaynakla (120 örnek, 12 GB VRAM) rekabetçi sonuçlara ulaşılması, çalışmanın önemli metodolojik katkıları olarak değerlendirilmektedir. Elde edilen bulgular, özellikle anatomik bilgi çıkarımı ile büyük dil modellerinin birlikte kullanılmasının, gelecekte açıklanabilir ve klinik açıdan daha güvenilir yapay zekâ tabanlı radyoloji sistemlerinin geliştirilmesine katkı sağlayabilecek önemli bir yaklaşım olduğunu göstermektedir. Çok merkezli veri toplama, ek görüntüleme sekanslarının entegrasyonu, gelişmiş füzyon stratejilerinin uygulanması ve ileri LLM eğitim tekniklerinin kullanılması, bu yaklaşımın klinik uygulanabilirliğini güçlendirecek gelecek araştırma yönleri olarak öne çıkmaktadır.

KAYNAKLAR

[1] L. Cui *et al.*, "Deep learning in ischemic stroke imaging analysis: a comprehensive review," *BioMed Research International,* vol. 2022, no. 1, p. 2456550, 2022.

[2] V. L. Feigin *et al.*, "Global, regional, and national burden of stroke and its risk factors, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021," *The Lancet Neurology,* vol. 23, no. 10, pp. 973-1003, 2024.

[3] M. Vijayan and P. H. Reddy, "Peripheral biomarkers of stroke: focus on circulatory microRNAs," *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease,* vol. 1862, no. 10, pp. 1984-1993, 2016.

[4] M. Vijayan, S. Kumar, J. Bhatti, and P. Reddy, "Molecular links and biomarkers of stroke, vascular dementia, and Alzheimer's disease," *Progress in molecular biology and translational science,* vol. 146, pp. 95-126, 2017.

[5] V. E. O'Collins, G. A. Donnan, M. R. Macleod, and D. W. Howells, "Animal models of ischemic stroke versus clinical stroke: comparison of infarct size, cause, location, study design, and efficacy of experimental therapies," in *Animal models for the study of human disease*: Elsevier, 2017, pp. 481-523.

[6] D. Y.-W. Fann, S.-Y. Lee, S. Manzanero, P. Chunduri, C. G. Sobey, and T. V. Arumugam, "Pathogenesis of acute stroke and the role of inflammasomes," *Ageing research reviews,* vol. 12, no. 4, pp. 941-966, 2013.

[7] M. A. Topçuoğlu, "Stroke epidemiology and near future projection in Turkey: analysis

of Turkey data from the Global Burden of Disease Study," *Turkish Journal of Neurology,* vol. 28, no. 4, pp. 200-211, 2022.
[8] A. SERÇE, E. K. UMAY, and F. A. ÇAKÇI, "Early Intensive Multi-faceted Rehabilitation in Stroke Patients: What is the Best Effective Rehabilitation Time?," *Journal of Physical Medicine & Rehabilitation Sciences,* vol. 24, no. 3, 2021.
[9] C. S. Kidwell and S. Warach, "Acute ischemic cerebrovascular syndrome: diagnostic criteria," *Stroke,* vol. 34, no. 12, pp. 2995-2998, 2003.
[10] A. E. Baird *et al.*, "Clinical correlations of diffusion and perfusion lesion volumes in acute ischemic stroke," *Cerebrovascular Diseases,* vol. 10, no. 6, pp. 441-448, 2000.
[11] M. G. Lansberg, G. W. Albers, C. Beaulieu, and M. P. Marks, "Comparison of diffusion-weighted MRI and CT in acute stroke," *Neurology,* vol. 54, no. 8, pp. 1557-1561, 2000.
[12] A. Löve, R. Siemund, G. Andsberg, M. Cronqvist, S. Holtås, and I. Björkman-Burtscher, "Comprehensive CT evaluation in acute ischemic stroke: impact on diagnosis and treatment decisions," *Stroke research and treatment,* vol. 2011, no. 1, p. 726573, 2011.
[13] H. Kuo, T.-W. Liu, Y.-P. Huang, S.-C. Chin, L.-S. Ro, and H.-C. Kuo, "Differential diagnostic value of machine learning–based models for embolic stroke," *Clinical and Applied Thrombosis/Hemostasis,* vol. 29, p. 10760296231203663, 2023.
[14] O. Y. Bang and W. Li, "Applications of diffusion-weighted imaging in diagnosis, evaluation, and treatment of acute ischemic stroke," *Precision and Future Medicine,* vol. 3, no. 2, pp. 69-76, 2019.
[15] S. Yaghi *et al.*, "Itemized NIHSS subsets predict positive MRI strokes in patients with mild deficits," *Journal of the neurological sciences,* vol. 358, no. 1-2, pp. 221-225, 2015.
[16] G. Schlaug, B. Siewert, A. Benfield, R. Edelman, and S. Warach, "Time course of the apparent diffusion coefficient (ADC) abnormality in human stroke," *Neurology,* vol. 49, no. 1, pp. 113-119, 1997.
[17] B. J. Kim *et al.*, "Diffusion-weighted image and fluid-attenuated inversion recovery image mismatch: unclear-onset versus clear-onset stroke," *Stroke,* vol. 45, no. 2, pp. 450-455, 2014.
[18] S. Gaddamanugu *et al.*, "Clinical applications of diffusion-weighted sequence in brain imaging: beyond stroke," *Neuroradiology,* vol. 64, no. 1, pp. 15-30, 2022.
[19] R. S. Barros *et al.*, "Automatic segmentation of cerebral infarcts in follow-up computed tomography images with convolutional neural networks," *Journal of neurointerventional surgery,* vol. 12, no. 9, pp. 848-852, 2020.
[20] N. Noor, N. Saad, A. Abdullah, and N. Ali, "Automated segmentation and classification technique for brain stroke," *International Journal of Electrical and Computer Engineering,* vol. 9, no. 3, p. 1832, 2019.
[21] H. Kuang, B. K. Menon, and W. Qiu, "Segmenting hemorrhagic and ischemic infarct simultaneously from follow-up non-contrast CT images in patients with acute ischemic stroke," *IEEE Access,* vol. 7, pp. 39842-39851, 2019.
[22] S. Uçkun, M. Ağralı, and V. Kılıç, "Deep Learning-Based Ischemic Stroke Segmentation on Brain Computed Tomography Images," *Avrupa Bilim ve Teknoloji Dergisi,* no. 50, pp. 105-112, 2023.
[23] E. Ruthra and A. R. Bevi, "Ischemic stroke lesion segmentation in the acute phase using automated deep U-Net model," *Health and Technology,* pp. 1-10, 2026.
[24] Z. Li *et al.*, "A novel multi-scale channel attention-guided neural network for brain stroke lesion segmentation," *Ieee access,* vol. 11, pp. 66050-66062, 2023.
[25] M. Angulakshmi and G. Lakshmi Priya, "Automated brain tumour segmentation

techniques—a review," *International Journal of Imaging Systems and Technology,* vol. 27, no. 1, pp. 66-77, 2017.
[26] A. M. A. A. A. Isa *et al.*, "A review of MRI acute ischemic stroke lesion segmentation," *International Journal of Integrated Engineering,* vol. 12, no. 6, pp. 117-127, 2020.
[27] D. L. Pham, C. Xu, and J. L. Prince, "Current methods in medical image segmentation," *Annual review of biomedical engineering,* vol. 2, no. 1, pp. 315-337, 2000.
[28] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical image analysis,* vol. 42, pp. 60-88, 2017.
[29] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature,* vol. 521, no. 7553, pp. 436-444, 2015.
[30] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-net and its variants for medical image segmentation: A review of theory and applications," *IEEE access,* vol. 9, pp. 82031-82057, 2021.
[31] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "Doubleu-net: A deep convolutional neural network for medical image segmentation," in *2020 IEEE 33rd International symposium on computer-based medical systems (CBMS)*, 2020: IEEE, pp. 558-564.
[32] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015: Springer, pp. 234-241.
[33] Y.-F. Ko and K.-S. Cheng, "Semi-Siamese U-Net for separation of lung and heart bioimpedance images: A simulation study of thorax EIT," *PloS one,* vol. 16, no. 2, p. e0246071, 2021.
[34] S. Oh, Y.-J. Kim, Y.-T. Park, and K.-G. Kim, "Automatic pancreatic cyst lesion segmentation on EUS images using a deep-learning approach," *Sensors,* vol. 22, no. 1, p. 245, 2021.
[35] M. Nishio, K. Fujimoto, and K. Togashi, "Lung segmentation on chest X-ray images in patients with severe abnormal findings using deep learning," *International Journal of Imaging Systems and Technology,* vol. 31, no. 2, pp. 1002-1008, 2021.
[36] J. Jin *et al.*, "Multiple U-Net-based automatic segmentations and radiomics feature stability on ultrasound images for patients with ovarian cancer," *Frontiers in Oncology,* vol. 10, p. 614201, 2021.
[37] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *ieee Computational intelligenCe magazine,* vol. 13, no. 3, pp. 55-75, 2018.
[38] J. Hirschberg and C. D. Manning, "Advances in natural language processing," *Science,* vol. 349, no. 6245, pp. 261-266, 2015.
[39] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems,* vol. 30, 2017.
[40] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171-4186.
[41] T. Brown *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems,* vol. 33, pp. 1877-1901, 2020.
[42] R. Bommasani *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258,* 2021.

[43] L. Montenegro, L. M. Gomes, and J. M. Machado, "What We Know About the Role of Large Language Models for Medical Synthetic Dataset Generation," *AI,* vol. 6, no. 6, p. 109, 2025.
[44] Q. Xie, E. J. Schenck, H. S. Yang, Y. Chen, Y. Peng, and F. Wang, "Faithful AI in medicine: a systematic review with large language models and beyond," *MedRxiv,* 2023.
[45] T. Osborne *et al.*, "Towards Inpatient Discharge Summary Automation via Large Language Models: A Multidimensional Evaluation with a HIPAA-Compliant Instance of GPT-4o and Clinical Expert Assessment," *medRxiv,* p. 2025.04. 03.25325204, 2025.
[46] P. Ahadian and Q. Guan, "A survey on hallucination in large language and foundation models," 2025.
[47] S. Chen *et al.*, "MSA-YOLOv5: Multi-scale attention-based YOLOv5 for automatic detection of acute ischemic stroke from multi-modality MRI images," *Computers in Biology and Medicine,* vol. 165, p. 107471, 2023.
[48] M. R. Hernandez Petzsche *et al.*, "ISLES 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset," *Scientific data,* vol. 9, no. 1, p. 762, 2022.
[49] Y. Yu, J. J. Heit, and G. Zaharchuk, "Improving ischemic stroke care with MRI and deep learning artificial intelligence," *Topics in Magnetic Resonance Imaging,* vol. 30, no. 4, pp. 187-195, 2021.
[50] O. Maier *et al.*, "ISLES 2015-A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI," *Medical image analysis,* vol. 35, pp. 250-269, 2017.
[51] S. Zhang, L. Tan, Q. Han, H. Wang, and J. Meng, "Automatic report generation on a large-scale stroke MRI dataset," in *2023 IEEE 6th International Conference on Electronic Information and Communication Technology (ICEICT)*, 2023: IEEE, pp. 123-128.
[52] R. Ahmed, A. Al Shehhi, N. Werghi, and M. L. Seghier, "Segmentation of stroke lesions using transformers-augmented MRI analysis," *Human brain mapping,* vol. 45, no. 11, p. e26803, 2024.
[53] R. Daoudi, A. Mouelhi, and M. Sayadi, "Automatic ischemic stroke lesions segmentation in multimodality MRI using mask region-based convolutional neural network," in *2020 4th International Conference on Advanced Systems and Emergent Technologies (IC_ASET)*, 2020: IEEE, pp. 362-366.
[54] Y. Xue *et al.*, "A multi-path 2.5 dimensional convolutional neural network system for segmenting stroke lesions in brain MRI images," *NeuroImage: Clinical,* vol. 25, p. 102118, 2020.
[55] M. Nouman, M. Mabrok, and E. A. Rashed, "Neuro-TransUNet: A comprehensive transformer-based architecture for precise segmentation of stroke lesions in 3D MRI," in *Proceedings of the 2024 9th International Conference on Multimedia and Image Processing*, 2024, pp. 152-156.
[56] R. Geetha, E. Priya, and M. Vijayakumar, "An approach for automated acute cerebral ischemic stroke lesion segmentation and correlation of significant features with modified Rankin Scale," *Biomedical Signal Processing and Control,* vol. 100, p. 106921, 2025.
[57] A. Rahman *et al.*, "Deep learning-driven segmentation of ischemic stroke lesions using multi-channel MRI," *Biomedical Signal Processing and Control,* vol. 105, p. 107676, 2025.
[58] B. P. Garcia-Salgado *et al.*, "Enhanced ischemic stroke lesion segmentation in MRI

using attention U-Net with generalized dice focal loss," *Applied Sciences,* vol. 14, no. 18, p. 8183, 2024.
[59] Y. Wang, W. Wang, Q. Chen, K. Huang, A. Nguyen, and S. De, "Zero-shot text classification with knowledge resources under label-fully-unseen setting," *Neurocomputing,* vol. 610, p. 128580, 2024.
[60] T. Schopf, D. Braun, and F. Matthes, "Evaluating unsupervised text classification: zero-shot and similarity-based approaches," in *Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval*, 2022, pp. 6-15.
[61] Y. Makhija, E. De Brouwer, and R. G. Krishnan, "Learning predictive checklists with probabilistic logic programming," *arXiv preprint arXiv:2411.16790,* 2024.
[62] H. Wang *et al.*, "Kmtlabeler: An interactive knowledge-assisted labeling tool for medical text classification," *IEEE Transactions on Visualization and Computer Graphics,* 2024.
[63] A. S. Alammary, "Investigating the impact of pretraining corpora on the performance of Arabic BERT models," *The Journal of Supercomputing,* vol. 81, no. 1, p. 187, 2025.
[64] N. Guan, K. Chen, and N. Koudas, "Can large language models design accurate label functions?," *arXiv preprint arXiv:2311.00739,* 2023.
[65] C. Huang, Y. Xie, Z. Jiang, J. Lin, and M. Li, "Approximating human-like few-shot learning with gpt-based compression," *arXiv preprint arXiv:2308.06942,* 2023.
[66] K. M. Sünnetci, "Biomedical text-based detection of colon, lung, and thyroid cancer: A deep learning approach with novel dataset," *Displays,* vol. 89, p. 103068, 2025.
[67] P. Divya, Y. Sravani, C. Vishnu, C. K. Mohan, and Y. W. Chen, "Memory guided transformer with spatio-semantic visual extractor for medical report generation," *IEEE Journal of Biomedical and Health Informatics,* vol. 28, no. 5, pp. 3079-3089, 2024.
[68] G. V. Magalhães, R. L. d. S. Santos, L. H. Vogado, A. C. de Paiva, and P. d. A. dos Santos Neto, "XRaySwinGen: Automatic medical reporting for X-ray exams with multimodal model," *Heliyon,* vol. 10, no. 7, 2024.
[69] X. Zeng, T. Liao, L. Xu, and Z. Wang, "AERMNet: Attention-enhanced relational memory network for medical image report generation," *Computer Methods and Programs in Biomedicine,* vol. 244, p. 107979, 2024.
[70] Z. Wang, L. Liu, L. Wang, and L. Zhou, "R2gengpt: Radiology report generation with frozen llms," *Meta-Radiology,* vol. 1, no. 3, p. 100033, 2023.
[71] S. Yang, X. Wu, S. Ge, Z. Zheng, S. K. Zhou, and L. Xiao, "Radiology report generation with a learned knowledge base and multi-modal alignment," *Medical Image Analysis,* vol. 86, p. 102798, 2023.
[72] R. Kikinis, S. D. Pieper, and K. G. Vosburgh, "3D Slicer: a platform for subject-specific image analysis, visualization, and clinical support," in *Intraoperative imaging and image-guided therapy*: Springer, 2013, pp. 277-289.
[73] I. Sobel and G. Feldman, "A 3x3 isotropic gradient operator for image processing," *a talk at the Stanford Artificial Project in,* vol. 1968, pp. 271-272, 1968.
[74] B. C. Lowekamp, D. T. Chen, L. Ibáñez, and D. Blezek, "The design of SimpleITK," *Frontiers in neuroinformatics,* vol. 7, p. 45, 2013.
[75] C.-F. Liu *et al.*, "Digital 3D brain MRI arterial territories atlas," *Scientific Data,* vol. 10, no. 1, p. 74, 2023.
[76] F. Isensee, P. F. Jäger, P. M. Full, P. Vollmuth, and K. H. Maier-Hein, "nnU-Net for brain tumor segmentation," in *International MICCAI Brainlesion Workshop*, 2020: Springer, pp. 118-132.
[77] S. Guo, Q. Chen, L. Wang, L. Wang, and Y. Zhu, "nnUnetFormer: an automatic

method based on nnUnet and transformer for brain tumor segmentation with multimodal MR images," *Physics in Medicine & Biology,* vol. 68, no. 24, p. 245012, 2023.
[78] B. Yu, Q. Zhou, L. Yuan, H. Liang, P. Shcherbakov, and X. Zhang, "3D medical image segmentation using the serial–parallel convolutional neural network and transformer based on cross-window self-attention," *CAAI Transactions on Intelligence Technology,* vol. 10, no. 2, pp. 337-348, 2025.
[79] Y. He, V. Nath, D. Yang, Y. Tang, A. Myronenko, and D. Xu, "Swinunetr-v2: Stronger swin transformers with stagewise convolutions for 3d medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2023: Springer, pp. 416-426.
[80] Y. Cai *et al.*, "Swin Unet3D: a three-dimensional medical image segmentation network combining vision transformer and convolution," *BMC medical informatics and decision making,* vol. 23, no. 1, p. 33, 2023.
[81] H. Cao *et al.*, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *European conference on computer vision*, 2022: Springer, pp. 205-218.
[82] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*, 2016: Springer, pp. 424-432.
[83] A. V. Dobshik *et al.*, "Acute ischemic stroke lesion segmentation in non-contrast CT images using 3D convolutional neural networks," *Компьютерная оптика,* vol. 47, no. 5, pp. 770-777, 2023.
[84] K. H. Zou *et al.*, "Statistical validation of image segmentation quality based on a spatial overlap index1: scientific reports," *Academic radiology,* vol. 11, no. 2, pp. 178-189, 2004.
[85] M. A. Rahman and Y. Wang, "Optimizing intersection-over-union in deep neural networks for image segmentation," in *International symposium on visual computing*, 2016: Springer, pp. 234-244.
[86] M. W. Woolrich *et al.*, "Bayesian analysis of neuroimaging data in FSL," *Neuroimage,* vol. 45, no. 1, pp. S173-S186, 2009.
[87] J. M. Goldstein *et al.*, "Hypothalamic abnormalities in schizophrenia: sex effects and genetic vulnerability," *Biological psychiatry,* vol. 61, no. 8, pp. 935-945, 2007.
[88] J. A. Frazier *et al.*, "Structural brain magnetic resonance imaging of limbic and thalamic volumes in pediatric bipolar disorder," *American Journal of Psychiatry,* vol. 162, no. 7, pp. 1256-1265, 2005.
[89] B. Hui *et al.*, "Qwen2. 5-coder technical report," *arXiv preprint arXiv:2409.12186,* 2024.
[90] A. Q. Jiang *et al.*, "Mistral 7b. arxiv," *arXiv preprint arXiv:2310.06825,* vol. 10, p. 3, 2023.
[91] *Kumru: A Turkish large language model*. (2025). Accessed: 2026 May 5. [Online]. Available: https://huggingface.co/vngrs-ai/Kumru-2B
[92] H. T. Kesgin *et al.*, "Optimizing large language models for Turkish: New methodologies in corpus selection and training," in *2024 Innovations in Intelligent Systems and Applications Conference (ASYU)*, 2024: IEEE, pp. 1-6.
[93] A. Sellergren *et al.*, "Medgemma technical report," *arXiv preprint arXiv:2507.05201,* 2025.
[94] A. Hurst *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276,* 2024.
[95] Anthropic, "Claude 3.5 Sonnet [Large language model]," 2024. [Online]. Available:

https://www.anthropic.com.
[96] G. Team *et al.*, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *arXiv preprint arXiv:2403.05530,* 2024.
[97] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," *Advances in neural information processing systems,* vol. 36, pp. 10088-10115, 2023.
[98] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311-318.
[99] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65-72.
[100] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675,* 2019.
[101] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74-81.
[102] H. Hsu and P. A. Lachenbruch, "Paired t test," *Wiley StatsRef: statistics reference online,* 2014.
[103] T. K. Kim, "T test as a parametric statistic," *Korean journal of anesthesiology,* vol. 68, no. 6, pp. 540-546, 2015.
[104] J. Cohen, *Statistical power analysis for the behavioral sciences*. routledge, 2013.
[105] M. Pinar, A. B. Altinel, and A. Aktas, "HeaRT: An Innovative Health Representation Transformer With Clustered Feature Augmentation for Biomedical Text Classification," *IEEE Access,* vol. 13, pp. 215748-215770, 2025.
[106] H. Christian, M. P. Agus, and D. Suhartono, "Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF)," *ComTech: Computer, Mathematics and Engineering Applications,* vol. 7, no. 4, pp. 285-294, 2016.
[107] M. Lan, C. L. Tan, J. Su, and Y. Lu, "Supervised and traditional term weighting methods for automatic text categorization," *IEEE transactions on pattern analysis and machine intelligence,* vol. 31, no. 4, pp. 721-735, 2008.
[108] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, 2019, pp. 3982-3992.
[109] N. O. F. Elssied, O. Ibrahim, and A. H. Osman, "A novel feature selection based on one-way anova f-test for e-mail spam classification," *Research Journal of Applied Sciences, Engineering and Technology,* vol. 7, no. 3, pp. 625-638, 2014.
[110] A. V. Ponce-Bobadilla, V. Schmitt, C. S. Maier, S. Mensing, and S. Stodtmann, "Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development," *Clinical and translational science,* vol. 17, no. 11, p. e70056, 2024.
[111] I. D. Mienye and Y. Sun, "A survey of ensemble learning: Concepts, algorithms, applications, and prospects," *Ieee Access,* vol. 10, pp. 99129-99149, 2022.
[112] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning,* vol. 63, no. 1, pp. 3-42, 2006.
[113] G. Ke *et al.*, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems,* vol. 30, 2017.
[114] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, "Online passive-

aggressive algorithms," *Journal of Machine Learning Research,* vol. 7, no. Mar, pp. 551-585, 2006.
[115] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random forests," in *Ensemble machine learning*: Springer, 2012, pp. 157-175.
[116] M. Yasir, A. M. Karim, S. K. Malik, A. A. Bajaffer, and E. I. Azhar, "Application of decision-tree-based machine learning algorithms for prediction of antimicrobial resistance," *Antibiotics,* vol. 11, no. 11, p. 1593, 2022.
[117] T. Chen *et al.*, "Xgboost: extreme gradient boosting," *R package version 0.4-2,* vol. 1, no. 4, pp. 1-4, 2015.
[118] T. Zhang, "Solving large scale linear prediction problems using stochastic gradient descent algorithms," in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 116.
[119] T. Zhou, S. Ruan, and S. Canu, "A review: Deep learning for medical image segmentation using multi-modality fusion," *Array,* vol. 3, p. 100004, 2019.
[120] Z. Zhu, X. He, G. Qi, Y. Li, B. Cong, and Y. Liu, "Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI," *Information Fusion,* vol. 91, pp. 376-387, 2023.

ÖZGEÇMİŞ

Adı Soyadı : Merve PINAR

Öğrenim Bilgileri

Derece	Bölüm/Program	Üniversite	Yıl
Lisans	Bilgisayar Mühendisliği	Çanakkale 18 Mart Üniversitesi	2009-2013
Yüksek Lisans	Bilgisayar Mühendisliği	Marmara Üniversitesi	2019-2022

Görevler

Görev Yeri	Görevi	Yıl
Marmara Üniversitesi	Araştırma Görevlisi	2021-Halen
Üsküdar Üniversitesi	Araştırma Görevlisi	2019-2020
Türk Telekom	Oracle DBA	2015-2016
Grid Teknoloji	Oracle DBA	2013-2015

Dergilerde Yayımlanan Makaleler

1. HeaRT: An Innovative Health Representation Transformer With Clustered Feature Augmentation for Biomedical Text Classification
PINAR M., Altinel A. B., AKTAŞ A.
IEEE Access , cilt.13, ss.215748-215770, 2025

2.TRimCapS: MAKİNE ÖĞRENMESİ İLE TÜRKÇE DİLİNDEKİ GÖRÜNTÜ ALT YAZILARINI SINIFLANDIRMA SİSTEMİ
Merve Pınar, Esra Yılmaz, Zeki Çıplak, Ayşe Berna Altınel Girgin
İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, cilt 24, sayı 48, sayfalar 438-464

3. Feature efficiency in IoMT security: A comprehensive framework for threat detection with DNN and ML.
Pınar M., Aktas A., Ülkü E. E.
COMPUTERS IN BIOLOGY AND MEDICINE , cilt.186, sa.186, ss.109603, 2025

4. k-medoids Kümeleme, Boosting ve BERT algoritmalarıyla Kredi Kartlarındaki Şüpheli Davranışların Tespit Edilmesi
Merve Pınar, Ayşe Berna Altınel Girgin
Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi, cilt 16, sayı 2, sayfalar 345-356

5. Deep Learning-Assisted Segmentation and Classification of Brain Tumor Types on Magnetic Resonance and Surgical Microscope Images
Cekic E., Pinar E., PINAR M., DAĞÇINAR A.
World Neurosurgery , cilt.182, 2024

6. Yazılım projelerinde fazla mesainin proje ekibi ve projenin yönetimine etkisi
Merve Pınar, Büşra Büyüktanır, Şura Emanet, Buket Doğan
International Journal of Advances in Engineering and Pure Sciences, cilt 32, sayı 4, sayfalar 420-429

Bildiriler

1. AI-Assisted Detection of Dental Findings in Panoramic Dental Radiographs
İrem Selen, Ümmügülsüm Erol, Abdulsamet Aktaş, Ayşe Berna Altınel, Merve Pınar
2025 10th International Conference on Computer Science and Engineering (UBMK), Sayfalar 883-888, Yayıncı IEEE

2. A Comparative Analysis of Convolutional Neural Networks, Vision Transformers and Hybrid Deep Learning Architectures for Lung CT Image Classification
Sude Nur Gülcan, Elif Tilki, Karun Acar, Abdulsamet Aktaş, Ayşe Berna Altınel, Merve Pınar
2025 10th International Conference on Computer Science and Engineering (UBMK), Sayfalar 946-951, Yayıncı IEEE

Projeler

2024 – 2026
Beyin MR Görüntülerinden Derin Öğrenme Modelleri Kullanarak Hastalık Teşhisi
Yükseköğretim Kurumları Destekli Proje , BAP Doktora
Altınel Girgin A. B. (Yürütücü), Pınar M.

2024 – 2026
Elektrofizyolojik Ve Kinematik Veriler Kullanılarak Kişiye Özgü Rehabilitasyon Cihazı İçin Makine Öğrenmesi Modellerinin Oluşturulması
TÜBİTAK Projesi , 1001 - Bilimsel ve Teknolojik Araştırma Projelerini Destekleme Programı
Çürük E.(Yürütücü), Pınar M., Sarıkaş A., Pınar E., Başpınar U.

2025 – 2027
Biyomedikal Görüntülerde Hareketli Hücre Analizleri
Yükseköğretim Kurumları Destekli Proje , BAP Araştırma Projesi
Aktaş A. (Yürütücü), Pınar M., İlhan H. O., Serbes G., Uzun H.

2026-2027
BRAINSİGHT: YAPAY ZEKA DESTEKLİ İNME TESPİTİ VE BÜYÜK DİL MODELLERİ İLE OTOMATİK RADYOLOJİK RAPORLAMA OTOMASYONU

TÜBİTAK Projesi , 1512 - Girişimcilik Aşamalı Destek Programı
Aktaş A. (Yürütücü), Pınar M. (Yürütücü), Midi İ., Altınel Girgin A. B.

2025 – 2026
Anatomik Atlas Destekli İnme Segmentasyonu Ve Büyük Dil Modelleriyle Radyolojik Raporlama Sistemi
TÜBİTAK Projesi , 1002 - Hızlı Destek Programı
Aktaş A., Pınar M. (Yürütücü), Altınel Girgin A. B., Midi İ.