Ekg Veriseti Anomali Tespiti

Github

Yusuf Üzeyir Kaya
Bilişim Sistemleri Mühendisliği
Kocaeli Üniversitesi
Kocaeli,Türkiye
ys.kaya1400@gmail.com

Buğra Menteş
Bilişim Sistemleri Mühendisliği
Kocaeli Üniversitesi
Kocaeli,Türkiye
bugramentes57@gmail.com

***Özet*-- Bu proje, Elektrokardiyogram (EKG) verileri üzerinde gerçek zamanlı anomali tespiti yapmayı amaçlayan bir büyük veri analiz uygulamasıdır. Kalbin elektriksel aktivitesini ölçen EKG, ritim bozuklukları ve diğer kardiyak problemleri belirlemek için hayati bir araçtır. Ancak, büyük miktarda veri ile çalışıldığında manuel analiz hem zaman alıcı hem de hata yapma olasılığı yüksektir. Bu projede, büyük veri teknolojilerinden Apache Spark ve Apache Kafka kullanılarak ölçeklenebilir ve yüksek performanslı bir analiz sistemi geliştirilmiştir.**

Abstract-- This project is a big data analysis application aimed at real-time anomaly detection on Electrocardiogram (ECG) data. ECG, which measures the electrical activity of the heart, is a vital tool for identifying arrhythmias and other cardiac issues. However, when dealing with large volumes of data, manual analysis is both time-consuming and prone to errors. In this project, scalable and high-performance analysis systems have been developed using big data technologies such as Apache Spark and Apache Kafka. Anahtar Kelimeler— Spark, Kafka, Anomali, Pytorch, Veri Seti, Tensorflow, Ekg, Ecg

I. GİRİŞ

Elektrokardiyogram (EKG), kalbin elektriksel aktivitesini kaydederek kardiyovasküler sistemin sağlık durumu hakkında önemli bilgiler sağlayan bir yöntemdir. Günümüzde, artan veri miktarı ve sağlık teknolojilerindeki gelişmeler, EKG verilerinin işlenmesi ve analizi için otomatik ve etkili çözümlerin geliştirilmesini zorunlu kılmaktadır.

Bu proje, EKG verileri üzerinde anomali tespiti gerçekleştirmeyi hedeflemektedir. Anomaliler, genellikle kardiyak ritim bozuklukları veya ciddi sağlık sorunlarının habercisi olabileceğinden, erken tespit büyük önem taşımaktadır. Bu bağlamda, projede büyük veri teknolojileri olan **Apache Spark** ve **Apache Kafka** entegre bir şekilde kullanılmaktadır. Bu teknolojiler, EKG verilerinin gerçek zamanlı işlenmesini ve analitik sonuçların hızlı bir şekilde elde edilmesini sağlayarak, sağlık sistemleri için kritik bir çözüm sunmaktadır.

Projenin amacı, gerçek zamanlı veri akışından anomalileri tespit etmek için yapay zeka destekli modeller geliştirmek ve bu modelleri büyük veri altyapısıyla birleştirmektir. Özellikle, Spark ile paralel işleme ve Kafka ile veri akışı yönetimi, projeye ölçeklenebilirlik ve esneklik kazandırmaktadır. Bu proje, sadece teknik bir uygulama olmanın ötesinde, sağlık alanında daha hızlı ve doğru teşhisler yapılmasına katkıda bulunmayı hedeflemektedir.

Bu raporda, projenin metodolojisi, kullanılan teknolojiler, elde edilen sonuçlar ve sağlık sektöründe olası etkileri detaylı bir şekilde ele alınacaktır.

II. SİSTEM MİMARİSİ

A. *Kafka Producer*

Kafka Producer, sistemin veri üretimini sağlayan ve verilerin gerçek zamanlı işlenmesi için başlangıç noktasını oluşturan bir bileşendir. Projemizde Kafka Producer, EKG sinyallerini veri kaynağından alarak, belirlenen formatta Kafka'nın ilgili topic'lerine gönderir. Her bir EKG sinyali, belirli bir formatta işlenerek Kafka'ya aktarıldığından, verinin düzenli ve hızlı bir şekilde iletilmesi sağlanır. Producer'ın yüksek hızlı veri aktarımı yeteneği, EKG verilerinin akıcı bir şekilde sisteme dahil olmasını mümkün kılar. Bu sayede, Spark Streaming gibi gerçek zamanlı işleme bileşenleri anlık olarak bu veriler üzerinde analiz yapabilir ve anomali tespit süreçleri kesintisiz şekilde ilerler.

B. *Kafka Consumer*

Kafka Consumer, projemizde işlenen veriyi Kafka topic'lerinden alarak analiz etmek veya başka bir sisteme aktarmak için kullanılan bir bileşendir. EKG sinyalleri üzerindeki anomali tespit sonuçları, Kafka Consumer tarafından belirlenen topic'ten alınır ve anomali tespit edilen durumlarda aksiyon alınmasını sağlayacak süreçleri tetikler. Örneğin, bir anomali tespit edildiğinde, bu bilgi bir uyarı sistemi iletilerek kullanıcıya bildirilebilir. Kafka Consumer, aynı zamanda verileri işleyerek farklı sistemlere entegrasyon sağlayabilir. Consumer'ın sürekli veri akışını destekleyen yapısı sayesinde, tespit edilen anomaliler hızla değerlendirilir ve sistemin gerçek zamanlı çalışması sağlanır. Kafka'nın ölçeklenebilir ve esnek mimarisi, bu sürecin verimli bir şekilde işlemesine olanak tanır.

C. *Spark Streaming*

Spark Streaming, projemizde gerçek zamanlı veri işleme ve analiz süreçlerinin merkezinde yer almaktadır. Kafka'dan alınan EKG verileri, Spark Streaming

aracılığıyla işlenir ve anomali tespiti algoritmaları bu veri akışı üzerinde uygulanır. Spark Streaming, veriyi küçük mikro partisyonlara bölerek her bir parçayı eş zamanlı olarak işler. Bu işlem, anormal sinyalleri tespit etme ve normal sinyallerden ayırma görevini üstlenir. Özellikle sürekli gelen verilerle çalışırken Spark Streaming, anlık analiz yaparak anomalilerin belirlenmesini sağlar. Apache Spark'ın güçlü paralel işleme yetenekleri, büyük veri kümelerinin hızlı bir şekilde analiz edilmesini mümkün kılar. Projemizde, Spark Streaming'in performansı sayesinde EKG verileri üzerinde yapılan anomali tespitleri doğru ve kesintisiz bir şekilde gerçekleştirilmiştir.

D. *Makine Öğrenmesi Modelleri*

1) *Autoencoder*

Autoencoder modeli, veri setindeki anormalliklerin tespiti için kullanılan temel makine öğrenmesi yöntemlerinden biridir. Bu model, verileri sıkıştırarak bir kodlama katmanına dönüştürür ve ardından bu kodlama katmanından yeniden orijinal veriyi oluşturmayı öğrenir. Projemizde, EKG sinyallerinin normal ve anormal olarak ayrılabilmesi için autoencoder kullanılmıştır. Model, normal verileri öğrenerek düşük yeniden oluşturma hatası üretirken, anormal veriler için yüksek yeniden oluşturma hatası üretir. Bu hatalar, anomali tespiti için bir metrik olarak kullanılmış ve belirli bir eşik değerinin üzerindeki hatalar "anomali" olarak sınıflandırılmıştır..

2) *VERI SETI DAĞILIMI*

Veri seti analizi, normal ve anormal verilerin dağılımlarını anlamak için kritik bir adımdır. Projede, veri setinin %58.4'ünün normal, %41.6'sının ise anormal verilerden oluştuğu tespit edilmiştir. Bu oranlar, EKG sinyallerinin doğal yapısını ve anormal durumların tespit edilme sıklığını göstermektedir. Pie chart gibi görselleştirme araçları kullanılarak bu dağılım açık bir şekilde ifade edilmiştir.

E. *Anomali Tespit Sonuçları*

Anomali tespit sonuçları, modelin başarımını değerlendirmek için görselleştirilmiştir. Normal ve anormal veriler için oluşturulan grafikler, modelin doğru sınıflandırma oranını ve hata dağılımlarını göstermektedir. Özellikle, modelin yüksek doğruluk oranı, geliştirilmiş sistemin güvenilirliğini kanıtlamaktadır.

1) *Isı Haritaları (Heatmap)*

Projedeki ilişkisel analizlerde, ısı haritaları kullanılmıştır. Bu haritalar, veriler arasındaki korelasyonları ve modelin yeniden yapılandırma hatalarının dağılımını göstermiştir. Isı haritaları, özellikle anormalliklerin belirli veri noktalarında yoğunlaştığını görsel olarak ifade ederek tespit sürecini desteklemiştir.

2) *Isı Haritaları (Heatmap)*:

Özellikle korelasyon analizlerini daha iyi anlayabilmek için ısı haritaları kullanılacaktır. Isı haritaları, veri setindeki özellikler arasındaki ilişkilerin görsel olarak gösterilmesini sağlar, bu da kullanıcıların veri özellikleri arasındaki bağlantıları daha rahat görmelerine yardımcı olur.

III. KULLANILAN TEKNOLOJİLER

A. *Apache Kafka*

Apache Kafka, büyük veri işleme projelerinde yüksek hızlı veri akışı sağlamak için kullanılan bir dağıtık mesajlaşma sistemidir. Projemizde, Kafka hem producer hem de consumer süreçlerinde etkin bir şekilde kullanılarak EKG sinyallerinin gerçek zamanlı işlenmesi sağlanmıştır. Kafka, veri güvenilirliğini ve işleme hızını artırarak projenin ölçeklenebilirliğini desteklemiştir.

B. *Apache Spark*

Apache Spark, projemizde büyük veri analitiği için kullanılan bir başka önemli teknolojidir. Spark Streaming modülü, Kafka'dan alınan gerçek zamanlı EKG verilerinin işlenmesi ve analiz edilmesi için kullanılmıştır. Bu teknoloji, hem hızlı veri işleme kapasitesi hem de veri paralelliği özellikleriyle projenin etkinliğini artırmıştır.

C. *Python ve Makine Öğrenmesi Kütüphaneleri*

Proje, veri seti üzerinde ön işleme, görselleştirme ve model geliştirme süreçlerinde ***Python*** *programlama dilini kullanmaktadır. Python, veri bilimi ve makine öğrenmesi alanında yaygın olarak kullanılan bir dildir ve çok sayıda güçlü kütüphane sunmaktadır.*

1) *Scikit-Learn:*

Scikit-Learn, projede temel makine öğrenmesi süreçleri ve model değerlendirme metriklerinin hesaplanması için kullanılmıştır. Precision, recall, f1-score gibi metriklerin hesaplanmasında etkili bir araç olarak rol oynamıştır.

2) *PyTorch*

Proje kapsamında alternatif model geliştirme için PyTorch kullanılmıştır. Dinamik grafik yapısı ve esnekliği sayesinde modelin deneysel çalışmaları kolaylaştırılmıştır.

3) **Pandas ve NumPy**:

Pandas ve NumPy, veri manipülasyonu ve hesaplama işlemlerinde projenin temel taşları olmuştur. Veri setinin temizlenmesi, yeniden yapılandırılması ve analiz edilmesi gibi süreçlerde etkin bir şekilde kullanılmıştır.

4) *Matplotlib ve Seaborn:*

Matplotlib ve Seaborn, proje sonuçlarının görselleştirilmesi için kullanılmıştır. EKG sinyallerinin zaman serisi grafiklerinden, anomali tespit sonuçlarının dağılımlarına kadar birçok görsel araç bu kütüphanelerle oluşturulmuştur. Bu sayede sonuçların görsel olarak ifade edilmesi ve anlaşılabilirliği artırılmıştır.

D. *Jupyter Notebook*

Jupyter Notebook, veri bilimcilerinin ve geliştiricilerin veri analizi, görselleştirme ve model geliştirme işlemlerini interaktif bir ortamda gerçekleştirmelerini sağlayan bir araçtır. Proje sürecinde, Jupyter Notebook kullanılarak her adım detaylı bir şekilde kaydedilmiştir. Verilerin görselleştirilmesi, modelin eğitilmesi ve sonuçların raporlanması gibi süreçler bu ortamda yapılmıştır. Ayrıca, Jupyter Notebook, Python kodlarının çalıştırılmasında, açıklamalar eklenmesinde ve sonuçların hızlı bir şekilde paylaşılmasında büyük kolaylık sağlamaktadır. Veri seti üzerinde yapılan tüm analizler ve geliştirme süreçleri, Jupyter Notebook üzerinden gerçekleştirildiği için proje geliştirme süreci daha verimli hale gelmiştir.

IV. Bazı Onemli Çiktilar

```
Model Doğruluğu: 0.9940294362677035

Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.93      0.94      1456
           1       1.00      1.00      1.00     27352

    accuracy                           0.99     28808
   macro avg       0.97      0.96      0.97     28808
weighted avg       0.99      0.99      0.99     28808
```

Figure 1

Modelin performans değerlendirmesi sonuçlarına göre, doğruluk oranı %99.4 olarak hesaplanmıştır. Pozitif sınıf (1) için Precision, Recall ve F1-Score değerleri %100 olarak ölçülmüş, bu da modelin bu sınıf üzerinde mükemmel bir performans sergilediğini göstermektedir. Negatif sınıf (0) için ise Precision %95, Recall %93 ve F1-Score %94 olarak kaydedilmiştir. Macro ve weighted ortalamalar da sırasıyla %97 ve %99 seviyelerinde olup, modelin tüm sınıflar üzerinde dengeli ve başarılı bir performans sergilediğini ortaya koymaktadır.

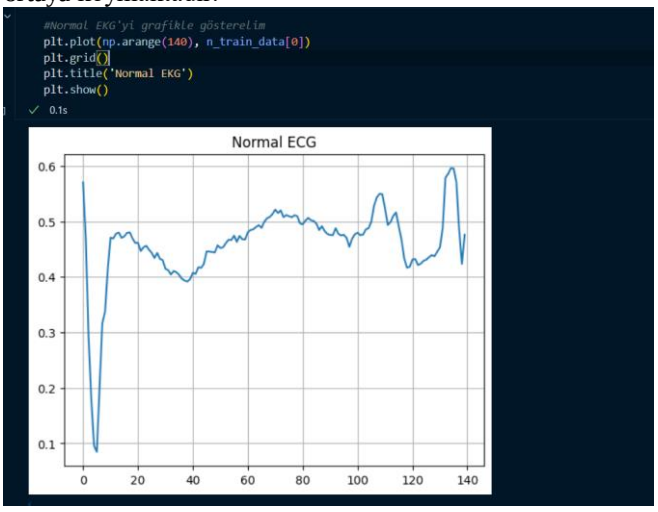

Figure 2

Grafikte, proje kapsamında analiz edilen normal bir EKG sinyali görülmektedir. Bu sinyal, eğitim verilerinden alınmış bir örneği temsil etmekte ve zaman serisine dayalı olarak kalp ritminin düzenini göstermektedir. Grafikteki dalgalanmalar, EKG'nin doğal bileşenlerini yansıtmakta ve normal bir kalp atışındaki elektriksel aktiviteyi ifade etmektedir. Bu grafik, projenin normal veri ile anormal veri arasındaki farkı tespit etmek için kullanılan temel verilerden biridir.

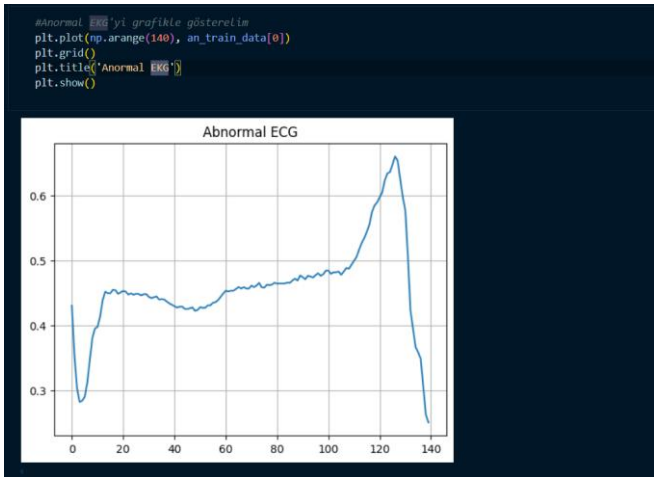

Figure 3

Grafikte, proje kapsamında analiz edilen anormal bir EKG sinyali görülmektedir. Bu sinyal, eğitim verilerinden alınmış bir anormal veri örneğini temsil etmektedir. Sinyalin dalgalanmalarındaki düzensizlikler ve ani değişimler, normal bir EKG sinyaline kıyasla farklılıkları ortaya koymaktadır. Bu grafik, anomali tespiti için modelin anormal veriler üzerindeki performansını değerlendirmek ve anomalileri daha net anlamlandırmak için kullanılmaktadır. Görsel, sistemin normal ve anormal EKG sinyalleri arasındaki ayrımı nasıl gerçekleştirdiğini göstermesi açısından önemlidir.

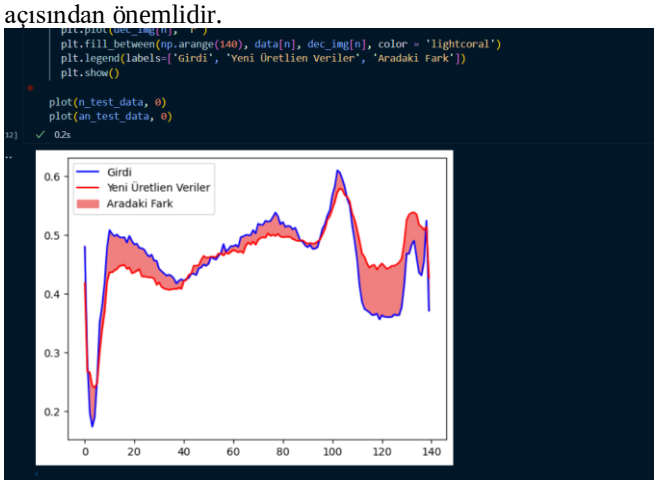

Figure 4

Bu görsel, orijinal EKG sinyali (mavi) ile model tarafından yeniden oluşturulan sinyalin (kırmızı) karşılaştırmasını göstermektedir. İki sinyal arasındaki fark (kırmızı gölge), modelin anormallikleri ne kadar iyi tespit edebildiğini ifade eder. Gölgeli alanın geniş olduğu bölgeler, modelin potansiyel anomalileri belirlediği yerlerdir. Bu, modelin performansını ve anomalilerin tespitindeki başarısını değerlendirmek için önemli bir görselleştirmedir.

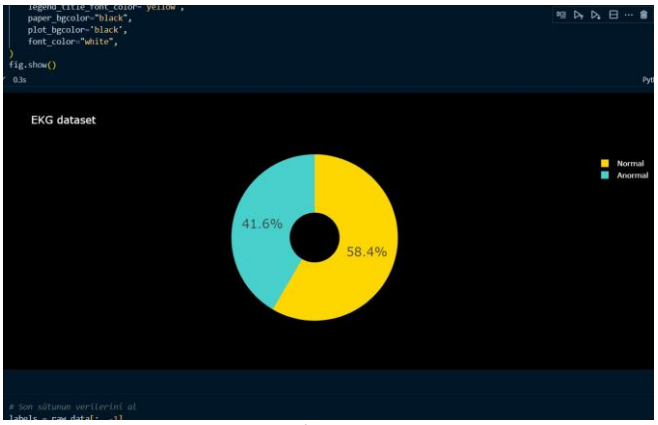

Figure 5

Bu pasta grafiği, kullanılan EKG veri setindeki normal ve anormal verilerin oranlarını göstermektedir. Verilerin %58,4'ü normal, %41,6'sı ise anormal olarak sınıflandırılmıştır. Bu dağılım, modelin hem normal hem de anormal durumları öğrenmesi ve tespit etmesi için dengeli bir veri seti sunmaktadır.

V. GELIŞTIRILMESI DÜŞÜNÜLENLER

Proje, gerçek zamanlı veri akışı ve anomali tespiti alanlarında önemli bir ilerleme kaydetmiş olsa da, gelecekte yapılacak geliştirmeler ile modelin doğruluğu ve genel işlevselliği daha da artırılabilir.

A. *Derin Öğrenme Modelinin İyileştirilmesi:*

Kullanılan derin öğrenme modeli daha karmaşık sinir ağı mimarileri ile geliştirilebilir. Örneğin, zaman serisi verilerinde oldukça başarılı olan LSTM veya CNN gibi mimarilerin kullanılması anomali tespitinde daha iyi sonuçlar verebilir. Modelin hiperparametre optimizasyonu ve veri augmentasyonu gibi yöntemlerle daha fazla genel performans artırılabilir.

B. *Anomali Tespit Performansı:*

Mevcut anomali tespit algoritmalarının performansı iyileştirilerek daha doğru ve hızlı sonuçlar elde edilebilir. Daha gelişmiş algoritmaların (örneğin, Autoencoder ve One-Class SVM) test edilmesi, modelin performansını karşılaştırmalı olarak değerlendirme imkânı sunar. Ayrıca, tespit edilen anomalilerin klinik doğruluğunun uzmanlar tarafından onaylanması, tıbbi açıdan anlamlı sonuçlar elde edilmesini sağlayacaktır.

C. *Özellik Mühendisliği:*

EKG sinyalleri karmaşık ve çok boyutlu yapıya sahiptir, bu yüzden doğru özelliklerin çıkarılması anomali tespiti için kritik önem taşır. Mevcut özelliklerin yanında, frekans alanında elde edilen parametreler, dalga formları ve kalp ritmiyle ilgili klinik anlam taşıyan özelliklerin eklenmesi performansı artırabilir. Domain uzmanlarıyla çalışarak anlamlı özellikler belirlenmesi sistemin doğruluğunu güçlendirebilir.

D. *Modelin Açıklanabilirliği:*

Sistemin kararlarının daha şeffaf hale getirilmesi için SHAP veya LIME gibi araçlar kullanılabilir. Bu araçlar, modelin neden belirli bir anomalinin var olduğunu düşündüğünü anlamaya yardımcı olur. Böylece, sağlık profesyonelleri modelin çıktılarından daha kolay yararlanabilir ve klinik kararlarda bu sonuçları kullanabilir.

VI. SÖZDE KOD

```
START

// 1. Kafka'dan EKG verilerini alma
CONNECT to Kafka Broker
SUBSCRIBE to EKG_Data_Topic
WHILE (Kafka has new messages)
   READ EKG signal data
   APPEND data to processing queue
END WHILE

// 2. Veri ön işleme
DEFINE function preprocess(data)
   REMOVE noise using bandpass filter
   NORMALIZE signal
   RETURN preprocessed data
END FUNCTION

FOR each EKG signal in processing queue
   PROCESSED_SIGNAL = preprocess(signal)
   ADD to Spark RDD
END FOR

// 3. Spark ile veri işleme
DEFINE Spark job process_EKG(RDD)
   MAP signal to extract_features(signal)
   APPLY anomaly_detection_model(features)
   RETURN (anomaly_status, processed_data)
END FUNCTION

RESULTS = process_EKG(Spark RDD)

// 4. Anomali tespiti
DEFINE function anomaly_detection_model(features)
   LOAD trained ML model
   PREDICT anomaly_status based on features
   RETURN anomaly_status
END FUNCTION

// 5. Sonuçların görselleştirilmesi
DEFINE function visualize_results(results)
   FOR each record in results
      IF anomaly_status == "Anomaly Detected"
         HIGHLIGHT signal in red
      ELSE
         PLOT signal normally
      END IF
   END FOR
   DISPLAY results in user dashboard
END FUNCTION

visualize_results(RESULTS)

// 6. Kullanıcıya geri bildirim
DEFINE function alert_user(results)
   FOR each anomaly in results
```

```
      IF anomaly_status == "Anomaly Detected"
         SEND alert to user
      END IF
   END FOR
END FUNCTION

alert_user(RESULTS)

END
```

VII. SONUÇ

Bu proje kapsamında, EKG sinyallerinin analizi ve anomali tespitine yönelik bir sistem geliştirilmiştir. Projenin ana hedefi, normal ve anormal EKG verilerini sınıflandırarak potansiyel sağlık sorunlarını erken tespit edebilecek bir çözüm sunmaktır. Kullanılan veri setindeki normal ve anormal dağılımlar dikkatle analiz edilmiş, sinyaller üzerinde ön işleme ve makine öğrenmesi modelleri uygulanmıştır. Anomali tespiti için kullanılan modelin doğruluk oranı %99.4 olarak hesaplanmış, bu da geliştirilen sistemin oldukça başarılı olduğunu göstermiştir. Projede, Kafka kullanılarak gerçek zamanlı veri akışı sağlanmış, Spark Streaming ile veriler işlenmiş ve anomali tespiti gerçekleştirilmiştir. Sonuçlar, anormal EKG sinyallerinin doğru bir şekilde tespit edildiğini ve sistemin, verilerin grafiksel analizleriyle desteklenen detaylı bir sınıflandırma sunduğunu ortaya koymaktadır. Bu sistemin sağlık alanında pratik bir uygulama potansiyeline sahip olduğu ve gerçek zamanlı olarak EKG verilerini analiz edebilme kapasitesiyle önemli bir katkı sağlayacağı öngörülmektedir.

KAYNAKÇA

- Spark
- Kafka
- StackOverflow1
- StackOverflow2
- StackOverflow3
- Spark Yapısı
- Çevresel Değişkenler
- Spark Eğitimi
- Büyük Veriye Giriş Kursu