**T.C.**
**AKDENİZ ÜNİVERSİTESİ**

# POLIKISTIK OVARYUM SENDROMU TANISI İÇİN YAPAY ZEKA DESTEKLİ KLİNİK ARAÇ TASARIMI

**Jacklyn Günce KAYA**

**FEN BİLİMLERİ ENSTİTÜSÜ**

**ELEKTRİK-ELEKTRONİK MÜHENDİSLİĞİ ANA BİLİM DALI**

**YÜKSEK LİSANS TEZİ**

**TEMMUZ 2025**

**ANTALYA**

**T.C.**
**AKDENİZ ÜNİVERSİTESİ**

# POLIKISTIK OVARYUM SENDROMU TANISI İÇİN YAPAY ZEKA DESTEKLİ KLİNİK ARAÇ TASARIMI

**Jacklyn Günce KAYA**

**FEN BİLİMLERİ ENSTİTÜSÜ**

**ELEKTRİK-ELEKTRONİK MÜHENDİSLİĞİ ANA BİLİM DALI**

**YÜKSEK LİSANS TEZİ**

**TEMMUZ 2025**

**ANTALYA**

**T.C.**
**AKDENİZ ÜNİVERSİTESİ**
**FEN BİLİMLERİ ENSTİTÜSÜ**

# POLIKISTIK OVARYUM SENDROMU TANISI İÇİN YAPAY ZEKA DESTEKLİ KLİNİK ARAÇ TASARIMI

**Jacklyn Günce KAYA**

**FEN BİLİMLERİ ENSTİTÜSÜ**

**ELEKTRİK-ELEKTRONİK MÜHENDİSLİĞİ ANA BİLİM DALI**

**YÜKSEK LİSANS TEZİ**

**TEMMUZ 2025**

**ANTALYA**

**T.C.**
**AKDENİZ ÜNİVERSİTESİ**
**FEN BİLİMLERİ ENSTİTÜSÜ**

# POLIKISTIK OVARYUM SENDROMU TANISI İÇİN YAPAY ZEKA DESTEKLİ KLİNİK ARAÇ TASARIMI

**Jacklyn Günce KAYA**

**FEN BİLİMLERİ ENSTİTÜSÜ**

**ELEKTRİK-ELEKTRONİK MÜHENDİSLİĞİ ANA BİLİM DALI**

**YÜKSEK LİSANS TEZİ**

Bu tez 21/07/2025 tarihinde jüri tarafından Oybirliği ile kabul edilmiştir.

| | |
|---|---|
| Prof. Dr. Süleyman BİLGİN (Danışman) | [imza] |
| Prof. Dr. Övünç POLAT | [imza] |
| Dr. Öğr. Üyesi Ufuk ÖZKAYA | [imza] |

# ÖZET

## POLIKISTIK OVARYUM SENDROMU TANISI İÇİN YAPAY ZEKA DESTEKLİ KLİNİK ARAÇ TASARIMI

**Jacklyn Günce Kaya**

**Yüksek Lisans Tezi, Elektrik-Elektronik Mühendisliği Anabilim Dalı**

**Danışman: Prof. Dr. Süleyman Bilgin**

**Temmuz 2025; 79 Sayfa**

Çalışmanın amacı, kadınların üreme sağlığını etkileyen önemli bir endokrin bozukluk olan Polikistik Over Sendromu (PKOS) tanısı için yapay zeka (YZ) destekli bir klinik karar destek aracı geliştirmektir. Literatürde, PKOS tanısı genellikle klinik semptomlar, ultrason görüntüleme ve hormonal analizler kullanılarak konulmaktadır. Bununla birlikte, bu süreç uzun bir süreçtir ve esas olarak hasta ve doktor etkileşimine bağlıdır. Bu tezde, hastaların metin tabanlı ifadelerine dayalı olarak eksik bilgileri belirleyen ve dinamik soru yönlendirmeleri yapabilen bir sistem tasarlanmıştır.

BioBERT ve SciBERT, biyomedikal alana özel olarak geliştirilmiş önceden eğitilmiş iki dil modeli olarak kullanılmıştır. Geleneksel büyük dil modellerinden (BDM) farklı olarak, geliştirilen sistem, hasta verilerine dayalı olarak PKOS riskini yüzde olarak tahmin eder ve makalelerden doğrulama yaparak karara yardımcı olmaktadır. Sistem ayrıca PKOS ile benzer semptomlara sahip NCCAH, hiperprolaktinemi ve Cushing sendromu gibi diğer hastalıkların ayırıcı tanısını da sunabilmektedir. Sistem çıktılarının açıklama kalitesini belirlemek için BLEU, ROUGE ve BERTScore gibi doğal dil işleme (DDİ) metrikleri kullanılmıştır.

Sonuç olarak, çalışma, DDİ tabanlı dinamik sorgulama ve risk tahmini özellikleri ile PKOS tanısında YZ temelli etkileşimli sistemlerin potansiyelini ortaya koymakta ve literatüre yenilikçi bir katkı sunmaktadır.



**ANAHTAR KELİMELER:** BioBERT, Büyük Dil Modelleri, Doğal Dil İşleme, Polikistik Over Sendromu, SciBERT, Yapay Zeka


**JÜRİ:**

Prof. Dr. Süleyman BİLGİN

Prof. Dr. Övünç POLAT

Dr. Öğr. Üyesi Ufuk ÖZKAYA

## ABSTRACT

## ARTIFICIAL INTELLIGENCE SUPPORTED CLINICAL TOOL DESIGN FOR THE DIAGNOSIS OF POLYCYSTIC OVARY SYNDROME

**Jacklyn Günce Kaya**

**MSc Thesis in Electrical and Electronics Engineering Department**

**Supervisor: Prof. Dr. Süleyman Bilgin**

**July 2025; 79 pages**

The aim of this study is to develop an artificial intelligence (AI)-assisted clinical decision support tool for the diagnosis of Polycystic Ovary Syndrome (PCOS), a significant endocrine disorder affecting women's reproductive health.

In the literature, PCOS diagnosis is generally made using clinical symptoms, ultrasound imaging, and hormonal analyses. However, this process is time-consuming and largely depends on patient-physician interaction. In this thesis, a system has been designed that identifies missing information based on patients' text-based expressions and dynamically generates follow-up questions. BioBERT and SciBERT, two pre-trained language models specifically developed for the biomedical domain, were utilized.

Unlike traditional large language models (LLMs), the proposed system estimates PCOS risk as a percentage based on patient data and supports decision-making by validating findings through scientific literature. The system can also suggest differential diagnoses for conditions such as NCCAH, hyperprolactinemia, and Cushing's syndrome, which present with symptoms similar to PCOS. To evaluate the quality of the system's generated explanations, natural language processing (NLP) metrics such as BLEU, ROUGE, and BERTScore were used.

In conclusion, this study demonstrates the potential of AI-based interactive systems in PCOS diagnosis through NLP-driven dynamic querying and risk prediction, offering an innovative contribution to the literature.



**KEYWORDS:** Artificial Intelligence, BioBERT, Large Language Models, Natural Language Processing, Polycystic Ovary Syndrome, SciBERT


**COMMITTEE:**

Prof. Dr. Süleyman BİLGİN

Prof. Dr. Övünç POLAT

Asst. Prof. Dr. Ufuk ÖZKAYA

# ÖNSÖZ

Bu tez çalışması, yapay zekâ destekli yaklaşımların kadın sağlığı alanında çok önemli olan Polikistik Over Sendromu tanı sürecine dahil edilmesi amacıyla gerçekleştirilmiştir. Bu çalışmanın temel amacı, tıpta dijitalleşmenin ve klinik karar destek sistemlerinin gelişimiydi. Hedefim, klinik uygulamada faydalı ve bilimsel olarak sağlam bir model geliştirmekti. Bu araştırmada yer alan tüm/kısmi nümerik hesaplamalar TÜBİTAK ULAKBİM, Yüksek Başarım ve Grid Hesaplama Merkezi'nde (TRUBA kaynaklarında) gerçekleştirilmiştir.

Çalışma süresince bana akademik rehberlik, teşvik edici yönlendirmeler ve değerli katkılarıyla her zaman yol gösteren değerli danışman hocam Prof. Dr. Süleyman BİLGİN'e en içten şükranlarımı sunarım. Kadın Hastalıkları ve Doğum Uzmanı Dr. Arif Can ÖZSİPAHİ'ye klinik aşamada destekleriyle çalışmamın gerçekçi ve uygulanabilir bir zemine oturmasına yardımcı olduğu için çok teşekkür ederim.

Bu süreç boyunca yanımda olan, bana inanan ve yardımlarını hiçbir zaman esirgemeyen sevgili aileme sonsuz şükranlarımı sunuyorum. Sevgileri ve sabırları olmasaydı bu yolculuk eksik kalırdı.

Tezimin, bu alanda yapılacak yeni çalışmalara ışık tutmasını ve özellikle kadın sağlığını iyileştirmek için çözüm odaklı teknolojik yaklaşımların geliştirilmesine katkıda bulunmasını umuyorum.

# İÇİNDEKİLER

## AKADEMİK BEYAN

Yüksek Lisans Tezi olarak sunduğum “Polikistik Ovaryum Sendromu Tanısı için Yapay Zeka Destekli Klinik Araç Tasarımı” adlı bu çalışmanın, akademik kurallar ve etik değerlere uygun olarak yazıldığını belirtir, bu tez çalışmasında bana ait olmayan tüm bilgilerin kaynağını gösterdiğimi beyan ederim.

21/07/2025

Jacklyn Günce KAYA

İmzası

## SİMGELER VE KISALTMALAR

Tez çalışmasında ondalık sayı gösterimlerinde nokta (.) sembolü kullanılmıştır.

**Kısaltmalar**

ADASYN :Adaptive Synthetic Sampling Method for Imbalanced Data (Dengesiz Veriler için Uyarlanabilir Sentetik Örnekleme Yöntemi)

AMH :Anti-Müllerian Hormon

BDM :Büyük Dil Modelleri

BioBERT :Bidirectional Encoder Representations from Transformers for Biomedical Text Mining

BLEU :Bilingual Evaluation Understudy

BP :Brevity Penalty (Kısalık Cezası)

catBoost :Categorical Boosting

CBOW :Continuous Bag of Words (Sürekli Kelime Torbası)

CKDKM :Cochrane Klinik Denemeler Kayıt Merkezi

ÇKA :Çok Katmanlı Algılayıcı

CLS :Classification Token

DDİ :Doğal Dil İşleme

DN :Doğru Negatif

DP :Doğru Pozitif

DPO :Doğru Pozitif Oranı

DVM :Destek Vektör Makineleri

EAA :Eğri Altı Alan

EMBASE :Excerpta Medica DataBASE

EUOAD :En Uzun Ortak Alt Dizi

ESA :Evrişimli Sinir Ağları

FAISS :Facebook YZ Similarity Search (Facebook YZ Benzerlik Araması)

GNB :Gaussyen Naif Bayes

GloVe :Global Vectors for Word Representation

GPU :Graphics Processing Unit (Grafik işlemci birimi)

HRFLR :Hibrit Random Forest ve Lojistik Regrasyon

JSON :JavaScript Object Notation (JavaScript Nesne Gösterimi)

KNN :K-En Yakın Komşu

LR :Lojistik Regrasyon

MEDLINE :Medical Literature Analysis and Retrieval System Online (Tıbbi Literatür ve Bilgi Erişim Sistemi)

MÖ :Makine Öğrenmesi

MTEB :Massive Text Embeddings Benchmark

NB :Naif Bayes

NC-CAH :Non-Classic Congenital Adrenal Hyperplasia

PKOM :Polikistik Over Morfolojisi

PKOS :Polikistik Over Sendromu

PDF :Portable Document Forma

PRISMA :Preferred Reporting Items for Systematic Reviews and Meta-Analyses (Sistematik İncelemeler ve Meta-Analizler için Tercih Edilen Raporlama Öğeleri)

PubMED :Public/Publisher MEDLINE

ResNET :Residual Network (Kalan Ağ)

RF : Rastgele Orman

RFLR :Lojistik Regrasyon Rastgele Orman

ROC :Receiver Operating Characteristic (Alıcı çalışma özelliği eğrisi)

ROUGE :Recall-Oriented Understudy for Gisting Evaluation

SHAP :SHapley Additive exPlanations

SMOTE :Synthetic Minority Over-sampling Technique

TF-IDF :Term Frequency - Inverse Document Frequency (Terim Frekansı - Ters Belge Frekansı)

TSA :Tekrarlayan Sinir Ağı

Turing-NLG : Turing Natural Language Generation (Turing Doğal Dil Üretimi)

TÜBİTAK :Türkiye Bilimsel ve Teknolojik Araştırma Kurumu

ULAKBİLİM :Ulusal Akademik Bilgi Sistemi

WaOEL :Walrus Optimization Ensemble Learning (Walrus Optimizasyonlu Topluluk Öğrenmesi)

Word2Vec :Word to Vector

WoS :Web of Science

XGBRF :Extreme Gradient Boosting Random Forest

YN :Yanlış Negatif

YP :Yanlış Pozitif

YPH :Yüksek Performanslı Hesaplama

YPO :Yanlış Pozitif Oranı

YSA :Yaklaşık En Yakın Komşu

YZ :Yapay Zeka

# ŞEKİLLER DİZİNİ

# ÇİZELGELER DİZİNİ

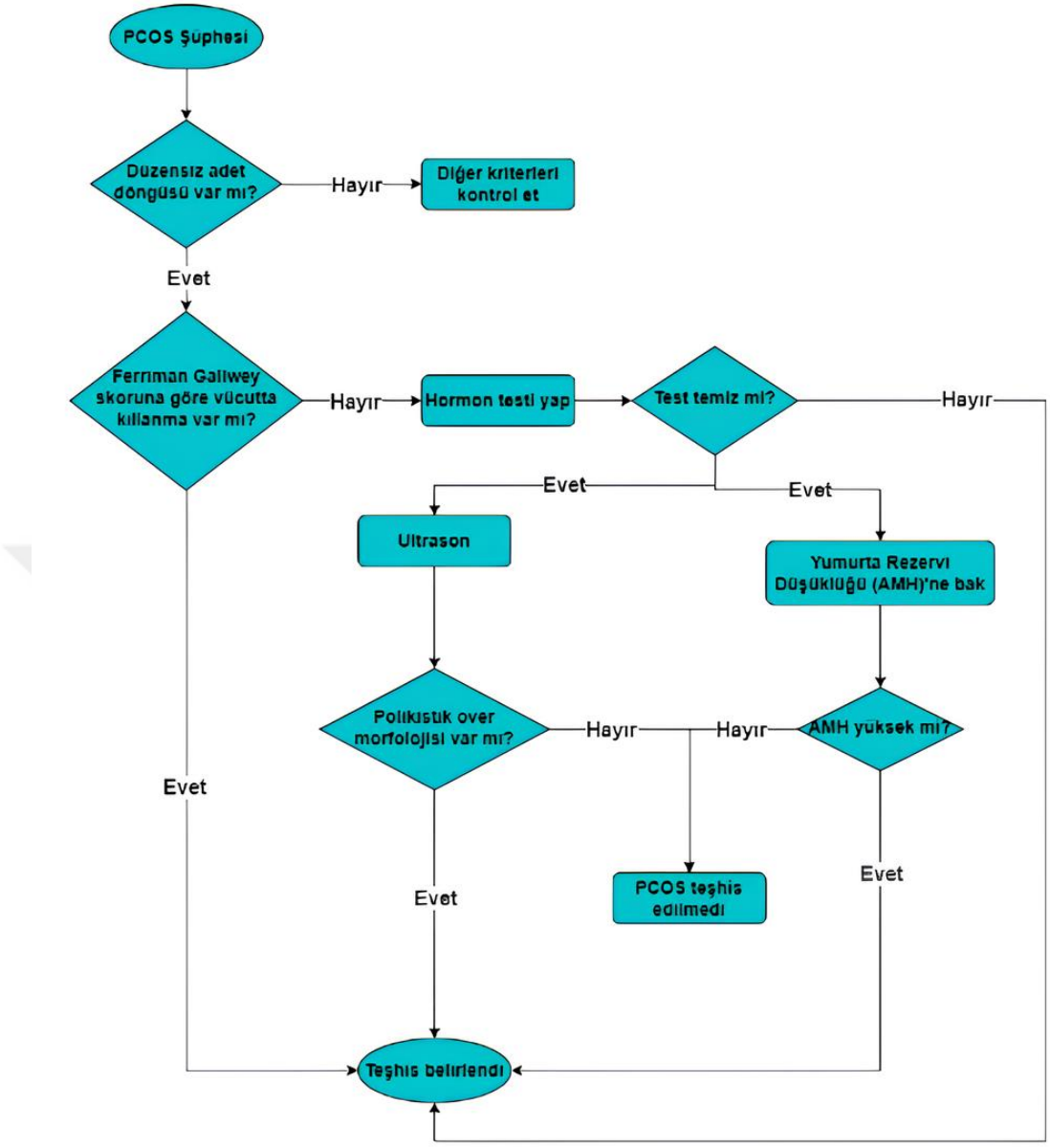
# 1. GİRİŞ

Dünya çapında milyonlarca kadın, üreme sağlığını doğrudan etkileyen karmaşık bir endokrin durum olan PKOS'tan muzdariptir. PKOS tanısı koymak için sıklıkla ultrason görüntüleme, hormon seviyesi analizi ve klinik semptomlar kullanılmaktadır. Ancak bu tanı teknikleri zahmetlidir, yoruma tabidir ve sıklıkla yeterince hassas değildir. Bu nedenle, erken ve doğru tanıyı garanti etmek için daha etkili, otomatik ve objektif karar destek sistemlerine ihtiyaç vardır.

Bu tezin amacı, PKOS teşhisi için BDM, DDİ ve YZ kullanan bir klinik karar destek sistemi oluşturmaktır. Oluşturulan sistem, eksik verileri tespit ederek dinamik bir şekilde ekstra sorular üretebilir, hastanın endişelerini metinsel olarak değerlendirebilir ve doktorların karar vermesine yardımcı olmak için riskleri öngörebilir. Bu sistem, özellikle kadın sağlığı alanında teşhis prosedürlerini desteklemenin yanı sıra, sezgisel kullanıcı ara yüzleri sunarak klinik ortamlarda faydalı olmayı amaçlamaktadır.

Üretilen sistemi düzenlemek için sağlık sektörüne uyarlanmış önceden eğitilmiş iki güçlü dil modeli olan BioBERT ve SciBERT kullanılmıştır. Sistem çıktılarını değerlendirmek için BLEU, ROUGE ve BERTScore gibi DDİ ölçütleri kullanılmıştır. Bu ölçütler sadece sınıflandırma doğruluğunu değil, aynı zamanda modelin yorumlanabilirliğini ve üretilen metinlerin kalitesini de dikkate almıştır. Bu şekilde çalışma, literatürde bulunan PKOS tabanlı YZ uygulamalarından daha kesin, anlaşılır ve modern bir yaklaşım sunmaktadır.

Şekil 1.1'de gösterilen yapı, tanı yönteminin temelini oluşturmaktadır. PKOS'un Değerlendirilmesi ve Yönetimi için dünya çapında kabul gören Uluslararası Kanıta Dayalı Kılavuz (Teede vd. 2023) bu metodoloji için model oluşturmuştur. Algoritma, PKOS tanısı konulabilmesi için aşağıdaki koşullardan en az ikisinin mevcut olması gerektiğini belirtmektedir: PKOM, hiperandrojenizm semptomları veya düzensiz adet döngüleri. AMH seviyeleri genç yaş gruplarında referans olarak kullanılırken, ultrason verileri genellikle 21 yaşın üzerindekilerde kullanılır. Bu karar verme prosedürlerini özetleyen bir akış şeması Şekil 1.1'de gösterilmektedir.




**Şekil 1.1.** PKOS teşhis diyagramı

Geleneksel hasta verilerinin ötesinde, bu tezde kullanılan veri kümeleri PubMed, UpToDate ve diğer karşılaştırılabilir literatür kaynaklarından bilgi çıkarmak için DDİ teknikleri kullanılarak oluşturulmuştur. Bu yaklaşım, modelin çeşitli endokrin bozuklukların ayırıcı tanısına ilişkin verilerin yanı sıra literatürdeki yeni materyalleri de içermesini sağlar. Oluşturulan teknik sadece bu çalışma için değil, aynı zamanda YZ'nın terapötik uygulamaları için bir temel olarak gelecekteki araştırmalar için de önem taşımaktadır.

Bu bağlamda, çalışmanın genel yapısı ve amacı ilk bölümde sunulmuştur. İkinci bölümde konuyla ilgili literatür kapsamlı bir şekilde incelenmiş, üçüncü bölümde veri setlerinin oluşturulması, modelleme prosedürü ve sistem tasarımı ayrıntılı bir şekilde detaylandırılmıştır. Deneysel sonuçlar dördüncü bölümde sunulmakta ve

değerlendirilmekte, beşinci bölümde tartışılmakta ve son bölümde genel sonuçlar ve daha ileri araştırmalar için öneriler sunulmaktadır.

## 2. KAYNAK TARAMASI

Ulusal Sağlık Enstitüleri tarafından yapılan yeni bir araştırmaya göre, YZ ve makine öğrenimi, genellikle 15 ila 45 yaş arasındaki kadınlar arasında en yaygın hormon bozukluğu olan PKOS'u etkili bir şekilde tespit edip teşhis edebilmektedir. Araştırmacılar, PKOS'u teşhis etmek ve sınıflandırmak amacıyla verileri analiz etmek için YZ/MÖ'ni kullanan yayınlanmış bilimsel çalışmaları sistematik olarak incelediler ve YZ/MÖ tabanlı programların PKOS'u başarılı bir şekilde tespit edebildiğini bulmuşlardır. Son yıllarda, birçok araştırmacı, klinik parametreler ve vital işaretlerden oluşan veri setleri ile PKOS teşhisi için YZ modelleri önermiştir.

Prathibanandhi vd. (2024) yaptıkları çalışmada, derin öğrenme modellerinin teşhis süreçlerine entegrasyonu sayesinde umut vadetmektedir. Özellikle, teşhis doğruluğunu artırmak için ön işleme ve segmentasyon tekniklerini kullanan ve ardından ESA ile sınıflandırma yapan kapsamlı bir yaklaşım önerilmiştir. Bu metodoloji, analiz için girdileri standartlaştırmak adına çevirme, döndürme ve yeniden ölçekleme gibi görüntü geliştirme tekniklerini içermektedir. Ayrıca, ilgi alanlarını izole etmek için boyuta dayalı filtreleme ve kontur algılama gibi segmentasyon teknikleri kullanılarak daha doğru özellik çıkarımı sağlanmaktadır. Yapılan çalışmalar, ESA modellerinin çoklu performans ölçümlerinde ResNet ve GNB gibi geleneksel modellerden daha iyi performans gösterdiğini ve yaklaşık %98 doğruluk oranına ulaştığını göstermektedir. Bu bulgular, YZ destekli araçların, PKOS'un erken tanı ve tedavisinde klinisyenlere önemli ölçüde yardımcı olabileceğini ve sonuç olarak hasta sağlığı üzerindeki olumlu etkilerini artırabileceğini göstermektedir.

PKOS teşhisi ile ilgili olarak, Abu Adla vd. (2021) yaptıkları çalışmada hastalığın çoğunlukla teşhis edilemeyen bir hastalık olduğunu belirtmişlerdir, muhtemelen çok faktörlü bir tezahürden kaynaklanabilir. Makine öğreniminin varlığı, PKOS durumunun teşhisi için veri kullanımı yoluyla nicel bir temel sağlar. Transvajinal ultrasonun kullanımı, yumurtalıkların kendisinin görüntülenmesi ve ardından makine öğrenimine dayalı tahmin makinelerinin eğitimi için verilerin kullanılması için görüntü işleme modellerine doğru tarama ve görüntü edinme olanakları sağlamıştır. Bu teşhis yöntemi, prosedürü gerçekleştirmek için tanı ekipmanının ve uzman personelinin bulunabilirliğine dayanır. Son zamanlarda, hastalardan elde edilen biyokimyasal ölçümlerden yaşam tarzı faktörlerine kadar uzanan bir dizi faktörün kullanımı, PKOS hastalığının tahmin ve teşhisi için yapılmıştır. NB, KNN, YSA, LR ve DVM gibi modeller, hastalığın modellenmesine çeşitli kapasitelerde araştırılmıştır (Nsugbe 2023).

DAISy-PKOS Phenome Study (Melson 2023) ve "Novel Subtypes of Polycystic Ovary Syndrome" çalışması (Zi-Jiang 2023), klinik kullanım için metabolik risk tahmin modelleri geliştirmek üzere multi-omik verileri ve centroid tabanlı kümeleme, bağlantı ve uyarlanabilir benzerlik ölçümleri gibi gelişmiş kümeleme tekniklerini entegre etmeye yönelik çok merkezli çabaları gösteren gözlemsel araştırma girişimlerine iki örnektir. YZ destekli tanı araçlarının nihai doğrulaması ve standart klinik uygulamaya dahil edilmesi bu araştırmalar sayesinde mümkün olmuştur (Zi-Jiang 2023; Melson 2023).

PKOS hastalığının erken tespiti ve tanısı için Bhat (2021) tarafından yeni bir teknik önerildi. Önerilen model XGBRF ve catBoost'a dayanıyordu. Verilerin ön işlenmesinden sonra, univariate özellik seçim yöntemiyle en üst 10 özellik seçildi.

Doğruluk sonuçlarını karşılaştırmak için uygulanan sınıflandırıcılar ÇKA, karar ağacı, DVM, HRFLR, RF, LR ve gradient boosting idi. Sonuçlar, XGBRF'nin %89 doğruluk skoru sergilediğini, catBoost'un ise %95 doğruluk skoru ile daha iyi performans gösterdiğini gösterdi. Diğer sınıflandırıcıların doğruluk skorları %76 ile %85 arasında yer aldı. catBoost tekniği, PKOS hastalığının erken tespiti için en iyi modeldi (Bhat, 2021).

Bharati vd. (2020), Kaggle PKOS veri setini kullanarak filtre tabanlı univariate özellik seçimi yöntemini uyguladı. Araştırmacılar, gradient boosting, RF, LR ve HRFLR modellerini kullanarak, seçilen on özellikle PKOS'u sınıflandırdı. Sonuçlar, Hybrid RFLR modelinin %91,01 doğruluk ve %90 hatırlama değeri ile diğer sınıflandırıcılara göre üstün olduğunu gösterdi (Bharati vd. 2020). Hassan ve Mirza (2020) çalışmasında, Kaggle PKOS veri setini sınıflandırmak için RF, DVM, karar ağaçları, NB ve LR gibi farklı makine öğrenimi yöntemlerini kullandı. Deneysel sonuçlar, RF modelinin %96 sınıflandırma doğruluğu ile diğer yöntemlere üstün olduğunu ortaya koydular. Neto vd. (2021), PKOS Kaggle veri setinin sınıflandırılmasında farklı sınıflandırıcıların performansını karşılaştı. RF, diğer sınıflandırıcılara göre %95 doğruluk oranı ve %96 hassasiyet oranı elde ederek öne çıktı (Hassan ve Mirza 2020).

Zhang vd. (2021), Çin Tıp Üniversitesi'ndeki Shengjing Hastanesi'nden Raman spektrum verilerini içeren iki veri kümesi tanımladılar. Yazarlar XGBoost, KNN ve RF modellerini birleştiren bir istifleme modeli kullanmışlardır. İlk katman KNN, RF ve XGBoost'u birleştirirken, ikinci katmanda yalnızca XGBoost kullanılmıştır. İki veri kümesi incelenmiştir: foliküler sıvı ve plazma örnekleri. Yığınlama modeli, foliküler sıvı verilerini kullanarak %89,32 doğruluk elde ederek diğer tekniklerden daha iyi performans göstermiştir.

YZ teknolojisinin sağlık sektöründeki potansiyeli, Florida Üniversitesi ve NVIDIA tarafından geliştirilen yeni bir YZ yazılımı ile bir kez daha gözler önüne serilmiştir. İki kurumun ortak çalışması sonucunda ortaya çıkan bu YZ yazılımı, doktorların hastalarının tıbbi notlarını yazmalarını taklit ederek, yazılan notları gerçek doktor notlarından ayırt edilemez kılmaktadır. Florida Üniversitesi Tıp Fakültesi'nde görev yapan araştırmacılar, 2 milyon hastanın özel bilgilerini içeren tıbbi kayıtları kullanarak bu YZ modelini eğitmiş ve geniş bir dil modeli olan GatorTronGPT'yi oluşturmuşlardır. Bu model, tıp doktorlarının notalarını taklit ederek klinik metinler üretebilmektedir. Yapılan literatür çalışması, bu gelişmenin sağlık sektöründe kayda değer bir etki yaratabileceğini öne sürerek, YZ destekli notlama ve belgeleme süreçlerindeki potansiyeli vurgulamaktadır (Cumhuriyet 2023).

Son zamanlarda sağlık sektöründe YZ uygulamalarının artan önemi, birçok şirketin ve araştırma kurumunun bu alana yönelik çözümler geliştirmesine yol açmıştır. Bu bağlamda, Ada Health, kullanıcıların semptomlarını değerlendirmesi ve kişiselleştirilmiş sağlık önerileri sunması amacıyla YZ tabanlı bir tıbbi sohbet robotu olarak öne çıkmaktadır. Aynı zamanda, Infermedica ve Isabel gibi uygulamalar, doktorlara hızlı ve doğru ayırıcı tanı koymalarında yardımcı olmak üzere tasarlanmıştır. Bu uygulamalar, semptomları analiz ederek önceki verilere dayalı öneriler oluşturmak için gelişmiş algoritmalar kullanmaktadır. Özellikle, Butterfly IQ gibi uygulamalar, sağlık profesyonellerine taşınabilir bir ultrason cihazı aracılığıyla gerçek zamanlı ultrason tarama imkanı sunarak, geleneksel ultrason ekipmanına ulaşımın sınırlı olduğu durumlarda kullanışlı bir çözüm sunmaktadır. Ayrıca, NVIDIA Clara gibi teknolojik

gelişmeler, tıbbi görüntüleme ve analiz alanlarında daha karmaşık uygulamalara odaklanarak sağlık sektöründe önemli bir rol oynamaktadır. Bu uygulamaların, tıp profesyonellerinin işini kolaylaştırmak, hastalara daha iyi hizmet sunmak ve sağlık sektöründe inovasyonu teşvik etmek adına önemli bir potansiyele sahip olduğu görülmektedir. Literatürdeki bu YZ tabanlı tıbbi araçlar, günümüzdeki sağlık hizmetleri konseptini temelden değiştirecek önemli adımlar olarak değerlendirilmektedir (DoktorTakvimi 2023).

Barrera vd. (2023) yaptıkları çalışmada, YZ ve MÖ'nin PKOS teşhisindeki potansiyelini değerlendiren bir sistematik inceleme sundular. Çalışmalar, YZ/MÖ yöntemlerinin, PKOS'u tanıma, sınıflandırma, stratifikasyon yapma ve tahmin etme konularındaki etkinliğini araştırmaktadır. PRISMA yönergelerini takip etmişler, 1 Ocak 2022'ye kadar yayımlanan tüm ilgili çalışmalar gözden geçirilmiş, YZ/MÖ tekniklerinin PKOS'u teşhis etme, sınıflandırma, stratifikasyon yapma veya tahmin etme kullanımları değerlendirmişlerdir. Bunu başarmak için MEDLINE, EMBASE, CKDKM, WoS ve IEEE Xplore dahil olmak üzere birden fazla veri tabanında kapsamlı bir literatür taraması gerçekleştirerek 1 Ocak 2022'ye kadar yayınlanmış ilgili çalışmaları tespit etmişlerdir. Yazarlar incelemelerinde, PKOS'un teşhisinde, sınıflandırılmasında veya öngörülmesinde YZ/MÖ kullanımını değerlendiren İngilizce dilinde, hakemli orijinal çalışmalara odaklanarak belirli dahil etme kriterleri oluşturmuşlardır. Çalışmalar, PKOS için standart tanı kriterleri kullanıp kullanmadıklarına veya sınıflandırma amacıyla bu kriterleri yalnızca kısmen kullanıp kullanmadıklarına göre kategorize edilmiştir.

Çeşitli model kombinasyonlarıyla doğruluk oranlarını iyileştirmek için, çağdaş yaklaşımlar topluluk ve istifleme yöntemlerini tercih etmektedir. Panjwani ve arkadaşları, PKOS semptom verilerinden elde edilen 12 özellikli bir veri kümesinde, yedi temel sınıflandırıcıdan oluşan bir toplulukla meta sınıflandırıcı olarak derin öğrenmeyi kullanan bir model geliştirmiştir; hiperparametre ayarlı modeli mors optimizasyonu ile WaOEL olarak adlandırmışlar ve %92.8 doğruluk ve %93 EAA elde etmişlerdir (Panjwani vd. 2025). Emara vd. (2025), Kaggle'daki 541 hastalık verisi üzerinde ADASYN/SMOTE dengesi ve BORUTA özellik seçimini kullanarak bir istifleme öğrenme mimarisi geliştirmiş ve %97'ye varan bir doğruluk oranı belirtmişlerdir. Çeşitli öğrenme tekniklerini birleştiren bu ve ilgili hibrit yaklaşımlar, PKOS sınıflandırmasında son derece iyi performans göstermiştir.

Güncel araştırmalar sadece model doğruluğuna değil, aynı zamanda bilimsel keşifleri hızlandırabilecek YZ altyapılarına da odaklanıyor. FutureHouse gibi yeni firmalar, araştırmacılar için bilimsel materyalleri tarayan “Crow” ve “Falcon” gibi özel YZ ajanlarıyla tamamlanan platformlar sağlıyor (FutureHouse 2025). Örneğin, bir tanıtım videosunda FutureHouse teknolojisi, PKOS için potansiyel yeni ilaç hedeflerini keşfetmek üzere binlerce yayını saniyeler içinde taradı. Bu tür gelişmeler, PKOS araştırmalarında YZ kullanımını standart klinik verilerin ötesine genişletmeye yönelik adımlar olarak görülebilir (InteligenAI 2025).

Sonuç olarak Barrera vd. (2023), PKOS'un tanı ve yönetimini geliştirmede YZ/MÖ teknolojilerinin önemli potansiyelini vurgulamıştır. Bununla birlikte, mevcut metodolojik boşlukları gidermek, bulguların sağlamlığını artırmak ve alandaki paydaşlar arasında işbirliğini teşvik etmek için daha fazla araştırma yapılması çağrısında bulundular. Bu sistematik derleme, PKOS tanısında YZ/MÖ uygulamalarının mevcut

durumunu anlamak için temel bir kaynak görevi görmekte ve bu alandaki araştırmaların ilerletilmesinin önemini vurgulamaktadır.

Bu çalışma, PKOS tanı sistemlerini yaygın olarak kullanılan sabit yapılandırılmış anketlerin, görüntü işleme tabanlı modellerin veya yapılandırılmış klinik verilerin ötesine taşıyarak mevcut literatürdeki önemli bir boşluğu kapatmaktadır. Dinamik karar verme, hasta şikayetlerinin doğal dilde işlenmesi ve eksik veri yönetimi gibi niteliksel özellikler literatürdeki modellerin çoğunda yeterince yer almamaktadır. Ancak bu çalışma, eksik verilere dayalı soruları otomatik olarak oluşturabilen, hasta şikayetlerini kendi kelimeleriyle analiz edebilen ve BioBERT ve SciBERT gibi biyomedikal dil modellerini kullanarak risk tahmini yapabilen entegre bir karar destek sistemi sunarak teknik ve klinik açıdan daha eksiksiz bir çözüm geliştirmektedir. Ayrıca, sadece sınıflandırma değil, aynı zamanda doğal dilde içgörülü açıklamalar üreterek, sağlık uygulayıcılarına açık ve anlaşılır çıktılar sunma açısından literatürdeki diğer araştırmalardan ayrılmaktadır. Bu şekilde çalışma, PKOS teşhisi için YZ uygulamalarının yeteneklerini geliştirmekte ve klinik karar verme sürecinin dijitalleştirilmesi için son teknoloji bir model sunmaktadır.

## 2.1. Yapay Zeka

İnsan zekasının çeşitli yönlerini makineler, özellikle bilgisayar sistemleri tarafından taklit etme ve uygulama yeteneği, YZ olarak bilinir. YZ, derin öğrenme, konuşma tanıma, makine görüşü, uzman sistemler ve DDİ gibi uygulamaları içerir. YZ sistemleri genellikle çok sayıda etiketli veri ile eğitilir. Daha sonra, bu verilerdeki kalıpları ve korelasyonları analiz ederek yeni veriler üzerinde tahminler yapar (Russell ve Norvig 2021). Örneğin, bir görüntü tanıma sistemi, fotoğraflardaki nesneleri veya desenleri milyonlarca etiketli görsel üzerinden öğrenerek tanıyabilir. Benzer şekilde, bir chatbot, büyük metin verilerinden öğrenerek insanlarla gerçek zamanlı, doğal bir şekilde konuşabilir (Goodfellow vd. 2016).

YZ, özellikle teşhis ve tedavi süreçlerinde devrim yaratabilir. YZ destekli klinik araçlar, PKOS gibi karmaşık ve çok faktörlü hastalıkların tanısında kritik olabilir. YZ destekli klinik araçlar, PKOS tanısı için şu avantajları sağlar: İlk olarak, derin öğrenme algoritmaları, insan gözünün fark edemeyeceği ince desenleri büyük veri kümelerinden öğrenebilir. Örneğin, ESA, ultrason görüntülerinde folikül sayısını ve boyutlarını analiz ederek PKOS tanısını destekleyebilir (Goodfellow vd. 2016). İkincisi, klinisyenler, YZ sistemlerinin gerçek zamanlı olarak işlediği hasta verilerini kullanarak karar vermelerine yardımcı olur. Bu, zaman kısıtlaması olan klinik ortamlarda özellikle faydalıdır. Üçüncüsü, YZ, hastanın geçmişini, genetik bilgilerini ve yaşam tarzı değişkenlerini kullanarak kişiselleştirilmiş teşhis ve tedavi önerileri sağlayabilir (Topol 2019). Bununla birlikte, YZ sağlıkta kullanılırken veri gizliliği, model açıklanabilirliği ve etik sorunlar gibi sorunlar da vardır. Örneğin, YZ modellerinin "kara kutu" şekli, klinisyenlerin sonuçlara güvenmesini zorlaştırabilir (Holzinger vd. 2019).

YZ destekli bir klinik araç tasarımı, PKOS tanısı için birden fazla veri kaynağını (ultrason görüntüleri, laboratuvar bulguları ve klinik semptomlar) entegre eden bir makine öğrenimi modeli gerektirir. Bu tür bir sistem, derin öğrenme tekniklerinin yanı sıra DVM veya rastgele ormanlar gibi geleneksel algoritmaları da kullanabilir.

Son olarak, YZ destekli araçların klinik doğruluğunu ve güvenilirliğini garanti etmek için kapsamlı validasyon süreçleri ve çok merkezli çalışmalar gereklidir (Azziz vd. 2016). Bu nedenle, YZ, PKOS gibi karmaşık hastalıkların tanısında klinik süreçleri değiştirebilir. YZ destekli klinik araçlar, doğru, hızlı ve kişiselleştirilmiş teşhisler sunarak sağlık hizmetlerinin etkinliğini artırabilir. Bununla birlikte, bu teknolojilerin yaygın olarak kullanılması için düzenleyici, ahlaki ve teknik zorlukların ele alınması çok önemlidir.

## 2.2. Doğal Dil İşleme

Bilgisayar bilimleri ve YZ alanlarında önemli bir kavram olan DDİ, bilgisayarların insan dilini anlamasını ve işlemesini sağlayan disiplinleri kapsar. Bu alan, bilgisayarların dilin karmaşıklığını ve çok katmanlı yapısını anlamak, yorumlamak ve üretmek için kullanılmasını içerir. Bilgisayarların dil işleme yeteneklerini geliştirmek amacıyla atılan ilk adımlar, dil işleme teknolojisinin kökenlerini oluşturmaktadır. Bilgisayarlı dil işleme alanında çeşitli gelişmeler olmuştur. Bu alan, sentaks, semantik, pragmatik, morfoloji, kelime dağarcığı ve morfoloji gibi temel kavramlar üzerine inşa edilmiştir. Sözcükleri ayırma, kök bulma, anlamlı hale getirme, özel isimleri tanıma ve duygu analizi gibi işlemler DDİ'nin temel tekniklerindendir. DDİ; metin madenciliği, dil modelleme, konuşma tanıma, duygu analizi ve çeviri sistemleri dahil olmak üzere çok sayıda uygulamada etkili bir şekilde kullanılmaktadır.

DDİ, günlük yaşamın ayrılmaz bir parçası olup, perakende ve tıp gibi çeşitli sektörlerde önemli bir rol oynamaktadır. Özellikle, Amazon'un Alexa'sı ve Apple'ın Siri gibi konuşma araçları, kullanıcı sorgularını anlamak ve etkili yanıtlar sağlamak için DDİ teknolojisini kullanmaktadır. Gelişmiş araçlar, örneğin ticari uygulamalar için kullanılan GPT-3, çok çeşitli konularda karmaşık düzyazılar üretebileceği gibi tutarlı konuşmalar gerçekleştirebilen güçlü sohbet robotları da ortaya çıkarabilir. Google, arama motoru sonuçlarını iyileştirmek amacıyla DDİ'yi kullanırken, sosyal medya platformları da nefret söylemini tespit etmek ve filtrelemek için bu teknolojiyi benimsemektedir.

### 2.2.1. Doğal dil işleme teknikleri

Metin tabanlı verileri anlamlandırmak, sınıflandırmak ve bilgi üretmek gibi görevleri yerine getirmek için DDİ teknikleri kullanılır. Genel olarak, derin öğrenme tabanlı yaklaşımlar ve geleneksel MÖ yöntemleri bu teknikleri ele alabilecek iki ana kategoridir (Young vd. 2018).

#### Geleneksel makine öğrenimi DDİ teknikleri

LR, NB, karar ağaçları ve DVM, geleneksel denetimli sınıflandırma algoritmalarıdır. Bu teknikler, TF-IDF veya kelime sıklığı gibi yüzeysel özelliklere dayanarak metinleri sınıflandırır ve temsil eder (Sebastiani 2002). Konu modellemede sıklıkla kullanılan Latent Dirichlet Allocation yöntemi, belgeleri konu-kelime dağılımları üzerinden analiz eden istatistiksel bir yöntemdir (Blei vd. 2003). Rabiner'e (2002) göre Gizli Markov Modelleri, konuşma bölütleme gibi ardışık dil görevlerinde kullanılan bir başka geleneksel yöntemdir.

Özellikle anlamsal bağlantıların, tıbbi terminolojinin ve hasta tanımlarındaki bağlamın çok önemli olduğu sağlık tabanlı uygulamalarda bu kısıtlama kayda değer bir performans kaybına neden olmaktadır. Ayrıca, standart modeller verileri doğrudan sınıflandırabilse de metin üretimi veya açıklama gibi görevler için yetersiz kalmaktadır.

Diğer yandan, bu çalışmada kullanılan BioBERT ve SciBERT gibi önceden eğitilmiş derin öğrenme tabanlı dil modelleri, sınıflandırma ve doğal dil üretimi görevlerinde daha iyidir ve biyomedikal metinlerdeki anlamsal bağlantıları daha iyi anlayabilmektedir. Dolayısıyla bu çalışma, mevcut DDİ yaklaşımlarının sınırlı kabiliyetinin üstesinden gelerek daha kapsamlı, bağlamsal ve klinik açıdan faydalı sonuçlar verebilecek bir strateji ile literatüre katkıda bulunmaktadır.

**Derin öğrenme DDİ teknikleri**

Derin öğrenme tabanlı DDİ teknikleri, metinlerin bağlamsal ve karmaşık yapılarını daha etkili bir şekilde analiz etmeyi mümkün kılmaktadır. ESA (Gabrilovich ve Markovitch 2007), sınıflandırma görevlerinde metinleri matris temsilleri kullanılarak değerlendirilmektedir. TSA gibi yaklaşımlar, n-gram veya pencere tabanlı yaklaşımlar kullanarak metni işler, ancak bağlamsal bütünlüğü yakalamakta yetersiz kalabilmektedir (Goldberg 2017). Seq2Seq mimarisi çeviri ve özetleme gibi görevlerde kullanılırken, otomatik kodlayıcılar metinleri daha düşük boyutlu gösterimlere dönüştürerek özellik öğrenimi ve boyut indirgeme için kullanılmaktadır (Hinton ve Salakhutdinov 2006). Son olarak, self-attention mekanizması ve Transformer mimarisi (Vaswani vd. 2017) uzun bağlamları yakalamakta ve performansı önemli ölçüde geliştirmektedir.

Çalışmada, bağlamsal bilgilerle geliştirilmiş ve klinik metinlerin anlamsal temsilini doğrudan öğrenebilen BioBERT ve SciBERT adlı önceden hazırlanmış dil modelleri kullanıldı. Kendi kendine dikkat mekanizmasına sahip bu dönüştürücü tabanlı modeller, uzun bağlamları modelleme yeteneğine sahiptir. Tıbbi terminolojiye uyarlanmış ön eğitim prosedürleri sayesinde, eğitim süreci daha optimize edilmiş ve hedef sınıflara daha önce tartışılan tasarımlara kıyasla daha yüksek hassasiyetle ulaşılmıştır. Sonuç olarak, çalışmada geleneksel derin öğrenme teknikleri yerine alan odaklı BERT türevleri seçilmiştir.

### 2.2.2. Gömme modelleri

MÖ'de gömme modelleri, karmaşık verileri daha küçük boyutlu modellere dönüştürerek işlemeyi kolaylaştırmıştır (Nayab Hassan 2024). Özellikle DDİ, bilgisayarlı görme ve öneri sistemleri gibi alanlarda önemli bir rol oynadılar. Gömme, yüksek boyutlu verilerin daha düşük boyutlu vektörlere yoğun (dense) dönüştürülmesidir. Bu değişiklik, verinin anlam ilişkilerini koruyarak daha etkili bir analize izin verir. MÖ sistemleri, gömme modelleri kullanarak daha iyi sonuçlar elde edebilir. Günümüzde, BERT, GPT ve Sentence-BERT gibi daha gelişmiş modeller bu modelleri geliştirmiştir.

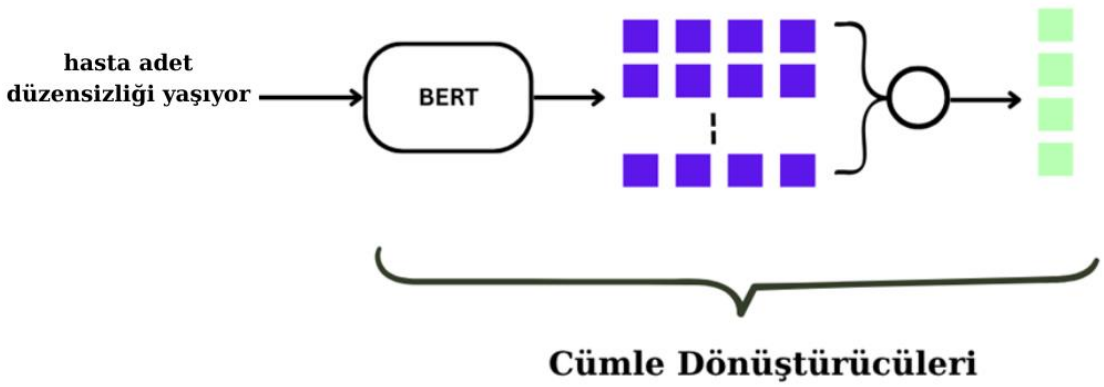
**Cümle dönüştürücüleri tekniği**

Cümle Dönüştürücüleri, en yeni metin ve görsel gömme modellerine erişmek, bunları kullanmak ve öğretmek için geliştirilmiş bir Python modülüdür (SBERT.net 2025). Bu modül, Cümle Dönüştürücü modelleri ile metin gömme işlemlerini yürütebilir

veya Cross-Encoder modelleri ile benzerlik skorlarını hesaplayabilir. Bu, semantik arama, metinsel benzerlik hesaplama ve anlamdaş cümle madenciliği gibi DDİ'deki çok sayıda uygulamayı mümkün kılmaktadır.

Hugging Face platformunda, MTEB liderlik tablosunda yer alan en güncel modeller de dahil olmak üzere doğrudan kullanılabilecek 5.000'den fazla Cümle Dönüştürücüleri modeli bulunmaktadır. Cümle Dönüştürücüleri modülü ayrıca kullanıcıların kendi kullanım alanlarına uygun modelleri kolayca eğitmesine ve ayarlarına izin vermektedir.

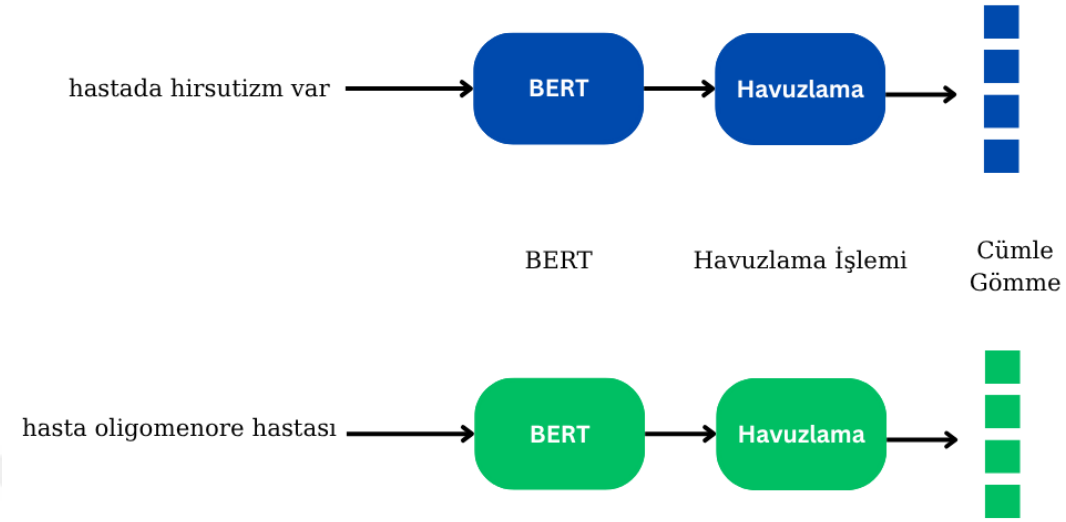
Tipik olarak, bir cümle dönüştürücü mimarisi, giriş cümlesini BERT tabanlı bir modele besleyerek her kelime için vektörler üretir. Bununla birlikte, BERT doğrudan cümle vektörleri üretmediği için bu kelime vektörlerinin ortalaması alınır (mean pool). Sonuç olarak, sabit boyutta bir cümle temsili üretilmektedir (SBERT.net 2025). Aşağıdaki Şekil 2.1 ve 2.2'de bu temel yapı gösterilmektedir. Anlamsal karşılaştırma faaliyetleri sıklıkla bu Cümle Dönüştürücüsü yapısını kullanmaktadır çünkü en basit ve verimli olanıdır.




**Şekil 2.1.** Cümle dönüştürücü mimarisiyle cümle vektörü üretimi

PKOS ile ilgili klinik belgelerin işlenmesi için dönüşüm tabanlı bir yaklaşım Şekil 2.1'de gösterilmektedir. Örneğin, bir klinisyen sisteme bir tanı gönderdiğinde, "hasta adet düzensizliği yaşıyor" gibi ifadeler önce her biri bağlamsal bir vektörle temsil edilen kelime parçalarına ayrılır. Şekilde bu vektörler mor renkli kare bloklar olarak gösterilmiştir. Her bir mor kare, bir kelimenin anlamını ve bağlamını temsil eden yüksek boyutlu bir vektördür. BERT tarafından oluşturulan bu çok sayıda kelime parçacığı vektörünün ardından, her şeyin tek bir sabit boyutlu vektörde özetlenmesi gerekmektedir. Görselde mor karelerin ardından gelen daire ve parantezli grup, "havuzlama" katmanı tarafından gerçekleştirilir. Ortalama alma, maksimum alma ya da BERT'in [CLS] kelime parçacığı tipik olarak pooling işlemini gerçekleştirmektedir. Bu katman, her cümlenin anlamını gösteren tek bir vektör sağlamaktadır. Görselin en sağında yer alan yeşil kareler, bu çıktıyı temsil etmektedir. Bu yapı, "Cümle Dönüştürücüleri" olarak adlandırılmaktadır. Cümle dönüştürücü sistemler, benzerlik ölçümü, sınıflandırma, bilgi erişimi ve soru-cevap sistemleri gibi DDİ ile ilgili çok sayıda görevde kullanılabilmektedir. Bu işlemler, doğal dildeki cümleleri anlamlı ve sabit boyutlu sayısal temsillere çevirerek gerçekleştirilmektedir. Bu tür sistemlerin temel amacı, makinelerin

dili daha iyi anlayabilmesini sağlamak için cümlelerin anlamsal yapısını vektör uzayında göstermektir.




**Şekil 2.2.** İki farklı cümle için havuzlama işlemi sonrası elde edilen gömmeler

BERT modeli, "hastada hirsutizm var" ve "hasta oligomenore hastası" cümlelerinden kelime vektörleri oluşturmak için kullanılmaktadır. Bu kelime vektörleri daha sonra Şekil 2.2'de görüldüğü gibi tek bir cümle gömme oluşturmak için birleştirilmektedir. Böylece, ifadeler anlamsal olarak bağlantılı vektörler oluşturulmaktadır. Bu işlemi daha soyut düzeyde betimleyen Şekil 2.1, cümle dönüştürücü yapının genel mimarisini sunmakta ve BERT çıktılarının nasıl tek bir vektöre indirgenerek cümle gömmesine dönüştüğünü göstermektedir. Her bir cümle için ayrı bir BERT ve havuzlama işlemi sonucunda oluşturulan gömüler, iki cümleyi karşılaştıran ve sürecin bağımsızlığını ve tekrarlanabilirliğini vurgulayan Şekil 2.2'de yan yana gösterilmiştir. Bu şekilde her iki tür birleştirildiğinde teorik ve pratik bütünlük sağlanmış olur.

Bu yapı sayesinde BDM daha hızlı ve etkili bir şekilde kullanılabilmektedir. Cümle dönüştürücüler, anlamsal benzerlik görevlerine ek olarak sağlık sektöründe bilgi çıkarma, tavsiye sistemleri ve klinik karar destek sistemlerinde de faydalıdır (Sleightholm 2025).

### GloVe tekniği

GloVe yöntemi, kelimeleri vektör temsilleri olarak canlandırmaktadır, Pennington vd. (2014) tarafından bu model ilk kez sunulmuştur. Model, kelimelerin geçme istatistiklerini kullanarak kelimeler arasındaki anlam ilişkilerini belirler. GloVe, kelime-kelime birlikte geçiş matrisindeki yalnızca sıfır olmayan bileşenleri kullanarak global istatistikleri verimli bir şekilde değerlendirir. Bu, önceki yerel pencere tekniklerinden (örneğin, skip-gram) farklı bir yaklaşımdır. GloVe, kelime analoji görevinde %75 doğruluk sağlarken adlandırılmış varlık tanıma benchmark'larında da üstün performans gösterdi. Modelin kaynak kodu ve eğitilmiş kelime vektörleri de Stanford DDİ projesi

aracılığıyla erişilebilir. GloVe, dil işleme, soru yanıtlama, belge sınıflandırma ve bilgi getirme gibi çok çeşitli uygulamalarda kullanılabilir.

**TF-IDF tekniği**

Bir belgedeki kelimelerin önemini belirlemek için TF-IDF yöntemi kullanılır. Bu yöntem, bir kelimenin bir belgedeki frekansını ve tüm belge kümesinde bu kelimenin kaç belgede bulunduğunu dikkate alarak kelimenin önemini belirler. TF-IDF'nin amacı, sıklıkla kullanılan ancak yaygın olmayan kelimeleri vurgulamaktır (Çelik vd. 2021). Bu değer, Denklem 2.1'de gösterildiği gibi hesaplanmaktadır.

$$w_{i,j} = tf_{i,j} \times \log \frac{N}{df_i} \quad (2.1)$$

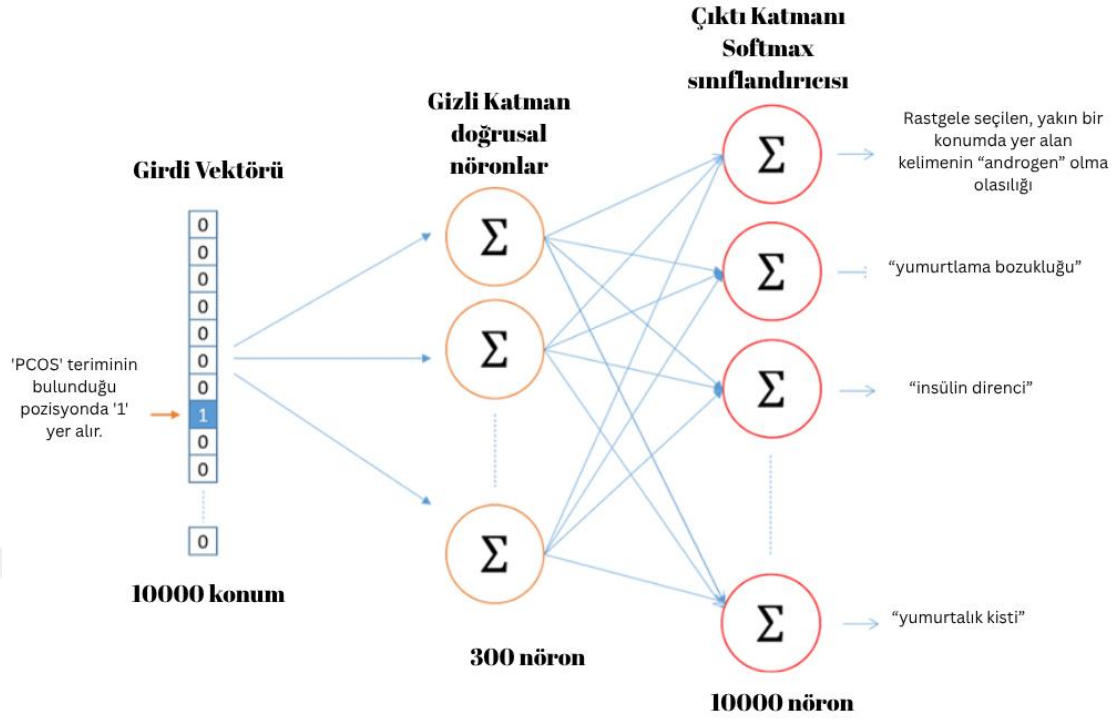
Burada, $tf_{i,j}$ ifadesi, $i$ kelimesinin $j$ numaralı belgede geçme sıklığını belirtir. $df_i$ kelimesinin yer aldığı belge sayısını ifade ederken, $N$ ise toplam belge sayısını temsil eder. Sezgisel olarak, belirli bir kelimenin belirli bir belgeyle ne kadar ilgili olduğunu ortaya koyabilir. Sadece birkaç belgede geçen kelimeler genellikle daha yüksek TF-IDF değerlerine sahiptir.

**Word2Vec tekniği**

Word2Vec, kelimeleri vektör temsillerine dönüştürerek bağlam ilişkilerini ve anlamsal benzerlikleri öğrenen bir yöntemdir. Bu yöntem, bilgisayarların büyük metin koleksiyonları üzerinden kelimelerin anlamlarını ve kullanım bağlamlarını öğrenmesini sağlar. Word2Vec, yüksek boyutlu bir uzayda her kelimeyi bir vektör olarak gösterir ve bu vektörlerin konumlarının kelimelerin anlamlarını yansıtmasını sağlar (Logunova vd. 2023).

Word2Vec algoritması, sığ yapay sinir ağları kullanır ve büyük bir metin korpusundan kelimelerin anlamlarını öğrenir. Sığ sinir ağları, derin sinir ağlarından farklı olarak yalnızca bir veya iki gizli katman içerir. Word2Vec, bu durumda hem modelin eğitilmesini daha şeffaf hale getirir hem de hesaplama sürecini hızlandırır. Model, hızlı bir şekilde kelimeler arasındaki anlamsal benzerlikleri tanıyabilir ve eş anlamlı kelimeleri bulabilir. Bir girdi olarak büyük bir metin kümesini kullanır ve her kelimeye yüksek boyutlu bir vektör atayarak bir vektör uzayı oluşturur. Word2Vec, kelimeler arasındaki ilişkileri analiz etmek ve manipüle etmek için öğrenilen vektörleri kullanarak kullanılabilir.

Öğrenilen vektörler, kelime ilişkilerini incelemek ve değiştirmek için Word2Vec ile birlikte kullanılabilir.




**Şekil 2.3.** Word2Vec – Skip-gram modelinin temsili yapısı

Şekil 2.3 temsili bir Skip-Gram model yapısını göstermektedir. Bu yapıda, sisteme girdi olarak tek bir merkez kelime verilir ve sistem bu kelimenin bağlamını kullanarak çevresindeki kelimeleri tahmin eder. Örneğin, giriş cümlesi "PKOS" ise, model yüksek bir olasılıkla "yumurtalama bozukluğu", 'androjen' ve "insülin direnci" gibi ilgili tıbbi terimleri tahmin edebilmelidir.

Gizli katmanda, girdi katmanından gelen kelimenin tek atışlık vektör temsili, ağırlık matrisleri ile sabit boyutlu bir kelime gömülmesine dönüştürülür. Softmax işlevi daha sonra bu temsili çıktı katmanına taşımak için kullanılır ve burada model potansiyel bağlam sözcükleri için bir olasılık dağılımı oluşturur.

Klinik metinlerde birlikte geçen veya anlamsal ilişkileri olan kelimeler ("folikül" ve "ultrason" veya 'hirsutizm' ve "androjen" gibi) bu öğrenme sürecinin sonunda vektör uzayında birbirlerine yakın yerleştirilir. Sonuç olarak, tıbbi fikirleri anlamsal olarak karşılaştırabilecek bağlama duyarlı bir anlamsal temsil üretilir.

Bu yapı, otomatik etiketleme, literatür arama önerileri ve anlamsal benzerlik analizi gibi bilgi erişim uygulamalarına uygulanabilme avantajına sahiptir. Word2Vec bu çalışmada FAISS gibi vektör tabanlı arama algoritmaları için bir ön işleme aracı olarak ve eğitim verileri üzerinde kelime gömme için kullanılmıştır.

### FastText tekniği

FastText, özellikle büyük veri setlerinde iyi çalışan, metin kategorizasyonu için hafif ve hızlı bir tekniktir (Joulin vd. 2016). Sınıflandırmak ve tahmin etmek için bu vektöre doğrusal bir katman uygulanır. Model, her kelimeyi sabit boyutlu bir vektörle

temsil eder. Giriş cümlesi veya paragraf, bu kelime vektörlerinin ortalaması alınarak $\overrightarrow{v_{text}} = \frac{1}{n}\sum_{i=1}^{n}\overrightarrow{v_{wi}}$ şeklinde tek bir metin vektörüne dönüştürülür. Sınıflandırmak ve tahmin etmek için bu vektöre doğrusal bir katman uygulanır.

FastText, etiketleri açıkça sınıflandırmak yerine hiyerarşik softmax olarak bilinen bir teknik kullanır. Bu yaklaşım, her etiketin bir yaprak düğüm olarak temsil edildiği Huffman algoritması tarafından üretilen ikili bir ağaç kullanır. Bir etiketin ait olduğu yoldaki düğümler için oluşturulan sigmoid fonksiyonları, etiketin olasılığını belirlemek için çarpılır:

$$P\,(y|x) = \prod_{n \in path(y)} \sigma(sign(n)\,.\,\overrightarrow{v_n}\;\overrightarrow{v_{text}} \tag{2.2}$$

Burada $\overrightarrow{v_{text}}$, girdiye ait metin vektörüdür; $\overrightarrow{v_n}$, her bir iç düğümün ağırlık vektörüdür; σ sigmoid fonksiyonudur; `sign(n)` ise yol üzerindeki yönü belirtir

Huffman ağacı nedeniyle daha sığ düğümler daha sık etiketlere sahiptir. Bu yapı sayesinde hem eğitim hem de çıkarım süresi optimize edilir. Modelin, özellikle geniş ve düzensiz sınıf dağılımlarında, yaygın olmayan etiketlerle bile iyi çalışmasını mümkün kılar. Ayrıca, bağlamsal verileri kaydetmek için FastText tarafından N-gram teknikleri de kullanılmaktadır. Kelimelerin hem tekil hem de sıralı biçimlerine odaklanarak, kısa kelime dizilerinin bağlamını daha iyi anlayabilmektedir.

Özetle FastText, doğruluk ve verimlilik açısından büyük ölçekli metin sınıflandırma sorunlarını çözmek için kullanışlı ve verimli bir yaklaşımdır. Bu bağlamda, özellikle çok sayıda sınıf içeren görevler için alternatif yaklaşımlara göre kayda değer faydalar sağlamaktadır.

### 2.2.3. Kosinüs benzerliği

Kosinüs benzerliği, çok boyutlu bir uzayda, özellikle de yüksek boyutlu uzaylarda iki vektör arasındaki açının kosinüsünü hesaplayarak ne kadar benzer olduklarını belirleyen matematiksel bir metriktir. Basitçe söylemek gerekirse, iki öğeyi sadece bireysel değerlerine göre karşılaştırmak yerine, işaret ettikleri "yönü" inceleyerek aralarındaki ilişkiyi kavramamızı sağlar (Miesle 2025).

DDİ ve veri analizi gibi alanlarda kosinüs benzerliği oldukça kullanışlıdır. DDİ'de belge kümeleme, duygu analizi ve metin madenciliği gibi uygulamalar için yaygın olarak kullanılır. Kesin öneriler veya sınıflandırmalar yapmak için metrik, anlamsal benzerliklerini belirlemek üzere iki metnin karşılaştırılmasına yardımcı olur. Kosinüs benzerliği, iki vektör arasındaki açının kosinüsünü temel alarak benzerliği ölçen bir yöntemdir ve Denklem 2.3'te gösterildiği gibi hesaplanmaktadır.

$$cosine_{similarity(A,B)} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} * \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{2.3}$$

Benzerlik oranı, iki vektör arasındaki açının kosinüsünü hesaplamak için formül[1] kullanılarak bulunur. 1 değerine yaklaştıkça vektörlerin yönleri birbirine benzer hale gelir, 0 değerine yaklaştıkça yönleri ortogonal hale gelir ve -1 değerine yaklaştıkça yönleri zıt yönlüdür.

### 2.2.4. FAISS

Meta YZ Research tarafından geliştirilen FAISS, yüksek boyutlu verilerle çalışırken benzerlik aramaları ve kümeleme işlemleri için mükemmel bir açık kaynaklı kütüphanedir. Özellikle derin öğrenme ve makine öğrenimi uygulamaları, büyük veri kümelerindeki benzer öğeleri hızlı ve doğru bir şekilde bulmak için bu kütüphaneyi kullanır (Johnson vd. 2019). FAISS, geleneksel arama motorlarının içerik tabanlı aramalarda yetersiz kaldığı durumlarda metin, görüntü ya da ses gibi yapılandırılmamış verilerle çalışır. FAISS, PKOS tanısı gibi karmaşık tıbbi verilerin, ultrason görüntüleri, laboratuvar sonuçları veya klinik semptomlar gibi çok boyutlu bilgilerin analizini vektör temsillerine dönüştürüp karşılaştırarak teşhis süreçlerini hızlandırabilir.

FAISS, Brute Force yöntemlerini ve YSA algoritmalarını veri arama işlemlerinde birleştirir. Bu nedenle, özellikle büyük veri kümeleriyle çalışırken işlem süresi önemli ölçüde kısaltılabilir. Ek olarak, FAISS'in paralel işleme desteği ve GPU hızlandırması, büyük veri setlerini hızlı bir şekilde analiz etmeyi mümkün hale getirir (Johnson vd. 2019).

FAISS yalnızca veri aramakla kalmaz, aynı zamanda benzerlik analizi yaparak verileri daha küçük, düşük boyutlu vektörler halinde depolar. Veri analizini hızlandırırken bellek kullanımını optimize eder. Bu özellik, FAISS'in özellikle sınırlı kaynaklara sahip cihazlarda son derece verimli olmasını sağlar. FAISS, arama doğruluğunu ve hızını kullanıcı ihtiyaçlarına göre özelleştirme yeteneğine sahiptir, bu da bir diğer önemli avantajdır. FAISS, bu esnekliği nedeniyle çok çeşitli amaçlarla uyumlu bir araçtır.

FAISS kütüphanesi bu araştırmada PCOS Tanı Destek Sisteminin temel bir parçası olarak eksik bilgilerin tespiti ve literatüre dayalı bilgi erişimini içeren faaliyetler için kullanılmıştır. Klinik notlar veya doktor tarafından girilen serbest metin ifadeleri önceden vektörize edilmiş tıbbi referans cümleleriyle karşılaştırılarak anlamsal olarak en ilişkili ifadeler bulunur. Doktor tarafından sağlanan bilgilerdeki boşlukları tespit ederek, sistem dinamik olarak sorular üretebilmiş ve doktora öneriler sunabilmiştir. Örneğin, hasta ifadesinde "adet düzensizliği" geçtiği ancak "androjen fazlalığı" hakkında bilgi verilmediği durumlarda, FAISS benzerlik skoruna göre eksik kalan semptom grupları tespit edilip, bu eksiklikler modelin literatürle uyumlu önerileri doğrultusunda tamamlanmıştır.

[1] $A_i$ve $B_i$sırasıyla A ve B vektörlerinin i. bileşenlerini temsil eder.
$\sum_{i=1}^{n} A_i B_i$, iki vektörün iç çarpımını (noktasal çarpım) ifade eder.
$\sum_{i=1}^{n} A_i^2$ ve $\sum_{i=1}^{n} B_i^2$, sırasıyla A ve B vektörlerinin normlarının kareleridir. Bu ifadelerin karekökü alınarak Öklidyen norm elde edilir.

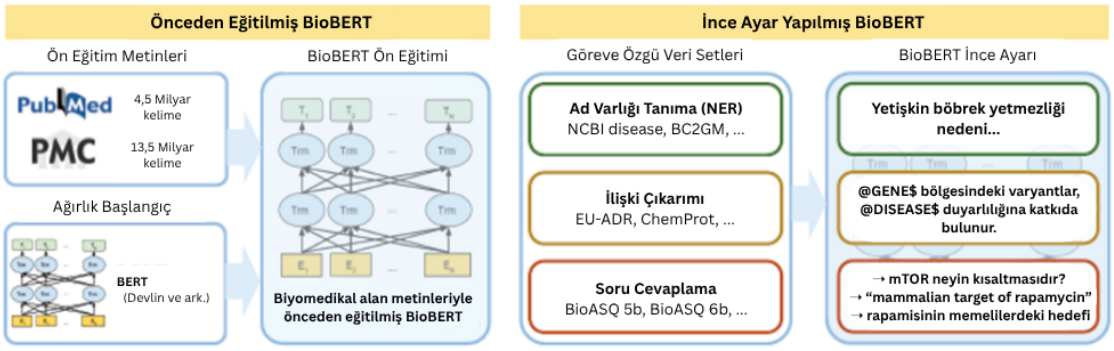
Bu modelin yapısı, geleneksel sabit soru-cevap sistemlerinin aksine hastaya özgü bilgileri tamamlamakta ve hekime özgü geri bildirim sunmaktadır. FAISS'in hızlı sorgulama hızı ve yüksek boyutlu vektörler arasında etkili karşılaştırmalar yapabilme kapasitesi sayesinde sistemin cevap üretme süresi son derece optimize edilmiştir. Ayrıca, her FAISS sorgusu için en yakın üç ila beş referans belgeyi özetlemek ve açıklığa kavuşturmak için Gemini modeli kullanılmıştır. Sonuç olarak, algoritma sadece eksik verileri tespit etmekle kalmamış, aynı zamanda doktora bilime dayalı açıklamalar da sunmuştur.

Bu bağlamda, FAISS uygulaması sistemin açıklayıcı ve bilgi tabanlı bileşenlerini geliştirmiş ve gerçek zamanlı klinik karar destek mekanizmasının dinamik bir yapıya kavuşmasını mümkün kılmıştır.

### 2.2.5. Doğal dil işleme modelleri

DDİ'deki gelişmeler, özellikle biyomedikal ve sağlık metinlerinin analizi için tasarlanmış modeller üretmiştir. Genel amaçlı dil modellerinin dezavantajlarını aşmak ve biyomedikal literatürün kendine özgü kelime dağarcığını daha doğru bir şekilde yakalamak için, alan odaklı önceden eğitilmiş modeller geliştirilmiştir. Aşağıda bu bağlamda sıklıkla kullanılan ve bu çalışmada karşılaştırılan iki önemli model yer almaktadır: BioBERT ve SciBERT.

**BioBERT:** BioBERT, (Lee vd. 2020) adı verilen önceden eğitilmiş bir dil temsil modelidir, özellikle biyomedikal alandaki metin madenciliği uygulamaları için oluşturulmuştur. Temelini Google'ın çift yönlü dönüştürücü mimarisi BERT oluşturmaktadır. Wikipedia (yaklaşık 2,5 milyar kelime) ve BooksCorpus (yaklaşık 0,8 milyar kelime) verileri üzerinde önceden eğitilen BioBERT, genel dil modeli BERT'in ağırlıkları kullanılarak başlatılır. Daha sonra PubMed CKDKM (PMC) tam metin makalelerinden 13,5 milyar kelime ve PubMed özetlerinden 4,5 milyar kelimeden oluşan bir biyomedikal derlem kullanılarak bir kez daha ön eğitime tabi tutulmuştur. Transfer öğrenimi görüntü işlemede sıklıkla kullanılmaktadır ve bu yöntemin DDİ'deki karşılığı olduğu düşünülmektedir. Şekil 2.4'te BioBERT'in genel hali gösterilmiştir.




**Şekil 2.4.** BioBERT'in ön eğitimi ve ince ayar süreci

Bu modelin temel amacı, genel alanlarda eğitilen dil modellerinin karşılaştığı biyolojik metinlerde görülen dilsel farklılıkların ve terminolojik özelliklerin üstesinden gelmektir. Özellikle, biyomedikal yayınlarda sıklıkla kullanılan benzersiz terminoloji, kısaltmalar ve genetik ifadelerin genel dil modelleri tarafından yeterince tanımlanması genellikle imkansızdır. Sonuç olarak, BERT gibi güçlü bir dil modeli, alana özgü veriler kullanılarak yeniden eğitildiğinde bu tür sınırlı alanlarda çok daha iyi performans göstermektedir.

Modelin ön eğitiminin ardından, adlandırılmış varlık tanıma, ilişki çıkarımı ve soru-cevaplama gibi alan-özgü görevler için küçük ayarlamalar yapılmıştır. Bu görevlerde kullanılan veri setleri arasında NCBI Disease ve BC2GM (adlandırılmış varlık tanıma için), EU-ADR ve ChemProt (ilişki çıkarımı için) ve BioASQ (soru-cevaplama için) yer almaktadır. Yapılan araştırmalar, BioBERT'in her üç görevde de hem temel BERT modeline hem de o zamanlarda kullanılan yöntemlere kıyasla önemli ölçüde daha iyi performans gösterdiğini göstermiştir. Modelin dil temsillerini geliştirdiği ve görev performansını artırdığı, özellikle PubMed ve PMC verileriyle yapılan ek ön eğitimle kanıtlanmıştır (Raghudeep 2019).

Ek olarak, modelin sözcük temsilinde kullanılan WordPiece Tokenizer yöntemi, nadir veya alan-özgü kelimelerin parçalanması yoluyla öğrenilebilirliğini geliştirmiştir. Örneğin, "Immunoglobulin" kelimesi "I ##mm ##uno ##g ##lo ##bul ##in" olarak bölünmüştür. Bu, sözlük dışı kelime sorununu önemli ölçüde azalttı ve modelin biyomedikal metinlerdeki geniş terminolojik çeşitliliğe karşı daha esnek hale gelmiştir.

BioBERT'in ön eğitimi 200 bin ila 700 bin adım arasında gerçekleşmiştir ve güncel GPU donanımları (örneğin; NVIDIA V100 32GB) ile modelin ince ayar süreci yaklaşık bir saat içinde tamamlanabilir. Sonuç olarak, BioBERT modeli, büyük miktarda etiketsiz alan verisi ile etkili dil temsillerini öğrenerek, biyomedikal alandaki etiketli veri eksikliğine rağmen çok sayıda görevin başarısını önemli ölçüde iyileştirir. Bu, gelecekte farklı sektörler için oluşturulacak alan-özgü BERT türevleri için iyi bir örnektir.

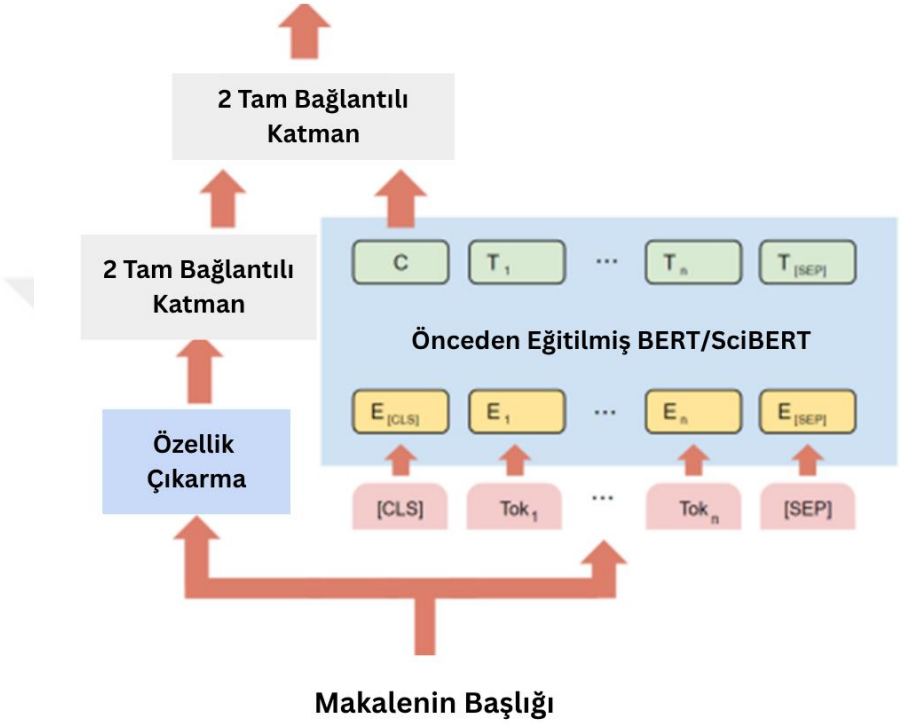
**SciBERT:** SciBERT, (Beltagy vd. 2019), BERT mimarisi üzerine kurulan ve özellikle bilimsel literatür için önceden eğitilmiş bir dil modeli olarak bilimsel metin madenciliğinde büyük başarılar elde etti. Model, genel dil modellerinden farklı olarak, bilimsel bağlamda daha hassas ve anlamlı dil temsilleri elde etmeyi amaçlamaktadır. Bunun nedeni, 1,14 milyon bilimsel makaleden oluşan geniş bir seçki üzerinde ön eğitim almaktır. Bu, DDİ ile ilgili görevlerde bilimsel içeriğe özgü terminolojiyi ve yapıları daha iyi öğrenmelerine olanak tanır.

SciBERT, bilimsel metinlerde etiketlenmiş veri miktarının sınırlı olması nedeniyle yalnızca genel dil verileriyle değil, bilimsel makalelerden oluşan özel bir derlemle önceden eğitilmiş bir modeldir. Bu nedenle, adlandırılmış varlık tanıma, belge sınıflandırma, metin benzerliği analizi ve ilişki çıkarımı gibi DDİ görevlerinde daha iyi performans göstermiştir. Özellikle BioBERT gibi alan özelleştirilmiş BERT türevlerine kıyasla daha geniş bir bilim alanı kapsayan bir model olması, çeşitli bilim alanlara kolayca uyarlanmasını sağlamaktadır.

SciBERT'in kullanıldığı örnek sınıflandırma mimarisi Şekil 2.5'te gösterilmektedir. Modelin temel amacı, bilimsel tıbbi makalelerden elde edilen metinlerin, PKOS ile ilişkili içeriklerine göre sınıflandırılmasıdır. Bu yapı, SciBERT

modelinin önceden eğitilmiş gömülü temsil yeteneklerinden faydalanarak, metinleri üç farklı başlık altında otomatik olarak etiketleyen çok katmanlı bir sinir ağı mimarisi içermektedir:

- PKOS'un Tanısı ve Klinik Özellikleri
- PKOS'un Ayırıcı Tanısı
- PKOS'un Epidemiyolojisi, Patogenezi ve Tedavisi




**Şekil 2.5.** SciBERT tabanlı sınıflandırma modeli

Model, bilimsel yayınların başlık ve içerik metinlerini işleyerek bağlamsal temsiller oluşturmuş, ardından bu temsilleri yayınların ait olduğu sınıfı belirlemek için kullanmıştır. SciBERT, PKOS gibi çeşitli tıbbi uzmanlık alanlarını içeren senaryolar için mükemmel bir seçimdir, çünkü bilimsel terminoloji içeren metinleri okumada başarılıdır. SciBERT modeli metinler arasındaki hem yüzeysel benzerlikleri hem de derin anlamsal ilişkileri öğrenebildiğinden, bu çalışmada özellikle "Ayırıcı Tanı" gibi klinik açıdan zor sınıflandırmalarda daha güvenilir ve doğru sonuçlar üretmiştir.

Bilimsel materyallerle önceden eğitildiği için, bu çalışmadaki SciBERT tabanlı model - BERT mimarisi üzerine inşa edilmiş olmasına rağmen - tıbbi metinler için daha iyi sınıflandırma performansı elde etmiştir. Bu yapı, modelin klinik notlara ek olarak PubMed ve UpToDate gibi kaynaklardan gelen literatür paragraflarını anlamasını ve doğru bir şekilde sınıflandırmasını sağlamıştır.

Sonuç olarak, SciBERT, önceden eğitilmiş bir model olarak öne çıkıyor ve bilimsel alanlarda DDİ ihtiyaçlarını karşılamak için son derece uygundur. Bu nedenle,

sadece klasik metin madenciliği görevlerinde değil, aynı zamanda yaratıcı içerik tanıma ve sınıflama gibi alanlarda da kullanılabilmektedir.

### 2.2.6. Önceden eğitilmiş dil modellerinin ince ayarı (Fine-Tuning)

Büyük ölçekli önceden eğitilmiş dil modellerinin geliştirilmesi, son yıllarda DDİ alanında önemli ilerlemelere yol açtı. RoBERTa gibi modeller, milyonlarca cümle üzerinden dil yapısına genel bir bakış sağlar; ancak, sağlıkla ilgili metinleri sınıflandırmak gibi daha sınırlı görevlerde, doğrudan kullanıldıklarında BERT'in başarısı sınırlıdır. Bu nedenle, bu modeller hedef görev için özel olarak yeniden eğitilmelidir. Literatürde ince ayar veya ince ayar olarak adlandırılan bu prosedür, "fine-tuning" olarak adlandırılmaktadır (Devlin vd. 2019).

Model, önceden öğrendiği dil temsillerini korur ve görev verisi üzerinden ağırlıkları yeniden optimize ederek yeni bir bağlama girer. Eğitim genellikle göreve özgü etiketli veriler kullanılarak gerçekleştirilir ve modelin tüm parametreleri bu verilere göre güncellenir. Özellikle sağlık alanında, bu yöntem, az sayıda veri ile yüksek doğruluk elde edilmesini sağladığı için yaygın olarak kullanılmaktadır. Çeşitli çalışmalar, BioBERT ve SciBERT gibi alan özelinde önceden eğitilmiş modellerin genel modellerden daha iyi performans gösterdiğini göstermiştir.

Bu çalışmada, görev verilerini kullanarak BioBERT ve SciBERT gibi büyük, önceden eğitilmiş dil modelleri, PKOS ile ilgili bilimsel metinleri sınıflandırmak için yeniden eğitilmiştir. Eğitim süreci boyunca her model, klasik tam ince ayar veya tam ince ayar yöntemiyle optimize edildi; yani, eğitim verisi tüm model ağırlıklarını değiştirdi. Her modelin beş eğitim dönemi (epoch) sonunda gösterdiği doğruluk ve F1 performansları, ilk değerlendirmenin temelini oluşturdu; bu aşamada SciBERT'in genelleme yeteneği açısından öne çıktığı belirlendi. Bununla birlikte, daha fazla eğitimin performansı nasıl etkilediğini belirlemek için süreç yediye kadar uzatılmıştır. Bu eğitimin sonunda, modellerin test verilerinin sonuçları yeniden incelenmiştir. Böylece, modelin son hali ile erken öğrenme dönemindeki davranışlar arasındaki performans değişimleri karşılaştırmalı olarak ortaya konmuştur. Bu yöntem sadece en iyi sonuçları sağlamakla kalmaz, aynı zamanda modellerin öğrenme eğrilerini inceleyerek hangi modellerin daha hızlı istikrar kazandığını da göstermiştir. Doğruluk, karışıklık matrisleri ve sınıf bazlı F1 puanları eğitim çıktılarını inceledi. Sonuçlar, görev özelinde yapılan bu iyileştirme sürecinin, özellikle PKOS gibi semptom çeşitliliği yüksek bir tıbbi konuda önemli faydalar sağladığını göstermektedir.

## 2.3. Büyük Dil Modelleri

BDM, derin öğrenme tabanlı sinir ağlarını tasarlar ve milyarlarca parametreye sahiptir. Bu modeller, dilin yapısal ve anlamsal örüntülerini belirlemek için çok sayıda yazılı veri analiz eder. Bununla birlikte, girdilere karşılık, en olası kelimeyi veya kelime dizisini oluşturma yetenekleri, istatistiksel örüntü tanıma yeteneklerine ve bağlamdan çıkarım yapma yeteneklerine bağlıdır. Bu yönüyle, BDM'ler insan benzeri metin yazmada oldukça iyidir.

Bu modeller, her kelimenin çevresindeki bağlamı dikkate alarak, dilin doğal akışına uygun bir şekilde yanıt verir (Vaswani vd. 2017). BDM'ler, hatırladıkları belirli ifadeleri veya cümleleri doğrudan saklamak yerine, öğrendikleri metinlerdeki istatistiksel

bağlantıları kullanarak en iyi sonuçları üretmeye çalışırlar. Bu, önceki ifadeleri birebir hatırlamak ya da kopyalamak yerine, mevcut durumda en uygun kelimeyi tahmin etmelerini gerektirir (Suri vd. 2024).

BDM'ler, sadece dil kalıplarını ezberlemekten öteye geçerek, bu bağlamsal üretim süreci aracılığıyla yeni ve tutarlı metinler oluşturabilirler. Bu nedenle, özetleme, çeviri, soru yanıtlama ve içerik oluşturma gibi DDİ görevlerinde üstün performans gösterebilirler.

### 2.3.1. Gradio kütüphanesi

Gradio, herkesin makine öğrenimi modellerini kullanmasına izin veren açık kaynaklı bir Python kütüphanesidir. Tek bir satır kodla web tabanlı arayüzler oluşturmak için teknik bilgi gerektirmez. Bu nedenle, makine öğrenimi modelleri, API'ler veya basit Python fonksiyonları hızla interaktif demolara dönüşebilir ve teknik olmayan kullanıcılarla paylaşılabilir (Abid vd. 2019). Gradio'nun en büyük gücü, çeşitli alanlardan insanların bir araya geldiği projelerde iletişimi kolaylaştırmasıdır. Örneğin, bir veri bilimcisi ile bir doktorun işbirliğinde modeller daha kolay erişilebilir hale gelir. Gradio, Jupyter ve Google Colab gibi ortamlarda çalışır ve Hugging Face Spaces gibi platformlarda barındırılabilir. Python 3.10 ve üstü sürümlerle uyumludur. Böylece modeller sürekli olarak erişilebilir kalır ve otomatik olarak oluşturulan bağlantılar sayesinde dünyanın her yerinde paylaşılabilir (Gradio 2024).

Gradio, metin ve webcam gibi çeşitli giriş-çıkış türlerini desteklediği için çok yönlü bir araçtır. Bir vaka çalışması, Gradio'nun endüstride ve akademik çalışmalarda önemli bir rol oynadığını göstermiştir. Örneğin, bir makine öğrenimi uzmanı ile bir kardiologun işbirliğini kolaylaştırmıştır (Abid vd. 2019).

### 2.3.2. Gemini dil modeli

Google DeepMind tarafından geliştirilen Gemini, multimodal BDM ailesinin bir üyesi olan bir YZ sistemidir. Bu model ailesi, bilgi çıkarma, içerik oluşturma, DDİ ve soru yanıtlama gibi görevlerde mükemmel doğruluk ve açıklanabilirlik sunar. Metin, görsel ve diğer veri türleriyle entegre edilebilir. Klinik karar destek sistemleri, Gemini'nin özellikle tıp gibi bilgi yoğun alanlarda kaynak metinlerden açıklamalar sağlama kapasitesi sayesinde devrim yaratmaktadır (Google DeepMind 2023).

Gemini, bu çalışmada PKOS ile ilgili tıbbi makalelerden materyallerle etkileşime girmek, kullanıcı teşhis sorgularına anlayışlı yanıtlar vermek ve cesaret verici YZ tabanlı öneriler üretmek için kullanılmıştır. Modelin açıklamaları, ROUGE, BLEU ve BERTScore gibi metin benzerliği ölçütleri kullanılarak değerlendirildi; bu da YZ'nın klinik bir ortamda güvenilir ve aydınlatıcı bilgi üretme yeteneğinin ölçülmesini sağladı. Gemini'nin bu konudaki başarısı, gelecekte klinik karar destek sistemleri ve tıbbi metin madenciliği için olumlu bir model teşkil etmektedir.

## 2.4. Değerlendirme Teknikleri

YZ destekli DDİ sistemleri, otomatik olarak metin üretebilen ve bu metinleri anlamlı bir şekilde değerlendirebilen araçlar oluşturmayı hedefliyor. Bununla birlikte, bu metinlerin tutarlılığını, doğruluğunu ve kalitesini değerlendirmek çok önemlidir.

Özellikle klinik karar destek sistemleri gibi hassas uygulamalarda, yazılan metinlerin doğruluğu ve güvenilirliği çok önemlidir. Bu bölümde, YZ destekli bir klinik araç, PKOS teşhisi için kullanılan metin değerlendirme tekniklerini ele almaktadır. BLEU, ROUGE ve BERTScore gibi metrikler, sistemin ürettiği metinlerin doktorlar tarafından hazırlanan referans metinlerle ne kadar uyumlu olduğunu ölçmek için kullanılır. Bu değerlendirme yaklaşımları, metinlerin yüzeysel benzerliklerini ve anlamsal doğruluğunu analiz ederek sistemin etkinliğini ve güvenilirliğini garanti eder.

### 2.4.1. BLEU skoru

BLEU, DDİ alanında, özellikle makine çevirisi ve metin üretimi gibi uygulamalarda, otomatik olarak üretilen metin kalitesini değerlendirmek için kullanılan bir ölçüttür. Sohbet robotları, metin özetleme ve klinik rapor üretimi gibi çeşitli DDİ görevlerinde günümüzde yaygın olarak kullanılmaktadır. Bu robotlar ilk olarak insan çeviri çıktılarını karşılaştırmak için geliştirilmiş olsa da (Papineni vd. 2002). Bu çalışma, PKOS teşhisi için YZ destekli bir klinik araç geliştiriyor. BLEU skoru, sistemin ürettiği metin tabanlı çıktıların insan tarafından yazılan referans metinlerle karşılaştırılarak değerlendirilebilir.

BLEU skoru, otomatik olarak üretilen bir metnin referans metinlere ne kadar benzediğini değerlendirmek için kullanılan bir değerlendirme ölçütüdür. Bir gramlık tek kelime, iki kelimelik gruplar veya daha uzun kelime dizilerinin referans metinlerdeki eşleşmelerini hesaplar. BLEU, kelime dizisi benzerliklerini incelerken, yazılan metnin uzunluğunu da inceler. Çok kısa veya çok uzun çıktılar, BP ile cezalandırılır. Referans metne çok benzeyen metinler, 0 ile 1 arasında bir skora sahip değerler alır. PKOS teşhisi için tasarlanmış bir YZ sistemi, örneğin, sistemin teşhis özetini doktorların referans özetleriyle karşılaştırarak BLEU skoruyla ölçebilir. BLEU skoru, Denklem 2.4'te gösterildiği gibi hesaplanmaktadır (Doshi 2021).

$$\mathrm{BLEU(N)} = \mathrm{BP} * \exp\left(\sum_{n=1}^{N} w_n \, log \, p_n\right) \tag{2.4}$$

$p_n$, n-gram hassasiyetini gösterir, yani üretilen metindeki n-gram kelime dizilerinin referans metinlerde ne kadar yer aldığını gösterir. 1-gram tek kelimeleri ifade ederken, 2-gram iki kelimelik grupları ifade eder. Kısacası, bu oran, sistemin yazdığı yazıların doktorların yazdıkları yazılara içerik olarak ne kadar benzediğini gösteriyor.

$w_n$, her n-gram seviyesinin ağırlığını gösterir. Örneğin, 1 gramdan 4 grama kadar her biri için $w_n$= 1/4 kullanılır ve bu ağırlıklar tipik olarak eşit olarak dağıtılır. Bu, farklı uzunluktaki kelime gruplarının skora katkıda bulunmasının nasıl dengelendiğini sağlar.

BP, metnin referans metne uygunluğunu kontrol eder. Sistemin çok kısa bir metin üretmesi durumunda bu ceza uygulanır ve skor düşüyor. Bu nedenle, eksik veya yetersiz bilgi içeren metinler daha az puan alır. BP, Denklem 2.5'te gösterildiği gibi hesaplanmaktadır (Doshi 2021).

$$BP = \begin{cases} 1, & if \; c > r \\ e^{\left(1-\frac{r}{c}\right)}, & if \; c \leq r \end{cases} \tag{2.5}$$

Burada c, sistemin ürettiği metnin kelime sayısını yani uzunluğu; r, referans metnin kelime sayısını göstermektedir. Formül, üretilen metinlerin referans metne göre daha kısa olması durumunda, yani $c \leq r$, skorun düşmesini sağlar. Bunun temel nedeni, kısa metinlerin kritik bilgileri atlama olasılığının yüksek olmasıdır. Örneğin, PKOS tanısı için geliştirilen bir YZ sistemini ele alalım. Bu sistem, hastaların klinik bulgularını özetleyerek teşhis raporları oluşturabilir. Ancak, oluşturulan bu raporların, klinik gereklilikleri ve tıbbi standartları karşılayabilmesi için içerdikleri bilgilerin kapsamlı ve eksiksiz olması gerekmektedir. Eğer YZ sistemi tarafından üretilen metinler, önemli klinik detayları eksik bırakırsa, bu durum tıbbi karar alma sürecini olumsuz etkileyebilir. Bu bağlamda, değerlendirme formülünün üretilen metnin uzunluğunu göz önünde bulundurarak skoru düşürmesi, yalnızca dil bilgisel doğruluğu değil, aynı zamanda metnin bilgi yoğunluğunu ve klinik standartlara uygunluğunu da sağlamayı amaçlar. PKOS gibi hassas bir klinik alanda, bu tür bir değerlendirmenin uygulanması, üretilen raporların güvenilirliğini ve doğruluğunu artırarak klinik pratikte güvenli ve etkili bir karar destek sistemi sunar.

### 2.4.2. ROUGE skoru

DDİ alanında otomatik olarak üretilen metinlerin kalitesini değerlendirmek için ROUGE skoru kullanılır. Metin özetleme, makine çevirisi ve otomatik raporlama gibi görevlerde sistemin ürettiği metinlerin referans metinlerle ne kadar örtüştüğünü ölçer. ROUGE, özellikle geri çağırma odaklı bir yaklaşımla metin benzerliğini inceler (Lin 2004). PKOS tanısı için tasarlanan YZ destekli klinik araçta, ROUGE skoru, sistemin ürettiği teşhis özetleri veya hasta raporlarının içeriğe ne kadar uygun olduğunu değerlendirmek için kullanılabilir.

ROUGE, temel olarak n-gram'lar yani en uzun ortak alt diziler, kelime sıraları veya kelime dizileri gibi metin birimlerini karşılaştırır. ROUGE-N, ROUGE-L ve ROUGE-S en yaygın türleri arasındadır. ROUGE-1 tek kelimelik dizileri ve ROUGE-2 iki kelimelik dizileri karşılaştırır. ROUGE, referans metindeki kelime dizilerinin üretilen metinde ne kadar yer aldığını ölçer ve bazen hassasiyet ve F1 skoru gibi ek metriklerle desteklenir. Lin'e (2004) göre, bu, sistemin kritik bilgileri kaçırıp kaçırmadığını anlamaya yardımcı olur. Örneğin, bir PKOS projesinde bir YZ sisteminin ürettiği teşhis raporunun, doktorların yazdığı rapordaki kritik klinik terimleri ve ifadeleri içerip içermediğini ROUGE skoruyla kontrol edilebilmektedir.

Çalışmada oluşturulan yapay zeka destekli sistem, her sınıfa özgü açıklayıcı klinik cümleler oluşturur ve kullanıcının serbest metinlerini sınıflandırır. Bu açıklamalar ile tıbbi referans metinleri arasındaki benzerlik derecesi ROUGE-N ve ROUGE-L metrikleri kullanılarak değerlendirilmektedir. Bu metrikler, özellikle anlatı yapısı (hangi terimler hangi sırayla dahil edilmiştir?) ve içerik doğruluğu (hangi terimler dahil edilmiştir?) söz konusu olduğunda modelin performansını değerlendirmek için çok önemlidir.

#### ROUGE-N: Tıbbi Anahtar Terimlerin Kelime Düzeyinde Benzerliği

Oluşturulan ek açıklama ve referans ek açıklamanın n-gram eşleşmeleri ROUGE-N metriği kullanılarak değerlendirilir. Bir, iki veya daha fazla ardışık kelimeden oluşan gruplar bu bağlamda "n-gram" olarak adlandırılır. Örneğin "Oligomenore teşhisi" 2 gramlık (bigram) bir cümledir. Bu ifade hem model çıktısında hem de doktorun

açıklamasında yer alıyorsa bir eşleşme olarak kabul edilir. ROUGE-N değeri, modelin tıbbi açıdan doğru açıklamalar üretip üretmediğini belirlemek için Denklem 2.6 kullanılarak belirlenir.

$$ROUGE - N = \frac{\text{Referans doktor açıklamasındaki n − gram'ların, model açıklamasında eşleşme sayısı}}{\text{Referans açıklamadaki toplam n − gram sayısı}} \quad (2.6)$$

Bu değer, modelin "hiperandrojenizm", "ovulasyon bozukluğu", "insülin direnci" gibi tıbbi anahtar terimleri kelime düzeyinde ne kadar doğru kullandığını ortaya koymaktadır.

**ROUGE-L: Klinik Anlatımın Sırasal Doğruluğu**

ROUGE-L metriği, içeriğin kelimelerinin mevcut olmasının yanı sıra doğru bir şekilde sunulup sunulmadığını da değerlendirir. EUOAD fikri bu değerlendirmenin temelini oluşturmaktadır. ROUGE-L puanı, hem referans açıklamada hem de model çıktısında aynı sırada görünen en uzun kelime dizisinin uzunluğu ile artar (örneğin, "önce klinik değerlendirme, sonra laboratuvar testi").

$$ROUGE - L = \frac{\text{Model ve referans açıklama arasındaki EUOAD uzunluğu}}{\text{Referans açıklamanın toplam uzunluğu}} \quad (2.7)$$

Modelin tıbbi içeriği klinik bir akış içinde aktarıp aktarmadığını belirlerken bu metrik özellikle yardımcı olur. Önce semptomun tartışılması ve ardından tanısal testlere geçilmesi gibi doğal sıralama bu durumda faydalıdır.

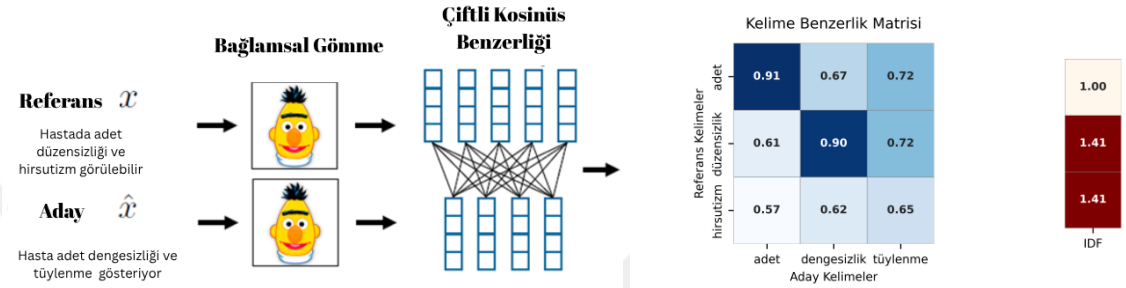
### 2.4.3. BERTScore

BERTScore, bağlamsal anlamsal benzerliğe odaklanan bir değerlendirme ölçütüdür. Özellikle dil üretimi görevlerinde sistem çıktılarının insan yazımı metinlerle ne kadar örtüştüğünü analiz etmek için geliştirilmiştir. BERTScore, kelimelerin bağlam içindeki anlamlarını karşılaştırır. Geleneksel metrikler olan BLEU veya ROUGE, yüzeysel n-gram eşleşmelerine dayanır ve anlam bilgisini yeterince yansıtamazlar (Zhang vd. 2019).

Bu yöntem, önceden eğitilmiş BERT modeli aracılığıyla referans ve aday cümlelerdeki her bir kelimeyi bağlamsal vektörlere dönüştürmektedir. Bu vektörler arasında çift yönlü kosinüs benzerliği hesaplandıktan sonra, her bir kelimenin en yakın eşleniği bulunmaktadır. Bu prosedür, hem aday cümledeki kelimelerin referansa göre karşılaştırılmasıyla geri çağırma hem de aday cümledeki kelimelerin referansa göre karşılaştırılmasıyla hassasiyet metriklerini oluşturmaktadır. Daha sonra bu iki metrik harmonik ortalama ile birleştirilerek F1 puanı elde edilmektedir.

Çalışmada kullanılan BERTScore metriğinin işleyişi Şekil 2.6'da gösterilmektedir. Örneğin, literatürde yer alan "Hastada adet düzensizliği ve hirsutizm görülebilir" referans cümledir, ancak hekimin sisteme girdiği "Hasta adet dengesizliği ve tüylenme belirtileri gösteriyor" model tarafından üretilen aday cümledir. BERT modeli ilk aşamada her iki cümledeki kelimelerin bağlamsal vektörlerini elde etmek için kullanılır. Daha sonra bu vektörlerin ikili kosinüs benzerliği hesaplanır. Her bir referans terime en çok benzeyen aday cümledeki kelime bulunur.

Örneğin, “hirsutizm” terimi bu işlem sonucunda ‘tüylenme’ ile “adet düzensizliği” terimi ise “adet dengesizliği” ile çok yüksek benzerlik skoruna sahiptir. Kelime benzerliğine dayalı maksimum skorlar şeklin sağ bölümünde gösterilmektedir. Bu skorlar hesaplanırken klinik açıdan önemli terimlerin katkısını artırmak ve “ve” ve “hasta” gibi yaygın kelimelerin etkisini azaltmak için IDF (ters doküman frekansı) ağırlıkları da dahil edilmiştir. Bu yapı, modelin hem yüzeysel hem de bağlamsal düzeyde doğru açıklamalar sağlama kabiliyetinin ölçülmesine olanak tanıyor. Çalışmaya göre, SciBERT modeli bu ifadeleri daha hassas bir şekilde eşleştirebildiği için BERTScore'da daha iyi performans göstermiştir.




**Şekil 2.6.** BERTScore hesaplama süreci

BERTScore sonuçlarını daha anlaşılır hale getirmek için, benzerlik skorları ampirik olarak elde edilmiş bir alt sınır değeri ile yeniden ölçeklendirilmiştir. Bu alt sınır, gerçek anlamda benzemeyen cümle çiftleri kullanılarak hesaplanmıştır. Bu, skorların daha anlamlı bir aralıkta değerlendirilmesini sağlamaktadır (Tsang 2024).

BERTScore, PKOS teşhisine yardımcı olacak YZ tabanlı sistemlerde oluşturulan teşhis özetlerinin uzman doktorlar tarafından yazılmış metinlerle anlamsal uyumunu değerlendirmek için etkili bir ölçüttür. Örneğin, sistem tarafından otomatik olarak oluşturulan bir hasta raporunun, tıbbi terminoloji ve semptom ifadeleri açısından insan yazımı bir raporla nasıl karşılaştırıldığını doğru bir şekilde ölçmek mümkündür.

Bununla birlikte, BERTScore bazı kısıtlamaları vardır. Özellikle BDM kullanılarak hesaplama maliyetleri artabilmektedir. Ek olarak, metinlerin sadece anlamsal benzerliklerini incelediği için yapısal sorunları veya dilbilgisi hatalarını dikkate almaz. Bu nedenle, dilbilgisel doğruluk ve terminolojik tutarlılık açısından daha fazla kontrol araçlarının klinik uygulamalarda BERTScore'a ek olarak kullanılması önerilmektedir (Reiter 2018).

Sonuç olarak, BERTScore, bağlamsal anlam ölçümü için etkili bir ölçüm aracı olarak hizmet eder ve klinik metinlerin kalitesinin değerlendirilmesinde önemli bir rol oynayabilmektedir.

## 2.5. Model Performans Kriterleri

Oluşturulan modelin güvenilirliğini ve kullanışlılığını göstermek için, YZ destekli kategorizasyon sistemlerinin performansını çeşitli kriterler kullanarak değerlendirmek çok önemlidir. Değerlendirme ölçütlerinin doğru seçilmesi, özellikle sınıflandırma hatalarının büyük yansımaları olabileceği sağlık sektöründe çok önemli hale gelmektedir.

Bu bağlamda doğruluk, geri çağırma, kesinlik, F1 puanı ve EAA gibi ölçütler, modelin genel performansının yanı sıra pozitif sınıfları yakalama kapasitesini de kapsamlı bir şekilde inceleme şansı sunar. Bu ölçütler, modelin teknolojik performansının yanı sıra klinik güvenilirliğini de değerlendirerek PKOS gibi ciddi hastalıkların tanımlanması için karar destek sistemlerinin sağlık hizmetlerine entegre edilmesini kolaylaştırır.

### 2.5.1. Eğri altı alan

EAA olarak bilinen performans kriteri, bir modelin sınıflandırma durumlarında pozitif ve negatif sınıflar arasında ne kadar iyi ayrım yapabildiğini ölçer. ROC eğrisinin eğri altındaki alanı olarak hesaplanır. Çeşitli sınıflandırma eşiklerinde modelin DPO ve YPO arasındaki bağlantı ROC eğrisi tarafından görsel olarak gösterilir. EAA, bu eğrinin altındaki alanı sayısal bir değere dönüştürerek modelin genel başarı seviyesinin bir özetini sunar (Fawcett, 2006). Bu değer, Denklem 2.8'de gösterildiği gibi hesaplanmaktadır.

$$EAA = \int_0^1 DPO(YPO)dYPO \tag{2.8}$$

DPO dikey eksende gösterilirken, YPO yatay eksende gösterilir, bu da 0 ile 1 arasında normalize edilmiş bir eğri oluşturur. Elde edilen alan, modelin ne kadar farklı olduğunu göstermektedir. EAA değeri 1'e yaklaştıkça modelin pozitif ve negatif kategorileri doğru bir şekilde ayırt ettiği anlaşılır. EAA'nin 0,5'den küçük olması, sınıfların tahminlerini yanlış yaptığı anlamına gelir (Hanley vd. 1982).

EAA, doğruluk gibi genel doğruluk oranı yerine daha güvenilir bir ölçümdür çünkü sağlık alanında veriler genellikle dengesiz dağılır. EAA, doğru tahmin sayısını değil, modelin tüm eşiklerdeki genel ayırt ediciliğini temsil eder (Saito vd. 2015). Bu nedenle, modelin güvenilirliğini değerlendirmek için EAA, PKOS gibi klinik açıdan kritik sınıflandırmalarda sıklıkla tercih edilen bir ölçüttür.

### 2.5.2. Doğruluk

Doğruluk, bir sınıflandırma modelinin ne kadar doğru tahmin yaptığını belirleyen en temel kriterlerden biridir. Model tarafından doğru sınıflandırılan vaka sayısı ile toplam örnek sayısı arasındaki oran olarak tanımlanır. En geniş anlamıyla, "modelin ne kadar doğru yaptığını" özetleyen bir orandır. Bu oran, Denklem 2.9'da gösterildiği gibi hesaplanmaktadır.

$$Doğruluk = \frac{DP + DN}{DP + DN + YN + YP} \tag{2.9}$$

Doğruluk 0 ile 1 arasında bir sayı olarak belirlenir ve genellikle yüzde olarak ifade edilir. Örneğin, %90 doğru bir model her 100 örnekten 90'ını doğru şekilde tahmin eder. Ancak kesinlik birçok durumda güvenilir bir gösterge olmayabilir. Özellikle eşit olmayan veri setlerinde yanıltıcı olabilir. Örneğin, PKOS gibi nadir bir durumu içeren veri kümelerinde pozitif örnek sayısı azsa, model çoğunluk sınıfını negatif tahmin ederek yüksek doğruluk elde edebilir. Ancak bu, iyi vakaları gözden kaçırdığı için sağlık açısından zararlıdır. Sonuç olarak, kesinlik, geri çağırma, F1 puanı ve EAA gibi doğruluk dışındaki metrikler kullanılarak değerlendirilmesi tavsiye edilmektedir (Saito vd. 2015).

Doğruluk, genel performansı gösteriyor gibi görünse bile, sağlık verilerinde karar destek sistemlerini değerlendirmek için tek başına yetersizdir. Özellikle önemli vakaları kaçırma riski olan sistemlerde, hatırlama ve özgüllük gibi ölçütlerle birlikte yorumlanması daha yararlıdır.

### 2.5.3. Geri çağırma

Sınıflandırma probleminin başarısının önemli bir göstergesi, modelin pozitif sınıfa ait örnekleri ne kadar iyi tanımlayabildiğini gösteren geri çağırmadır. Yüksek geri çağırma değerlerine sahip modeller, özellikle tıbbi teşhis gibi önemli alanlarda tercih edilir, çünkü bu metrik sistemin gerçekten hasta olan insanları ne kadar iyi yakaladığını açıkça gösterir. Doğru tahmin edilen pozitif örneklerin tüm gerçek pozitif örneklere oranı geri çağırma olarak bilinir. Başka bir deyişle, geri çağırma, modelin "sağlıklı" olarak tanımlayamadığı hasta insan sayısından doğrudan etkilenir. Bu oran, Denklem 2.10'da gösterildiği gibi hesaplanmaktadır.

$$Geri\ Çağırma = \frac{DP}{DP + YN} \tag{2.10}$$

Model, bu değer 1'e veya %100'e yaklaştıkça azalan hata ile pozitif sınıfı tanımlamıştır. Örneğin düşük bir hatırlama değeri, PKOS'lu birçok bireyin sistem tarafından sağlıklı olarak sınıflandırıldığını gösterir ki bu da önemli terapötik tehlikeler sağlayabilir. Sonuç olarak, hafıza hem modelin başarısı hem de sistemin sosyal veya klinik etkisi için bir gösterge görevi görür (Sokolova vd. 2009; Powers, 2020).

### 2.5.4. Kesinlik

Kesinlik adı verilen bir başarı ölçütü, bir sınıflandırma modelinin pozitif olarak öngördüğü vakaların kaçının aslında doğru olduğunu gösterir. Başka bir deyişle, modelin "bu kişi hasta" ifadesi, hasta olma olasılığını temsil eder. Kesinlik, özellikle gereksiz tedavi veya yanlış pozitiflerin ortaya çıkma olasılığının yüksek olduğu psikolojik strese yol açabilecek tıbbi prosedürler gibi alanlarda çok önemlidir. Modelin gerçek pozitif tahminlerinin tüm pozitif tahminlere oranı kesinlik olarak bilinir. Bu oran, Denklem 2.11'de gösterildiği gibi hesaplanmaktadır.

$$Kesinlik = \frac{DP}{DP + YP} \tag{2.11}$$

Kesinlik parametresi 1'e yaklaştıkça modelin pozitif tahminleri daha doğru hale gelir. Örneğin, PKOS için düşük tanısal hassasiyete sahip bir sistemin birçok sağlıklı insanı "hasta" olarak sınıflandırması ve belki de gereksiz prosedürlere yol açması muhtemeldir. Bu bakış açısına göre doğruluk, teknolojik hususların yanı sıra klinik duyarlılık ve etik meselesidir (Sokolova vd. 2009; Powers, 2020).

### 2.5.5. F1 puanı

Bir sınıflandırma modelinin performansı, hassasiyet ve geri çağırmanın harmonik ortalamasını hesaplayan F1 puanı kullanılarak değerlendirilir. F1 puanı, hassasiyet ve geri çağırmayı tek bir çatı altında toplayarak modelin toplam performansının daha dengeli bir resmini verir çünkü bu iki kriter genellikle birlikte çalışır. YP veya YN tahminlerin farklı

maliyetlere neden olabileceği klinik uygulamalarda veya dengesiz veri setlerinde, F1 puanı özellikle yararlıdır. F1 puanı, Denklem 2.12'de gösterildiği gibi hesaplanmaktadır.

$$F1\ puanı = 2\ *\ \frac{Kesinlik * Geri\ çağırma}{Kesinlik + Geri\ çağırma} \tag{2.12}$$

Bu algoritma, hassasiyet ve geri çağırmanın önemli ölçüde farklılık göstermesi durumunda F1 puanının cezalandırılmasına ve dengeleyici bir ölçü vermesine olanak tanır. Örneğin, bir PKOS tespit sistemi yüksek hassasiyete ancak düşük geri çağırmaya sahipse hastaları kaçırıyor veya sağlıklı kişilere yanlış tanı koyuyor olabilir ya da tam tersi olabilir. F1 puanı bu durumlarda bu uyumsuzluğu kabul eder ve daha kapsamlı bir değerlendirme sunar. Bu nedenle, sağlık gibi yüksek hassasiyetli alanlarda, F1 puanı sınıflandırma modellerinin doğruluğunu değerlendirmek için çok önemli bir araçtır (Sokolova vd. 2009; Powers, 2020).

## 2.7. Polikistik Over Sendromu Tanısı

PKOS olarak bilinen karmaşık ve çeşitli endokrin durum, kadınların üreme sağlığını önemli ölçüde etkilemektedir. Hastalığın birçok çeşidinin olması ve bazı semptomlarının diğer hastalıklarla benzerlik göstermesi, klinik gözlemlere, laboratuvar testlerine ve görüntüleme tekniklerine dayansa bile net bir teşhis koymayı zorlaştırmaktadır. Sonuç olarak, son zamanlarda MÖ ve YZ gibi en yeni teknolojileri PKOS tanı prosedürlerine dahil etmek için önemli girişimlerde bulunulmuştur. Hem geleneksel tanı standartları hem de bu prosedürde kullanılabilecek teknoloji yöntemleri aşağıda açıklanmaktadır.

### 2.7.1. Polikistik over sendromu

PKOS, üreme çağındaki kadınlarda en yaygın endokrinopatidir ve tahmini prevalansı %4 ila %20 arasında değişmektedir; 2019'da dünya genelinde 66 milyondan fazla kadını etkilemiştir (Barrera 2023). Kamu sağlığındaki yükü büyüktür ve yalnızca 2020'de, sadece Amerika Birleşik Devletleri'nde, PKOS ile ilişkili semptomları yönetmek için neredeyse sekiz milyar dolar harcanmıştır (Riestenberg 2022). Hastanın PKOS'a sahip olup olmadığını belirlemek, mevcut tüm sınıflandırmalarda kullanılan kriterlerin aynı olması ve yalnızca kriterlerin kombinasyonunun tartışmalı olması nedeniyle basit bir süreç olmalıdır. PKOS teşhisinin zorlu yanlarından birisi birkaç hastalık ile karışmasıdır. 2023 yılına kadar PKOS teşhisi, Rotterdam kriterleri/Uluslararası PKOS kriterleri temel alınarak klinik kriterlere dayanmaktaydı. Şu anki adı Modifiye Rotterdam Kriterleri olarak adlandırılmaktadır. En yaygın kabul gören teşhis kriteri olarak görülmekteydi. "International Evidence-based Guideline for the Assessment and Management of Polycystic Ovary Syndrome" (Teede vd. 2023) çalışmasının 2023 yılında yayınlanmasıyla AMH kriteri de PKOS teşhis kriterlerine eklenmiş oldu.

### 2.7.2. Karışan diğer hastalıklar

PKOS, bir dışlama tanısıdır ve ayırıcı tanıda karışabilecek diğer hastalıklar dışlandıktan sonra tanı konulabilir. PKOS, hiperandrojenizm, idiopatik hirsutizm ve NC-CAH gibi diğer durumlarla karışabilir. Bu durumları ayırt etmek genellikle bir sağlık profesyoneli tarafından detaylı bir değerlendirme gerektirmektedir. Hiperandrojenizm, vücutta normalden daha fazla erkek cinsiyet hormonu olan androjenlerin bulunması

durumunu ifade eder. PKOS, hiperandrojenizme neden olan bir durumdur, ancak hiperandrojenizm tek başına PKOS tanısını koymak için yeterli değildir. İdiopatik hirsutizm, hormon seviyelerinde belirgin bir değişiklik olmadan anormal derecede fazla vücut kıllanması durumunu ifade eder. PKOS, tipik olarak hirsutizme neden olan hormonal değişikliklerle ilişkilidir. Ancak, hirsutizm yaşayan bir bireyde sadece PKOS olmayabilir, bu nedenle diğer durumlar da değerlendirilmelidir. NC-CAH, adrenal bezlerin normalden fazla androjen ürettiği genetik bir durumdur. Bu durum, hirsutizm, adet düzensizlikleri ve diğer hormonal belirtilere neden olabilir.

YZ'da gerçekleşen son on yıldaki devrim niteliğindeki ilerlemeler, PKOS'u teşhis etme ve yönetme yeteneğimizi hızla geliştirmeyi vaat etmektedir. Bu, YZ'nın PKOS gibi heterojen bozuklukların teşhisi için ideal bir yardımcı yapma yeteneğinden kaynaklanmaktadır, çünkü YZ, farklı veri miktarlarını işleme yeteneğine sahiptir. Birkaç çalışma, aile genetik geçmişi, biyobelirteçler ve demografik bilgiler gibi farklı verileri birleştirip PKOS teşhisi için birleşik bir algoritma oluşturabilen ve teşhis tahminleri yapabilen makine öğrenimi modellerinin yeteneğini araştırmıştır. Bu çalışmaların bazı eksiklikleri, küçük ölçekli olmaları, ilgili karşılaştırıcıların eksikliği, çeşitli teşhis kriterlerinin kullanımı ve raporlama yapılarındaki heterojenlik gibi durumları içermektedir (Barrera vd. 2023). Bu nedenle, PKOS teşhisindeki YZ'nın/MÖ'nin gerçek bilgi boşlukları ve kapsamı belirsiz kalmaktadır (Silva 2022).

Günümüzde sağlık alanında YZ uygulamalarının kullanımı giderek artmaktadır ve bu teknolojilerin sunduğu potansiyel, bir dizi sağlık sorununun daha hızlı ve etkili bir şekilde çözülmesini sağlamaktadır. Bu bağlamda, önerdiğimiz proje, YZ teknolojilerini kullanarak PKOS'nun daha hızlı, doğru ve erken bir şekilde teşhis edilmesini sağlamaktır. PKOS, kadın üreme sağlığını etkileyen yaygın bir durumdur ve erken teşhis, hastaların daha etkili tedavi stratejilerine erişmelerine olanak tanır.

Geleneksel teşhis yöntemleri zaman alıcı olabilir ve sonuçlarında belirsizlik bulunabilir. Bu bağlamda, YZ destekli tanı yöntemleri, büyük veri analizi ve makine öğrenimi algoritmaları kullanarak hastalığın belirtilerini daha hassas bir şekilde değerlendirmeyi ve teşhis sürecini iyileştirmeyi amaçlar. Projenin önemi, PKOS tanısında geçen süreyi kısaltarak hastaların yaşam kalitesini artırmak, sağlık sistemine yük getiren uzun süreli tedavi ve yönetim maliyetlerini azaltmak ve toplumsal farkındalığı artırmaktır. PKOS'un kadınların hayatlarını tehdit etmesi, PKOS'un her 5 kadından 1'ini etkilemesi ve PKOS teşhis zorluğu sorunları ile başa çıkmak adına, bu projenin önemi bir kat daha artmaktadır. Ayrıca, bu projenin bilimsel açıdan değeri, büyük veri analizi ve YZ teknolojilerinin kadın sağlığı alanındaki uygulamalarını genişleterek, bilim dünyasına yeni bilgiler sunma potansiyelinde yatar. Proje başarıyla uygulandığında, yüksek doğruluk oranlarına sahip tanı sonuçları elde edilmesi beklenmekte ve bu da hastaların güvenilir ve etkili tedaviye daha erken ulaşmalarını sağlayacaktır. Bunun yanı sıra, projenin ekonomik katkıları da göz önüne alınmalıdır. YZ destekli teşhis süreci, geleneksel yöntemlere göre daha hızlı ve verimli olacak, bu da sağlık hizmetlerindeki maliyetleri düşürebilecek ve kaynakların daha etkin kullanılmasına olanak tanıyabilecektir. Bu, toplumsal anlamda daha geniş bir kesimin sağlık hizmetlerine erişimini artırabilir ve kadın üreme sağlığına yönelik bilinci artırabilir. Ayrıca, projenin bilimsel katkıları da göz ardı edilmemelidir. Proje kapsamında geliştirilen algoritmalar ve elde edilen sonuçlar, bilimsel yayınlarda paylaşılabilir ve bu alandaki araştırmalara önemli bir katkı sağlayabilir. Bu, bilimsel topluluğun bu teknolojiye olan ilgisini

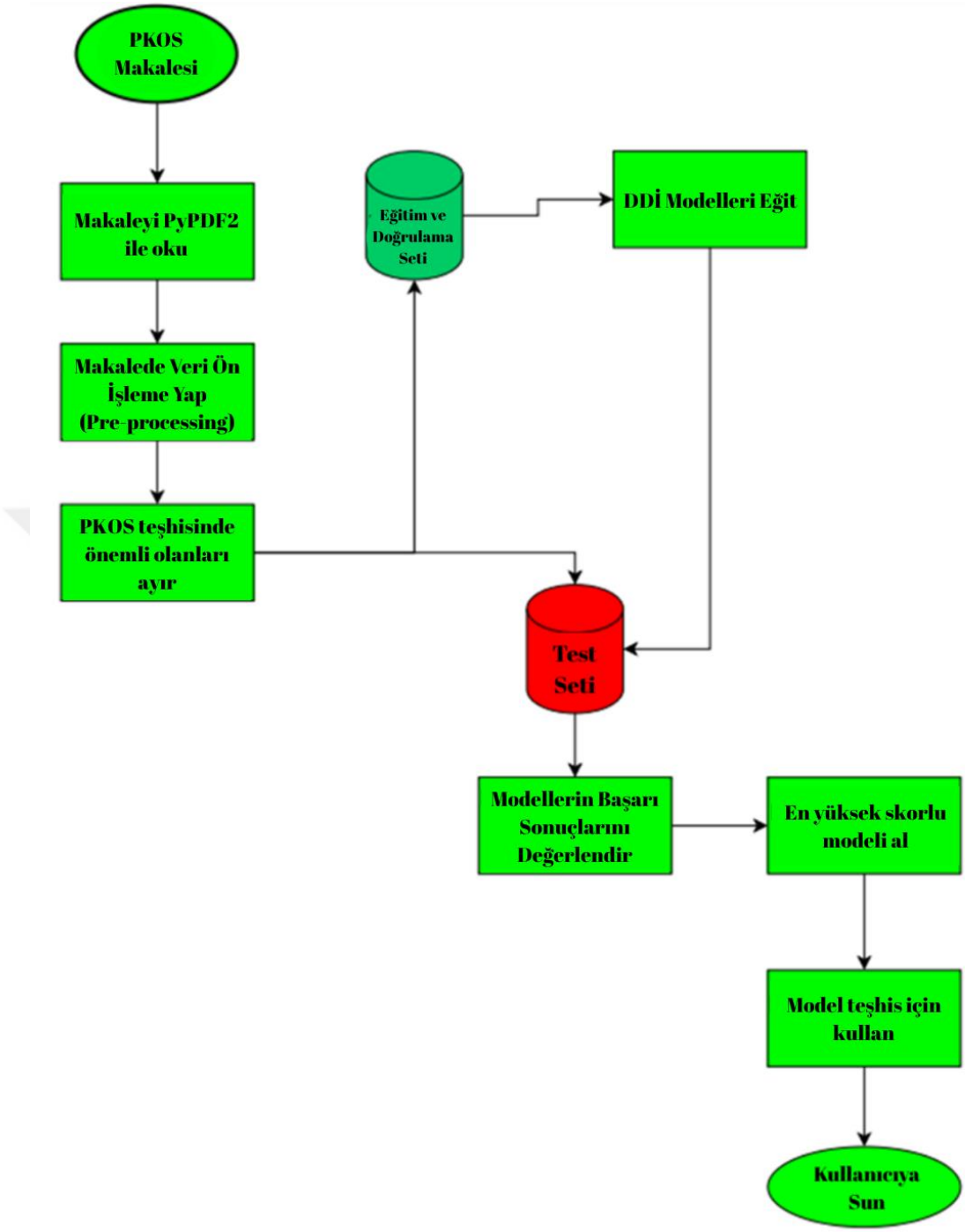
artırabilir ve benzer projelerin geliştirilmesine ilham verebilir.

Bu amaç doğrultusunda projemiz, hastaların erken aşamada doğru bir tanı almasını sağlayarak sağlık sonuçlarını iyileştirmeyi, toplumsal farkındalığı artırmayı ve bilimsel araştırmalara değerli katkılarda bulunmayı hedeflemektedir. Projenin temel sorularına cevap arayarak, YZ PKOS'u %100 teşhis edebilir mi, doktorlara yardımcı olabilecek bir araç var mı, YZ modelleri yüzde kaç başarı oranı ile yardımcı olabilir sorularına pratik çözümler sunmayı amaçlamaktadır.

# 3. MATERYAL VE METOT

Çalışmaya gerekli verilerin toplanması ile başlanılmıştır. Verilerin güncel olabilmesi adına ASRM Tıp Kongresi'ndeki (Teede 2023) tıbbi makalelerin olduğu sitelerden PKOS teşhisi konusundaki makaleler toplanılmıştır. Bu çekilen makaleler veri ön işleme yapılarak temizlenmiştir. Bu işlem, verideki önemli kelimelerin seçilmesini sağlamaktadır. Örneğin; PKOS için hormonlar önemlidir ve bu bilgiler ön işleme ile elde tutulmaktadır. PKOS teşhisinde önemli olan kelimeler ayrıldıktan sonra Eğitim, Doğrulama ve Test set olmak üzere üç ayrı veri kümesi oluşturulmaktadır. Eğitim set veri kümesi, makine öğrenimi modelinin öğrenme sürecinde kullanılan verileri içermektedir. Test set veri kümesi, eğitim süreci tamamlandıktan sonra modelin performansını değerlendirmek için ayrılan bağımsız bir veri kümesidir. Modelin çeşitli parametre kombinasyonlarıyla nasıl çalıştığını test ederek doğrulama seti, en uygun modelin seçilmesine yardımcı olmaktadır. Bu üç setin ayrılması, modelin gerçek dünya verilerine daha iyi uyum sağlamasını ve aşırı öğrenme gibi sorunların önlenmesini sağlar. Eğitim seti, modelin öğrenme sürecinde kullanılmakta; doğrulama seti, modelin eğitim sırasında hiperparametrelerini ayarlar ve aşırı öğrenmeyi önlemekte ve test seti, modelin gerçek dünya verilerine ne kadar iyi uyarlandığını bağımsız olarak değerlendirmektedir. Ayrıca, modelin doğruluğunu ölçmek ve genelleme yeteneğini değerlendirmek için kullanılmaktadır.

Eğitim seti DDİ modelleri ile eğitilerek doğrulama ve test setleriyle karşılaştırılmaktadır. Ardından, eğitilmiş YZ modelleri başarı sonuçlarını ekrana getirmektedir. En yüksek skorlu model seçilir ve bu model PKOS tanısı konulması için kullanılır. Yazım olarak anlatılan işleyiş, Şekil 3.1'de blok diyagramı olarak gösterilmiştir.




**Şekil 3.1.** Çalışmanın blok diyagramı

### 3.1. Programlama Dili ve Entegre Geliştirme Ortamları

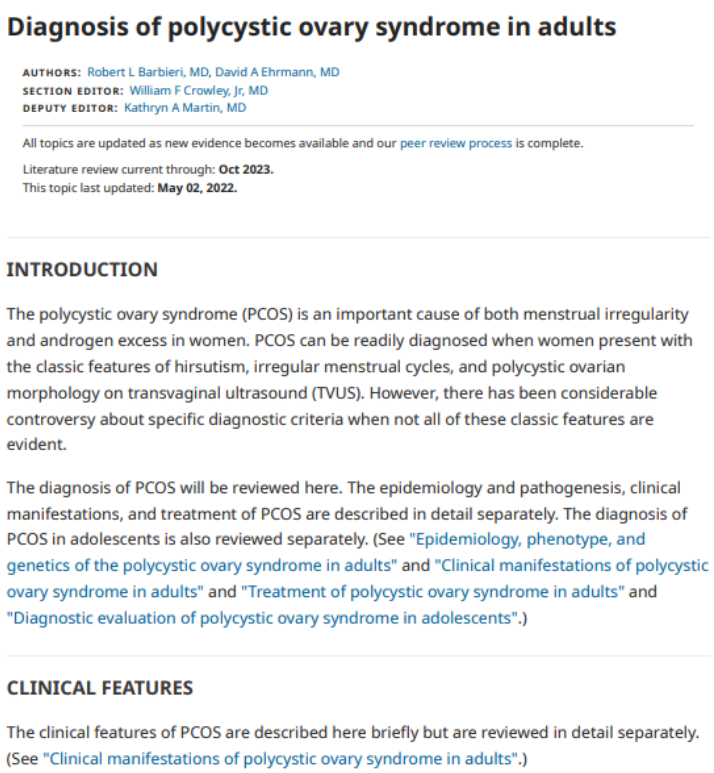
PKOS teşhisine yardımcı olacak YZ modelinin geliştirilmesi için Python 3.9.7 programlama dili kullanılmıştır. Python, geniş kütüphane desteği ve veri bilimi alanındaki güçlü araçları sayesinde proje için tercih edilmiştir (Rossum 2009). Çalışmada geliştirme ortamı olarak, hafif ve modüler yapısı, geniş uzantı desteği ve hata ayıklama özellikleri ile Python projelerinin yönetimini kolaylaştıran Visual Studio Code kullanılmıştır. (Microsoft 2025).

Model eğitimi Visual Studio Code ortamında gerçekleştirilmiş olup eğitilen model ise Trae platformu üzerinde entegre edilerek test edilmiştir. Trae, YZ destekli bir yazılım geliştirme platformudur. Gerçek zamanlı kod tamamlama, YZ ile sohbet, proje dosyalarını otomatik oluşturma ve kod hatalarını giderme gibi özellikler sunarak geliştiricilere üretken bir geliştirme ortamı sağlamaktadır (Trae 2025).

Ayrıca, model eğitimi sürecinde GPU hızlandırma gerektiren işlemler için TRUBA süper bilgisayar altyapısı kullanılmıştır. TRUBA, TÜBİTAK ULAKBİM tarafından sağlanan YPH sistemidir (TÜBİTAK 2025). TRUBA'nın yüksek işlem gücü ve paralel hesaplama özellikleri, büyük veri kümelerinin işlenmesi ve derin öğrenme modellerinin eğitimi için kritik öneme sahiptir.

### 3.2. Veri Seti

Çalışmada PKOS teşhisi için kullanılan veri seti; "International Evidence-Based Guideline for the Assessment and Management of Polycystic Ovary Syndrome" (Teede vd. 2023) başlıklı makaleden elde edilmiştir. Bu kaynak, PKOS konusundaki uluslararası, kanıta dayalı yönergeleri içermesi bakımından öne çıkmaktadır. Güncel ve uzmanlar tarafından değerlendirilen bu kaynak, proje için güvenilir veri setini sağlayarak PKOS teşhisi, değerlendirmesi ve yönetimi konularındaki en son bilgilere erişim imkanı sunmaktadır. Bu veri seti, projenin başarılı bir şekilde ilerlemesi ve YZ modelinin güvenilirliğinin artırılması açısından önemli bir temel oluşturacaktır. Şekil 3.2’de, kullanılan makalenin bir ekran görüntüsü yer almaktadır.



Diagnosis of polycystic ovary syndrome in adults

AUTHORS: Robert L Barbieri, MD, David A Ehrmann, MD
SECTION EDITOR: William F Crowley, Jr, MD
DEPUTY EDITOR: Kathryn A Martin, MD

All topics are updated as new evidence becomes available and our peer review process is complete.

Literature review current through: Oct 2023.
This topic last updated: May 02, 2022.

INTRODUCTION

The polycystic ovary syndrome (PCOS) is an important cause of both menstrual irregularity and androgen excess in women. PCOS can be readily diagnosed when women present with the classic features of hirsutism, irregular menstrual cycles, and polycystic ovarian morphology on transvaginal ultrasound (TVUS). However, there has been considerable controversy about specific diagnostic criteria when not all of these classic features are evident.

The diagnosis of PCOS will be reviewed here. The epidemiology and pathogenesis, clinical manifestations, and treatment of PCOS are described in detail separately. The diagnosis of PCOS in adolescents is also reviewed separately. (See "Epidemiology, phenotype, and genetics of the polycystic ovary syndrome in adults" and "Clinical manifestations of polycystic ovary syndrome in adults" and "Treatment of polycystic ovary syndrome in adults" and "Diagnostic evaluation of polycystic ovary syndrome in adolescents".)

CLINICAL FEATURES

The clinical features of PCOS are described here briefly but are reviewed in detail separately. (See "Clinical manifestations of polycystic ovary syndrome in adults".)

**Şekil 3.2.** Veri setinden ekran görüntüsü

Bununla birlikte, veri setine Smet vd. (2018) tarafından yazılan "Rotterdam Criteria, the End" başlıklı bir makale de dahil edilmiştir. Bu makale, PKOS tanısı için sıklıkla kullanılan ve literatürde sıklıkla referans gösterilen Rotterdam kriterlerini içermektedir. Bu makale, 2003 yılında oluşturulan Rotterdam kriterlerinin klinik geçerliliğini incelemektedir ve tanı ölçütlerinin zaman içinde nasıl değiştiğini incelemektedir. Over morfolojisi, antral folikül sayısı ve hiperandrojenizm gibi temel tanı bileşenleri üzerinde durması nedeniyle, veri setinin tanısal çeşitliliğini ve klinik gerçekçiliğini artırmaktadır. Bu makaleye ait ekran görüntüsüne Şekil 3.3'te yer verilmiştir.

Editorial

Rotterdam criteria, the end

Maria-Elisabeth Smet, MD[1] and Andrew McLennan, BSc, MBBS (Hons), FRANZCOG, FRCOG, CODU[1,2]

[1]Sydney Ultrasound for Women, Level 1, 56 Neridah Street, Chatswood, New South Wales 2067, Australia
[2]Discipline of Obstetrics Gynaecology and Neonatology, Faculty of Medicine, University of Sydney, Sydney, New South Wales 2067, Australia

Rotterdam criteria

According to the Rotterdam consensus,[1] polycystic ovarian syndrome (PCOS) is defined by the presence of two of three of the following criteria: oligo-anovulation, hyperandrogenism and polycystic ovaries (≥ 12 follicles measuring 2-9 mm in diameter and/or an ovarian volume > 10 mL in at least one ovary). The sonographic criteria were based on a study published in 2003 by Jonard *et al.*,[2] where patients with PCOS were found to have significantly more follicles in the 2-5 mm range than a control group comprised of women with tubal or male factor infertility. The decision to set the cut-off at 12 follicles measuring between 2 and 9 mm resulted from a compromise between specificity (99%) and sensitivity (75%), noting that narrowing the range to between 2 and 5 mm did not improve the diagnostic power. A suggestion was made to repeat the assessment in a subsequent cycle if the ovary was enlarged and its antral follicle count obscured by a dominant follicle (>10 mm) or a corpus luteum.[3] There is no consensus on how to classify ovaries with a high follicle count when using oral contraception or other exogenous hormones or when there is evidence of a dominant follicle or corpus luteum in successive cycles.

Is the Rotterdam consensus still appropriate?

Over the 15 years since the Rotterdam consensus, increasing use of transvaginal assessment and technological improvements in ultrasound resolution have resulted in 20%–30% of regularly cycling, normo-ovulatory women satisfying the Rotterdam criteria for polycystic ovarian morphology[4] and even more in a younger demographic.[5] The sonographic assessment has, over

management and stigmatisation. The same is true for women with regular, ovulatory cycles attending ultrasound for reasons other than fertility assessment, who happen to have 15 visible antral follicles and a dominant follicle or corpus luteum. Being wrongfully labelled 'polycystic' may also adversely affect self-esteem and influence women's choices around their diet and use of contraception.

So, is there a better way?

Over the past few years, the sonographic criteria embedded in the Rotterdam Consensus have been rightly challenged. In 2011, Dewailly[6] studied 240 consecutive patients with mixed symptoms and recommended an ovary not be considered to have polycystic ovarian morphology until ≥ 19 follicles were noted, whilst Lujan *et al.*[7] in 2013 studying 98 women with NIH classified PCOS and 70 normo-ovulatory volunteers recommended a threshold at ≥ 26 follicles. The difference in recommended follicle numbers is explained by Dewailly excluding clinically normal patients with high follicle numbers from the control group. A meta-analysis performed by the Androgen Excess and Polycystic Ovary Syndrome Society in 2014[8] found that the median follicle number per ovary (FNPO) in women of reproductive age is between 13 and 16 and strongly advocated increasing the threshold for polycystic ovarian morphology to ≥ 25 follicles. Use of ≥ 25 follicles is also supported by the International Society of Ultrasound in Obstetrics and Gynaecology Consensus Group.[9,10]

Labelled or numbered?

Strictly, use of the term 'polycystic' in most contexts is incor-

**Şekil 3.3.** "Rotterdam Criteria, the End" başlıklı makaleden alınan ekran görüntüsü

Bu iki temel kaynağın kombinasyonu, modelin hem semptomatik açıklamalara hem de klinik kriterlere dayalı ifadeleri öğrenmesine izin verdi. Bu nedenle, yeni sınıflandırma sisteminin dil yapısı hakkında daha fazla bilgi edinmesi ve gerçek klinik durumlara daha yakın sonuçlar üretmesi sağlanmıştır.

Kapsamını ve genelleme kabiliyetini genişletmek için PubMed veri tabanından PKOS başlıklı 100 bilimsel yayın daha veri setine eklenmiştir. Bu makaleler, birçok klinik senaryo, tedavi yöntemi, patofizyolojik açıklamalar ve ayırıcı tanı tartışmalarını içererek makine öğrenimi modellerinin tıbbi dili kapsamlı bir şekilde kavramasına yardımcı olmaktadır. Bu kapsamlı materyalin amacı, PKOS ile karıştırılabilecek diğer endokrin hastalıkların hem teşhisinde hem de ayırt edilmesinde YZ tabanlı sınıflandırıcıların doğruluğunu artırmaktı. Projenin literatüre dayalı YZ sistemi, PubMed'den alınan bu 100 makale ile birlikte daha güvenilir hale gelmiştir.

### 3.3. Ön İşleme

Bu çalışma, YZ modelini eğitmek ve test etmek için kullanılacak metin tabanlı veri kümeleri için çok aşamalı bir hazırlık prosedürünü içeriyordu. Bu prosedür sırasında, açık erişimli internet kaynaklarından toplanan makaleler ve PDF belgelerinden çıkarılan içerikler hem yöntemsel olarak temizlenmiş hem de düzenlenmiştir.

#### 3.3.1. PDF dokümanlarından veri elde etme

PyPDF2 kütüphanesi kullanılarak, PDF formatındaki birincil kaynak olan "International Evidence-Based Guideline for the Assessment and Management of Polycystic Ovary Syndrome" sayfa sayfa okunmuş ve düz metin formatına

dönüştürülmüştür. Metin satır satır incelendikten sonra başlık ve içerik bölümleri ayrı ayrı tanımlanmış ve etiketlenmiştir. Bu şekilde metin JSON formatına dönüştürüldükten sonra başlık tabanlı bir yapı kullanılarak düzenlenmiştir.

### 3.3.2. Metin temizleme ve standartlaştırma

Veri hazırlama sürecinin bu noktasında, nihai metinlerin okunabilirliğini artırmak ve YZ modeli için daha anlamlı olmalarını sağlamak için kapsamlı bir temizleme prosedürü kullanılmıştır.

İlk olarak, PDF kaynaklarında sıklıkla görülen yazım hataları ve bozuk karakterler (\u0000 gibi) düzeltilmiştir. Daha sonra yazar adları, sayfa numaraları, internet adresleri, tarih ve saat bilgileri ve e-posta adresleri de dahil olmak üzere model için anlam ifade etmeyen materyaller çıkarılmıştır. Bu prosedürün ardından, fix_text ve düzenli ifadeler gibi düzeltici araçlar kullanılarak materyal daha tutarlı hale getirildi ve gereksiz metinler elenmiştir.

Ayrıca, "REFERANS" ve "SONUÇ VE ÖNERİLER" gibi tanı ile ilgisi olmayan veya YZ eğitimi için zayıf bilgi yoğunluğuna sahip pasajlar metinden çıkarılmıştır. Sonuç olarak, model öncelikle önemli klinik verilere ve teşhisle ilgili içeriğe odaklanabildi. Bu, metinlerin içerik anlamını iyileştirmiş ve modelin öğrenme süreci için daha etkili bir veri yapısı üretilmiştir.

### 3.3.3. PubMed kaynağından makale toplama

PKOS ile ilgili bilimsel literatürü genişletmek için PubMed veritabanından alınan yaklaşık 100 makalenin başlık ve özet bilgilerini içeren bir ".csv" dosyası kullanılmıştır. Bu dosyayı ayrıştırmak için Python programlama dili kullanılmış, her bir makalenin başlığı ve özeti tek ve kapsamlı bir metin halinde birleştirilmiştir. Verileri ilgili ve model eğitimine uygun hale getirmek için çeşitli temizleme prosedürleri kullanılmıştır. Burada, metnin referans numaraları elenerek, bozuk HTML karakterleri düzeltilerek ve gereksiz boşluklar yoğunlaştırılmıştır.

Yetersiz yapısal bütünlüğe veya yetersiz bilgi yoğunluğuna sahip kayıtlar veri setinden çıkarılmıştır. Veriler temizlendikten sonra JSON formatında düzenlenmiş ve ayrı bir PubMed ID atanmıştır. Bu prosedür, literatürden toplanan verilerin işlenmesini kolaylaştırmakta ve YZ modeli eğitimi için uygun bir temel sunmaktadır.

### 3.3.4. Veri birleştirme ve nihai temizleme

Çeşitli kaynaklardan toplanan tüm veri kümeleri birleştirilerek genişletilmiş bir metin veri kümesi oluşturulmuştur. text_only_dataset.json adlı bir ana veri kümesi oluşturmak için her veri noktası bir kez daha temizlendi. Sınıflandırma modeli bu veri kümesi doğrudan girdi olarak kullanılarak eğitilmiştir.

Metinlerin gürültüsü bu çok aşamalı hazırlık prosedürü ile azaltılmış ve aynı zamanda oldukça bilgi yoğun, anlamsal olarak anlamlı bir yapı oluşturulmuştur. Bu da modelin ayırt edici teşhis kabiliyetini ve doğruluğunu artırmıştır. Kapsamlı veri temizliği özellikle tıbbi ortamda modellenen görevler için önemlidir.

### 3.4. Vektörleştirme ve Özellik Çıkartma

Ön işleme aşamalarından sonra, PKOS teşhisi ile ilgili metinlerin analiz edilmesi ve modelin eğitilmesi için metin verilerinin sayısal biçime dönüştürülmesi gerekmektedir. Bu süreçte, metin özelliklerini çıkarmak ve sayısal temsiller elde etmek amacıyla çeşitli DDİ teknikleri uygulanmıştır. Temizlenmiş metin verileri önce JSON formatında yapılandırılmış, ardından klasik ve bağlama duyarlı vektörleştirme yöntemleriyle sayısallaştırılmıştır. Bu süreçte hem klasik gömme yöntemlerinin hem de bağlama duyarlı (contextual) modellerin değerlendirilmesi yapılmıştır.

TF-IDF yöntemi, metinlerdeki kritik kelimeleri belirlemek için kullanılmıştır. Bu yöntem, bir kelimenin belirli bir metinde ne kadar sık kullanıldığını ve genel olarak veri kümesinde ne kadar nadir bulunduğunu dikkate alarak kelimenin önemini ölçmektedir. Bu yaklaşım, popüler ancak düşük bilgi içeren terimleri (stop-words) otomatik olarak filtreleyerek yalnızca metne özgü önemli kelimeleri vurgulamaktadır.

Word2Vec modeli, kelimeleri vektörlere çevirmek için iki farklı yöntem kullanır: Skip-Gram ve CBOW. Skip-Gram yöntemi, bir kelimenin etrafındaki bağlamı tahmin ederken, CBOW yöntemi belirli bir bağlamdaki eksik kelimeyi tahmin etmektedir. Eğitim için metinler kelime parçacıklarına ayrıldı ve kelimeler arası ilişkileri öğrenen vektörler oluşturulmuştur.

FastText modeli de Word2Vec'e benzer bir şekilde çalışmakta, ancak kelimeleri karakter n-gramlarına bölerek öğrenmeyi kolaylaştırmaktadır. Bu nedenle, model bilinmeyen veya nadir kelimeleri de tahmin edebilmektedir.

Kelime vektörleri oluşturmak için GloVe modeli büyük metin kümelerinden kelimeler arası ilişkileri öğrenmektedir. Bu yaklaşım, kelimelerin bir arada bulunma sıklığını göz önünde bulundurarak anlam temsilleri oluşturmaktadır. Bu araştırma için önceden eğitilmiş GloVe vektörleri yüklenmiş ve veriler her bir kelimeye karşılık gelen vektörlerle karşılaştırılmıştır.

Bu modellerin sonuçlarını değerlendirmek için Kosinüs Benzerliği kullanılmıştır. Çizelge 3.1, elde edilen gömme vektörleri arasındaki benzerlikleri göstermektedir. Karakter düzeyinde bilgi işleyen FastText ve bağlam temelli öğrenme yapan Word2Vec “Skip-Gram” yöntemleri, over ve hormon kelimeleri arasındaki ilişkiyi daha iyi modellemiştir. Bağlamsal bilgi içermeyen GloVe yöntemi, bununla birlikte daha düşük bir benzerlik oranı sağladı. Bu bulgular, bağlamsal anlamın özellikle tıbbi metinler için ne kadar önemli olduğunu göstermektedir ve bağlam duyarlılığı yüksek modellerin analiz için daha iyi sonuçlar verdiğini göstermektedir. Bağlama duyarlı yaklaşımlardan biri olan Cümle Dönüştürücüleri modeli de bu bağlamda değerlendirilmiş ve karşılaştırılmıştır.

**Çizelge 3.1.** Over ve Hormon Kelimeleri Arasında Farklı Gömme Yöntemleriyle Elde Edilen Kosinüs Benzerlik Skorları

| **Gömme Yöntemi** | **Kosinüs Benzerliği** |
|---|---|
| Word2Vec (CBOW) | 0.7005 |
| Word2Vec (Skip-Gram) | 0.5982 |
| FastText | 0.6986 |
| GloVe | 0.3095 |
| Cümle Dönüştürücüleri | 0.4974 |

Bununla birlikte, daha fazla araştırma, klasik modeller tarafından üretilen gömme vektörlerinin 0.97 ila 0.99 arasında değişen ortalama kosinüs benzerlik değerlerine sahip olduğunu ortaya koymaktadır. Bu veriler, nispeten benzer vektörlerin farklı metin eşleşmelerini temsil ettiğini göstermektedir. Bu durum, modelin anlamsal olarak yanlış benzerlikler ürettiğini ve metinler arasındaki anlam farklılıklarını ayırt etmekte zorlandığını göstermektedir.

### 3.4.1. Gelişmiş vektör temsilleri için dönüştürücü tabanlı sınıflandırma modellerinin kullanılması

Geleneksel metin vektörleştirme yaklaşımları, kelimeleri bağlamdan bağımsız olarak sabit vektörler olarak göstermesine rağmen, bazı avantajlar sunmuştur. Bu nedenle, bağlama bağlı anlam farklılıkları ve çok anlamlılığı ortaya çıkarma yeteneği, özellikle tıbbi metinlerde sıklıkla görülmektedir. PKOS ile ilgili yazılarda, kelimelerin yalnızca bireysel anlamları değil, cümleler ve paragraflar içindeki bağlamsal ilişkileri de önemlidir. Sonuç olarak, bağlam duyarlılığı yüksek ve anlamsal ilişkileri daha iyi modelleyen yöntemler ihtiyaç duyulmuştur.

Bu modellere ek olarak, Cümle Dönüştürücüleri (MiniLM-L6-v2) modeli kullanılarak cümle düzeyinde bağlama duyarlı vektör temsilleri elde edilmiş ve FAISS paketi kullanılarak bu vektörler üzerinde hızlı benzerlik aramaları yapılmıştır. Cümle Dönüştürücüleri modeli, bağlamı dikkate alarak cümle düzeyinde vektörler oluşturduğu için farklı ifadeler arasında düşük benzerlik puanları ürettiğinden metinleri daha etkili bir şekilde ayırt edebilmiştir. Bu sayede geleneksel yaklaşımlarda ortaya çıkan sahte benzerliklerin önüne geçilmiştir.

**Çizelge 3.2.** Tüm Metinler Arasındaki Ortalama Kosinüs Benzerlik Skorları

| **Gömme Yöntemi** | **Kosinüs Benzerliği** |
|---|---|
| Word2Vec (CBOW) | 0.9971 |
| Word2Vec (Skip-Gram) | 0.9796 |
| FastText | 0.9981 |
| GloVe | 0.9333 |
| Cümle Dönüştürücüleri | 0.6736 |

Cümle Dönüştürücüleri sınıflandırma problemleri için çok kullanışlıdır çünkü minimum benzerliklerine rağmen metinler arasındaki anlamsal farklılıkları doğru bir şekilde yakalarlar. Aslında, az miktarda etiketli veriyle bile, BioBERT ve SciBERT modelleri bu bağlam duyarlılığı sayesinde sınıflandırma görevlerini mükemmel doğrulukla tamamlayabilmiştir. Ayrıca, bu algoritmalar PKOS teşhislerini daha etkili bir şekilde sınıflandırabilmiş ve kelime anlamlarını bağlam içinde yorumlayabilmiştir.

Çalışmada BioBERT ve SciBERT modelleri dönüştürücü mimarisine dayalı olarak kullanılmıştır. BioBERT, büyük ölçekli biyomedikal metinler üzerinde önceden eğitilmiş bir model olarak tıbbi verileri analiz etmiştir, ancak SciBERT, bilimsel yayınlar üzerine eğitilmiş olması sayesinde araştırma makalelerindeki verilerin bağlamsal analizini başarıyla gerçekleştirilmiştir (Beltagy ve Cohan 2019; Lee vd. 2020). Her iki model de tıbbi alanda derinlemesine bilgi edinmiş ve PKOS gibi karmaşık medikal kavramları anlamada daha iyi sonuçlar sunmuştur. Bu modellerin tercih edilme nedenleri, bağlam duyarlılığı, semantik farkındalık ve uçtan uca öğrenme ve transfer öğrenme avantajları olmuştur. BERT tabanlı modeller, tıbbi metinlerdeki anlam kaymalarını doğru bir şekilde modellemiş ve her cümledeki kelimelerin bağlamsal anlamlarını analiz etmiştir. Bu modeller, daha önce kapsamlı biyolojik veriler kullanılarak eğitildikleri için az miktarda etiketli veriyle bile iyi performans göstermiştir.

Gömme ve sınıflandırma adımlarını baştan sona düzene sokarak, standart yaklaşımlara kıyasla genel doğruluğu artırmıştır. Ayrıca tüm cümleleri dikkate alarak teşhis cümlelerini daha doğru bir şekilde sınıflandırdı. Böylece, klasik yaklaşımların sabit vektör kısıtlamalarını aşarak bağlamı doğru analiz etmiş ve daha başarılı bir sınıflandırma süreci sunmuştur. Tüm karşılaştırmalar, PKOS teşhisine ilişkin metinlerin değerlendirilmesi söz konusu olduğunda, bağlama duyarlı modellerin etkinlik ve güvenilirlik açısından klasik modellerden daha iyi performans gösterdiğini ortaya koymaktadır.

### 3.5. Model Eğitimi

Hazırlık ve veri temizleme prosedürlerinin tamamlanmasının ardından, PKOS ile ilgili belgeler derin öğrenme tabanlı dil modeli mimarileri kullanılarak otomatik olarak

sınıflandırılmıştır. Literatürde yaygın olarak tercih edilen ve tıbbi ortamda iyi performans gösteren iki farklı Dönüştürücü tabanlı model BioBERT ve SciBERT bu amaç için seçilmiştir.

### 3.5.1. Etiketleme süreci ve otomatik kategorizasyon

Her belgenin başlığı (section_key) ve içeriği (text), model eğitimi için kullanılan text_only_dataset.json dosyasında temsil etmek için kullanılır. Özel etiketleme fonksiyonu auto_label_text() yardımıyla bu veriler aşağıda listelenen üç ana sınıfa ayrılır:

- PKOS'un Tanısı ve Klinik Özellikleri
- PKOS'un Ayırıcı Tanısı
- PKOS'un Epidemiyolojisi, Patogenezi ve Tedavisi

Başlık bilgilerinin yanı sıra, metin içeriğindeki tıbbi ifadelerden türetilen anahtar kelimeler de etiketleme için kullanılmıştır. Örneğin, bir belge "cushing", "prolaktinoma", 'nccah' veya "tiroid disfonksiyonu" gibi terimler içeriyorsa "Ayırıcı Tanı" sınıfına yerleştirilmiştir. Benzer şekilde, "Tedavi/Patogenez" sınıfına "insülin direnci", "metformin", "epidemiyoloji" ve "obezite" gibi terimler verilirken, "Tanı ve Klinik Özellikler" sınıfına "hirsutizm", 'oligomenore' ve "rotterdam kriterleri" gibi klinik tanı belirteçleri verilmiştir.

Veri kümesinin yüksek doğrulukta etiketlenmesini sağlamak için bu kurallar, uzman bilgisini yansıtan kural tabanlı bir ön sınıflandırıcı olarak işlev görmüştür.

### 3.5.2. Model mimarisi ve eğitim süreci

Model eğitimi sürecinde kullanılan mimari yapı Şekil 3.4'te gösterilmektedir. Bu yapı, dönüştürücü tabanlı bir metin sınıflandırma sisteminin tüm bileşenlerini kapsamaktadır.


**Şekil 3.4.** SciBERT/BioBERT tabanlı metin sınıflandırma modelinin mimarisi

Her metin önce 500 karakterlik bölümlere ayrılır, ardından ilgili dönüşüm modeline girilebilecek belirteçlere, dikkat maskelerine ve bölüm kimliklerine dönüştürülür. BioBERT ve SciBERT modelleri, sırasıyla bilimsel ve biyomedikal makalelere özgü önceden eğitilmiş gömülü bilgilere sahip olduğundan, bu dönüştürme işlemini bağlama duyarlı bir şekilde gerçekleştirebilirler. [CLS] belirtecinden türetilen bağlamsal vektör, sınıflandırma işlemi için birincil bilgi kaynağı olarak hizmet eder ve özellikle modelin girdisi olarak sağlanan metni temsil eder. Bu vektör, modelin aşırı uyum sağlamasını durdurmak için bir bırakma katmanı uygulandıktan sonra üç sınıf için doğrusal bir sınıflandırıcı katmanına aktarılır. Bu katmandan çıkan skorlar softmax fonksiyonu aracılığıyla olasılıklara dönüştürülerek nihai sınıf tahmini yapılmaktadır.

### 3.5.3. Eğitim parametreleri ve uygulama detayları

Eğitim prosedürünü yürütmek için Hugging Face Transformers kütüphanesinin Trainer API'si kullanılmıştır. Model 2e-5 sabit öğrenme oranı, 8 sabit yığın boyutu, 0,01 sabit ağırlık azalma oranı ve 8 sabit eğitim periyodu ile eğitilmiştir. Modelin performansı, her eğitim oturumunun sonunda doğrulama verileri üzerinde değerlendirilmiş ve en iyi performansı gösteren model, doğruluk ve kayıp değerleri hesaplanarak otomatik olarak kaydedilmiştir. Ayrıca, anlamsız eğitim adımlarını ortadan kaldırmak için bir erken durdurma mekanizması (EarlyStoppingCallback) kullanılmıştır. Eğitimin ardından modeller farklı bir test veri kümesi üzerinde değerlendirilmiş ve her modele bir

sınıflandırma raporu (classification_report.json) ve karışıklık matrisi (confusion_matrix.png) verilmiştir. Performansı değerlendirmek için doğruluk, sınıf tabanlı F1 puanları ve makro ortalama F1 puanları gibi ölçütler kullanılmıştır.

Bu çalışmada, parametre-etkin ince ayar (PEFT) teknikleri yerine tüm model ağırlıklarının değiştirildiği geleneksel ince ayar (Fine Tuning) yöntemi kullanılmıştır. PEFT teknikleri sadece belirli katmanları değiştirir. Bu yöntem, modelin tüm parametrelerinin eğitim verisine göre optimize edilmesini sağlarken, özellikle orta boyutlu veri kümeleriyle başarılı olmayı mümkün kılmıştır. Eğitim süresi boyunca modelin doğrulama başarısı izlenmiştir ve en iyi ağırlıklar kaydedilmiştir. Bu, aşırı öğrenme riskini azaltmıştır. Bu ince ayar, SciBERT ve BioBERT gibi önceden eğitilmiş büyük modellerin klinik metin sınıflandırması gibi görevlerde başarılı olmuştur. PEFT gibi yaklaşımların, özellikle daha az etiketli veriyle çalışan durumlarda, daha düşük hesaplama maliyetleri ve daha hızlı eğitim avantajlarıyla sistemin verimliliğini artırabileceği öngörülmektedir.

Bulgulara göre, her iki model de mükemmel sınıflandırma doğruluğu göstermiştir. Ancak SciBERT'in bilimsel terminolojiyle ilgili ön eğitim geçmişi, özellikle bağlamsal olarak karmaşık olan "Ayırıcı Tanı" konusunda daha iyi sonuçlar elde etmesine yardımcı olmuştur. Buna karşılık, BioBERT "Teşhis ve Klinik Özellikler" sınıfında daha yüksek bir hatırlama oranı göstermiş ve genel teşhis ve klinik ifade kalıplarını daha iyi genelleştirebilmiştir. Eğitimin ardından, bu modeller gerçek zamanlı bir klinik karar destek sistemi olarak kullanıldı ve Gradio tabanlı bir kullanıcı arayüzüne dahil edildi. Kullanıcı tarafından sağlanan serbest metin klinik notların analizi yoluyla sistem, modelin sınıflandırmasına dayalı olarak eksik verileri belirleyebilir, Gemini'ye dayalı olarak açıklayıcı yanıtlar üretebilir ve gerekirse dinamik tanı sorularını sırayla sorabilir. Bu bütüncül yapı sayesinde sistem yalnızca metin sınıflandırması yapmakla kalmamakta, aynı zamanda ayırıcı tanı uyarıları ve laboratuvar değerlendirmeleri ile desteklenmiş, açıklamalı bir teşhis süreci sunarak klinisyenlere etkin bir karar destek aracı sağlamaktadır.

### 3.6. Uygulama Ara Yüzü ve Değerlendirme Sistemi

Kullanıcı katılımını kolaylaştırmak için Gradio kütüphanesi, bu çalışmada oluşturulan YZ karar destek sistemini bir web ara yüzüne entegre etmek için kullanılmıştır. Doktorlar gibi kullanıcılar, ara yüzü kullanarak hasta hakkında serbest metin klinik notlar girebilir. Sistemin DDİ modülü, sağlanan materyali analiz eder ve PKOS semptomlarının varlığına ilişkin bir ilk belirleme yapar.

Sistem, önceden ayarlanmış bir belirti kontrol listesinden geçer. Literatürde PKOS teşhisi için klinik olarak önemli olduğu gösterilen dört ana semptom bu listenin temelini oluşturmaktadır:

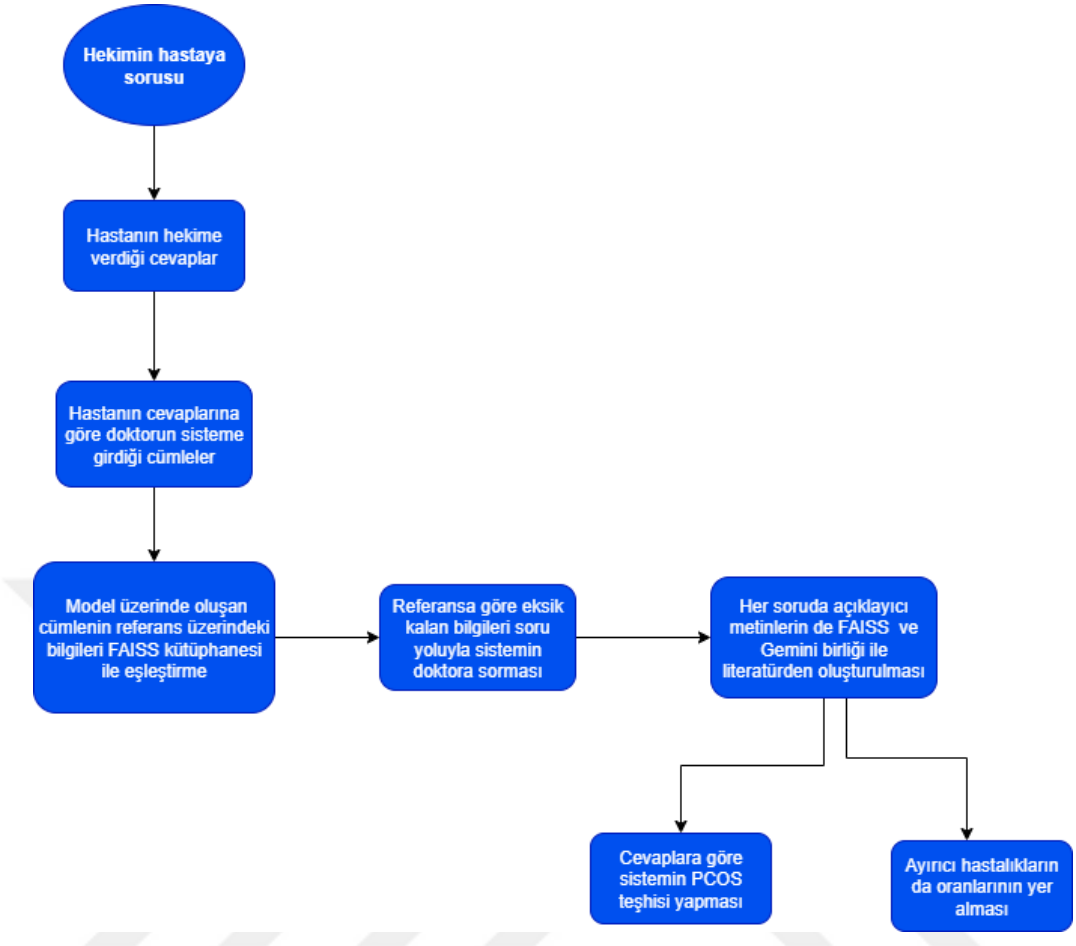
- Oligomenore veya anovulasyon (adet düzensizliği)
- Hirsutizm (aşırı tüylenme)
- Akne veya saç dökülmesi gibi hiperandrojenizm bulguları
- Polikistik over morfolojisi (ultrason bulgusu)

İlk olarak, kullanıcının serbest metni semptomlar hakkında herhangi bir ipucu aramak için incelenir. Bu çalışmayı yapmak için BioBERT veya SciBERT modeli kullanılır. Giriş metnini sınıflandırdıktan sonra model, ifadenin PKOS tanısı, ayırıcı tanı veya tedavi/epidemiyoloji ile ilgisi olup olmadığını tespit eder. Eş zamanlı olarak, sistemin FAISS tabanlı semantik arama modülü, kullanıcının ifadesine en yakın materyali bulmak için ilgili tıbbi makalelerden alınan metin parçacıklarını kullanır.

Kullanıcı metninde önemli bir semptomun yeterince tanımlanmaması durumunda, algoritma eksik olan bilgiyi belirler ve kullanıcıya açık uçlu sorular yöneltir. Şekil 1.1'deki algoritma temel alınarak, makaleden kullanıcıya sorulacak sorular JSON formatında oluşturulmuştur. Sorular EK-1'de gösterilmiştir. Gemini yaklaşımı kullanılarak, her bir semptomun tıbbi bağlamını açıklayan kısa bilimsel özetler de her bir sorgulama ile birlikte üretilir. Dolayısıyla sistem, soru sormanın yanı sıra kullanıcıya açıklamalar da sunmaktadır.

Sistem, kullanıcının yanıtlarını iyi ya da olumsuz olarak yorumlamakta ve iç değerlendirme prosedürüne dahil etmektedir. Sorulara verilen olumlu yanıtların sayısı belirli bir seviyeyi aştığında sistem kullanıcıya "yüksek PKOS olasılığı" çıktısını verir. Eş zamanlı olarak, sistem ayırıcı tanı seçeneklerini değerlendirir. Girilen metne dayanarak Cushing sendromu, hiperprolaktinemi, klasik olmayan konjenital adrenal hiperplazi, tiroid disfonksiyonu ve androjen salgılayan tümörler gibi durumlarla semantik bir benzerlik puanı belirlenir. Böylece tüketici PKOS dışındaki potansiyel hastalıklardan da haberdar edilmektedir.

Bir sohbet arayüzü aracılığıyla, bu alışverişlerin her biri adım adım gerçekleşir. Her kullanıcı girdisi değerlendirilir, model bir yanıt üretir, değerlendirme revize edilir ve gerekirse daha fazla sorgu ortaya çıkarılır. Bu yapı, sadece sabit bir teşhis sistemi olmak yerine, dinamik olarak bilgi toplayan ve açıklamalarla desteklenen bir tıbbi karar destek platformu olarak konumlandırılmıştır. Sistemin çalışması adım adım Şekil 3.5'te gösterilmiştir.




**Şekil 3.5.** Geliştirilen karar destek sisteminin kullanıcı girdisine dayalı teşhis ve açıklama süreci akışı

BioBERT ve SciBERT modelleri bu çalışmada PKOS ile ilgili tıbbi metin sınıflandırması için temel olarak kullanılmıştır. Bu karar, her iki modelin de literatürde güçlü bir geçmişe sahip olması ve bilimsel ve biyolojik kitaplar üzerinde önceden eğitilmiş olması nedeniyle verilmiştir. Lee vd. (2020), PMC ve PubMed gibi biyomedikal literatür kaynaklarından elde edilen büyük veri kümeleri üzerinde eğitildikten sonra, adlandırılmış varlık tanıma, ilişki çıkarma ve metin sınıflandırma gibi görevlerde geleneksel BERT modellerinden %3 ila %10 daha iyi performans gösteren BioBERT'i geliştirmiştir. Buna benzer şekilde, Beltagy vd. (2019) bilimsel dergilerden alınan metinler kullanılarak öğretilen ve bilimsel literatürün dilbilgisi yapısını ve özel kelime dağarcığını öğrenebilen SciBERT'i göstermiştir.

Bu çalışma kapsamında gerçekleştirilen deneysel incelemelerde BioBERT ve SciBERT'in diğer tıbbi modellerden daha etkili olduğu gösterilmiştir. Beş eğitim döneminden sonra her modeli değerlendirmek için aynı veri kümesi ve hiperparametreler kullanılmıştır. Test seti bulgularına göre SciBERT modeli, 0.6728 makro ortalama F1 puanı ve %97.27 doğruluk oranıyla en iyi performansı göstermiştir. Ayrıca, BERTScore (0.89), ROUGE ve BLEU (17.72) dahil olmak üzere açıklama kalitesi kriterlerinde mükemmellik göstermiştir. BioBERT modeli, özellikle genel tanı sınıfında %96.44 doğruluk ve 0.5869 F1 puanı ile tutarlı bir şekilde iyi performans göstermiştir.

PubMedBERT (%96.26 doğruluk, 0.6994 F1), BlueBERT (%96.26 doğruluk, 0.5175 F1) ve ClinicalBERT (%96.44 doğruluk, 0.5128 F1) gibi diğer tıbbi modeller, özellikle bağlama duyarlı ve küçük örneklem sınıflarında daha düşük F1 puanları sergilemiş, ancak diğer sınıflarda iyi doğruluk oranlarına sahip olmuştur.

Bu sonuçlara göre BioBERT ve SciBERT, bağlamsal anlam çıkarımı ve genel sınıflandırma performansı açısından çalışmanın hedeflerine en uygun modellerdir. Bu nedenle modellerin seçiminde hem deneysel sonuçlar hem de literatür desteği dikkate alınmış ve bu iki model karar destek sisteminin ana yapı taşları olarak tasarlanmıştır.

### 3.6.1. Klinik geri bildirime dayalı sistem revizyonu

Hedef kullanıcı olan hekimlerin ihtiyaçlarını karşılamak için geliştirilen sistemin tasarımının uzman görüşü ile değerlendirilmesi gerekiyordu. Sistem tasarımının ilk aşamasında, kullanıcıdan alınan şikayet metnine dayanarak eksik bilgileri otomatik olarak belirleyen ve ardından açık uçlu sorular yönelten bir yapı tasarlanmıştır. Bu sistemin amacı, kullanıcının verdiği cevaplara göre PKOS olasılığını yüzde cinsinden sunmaktır.

Ancak uzman hekim ile yapılan değerlendirme sonucunda, bu sistemin doğrudan PKOS olasılığı sunmasının tek başına yeterli olmayabileceği, zira çoğu hastada zaten PKOS şüphesinin bulunduğu; asıl amacın ise PKOS'nun dışındaki olasılıkların elenmesi veya ayırıcı tanıların göz önüne alınması olması gerektiği ifade edilmiştir.

Sistem, sadece PKOS olasılığını belirlemekle kalmayacak şekilde yeniden tasarlanmıştır. Hastadan alınan yanıtlara göre hem PKOS tanısı yönünde bir değerlendirme yapılmakta hem de ayırıcı tanı (örneğin, tiroid disfonksiyonu, Cushing sendromu, vb.) için kullanıcı bilgilendirilmektedir. Bu nedenle sistem, açık uçlu sorular sorarak teşhis sürecini destekleyen ve seçenekleri adım adım ele alan interaktif bir karar destek platformuna dönüştürülmüştür.

# 4. BULGULAR

Bu çalışmada, bilimsel metinler üzerinde önceden eğitilmiş derin öğrenme tabanlı iki farklı model olan BioBERT ve SciBERT, PKOS teşhisine yardımcı olabilecek bir sınıflandırma sistemi geliştirmek amacıyla karşılaştırmalı olarak eğitilmiştir. Yapılan deneysel çalışmalar sonucunda, modellerin doğruluk ve kayıp değerleri analiz edilerek performansları değerlendirilmiştir. Çizelge 4.1'de her iki modelin 5 eğitim dönemi boyunca elde ettiği kayıp ve doğruluk değerleri gösterilmektedir.

**Çizelge 4.1.** BioBERT ve SciBERT Modellerinin 5 Eğitim Dönemli Fine-Tuning Sonuçları

| ***Model*** | ***Eğitim Dönemi*** | ***Kayıp (Loss)*** | ***Doğruluk*** |
|---|---|---|---|
| *BioBERT* | 1 | 1.3013 | 0.1667 |
| | 2 | 1.2366 | 0.1667 |
| | 3 | 0.9554 | 0.6667 |
| | 4 | 0.8782 | 0.6667 |
| | 5 | 0.7954 | 0.6667 |
| *SciBERT* | 1 | 1.3700 | 0.1667 |
| | 2 | 0.9490 | 0.5000 |
| | 3 | 0.8560 | 0.8333 |
| | 4 | 0.7029 | 0.8333 |
| | 5 | 0.7074 | 0.8333 |

3. eğitim dönemi itibarıyla, SciBERT modeli sabit bir başarı göstererek doğruluk oranını %83.33'e yükseltti (bkz. Çizelge 4.1). Bununla birlikte, BioBERT modeli en yüksek %66.67 doğruluk oranına ulaştı. Her iki modelin başlangıçta düşük performans göstermesinin nedeni, küçük ve düzensiz bir veri kümesidir. Bununla birlikte, model öğrenmesi ilerledikçe, özellikle SciBERT'in daha istikrarlı olduğu açıkça görülmektedir.

Ek olarak, her iki model 8 eğitim dönemi boyunca eğitilerek daha detaylı bir performans analizi yapılmıştır. Bu deneysel süreç sonucunda elde edilen doğruluk ve kayıp değerleri Çizelge 4.2'de sunulmuştur.

**Çizelge 4.2.** BioBERT ve SciBERT Modellerinin 8 Eğitim Dönemli Fine-Tuning Sonuçları

| *Model* | *Eğitim Dönemi* | *Kayıp (Loss)* | *Doğruluk* |
|---|---|---|---|
| *BioBERT* | 1 | 0.9901 | 0.5000 |
| | 2 | 0.9643 | 0.6667 |
| | 3 | 0.8357 | 0.6667 |
| | 4 | 0.7631 | 0.6667 |
| | 5 | 0.7074 | 0.6667 |
| | 6 | 0.6386 | 0.8333 |
| | 7 | 0.5194 | 0.8333 |
| | 8 | 0.5328 | 0.8333 |
| *SciBERT* | 1 | 1.1592 | 0.1667 |
| | 2 | 0.9407 | 0.8333 |
| | 3 | 0.7400 | 0.8333 |
| | 4 | 0.5796 | 0.8333 |
| | 5 | 0.5050 | 0.8333 |
| | 6 | 0.3794 | 1.0000 |
| | 7 | 0.3623 | 0.8333 |
| | 8 | 0.3038 | 1.0000 |

Her iki modelin eğitim süresi uzatıldığında, özellikle 6. eğitim dönemi sonrasında anlamlı bir performans artışı gözlemlenmiştir. BioBERT modeli, 6. eğitim dönemi itibarıyla doğruluk oranını %83.33'e yükseltmiş ve bu oranı eğitim sonuna kadar korumuştur. SciBERT modeli ise 6. ve 8. eğitim dönemlerinde %100 doğruluk oranına ulaşarak dikkat çekici bir başarı göstermiştir.

Ancak bu başarı, özellikle SciBERT tarafında overfitting (aşırı öğrenme) riski barındırmaktadır. 2. eğitim döneminden sonra doğruluk oranında ani bir sıçrama ve 6. eğitim döneminde %100 doğruluk elde edilmesi, modelin sınırlı veri kümesini ezberlemiş olabileceğini göstermektedir. Bu, modelin eğitim verisine yüksek düzeyde uyum sağlayıp genelleme kabiliyetini kaybettiği anlamına gelebilir. Bunun nedeni küçük veri kümesi, eğitim dönemi sayısının fazlalığı ve doğrulama seti kullanılmadan sadece eğitim verisi üzerinde değerlendirme yapılmış olması olabilir. Bu durumlar göz önüne alındığında, modelin performansının daha gerçekçi değerlendirilmesi için doğrulama seti kullanımı

önerilmektedir. Buna ek olarak, modelin eğitim verisine fazla uyum sağlamasını engellemek için dropout ve regularizasyon gibi önleyici yöntemler kullanılabilmektedir.

Bu gelişmeler, modeli daha uzun süre eğitmenin başarı oranını artırabileceğini ve daha güçlü temsil yetenekleri kazandırdığını göstermektedir. Ayrıca, küçük boyutlu ve sınırlı örnek içeren veri kümelerinde bile modelin öğrenmeye devam ettiğini ve genel performansını artırabildiğini ortaya koymaktadır. Çalışma, DDİ tekniklerinin sağlık alanında nasıl kullanılabileceğine odaklanıyor. Özellikle, uzmanlara PKOS gibi karmaşık hastalıkların tanısında yardımcı olabilecek bir ön tarama mekanizması sunmaktadır. Geliştirilen sistemin, hastaların ifadelerinden anlam çıkarma ve ön tanı önerme yeteneği sayesinde klinik karar destek sistemlerine entegre edilme potansiyeli oldukça yüksektir.

### 4.1. UpToDate + Rotterdam Verisi ile Eğitilen Modellerin Karşılaştırmalı Analizi

Tanı kriterlerinin temelini oluşturan Rotterdam makalesi, PKOS ile ilgili UpToDate kaynağına ait metne ek olarak ikinci deneysel çalışmanın veri kümesine dahil edildi. Böylece model hem semptomlara hem de tanı ölçütlerine ilişkin ifadelerle karşılaştırıldı; bu durumun öğrenme süreci üzerindeki etkisi incelenmiştir. Eğitim, beş eğitim dönemi ve sekiz eğitim dönemi olarak iki farklı senaryoda uygulandı ve sonuçları karşılaştırıldı.

Rotterdam verilerinin eklendiği ve beş eğitim dönemi ile sınırlandırılmış eğitim sürecinde, SciBERT modeli ikinci eğitim döneminde %88.89 doğruluk oranına ulaştı (Bkz. Çizelge 4.3). Bu oran, bu araştırmanın en yüksek doğruluk değerlerinden biri olarak öne çıkmaktadır. SciBERT'in bu başarısı, bilimsel metinlerde kullanılan bağlamsal ifadeler ve teknik terimlere olan yatkınlığının bir göstergesidir. Aynı verilerle eğitilen BioBERT modeli, üçüncü eğitim döneminde %77.78 doğrulukla en iyi performansını gösterdi, ancak daha sonra bu seviyenin altında kaldı. Bu, BioBERT'in çok hızlı öğrendiğini ve geliştikçe bunu sürdüremediğini gösteriyor.

**Çizelge 4.3.** Rotterdam Verisinin Eklenmesiyle Eğitilen Modellerin Performansı (Eğitim Dönemi = 5)

| *Model* | *Eğitim Dönemi* | *Kayıp (Loss)* | *Doğruluk* |
|---|---|---|---|
| *BioBERT* | 1 | 1.1325 | 0.2222 |
| | 2 | 1.0247 | 0.6667 |
| | 3 | 0.8246 | 0.7778 |
| | 4 | 0.7507 | 0.6667 |
| | 5 | 0.6444 | 0.5556 |
| *SciBERT* | 1 | 1.2994 | 0.2222 |
| | 2 | 1.0851 | 0.8889 |

**Çizelge 4.3**'ün devamı

| | | | |
|---|---|---|---|
| | 3 | 0.8197 | 0.7778 |
| | 4 | 0.7753 | 0.7778 |
| | 5 | 0.9280 | 0.6667 |

Modellerin genel seyirleri, eğitim süresini sekiz eğitim dönemine çıkarıldığında biraz daha farklılık göstermiştir (Çizelge 4.4). BioBERT, bu süreçte en yüksek %66.67 doğruluk düzeyine ulaştı ve bu oranı birkaç dönem boyunca istikrarlı bir şekilde korudu. Modelin kayıp değerlerinde görülen değişiklikler, veri kümesinin küçük olmasından kaynaklı olarak modelin öğrenme sürecinde kararsızlaştığını göstermektedir. Bununla birlikte, SciBERT, dördüncü eğitim döneminden itibaren %88.89 doğruluğa ulaşarak bu eğitim düzeyini sonuna kadar korumayı başardı. Bu, modelin istikrarlı ve yüksek başarılı olduğunu gösteriyor.

**Çizelge 4.4.** Rotterdam Verisinin Eklenmesiyle Eğitilen Modellerin Performansı (Eğitim Dönemi = 8)

| ***Model*** | ***Eğitim Dönemi*** | ***Kayıp (Loss)*** | ***Doğruluk*** |
|---|---|---|---|
| *BioBERT* | 1 | 0.8823 | 0.5556 |
| | 2 | 1.1922 | 0.5556 |
| | 3 | 0.6977 | 0.5556 |
| | 4 | 0.8211 | 0.5556 |
| | 5 | 1.0518 | 0.5556 |
| | 6 | 0.6329 | 0.6667 |
| | 7 | 1.0163 | 0.6667 |
| | 8 | 0.5233 | 0.6667 |
| *SciBERT* | 1 | 1.5728 | 0.2222 |
| | 2 | 0.9780 | 0.6667 |
| | 3 | 1.1686 | 0.6667 |
| | 4 | 0.6564 | 0.8889 |
| | 5 | 0.8626 | 0.7778 |
| | 6 | 0.5741 | 0.8889 |

**Çizelge 4.4**'ün devamı

| | | | |
|---|---|---|---|
| | 7 | 0.5405 | 0.8889 |
| | 8 | 0.6948 | 0.8889 |

Bu sonuçlar dikkate alındığında, Rotterdam kriterlerine ait bilimsel içeriğin eklenmesinin, özellikle SciBERT modeli üzerinde olumlu bir etkisi olduğu söylenebilir. Modelin, yapılandırılmış tanı bilgileriyle desteklendiğinde daha isabetli sınıflamalar yaptığı gözlenmiştir. BioBERT, tıbbi içeriklere özel olarak eğitilmiş olmasına rağmen, bu tür küçük, dar veri kümelerinde SciBERT'i geride bıraktı.

### 4.2. Genişletilmiş Veri Kümesi ile Model Performansının Değerlendirilmesi

Veri seti, PubMed üzerinden otomatik olarak çekilen 100 PKOS makalesini içeren son deneysel çalışmada kullanılmıştır. Ayrıca UpToDate ve Rotterdam kaynakları da kullanılmıştır. Modelin çeşitliliğine ve miktarına bağlı olarak performansının nasıl değiştiği, bu genişletilmiş veri kümesiyle oluşturulan daha gerçekçi bir öğrenme ortamında incelenmiştir. Eğitim verisinin %70'i eğitim, %10'u doğrulama (validation), %20'si ise test olarak ayrılmıştır. Veri kümesi, önce %80 eğitim+doğrulama ve %20 test olacak şekilde bölünmüş, ardından eğitim kısmının %12.5'i doğrulama verisi olarak ayrılarak %70 eğitim ve %10 doğrulama oranı elde edilmiştir. Her iki model 5 ve 8 eğitim dönemiluk ayrı eğitimler yapılmıştır.

Beş eğitim dönemlik ilk eğitimden sonra modellerin doğruluğu %91.89 olmuştur. Bununla birlikte, eğitim süresi 8 eğitim dönemi çıkarıldığında her iki modelin doğruluğu %97.27'ye yükselmiştir. Bu, modelin daha tutarlı yanıtlar ürettiğini ve daha fazla veriyle bağlamı daha iyi öğrendiğini göstermektedir. Çizelge 4.5'te, 5 eğitim dönemi boyunca eğitilen BioBERT ve SciBERT modellerinin doğruluk ve kayıp değerleri yer almaktadır.

**Çizelge 4.5.** PubMed + UpToDate + Rotterdam Veri Seti ile Eğitilen Modellerin Performans Sonuçları (Eğitim Dönemi = 5)

| ***Model*** | ***Eğitim Dönemi*** | ***Kayıp (Loss)*** | ***Doğruluk*** |
|---|---|---|---|
| *BioBERT* | 1 | 0.9174 | 0.5315 |
| | 2 | 0.7408 | 0.7117 |
| | 3 | 0.6438 | 0.7568 |
| | 4 | 0.4334 | 0.8468 |
| | 5 | 0.2600 | 0.9189 |
| *SciBERT* | 1 | 0.9875 | 0.4505 |

**Çizelge 4.5**'in devamı

| | | | |
|---|---|---|---|
| | 2 | 0.8021 | 0.6667 |
| | 3 | 0.6673 | 0.7748 |
| | 4 | 0.5246 | 0.8468 |
| | 5 | 0.3586 | 0.9189 |

Modellerin sekiz eğitim dönemi sonunda elde ettikleri sonuçlar, çizelge 4.6'da gösterilmektedir. Bu çizelge, modelin daha uzun eğitim süresiyle birlikte daha yüksek doğruluk oranlarına ulaştığını göstermektedir.

**Çizelge 4.6.** PubMed + UpToDate + Rotterdam Veri Seti ile Eğitilen Modellerin Performans Sonuçları (Eğitim Dönemi = 8)

| ***Model*** | ***Eğitim Dönemi*** | ***Kayıp (Loss)*** | ***Doğruluk*** |
|---|---|---|---|
| *BioBERT* | 1 | 0.1913 | 0.9617 |
| | 2 | 0.1943 | 0.9672 |
| | 3 | 0.1360 | 0.9672 |
| | 4 | 0.1166 | 0.9727 |
| | 5 | 0.1189 | 0.9727 |
| | 6 | 0.1271 | 0.9727 |
| | 7 | 0.1181 | 0.9781 |
| | 8 | 0.1203 | 0.9727 |
| *SciBERT* | 1 | 0.2014 | 0.9617 |
| | 2 | 0.1447 | 0.9672 |
| | 3 | 0.0945 | 0.9781 |
| | 4 | 0.1458 | 0.9672 |
| | 5 | 0.1552 | 0.9727 |
| | 6 | 0.1580 | 0.9727 |
| | 7 | 0.1724 | 0.9727 |
| | 8 | 0.1762 | 0.9727 |

Elde edilen bu sonuçlar, büyük ve yüksek kaliteli veri setlerinin dil modeli performansı üzerinde ne denli etkili olduğunu göstermektedir. Modeller, özellikle PubMed'den elde edilen makalelerin hem klinik hem de akademik dil örüntülerini içermesi nedeniyle bağlamı daha iyi öğrenmiştir. Daha önceki deneylerle karşılaştırıldığında, SciBERT ve BioBERT bu geniş veri setinde daha hızlı öğrenme ve yüksek doğruluk oranları elde etti.

Bu çalışma, eğitim sürecinde veri çeşitliliğinin artırılmasının doğruluk oranını ve modelin genellenebilirliğini artırabileceğini göstermektedir. Gelecekte, modelin bu başarısı, daha kapsamlı tıbbi metinlerin entegre edildiği çok dilli veya çok modelli sistemlerin geliştirilmesi için sağlam bir temel oluşturacaktır.

Gelecekte yapılacak çalışmalar, daha büyük ve dengeli veri kümeleriyle sistemin eğitilmesini sağlayarak model doğruluğunu daha da artırabilir. Ayrıca farklı dil modelleri, gelişmiş veri ön işleme teknikleri ve alternatif sınıflandırma algoritmalarının denenmesi, sistemin hem doğruluğunu hem de genellenebilirliğini iyileştirebilir.

Bununla birlikte, modelin performansını etkileyen önemli bir faktör de kullanıcıdan gelen yanıtların doğrudan doğal dil yerine “Evet/Hayır” şeklinde basitleştirilmiş olmasıdır. Bu durum, sınıflandırma sisteminin bağlamı tam olarak anlayabilme yeteneğini sınırlayabilir. Bu nedenle, gelecekte doğal dilin daha zengin bir şekilde işlenmesini sağlayacak yöntemler tercih edilmelidir.

### 4.3. Model Performans Değerlendirmesi: Açıklama Kalitesi Ölçümü

YZ destekli klinik karar destek sistemleri tarafından üretilen verilerin kalitesi çok önemlidir, ancak sistemlerin güvenilirliğini artırmak için orijinal metinle uyumluluğu da önemlidir. BioBERT ve SciBERT olmak üzere önceden eğitilmiş iki farklı dil modeli kullanılarak, çalışmada oluşturulan sistemin ek açıklama kalitesi bu çerçevede değerlendirilmiştir. Her iki modeli de örneklemek için belirli klinik metin alıntıları kullanılmış ve sonuçta elde edilen ek açıklamalar metodik olarak karşılaştırılmıştır.

Ancak değerlendirmede BLEU puanı da dikkate alınmıştır. N-gram tabanlı BLEU ölçütü, özellikle makine çevirisi kalitesini değerlendirmek için oluşturulmuştur. Bu ölçüt, modelin çıktısının anlamsal bütünlük ve dilsel akıcılık açısından referans metne ne kadar benzediğini değerlendirmeye yardımcı olur. Değerlendirmede ayrıca anlamsal benzerliği ölçmek için kullanılan çağdaş bir yöntem olan BERTScore'dan da yararlanılmıştır. Ek açıklama ile kaynak metnin anlam bakımından ne kadar yakın eşleştiğini belirlemek için BERTScore, ifadelerin bağlamsal benzerliğini dikkate alır. Bu ölçüm, geleneksel n-gram tabanlı yaklaşımların ötesinde, bağlamı doğru bir şekilde aktarıp aktarmadığına odaklanarak daha kapsamlı bir inceleme sağlar.

Her iki model de aynı 100 örnek üzerinde değerlendirilmiş ve modelin metin üretim performansını değerlendirmek için beş ve sekiz eğitim dönemi için eğitilmiştir. Üretilen ek açıklamalar BERTScore, BLEU ve ROUGE (ROUGE-1, ROUGE-2 ve ROUGE-L) metrikleri kullanılarak değerlendirilmiş ve her bir epok için referans ek açıklamalarla karşılaştırılmıştır. Model performansının istatistiksel olarak daha anlamlı bir şekilde karşılaştırılması, her bir metrik için puanların 100 örnek üzerinden ortalama

olarak hesaplanmasıyla mümkün olmuştur. Çizelge 4.7 ve Çizelge 4.9, 5 ve 8 eğitim döneminden elde edilen sonuçlar için karşılaştırmalı değerleri göstermektedir.

**Çizelge 4.7.** Klinik Metin Açıklamalarında BioBERT ve SciBERT Performanslarının Çoklu Metriklerle Değerlendirilmesi (5 Eğitim Dönemi)

| Metrik | BioBERT (ortalama) | SciBERT (ortalama) |
|---|---|---|
| **ROUGE-1** | 0.533 | 0.553 |
| **ROUGE-2** | 0.257 | 0.274 |
| **ROUGE-L** | 0.391 | 0.404 |
| **BLEU** | 15.82 | 17.72 |
| **BERTScore** | 0.89 | 0.89 |

ROUGE kriterleri açısından SciBERT modeli BioBERT'ten daha iyi performans göstermiştir. Özellikle, BioBERT ve SciBERT için ROUGE-1 değerleri sırasıyla 0.533 ve 0.553 olarak belirlenmiştir. Bu bulgu, ROUGE-1 metriğinin model çıktısındaki kelimeler ile referans açıklamasındaki anahtar kelimeler arasındaki örtüşmeyi yakalaması nedeniyle SciBERT'in kaynak metindeki önemli terimleri daha hassas bir şekilde kullandığını göstermektedir. Üstünlük ROUGE-2 ve ROUGE-L skorları için karşılaştırılabilir düzeydedir. BioBERT ve SciBERT'in ROUGE-2 puanları sırasıyla 0.257 ve 0.274'tür. ROUGE-L'de BioBERT 0.391 puan, SciBERT ise 0.404 puan almıştır. Bu sonuçlar, SciBERT'in ifade, yapısal ve sözcük düzeyinde daha etkili açıklamalar ürettiğini göstermektedir.

BLEU puanları söz konusu olduğunda, SciBERT en iyisi olmaya devam etmektedir. SciBERT ek açıklamaları için ortalama BLEU puanı 17.72 iken, BioBERT ek açıklamaları 15.82 puan almıştır. Model çıktısının n-gram dizilerinin referans metne benzerliği BLEU metriği ile ölçülür. Buradan SciBERT'in daha tutarlı ve sözdizimsel olarak daha sağlam cümle yapıları oluşturduğu açıkça görülmektedir. BERTScore analizine göre modellerin performansı daha dengeli dağılmıştır. Bağlamsal anlam benzerliğini ölçen bu metrik, her iki modele de 0.89 F1 puanı vermiştir. Bu sonuç, BioBERT'in anlamı kaybetmeden açıklamalar sunabildiğini ve bağlamsal bilginin bütünlüğünü korumada SciBERT kadar etkili olduğunu göstermektedir.

Genel olarak, SciBERT modeli biçimsel ve yüzeysel benzerliği ölçen ROUGE ve BLEU gibi kriterlerde daha iyi performans göstererek açıklamalarının dilsel ve yapısal olarak daha sağlam olduğunu göstermektedir. Bu da özellikle klinik ortamda okunabilirlik ve doğruluk açısından önemli bir fayda sağlamaktadır. Bununla birlikte, BioBERT'in bağlamsal anlamsal bütünlüğü koruma kapasitesi göz önünde bulundurulduğunda, her iki modelin de anlamsal doğruluk açısından tatmin edici bir şekilde işlediği iddia edilebilir. Özetle, SciBERT modeli, PKOS teşhisi için tasarlanan sohbet robotu sisteminde açıklama üretmedeki genel etkinliği ve tutarlılığı nedeniyle zorlayıcı bir rakiptir.

Çizelge 4.8 sekiz eğitim dönemi ardından elde edilen değerlendirme metriklerini göstermektedir. Model, sekiz eğitim dönemi sonra, önceki beş eğitim dönemine kıyasla genel olarak daha iyi performans göstermiştir. Özellikle ROUGE-L ve BERTScore gibi semantik tabanlı metriklerde önemli kazanımlar elde edilmiştir. Bu bulgular, modeli eğitmek için ne kadar çok zaman harcanırsa, oluşturulan metinlerin referans ek açıklamalarına o kadar çok benzediğini göstermektedir. Bununla birlikte, BLEU ve ROUGE-2 gibi bazı metriklerde sınırlı düzeyde değişim gözlenmiş olup, bu durum modelin belirli yönlerinin daha erken doyuma ulaştığını düşündürmektedir.

**Çizelge 4.8.** Klinik Metin Açıklamalarında BioBERT ve SciBERT Performanslarının Çoklu Metriklerle Değerlendirilmesi (8 Eğitim Dönemi)

| Metrik | BioBERT (ortalama) | SciBERT (ortalama) |
|---|---|---|
| **ROUGE-1** | 0.467 | 0.479 |
| **ROUGE-2** | 0.240 | 0.252 |
| **ROUGE-L** | 0.339 | 0.351 |
| **BLEU** | 11.81 | 12.41 |
| **BERTScore** | 0.887 | 0.893 |

ROUGE-1 ölçeğinde BioBERT 0.467 ve SciBERT 0.479 puan almıştır. Bu skorlar SciBERT'in referans ek açıklamalarında kullanılan terimleri daha hassas bir şekilde kullanabildiğini göstermektedir çünkü ROUGE-1 metriği ek açıklama çıktısının referans ek açıklamayla kelime kelime ne kadar yakın eşleştiğini değerlendirmektedir. Benzer şekilde, BioBERT ve SciBERT için ROUGE-2 değerleri sırasıyla 0.240 ve 0.252 olarak belirlenmiştir. Bu tutarsızlık, ROUGE-2 ardışık iki kelimelik kelime dizilerinin benzerliğini ölçtüğü için SciBERT'in daha akıcı ve yapısal olarak uyumlu cümleler ürettiğini göstermektedir. ROUGE-L metriği, SciBERT'in 0.351 ve BioBERT'in 0.339 puan almasıyla karşılaştırılabilir bir baskınlık göstermektedir. Yapısal benzerlik açısından önemli olan ROUGE-L, modelin referans açıklama ile örtüşen en uzun kelime dizisini dikkate alır. SciBERT'in bu konuda daha kapsamlı ve yapısal olarak daha sağlam yanıtlar verdiği gözlemlenmektedir.

SciBERT ve BioBERT sırasıyla 12.41 ve 11.81 BLEU değerlerine sahiptir. Bu farklılık, BLEU referans ve model çıktısı arasındaki benzerliği n-gram düzeyinde hesapladığından, SciBERT'in referans cümlelere daha yakın, dilbilgisi açısından daha doğru tanımlar ürettiğini göstermektedir. Daha yüksek bir BLEU puanı, özellikle tıbbi terminolojide doğruluğun gerekli olduğu durumlarda modelin güvenilirliğini artırmaktadır.

SciBERT, bağlamsal anlamsal bütünlüğü ölçen BERTScore istatistiğinde küçük bir avantaja sahipti. BioBERT ve SciBERT için puanlar sırasıyla 0.887 ve 0.893'tür. SciBERT'in anlamı korumada daha etkili olduğu görülmektedir çünkü BERTScore hem yüzey benzerliğini hem de ifadelerin bağlama özgü anlamını dikkate almaktadır. Çok fazla görünmese de, bu ayrım bağlamsal doğruluk için çok önemlidir.

Eğitim süresi artırıldığında her iki model de daha iyi performans gösterirken, SciBERT bu prosedürden daha fazla kazanç elde etmekte ve açıklama kalitesi daha çarpıcı bir şekilde artmaktadır. SciBERT, her istatistikte BioBERT'ten daha iyi performans göstermesinin de kanıtladığı gibi, bilimsel metinler, özellikle de klinik açıklamalar oluşturmak için daha iyi bir modeldir. Sonuç olarak, SciBERT klinik karar destek sistemleri veya tıbbi açıklamaların oluşturulması için daha iyi bir seçim olarak görülebilmektedir.

### 4.3.1. Klinik metin açıklamalarında kullanılan değerlendirme metriklerinin anlamı ve beklenen aralıkları

Metin üretimindeki doğruluk, akıcılık ve anlamsal benzerlik, bu çalışmada kullanılan değerlendirme ölçütleri kullanılarak ölçülmektedir. BLEU, üretilen metnin referans metne yüzeysel benzerliğini ölçen ROUGE-1, ROUGE-2 ve ROUGE-L'ye göre gramer yapılarına ve n-gram eşleşmelerine daha fazla önem vermektedir. Bu ölçümlerde 0,5 ila 0,6 aralığındaki ROUGE-1 puanları genellikle güçlü olarak kabul edilir, ancak 0,25'in üzerindeki ROUGE-2 değerlerinin daha kısa terimlerin örtüşmesini göstermek için yeterli olduğu düşünülmektedir. Bu çalışmada SciBERT'in ROUGE-1 ve ROUGE-2 skorları sırasıyla 0,553 ve 0,274 olup tahminlerle tutarlıdır. Çizelge 4.9'da genel anlamda metrikler değerlendirilmiştir.

**Çizelge 4.9.** Kullanılan Değerlendirme Metriklerinin Anlamı ve Beklenen Aralıklar

| Metrik | Normal Beklenen Aralık | Proje Değerleri (En Yüksek) | Yorum |
|---|---|---|---|
| **ROUGE-1** | 0.5 – 0.6 | **0.553** (SciBERT, 5 epok) | Referansla kelime düzeyinde benzerlik için kabul edilebilir ve iyi düzey. |
| **ROUGE-2** | 0.25 – 0.35 | **0.274** (SciBERT, 5 epok) | Kelime çiftleri düzeyinde tutarlılık, yeterli düzeyde; ileri iyileştirme yapılabilir. |
| **ROUGE-L** | 0.35 – 0.45 | **0.404** (SciBERT, 5 epok) | Cümle yapısı ve dizilim benzerliğinde yeterli başarı. |
| **BLEU** | 15 – 30 (**metin üretiminde**) | **17.72** (SciBERT, 5 epok) | N-gram benzerliği açısından kabul edilebilir, daha yüksek değerler insan benzeri metinleri gösterir. |
| **BERTScore** | ≥ 0.85 (**yüksek**), ≥ 0.90 (**çok yüksek**) | **0.893** (SciBERT, 8 epok) | Semantik benzerlik için çok yüksek düzeyde. Bağlamsal anlam güçlü şekilde yakalanmış. |

Özellikle metin üretimi veya makine çevirisi içeren görevlerde, 15'ten büyük BLEU puanları kabul edilebilir kaliteyi gösterir. Bu araştırmada SciBERT, beş eğitim epokundan sonra 17,72 BLEU puanı elde ederek modelin doğru ve dilsel açıdan akıcı ifadeler ürettiğini göstermiştir. 0,85'lik bir puan, güçlü anlamsal bütünlüğe sahip olarak kabul edilir. BERTScore, bağlamsal benzerliği hesaba katan çağdaş bir metriktir. SciBERT modeli, 0,893'lük BERTScore ile ödev üzerinde harika bir iş çıkarmıştır. BioBERT 0,887 ile karşılaştırılabilir derecede bağlamsal doğruluk sunmuştur.

Sonuç olarak, bu metriklerin projeye özgü değerleri, PKOS teşhisi için oluşturulan açıklayıcı metin oluşturma sisteminin biçimsel ve anlamsal sağlamlığını göstermektedir. Özellikle, SciBERT modeli anlamsal tutarlılık ve dilsel doğruluk açısından daha iyi performans göstererek klinik karar destek uygulamaları için üstün bir seçenek haline gelmektedir.

### 4.4. Otomatik Etiketleme ve Eğitilmiş Modellerin Sınıflandırma Performansı

Her iki modelin de test verileri üzerindeki nihai performansı model eğitiminin ardından değerlendirilmiştir. Test setinde, BioBERT modeli makro ortalama %52 F1 puanı ve %96.44 doğruluk elde etmiştir. Ancak, %64'lük ortalama F1 skoru ile SciBERT modeli BioBERT'i geride bırakmış ve test setinde %97'lik daha yüksek bir doğruluğa ulaşmıştır. Bu bulgular, SciBERT'in bağlamlar arasında ayrım yapma konusunda daha iyi performans gösterdiğini ortaya koymaktadır.

Özellikle "PKOS'un Ayırıcı Tanısı" sınıfında, sınıflandırma raporlarına göre SciBERT modeli sonuçları doğru tahmin etmede %100 doğruluk, %16.7 hatırlama ve %28 F1 puanı göstermiştir. Ancak SciBERT, her iki modelin de başarılı olduğu "PKOS'un Epidemiyolojisi, Patogenezi ve Tedavisi" sınıfında %98.5 F1 puanı ve %97.5 doğruluk oranıyla öne çıkmıştır. "Tanı ve Klinik Özellikler" sınıfında BioBERT %60 F1 skoru elde ederken, SciBERT Çizelge 4.10'da gösterildiği gibi %67 ile daha dengeli bir performans sergilemiştir.

**Çizelge 4.10.** BioBERT ve SciBERT Modellerinin Test Seti Sınıflandırma Performansları (5 Eğitim Dönemi)

| Sınıf | Model | Doğruluk | Duyarlılık | F1 Puanı |
|---|---|---|---|---|
| **PKOS Teşhisi ve Klinik Özellikler** | BioBERT | 0.75 | 0.50 | 0.60 |
| | SciBERT | 0.78 | 0.58 | 0.67 |
| **Ayırıcı Tanı** | BioBERT | 0.00 | 0.00 | 0.00 |
| | SciBERT | 1.00 | 0.17 | 0.29 |
| **Epidemiyoloji ve Tedavi** | BioBERT | 0.97 | 1.00 | 0.98 |
| | SciBERT | 0.97 | 1.00 | 0.99 |
| **Genel Ortalama (Macro)** | BioBERT | 0.57 | 0.50 | 0.53 |
| | SciBERT | 0.92 | 0.58 | 0.65 |
| **Ağırlıklı Ortalama (Weighted)** | BioBERT | 0.95 | 0.96 | 0.95 |
| | SciBERT | 0.97 | 0.97 | 0.96 |

BioBERT'in "PKOS'un Teşhisi ve Klinik Özellikleri" sınıfındaki ortalama performansı %75 kesinlik, %50 geri çağırma ve %60 F1 puanı olmuştur. Aynı sınıfta %77.8 hassasiyet, %58.3 geri çağırma ve %66 F1 skoru ile SciBERT daha dengeli bir bağlamsal teşhis performansı göstermiştir.

İki modelin ağırlıklı ortalama puanları da karşılaştırıldığında BioBERT ve SciBERT sırasıyla %95.4 ve %96.3 F1 puanları elde etmiştir. Bu farklılık, küçük örneklem boyutlarına sahip sınıflarda bile SciBERT'in genellemede nispeten daha iyi performans gösterdiğini ve daha eşit dağılımlı sınıfları tahmin ettiğini göstermektedir.

Beş epoğun ardından elde edilen bu bulgulara göre, SciBERT modeli "Ayırıcı Tanı" sınıfında sınırlı performans göstermiş, ancak SciBERT modeli genelleme başarısını küçük verilerde bile sürdürebilmiştir. Bununla birlikte, her iki model için de en iyi performansa sahip sınıf "Epidemiyoloji ve Tedavi" olmuştur. SciBERT, genel ortalamalardaki daha yüksek F1 puanının da gösterdiği gibi, çeşitli durumlardaki derslerde daha dengeli yargılarda bulunabilmektedir.

Model eğitim süresi sekiz epoğa uzatıldığında sınıflandırma performansında değişiklikler görülmüştür. "Teşhis ve Klinik Özellikler" sınıfında, BioBERT %86 doğruluk ve %63 F1 puanı ile SciBERT'ten daha iyi performans gösterirken, SciBERT %100 doğruluk ve %59 F1 puanı ile daha düşük duyarlılık göstermiştir. Bu, SciBERT'in zayıf kapsama alanına ancak iyi tahmin doğruluğuna sahip olduğu anlamına gelmektedir.

**Çizelge 4.11.** BioBERT ve SciBERT Modellerinin Test Seti Sınıflandırma Performansları (8 Eğitim Dönemi)

| Sınıf | Model | Doğruluk | Duyarlılık | F1 Puanı |
|---|---|---|---|---|
| **PKOS Teşhisi ve Klinik Özellikler** | BioBERT | 0.86 | 0.50 | 0.63 |
| | SciBERT | 1.00 | 0.42 | 0.59 |
| **Ayırıcı Tanı** | BioBERT | 1.00 | 0.17 | 0.29 |
| | SciBERT | 0.67 | 0.33 | 0.44 |
| **Epidemiyoloji ve Tedavi** | BioBERT | 0.97 | 1.00 | 0.99 |
| | SciBERT | 0.97 | 1.00 | 0.99 |
| **Genel Ortalama (Macro)** | BioBERT | 0.97 | 0.97 | 0.96 |
| | SciBERT | 0.88 | 0.58 | 0.67 |
| **Ağırlıklı Ortalama (Weighted)** | BioBERT | 0.97 | 0.97 | 0.96 |
| | SciBERT | 0.97 | 0.97 | 0.96 |

Çizelge 4.11, sekiz eğitim dönemi boyunca BioBERT modelinin sınıf düzeyinde daha iyi tutarlılık ve "Ayırıcı Tanı" gibi düşük örneklemli sınıflarda tahmin performansında en azından mütevazı bir iyileşme gösterdiğini ortaya koymaktadır. Buna karşılık, SciBERT F1 puanı açısından liderliğini sürdürmüş, ancak "Teşhis ve Klinik Özellikler" sınıfındaki hassasiyeti azalmıştır.

Sonuç olarak, SciBERT modeli F1 puanları ve genel sınıf dengesi açısından daha iyidir. Bu, her iki modelin de sınıflandırma doğruluğu açısından mükemmel performans göstermesine rağmen, SciBERT'in PKOS ile ilgili metinleri sınıflandırmak için daha güvenilir ve sağlam bir model olduğunu göstermektedir. Bununla birlikte, BioBERT, özellikle tanı hassasiyetinin önemli olduğu durumlarda, düşük varyanslı ve tutarlı sonuçları nedeniyle tercih edilebilmektedir.

# 5. TARTIŞMA

Çalışma kapsamında geliştirilen YZ tabanlı sohbet robotu, PKOS teşhisi sürecinde doktor-hasta etkileşimini taklit eden, DDİ temelli özgün bir sistem olarak tasarlanmıştır. Geliştirilen model, hastanın semptomlarını metin üzerinden alarak, eksik veya yetersiz bulduğu bilgiler için soru üretme yeteneğine sahiptir. Böylece doktorun doğrudan sorması gereken detaylı soruları otomatik olarak yöneltebilmekte ve toplanan veriler doğrultusunda PKOS riskini yüzde cinsinden tahmin edebilmektedir. Bu yaklaşım, hem teknik alt yapısı hem de kullanıcıyla kurduğu etkileşim bakımından literatürdeki önceki çalışmalardan farklılık göstermektedir.

Literatürde, PKOS tanısı genellikle klinik bulgulara, hormon test sonuçlarına ve ultrason verilerine dayanmaktadır. Bununla birlikte, YZ modelleri tarafından işlenen metin tabanlı hasta şikâyetlerinin ön değerlendirmesi, sağlık profesyonellerinin iş yükünü azaltabilir ve hastaya zaman kazandırabilir. Bu tür bir sohbet robotu tasarımıyla doğrudan ilgili çok az araştırma vardır, ancak benzer tekniklerin çeşitli hastalık gruplarında kullanıldığı görülmektedir.

Chang vd. (2024), ChatGPT tabanlı chatbotların psikiyatrik danışmanlık alanında empatik yanıt verebilme yeteneğini test etmiş ve kullanıcılar tarafından yüksek oranda tatmin edici bulunduğunu belirtmişlerdir. Rasch analizine göre modelin toplam yetenek seviyesi 2.23'tür. Bu skor, modelin genel performansını ve zorlu sorular karşısındaki başarısını göstermiştir. Sohbet robotlarının verdiği yanıtların doğruluk oranı yaklaşık %85.6 olarak bildirilmiş ve kullanıcılar tarafından tatmin edici bulunmuştur ve itibar puanı 4.5/5 olarak hesaplanmıştır. Bu çalışma, DDİ modellerinin yalnızca bilgi sunmakla kalmayıp, empatik bir biçimde insanlarla iletişim kurabilme potansiyelini de ortaya koymaktadır. Benzer şekilde, Sorin vd. (2024) tarafından yürütülen bir başka araştırmada, ChatGPT'nin empatik yanıt üretim performansı insan uzmanlarla kıyaslandığında anlamlı derecede yüksek bulunmuş ve ortalama 4.64/5 kullanıcı memnuniyeti puanı aldığı raporlanmıştır. Bu çalışmalar, tez çalışmasında empatik soru-cevap yapısını modelleme girişimini desteklemektedir. Bununla birlikte, mevcut uygulamaların çoğu hastalık tespit etmek yerine bilgi verme veya yönlendirme amaçlıdır. Çalışmada, hastanın semptomlarına göre risk yüzdesi çıkarılması ve model tarafından eksik bilgi sorgulanması, mevcut literatüre yenilikçi bir katkı sunmaktadır. Son yıllarda, birçok araştırmacı, klinik parametreler ve vital işaretlerden oluşan veri setleri ile PKOS teşhisi için YZ modelleri önermiştir Bu, YZ'nın yalnızca statik bilgi sağlayıcı olmaktan çıkıp interaktif bir karar destek sistemine dönüşmesinin bir örneğidir.

## 5.1. Örnek Vaka Uygulaması: Sistem Yanıtı ile Gerçekleştirilen Dinamik Tanı Süreci

Bu çalışmada geliştirilen YZ tabanlı karar destek sistemi, hastanın yazdığı metni analiz ederek eksik bilgileri bulmak ve makale temelli sorular oluşturmak için kullanılabilir. Aşağıdaki uygulama, sistemin adım adım nasıl çalıştığına dair bir klinik vaka senaryosudur. Bu örnek, sistemin tanı sürecine aktif olarak katıldığını göstermektedir, yani sadece bilgi dağıtan bir yapıdan ileri bir aşamaya geçmiştir.

Örnek hasta metni (Girdi):

"*The patient is a 25-year-old woman presenting with concerns about irregular menstrual cycles and excessive facial hair. She reports having periods every 45 to 60 days for the past year. In addition, she has noticed increased terminal hair growth on her chin, upper lip, and lower abdomen. She also mentions persistent acne, especially on her jawline and cheeks. Physical examination revealed moderate hirsutism and a BMI of 28. The patient denies any significant recent weight changes, and there is no known family history of similar endocrine issues.*"

Bu yazıya dayanarak sistem, eksik tanı gereksinimlerini belirlemiş ve bunlara ilişkin on bir dinamik soru sormuştur. Her soruya verilen yanıtlara göre sistem, Gemini modeli üzerinden bilimsel makalelerden alınan bağlama dayalı tıbbi açıklamalar üretti. Bu değerlendirme sonuçları aşağıdakilerdir:

**Dinamik Tanı Özeti:**

- **PKOS riski:** %53
- **Sorulan soru sayısı:** 11
- **Kritik bulgular:**
  - Transvaginal ultrasonografi yapıldı: ✓
  - Polikistik over görünümü mevcut: ✓
  - AMH yüksekliği mevcut: ✓
  - Hirsutizm gözlemleniyor: ✓
  - Galaktore yok: ✗

Sistem, sonuçların yanı sıra her örnek için literatüre dayalı testler önermiş ve potansiyel ayırıcı tanıları belirlemek için kullanıcı verilerini analiz etmiştir. Bu vaka çalışması, oluşturulan sistemin sadece bilgi veren bir yapının ötesine geçerek literatür destekli bilgi çıkarma, eksik veri sorgulama ve ayırıcı tanı oluşturma gibi çok katmanlı faaliyetleri nasıl gerçekleştirdiğini göstermektedir. Bu şekilde sistem, YZ destekli klinik karar verme sürecine aktif olarak katılmaktadır

Sistemin ürettiği açıklamalar ve teşhis sorguları için birkaç metinsel değerlendirme ölçütü hesapladığı görülmektedir. Bunlardan ilki olan "Soru-Metin Uyum Skoru" %6 olarak bulunmuştur. Bu oran, sistemin ürettiği sorular ile hastanın yazılı metninin ne kadar örtüştüğünü ölçmektedir. Düşük bir puan, algoritmanın metnin doğrudan içeriğine ek olarak dolaylı veya eksik tanısal olarak önemli bilgilere odaklandığı anlamına gelir. Başka bir deyişle, sistem aktif olarak tanı için gerekli olan ancak metinle eşleşmeyen bilgileri aramaktadır. Sistemin ürettiği ve hastanın metniyle tam olarak eşleşen soruların yüzdesi "Tam Eşleşen Soru Oranı" olarak bilinir ve bu oran %8'dir. Bu oranın da aynı şekilde düşük olması, yöntemin etkisiz olduğu anlamına gelmemektedir; daha ziyade, sadece gerçekleri tekrarlamak yerine, eksik olan ve doktorun gözden kaçırabileceği bilgileri vurgulamaktadır. Son olarak, %15 "Metinsel Değerlendirme Yeterlilik Puanı" olarak belirlenmiştir. Bu puan, sistemin ek açıklamalarının bağlamsal doğruluğunu değerlendirmek için BERTScore tabanlı anlamsal benzerlik tekniğini kullanır. Skora göre, sistemin ek açıklamaları bağlamı yakalamak için büyük bir umut vaat ediyor ve esasen literatürle tutarlı. Sistem, yalnızca

verilen verileri değerlendirmekle kalmayan, aynı zamanda bu metriklerin genel yorumuna göre daha ilgili bir klinik değerlendirme üretmek için boşlukları doldurabilen bir YZ tekniği sunmaktadır. Tam ekran çıktı ve değerlendirme sürecine ilişkin detaylar EK-2'de verilmiştir.

Sistemin İngilizce dilinde çalışması da teknik bir zorunluluğun ötesinde, bilinçli bir tercihtir. Bunun nedeni, sistemde kullanılan BioBERT/SciBERT gibi önceden eğitilmiş dönüştürücü tabanlı biyomedikal modellerin ve Gemini gibi üretici dil modellerinin İngilizce yazılmış tıbbi literatür üzerinde eğitilmiş olmasıdır. Sonuç olarak, bu modeller İngilizce metinlerde çok daha kesin, bağlamsallaştırılmış ve bilime dayalı yanıtlar üretebilir. Ayrıca, sistemin bilgi almak için kullandığı veri kaynaklarının (PMC ve PubMed gibi) büyük çoğunluğu da aynı şekilde İngilizcedir. Modellerin doğal çalışma dilinin avantajlarından faydalanmak ve küresel literatürle tutarlılığı garanti etmek için vaka çalışması İngilizce olarak oluşturulmuştur. Sistemin adım adım algoritmik süreci EK-3'te ayrıntılı olarak verilmiştir.

## 5.2. Sınırlılıklar ve Geliştirme Önerileri

Günümüzde klinik metin analizi genellikle ChatGPT, Claude, Gemini, Grok ve DeepSeek gibi genel amaçlı BDM'leri kullanılarak yapılmaktadır. Bu modeller, PKOS gibi karmaşık teşhis prosedürlerinde bilgi sağlama yeteneklerine rağmen, tıbbi karar destek sistemlerine entegrasyon söz konusu olduğunda önemli sınırlara sahiptir. Bu modellerin çoğu kullanıcıyı yönlendiren aşamalı bir diyalog yapısından yoksundur, eksik bilgileri tanımlayamaz ve tanı kriterlerini rutin olarak değerlendirmez.

Çalışmanın DDİ tabanlı karar destek sistemi, bilgi sağlamaktan daha fazlasını yapmak üzere tasarlanmıştır; kullanıcı tarafından sağlanan klinik metinleri analiz eder, eksik verileri tanımlar, dinamik sorular oluşturur, ilgili literatüre dayalı açıklamalar sunar ve sonunda bir tanı risk yüzdesi belirleyerek etkileşimli bir tanı süreci sağlamaktadır.

Çizelge 5.1, soru oluşturma, risk tahmini, tanı kriterleri kontrolü, diyalog yönetimi ve literatüre dayalı doğrulama açısından, oluşturulan karar destek sistemini ChatGPT, Claude, Gemini, DeepSeek ve Grok gibi iyi bilinen BDM tabanlı sistemlerle karşılaştırmaktadır.

**Çizelge 5.1.** ChatGPT ile Çalışmanın Karşılaştırılması

| Özellik | Geliştirilen Sistem | ChatGPT | Gemini | Claude | DeepSeek | Grok |
|---|---|---|---|---|---|---|
| **Soru Üretme** | Eksik bilgiye göre dinamik | ✕ Kullanıcı sormazsa üretmez | ✕ | ✕ | ✕ | ✕ |
| **Risk Yüzdesi Tahmini** | ✓ Makale temelli % hesaplama | ✕ | ✕ | ✕ | ✕ | ✕ |
| **Tanı Kriteri Kontrolü** | ✓Sistematik, otomatik | ✓Üç kriteri açıkça işler | ⚠ Belirsiz | ✓Üç kriteri açıkça işler | ✓Üç kriteri açıkça işler | ✓Üç kriteri kısmen işler |

**Çizelge 5.1**'in devamı

| | | | | | | |
|---|---|---|---|---|---|---|
| **Diyalog Yönetimi** | Adım adım, yönlendirici | Açık uçlu | Açık uçlu | Açık uçlu | Açık uçlu | Açık uçlu |
| **Makale Temelli Doğrulama** | ✓ FAISS + Gemini ile | ✕ | ⚠ Kısmen (içerik kullanımı) | ✕ | ✕ | ✕ |
| **Açıklama Kalitesi Skorları (ROUGE, BERTScore)** | ✓ Otomatik hesaplanır | ✕ | ✕ | ✕ | ✕ | ✕ |
| **Ayırıcı Tanı Önerisi** | ✓ Liste + test önerisi | ✕ | ✕ | ✕ | ⚠ Genel öneriler | ✕ |
| **Klinik Veri Analizi (Hormon, Ferriman vb.)** | ✓ Metinden analiz yapar | ✕ Elle yorum gerekir | ✕ | ✕ | ✕ | ✕ |
| **Tanı Açıklaması Stili** | Madde madde + bilimsel gerekçeli | Madde madde + sade | Paragraf odaklı | Direkt tanılayıcı | Protokol stili | Basit/sohbet tarzı |
| **Şeffaflık / Gerekçelendirme** | ✓ Kanıta dayalı açıklama + kaynak | ✓ Metne dayalı gerekçe | ✓ Kısmi | ✕ | ⚠ Yüzeysel | ⚠ Basit uyarı içerir |
| **Etik Uyarı ve Sorumluluk Reddi** | ✓ Tıbbi uyarı içerir | ✓ | ✓ | ✓ | ✓ | ✓ |
| **Klinik Karar Destek Sistemi Rolü** | ✓ Karar destekleyici sistem | ✕ Bilgi destekleyici | ✕ | ✕ | ⚠ Öneri verir | ✕ |
| **İnteraktif Kullanıcı Deneyimi** | Dinamik soru-cevap + analiz | Tek seferlik cevap | Tek seferlik cevap | Tek seferlik cevap | Tek seferlik cevap | Tek seferlik cevap |
| **Makaleden Bilgi Alma** | ✓ FAISS ile ilgili bölümü çeker | ✕ | ✕ | ✕ | ✕ | ✕ |
| **Açıklanabilirlik** | ✓ SHAP tarzı metriklerle destekler | ✕ | ✕ | ✕ | ✕ | ✕ |

Mevcut genel amaçlı BDM'ler ile (ChatGPT, Claude, Gemini, DeepSeek, Grok) karşılaştırıldığında, Çizelge 5.1 önerilen DDİ tabanlı karar destek sisteminin PKOS teşhis sürecinde daha klinik olarak yönlendirilmiş ve hedeflenmiş bir performans sergilediğini göstermektedir. Oluşturulan sistemin eksik verileri tanımlama ve yanıt olarak dinamik sorgular sağlama yeteneği en dikkat çekici özelliklerinden biridir. Eksik teşhis kriterlerini sorarak, sistem yalnızca sağlanan verileri değerlendirmekle kalmaz, aynı zamanda

kullanıcının teşhis prosedürüne aktif katılımını da garanti etmektedir. Ayrıca sistem, makale tabanlı doğrulama özelliği sayesinde ürettiği açıklamaları desteklemek için literatürden kanıta dayalı bilgiler sağlayarak klinik güvenilirliği ve açıklama kalitesini artırmaktadır.

Diğer modeller genellikle teşhis sürecini sistematik olarak yönetmez; bunun yerine açık uçlu konuşmalar yapar ve kullanıcı girdisine dayalı yanıtlar üretmektedir. Örneğin, ChatGPT veya Gemini, ultrason sonuçları veya hirsutizm varlığı gibi diğer tanı kriterlerini aktif olarak sormaz, ancak bir hasta sadece "adet düzensizlikleri" bildirirse hemen PKOS tanısı ile ilgili genel bilgiler sunmaktadır.

Kullanıcıya interaktif ve rehberli bir deneyim sunmak için, tasarlanan sistem ayrıca risk yüzdesi tahmini, klinik tanı kriterlerinin kontrolü, ayırıcı tanı önerileri ve açıklama puanlaması (ROUGE, BERTScore) gibi sofistike özellikler içermektedir.

Bu şekilde sistem, bilgi sunan bir dil modeli olmanın yanı sıra karar verme sürecini destekleyen, organize eden ve netleştiren bir yardımcı hekim aracı olarak konumlandırılmaktadır. Klinik tanı desteği söz konusu olduğunda, bu karşılaştırma tablosunun da ortaya koyduğu gibi, alana özgü dil modeli destekli sistemler, devasa genel amaçlı modellerden daha güvenilir, etkili ve yol gösterici sonuçlar verebilmektedir.

Çalışmanın BioBERT ve SciBERT modelleri, test setindeki daha iyi performanslarının yanı sıra literatürdeki sağlam itibarları nedeniyle seçilmiştir. Bağlamsal anlamsal tutarlılığa dayalı açıklamalar üretme hedefi doğrultusunda SciBERT modeli özellikle güçlü BERTScore ve BLEU dereceleri almıştır. Öte yandan BioBERT, genel kategorizasyon görevlerinde sürekli olarak mükemmel doğruluk sağlayarak sistemin güvenilirliğini artırmıştır.

Benzer sınıflandırma görevi doğruluk oranlarına diğer tıbbi modeller (PubMedBERT, ClinicalBERT ve BlueBERT) tarafından da ulaşılmıştır; bununla birlikte, özellikle bağlama duyarlı küçük sınıf vakalarında F1 puanlarında kayda değer düşüşler göstermişlerdir. Bu durum, bu modellerin genel olarak doğru olsalar bile, karar destek sisteminin açıklanabilirlik ve bağlam temelli değerlendirme ihtiyaçlarını karşılayamayabilecekleri anlamına gelmektedir.

Bu karşılaştırma tablosu, alana özgü dil modeli destekli sistemlerin, klinik teşhis yardımı alanında, genel amaçlı büyük modellerden daha güvenilir, etkili ve yönlendirici çıktılar üretebileceği sonucuna varmaktadır. Bu bulgu, BDM'nin klinik bağlamda daha hedefe yönelik hale getirilmesi gerektiğini de göstermektedir.

Bu bağlamda, özellikle biyomedikal alanlara özgü olarak önceden eğitilmiş dil modelleri öne çıkmaktadır. BioBERT ve SciBERT modelleri, genel dil modellerine kıyasla biyomedikal metinlerde daha yüksek doğruluk düzeylerine sahiptir. BioBERT, Lee ve arkadaşları tarafından (Lee vd. 2020) yapılan kapsamlı bir araştırmada, klinik belge sınıflandırma ve adlandırılmış varlık tanıma gibi görevlerde temel BERT modeline göre %3 ila 10'luk bir performans artışı sağladı. Bu nedenle, PKOS gibi klinik terimlerin yer aldığı metinlerde doğru anlam çıkarmak için tercih edilmesini desteklemektedir.

BioBERT ve SciBERT modelleri tarafından üretilen açıklamaların niteliğine ilişkin ölçütlerde anlamlı bir fark olup olmadığını tespit etmek için bu araştırmada eşleştirilmiş örneklem t-testi kullanılmıştır. Analiz bulguları, SciBERT modelinin ROUGE, BLEU ve BERTScore gibi değerlendirme ölçütlerindeki ortalama puanlarının daha yüksek olmasına rağmen, bu farkların istatistiksel olarak anlamlı olmadığını göstermiştir. Tüm ölçümlerin p-değerlerinin 0.05 anlamlılık düzeyinden yüksek olması da bunu desteklemektedir (bkz. Çizelge 4.8, Çizelge 4.10). Bu sonuç, iki model arasındaki farklılıkların tesadüfi olabileceğini ve açıklama üretme açısından benzer olduklarını ima etmektedir. Dolayısıyla, SciBERT teknik olarak daha üstün olsa da, BioBERT de açıklama kalitesi açısından güvenilir bir seçim olarak kabul edilebilir.

Metin tabanlı modelleme yoluyla hastalık riski tahmini yapıldığında, dikkat edilmesi gereken bazı sınırlamalar vardır. Öncelikle, hastanın ifadeleri bağlama ve kültürel olarak farklı olabilir. Model bu tür değişikliklere dayanmalıdır. Ek olarak, modelin çıkardığı sonucun güvenilirliği, verilen bilgilerin doğruluğu ve eksiksizliğine doğrudan bağlıdır. Modelin soru sorma stratejilerini geliştirmesi ve pekiştirmeli öğrenme bu konuda önemli bir katkıda bulunmuştur. Belirli örneklerde doğru yerde soru sormak, modelin gelecekte benzer teşhis sistemlerinin daha akıllı ve kullanıcıya duyarlı hale getirilebileceğini göstermektedir.

Genel olarak bu çalışma, DDİ tabanlı teşhis destek sistemlerinin PKOS gibi klinik tanıda kullanılabilirliğini göstermesi açısından özgün bir örnek teşkil etmektedir. Önerilen sistem, doktorlara ilk değerlendirme aşamasında yardımcı olabilecek bir ön tarama aracı olma potansiyeli taşımaktadır. Ancak, gerçek dünyada klinik kullanım için daha fazla veriyle eğitilmesi, çeşitli hasta profilleri üzerinde test edilmesi ve etik sorumlulukların dikkatle değerlendirilmesi gereklidir.

Deneysel değerlendirmelere göre, SciBERT modeli bağlamsal uyum, açıklama üretme kalitesi ve sınıflandırma doğruluğu açısından ortalama olarak BioBERT'ten daha iyi performans göstermiştir. Bu sonuçlar göz önüne alındığında, SciBERT modelinin özellikle PKOS gibi bağlama son derece duyarlı klinik durumlarda daha iyi bir seçim olduğu açıktır. Bununla birlikte, bu analizlerin küçük bir veri kümesine ve örneklem büyüklüğüne dayandığı belirtilmelidir; sonuç olarak, bu bulguların daha ileri araştırmalarla doğrulanması kritik önem taşımaktadır.

# 6. SONUÇLAR

Bu çalışma, PKOS için YZ destekli yeni bir klinik karar destek sistemi oluşturmak için makale tabanlı bilgi çıkarma yaklaşımlarını, BDM ve DDİ birleştirmektedir. Sistem, hastanın serbest metnini inceleyerek eksik bilgileri belirleyebilir, dinamik sorgular oluşturabilir ve bilimsel literatüre dayalı iyi gerekçelendirilmiş açıklamalar sağlayabilir. Bu çerçeve, geleneksel BDM sistemlerinin ötesine geçerek klinik yargıları destekleyen etkileşimli bir mimari sunmaktadır.

Örnek bir vaka uygulamasında sistem, hastanın yazdığı metne dayanarak PKOS ihtimalini %53 olarak hesaplamış, tanısal açıdan eksik görülen noktalar için 11 dinamik soru yöneltmiş ve kritik bulgulara ilişkin bilimsel makalelerden elde edilen bağlamlarla desteklenmiş açıklamalar üretmiştir. Ancak bu oran, sistemin genel doğruluğunu değil, hastaya özgü verilerden türetilmiş bireysel bir risk tahmini oranını temsil etmektedir. Sistem, tanı kriterlerini dinamik olarak sorgulayıp değerlendirebildiğinden, sabit bir karar modeli yerine kişiselleştirilmiş bir tanısal yaklaşım sergilemektedir. Ayırıcı tanı bağlamında, sistem ayrıca çeşitli hastalık senaryolarını incelemiş ve her biri için uygun testler önermiştir. Sonuç olarak, PKOS tanısını değerlendirmenin yanı sıra hiperprolaktinemi, Cushing hastalığı ve tiroid disfonksiyonu gibi diğer potansiyel nedenleri de hesaba katarak kapsamlı bir klinik değerlendirme gerçekleştirmiştir.

Hem BioBERT hem de SciBERT modellerinin deneysel aşama boyunca metin üzerinde doğru ek açıklamalar üretebildiği ve performanslarında istatistiksel olarak anlamlı bir fark olmadığı görülmüştür. Ayrıca, BioBERT ve SciBERT modellerinin sınıflandırma ve açıklama kalitesi değerlendirmeleri, her iki modelin de özellikle büyük veri kümeleri üzerinde eğitildiğinde %96'nın üzerinde doğruluk elde edebileceğini göstermiştir. Metin sınıflandırma görevlerinde makro ortalama F1 puanı %64 olan SciBERT, BioBERT'ten daha iyi performans göstermiştir. Ortalama olarak, BLEU ve ROUGE gibi ölçütlerde de daha yüksek puanlar elde edilmiştir. Bu sonuç, her iki modelin de sistem tarafından güvenilir bir şekilde açıklama motoru olarak kullanılabileceğini göstermektedir.

Deneysel araştırmalarda diğer tıbbi dil modelleri de değerlendirilmiştir, ancak BioBERT ve SciBERT modelleri açıklama kalitesi ve sınıflandırma doğruluğu açısından en tutarlı ve güvenilir sonuçları üretmiştir. PubMedBERT, ClinicalBERT ve BlueBERT gibi modeller de yüksek doğruluk elde etmiştir; ancak küçük örnek gruplarındaki dengesizlik nedeniyle bu modellerin F1 değerleri daha düşük olmuştur. Sonuç olarak, bulgular bölümünde yalnızca birincil araştırma yapı taşları olarak kabul edilen BioBERT ve SciBERT modelleri vurgulanmıştır.

Önceden eğitilmiş tüm biyomedikal dil modelleri (BioBERT ve SciBERT gibi) ve bilgi erişim motorları (FAISS ve Gemini gibi), modelin İngilizce olarak oluşturulmasının temel gerekçesi olan İngilizce veriler üzerinde eğitilmiştir. Sonuç olarak, sistem bilimsel yayınlarla daha tutarlı, daha rasyonel ve bağlama uygun yanıtlar sunabilmektedir.

Sistemin kullanıcı metnindeki eksik teşhis bileşenlerini belirleme, dinamik olarak bir soruya dönüştürme ve kullanıcının yanıtlarına göre ilerleme yeteneği, en yaratıcı özelliklerinden biridir. Bu da onu geleneksel sohbet robotlarından ayırıyor. Ayrıca,

ROUGE, BLEU ve BERTScore gibi metrikler kullanılarak oluşturulan yanıtların kalite kontrolü sayesinde yöntem artık daha şeffaf ve anlaşılmaktadır. Sistem, geliştirme süreci boyunca kadın hastalıkları ve doğum hekimi tarafından verilen tavsiyelerle yeniden gözden geçirilmiştir. İlk etapta yalnızca PKOS olasılığını tahmin etmeye odaklanan sistem, hekim tarafından ayırıcı tanıların da değerlendirilmesini sağlayacak şekilde yeniden düzenlenmiştir. Bu yöntem sayesinde sistem, yalnızca pozitif olasılığı artırmakla kalmaz, aynı zamanda diğer olasılıkları dışlayarak daha doğru klinik yönlendirmeler sunabilir hale geldi.

Bulgular, DDİ ve BDM tabanlı sistemlerin PKOS gibi karmaşık tanılı sendromların değerlendirilmesinde etkili, güvenilir ve zaman kazandıran klinik araçlar olarak kullanılabileceğini göstermektedir. Bununla birlikte, sistemin klinik uygulamada yaygın olarak kullanılabilmesi için daha büyük hasta grupları üzerinde test edilmesi, etik denetimlerden geçmesi ve kullanıcı eğitimi alması büyük önem taşımaktadır.

Bu çalışma, hem akademik hem de klinik uygulamalar açısından literatüre önemli katkılar sağlamaktadır. Bilgisayar mühendisliği, biyomedikal ve tıbbın kesiştiği noktada disiplinler arası bir çabanın sonucudur. Oluşturulan teknolojinin gelecekte çeşitli hastalık kategorilerine uyarlanması, daha kapsamlı ve güvenilir karar destek çözümleriyle sonuçlanabilmektedir.

## 7. KAYNAKLAR

Abid, A., Abdalla, A., Abid, A., Khan, D., Alfozan, A., and Zou, J. 2019. Gradio: Hassle-free sharing and testing of MÖ models in the wild. arXiv preprint arXiv:1906.02569.

Adla, Y.A.A., Raydan, D.G., Charaf, M.Z.J., Saad, R.A., Nasreddine, J. and Diab, M.O. 2021, October. Automated detection of polycystic ovary syndrome using machine learning techniques. In 2021 Sixth international conference on advances in biomedical engineering (ICABME) (pp. 208-212). IEEE.

Afifah, S., Mudzakir, A., and Nandiyanto, A. B. D. 2022. How to calculate paired sample t-test using SPSS software: From step-by-step processing for users to the practical examples in the analysis of the effect of application anti-fire bamboo teaching materials on student learning outcomes. Indonesian Journal of Teaching in Science, 2(1), 81-92.

Azziz, R., Carmina, E., Chen, Z., Dunaif, A., Laven, J. S., Legro, R. S., Lizneva, D., Natterson-Horowtiz, B., Teede, H. J. and Yildiz, B. O. 2016. Polycystic ovary syndrome. Nature reviews Disease primers, 2(1), 1-18.

Barrera, F.J., Brown, E.D., Rojo, A., Obeso, J., Plata, H., Lincango, E.P., Terry, N., Rodríguez-Gutiérrez, R., Hall, J.E. and Shekhar, S. 2023. Application of machine learning and artificial intelligence in the diagnosis and classification of polycystic ovarian syndrome: a systematic review. Frontiers in Endocrinology, 14, p.1106625.

Beltagy, I., Lo, K., and Cohan, A. 2019. SciBERT: A pretrained language model for scientific text. arXiv preprint arXiv:1903.10676.

Bharati, S., Podder, P. and Mondal, M.R.H. 2020, June. Diagnosis of polycystic ovary syndrome using machine learning algorithms. In 2020 IEEE region 10 symposium (TENSYMP) (pp. 1486-1489). IEEE.

Bhat, S.A. 2021. Detection of polycystic ovary syndrome using machine learning algorithms (Doctoral dissertation, Dublin, National College of Ireland).

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Chang, Y., Huang, S. S., Hsu, W. Y., and Liu, Y. C. 2024. Evaluating Chatbots in Psychiatry: Rasch-Based Insights into Clinical Knowledge and Reasoning. medRxiv, 2024-12.

Choi, R.Y., Coyner, A.S., Kalpathy-Cramer, J., Chiang, M.F. and Campbell, J.P. 2020. Introduction to machine learning, neural networks, and deep learning. Translational vision science & technology, 9(2), pp.14-14.

Cumhuriyet, 2023. Doktor yapay zekanın yazdığı notlar, gerçeğinden ayırt edilemiyor, https://www.cumhuriyet.com.tr/bilim-teknoloji/doktor-yapay-zekanin-yazdigi-notlar-gerceginden-ayirt-edilemiyor-2146856#google_vignette [Son erişim tarihi: 13.03.2025].

Çelik, Ö., ve Koç, B. C. 2021. TF-IDF, Word2vec ve Fasttext vektör model yöntemleri ile Türkçe haber metinlerinin sınıflandırılması. *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi*, 23(67), 121-127.

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. 2019, June. 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)* (pp. 4171-4186).

DoktorTakvimi, 2023. Tıp Sektörü İçin En Kullanışlı YZ Uygulamaları, pro.doktortakvimi.com [Son erişim tarihi: 13.03.2025].

Doshi K. 2021. Foundations of NLP Explained — BLEU Score and WER Metrics. Towards Data Science. https://towardsdatascience.com/foundations-of-nlp-explained-bleu-score-and-wer-metrics-1a5ba06d812b [Son erişim tarihi: 14.05.2025].

Emara, H. M., El-Shafai, W., Soliman, N. F., Algarni, A. D., Alkanhel, R., and Abd El-Samie, F. E. 2025. A stacked learning framework for accurate classification of polycystic ovary syndrome with advanced data balancing and feature selection techniques. *Frontiers in Physiology*, 16, 1435036.

Escobar-Morreale, H.F., 2018. Polycystic ovary syndrome: definition, aetiology, diagnosis and treatment. *Nature Reviews Endocrinology*, 14(5), pp.270-284.

Fawcett, T. 2006. An introduction to ROC analysis. *Pattern recognition letters*, 27(8), 861-874.

FutureHouse, 2025. Launching FutureHouse Platform YZ Agents. https://www.futurehouse.org/research-YSAouncements/launching-futurehouse-platform-YZ-agents [Son erişim tarihi: 16.05.2025].

Gabrilovich, E., and Markovitch, S. 2007, January. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *IJcAI* (Vol. 7, No. 1, pp. 1606-1611).

GeeksforGeeks, 2024. FastText: Working and Implementation. https://www.geeksforgeeks.org/fasttext-working-and-implementation/ [Son erişim tarihi: 31.05.2025].

Goldberg, Y. 2017. *Neural network methods in natural language processing*. Morgan & Claypool Publishers.

Goodfellow, I., Bengio, Y., and Courville, A. 2016. Deep Learning. MIT Press. https://www.deeplearningbook.org/ [Son erişim tarihi: 10.05.2025].

Google DeepMind., 2023. Gemini: Our largest and most capable YZ model. Retrieved from https://deepmind.google/technologies/gemini/ [Son erişim tarihi: 29.05.2025].

Gradio. 2024. Gradio: Build and share delightful machine learning apps. https://www.gradio.app/ [Son erişim tarihi: 10.05.2025].

Hanley, J. A., and McNeil, B. J. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1), 29-36.

Hassan, M.M. and Mirza, T. 2020. Comparative analysis of machine learning algorithms in diagnosis of polycystic ovarian syndrome. *Int. J. Comput. Appl*, 975, p.8887.

Hinton, G. E., and Salakhutdinov, R. R. 2006. Reducing the dimensionality of data with neural networks. *science*, *313*(5786), 504-507.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., and Müller, H. 2019. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4), e1312. https://doi.org/10.1002/widm.1312

InteligenAI, 2025. InteligenAI LinkedIn Sayfası. https://www.linkedin.com/company/inteligenai/ [Son erişim tarihi: 16.05.2025].

Johnson, J., Douze, M., and Jégou, H. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3), 535-547.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv*:1607.01759.

Kim, T. K., 2015. T test as a parametric statistic. *Korean journal of anesthesiology*, 68(6), 540-546.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234-1240.

Lin, C. Y. 2004, July. Rouge: A package for automatic evaluation of summaries. *In Text summarization branches out* (pp. 74-81).

Logunova, I. and Bolgurtseva, O. 2023. Word2Vec: Why Do We Need Word Representations?. https://serokell.io/blog/word2vec [Son erişim tarihi: 19.03.2025].

Melson, E., Rocha, T. P., Veen, R. J., Abdi, L., McDonnell, T., Tandl, V., and Arlt, W. 2023, October. Unsupervised steroid metabolome cluster analysis to dissect androgen excess and metabolic dysfunction in 488 women with polycystic ovary syndrome-results from the prospective DAISy-PCOS study. *In Endocrine Abstracts* (Vol. 94). Bioscientifica. doi: 10.1530/endoabs.94.OC4.1

Microsoft, Visual Studio Code. https://code.visualstudio.com/ [Son erişim tarihi: 13.03.2025].

Miesle, P. 2025. What is Cosine Similarity: A Comprehensive Guide. https://www.datastax.com/guides/what-is-cosine-similarity [Son erişim tarihi: 16.04.2025].

Nayab, H. 2024. Embedding Models Explained: A Guide to NLP's Core Technology. https://medium.com/@nay1228/embedding-models-a-comprehensive-guide-for-beginners-to-experts-0cfc11d449f1 [Son erişim tarihi: 18.03.2025].

Neto, C., Silva, M., Fernandes, M., Ferreira, D. and Machado, J. 2021, February. Prediction models for Polycystic Ovary Syndrome using data mining. *In International Conference on Advances in Digital Science* (pp. 210-221). Cham: Springer International Publishing.

Nsugbe, E. 2023. An artificial intelligence-based decision support system for early diagnosis of polycystic ovaries syndrome. *Healthcare Analytics*, 3, p.100164.

Panjwani, B., Yadav, J., Mohan, V., Agarwal, N., and Agarwal, S. 2025. Optimized Machine Learning for the Early Detection of Polycystic Ovary Syndrome in Women. *Sensors*, 25(4), 1166. https://doi.org/10.3390/s25041166

Papineni, K., Roukos, S., Ward, T., and Zhu, W. J., 2002, July. Bleu: a method for automatic evaluation of machine translation. *In Proceedings of the 40th YSAual meeting of the Association for Computational Linguistics* (pp. 311-318).

Pennington, J., Socher, R., and MYSAing, C. D. 2014. Glove: Global vectors for word representation. *In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Powers, D. M., 2020. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:*2010.16061.

Prathibanandhi, J., and Vimala, G. A. G. 2024. Diagnosis of Polycystic Ovarian Syndrome with the Implementation of Deep Learning Models. *In 2024 Third International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)* (pp. 1-7). IEEE.

Rabiner, L. R. 2002. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257-286.

Raghudeep, 2019. Review: BioBERT paper. https://medium.com/@raghudeep/biobert-insights-b4c66fde8fa7 [Son erişim tarihi: 01.06.2025].

Reiter, E. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3), 393-401.

Riestenberg, C., Jagasia, A., Markovic, D., Buyalos, R.P. and Azziz, R. 2022. Health care-related economic burden of polycystic ovary syndrome in the United States: pregnancy-related and long-term health consequences. *The Journal of Clinical Endocrinology & Metabolism*, 107(2), pp.575-585.

Rietveld, T., and van Hout, R. 2017. The paired t test and beyond: Recommendations for testing the CKDKM tendencies of two paired samples in research on speech, language and hearing pathology. *Journal of communication disorders*, 69, 44-57.

Rossum, V. 2009. Python 3 reference manual.

Russell, S., and Norvig, P. 2021. *Artificial Intelligence*: A Modern Approach (4th ed.). Pearson.

Saito, T., and Rehmsmeier, M. 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3), e0118432.

SBERT.net, 2025. SentenceTransformers Documentation. https://sbert.net/ [Son erişim tarihi: 18.03.2025].

Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, *34*(1), 1-47.

Silva, I.S., Ferreira, C.N., Costa, L.B.X., Sóter, M.O., Carvalho, L.M.L., de C. Albuquerque, J., Sales, M.F., Candido, A.L., Reis, F.M., Veloso, A.A. and Gomes, K.B. 2022. Polycystic ovary syndrome: Clinical and laboratory variables related to new phenotypes using machine-learning models. *Journal of Endocrinological Investigation*, pp.1-9.

Sleightholm, E. 2025. Marqo. Introduction to Cümle Dönüştürücüleri. https://www.marqo.YZ/course/introduction-to-sentence-transformers [Son erişim tarihi: 31.05.2025].

Smet, M. E., and McLennan, A. 2018. Rotterdam criteria, the end. *Australasian Journal of Ultrasound in Medicine*, 21(2), 59-60. doi: 10.1002/ajum.12096

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

Sorin, V., Brin, D., Barash, Y., Konen, E., Charney, A., Nadkarni, G., and Klang, E. 2024. Large Language Models and Empathy: Systematic Review. *Journal of Medical Internet Research*, 26, e52597. doi: 10.2196/52597

Suri, G., Slater, L. R., Ziaee, A., and Nguyen, M. 2024. Do large language models show decision heuristics similar to humans? A case study using GPT-3.5. *Journal of Experimental Psychology: General*, *153*(4), 1066.

Sütçü, C. S. 12Punto., 2024. William Sealy Gosset ve Student'in hikayesi. https://12punto.com.tr/yazarlar/cem-sefa-sutcu/william-sealy-gosset-ve-studentin-hikayesi-45965 [Son erişim tarihi: 28.05.2025].

Teede, H.J., Tay, C.T., Laven, J.J., Dokras, A., Moran, L.J., Piltonen, T.T., Costello, M.F., Boivin, J., Redman, L.M., Boyle, J.A. and Norman, R.J. 2023. Recommendations from the 2023 international evidence-based guideline for the assessment and management of polycystic ovary syndrome. *European journal of endocrinology*, 189(2), pp.G43-G64. https://doi.org/10.1093/ejendo/lvad096

Topol, E. J. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1), 44-56.

Trae, 2025. What is Trae?. https://docs.trae.YZ/docs/what-is-trae?_lang=en [Son erişim tarihi: 08.04.2025].

Tsang, S. 2024. Brief Review — BERTScore: Evaluating Text Generation with BERT. https://sh-tsang.medium.com/brief-review-bertscore-evaluating-text-generation-with-bert-0bc5fc889d7b [Son erişim tarihi: 01.06.2025].

TÜBİTAK, 2018. https://www.truba.gov.tr [Son erişim tarihi: 13.03.2025].

Wilkerson, S. 2008. Application of the Paired t-test. *XULAneXUS*, *5*(1), 7.

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

Young, T., Hazarika, D., Poria, S., and Cambria, E. 2018. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, *13*(3), 55-75.

Zi-Jiang, C. 2023. Novel Subtypes of Polycystic Ovary Syndrome, https://clinicaltrials.gov/study/NCT06124391 [Son erişim tarihi: 13.03.2025].

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhang, X., Liang, B., Zhang, J., Hao, X., Xu, X., Chang, H.M., Leung, P.C. and Tan, J. 2021. Raman spectroscopy of follicular fluid and plasma with machine-learning algorithms for polycystic ovary syndrome screening. *Molecular and cellular endocrinology*, 523, p.111-139. https://doi.org/10.1016/j.mce.2020.111139

## 8. EKLER

### EK-1: Soruların Yer Aldığı JSON


```
[
  {
    "question": "Does the patient have irregular menstrual cycles?",
    "expected_answer": "Yes",
    "evidence_keywords": ["irregular menstrual cycles", "oligomenorrhea", "amenorrhea"],
    "positive_when": "yes"
  },
  {
    "question": "Please enter the Ferriman-Gallwey score (0–36) indicating the level of body hair:",
    "expected_answer": "numeric",
    "evidence_keywords": ["hirsutism", "facial hair", "body hair", "Ferriman-Gallwey"],
    "answer_type": "scale",
    "scale_min": 0,
    "scale_max": 36
  },
  {
    "question": "Have hormone tests been performed?",
    "expected_answer": "Yes",
    "evidence_keywords": ["testosterone", "hormone test", "androgen levels", "17-hydroxyprogesterone"],
    "positive_when": "yes"
  },
  {
```

"question": "Were the hormone test results within normal ranges?",

"expected_answer": "Yes",

"evidence_keywords": ["normal hormone levels", "testosterone normal", "androgen normal"],

"positive_when": "yes"

},

{

"question": "Was a transvaginal ultrasound performed?",

"expected_answer": "Yes",

"evidence_keywords": ["transvaginal ultrasound", "TVUS", "ultrasound shows"],

"positive_when": "yes"

},

{

"question": "Is polycystic ovary morphology present?",

"expected_answer": "Yes",

"evidence_keywords": ["polycystic ovaries", "increased follicle count", "ovarian volume"],

"positive_when": "yes"

},

{

"question": "Was Anti-Müllerian hormone (AMH) evaluated?",

"expected_answer": "Yes",

"evidence_keywords": ["AMH", "Anti-Müllerian hormone", "ovarian reserve"],

"positive_when": "yes"

},

{

"question": "Are there physical features such as moon face, truncal obesity, or purple abdominal striae?",

"expected_answer": "No",

"evidence_keywords": ["moon face", "striae", "buffalo hump", "central obesity", "cushing"],

"positive_when": "no"

},

{

"question": "Has the patient experienced fatigue, weight gain, or cold intolerance recently?",

"expected_answer": "No",

"evidence_keywords": ["tsh", "hypothyroidism", "weight gain", "cold intolerance"],

"positive_when": "no"

},

{

"question": "Is there early onset of pubarche or elevated 17-OHP?",

"expected_answer": "No",

"evidence_keywords": ["17-OHP", "early puberty", "nccah", "21-hydroxylase deficiency"],

"positive_when": "no"

},

{

"question": "Is there rapid onset of hirsutism or deepening of the voice?",

"expected_answer": "No",

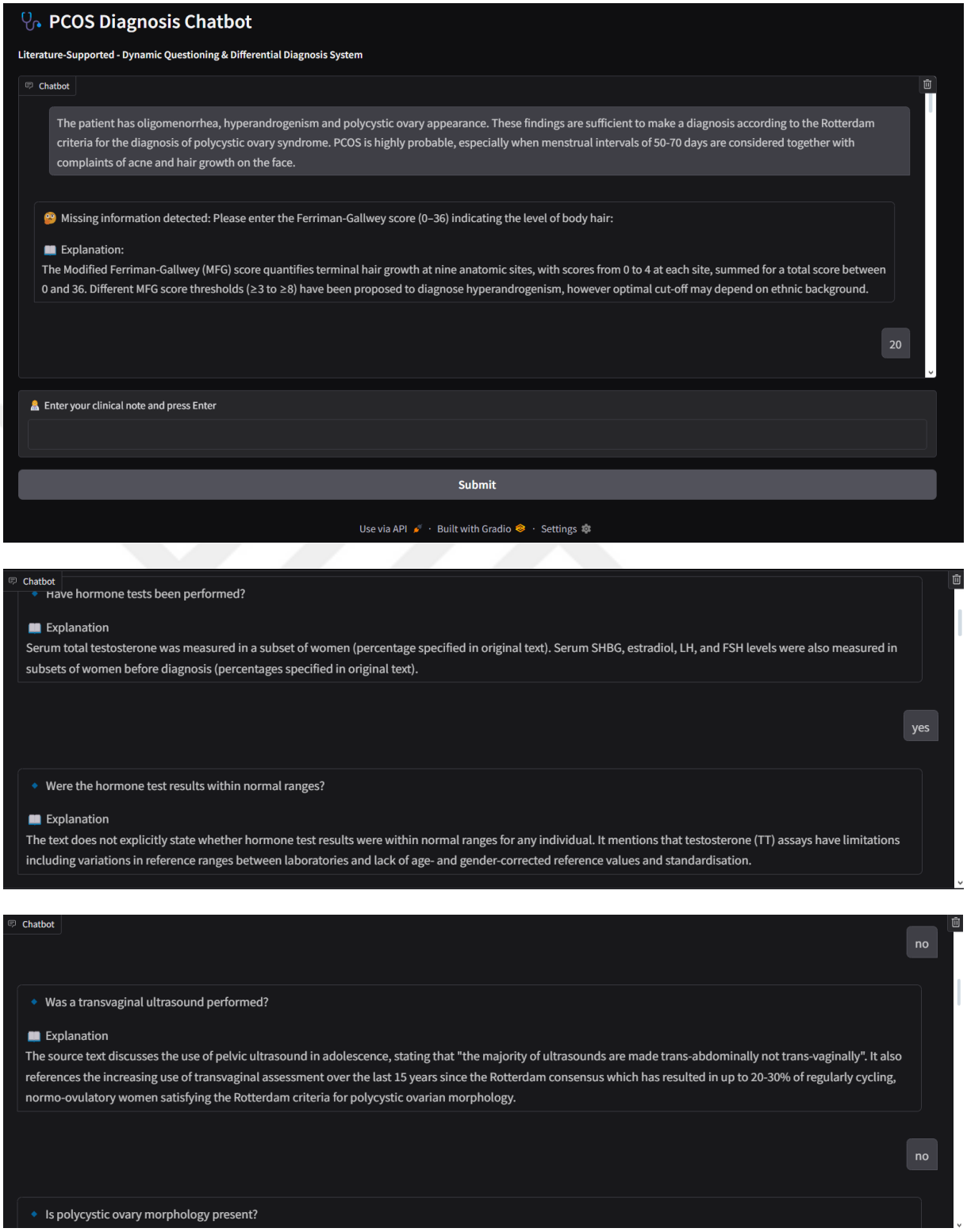
"evidence_keywords": ["rapid hirsutism", "deep voice", "balding", "virilization", "androgen-secreting tumor"],

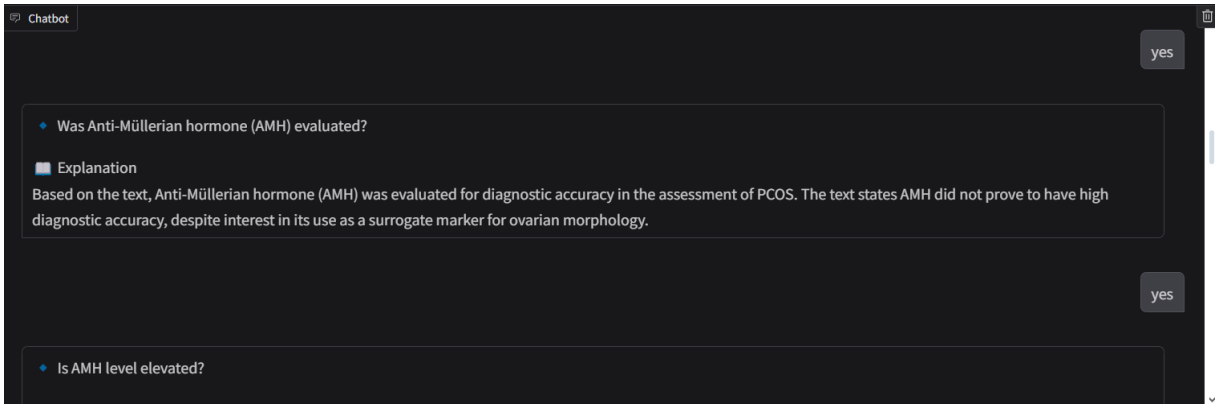
"positive_when": "no"

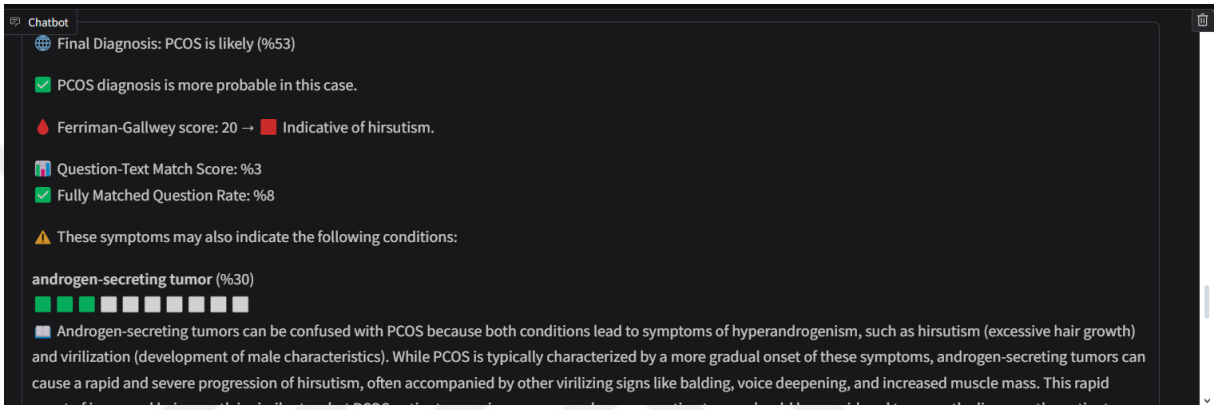
},

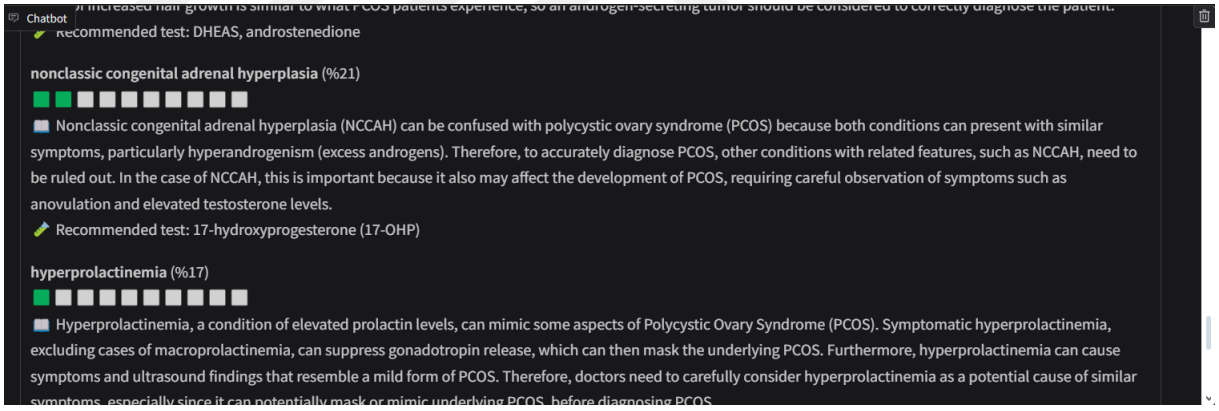
```
  {
    "question": "Is there galactorrhea or has the patient reported breast discharge?",
    "expected_answer": "No",
    "evidence_keywords": ["prolactin", "galactorrhea", "serum prolactin"],
    "positive_when": "no"
  }
]
```

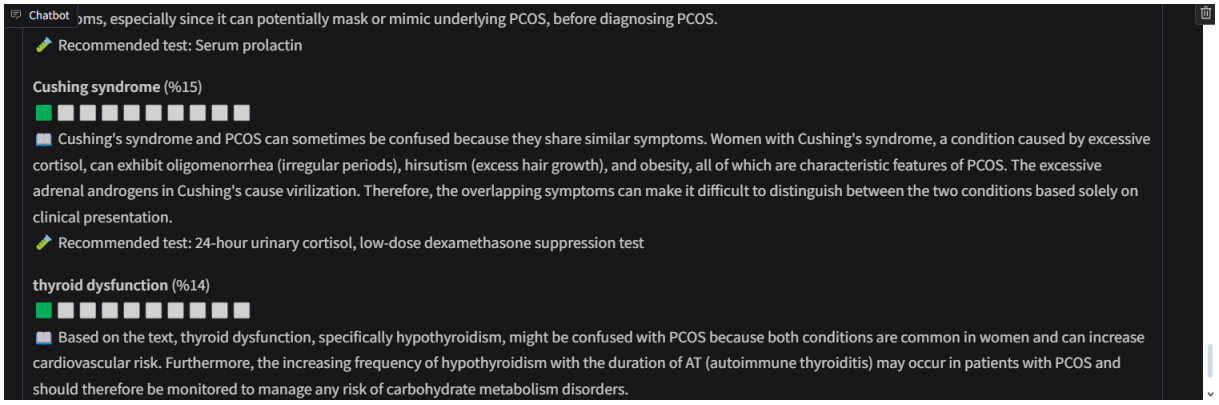
## EK-2: Sistemin Ekran Görüntüsü




Chatbot
yes
Was Anti-Müllerian hormone (AMH) evaluated?
Explanation
Based on the text, Anti-Müllerian hormone (AMH) was evaluated for diagnostic accuracy in the assessment of PCOS. The text states AMH did not prove to have high diagnostic accuracy, despite interest in its use as a surrogate marker for ovarian morphology.
yes
Is AMH level elevated?


Chatbot
Final Diagnosis: PCOS is likely (%53)
PCOS diagnosis is more probable in this case.
Ferriman-Gallwey score: 20 → Indicative of hirsutism.
Question-Text Match Score: %3
Fully Matched Question Rate: %8
These symptoms may also indicate the following conditions:
androgen-secreting tumor (%30)
Androgen-secreting tumors can be confused with PCOS because both conditions lead to symptoms of hyperandrogenism, such as hirsutism (excessive hair growth) and virilization (development of male characteristics). While PCOS is typically characterized by a more gradual onset of these symptoms, androgen-secreting tumors can cause a rapid and severe progression of hirsutism, often accompanied by other virilizing signs like balding, voice deepening, and increased muscle mass. This rapid


Chatbot
Recommended test: DHEAS, androstenedione
nonclassic congenital adrenal hyperplasia (%21)
Nonclassic congenital adrenal hyperplasia (NCCAH) can be confused with polycystic ovary syndrome (PCOS) because both conditions can present with similar symptoms, particularly hyperandrogenism (excess androgens). Therefore, to accurately diagnose PCOS, other conditions with related features, such as NCCAH, need to be ruled out. In the case of NCCAH, this is important because it also may affect the development of PCOS, requiring careful observation of symptoms such as anovulation and elevated testosterone levels.
Recommended test: 17-hydroxyprogesterone (17-OHP)
hyperprolactinemia (%17)
Hyperprolactinemia, a condition of elevated prolactin levels, can mimic some aspects of Polycystic Ovary Syndrome (PCOS). Symptomatic hyperprolactinemia, excluding cases of macroprolactinemia, can suppress gonadotropin release, which can then mask the underlying PCOS. Furthermore, hyperprolactinemia can cause symptoms and ultrasound findings that resemble a mild form of PCOS. Therefore, doctors need to carefully consider hyperprolactinemia as a potential cause of similar


Chatbot
Recommended test: Serum prolactin
Cushing syndrome (%15)
Cushing's syndrome and PCOS can sometimes be confused because they share similar symptoms. Women with Cushing's syndrome, a condition caused by excessive cortisol, can exhibit oligomenorrhea (irregular periods), hirsutism (excess hair growth), and obesity, all of which are characteristic features of PCOS. The excessive adrenal androgens in Cushing's cause virilization. Therefore, the overlapping symptoms can make it difficult to distinguish between the two conditions based solely on clinical presentation.
Recommended test: 24-hour urinary cortisol, low-dose dexamethasone suppression test
thyroid dysfunction (%14)
Based on the text, thyroid dysfunction, specifically hypothyroidism, might be confused with PCOS because both conditions are common in women and can increase cardiovascular risk. Furthermore, the increasing frequency of hypothyroidism with the duration of AT (autoimmune thyroiditis) may occur in patients with PCOS and should therefore be monitored to manage any risk of carbohydrate metabolism disorders.

**EK-3: Sistem Algoritması**

```
ALGORITMA PKOS_TANI_DESTEK_SISTEMI

Girdi:
   - PDF ve PubMed makaleleri
   - Kullanıcıdan alınan klinik metin
Çıktı:
   - PKOS tanı değerlendirmesi
   - Açıklamalar (BDM destekli)
   - Ayırıcı tanı ihtimalleri

BAŞLA

1. VERİ ÖN İŞLEME
   FOR her veri kaynağı (PDF, PubMed CSV) DO
      metin ← METIN_ÇIKAR(veri)
      temiz_metin ← METNI_TEMIZLE(metin)
      json_aktarımı ← BAŞLIK_BAZLI_JSON_YAPILANDIR(temiz_metin)
      VERİ_SETİNE_EKLE(json_aktarımı)
   END FOR

   birleşik_veri ← TÜM_JSONLARI_BİRLEŞTİR()
   temiz_veri ← HER_METNI_TEMIZLE(birleşik_veri)
   JSON_KAYDET(temiz_veri, "text_only_dataset.json")

2. EMBEDDING VE FAISS INDEKSİ
   corpus ← METİNLERİ_AL("text_only_dataset.json")
   embed_model ← YÜKLE("sentence-transformers/all-MiniLM-L6-v2")
   vektörler ← embed_model.encode(corpus)
   faiss_index ← FAISS.İNDEKS_OLUŞTUR(vektörler)
   KAYDET_FAISS_INDEX(faiss_index, "retrieval_index.faiss")

3. OTOMATİK ETİKETLEME

   // Amaç: Etiketsiz literatür verilerinden, konusuna göre sınıf tahmini yapmak

   FOR her belge IN temiz_veri DO
      başlık ← belge["başlık"]
      metin ← belge["text"]

      IF başlık VEYA metin içerisinde:
         "tanı", "semptom", "klinik özellik" geçiyorsa → etiket ← "PKOS_TANI"
      ELSE IF "epidemiyoloji", "prevalans", "yaygınlık" varsa → etiket ←
"EPIDEMIYOLOJI"
      ELSE IF "tedavi", "ilaç", "progesteron", "doz" varsa → etiket ← "TEDAVI"
      ELSE IF "ayırıcı tanı", "diferansiyel" varsa → etiket ← "AYIRICI_TANI"
      ELSE → etiket ← "GENEL_BILGI"  // Varsayılan veya belirsiz durumlar
```

```
    BELGEYE_EKLE("label", etiket)
  END FOR

4. BERT TABANLI METİN SINIFLANDIRMA
  veriler ← CHUNKLA_VE_ETİKETLE(temiz_veri, max_length=500)
  dataset ← TransformersDataset(veriler)
  train, val, test ← SPLIT(dataset, oran=80/10/10)

  model ← SciBERT veya BioBERT
  tokenizer ← MODEL_TOKENIZERINI_YÜKLE(model)

  train_encoded ← TOKENIZE(train, tokenizer)
  model_eğit ← TransformersTrainer(model, train_encoded, val_encoded)
  EĞİTİMİ_BAŞLAT(model_eğit)
  EĞİTİLMİŞ_MODELİ_KAYDET()

5. GEMINI DESTEKLİ AÇIKLAMA ÜRETİMİ
  FOR her test_chunk IN test DO
    benzer_paragraf ← FAISS_ARA(test_chunk, faiss_index)
    explanation ← GEMINI_YANITLA(soru, benzer_paragraf)
    değerlendirme ← ROUGE / BERTScore ile karşılaştır
    KAYDET_LOG(explanation, değerlendirme)
  END FOR

6. DİNAMİK CHATBOT DESTEĞİ
  kullanıcı_metni ← INPUT()
  eksik_bilgiler ← SORULARLA_TESPİT_ET(kullanıcı_metni)
  hormon_analizi ← HORMON_KELİMELERİ_ARA(kullanıcı_metni)

  WHILE eksik_bilgiler VARSA DO
    soru ← SIRADAKİ(eksik_bilgiler)
    kontekst ← FAISS_ARA(soru, faiss_index)
    açıklama ← GEMINI_YANITLA(soru, kontekst)
    kullanıcı_cevabı ← INPUT(soru)
    CEVABI_KAYDET()
  END WHILE

  final_skor ← CEVAPLARA_GÖRE_TANI_HESAPLA()
  AYIRICI_TANI_SKORLARINI_HESAPLA(kullanıcı_metni)
  ÖNERİLERİ_VE_SONUCU_GÖSTER()

BİTİR
```

# ÖZGEÇMİŞ

**Jacklyn Günce KAYA**

## ÖĞRENİM BİLGİLERİ

| | |
|---|---|
| Yüksek Lisans<br>2023-2025 | Akdeniz Üniversitesi<br>Fen Bilimleri Enstitüsü, Elektrik-Elektronik Mühendisliği Anabilim Dalı, Antalya |
| Lisans<br>2019-2023 | Ege Üniversitesi<br>Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Antalya |

## MESLEKİ VE İDARİ GÖREVLER

| | |
|---|---|
| Proje Çalışanı<br>2025-2025 | Akdeniz Üniversitesi<br>Yabancı Diller Bölümü, Antalya |
| Stajyer Yazılım Mühendisi<br>2023-2023 | Yapı Kredi Teknoloji<br>Gebze, Kocaeli |
| Stajyer Yazılım Mühendisi<br>2022-2022 | OBSS Teknoloji<br>İstanbul |

## ESERLER

### Ulusal bilimsel toplantılarda sunulan ve bildiri kitaplarında basılan bildiriler

1- Kaya, J.G., Yeşil, L.E., Kaçmaz, B., Ünalır, M.O., Sezer, E., Güven, Z.A. & Uslu, E.E. (2023). Development of Text Analysis Framework for Radiology Reports. Proceeding of 14th Turkish Congress of Medical Informatics Association, 16-18 March 2023, Balçova-İzmir-Turkey, 10-18.