T.C.

MİMAR SİNAN GÜZEL SANATLAR ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

OLASILIKSAL SEMBOLİK MOTİF TANIMA

DOKTORA TEZİ

Oğuz AKBİLGİÇ
(20057203)

İstatistik Anabilim Dalı

İstatistik Programı

Tez Danışmanı: Prof. Dr. Eylem Deniz

ARALIK 2022

BEYAN

Mimar Sinan Güzel Sanatlar Üniversitesi Fen Bilimleri Enstitüsü tez yazım kılavuzuna uygun olarak hazırladığım bu tez çalışmasında;

- tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel etik kurallarına uygun olarak sunduğumu,
- başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi,
- kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- ücret karşılığı başka kişilere yazdırmadığımı (dikte etme dışında), uygulamalarımı yaptırmadığımı,
- ve bu tezin herhangi bir bölümünü bu üniversite veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim.

Oğuz AKBİLGİÇ

TEŞEKKÜR

Bu tezin yazım aşamasında bana her türlü desteğini esirgemeyen tez danışmanım Prof. Dr. Eylem Deniz'e teşekkür ederim. Sayın hocam Prof. Dr. Nalan Cinemre'ye akademik kariyerimin her aşamasındaki önemli katkıları için çok teşekkür ederim. Tez izleme komitesinin değerli üyeleri Prof. Dr. Emrah Önder, Doç. Dr. Ayça Çakmak Pehlivanlı ve Doç. Dr. Öyküm Esra Yiğit'e tüm tavsiyeleri ve destekleri için teşekkür ederim.

Bu çalışma sırasında özellikle uygulama aşamasında destek veren değerli çalışma arkadaşım Dr. Fatma Aydınlık Güntürkün'e teşekkür ederim.

Kesintilerle de olsa yaklaşık on sekiz yıl suren bu doktora serüveninde kendilerine ayırmam gereken zamanı bu çalışmaya ayırmam konusunda anlayış ve desteklerini esirgemeyen aileme sonsuz teşekkürlerimi sunarım.

Kaan'a

OLASILIKSAL SEMBOLİK MOTİF TANIMA

ÖZET

Yapay Zeka (YZ) hayatımızın her alanında yer bulmaya başlayan bir kavram olarak son yılların en popüler terimlerinden birisi haline gelmiştir. Özünde, insan beyninin kavrama ve yorumlama fonksiyonlarını yürütebilecek makinaların yaratılması olarak daha çok robotik alanı ile ilgilidir. Ancak, uygulamada, makine öğrenmesi ve derin öğrenme destekli istatistiksel karar verme yöntemlerinin bir uzantısı olarak biraz da yanlış bir şekilde bilimsel literatüre yerleşmiştir. Dolayısı ile, makine örenmesi temelli herhangi bir sınıflandırma ya da regresyon probleminin YZ olarak adlandırıldığı sayısız bilimsel çalışmaya rastlamak mümkündür. Bu çalışmada da YZ terimi, makina örenmesi, derin öğrenme ve istatistiksel modellemeyi kapsayan bir disiplini işaret edecek şekilde kullanılmıştır.

Tıp bilimi, YZ'nın en yaygın kullanıldığı bir alan olarak öne çıkmaktadır. Bunda, biyomedikal mühendisliği ve bilgisayar teknolojilerindeki ilerlemeler ve makina öğrenmesi alanında ortaya atıla sayısız yeni algoritmaların katkısı büyüktür. Biyomedikal alandaki gelişmeler sayesinde insan fizyolojisini temsil eden daha fazla saha, yüksek frekans ve çözünürlükte ve daha uzun süreli veri üretmenin yolu açılmaktadır. Bu teknolojileri takiben oluşan daha yüksek hacimdeki verinin saklaması ve işlenmesi konusunda bilgisayar teknolojileri de sürekli gelişmektedir. Örneğin kan basıncını saniyede 2000 örnek (2000 Hz) alarak raporlayan bir yatak başı monitörü, tek bir hasta ve sadece kan basıncı için, sadece bir günde 172.600.000 veri üretmektedir. Dolayısı ile, son yirmi yılda, elektronik hasta kayıtlarında saklanabilen verini hacmi gigabayt seviyelerinden petabayt seviyelerine çıkmıştır. Buna rağmen, birçok hastanede, yatak başı monitörlerde üretilen veri saklanamamakta ver her 48 saate bir silinmektedir. Bu tür

verilerin özetlenerek, temsili değişkenlerin saklandığı veri tabanları bir şekilde çözüm olabilmektedir.

Hastanelerde üretilen yüksek hacimli verilerin depolanmasını dışında analizi de oldukça zorludur. Çünkü, elektronik hasta kayıtlarında derlenen veriler genellikle geleneksel parametrik istatistiksel yöntemlerin varsayımlarını sağlamayacak şekilde doğrusal ve normal olmayan, yüksek korelasyonlu, eksik ve yüksek oranda gürültü içeren türdendir. Bu tür verilerin analizlerinden doğan ihtiyacı karşılamak üzere son yıllarda makine öğrenmesi alanında birçok yeni algoritma geliştirilmiştir. Ancak, makine örenmesi algoritmaları, doğrudan ham veri ile çalışan derin öğrenme yöntemleri dışında, bu tür boylamsal verileri temsilen değişkenlerin çıkarılmasını gerektirir. Bu aşamada, sinyal işleme yöntemleri yüksek frekanslı boylamsal verilerden değişken çıkarılması için sıklıkla kullanılmaktadır ve bu alandaki literatürde hızla gelişmektedir.

Yüksek frekanslı boylamsal verilerden temsili değişken çıkarılması için geliştirilmiş yeni algoritmalardan birisi, Olasılıksal Sembolik Motif Tanıma (OSMT) yöntemidir. OSMT yöntemi, boylamsal verilerin zaman içerisindeki değişimlerini olasılıksal modellerinin elde edilmesi ve bu olasılıksal modeller arasındaki uzaklıkların kullanılarak birden fazla serinin birbirlerine benzerliklerini karakterize eden değişkenler elde edilmesine dayanmaktadır. Bu çalışmanın amacı, OSMT yöntemini Türkçe literatüre kazandırmak ve kardiyovasküler hastalıkların tahmini üzerine uygulamasını göstermektir.

Anahtar Kelimeler: Olasılıksal sembolik motif tanıma, sinyal isleme, yapay zeka, makine öğrenmesi, EKG, elektrokardiyogram, kalp yetersizliği, kardiyomiyopati

PROBABILISTIC SYMBOLIC PATTERN RECOGNITION

SUMMARY

Artificial Intelligence is the buzz word of the century and finding an application in almost everywhere in our life. Simply, the goal of AI is to create machines that can do what human does there it is more relevant to robotics. However, the term 'AI' is interchangeably used to refer machine and deep learning. As a result, there are numerous publications in the literature where AI is used for machine learning based predictive models. In this study, we have used AI referring to a domain including machine learning and statistical modeling.

Medicine is one of the area of science where AI is most frequently used. This is partially because of the advancements in the biomedical and computer science as well as the methodological advancements in machine learning. Advancements in biomedical engineering technologies result in medical devices that can generate very large data representing human physiology at higher resolution and higher sampling frequency. For example, a bedside monitor measuring blood pressure at 2000Hz would generate 172,600,000 data points per patient, per day. Considering several other physiological data, several patients with years of data, the amount of data generated at a single hospital converges to petabytes. Yet, unfortunately, such rich data from bedside monitors typically purged every 48 hours in most institutions due to the challenges in streaming and storing such data. Thus, there are still need for more efficient ways of utilizing such rich data.

Besides the challenges in storing large volumes of data in healthcare systems, it is also challenging to analyze such high volume data. This is because the data in electronic health records (EHR) are typically with high correlation and include large number of missing data as well as error. Further, there are several different types of the data modalities in

EHR such as tabular format structured data, image data, text data, signal date, genomic data etc. explaining complex biological mechanism. It is very challenging to apply parametric statistical methods directly on such data. Recently, machine learning methods have emerged as an alternative solution to analyze data EHR data as nonlinear models with few assumptions about data. However, classical machine learning models, except the deep learning models typically allowing working with raw data, require extracting features representing the raw unstructured data such as signal, image, or text. These features are then used as predictors in machine learning models. When processing raw streaming physiologic data, signal processing algorithms are frequently used to extract features. Commonly used such methods are Fourier Transformation, Sample Entropy, Wavelet Transformation. On the other hand, despite deep learning models such as convolutional neural networks, directly can be used to predict or classify an outcome, they are also used as a way of extracting features.

In this study, we introduce a novel signal processing method, Probabilistic Symbolic Pattern Recognition (PSPR), to analyze longitudinal data points. We propose PSPR as a way of analysis of any kind of longitudinal or time series data where the data points does not need to be numerical as long as they are of a member of a finite set. The numerical series and time series, in this definition, are a subset of the domains that PSPR can be implemented. We designed PSPR to learn pattern transition behavior of given symbolic series to predict the future behavior or to compare the behavior of multiple series to carry out clustering and classification tasks. We specifically aimed to implement PSPR method on electrocardiogram (ECG) data to classify and predict risk for cardiovascular diseases.

This study introduces PSPR methods and showed its efficiency in extracting features from ECG data. PSPR can be implemented and tested on analysis of any kind of symbolic series beyond medicine.

Keywords: Probabilistic symbolic pattern recognition, signal processing, artificial intelligence, machine learning, ECG, electrocardiogram, heart failure, cardiomyopathy

İÇİNDEKİLER

KISALTMALAR

EKG: Elektrokardiyogram

SD: Sinüs düğümü

NSR: Normal sinüs ritim

YZ: Yapay Zeka

KSFD: Kısa zamanlı Fourier Dönüşümü

FD: Fourier Donuşumu

KDD: Kesikli Dalgacık Dönüşümü

SDD: Sürekli Dalgacık Dönüşümü

ESA: Evreşimli Sinir Ağları

YakEnt: Yaklaşık Entropisi

ÖrEnt: Örnek Entropisi

GA: Genetik Algoritma

OSMT: Olasılıksal Sembolik Motif Tanıma

GFM: Geçiş Frekansları Matrisi

GOM: Geçiş Olasılıkları Matrisi

GBM: Geçiş Benzerlik Matrisi

GOB: Geçiş Olasılıkları Benzerliği

KKY: Konjestif Kalp Yetersizliği

KAD: Kalp Atisi Değişkenliği

SDRR: RR standart sapması

TABLO LİSTESİ

ŞEKİL LİSTESİ

1. GİRİŞ

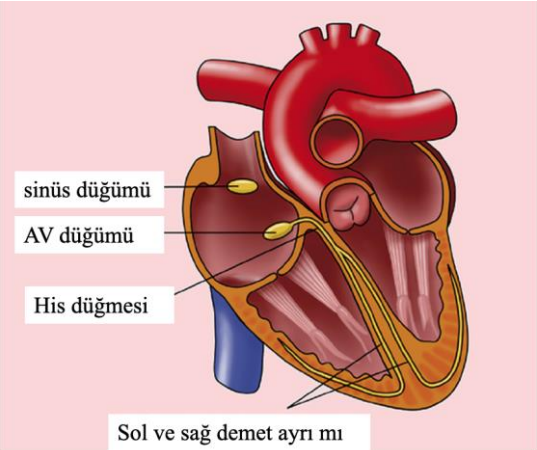
Kalp, anne karnında ilk atmaya başlaması ile hem etik hem hukuksal açıdan hayatı başlatan ve durması ile hayatı sonlandıran en temel organdır. Temel işlevi, vücuda gerekli kanın, dolayısı ile oksijenin, düzenli olarak iletilmesi olduğu için en hayati organların başında gelmektedir. Doğrudan kalp ile ilgili olan veya olmayan birçok etmen, kalp atışındaki düzensizliklere yol açabilmektedir. Kalp atışındaki bu düzensizliklerin teşhis edilmesinde kullanılan en temel araç elektrokardiyografi (EKG) sıklıkla kullanılır. Bu kısımda EKG ile genel bilgiler mevcut literatürden ve Elektrokardiyogram (Akbilgic et al., 2023) isimli kitap bölümünden alıntılardan oluşmaktadır.

1.1. Elektrokardiyogram

İnsan bedeninde sinir ve kas hücreleri birbirleri ile iletişiminde kimyasal ve elektrik sinyallerini kullanırlar. Bu elektrik sinyalleri aynı zamanda kalp atışımızı da düzenler (2006). EKG, kalp atışı sürecinde üretilen elektrik aktivitenin kayıt altına alınmasından ibarettir. Kalbin elektrik ürettiği 1856 yılında Von Kölliker ve Müller tarafından keşfedilmiştir (Silverman, 1992) ve ilk defa 1887 yılında Augustus D Waller tarafından kaydedilmiştir (Lüderitz & de Luna, 2017). 1887 yılında Uluslararası Fizyoloji Kongresinde Waller'in sunumundan esinlenen William Einthoven, Waller'in geliştirdiği yöntemi ilerleterek 1902 yılında insan kalbinin elektrik aktivitesini detaylı olarak kaydetmeyi başarmış ve 1908 yılında bu çalışması ile Nobel ödülü almıştır (Lüderitz & de Luna, 2017; Silverman, 1992).

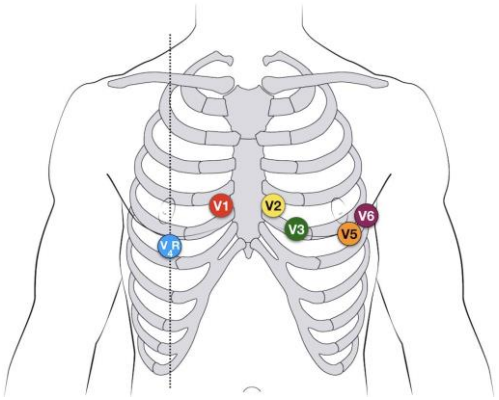
1.1.1. Elektrokardiyografinin İşlevi ve PQRST Kompleksi

Kalpte elektrik sinyalleri Şekil 1'de görüldüğü gibi sağ kulakçıkta yer alan ve sinüs düğümü (SD) adı verilen bazı hücreler tarafından yaratılır ve kalp kasları üzerinde çok düşük voltajlı akım şeklinde yayılır (Şekil 1) (Bilt, 2022).



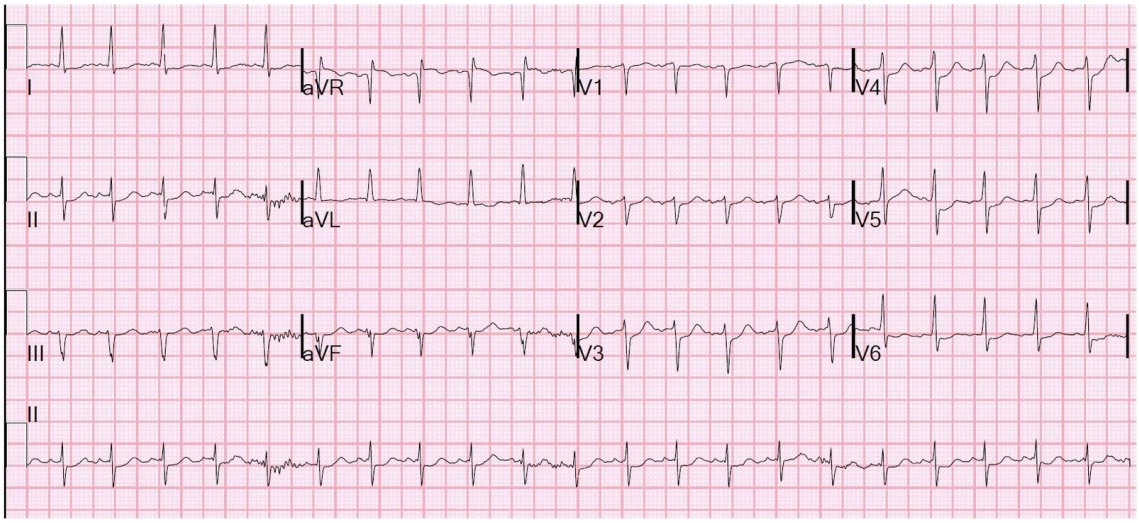
Şekil 1 Kalpte elektrik enerjisinin dağılım kanalları

Kalp kaslarına gönderilen bu uyarı sayesine önce kalbin kulakçıkları daha sonra karıncıkları kasılarak kalp içinde yer alan kanın hareketini sağlar. Kalpte meydana gelen bu elektrik akımları düşük seviyede olsada, vücut yüzeyinden ölçülebilecek düzeydedir. ECG ile, kalpteki elektrik yüklerinin insan teninin farklı açı ve Şekil 2'de gösterilen noktalardan noktalarından elektrotlar yardımı ile ölçülmesi sağlanır (Şekil 2) (Bilt, 2022).



Şekil 2 Elektrot yerleştirme noktaları

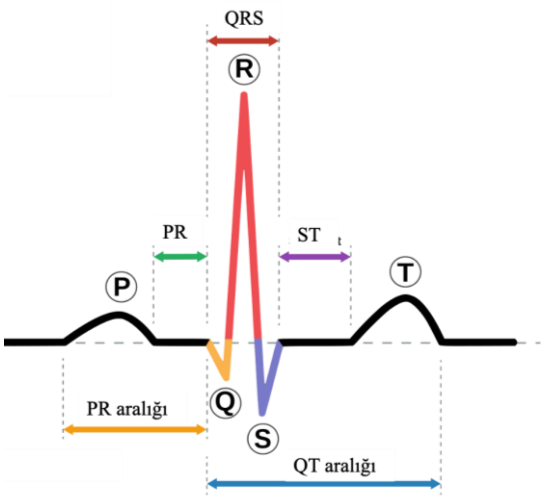
Bu ölçümlerin çizdirilmesi ile oluşan grafiklere elektrokardiyografi denir (InformedHealth.org, 2006; Prineas et al., 2010). En sık kullanılan EKG yöntemi, 12 derivasyonlu 10 saniyelik kayıt yöntemidir olup elde edilen EKG kaydının örneği Şekil 3 ile verilmiştir.



Şekil 3 12 derivasyon 10 saniyelik EKG örneği

1.1.2. Normal Sinüs Ritim

Sağlıklı bireylerde kalbin elektrik aktivitesinin belirli normlara uyması beklenir. Bu normlar genel olarak normal sinüs ritim (NSR) olarak adlandırılır ve Şekil 4 ile verilen temel aralıklardan oluşur. Sağlıklı bireylerde bazı temel EKG kompleksi aralıkları ve bu aralıklar için beklenen normal değerler Tablo 1 Normal sinüs ritim karakteristikleri ile verilmiştir (Sauer et al., 2022).



Şekil 4 Temel EKG aralıkları

Tablo 1 Normal sinüs ritim karakteristikleri

EKG Karakteristikleri		**Normal Değerler**
Kalp Atışı		60-100/dakika
P dalgası süresi		<120 ms
P dalgası yüksekliği		0.15-0.20 mV
P-R aralığı süresi	sinüs oranı>130/d ise	170
	130/d>sinüs oranı >100/d ise	180-190
	100/d>sinüs oranı>70/d ise	200
	70/d>sinüs oranı ise	210

1.1.3. Normal Sinüs Ritimden Sapmalar

Normal sinüs ritimden meydana gelen sapmalar kalp yetersizliği (Soliman et al., 2017), kalp krizi (Chang et al., 2019), inme (Agarwal & Soliman, 2013; Maheshwari et al., 2019) ve ritim bozukluğu (Heckbert et al., 2018) gibi kardiyovasküler hastalıkların işareti olabilecekleri gibi (Maron et al., 2014) ölüm riski (Afify et al., 2018) ile ilgili de bilgi

verebilirler. EKG okuma eğitimi almış bir doktor genel olarak bu tip EKG anomalilerini kolaylıkla belirleyebilir. Ancak bu tip hastalıkların EKG üzerinden tespit edilebilmesi için PQRST dalgalarının voltaj yükseklikleri ve sürelerine ilişkin bilgilerin çok doğru bir şekilde ortaya koyulabilmesi gereklidir. Diğer taraftan, doktor başına düşen aşırı hasta sayısı nedeni ile doktorlar sıklıkla bu EKG'lerin analizine yeterli vakit bulmakta zorlanmaktadır. Bununla beraber, hemen hemen her tecrübe seviyesindeki doktor, EKG okuma eğitim seviyesine bakmaksızın, hatalı okuma yapabilmektedir (Begg et al., 2016; Cook et al., 2020; Masoudi et al., 2006; Todd et al., 1996; Westdrop et al., 1992). Bu aşamada, EKG verisinin Yapay Zeka (YZ) ile otomatik olarak analizi üzerine yoğun bir literatür gelişmiştir.

Manuel EKG yorumlanmasından kaynaklanan diğer bir sakınca ise normal sinüs ritminin her zaman kardiyovasküler açıdan bir sorun olmadığı anlamına gelmemesidir. Örneğin, YZ tekniklerinin normal sinüs ritmi gösteren EKG verilerine uygulanması ile atriyal fibrilasyon hastalığının belirlenmesi mümkündür (Kamaleswaran, Mahajan, et al., 2018; Perez et al., 2019). Daha ilginci, Parkinson hastalığı gibi temelde kardiyovasküler olmayan bazı hastalıkların bile EKG verisine YZ uygulanması ile belirlenmesi mümkündür (Oguz Akbilgic et al., 2020).

1.2. Elektrokardiyografiden Değişken Elde Etme Yöntemleri

Klinik ortamda geleneksel EKG cihazları yoğun bir şekilde kullanılmaktadır. Klinik ortamda uygulanan standart 12-lead EKG 10 saniye uzunluğundadır. Örneklem frekansı cihazdan cihaza değişmekle beraber, genel olarak 500Hz kullanılmaktadır. Dolayısı ile tipik bir standart 12-derivasyon 10 saniye EKG verisi 5000x12 boyutunda bir matris oluşturmaktadır. Burada 12 sütunun her biri bir derivasyona karşılık gelmektedir. 5000 ise zamanı göstermektedir. 5000x12 boyutundaki matrisin elementleri ise ilgili zamanda ilgili derivasyonda ölçülen elektrik akımını mili volt (mv) cinsinden göstermektedir. Standart 10 saniye EKG dışında, Holter cihazları yardımı ile uzun süreli EKG kaydı da yapılmaktadır. Örneğin 500 Hz örneklem frekansında 7 günlük bir 12 derivasyon ECG kaydı, 302.400.000x12 boyutunda veri üretmektedir.

Giyilebilir teknolojiler ve alıcı teknolojilerindeki gelişmelere paralel olarak artık EKG klinik ortam ve cihazlardan bağımsız olarak, akıllı saat gibi kişisel cihazlar ile de toplanabilmektedir. Bu cihazda klinik kalite de genellikle 30 saniye uzunluğunda tek

derivasyon EKG verisini üretebilmektedir. Bu EKG verilerinin makine öğrenmesi algoritmalarında kullanımı genellikle değişken çıkarımı yöntemleri ile olmaktadır. Burada değişken çıkarımı, EKG verisinin belirli özelliklerini özetleyen değişken yaratma işlemlerinin bütünüdür.

EKG verisinden değişken çıkarma işlemi aşağıda özetlenen üç başlık altında toplanabilir.

1.2.1. Geleneksel EKG Değişkenleri

Geleneksel EKG değişkenleri, genel olarak Şekil 4 ile verilen EKG'yi oluşturan dalgaların boyu ve yüksekliği ile ifade edilen değişkenlerdir. Sıklıkla kullanılan geleneksel EKG değişkenlerine örnek olarak P dalgası boyu ve yüksekliği, QT süresi, iki R dalgası arası zaman, ST süresi, T dalgası yüksekliği değişkenleri verilebilir.

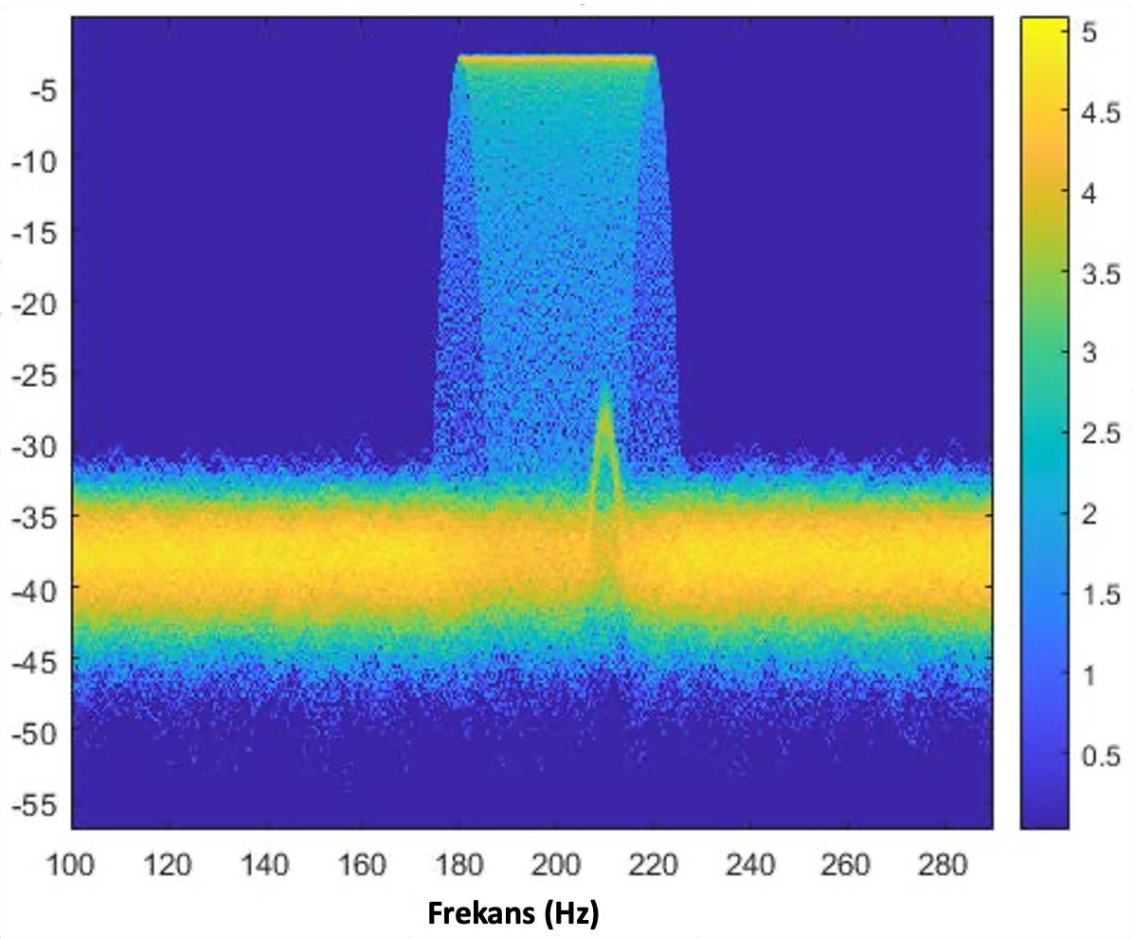
1.2.2. Sinyal İşleme Yöntemleri EKG Verisinden Değişken Çıkarımı

Sinyalden değişken çıkarımı yöntemleri temel olarak, ilgili sinyal verisinin belirli karakteristik özelliklerini temsil eden bilgilerin elde edilmesi işlemlerini kapsar. Bu işlemler ortalama veya standart sapma gibi basit istatistiksel yöntemlerle yapılabileceği gibi sinyal içerisindeki düzensizliği ve kaotik yapıyı temsil edecek şekilde geliştirilmiş yöntemlerle de yerine getirilebilmektedir. Sinyal işleme ile daha çok, bu tip ileri analizler işaret edilmektedir. Sinyal işleme yöntemleri genel olarak zaman boyutu, frekans boyutu, entropi ve doğrusal olmayan yöntemler olarak sınıflandırılır.

1.2.2.1. Fourier Dönüşümü

Kısa zamanlı Fourier Dönüşümü (KSFD), ya da İzge Grafikleri, çok bileşenli ve durağan olmayan zaman serilerinin doğrusal zaman-frekans dönüşümlerinde kullanılan bir yöntemdir. KSFD dönüşümlerinin karesi alınarak izge grafikleri elde edilir. Birden fazla sinyalin izge grafiklerinin eşanlı görselleştirilmesi ile sinyaller arasındaki zaman-frekans boyutlarındaki benzerlikler görselleştirilebilir. Bir sinyalin tekrarlayan izgeleri, o sinyalin zaman-frekans uzayında ne sıklıkta görüldüğünü ifade eder (Şekil 5 Tekrarlayan Sinyal motifinin zaman-frekans uzayında KZFD ile gösterimine bir örnek). Dolayısı ile, belirli bir frekans ne kadar uzun süre tekrarlıyorsa, izge grafiğinde o kadar yoğun olarak renklendirilir. Bu nedenlerle, KZFD yöntemi kullanılarak çıkarılan değişkenler, sinyal içerisinde sıklıkla tekrar eden motifleri modellemekte oldukça başarılıdır. Başlıca uygulama alanları ses sinyallerinin işlenmesi, metal plakalar üzerindeki çatlakların

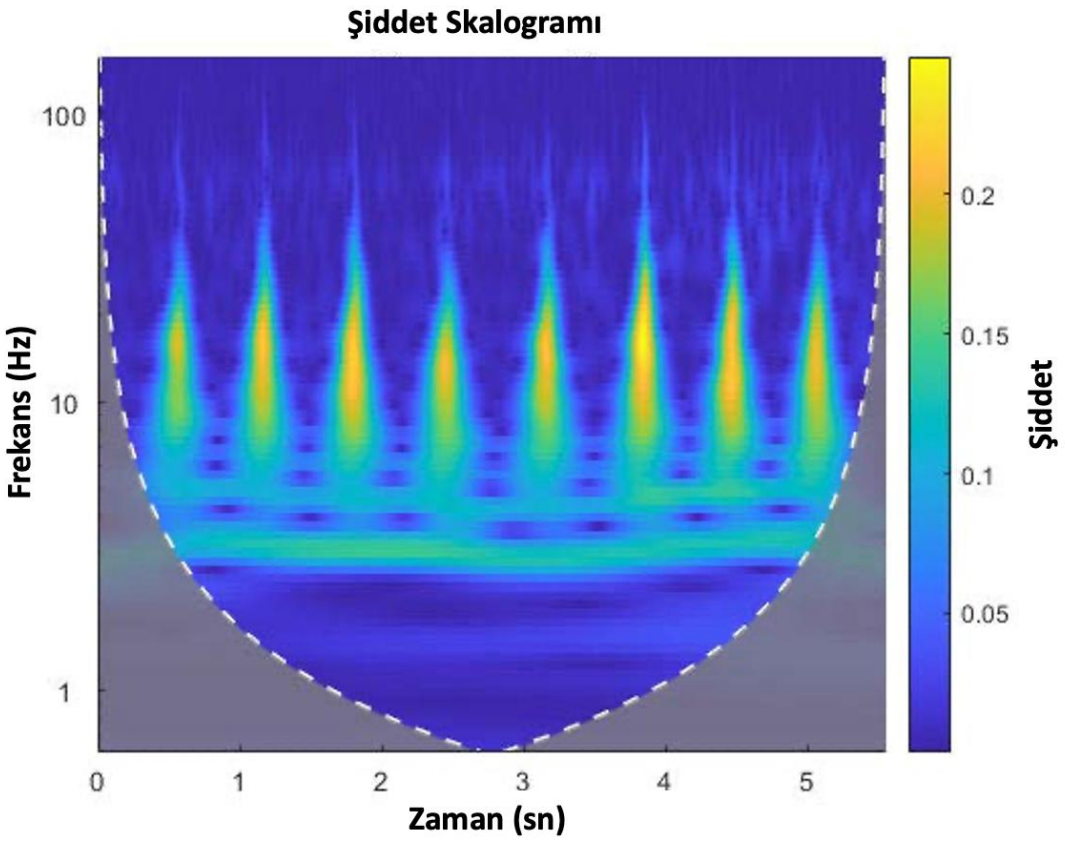
belirlenmesi, alıcıların dizilimlerinin düzenlenmesi, online telefon ve video görüşmelerinde sinyal kalitesinin belirlenmesi ve daha birçok alanda kullanılmaktadır.



Şekil 5 Tekrarlayan Sinyal motifinin zaman-frekans uzayında KZFD ile gösterimine bir örnek

1.2.2.2. Dalgacık Dönüşümü

Sürekli Dalgacık Dönüşümü (SDD), durağan olmayan sinyallerdeki geçişlerin ve anlık yüksek frekanslı değişimlerin modellenmesinde kullanılır. SDD zaman-frekans uzayını değişken zaman aralıkları ile tarar (Şekil 6 Sürekli Dalgacık Değişimi ile EKG üzerinde gürültü belirlenmesi). Zaman aralıkları otomatik olarak zamana bağlı olarak genişler. Böylece, düşük frekanslı motiflerin modellenmesine uygun bir dönüşüm gerçekleştirir. SDD çok geniş bir uygulama alanına sahiptir. Bu alanlardan bazıları, kalbin ve beynin elektrik aktivitesini temsil eden EKG ve elektroencephalogram sinyallerinin analizi ve daha genel olarak sinyallerin evreşimli sinir ağları ile analizi için dönüştürülmesidir (MATLAB, 2020).



Şekil 6 Sürekli Dalgacık Değişimi ile EKG üzerinde gürültü belirlenmesi

1.2.2.3. Entropi

Genel olarak entropi, bir sistem içerisindeki belirsizlik ya da düzensizlik olarak tanımlanabilir. Sıklıkla kullanıldığı termodinamik disiplininde ise, kapalı bir termodinamik sistemi içerisinde kullanıma elverişli olmayan enerjinin bir ölçüsüdür. Dolayısı ile ilgili sistemin mevcut durumunun bir ölçüsü olarak da düşünülebilir ve sistem ısısındaki dönüştürülebilir değişimler ile doğru sistem ısısının kendisi ile ters orantılıdır. Entropi, fizik alanındaki kullanımının dışında, bilgi kuramı ve zaman serileri arasındaki benzerliklerin ölçülmesinde de uygulama alanı bulmuş olan bir kavramdır.

Entropi, Cladue E. Shannon (1951) tarafından bilgi kuramına uyarlanmıştır. Bilgi kuramında entropi, bir rastlantı değişkeni için belirsizlik ölçüsü olarak tanımlanmaktadır (Wang, 2008, s. 1). Herhangi bir X rastlantı değişkeni için hesaplanan ve H ile gösterilen entropi, rastlantı değişkeninin taşıdığı bilginin bir ölçüsüdür. Diğer bir ifade ile, $P_X(X)$

dağılım fonksiyonuna sahip olduğu varsayılan bir X rastlantı değişkeni için Eşitlik 1 ile verilen eşitlik yardımı ile hesaplanır.

$$H(X) = -E[\ln P_X(X)] \quad (1)$$

Bu bağlamda entropi, rastlantı değişkeninin gözlemlenen değerlerinin, ilgili rastlantı değişkenin teorik dağılımına ne kadar uyduğu bilgisini vermektedir. Buradan hareketle, örneğin, normal dağılıma uyan bir rastlantı değişkeni için Shannon Entropisi Eşitlik 2 ile verilen formül kullanılarak hesaplanır.

$$H(X) = -E[\ln P_X(X)] = -E\left[\ln \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\frac{X-\mu}{\sigma}\right)^2}\right] = -E\left[\ln 1 - \ln \sqrt{2\pi}\sigma - \ln e^{-\left(\frac{X-\mu}{\sigma}\right)^2}\right] = \ln \sqrt{2\pi}\sigma \quad (2)$$

Yukarıdaki eşitlikte yer alan $\ln \sqrt{2\pi}\sigma$ entropi değeri bilinmeyen parametre değerlerine sahiptir ve örneklemden hesaplanır. Örneğin, tam olarak standart normal dağılımı takip eden bir örnek için, örnek standart sapması 1 olacaktır ve entropi değeri $\ln \sqrt{2\pi}$ olarak hesaplanır.

Bilgi kuramının bir uzantısı olarak da sayılabilecek şekilde, entropi, sonlu değerli zaman serilerinin içerisindeki düzensizliğin ve birden fazla zaman serisi arasındaki benzerliklerin bir ölçüsü olarak da kullanılmakladır.

Bununla birlikte, zaman serisi şeklinde ifade edilen sistemler üzerinden entropi hesaplamakta kullanılan yöntemler, EKG gibi kısa süreli ve yüksek oranda gürültü içeren seriler için uygun değillerdir. Bu soruna çözüm olarak, yaklaşık entropi (YakEnt) kavramını ortaya atmıştır (Pincus, 1995; Pincus & Singer, 1996; Pincus, 1991). YakEnt, entropi ile yakından ilişkili olan sistem karmaşıklığı ölçülerini kullanarak hesaplanmaktadır ve EKG ve benzeri fizyolojik verilerin oluşturduğu zaman serilerinin analizine uygundur. Ancak, YakEnt hem birbiri ile tutarsız sonuçlar üretmekte hem de uzunlu eşit zaman serilerine uygulanabildiği için kullanımı oldukça sınırlı kalmıştır.

Örneklem Entropisi (ÖrEnt) (Richman & Moorman, 2000), YakEnt'in bu eksikliklerini giderecek şekilde geliştirilmiş olan diğer bir entropi hesaplama yöntemidir.

1.2.3. Evreşimli Sinir Ağları (ESA)

Evreşimli sinir ağları (Lecun et al., 1998), yada İngilizcesi ile Convolutional Neural Networks, 1990'larda ortaya atılmıştır. Ancak, model mimarisinde öğrenilmesi gereken parametre sayısının milyonları bulabilmesi ve zamanın teknolojilerini bu hesaplamaları yapabilecek kapasitede olamaması nedeni ile yaygın kullanımı 2010 sonrasına kalmıştır. ESA, ham verinin işlenmesinde sıklıkla bir derin öğrenme algoritmasıdır. ESA asıl olarak 2 boyutlu resimlerin islenmesi için ortaya atılmış olmakla beraber, 1 boyutlu sinyallerin ve 3 boyutlu videoların islenmesinde de kullanılabilmektedir. ESA yönteminin temel amacı, evreşimli katmanlar aracılığı ile girdi olarak verilen resmin (ya da sinyalin) içerisinde, çıktı değişkenini tahmin etmekte kullanılabilecek motifleri yakalamak ve bu motifleri temsil eden nöronları normal ileri beslemeli ağlar yardımı ile çıktı değişkeninin kestirilmesinde kullanmaktır. Bu bağlamda, ESA girdi olarak verilen resmi islerken, her defasında, yani birden fazla evreşimli katmanlarda, orijinal resmin soyut temsili olan yeni resimler yaratır. Bu yeni resimler yaratılırken çıktı ile ilişkili olan motiflerin öne çıkarılması, ilişkili olmayan alanların ise geri planda kalması (etkisinin azaltılması) amaçlanır. Diğer bir ifade ile, evreşimli katmanlar aracılığı ile girdi resmi içerisinde çıktı ile ilişkisi olmayan gereksiz kısımlar filtrelenir. Bu filtrelenmiş soyut resimler, diğer ismi ile değişken haritaları, normal ileri beslemeli yapay sinir ağı katmanları ile çıktı katmanına bağlanır. Diğer bir ifade ile, verilen resimden evreşimli katmanlar aracılığı ile değişken çıkarılır ve bu değişkenler bilinen ileri beslemeli yapay sinir ağları modelleri ile islenir. Dolayısı ile, ESA değişken çıkarma amacı ile kullanıldığında tipik olarak son evreşimli katmanın düzleştirilmiş çıktı nöronları en son eğitilmiş modelden çıkarılarak değişken olarak kullanılır.

1.3. Değişken Secimi: Genetik Algoritma

Genetik Algoritma (GA), doğal seçilim yasaları matematiğe uyarlayan bir stokastik optimizasyon yöntemidir. Genetik algoritma ile değişken seçiminde bilgiler ikili sistemde kodlanır ve her bir sayı bir değişkenin modelde yer alıp almadığını gösterir. Amaç, Charles Darvin tarafından ortaya atılan en güçlü olanın hayatta kalması ilkesine göre, ilgili hata fonksiyonunun en iyi değerini veren değişken alt kümesinin bulunmasıdır. GA

diğer optimizasyon algoritmalarına kıyasla oldukça başarılı sonuçlar üretmektedir (Eiben & Smith, 2015). GA, sağlam regresyon ve deney tasarımı gibi istatistiksel yöntemlerin parametrelerinin en iyi değerlerinin bulunması amacı ile de kullanılmıştır (Hamada et al., 2001). Bu çalışmada GA, daha önceki başarılı uygulamalarından esinlenerek (Akbilgiç, 2011; Akbilgic et al., 2014), EKG verilerinden elde edilen değişkenler arasından ilgili çıktı değişkenindeki değişimi en iyi açıklayan değişkenlerin belirlenmesinde kullanılacaktır.

Bu çalışmanın ikinci bölümünde tıp alanında sinyal türü veriler ve bu verilerin ne tür hastalıkların teşhisinde kullanıldığı yer almaktadır. Üçüncü bölümde, sinyale tipi verilerden değişken çıkarmakta kullanılan başlıca yöntemler ifade edilmiştir. Dördüncü bölümde, geliştirdiğimiz OSMT yönteminin teorik altyapısı ve beşinci bölümde ise ÖSMT yöntemi, kardiyovasküler hastalıkların önceden tahminine uygulanmıştır. Sonuç bölümünde çalışmamızın çıkarımları özetlenmiş ve ilerili çalışma planlarına yer verilmiştir.

2. OLASILIKSAL SEMBOLİK MOTİF TANIMA

Zaman serileri analizi, istatistiğin, literatürü oldukça olgunlaşmış bir alt dalıdır. Sıklıkla, zamana bağlı bir değişkenin gelecekteki değerlerinin tahminlenmesine yönelik analizleri içerir. Birçok zaman serisi, değerleri, geçmiş değerleri ile açıklanan otoregresif yapıdadır. Otoregresif Hareketli Ortalama ve türevi olan hemen hemen tüm zaman serileri teknikleri, bağımlı değişkenin gerçel değerli sürekli bir olasılık dağılımına sahip olduğu varsayımına dayanır (Zhang & Moore, 2014).

Zaman serilerinin analizine dair diğer bir ilgi alanı ise, farklı serilerin benzerliklerine göre kümelemesidir. Dinamik Zaman Bükümü (Li & Clifford, 2012), bu fikre dayanan algoritmalara bir örnektir. Diğer taraftan, Kesikli Fourier Dönüşümü (KFD) (Harris, 1978; Ream, 1977), Dalgacık Dönüşümü (Schiff et al., 1994; Shima & Nakayama, 2009), Parçalı Toplamsal Yakınsama (Ren et al., 2018) ve Sembolik Toplamsal Yakınsama (STY) (Lkhagva et al., 2006a, 2006b) gibi zaman serilerini birbirini takip eden parçalı zaman dilimlerinde gerçekleşen olaylar olarak ifade etmeye dayanan yöntemler de mevcuttur. Bu yöntemlerin tamamı, zaman serisi verisinin mevcut kümeleme analizlerinin uygulanabileceği türden veriye dönüştürülmesine dayanır.

Aslında zaman serileri, örüntü ve fonksiyonel ilişkilerin, birbirini izleyen olaylar zincirine bağlı olduğu, sıralı veri türünün özel bir halidir. DNA üzerine kodlanmış olan genetik bilgiler, sıralı verinin zamana bağlı olmayan haline bir örnek olarak verilebilir. Zaman serileri sıralı verilerin özel bir hali olduğu için, genel kapsayıcı bir çerçeve içerisinde sıralı veriler için geliştirilmiş modeller, zaman serilerine de uyarlanabilir.

Bu çalışma, sıralı verilerin analizinde yeni bir yöntem olan Olasılıksal Sembolik Motif Tanıma (OSMT) (Akbilgic & Howe, 2017) tekniğinin tanıtılmasını amaçlamaktadır. OSMT, türlerin DNA dizilimlerine göre sınıflandırılması (Akbilgic & Howe, 2017), elektrokardiyografi (EKG) verilerin den ritim bozukluğu belirlenmesi (Kamaleswaran, Mahajan, et al., 2018; Mahajan, Kamaleswaran, Howe, et al., 2017) ve tahmini (Mahajan, Kamaleswaran, & Akbilgic, 2017; Sutton et al., 2019), yoğun bakım ünitelerinde kan zehirlenmesi riskinin erken tahminlenmesi (Kamaleswaran, Akbilgic, et al., 2018; van Wyk et al., 2017) ve Parkinson Hastalığının (O Akbilgic et al., 2020) hareket kabiliyetini sınırlayan belirtilerinin ortaya çıkmasından evvel, EKG verisinden teşhisi gibi çok geniş bir yelpazede kullanılabilir.

OSMT modelinin tek kapsayıcı varsayımı, seriyi oluşturan verilerin (sembollerin) sonlu sayıdaki sembolden oluşan bir alfabeden geliyor olmasıdır. İlgili alfabenin sembolleri bilinen ya da yaratılabilecek herhangi bir sembol kümesinden oluşabilir. Ancak, dil birliği sağlamak açısından, bu çalışmada kullanılacak semboller, İngiliz Alfabesini oluşturan terimlerinden seçilecektir. Bu bölümdeki tanımlar ve teorik gösterimler Oguz Akbilgiç'in Sequential Analysis Dergisinde yayınlanan 'Symbolic Pattern Recognition' isimli çalışmamsından alıntılanmıştır (Akbilgic & Howe, 2017).

Tanım 1. Olasılıksal Sembolik Motif Tanıma (OSMT): OSMT, birbirini takip eden olayların (sembollerin) aralarındaki geçiş davranışının modellenmesine dayanan bir sinyal işleme yöntemidir. Bir serinin gelecek değerlerinin tahminlenmesine ya da birden fazla serinin sınıflandırılması ve kümelenmesi analizlerinde kullanılabilir.

Yukarıda belirtildiği üzere, OSMT yöntemi, değerleri belirli bir sembol kümesinin elemanı olan sıralı verilerin analizinde kullanılır. Alfabe olarak adlandıracağımız bu küme, gösterim kolaylığı açısından, 26 harfli İngiliz Alfabesi, I=$\{a, b, c, \dots, z\}$, kullanılarak Eşitlik 3'de verildiği gibi gösterilecektir.

$$A_s = \{I_i\}_{i=1}^{s} \qquad (3)$$

Eşitlik 3'de s ile alfabe uzunluğu gösterilmektedir. Örneğin $s = 3$ ile tanımlanan seriler, İngiliz Alfabesi'nin ilk üç harfleri olan a, b ve c de oluşacaktır ve $A_3 = \{a, b, c\}$ ile gösterilecektir.

OSMT, elemanları sembollerden oluşan sıralı serilerin, olasılıksal modellerinin, gözlemlenen geçiş frekansları yardımı işe oluşturulmasına dayanır. Burada semboller arasındaki geçişler, ***k*-sembollü motif geçişi** olarak ifade edilmektedir. Örneğin beş elemanlı '*abcbb*' serisinde ilk sembolün '*a*' dan '*b*' ye dönüşmesi ('***ab***cbb') 1-sembollü motif geçişi, '*bcb*' nin '*b*' ye ('*a****bccc***') dönüşmesi ise üç-sembollü motif geçişine örnektir. Burada modellenecek olan en çok motif geçiş uzunluğu, *k* ile gösterilecektir. Geçiş olasılıklarının tek sembol geçişlere sınırlandırılmaması sayesinde, klasik Markov Süreçlerinden farklı olarak, bir sembolden diğerine geçişlerin sadece son gözlemlenmiş sembole değil, belirlenen sayıdaki gözlemenmiş olaylara bağlı olduğu göz önüne

alınmaktadır. Diğer bir ifade ile, daha uzun süreli bir bellek kullanılarak, serilerin olasılıksal modelinin oluşturulması hedeflenmektedir.

k-sembollü geçiş olasılıklarının hesaplanması için öncelikle, her bir geçişin gözlemlenen frekanslarının tutulduğu *k*-sembollü geçiş frekansları matrisleri (*GFM*), alfabe uzunluğu *s* olmak üzere aşağıdaki gibi ifade edilir.

$$GFM_k = F_{i,j} \quad , i = 1, \dots, s^k \; ve \; j = 1,2, \dots, s \tag{4}$$

Eşitlik 4'de yer alan GFM_k nin her bir elemanının, bulundukları satırın toplamına bölünmesi ile, *k*-sembollü geçiş olasılıkları matrisleri (GOM_k) elde edilir.

$$GOM_k = O_{i,j} \quad , i = 1, \dots, s^k \; ve \; j = 1,2, \dots, s, O_{i,j} = \frac{F_{i,j}}{\sum_{j=1}^{s} F_{i,j}} \tag{5}$$

Örnek 1: $A_3 = \{a, b, c\}$ alfabesinin elemanları ile oluşturulmuş $S =' aabcabccbabcabcbaabc'$ sembolik serisi için 2-sembollü geçiş frekansları ve olasılıkları Tablo 2 ile verildiği gibi hesaplanır.

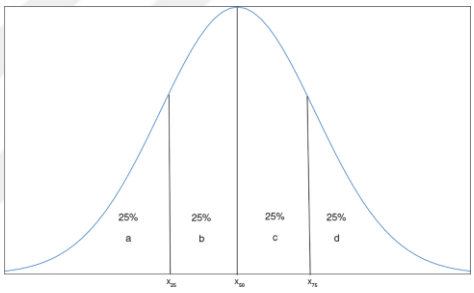
Tablo 2 Geçiş Frekansları Matrisinden Geçiş Olasılıkları Matrisinin elde edilmesi

	GFM_2			Total	GOM_2		
	a	*b*	*c*		*a*	*b*	*c*
aa	0	2	0	2	0,0	1,0	0,0
ab	0	0	5	5	0,0	0,0	1,0
ba	1	1	0	2	0,5	0,5	0,0
bc	2	1	1	4	0,5	0,25	0,25
ca	0	2	0	2	0,0	1,0	0,0
cb	2	0	0	2	1,0	0,0	0,0
cc	0	1	0	1	0,0	1,0	0,0

3.1. OSMT Yönteminin Sayısal Değerli Serilere Uygulanması

Daha önce ifade edildiği gibi OSMT yöntemi, sonlu sayıda bir alfabenin elemanları ile oluşturulmuş sembolik serilere uyarlanabilmektedir. Değerleri sonsuz elemanlı bir kümenin elemanları ile oluşturulmuş seriler, örneğin gerçel değerli sayısal seriler, kesikli hale getirildikten sonra OSMT yöntemi ile analiz edilebilirler.

Gerçel değerli X rastlantı değişkenin değerleri ile oluşturulmuş S sayısal serisi, tüm gözlemlenmiş değerlerine bir sembol karşılık getirilerek kesikli hale getirilebilir. Problemin yapısına bağlı olarak bu süreç, veriden çıkarım ya da uzman tavsiyesine bağlı olarak gerçekleştirilebilir. Örneğin, $X \sim N(\mu, \sigma)$ varsayımı altında, S serisinin her bir elamanı, aşağıdaki kurala bağlı olarak, $A_5 = \{a, b, c, d\}$ alfabesinin elemanları ile ifade edilerek kesikli hale getirilebilir (Tablo 7).

a: $* -\infty < x < \Phi^{-1}(0{,}25, \mu, \sigma)$ b: $* \Phi^{-1}(0{,}25, \mu, \sigma) < x < \Phi^{-1}(0{,}50, \mu, \sigma)$ c: $* \Phi^{-1}(0{,}50, \mu, \sigma) < x < \Phi^{-1}(0{,}75, \mu, \sigma)$ d: $* \Phi^{-1}(0{,}75, \mu, \sigma) < x < +\infty$	

Şekil 7 Sayısal değerli serilerin semboller ile ifade edilmesi

Bazı durumlarda kesikli hale dönüştürme işlemi uzman görüşü alınarak da yapılabilir. Örneğin, eldeki seri belirli bir zaman aralığında bir hastaya ait kaydedilmiş diyastole kan basıncı değerleri ise, bu değerleri, örneğin, 65 in altı (düşük), 65 ve 75 arası (normal) ve 75 üzeri (yüksek) için üç ayrı sembol kullanarak da kesikli hale getirmek uygun olabilir.

3.2. OSMT ile Değişken Çıkarımı

Geleneksel istatistiksel modelleme ve makine öğrenmesi yöntemlerinde kullanılan tahminleyici değişkenler her bir örnek için tek bir sayısal değer ile ifade edilebilen türden değişkenlerdir. Örneğin, yaş, eğitim seviyesi, cinsiyet gibi değişkenler her biri için tek bir değer alırlar. Ancak, aynı örnekten alınmış 200 Hz frekansında kaydedilmiş 10 saniyelik derivasyon I ECG, kalbin elektrik aktivitesine dair 2000 ölçüm değer üretmektedir. Bu ve benzeri boylamsal veriler, değişken çıkarımı işlemlerine tabi tutularak modellenirler.

Çıkarılan bu değişkenler, ilgili modelin girdi değişkenleri olarak kullanılır. Bu bağlamda, OSMT, sembolik serilerin birbirlerine benzerliklerinin bir ölçüsü olarak değişken çıkarımında kullanılabilir.

OSMT yönteminde sembolik serilerin geçiş davranışları geçiş olasılık matrisleri ile ifade edilmektedir. Bu nedenle, geçiş davranışları birbirine benzerlik gösteren serilerden elde edilen geçiş olasılık matrislerinin değerlerinin de birbirine benzer olması beklenir. Buradan hareketle, OSMT ile değişken çıkarımı, farklı serilerin olasılık geçiş matrisleri arasındaki uzaklığın ölçülmesi ile oluşturulan geçiş benzerlik matrislerinin (*GBM*) hesaplanmasını temel almaktadır. Doğal olarak, birbirinin aynısı i ve j serilerinin karşılaştırılması ile elde edilen uzaklık değerinin, $GBM_{ij}=0$, mükemmel benzerlik değerini vermesi beklenir.

GBM matrisinin hesaplanabilmesi için, analize konu olan seriler için hesaplanan *GOM*'ların aynı boyutta olması gerekir. Bunu sağlamak için, her bir seride gözlemlenmemiş motifler, satır değerleri *0* olacak şekilde *GOM*'lere dahil edilmelidir.

$k = 1,2, \dots, \max(n_p)$ ve D_i ve D_j serilerine ait geçiş olasılık matrisleri GOM_k^i ve GOM_k^j olmak üzere, D_i ve D_j arasındaki mutlak uzaklığı ifade eden GBM_{ij} Eşitlik 6'da verilen eşitlik ile hesaplanır.

$$GBM_{ij} = d\left(GOM_k^i, GOM_k^j\right) = \sum_{r=1}^{N_p k} \sum_{c=1}^{N_s} \left| GOM_k^i(r,c) - GOM_k^j(r,c) \right| \quad (6)$$

Eşitlik 6'da, $N_p k$, GOM_k matrisindeki motif sayısını (yada satır sayısını), N_s ise en büyük alfabe **boyutunu (uzunluğunu) göstermektedir**. Açıktır ki, her bir *k*-sembol geçiş için ayrı bir benzerlik matrisi elde edilmektedir.

Yukarıda ifade edildiği biçimde hesaplanan uzaklıklara dair bazı özellikler aşağıdaki gibi sıralanabilir.

- Farklı uzunluktaki seriler karşılaştırılabilir.
- Farklı alfabeler ile ifade edilmiş seriler karşılaştırılabilir.
- Eğer seriler zaman serisi ise, verilerin derleme frekansı eşit olmalıdır.

Burada hesaplanan *GOM* matrislerinin sütunları geleneksel istatistiksel modelleme ve makina öğrenmesi yöntemlerinde birer değişken olarak kullanılabilecekleri gibi bu matrisler danışmansız öğrenme yöntemlerinde kümeleme analizi amacı ile de kullanılabilir.

3.3. OSMT ile Tahminleme

Olasılıksal Sembolik Motif Tanıma yöntemi kullanılarak kesikli sembolik verilerin henüz gözlemlenmemiş davranışı tahmin edilebilir. D, n uzunluğunda elemanları d_j sembolleri ile ifade edilen bir kesikli sembolik seri olsun. Ayrıca d_j sembollerinin $A = \{a_1, a_2, \dots, a_{ns}\}$ alfabesinden alındığını varsayalım. Bu durumda d_{n+1} sembolün a_k olması olasılığı Eşitlik 7 de verilen formül ile hesaplanır.

$$P(d_{n+1} = a_k|D) = w_1 \times P(a_k|d_n) + w_2 \times P(a_k|d_{n-1}d_n) + w_3 \times P(a_k|d_{n-2}d_n) + \cdots + w_{np} \times P\left(a_k|d_{n-n_{p+1}}d_{n-n_{p+2}}d_n\right) \quad (7)$$

Eşitlik 7'de yer alan koşullu olasılıklar, uygun GOM_i matrisleri kullanılarak hesaplanır. Örneğin, $P(a_k|d_n)$ olasılığı GOM_1 ve $P(a_k|d_{n-1}d_n)$ olasılığı GOM_2 matrisi kullanılarak hesaplanır. Burada $P(d_{n+1} = a_k|D)$, ağırlıklı ortalama hesabından ibarettir. $w_k \geq 0$ ağırlık değerleri k-sembollü geçiş olasılıklarını ifade eder, ve bu ağırlık değerlerinin toplamı 1 e eşittir, $\sum_k w_k = 1$.

OSMT'nin tahminlemede kullanımı daha önce verilen $S =' aabcabccbabcabcbaabc'$ serisi üzerinden yapılabilir. Gösterim kolaylığı için, $n_p = 3$ seçecek olursak en fazla 3 sembollü geçiş olasıklarını hesaba katarak *S* serisinin hangi sembol ile devam edeceğini tahmin edebiliriz. Bu hesaplamayı yapabilmek için öncelikle Tablo 3, Tablo 4 ve Tablo 5'deki gibi geçiş olasılıkları oluşturulur.

Tablo 3 S serisi için 1-sembollu geçiş olasılıkları matrisi

	a	b	c
a	0,29	0,71	0,00
b	0,29	0,00	0,71
c	**0,40**	**0,40**	**0,20**

Tablo 4 S serisi için 2-sembollu geçiş olasılıkları matrisi

	a	b	c
aa	0,00	1,00	0,00
ab	0,00	0,00	1.00
ba	0,50	0.50	0.00
bc	**0,50**	**0,25**	**0,25**
ca	0,00	1,00	0,00
cb	1,00	0,00	0,00
cc	0,00	1,00	0,00

Tablo 5 S serisi için 3-sembollu geçiş olasılıkları matrisi

	a	b	c
aab	0,00	0,00	1,00
abc	**0,50**	**0,25**	**0.25**
baa	0,00	1.00	0.00
bab	0,00	0,00	1,00
bca	0,00	1,00	0,00
bcb	1,00	0,00	0,00
bcc	0,00	1,00	0,00
cab	0,00	0,00	1,00
cba	0,50	0,50	0,00
ccb	1,00	0,00	0,00

Tablo 3, Tablo 4 ve Tablo 5 ile verilen motif geçiş olasılıklarını $d_{n-2}d_n = abc$ bilgisi kullanılarak *S* serisini takip etmesi muhtemel tüm sembollerin olasılıkları sırası ile Eşitlik 8, Eşitlik 9 ve Eşitlik 10 ile hesaplanır.

$$P(a|S) = \frac{P(a|c)+P(a|bc)+P(a|abc)}{3} = \frac{0.40+0.50+050}{3} = 0.47 \quad (8)$$

$$P(b|S) = \frac{P(b|c)+P(b|bc)+P(b|abc)}{3} = \frac{0.40+0.25+0.25}{3} = 0.30 \quad (9)$$

$$P(c|S) = \frac{P(c|c)+P(c|bc)+P(c|abc)}{3} = \frac{0.20+0.25+0.25}{3} = 0.21 \quad (10)$$

Eşitlik 8, Eşitlik 9 ve Eşitlik 10 ile elde edilen sonuçlarına göre $S =' aabcabccbabcabcbaabc'$ serisinin 'a' sembolü ile devam etmesi en olası seçenek olarak belirlenmiştir.

Burada verilen örnekte 1-, 2, ve 3-sembollü geçiş olasılıklarının eşit olduğu varsayılmıştır. Ancak bu ağırlıklar probleme özel, verilen bir çıktı değerini optimize edecek şekilde de belirlenebilir.

3.3. OSMT ile Sınıflama Analizi

OSMT yöntemi kullanılırken hesaplanan geçiş olasılıkları matrisleri bir bakıma kesikli sembollerden oluşan bir dizinin olasılık matrislerine indirgenmesi işlemleridir. Bu nedenle, geçiş olasılıkları matrisleri birbirine benzeyen serilerin sembolleri arasında benzer geçiş davranışı gösterdikleri ve dolayısı ile birbirlerine benzer oldukları beklenir. Buradan hareketle, iki serinin birbirine benzerliklerinin bir ölçüsü olarak Geçiş Olasılıkları Benzerlik (*GOB*) kavramını ortaya atıyoruz.

Tanım 2: İki sembolik serinin olasılık geçiş matrisleri arasındaki matematiksel uzaklığa Geçiş Olasılıkları Benzerliği denir ve *GOB* ile gösterilir.

Geçiş Olasılıkları Matrisinin matematiksel formülasyonu, birbirinin aynı iki seri arasındaki uzaklığın sıfır olacağı şekilde tanımlanacaktır. Bunun dışında aşağıda verilen özellikleri de sağlaması beklenmektedir.

- Karşılaştırılan serilerin uzunluklarının aynı olması gerekmemektedir.
- Serilerin aynı alfabenin sembollerinden oluşması gerekmemektedir.

Yukarıda verilen özellikler OSMT ile iki serinin karşılaştırılmasında büyük esneklik sağlamaktadır. Ancak, karşılaştırılan serilerin oluşturulduğu alfabelerin ortak sembolleri aynı anlamı ifade etmektedir. Buradan hareketle:

- Karşılaştırılan serilerin zaman serisi olması durumunda, verilerin toplanma periyodunun aynı olması gerekir. Örneğin, günlük telefon satış rakamları serisi ile aylık mobilya satışları serilerini karşılaştıramayız.
- Eğer sembolik seriler gerçel değerli serilerin kesikli hale getirilmesi ile elde edilmiş ise, her iki seri de aynı yöntem ile kesikli hale getirilmiş olmalıdır.

Hesaplamalarda kolaylık sağlamak açısından, karşılaştırılan iki seriye ait geçiş olasılıkları matrislerinin aynı boyuta getirilmesi beklenir. Birbirinden farklı sembolleri olan serilerde her iki matrise de eksik olan semboller tüm satır elemanları *0* olan yeni sütun eklenmesi ile gerçekleştirilir. Diğer taraftan, bir seride gözlenmiş diğerinde gözlenmemiş geçiş motifleri, tüm sütun elemanları *0* olacak şekilde yeni bir satır olarak eklenir. Buradan hareketle, D_i ve D_j serilerinin k-sembollü geçiş olasılıkları matrisleri olan GOM_k^i ve GOM_k^j arasındaki uzaklık Eşitlik 11 ile verile formül yardımı ile hesaplanır.

$$uzaklık\left(GOM_k^i, GOM_k^j\right) = \sum_{r=1}^{N_p k} \sum_{c=1}^{N_s} \left| GOM_k^i(r,c) - GOM_k^j(r,c) \right| \quad (11)$$

Eşitlik 11 de $N_p k$ ile k'ıncı geçiş olasılıkları matrisinde yer alan motif sayısı (satır sayısı) ve N_s ile de en uzun alfabe uzunluğu ifade edilmektedir. Dolayısı ile uzaklık, GOM matrisleri arasındaki elemanter farkların mutlak değerlerinin toplamı olarak hesaplanmaktadır. Burada k değerinin birden büyük olacağı, yani sadece tek sembollü geçişlere bakılmayacağı varsayımı altında, *GOB* Eşitlik 12'de verildiği gibi hesaplanır.

$$GOB_{i,j} = \sum_{k=1}^{N_p} uzaklık\left(OGM_k^i, OGM_k^j\right) \quad (12)$$

Birbirinden faklı m serinin karşılaştırılması, elemanları $GOB_{i,j}$ $(i,j = 1,2,\dots m)$ olacak şekilde ve diyagonal elemanları *0* olan $m \times m$ boyutunda bir kare matris oluşturacaktır. Burada diyagonal elemanların *0* olma nedeni, ayni serinin kendisi ile karşılaştırılmasını ifade etmesi, yani *0* uzaklık ve maksimum benzerlik ifade etmesidir.

Örnek 2. Sınıflandırma analizine bir örnek olması açısından $S_1 = \{abcccba\}$, $S_2 = \{abcccba\}$ ve $S_3 = \{abcccba\}$ serilerini ele alalım. Bu seriler arasında bir ve iki sembollü geçiş olasılıkları dikkate alınarak elde edilen geçiş olasılıkları benzerlik değerleri Eşitlik 13 ile ifade edilen matris ile verilmiştir.

$$GOB = \begin{pmatrix} 0 & 1 & 3 \\ 1 & 0 & 2 \\ 3 & 2 & 0 \end{pmatrix} \quad (13)$$

Eşitlik 13 ile verilen GOB değerlerine göre, semboller arası geçiş davranışı açısından, S_1 ve S_2 serileri birbirine en çok benzeyen seriler S_1 ve S_3 serileri ise birbirinden en farklı serilerdir.

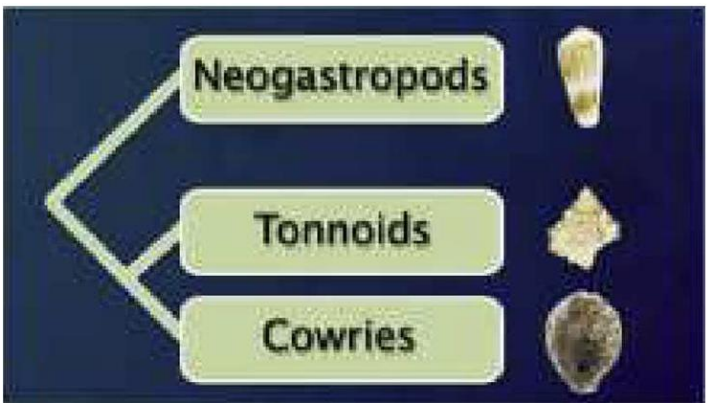
4. UYGULAMA

Bu bölümde OSMT yönteminin gerçek hayat problemlerinde kullanımı için yapılan örnek projeler yer almaktadır.

4.1. OSMT ile Sınıflandırma: DNA Dizilimi Sınıflama Uygulaması

Bu alt bölümde OSMT yönteminin sınıflandırma analizinde kullanımı üzerine bir uygulama sunmak amaçlanmıştır. Sınıflama uygulama alanı olarak, türlerin kısmi DNA dizilimleri kullanılarak uygun biyolojik sınıflara ayrılması problemi ele alınmıştır. Bu uygulamanın seçilmesinin amaçlarında biri, OSMT'nin sadece zaman serileri için tasarlanmadığını göstermektir. Bu kısımda gösterilen uygulama, OSMT'nin ilk çıkış makalesinden alıntılanmıştır (Akbilgic & Howe, 2017).

Bu uygulama da amaç, kısmi DNA dizilimleri verilen 13 salyangoz türünün, OSMT yöntemi ile uygun biyolojik sınıflara ayrılıp ayrılmadığını değerlendirilmesidir. Amacımıza yönelik 13 salyangoz türüne ait kısmi DNA dizisi verileri Amerikan Ulusal Biyoteknoloji Enformatik Merkezi'nin internet sayfasından indirilmiştir. ('http://www.ncbi.nlm.nih.gov/nuccore'). Bu üç salyangoz türü Şekil 8'de verilen taksonomi hiyerarşisine uyan Neogastropods (ya da koni salyangozları) Tonnaids ve Cowries türleridir. Şekil 8'den görüldüğü üzere, Tonnaids ve Cowries türleri birbirine Neogastropods e olduğundan daha yakındır.



Şekil 8 Ele alınan üç salyangoz turunun taksonomi yapısı

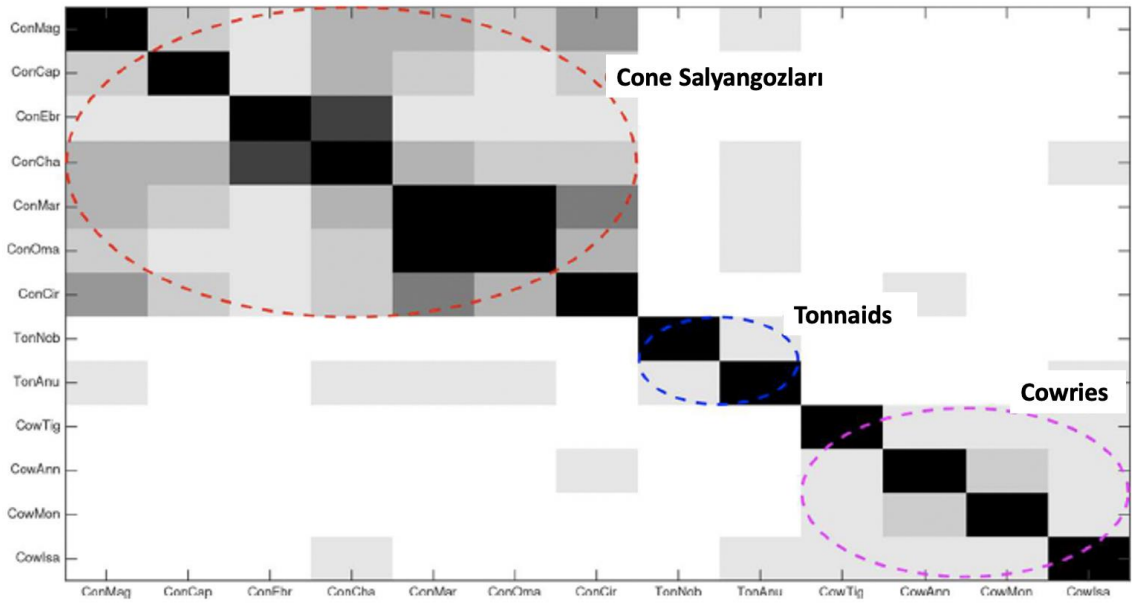
Elimizdeki 13 DNA diziliminden 7 tanesi (ConMag, ConCap, ConEbr, ConCha, ConMar, ConOma, ve ConCir) Neogastropods in alt türleri, iki tanesi (TonNob ve TonAnu) Tonnaids in alt türleri ve dört tanesi ise (CowTig, CowAnn, CowMon ve CowLsa) Cowries in alt türleridir.

Kısmi DNA dizileri verilen 13 salyangozun öncelikle olasılık geçiş matrisleri n_p=17 olacak şekilde 17-sembollü geçiş olasılıklarına kadar hesaplanmıştır. Daha sonra bu serilerin ikili karşılaştırmalarını ifade etmek üzere olasılık geçiş benzerlik değerleri hesaplanmıştır. Birbirleri ile ikili olarak karşılaştırılan 17 seri dolayısı ile 17x17 boyutunda Tablo 6 ile verilen uzaklık (ya da benzemezlik) matrisini oluşturmuştur.

Tablo 6 13 salyangoz türünün DNA'ları arası benzemezlik matrisi

		Cone Snails							Tonnaids					
		Con Mag	Con Cap	Con Ebr	Con Cha	Con Mar	ConO ma	Con Cir	Ton Nob	Ton Anu	Co wTi g	Cow Ann	Cow Mon	CowL SA
Cone Snails	ConMg	0	11.8	12,1	11,3	11,4	11,8	12,2	12,6	12,2	12,7	12,5	12,8	12,5
	ConCap	11.8	0	11,1	11,6	11,8	12,1	11,8	12,6	12,4	12,7	12,5	12,7	13,6
	ConEbr	12.1	12,1	0	10,1	12,2	12,1	12,1	12,5	12,6	12,6	12,6	12,5	12,4
	ConCha	11.3	11,6	10,1	0	11,6	12,0	11,7	12,7	12,2	12,6	12,6	12,5	12,4
	ConMar	11.4	11,8	12,2	11,6	0	9,1	10,6	12,6	12,4	12,6	12,5	12,6	12,6
	ConOma	11.8	12,1	12,1	12,0	9,1	0	11,5	12,5	12,4	12,6	12,6	12,6	12,6
	ConCir	11.2	11,8	12,1	11,7	10,6	11,5	0	12,7	12,5	12,5	12,4	12,7	12,5
Tonnaids	TonNob	12,6	12,6	12,5	12,7	12,6	12,5	12,6	0	12,2	12,6	12,5	12,5	12,7
	TonAnu	12,2	12,4	12,6	12,2	12,4	12,4	12,5	12,2	0	12,6	12,4	12,5	12,4
Cowries	CowTig	12.7	12,7	12,6	12,6	12,6	12,6	12,5	12,6	12,6	0	12,3	12,4	12,2
	CowAnn	12.5	12,5	12,6	12,6	12,5	12,6	12,4	12,5	12,4	12,3	0	12,0	12,3
	CowMon	12.8	12,7	12,5	12,5	12,6	12,6	12,7	12,5	12,5	12,4	12,0	0	12,1
	CowLsa	12.5	12,6	12,4	12,4	12,6	12,6	12,5	12,7	12,4	12,2	12,3	12,1	0

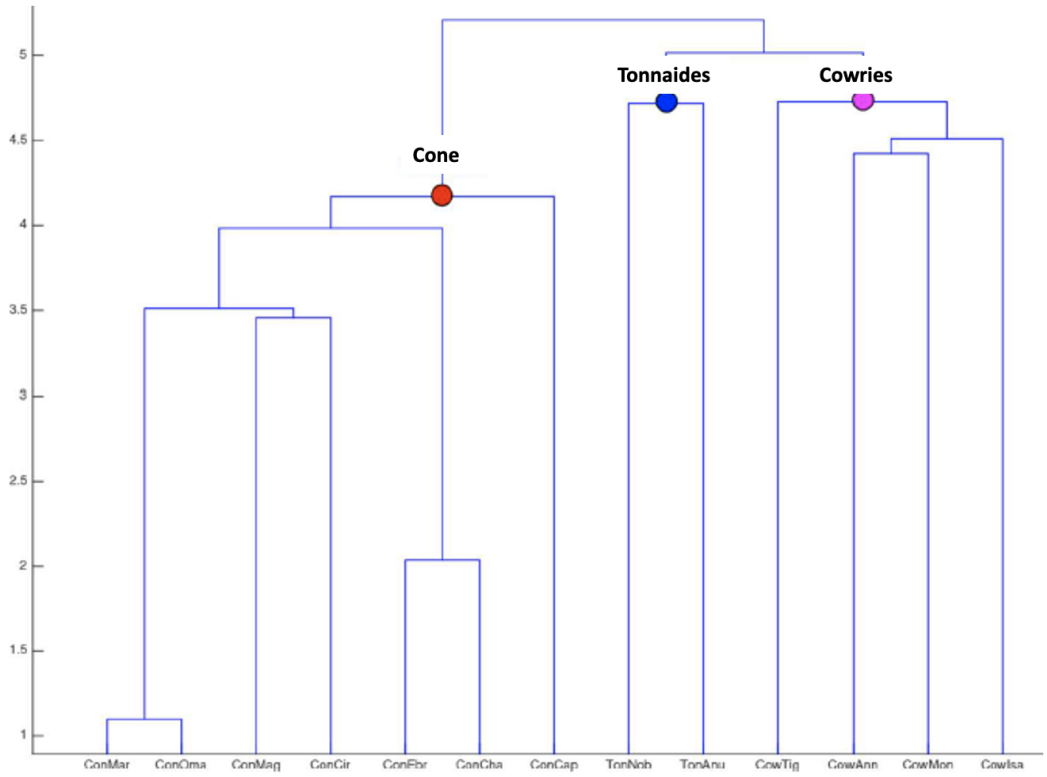
Tablo 6'da yer alan uzaklık değerleri kullanılarak Şekil 9'da yer alan ısı haritası çizilmiştir. Bu ısı haritasında mavi hiç 0 uzaklık değerini, dolayısı ile benzerliği ifade etmektedir. Kırmızı ise en yüksek uzaklığı, dolayısı ile benzemezliği ifade etmektedir.



Şekil 9 Salyangozların DNA benzemezlik değerlerinden üretilen sıcaklık haritası

Şekil 9'da yer alan sıcaklık haritasından anlaşılacağı üzere OSMT yöntemi aynı taksonomik yapıdan gelen türler arasında düşük uzaklık (benzemezlik) değerlerini başarı ile üretmiştir.

Şekil 10 ile yine Tablo 6'da yer alan uzaklık değerleri kullanılarak ağaç diyagramı çizilmiştir.



Şekil 10 Salyangozların DNA dizileri arasındaki benzemezlikten elde edilen ağaç diyagramı

Şekil 10'da yer alan ağaç diyagramı, OSMT'nin danışmansız sınıflandırma işlemini DNA dizileri üzerinden mükemmel doğruluk oranı ile gerçekleştirdiği görülmektedir. Howard Hughes Medical Institute (Institute, 2009) tarafından yayınlanan bu uygulama probleminin orijinal halinde hizalama yöntemi kullanılmıştır. Hizalama yöntemi, DNA analizinde çok sıklıkla kullanılmaktadır (Altschul et al., 1990). Bizim çalışmamızda ise herhangi bir hizalama yöntemi kullanmada, doğrudan DNA dizilimlerine OSMT yöntemi uygulayarak türleri doğru bir şekilde sınıflandırabildik. Buradan hareketle DNA dizilimlerinde nükleotidlerin birbirini takip etme davranışlarının tesadüfi olmadığı hipotezi ortaya atılabilir.

4.2. OSMT Yöntemi ile Sınıflandırma: Konjestif Kalp Yetersizliği Uygulaması

4.2.1. Problemin Tanıtımı

Konjestif Kalp Yetersizliği (KKY) ileri evre koroner hastalık olarak ortaya çıkan kronik bir hastalıktır. KKY, kalbin vücuda yeterli kan pompalayamaması şeklinde gelişir (Mudd & Kass, 2008). KKY en önde gelen halk sağlığı sorunlarından birisidir ve dünya genelinde 23 milyondan fazla kişiyi etkilemektedir (Mozaffarian et al., 2016; Roger, 2013). KKY ilk evrelerinde genel olarak asemptomatik ancak ilerleyici bir hastalıktır. Erken teşhis edilmesi halinde kalbin kan pompalama fonksiyonunu güçlendirecek müdahaleler mümkündür. Bu yolla hayati tehlike yaratabilecek atriyal fibrilasyon ve karaciğer hastalıklarının da önüne geçilebilir. KKY hastalığının yaygınlığı ve dolayısı ile toplum sağlığı üzerindeki olumsuz etkisi, erken teşhiste kullanılacak sinyal işleme yöntemleri yardımı ile düşürülebilir (Saykrs, 1973).

Yüksek KKY riskinde olan hastaların EKG'leri genellikle genişlemiş QRS kompleks gösterirler. EKG üzerindeki R-R aralıkları arasındaki değişkenliğin ölçüsü olan kalp atışı değişkenliği (KAD), kalbin dinamik yapısı hakkında önemli bilgiler vermektedir (Akselrod et al., 1985; Coumel et al., 1994; Rajendra Acharya et al., 2006; Saykrs, 1973). Son on yılda, KKY teşhisi için R-R aralıklarının analizine dayalı birçok yöntem ileri sürülmüştür. Bu teknikler genel olarak aşağıda verilen üç kategoride toplanabilir.

- Zaman-alanı metrikleri: RR süreleri ortalaması (RRO), R-R sürelerinin standart sapması (RRSS), Birbirini takip eden R-R süreleri farkının kareli ortalaması (RRKO), 50 mili saniyeden çok farkı olan komşu R-R aralıkları sayısı (RR50) (Rajendra Acharya et al., 2006).
- Frekans alanı metrikleri: Bu metrikler kesikli otonom sinir sistemin Fourier ve Dalgacık dönüşümü yöntemleri ile ECG verisinin 0-0,40 frekans güç spektral bandında analizi ile elde edilir (Akselrod et al., 1981; Malliani et al., 1991; Montano et al., 1994).
- Diğer Yöntemler: Kesikli Fourier ve Dalgacık dönüşümlerinin yanı sıra Örnek Entropi, Örnek Asimetri, R-R aralıklarının Lorenz (or Poincare) şemaları be HRV üçgensel indeks gibi diğer metriklerde KKY hastalığının teşhisinde kullanılmıştır (Guzzetti et al., 2000; Ho et al., 1997; Işler & Kuntalp, 2007; Kamen & Tonkin, 1995; Poon & Merrill, 1997; Yee et al., 2001).

KKY teşhisi alanında EKG kullanımına dair zengin bir literatür bulunmaktadır. Bunlardan birisinde, KKY hastalarının EKG'den belirlenmesi amacı ile, R-R aralıklarındaki karmaşıklığı ölçecek yeni bit Çok Ölçekli Örnek Entropi yöntemi önerilmiştir. Bu yöntemle elde edilen değişenlerin Fisher Ayırma analizi ile analizi sonucu eğitim veri setinde %92 doğru sınıflandırma oranı elde edilmiş olsa da, önerilen model test verisinde sadece %76 doğruluk oranını yakalamıştır (Costa & Healey, 2003). Bu konuda yapılan diğer bir çalışmada ise ölçeğe bağımlı Lypanov kuvveti ile HRV dinamiklerinin karakterize edilmesi amaçlanmış ve %100 sensitivite ve 95% spesifite elde edilmiştir (Hu et al., 2010). Diğer bir çalışma da ise üçüncü dereceye kadar kümülatife spektrumlar hesaplanarak HRV değerlerinin ikili spektrum değişkenleri hesaplanmıştır. Bu değişkenlerin boyutu daha sonra Genetik Algoritma ile indirgenmiş ve seçilen değişkenler destek vektör makineleri yardımı ile analiz edilerek KKY teşhisinde %97.5 doğruluk oranı yakalanmıştır (Yu & Lee, 2012). Uzun sureli HRV kayıtlarından elde edilen ölçümlerinin sınıflandırma ve regresyon analizi yöntemleri ile analiz ise %63.6 spesifite ve %93.3 sensitivite ile KKY teşhisi yapabilmiştir. Son olarak, R-R serilerine bulanık örnek entropi ve örnek entropi uygulanması ile elde edilen değişkenler ile KKY sınıflandırması %90 spesifite ve %87 sensitivite sonuçlarını vermiştir.

Bu uygulama da bizim amacımız, geliştirdiğimiz OSMT yöntemi ile önce EKG verisinden değişken çıkarmak ve daha sonra bu değişkenler ile KKY teşhisi yapmaktır.

4.2.2. Problemin Amacı

Bu uygulamadaki amaç, EKG üzerinden elde edilmiş R-R sureleri serisine OSMT uygulayarak KKY teşhisi yapacak bir model geliştirmektir.

4.2.3. Problemde Kullanılan Veri

Bu çalışmada kullanılan veri Columbia Presbyterian Medical Center, Washington School of Medicine ve New York Heart Association tarafından sağlanmış olup PhysioNet veri tabanı üzerinden erişime açıktır (Goldberger et al., 2000). Verimiz toplam 116 kişinin 128 örneklem frekansında 24 saat süreli Holter kayıtlarından elde edilen R-R sureleri serilerinden oluşmaktadır. 116 verinin 72 si sağlıklı bireylerden (35 erkek, 37 kadın, yas ortalaması 54.6) (PhysionNet, 2022a, 2022b) ve 44'u ise KKY hastalarından (28 erkek,

15 kadın, bir cinsiyeti bilinmeyen birey, yas ortalaması 55.6) derlenmiştir (Mietus et al., 2002).

R-R aralıkları serilerinden elde edile KAD değişkenleri genel olarak veriler arası tekrarlanabilir değişkenlerdir ve kardiyovasküler sağlık ile ilgili önemli bilgi içermektedir. Diğer taraftan, R-R serilerinden elde edilen KAD değişkenleri yanlış klinik yorumlara neden olabilir. Bunun nedeni, R-R serilerinin birçok yapay olguların EKG üzerinde etkisine çok hassas olmasıdır. Bu yapay olgular ECG kaydı sırasında kişinin aşırı hareket etmesi, yakında bulunan elektrikli cihazların yarattığı elektromanyetik alan, elektrotların tene tam temas etmemesi olarak özetlenebilir. Bu nedenle literatürde yapay olgulardan etkilenmemiş ECG kesitlerinden elde edilen R-R aralıkları serilerinin analizi tavsiye edilmektedir (Berntson et al., 1997; Peltola, 2012). Bizim veri setimizde yer alan 72 sağlıklı hasta R-R serilerinin üçünde, ver 44 KKY hastası R-R serilerinin dördünde yapay olgulara rastlanmıştır. Bu yapay olgu içeren EKG'lerin çıkarılması ile bu çalışmada kullanılan veri seti toplamda 69'u sağlıklı bireylerden ve 38'i KKY hastalarından olmak üzere toplamda 107 kişinin R-R aralık serilerini içermektedir.

4.2.4. Değişken Çıkarımı

R-R aralıkları serileri sayısal değerli serilerdir. OSMT yöntemini uygulanabilmesi için, öncelikle bu sayısal değerli seriler sembollere dönüştürülmüştür. Bu amaçla $n_s=8$ olacak şekilde sekize sembollü $A = \{a, b, c, d, e, f, g, h\}$ alfabesinin sembolleri kullanılmıştır. Hangi sayısal R-R değerlerinin hangi sembollere karşılık geleceği konusunda isi alan bilgisi kullan ilmiştir. Buna göre, normal sinüs ritim gösteren bir EKG de R-R suresi 0,6 saniye ile 1,2 saniye arasındadır. Buradan hareketle, Tablo 7 Sayısal R-R değerlerinin semboller ile ifade edilmesi ile verilen dönüştürme yöntemi kullanılarak R-R aralık serilerimiz sembolik serilere dönüştürülmüştür.

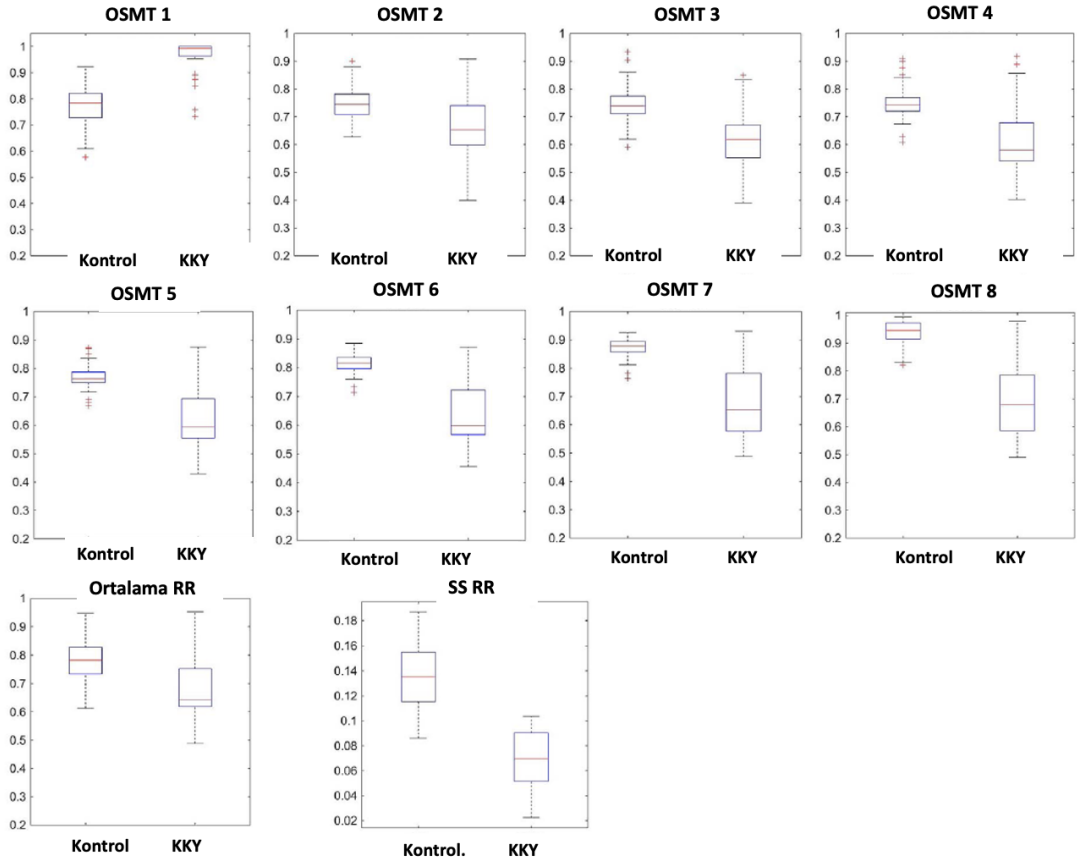
Tablo 7 Sayısal R-R değerlerinin semboller ile ifade edilmesi

R-R serisi değeri	Karşılık gelen sembol
0≤ R-R <0,30	a
0,30≤ R-R <0,60	b
0,60≤ R-R <0,750	c
0,75≤ R-R <0,90	d
0,90≤ R-R <1,05	e
1,05≤ R-R <1,20	f
1,20≤ R-R <1,50	g
1,50≤ R-R	h

R-R serilerinin sembolik serilere dönüştürülmesini takiben her bir seri için 9-sembollu motiflere kadar olasılık geçiş matrisleri hesaplanmıştır. Her bir seri için olasılık geçiş matrislerinin hesaplamasını takiben, her bir 107 serinin olasılık geçiş matrisini, 69 sağlıklı kişilerden alınan olasılık geçişleri ile karşılaştırdık. Burada sadece normal sağlıklı kişilerle karşılaştırma yapmanın amacı, normalden sapmaları ölçebilmektir. Dolaysı ile bu karşılaştırma sonucunda sekiz tane 107x69 boyutunda matris elde ettik. Mesela ilk matrisin birinci satir ikinci sütunu, 1-sembollu geçiş davranışları açısından birinci normal hastanın R-R serisinin ikinci normal hastanın R-R serisine benzerliğinin bir ölçüsüdür. Her bir matrisin sütun ortalamalarının alınması ile sekiz tane 107x1 boyutunda vektör elde edilmiştir. Ayni örnek üzerinden gidecek olursak, birinci matrisin birinci satiri, ilk normal sağlıklı kişinin, 69 sağlıklı kişiye 1-sembollu geçiş davranışları açısından ne kadar benzediğini ifade etmektedir. Sonuç olarak, bu sekiz vektörün her birisi, KKY sınıflandırılmasında kullanılacak birer değişken olarak belirlenmiştir.

OSMT ile çıkarılan sekiz değişkene ilave olarak KAD değişkenlerini de sınıflama analizinde kullanılmak üzere belirledik. Basta SDRR olmak üzere, R-R aralıkları üzerinden hesaplanan ortalama RR ve RR standart sapması (SDRR) indeks, toplam kuvvet temelli ölçülerin KKY belirlenmesinde kullanıldığı gösterilmiştir. Örneğin, sadece SDRR KKY ile normal hastaları %81,8 spesifisite ve %98,1 sensitivite ile ayırabilmektedir (Berntson et al., 1997). Bu nedenle, bu çalışmada için Uluslararası KAD Analizi Yönergesine (Breiman, 2001) bağlı kalarak SDRR ve AVRR değişkenlerini hesapladık.

Sonuç olarak sınıflandırma analizinde kullanılmak üzere sekiz tanesi OSMT ve iki tanesi zaman-alanı değişkenleri olacak şekilde toplam on değişken hesaplanmıştır. Bu 10 değişkenin normal ve KKY hastaların için değerlerini karşılaştıran kutu grafiği Şekil 11 OSMT değişkenlerinin Sağlıklı (kontrol) be KKY hastalarında dağılımı ile verilmiştir.



Şekil 11 OSMT değişkenlerinin Sağlıklı (kontrol) be KKY hastalarında dağılımı

4.2.6. Torbalanmış Karar Ağaçları ile KKY Sınıflama Analizi

Bu asamaya kadar R-R serilenden çıkarılmış olan on değişkeni girdi, R-R serisinin KKY hastasından gelip (KKY=1) gelmediğini (KKY=0) gösteren ikili değerli değişkeni çıktı olarak kullanan Torbalanmış Karar Ağaçları (TKA) modeli kurulmuştur. Modelin ezberlemesini önlemek ve tahmin varyansını azaltmak için sadece 30 zayıf tahminleyiciden oluşan TKA modeli kurduk (Breiman, 2001; Efron & Tibshirani, 1994). Her bir zayıf tahminleyici karar ağacının en fazla 20 defa dallanmasına izin verildi.

TKA modelinin KKY sınıflama performansı doğru pozitif, yanlış pozitif, doğru negatif, yanlış negatif, sensitivite, spesifite ve AUC bakımından ölçüldü.

Genelleştirilebilir bir TKA modeli elde edebilmek için çapraz doğrulama yöntemi uygulandı. Buna göre, elimizdeki berinin öncelikle %80 eğitim ve %20 si bağımsız test verisi olarak ayrıldı. %80 eğitim verisi kullanılarak 5-fold çapraz doğrulama yöntemi ile KKY sınıflandırma modeli kuruldu ve bu çapraz-doğrulanmış model %20 bağımsız test verisine uygulandı. Bu işlem ayrıca 10 defa %80:%20 model kurma: test verilerine rastgele yeniden seçerek tekrarlandı. Bu on tekrarda test verisinde elde ortalama %100 spesifite, %98.5 sensitivite ve genel olarak %99.5 doğruluk oranı elde edildi. OSMT yöntemi ile elde ettiğimiz bu sonuçların literatürde ayni veri kullanılarak önerilen yöntemlerle karşılaştırılması Tablo 8'de verilmiştir.

Tablo 8 Konjestif kalp yetersizliği tahmin modelleri karşılaştırması

Çalışma	Değişkenler	Sensitivite	Spesifisite	AUC
(Asyali, 2003)	SSRR	%81,8	%98,1	0.89
(Işler & Kuntalp, 2007)	Dalgacık, entropi, kısa donem KAD	%100	%94,4	0,97
(Hossen & Al-Ghunaimi, 2007)	Dalgacık	%82,4	%90,6	0,64
(Thuraisingham, 2009)	İkinci derece farklar diyagramı	%100	%100	-
(Melillo et al., 2013)	KAD, spektral güç	%89,7	%100	0,94
(Yu & Lee, 2012)	KAD ikili spektrum	%96,5	%100	-
(Melillo et al., 2011)	uzun donem KAD	%93,3	%63,3	0,78
(Acharya et al., 2017; von Tscharner & Zandiyeh, 2017)	EMD, doğrusal olmayan KAD	%97,0	%98,2	-
(von Tscharner & Zandiyeh, 2017)	faz karmaşıklığı	%87,0	%89,0	-
Bu çalışma	OSMT ve KAD	%94,7	%100	0,98

Tablo 8'den görüleceği üzere OSMT yöntemi ile değişkenler kullanılarak elde edilen KKY sınıflandırma başarısı, aynı veri kullanarak diğer yöntemlerle elde edile sınıflandırma başarılarından ya daha iyi ya da çok yakın sonuçlar vermektedir. Aslında, Tablo 8'de verilen sadece bir çalışma (Thuraisingham, 2009) mükemmel sınıflandırma başarısı ile bizim önerdiğimiz yöntemden daha iyi sonuç vermiştir. Bu çalışmada KKY sınıflandırması için R-R serisinin ikinci derece farklar serini kullanan bir istatistiksel yöntem önerilmiştir. Her ne kadar bu çalışma da ezberlemeyi engellemek amacı ile birini dışarıda bırakma yöntemi kullanılmış olsa da, parametre optimizasyonu sırasında verinin

tamamı kullanılmıştır. Bu nedenle yeni verilere genelleştirilebilip genelleştirilemeyeceği tartışma konusudur.

4.3. OSMT Yöntemi ile Tahminleme: Çocuk Kanserini Yenen Yetişkin Bireylerin Kardiyomiyopati (KMP) Riskinin Tahminlenmesi

4.2.1. Problemin Tanıtımı

Kanseri yenme kavramı birçok farklı anlamda kullanılsa da, bu çalışmada genel olarak kanser teşhisini takiben 5 yıl hatta kalan bireyleri ima etmektedir. 1980'lere kadar 30%'larda olan çocukluk kanserini yenme oranı geliştirilen etkin tedavi yöntemleri sayesinden günümüzde %85'lere ulaşmıştır. Sadece Amerika Birleşik Devletleri'nde 500,000 in üzerinde çocukluk kanserini yenmiş birey hayatını sürdürmektedir (Howlader et al., 2020). Her ne kadar çocukluk kanseri tedavisi alanında yeni geliştirilen antrasiklin gibi kemoterapi ilaçları hayatta kalma oranlarını artırmış olsa da, yüksek derece de kardiyotoksik etkiler göstermektedirler (Mulrooney et al., 2016).

Kardiyotoksik tedavi yöntemlerinin etkileri kısa, orta, ve hatta uzun vade de ortaya çıkabilmektedir. Kısa vadeli etkiler genellikle tedavi başladıktan 30 gün içerisinde ortaya çıkan komplikasyonlardır. Orta vadeli kardiyotoksik etkiler is genellikle tedaviyi takiben bir yıl içerisinde ortaya çıkmaktadırlar. Bu kısa ve orta vadede ortaya çıkan etkiler genellikle tedavinin gidişatını ve tedavide kullanılan yöntemin turunu ve dozunu değiştirmeyi gerektirebilmektedir. Diğer taraftan, kısa ve orta vade kardiyotoksisiteye bağlı etkilerin görülmüş olup olmasından bağımsız olarak, kanseri yenenler uzun donem kardiyotoksisiteye bağlı riskler altındadır (Mulrooney et al., 2016). Diğer bir ifade ile, çocukluk kanseri tedavisini takiben beş yıl hayatta kalmayı başarmış bireyler, yetişkin hayatlarında kanser tedavisine bağlı kardiyovasküler risk altındadır. Örneğin, çocukluk kanserini yenenler arasında konjestif kalp yetersizliği oranı yaşıt ve hemcinslerine göre 15 kat daha fazladır (Armenian & Bhatia, 2018). Kardiyovasküler risklerin en başında başlamalarından birisi is kardiyomiyopati (KMP) hastalığı olarak ortaya çıkmaktadır (Armenian et al., 2015).

Kanseri yenen hastaların geri kalan hayatlarında karşılaşabilecekleri hastalıkların erken teşhisi ve müdahalesi amacı ile farklı kurumlar tarafından geliştirilmiş kanseri yenenlerin uzun sureli takibine yönelik sürveyans yönergeleri mevcuttur (Armenian et al., 2015). Bu yönergeler, örneğin Children's Oncology Group Yönergeleri, tedavi surecinde maruz kalınan antrasiklin ve kalbe yöneltilmiş radyasyon miktarına göre her üç ila beş yılda bir ekokardiyogram ile kontrol önermektedir (Armenian et al., 2015). Her ne kadar bu

yönergeler oluşmuş hastalığın erken teşhisi için olanak sağlasa da, bu yönergeler hali hazırda hata olmamalarına rağmen yakın zamanda hasta olma riskinde olan kişileri belirlemekte yetersiz kalmaktadır. Bununla beraber, maalesef, çocukluk kanserini yenenleri yetişkin hayatlarında takip etme konusunda birçok aksaklıklar yaşanmaktadır. Bu aksaklıkların kimisi kişiye özgü davranışsal etkenlere bağlı olmakla beraber çoğu zaman sağlık sistemine ve ekokardiyogram ve kardiyak manyetik rezonans gibi pahalı görüntüleme yöntemlerine sinirli erişim ile ilgilidir. Bu bağlamda, ucuz ver erişimi kolay araçlar ile çocukluk kanserini yenenlerin uzun sureli takibini sağlayacak sürveyans yöntemlerine ihtiyaç vardır(Alchin et al., 2022; Landier et al., 2006; Linge & Follin, 2021).

4.3.2. Uygulamanın Amacı

Bu uygulamadaki amaç, yüksek kardiyomiyopati riskinde olan çocukluk kanserini yenmiş hastaların, EKG verisi kullanılarak önceden belirlenmesidir. İlgili uygulama çalışması Journal of Clinical Informatics-Clinical Cancer Informatics (Gunturkun et al., 2021) dergisinde sorumlu yazar Oğuz Akbilgiç olacak şekilde tarafından basılmıştır.

4.3.3. Problemde Kullanılan Veri

Bu uygulamada SJLIFE olarak bilinen, Çocukluk Kanserini Yenmiş Yetişkinlerin Ömür boyu Gözlem Topluluğu verileri kullanılmıştır. SJLIFE topluluğu Amerika Birleşik Devletleri, Tennessee Eyaleti, Memphis şehrinde yer alan St. Jude Çocuk Araştırma Hastanesi tarafından yürütülmektedir. SJLIFE topluluğu katılımcıları, çocukluk kanseri tedavisini St. Jude Çocuk Araştırma Hastanesinde almış yetişkin bireylerden oluşmaktadır. SJLIFE katılımcıları yaklaşık olarak her beş yılda bir St. Jude Çocuk Araştırma Hastanesine gelerek geniş kapsamlı bir gözetimden geçmektedir. Bu gözetimler sırasında sağlık durumunu özetleyen anketler ve kan tahlillerinin yanında detaylı EKG ve Eko tetkikleri de yapılmaktadır.

EKG verileri GE MAC 1200 EKG makinesi ile kaydedilmiş 500Hz frekansında örneklenmiş, 10 saniyelik 12-derivasyonlu EKG kaydı olarak derlenmiştir. Bu EKG verileri GE MUSE EKY yönetim yazılımının veri tabanından XML formatında çıkarılmıştır. XML dosyalarında ham EKG verileri Base64 kodlanarak yer almaktadır. Bu base64 kodlanmış veriler Python programlama dili kullanılarak de-kod edilmiştir. Bu

şekilde elde edile her bir EKG verisi 5000x12 boyutunda bir matris olarak numpy dosyası olarak analize hazır bir halde kaydedilmiştir.

EKO verileri GE Medikal Sistemlerin 2-boyutlu Doppler ultrason cihazı kullanılarak kaydedilmiştir be sol karıncık hacmi üç boyutlu olarak resmedilmiştir. Bu Eko görüntülerinden elde edile sol karıncık enjeksiyon fraksiyonu (SKEF) KMP teşhisinde kullanılmıştır.

Bu uygulamaya dahil edilen SJLIFE katılımcıları, 18 yaşını geçmiş, çocukluk kanserinin üzerinden en az on yıl geçmiş ve en az iki SJLIFE takibini tamamlamış bireylerden oluşmaktadır. İlk SJLIFE ziyaretinde kardiyomiyopati teşhisi konulmuş hastalara çalışmaya dahil edilmemiştir. Diğer bir ifade ile, ilk ziyaretlerinde kardiyomiyopati hastalığı olamayan hastaların verileri kullanılarak, bu hastaların gelecekte kardiyomiyopati olma risklerinin hesaplanması amaçlanmıştır. Bu tanıma uygun olarak 1,217 SJLIFE katılımcısı belirlenmiş ve çalışmaya dahil edilmiştir. Bu hastaların %52'si erkek ve %86 beyaz ırktandır. Kanser teşhisindeki medyan yas 8.4 (0,0-22,7) ve SJLIFE grubuna katılma yaş medyan değeri 31.7 (18,6-66,4) olarak bulunmuştur. Bu grup içerisine 817 (%67.1) kişisi göğüs bölgesine radyasyon almış be 932 kişi (%76,6) kişi antrasiklin tedavisi görmüştür. Yarısı (n=609) en az bir kardiyovasküler riske sahip, 92 (%7,6) tanesi diyabet, 249 (%20.5) hypertriglyceridemia, 247 (%20.3) tanesi hipertansiyon ve 383 (%31.5) hypercholesterolemia hastasıdır.

Çalışmanın çıktı değişkeni kardiyomiyopatidir. Kardiyomiyopati tamimi olarak Eko görüntülemesinde elde edilen SKEF değerinin %50 den küçük olması veya önceki bir Eko ile karşılaştırıldığında SKEF değerinde %10'un üzerinde bir düşüş olması olarak belirlenmiştir. Çalışmaya dahil edilen 1,217 SJLIFE katılımcısının 117'sinin (%9,6) gelecekteki SJLIFE ziyaretlerinde kardiyomiyopati görülmüştür (Gunturkun et al., 2020).

Çalışmanın girdi değişkenleri tedavi yöntemlerine bağlı faktörler, biyolojik ve demografik faktörler ve EKG değişkenleri olarak sınıflandırılabilir.

Çalışmada dört faklı tedaviye bağlı değişken kullanılmıştır. Bunlar birincisi, toplam antrasiklin dozu (mg/m^2) ve andriamycin, daunorubicin, epirubicin, idarubicin, ve mitoxantrone turu ilaçların toplam uygulanan dozu olarak hesaplanmıştır. İkinci ve ucundu değişkenler, göğüs bölgesine uygulanan radyasyon tedavisinin iki farklı şekilde ölçülmesi ile edilmiştir. İlk ölçüm Evet-Hayır (1-0) seklinde ölçülen göğüs bölgesine

radyasyon uygulanıp uygulanmadığı sorunun cevabi ikinci ölçüm ise sürekli değişken olarak uygulanan maksimum doz olarak ifade edilmiştir. Son olarak, göğüs bölgesine uygulanmamasına rağmen, radyasyon cihazlarının hassaslık seviyesine bağlı olarak kalbe sızıntı yapmış olan radyasyon miktarı ayrı bir değişken olarak hesaplanmıştır. Bu kalbe sızan radyasyon miktarı Texas Eyaleti Houston şehrinde yer alan MD Anderson Kanser Merkezi tarafından hesaplanmıştır (Howell et al., 2019).

Çalışmada kullanılan toplam on iki biyolojik ve demografik faktör girdi değişkeni olarak kullanılmıştır. Bu değişkenler kanser teşhisinde ve çıktı değişkeninin alındığı Eko çekim tarihindeki yas (yıl), irk, cinsiyet, vücut yüzey alanı, temel kanser turu, kalp atisi (bir dakikadaki toplam atış sayısı), solunum oranı (bir dakikadaki toplam nefes alma sayısı), sistole ve diyastole kan basıncı (mm Hg), tutun ürünleri kullanma durumu (evet-hayır) değişkenleridir. Ayrıca kardiyovasküler hastalık risk faktörü olup olmaması da bir değişken olarak kullanılmıştır. Bu değişkenin belirlenmesinde diyabet (diyet veya ilaç ile tedavi edilen), hypertriglyceridemia (aç karnına triglycerides ≥ 150 mg/dL veya lipit kontrol tedavisi altında olan), hipertansiyon (sistole kan basıncı > 140) veya diyastole kan basıncı > 90 veya tansiyon kontrol ilacı kullanan) hastalıklarından birisinin var olup olmama durumuna bakılmıştır.

Buraya kara ifade edilen değişkenlerin frekans (n) ve yüzde (%) olarak ifade edilenleri Tablo 9 ile ve medyan-aralık olarak ifade edilenler ise Tablo 10ile verilmiştir. Kategorik değişkenlerin istatistiksel karşılaştırılmasında Pearson Chi-Square ve Fisher Exact Testi sayısal değerli değişkenlerin karşılaştırılmasında ise Student T-testi kullanılmıştır.

Tablo 9 Frekans ve yüzde ile özetlenen edilen değişkenler

Risk Faktörleri	Toplam		Kardiyomiyopati		Kontrol		p
	n	%	n	%	n	%	
Cinsiyet							0,10
Kadın	602	49.5	59	41,9	553	50,3	
Erkek	615	50.5	68	58,1	547	49,7	
Irk							0,06
Beyaz	1041	58,6	94	80,3	947	86,1	
Siyah	156	12,8	22	18,8	134	12,2	
Diğer	20	1,6	1	0,9	19	1,7	
Kanser Turu							<0,01
Lösemi	501	41,2	30	25,6	471	42,8	
Sarkoma	150	12,3	23	19,7	127	11,5	
Hodgkin Lemfoma	206	16,9	31	26,5	175	15,9	
Hodgkin Olmayan Lenfoma	88	7,2	15	12,8	73	6,6	

Tablo 9 Devam							
CNS	85	7,0	5	4,3	80	7,3	
Neuroblastoma	59	4,9	4	3,4	55	5	
Wills Tümör	99	8,1	6	5,1	93	8,5	
Diğer	29	2,4	3	2,6	26	2,4	
Kardiyovasküler risk							
Hypertriglyceridemia	249	20,5	29	24,8	220	20,0	
Hipertansiyon	247	20,3	19	16,2	228	20,7	
Hypercholesterolemia	383	31,5	42	35,9	341	31,0	
Diyabet	92	7,6	16	13,7	76	6,9	
Herhangi biri	609	50	65	55,6	544	49,5	
Tutun ürünleri kullanımı							0.93
Evet	479	39,4	47	40,2	432	39,3	
Hayır	738	60,6	70	59,8	668	60,7	
Göğüs bölgesine Radyasyon							<0,01
Evet	415	34,1	55	47,0	360	32,7	
Hayır	802	65,9	62	53,0	740	67,3	

Tablo 10 Medyan ve aralık değeri ile özetlenen edilen değişkenler

	Toplam		Kardiyomiyopati		Kontrol		
Risk Faktörleri	Medyan	Aralık	Medyan	Aralık	Medyan	Aralık	p
Kanser Teşhisindeki Yas	8.4	0.0-22.8	10.2	0.4-21.2	8.2	0.0-22.8	0.04
İlk SJLIFE ziyaretindeki Yas	31.7	18.4-66.4	33.9	19.7-54.7	31.4	18.4-66.4	0.01
Hayatta kalma suresi	22.7	10.4-49.8	23.6	11.4-45.3	22.6	10.4-49.4	0.36
İlk SJLIFE ziyaretinden sonra çıktı değişkenin belirlendiği Eko testine kadar olan sure	5.2	0.5-9.2	5.3	0.9-9.5	5.2	0.5-9.2	0.15
Kalp Atışı (bir dakikadaki atış sayısı)	76.0	41-138	78.0	46-115	76.0	41-138	0.56
Solunum Oranı (bir dakikadaki nefes alma sayısı)	18.0	12-28	18.0	14-24	18.2	12-28	0.04
Sistole kan basıncı (mm Hg)	123.0	84-224	124.0	93-200	122.0	84-224	0.27
Diyastole kan basıncı (mm Hg)	76.0	49-118	77.0	55-111	76.0	49-118	0.11
Vücut yüzey alanı	1.9	1.1-3.0	1.9	1.3-2.8	1.9	1.1-3.0	0.02
Toplam antrasiklin dozu (mg/m^2)	168.7	35.1-734.2	206.3	49.2-669.3	157.2	35.1-734.2	<0.01
Toplam göğse uygulanan radyasyon dozu (cGy)	2600.0	150-6200	2600.0	450-4500	2600.0	150-6200	<0.01
Kalbe sizan ortalama radyasyon dozu (cGy)	2070.0	50-4920	2320.0	170-4920	2025.0	50-4520	<0.01

Tablo 9 ve Tablo 10 ile özetlenen değişkenlerin yanında, ham EKG sinyallerinden de değişken çıkarılmıştır. Bu aşamada yapılan değişken çıkarma işlemleri ayrı bir alt başlık olarak devam eden kısımda raporlanmıştır.

4.3.4. EKG Verisinden Çıkarılan Değişkenler

Ham EKG verileri ilk SJLIFE ziyaretlerinden alinmiş olup 5000x12 boyutunda matris olarak elde edilmiştir. Burada her bir sütun, bir 12 EKG derivasyonundan birisine karşılık gelmektedir. Bu derivasyonun sütun sıralaması I, II, III, aVR, aVL, aVF, V1, V2, V3, V4, V5, ve V6 şeklindedir.

Tanımlayıcı değişkenler: Tanımlayıcı değişkenlerin çıkarılması işleminde öncelikle 500 Hz örnek frekansındaki EKG verileri 200Hz'e indirgenmiştir. Daha sonra, on iki derivasyonun her birisi için ortalama, medyan, kurtosis, çarpıklık, varyans, kareli ortalama ve ortalama değerin üzerindeki voltaj değerleri oranı değerlerini hesaplanmıştır.

Örnek Entropi Değişkenleri: Örnek Entropi değişkenleri her bir derivasyon için ayrı ayrı hesaplanmıştır. Burada örnek entropi değişkenleri, her bir derivasyondaki sinyallerin düzenlilik ölçüşünü ifade etmektedir. Bu amaçla EKG sinyalleri 1 saniyelik aralıklara bölünmüş ve her bir aralık için ayrı örnek entropi değişkenleri çıkarılmıştır.

Burada öncelikle EKG verisi 200Hz'e indirgendi ve normalleştirildi. Tüm EKG sinyalinin ÖrEnt'i m=2 boyutundaki tüm seriler için hesaplandı. Tolerans penceresi (r) olarak her bir derivasyonun standart sapmasının 0,3 kati kullanıldı. Bu şekilde toplam 12 değişken çıkardık. Daha sonra her bir EKG derivasyonu 1 saniyelik aralıklara bölünerek her bir aralığa ÖrEnt uyguladık. Bu aralıklardan elde edilen değerlerin ortalama, medyan, kurtosis, çarpıklık, standart sapma ve aralık genişliği olmak üzere toplan 72 değişken çıkardık.

Fourier Dönüşümü Değişkenleri: EKG sinyalleri FD yöntemi ile farklı frekanslarda basit sinüs dalgalarına dönüştürülmüştür. Bu donuşum işlemi yapılırken voltaj değerleri ve dalgaların fazları parametre olarak kullanılmıştır. Bu aşamada hem kesikli hem sürekli Fourier dönüşümler frekans spektrumundan değişken çıkarmaktadır. Kesikli Fourier değişkenleri beş seviyeli dalgacık dönüşümlerinden hesaplanan tanımlayıcı istatistikler olarak hesaplanmıştır. Sürekli Fourier dönüşümü değişkenleri hesaplamak için dalgacıkları EKG sinyali üzerinde hareket ettirerek elde iki boyutlu skalogramlarin elde

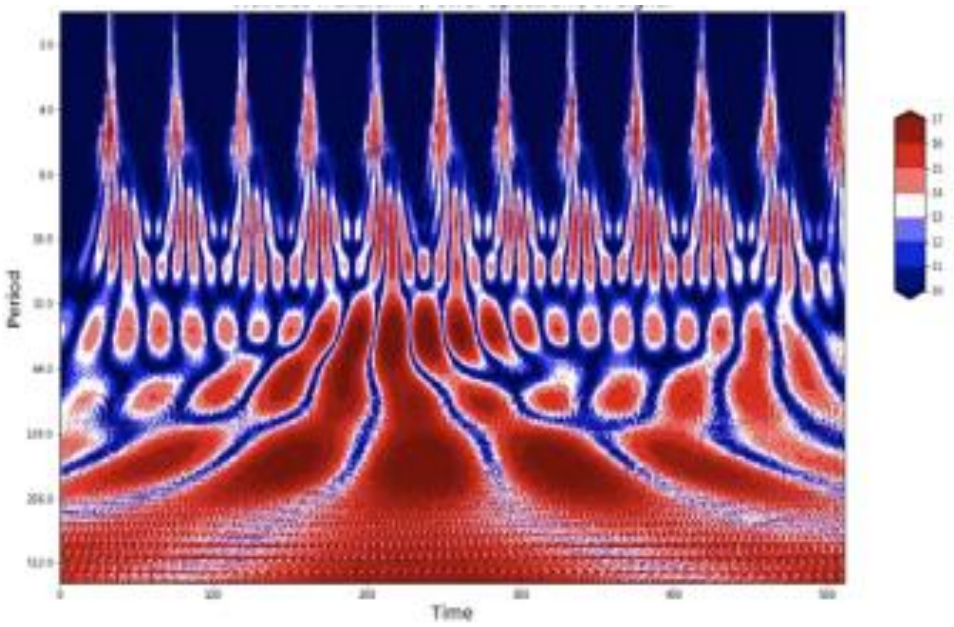
edilmiş ve bu skalogramlarin farklı zaman ve frekans değerlerindeki benzerliği kullanılmıştır. 200 Hz'e indirgenmiş EKG verisi üzerinden 0,1 ve 0,5 voltaj değerleri 1-20 Hz dalgacık faz aralıklarında toplam 528 değişken çıkardık.

Dalgacık Dönüşümü Değişkenleri: FD yöntemi ile çıkarılan değişkenler sinyale ait frekans temelli bilgileri yansıtmaktadır. Ancak, FD bu frekansların zaman ekseni ile ilgili bilgi vermemektedir. Bu noktada, DD yöntemi ile sinyale ait zaman alanı ile ilgili bilgilerin çıkarılması sağlanır. Bu bağlamda, 200 Hz'e indirgenmiş EKG verisine DD yöntemi aşağıdaki gibi SDD ve KDD olarak iki ayrı şekilde uygulanmıştır.

Sürekli dalgacık dönüşümü değişkenlerini çıkarmak amacı ile orijinal EKG verisi ile farklı ölçeklerdeki dalgacıklar arasındaki benzerli ölçülmüştür. Bu işlem yapılırken, dalgacıklar ECG verisinin zaman ekseninde Eşitlik 14 formülü kullanılarak kaydırılmış ve bu sayede farklı zaman aralıkları için olçum yapılmıştır.

$$X_w(a,b) = \frac{1}{|a|^{1/2}} \int_{-\infty}^{\infty} x(t)\psi\left(\frac{t-b}{a}\right) dt \qquad (14)$$

Sürekli dalgacık dönüşümünün EKG sinyaline uygulanması ile elde edilen çıktı 2 boyutlu skalogram olarak gösterilmiştir. Şekil 12 ile bu elde edilen skalogramlarin EKG Derivasyon-I için görüntüsü verilmiştir.



Şekil 12 Surekli Dalgacik Donusumunun EKG Derivasyon I'e uygulanmasi

Bu aşamada EKG verilerini 50Hz'e indirgedik. Her bir EKG derivasyonuna Morlet Dalgacıkları (Lin & Qu, 2000) uygulayarak her bir hasta için toplam 12 skalogram elde ettik. Bu skalogramlar bir araya getirilerek her bir hasta için 12 kanallı tek bir resim yaratık. Bu resimleri isleyecek 6 evreşimli katmanı olan bir ESA mimarisi oluşturduk. Bu ESA ağının eğitilmesi sonucu elde edilen optimum mimari üzerinden toplam 372 değişken çıkardık.

Dalgacık katsayılarının mümkün tüm ölçeklerde ve zaman noktalarında hesaplanması oldukça zordur. Bu bağlamda KDD dalgacık parametrelerinin kestirilmesinde kullanılan etkili but yöntemdir. KDD öncelikle ECG verisini Eşitlik 15 ile ifade edilen farklı frekans aralıklarına ayrıştırır. Bu sayede hem yaklaşık katsayılar hem de detay katsayıları elde edilir.

$$f[n] = \frac{1}{\sqrt{M}} \sum_k W_\varphi[j_0, k]\varphi_{j_0,k}[n] + \frac{1}{\sqrt{M}} \sum_{j=j_0}^{\infty} W_\varphi[j_0, k]\varphi_{j,k}[n] \qquad (15)$$

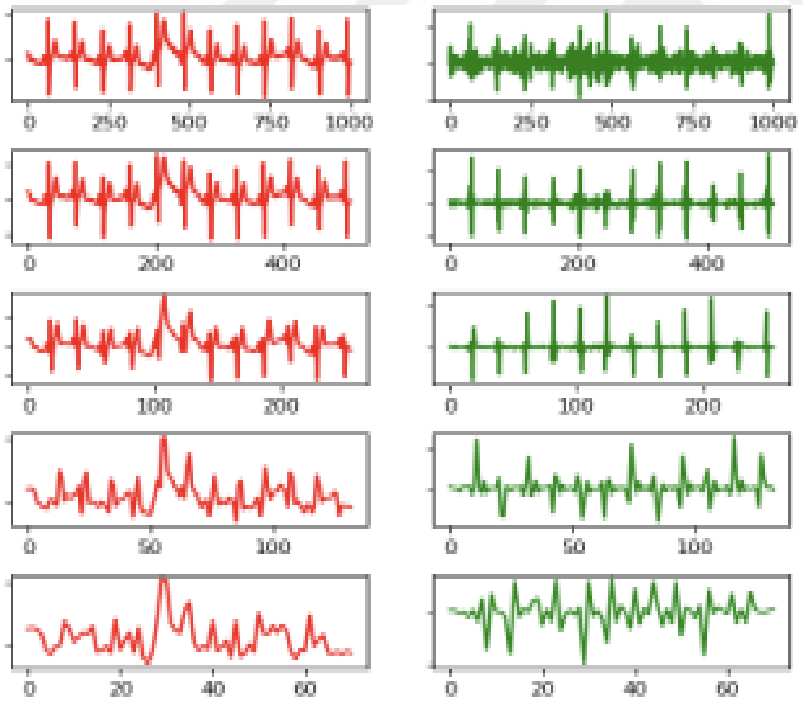
Burada kestirilen yaklaşık katsayılar bize düşük geçiş filtrelerini verir ve orijinal EKG verisini ölçekleme fonksiyonu sayesinde yumuşatır (Eşitlik 16).

$$W_\varphi[j_0,k] = \frac{1}{\sqrt{M}}\sum_n f[n]\varphi_{j_0,k}[n] \qquad (16)$$

Detay katsayısı ise yüksek geçiş filtresini verir ve dalgacık fonksiyonu kullanarak zaman ekseninde EKG de meydana gelen aşırı değişiklikleri ifade eder (Eşitlik 17).

$$W_\varphi[j_0,k] = \frac{1}{\sqrt{M}}\sum_n f[n]\varphi_{j,k}[n] \qquad (17)$$

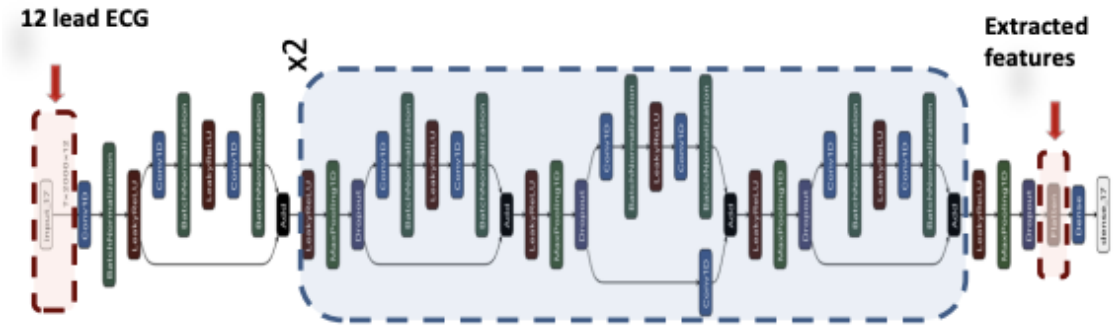
Bu kısımda EKG verisini 200 Hz'e indirgedik. Sym5 dalgacık fonksiyonu kullanarak 5 seviyeye kadar dalgacık ayrıştırması yaptık. Şekil 13 ile bir hastaya ait EKG için uygulanan dalgacık ayrıştırması örneklenmiştir.



Şekil 13 Bir hastanın Derivasyon I'e ait KDD dönüşümü örneği. Sol da Yaklaşık katsayılar ve sağda detay katsayıları

Detay katsayılarının otokorelasyonları 2, 14, 26, 38, 50, ve 62 gecikmeler için hesaplanarak toplamda 385 KDD değişkeni çıkardık.

Derin öğrenme değişkenleri: Bu kısımda derin öğrenme yaklaşımı olarak Evreşimli Sinir Ağları (ESA) kullanılmıştır. Derin öğrenme değişkenleri 12 derivasyonun eşanlı olarak ESA mimarisi ile işlenmesi ile elde edilmiştir. Burada kurduğumuz ESA mimarisinde EKG verisi 200 Hz'e indirgenerek kullanılmıştır. Kullanılan ESA mimarisi Şekil 14 ile verilmiştir.



Şekil 14 ESA ile değişken çıkarımı

Eğitimi tamamlanan modelin çıktıya bağlanmış olan son katmanındaki yer alan 448 nöron değerleri değişken olarak çıkarılmıştır.

OSMT ile değişken çıkarımı: Öncelikle orijinal EKG serileri 8 Hz'e indirgendi. Sayısal EKG voltaj serileri 5 harfli A={a,b,c,d,e} alfabesi kullanılarak sembolik seri haline getirildi. Bu işlem yapılırken her bir derivasyon için ayrı olarak yüze yirmilik dilimler kullanıldı. Bu bağlamda, voltaj değerleri açısından en düşük değerli yüzde yirmilik dilimde tüm voltaj değerleri 'a' ile, sonrakiler 'b' ile, be bu sekile en yüksek voltaj değerli yüzde yirmilik dilimdekiler 'e' ile değiştirildi. 6-sembolle kadar geçiş olasılıkları hesaplandı ve her bir kişinin geçiş olasılıkları matrisleri 117 kardiyomiyopati hastasının geçiş olasılıkları matrisi ile karşılaştırılarak toplamda 72 OSMT değişkeni hesaplandı.

4.3.5. Değişken Secimi

İlk aşamada çıkarılan değişken sayılarının çok fazla olması nedeni ile her bir yöntem ile çıkarılan değişkenlerden arasından en iyileri, her bir yöntem için ayrı işlemler yapılarak, Genetik Algoritma ile seçilmiştir. GA işlemi rastgele seçilmiş değişken alt kümeleri ile başlatılmış ve çapraz doğrulama ile elde edilen AUC değerlerinin performans ölçüsü

olarak kullanarak, yinelemeli olarak daha yüksek AUC veren değişken alt kümeleri belirlenmiştir. Her bir aşamada mevcut değişken alt kümesi kullanılarak XGBoost modeli kurulmuş ve bu modelden elde edilen çapraz doğrulama AUC değerleri kullanılmıştır. GA algoritması toplam 40 yineleme olarak kurulmuş, çaprazlama parametresi olarak 0,5 ve mutasyon parametresi olarak 0,05 kullanılmıştır.

Yukarıda ifade edilen GA algoritması her yöntemden elde edilen değişkenlere ayrı ayrı uygulanmış ve her yöntem için seçilen değişkenler ve karşılık gelen doğruluk oranları AUC, sensitivite ve spesifisite olarak Tablo 11de verilmiştir.

Tablo 11 GA ile değişken secimi

Yöntem	Çıkarılan Değişken sayısı	Seçilen değişken sayısı	AUC	Sensitivite	Spesifisite
Tanımlayıcı İstatistikler	168	8	0.74	65	67
Örnek Entropi	84	10	0.67	63	65
OSMT	72	10	0.80	68	72
Fourier Dönüşümü	528	14	0.72	63	66
Kesikli Dalgacık Dönüşümü	384	14	0.73	65	67
Sürekli Dalgacık Dönüşümü	372	15	0.73	66	69
Derin Öğrenme	448	15	0,76	71	71

Tablo 11 den görüleceği üzere OSMT yönteminden çıkarılan 84 değişken arasından seçilen 10 değişken ile model kurulduğunda, 0.80 AUC değeri ile en yüksek kardiyomiyopati tahmin başarası elde edilmiştir. OSMT yöntemini takiben ikinci en yüksek AUC, 0.76 ile derin öğrenme modelinden elde edilmiştir. Ele alınan değişken çıkarma yöntemleri arasında en düşük AUC ise, 0,62 ile Örnek Entropi yönteminden çıkarılan değişkenler ile elde edilmiştir.

4.3.6. Eksik Veriler

Çalışmada kullanılan EKG değişkenleri haricindeki toplam 20 değişken içerisinde eksik veri oranı %1'in altındadır. Tek tek değişkenlere bakıldığında, kalbe sızan radyasyon oranı %2,8, göğüs bölgesine uygulanan radyasyon oranı %0,9 ve göğüs bölgesine radyasyon uygulanıp uygulanmadığı bilgisi değişkeninde %0,9 oranında eksik veri vardır. Bu eksik veriler Multiple Imputation by Chained Equation yöntemini yürüten

MICE R-paketi kullanılarak doldurulmuştur (van Buuren & Groothuis-Oudshoorn, 2011).

4.3.7. Kardiyomiyopati Riski Tahmini

Tahminleyici modellerin gerçek hayata uyarlanabilirliği sadece bu modellerin doğruluk oranı ile değil, aynı zamanda tahmin yapmakta kullanılan değişkenlere ait verilerin elde edilebilme kolaylığı ile de belirlenir. Özellikle tip alanında, bazı veri türleri diğerlerine göre mevcut bir hastalık ile ilgili çok daha fazla bilgi içerse de elde, o bilgiyi üretecek cihazlar çok pahalı ve erişilemez olabilir. Örneğin, kalp kaslarındaki soruna ilişkin kardiyak MR, EKG verisine göre daha detaylı bilgi içeriyor olsa da, EKG daha kolay ulaşılabilir bir veri turudur. Buradan hareketle kardiyomiyopati tahmin modellerimizi üç aşamada kurduk. Her üç aşamada da Extreme Gradient Boosting (XGBoost) makine öğrenmesi yöntemi kullanılmıştır. Model kurma işlemi 5-fold çapraz doğrulama ile gerçekleştirilirmiştir ve Bayesci Hiperparametre optimizasyonu yöntemi kullanılmıştır. Optimize öğrenme oranı olarak 0,16 ve toplan zayıf tahminleyici sayısı olarak 1869 belirlenmiştir. Model performans AUC, sensitivite ve spesifisite olarak ölçülmüştür.

EKG Modeli: Bu aşamada sadece EKG verilerinden farklı yöntemlerle çıkarılan ve GA algoritması ile en iyi olarak seçilen toplam 86 değişken kullanılmıştır (bakiniz Tablo 11). Bu EKG değişkenleri ile kurulan XGBoost modeli 0,87 (0,83-0,90) AUC, 0,76 sensitivite ve 0.79 spesifisite sonucu vermiştir.

Klinik Değişken Modeli: Bu aşamada toplam 20 klinik değişkenler kullanılmıştır. Öncelikle bu değişkenlere benzer şekilde genetik algoritma uygulandığında, yedi tanesi seçilmiştir. Bu yedi klinik değişken ile kurulan XGBoost modeli 0,69 (0,64-0,74) AUC, 0,62 sensitivite ve 0.66 spesifisite sonucu vermiştir.

Karma Model: Son olarak, EKG değişkenleri ile klinik değişkenler birleştirilerek yeniden modelleme yapılmıştır. Bu aşamada kullanılan değişkenlerin listesi Tablo 12 ile verilmiştir.

Tablo 12 Karma modelde kullanılan değişkenler

Veri Turu (n)	EKH derivasyonu (n)	Değişken Çıkarma Metodu
	I (3)	ÖrEnt (1), OSMT (1), KDD (1)
	II (4)	ÖrEnt (1), OSMT (1), KDD (2)

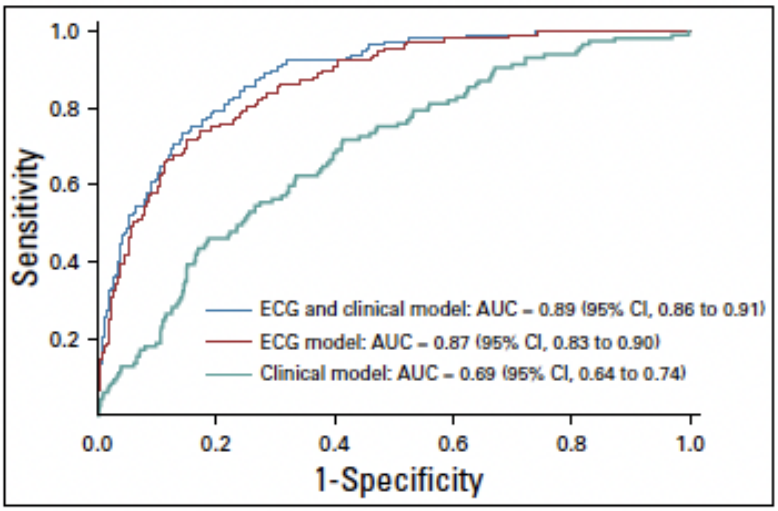
Tablo 12 Devam		
	III (4)	ÖrEnt (1), OSMT (1), FD (1), KDD (1)
	aVR (0)	-
	aVL (3)	OSMT (1), FD (1), KDD (1)
	aVF (4)	ÖrEnt (2), OSMT (1), FD (1)
	V1 (7)	Tanımlayıcı (2), ÖrEnt (1), OSMT (1), FD (3)
	V2 (6)	Tanımlayıcı (1), OSMT (1), FD (3), KDD (1)
	V3 (5)	Tanımlayıcı (2), ÖrEnt (2), KDD (1)
	V4 (1)	FT (1)
	V5 (10)	Tanımlayıcı (1), ÖrEnt (1), OSMT (1), FD (3), KDD (4)
	V6 (9)	Tanımlayıcı (2), ÖrEnt (1), OSMT (2), FD (1), KDD (3)
	12 derivasyon (30)	SDD (15), ESA (15)
Klinik veri (7)	Solunum Oranı	
	Vücut yüzey alanı	
	Antrasiklin dozu	
	Tütün ürünleri kullanımı	
	Göğüs bölgesine radyasyon alıp almama	
	Göğüs bölgesine alınan radyasyon dozu	
	Kalbe sızmış olan radyasyon dozu	

Tablo 12'de yer alan değişkenler ile kurulan XGBoost modeli 0,89 (0,86-0,91) AUC, 0,78 sensitivite ve 0.81 spesifisite sonucu vermiştir. Sınıflandırma performansı daha detaylı olarak Tablo 13ile verilmiştir.

Tablo 13 Kardiyomiyopati sınıflandırma matrisi

		Tahmin		
		Kontrol	Kardiyomiyopati	
Tahmin	Kontrol	892	208	Spesifisite= %81
	Kardiyomiyopati	26	91	Sensitivite= %78
		Negatif Tahminleme Oranı= %97	Pozitif Tahminleme Oranı= %30	

Her üç aşamada kurulan modellerin AUC'leri Şekil 15 ile karşılaştırılmıştır.



Şekil 15 Kardiyomiyopati modellerinin AUC karşılaştırması

4.3.7. Alt Grup Analizi

Elde ettiğimiz kardiyomiyopati tahmin modelinin ilk SJLIFE ziyaretini takiben kardiyomiyopatinin gerçekleşme suresi açısında da alt grup analizleri yaptık. İlk SJLIFE ziyaretinden 5 yıla kadar gerçeklesen kardiyomiyopatileri tahmin basarisi olarak 0,89 AUC, 0,79 sensitivite ve 0,81 spesifisite elde edilmiştir. Diğer taraftan, ilk SJLIFE ziyaretinden 5 ila 10 yıl sonra gerçeklesen kardiyomiyopati hastalığının tahmininde 0,88 AUC, 0,78 sensitivite be 0,81 spesifisite elde edilmiştir. Dolayısı ile elde ettiğimiz model, kardiyomiyopatinin on yıllık takip suresinde ne zaman gerçekleştiğine bakmaksızın basari ile tahmin yapmaktadır.

4.3.8. Değişken Önem Analizi

Sadece EKG değişkenleri kullanan modele nazaran az da olsa karma model 0,89 AUC ile en yüksek basariyi göstermiştir. Bu model üzerinden değişken önem analizi (Şekil 16) yaparak kardiyomiyopati tahminine en çok katkıda bulunan değişkenleri belirledik.

Şekil 16 Değişken Önem Analizi

Sekil 15'den görüleceği üzere, kardiyomiyopati tahmininden kullanılan en önemli on değişkenin beş tanesi OSMT yöntemi ile elde edilmiştir. Daha önemlisi, bu beş değişkenin ikisi, en önemli iki değişken olarak ortaya çıkmaktadır. OSMT değişkenlerinin dışında EKG verilerin oto korelasyon yapısı, alınan toplam radyasyon ve antrasiklin değerleri de önemli değişkenler olarak belirlenmiştir.

4.3.9. Uygulama Sonucu

Yüksek kardiyomiyopati riski altında olan çocukluk kanserini yenmiş bireyler yapay zeka yardımı ile yüksek doğruluk oranında belirlenebilirler. Sadece EKG kullanılarak kurulan model, klinik değişkenler eklendiğinde kurulan modele çok yakın sonuçlar vermektedir. EKG fonksiyonu olan giyilebilir teknolojilerdeki ilerlemeleri göz önüne aldığımızda, bu tip yöntemler kanseri yenen bireylerin hayat boyu düşük maliyetle be uzaktan takibini sağlama potansiyeline sahiptir. Diğer taraftan, analizlerimizin gösterdiği üzere, OSMT ile çıkarılan EKG değişkenleri diğer tüm yöntemler ile çıkarılan değişkenlerden daha önemli kardiyomiyopati riski bilgisi taşımaktadır.

5. SONUÇLAR VE TARTIŞMA

Bu çalışmada yeni bir sinyal isleme yöntemi olarak Olasılıksal Sembolik Motif Tanıma yöntemi tanıtıldı ver uygulamalarına yer verildi. OSMT yöntemi elemanları sonlu sayıda semboller olan bir kümeden oluşan serilerin, sembolleri arası geçiş olasılıklarının modellenmesine dayanır. Birden fazla seri, bu geçiş olasılıkları aralarındaki benzerliklerin bir ölçüsü kullanılarak karşılaştırılabilir. OSMT temel olarak, verilen bir serinin hangi semboller ile devam edeceğinin tahmin edilmesinde kullanılabilir. Bu bağlamda birçok farklı disiplinde uygulanması mümkündür. Örneğin eksik DNA bilgilerinin tamamlanmasında, eksik veri doldurulmasında, yapay olarak belirli bir tarzda ya da belirli bir müzisyenin eserlerine benzeyen yapay müzik üretme, finansal tahmin ve oyun teorisi gibi strateji belirleme gibi alanlarda kullanılabilir.

OSMT yönteminin bu çalışmada da üzerinde yoğunlaşılan uygulaması değişken çıkarımı olmuştur. OSMT yöntemi ile serilerin sembolleri arasındaki geçiş olasılıklarını temsil eden değişkenler çıkarılarak bu değişkenler makine öğrenmesi teknikleri ile analiz edilebilir. Bu çalışma da OSMT üç farklı probleme uygulandı.

İlk OSMT uygulaması DNA sınıflandırması üzerinedir. Kısmi DNA serileri verilen 3 farklı türe ait toplam 13 salyangoz, OSMT yöntemi ile doğru sınıflara mükemmel bir şekilde sınıflandırıldı. Burada dikkat çeken husus, OSMT'nin bu mükemmel sınıflandırma basarisini, DNA hizalama on işlemine gerek olmadan yapabilmesidir. Bu hem zaman kazanma hem de hizalama işlemi sırasından kaynaklı hataların önüne geçilmesi açısından önemlidir.

İkinci OSMT uygulaması, uzun sureli EKG kayıtlarından elde edilen R-R serileri yardımı ile konjestif kalp yetersizliğinin belirlenmesidir. OSMT yönteminden elde edilen sekiz değişken ve on tanımlayıcı değişken karar ağaçları ile analiz edildiğinde mükemmele yakın bir doğruluk ile konjestif kalp yetersizliği hastaları sağlıklı normal hastalardan ayrılabilmiştir. Kalp yetersizliğinin toplum sağlığı üzerindeki etkisi düşünüldüğünde, erken teşhis ve müdahalenin büyük önemi ortaya çıkmaktadır. Bu bağlamda burada önerilen ve benzeri modellerin hasta kayıt sistemlerine entegre edilmesi sayesinde konjestif kalp yetersizliği hastaları otomatik olarak belirlenebilir.

Ucuncu OSMT uygulamasi ise cocuklik kanserini yenmiş yetişkin bireylerin uzun donem kardiyomiyopati riskinin EKG verisinden tahmini üzerindedir. Bu calisma da ayrica

OSMT yöntemi diğer sinyal isleme yöntemlerime derin öğrenme yöntemleri ile de karşılaştırılmıştır. Elde ettiğimiz sonuçlar göstermiştir ki, OSMT yöntemi ile elde edilen değişkenler karşılaştırmada kullanılan Fourier Dönüşümü, Dalgacık Dönüşümü, Örnek Entropisi, Evreşimli Sinir Ağları gibi tüm yöntemlerden daha açıklayıcı değişkenler çıkarmıştır. Değişken önem analizi göstermiştir ki, en önemli on değişkenin beşi, OSMT değişkenleridir. Diğer taraftan, elde edilen sonuçlar klinik önem taşımaktadır. Çocukluk kanserini yenen bireyler yetişkin yaşantılarında yüksek kalp ve damar hastalıkları riski altındadır. Bu kişilerin sıklıkla takibi gerekirken genelde bu gerçek hayatta mümkün olmamaktadır. Bu çalışma göstermiştir ki, çocukluk kanseri yenen bireylerin kardiyomiyopati risk sadece EKG verileri kullanılarak yüksek doğruluk oranı ile belirlenebilir. Giyilebilir teknolojilerdeki hızlı ilerlemeler göz önüne alındığında, kanseri yenen bireyleri hayat boyu akıllı saat gibi EKG özelliği olan cihazlar yardımı ile uzaktan düşük maliyet ile takibi mümkün olabilir.

6. KAYNAKLAR

Acharya, U. R., Fujita, H., Sudarshan, V. K., Lih Oh, S., Muhammad, A., Koh, J. E. W., Hong Tan, J., Chua, C. K., Poo Chua, K., & San Tan, R. (2017). Application of empirical mode decomposition (EMD) for automated identification of congestive heart failure using heart rate signals. *Neural Computing and Applications*, *28*(10), 3073-3094. https://doi.org/10.1007/s00521-016-2612-1

Afify, H. M. A., Waits, G. S., Ghoneum, A. D., Cao, X., Li, Y., & Soliman, E. Z. (2018). Peguero Electrocardiographic Left Ventricular Hypertrophy Criteria and Risk of Mortality. *Front Cardiovasc Med*, *5*, 75. https://doi.org/10.3389/fcvm.2018.00075

Agarwal, S. K., & Soliman, E. Z. (2013). ECG abnormalities and stroke incidence. *Expert Rev Cardiovasc Ther*, *11*(7), 853-861. https://doi.org/10.1586/14779072.2013.811980

Akbilgiç, O. (2011). Hibrit radyal tabanlı fonksiyon ağları ile değişken seçimi ve tahminleme: menkul kıymet yatırım kararlarına ilişkin bir uygulama. *Unpublished doctoral dissertation). İstanbul University, İstanbul*.

Akbilgic, O., Bozdogan, H., & Balaban, M. E. (2014). A novel Hybrid RBF Neural Networks model as a forecaster. *Statistics and Computing*, *24*(3), 365-375. https://doi.org/10.1007/s11222-013-9375-7

Akbilgic, O., Butler, L., & Soliman, E. Z. (2023). Electrocardiogram. In A. Chang (Ed.), *Intelligence-Based Cardiology Artificial Intelligence and Human Cognition in Clinical Cardiology and Cardiac Surgery* (1st ed.). Elsevier Academic press.

Akbilgic, O., & Howe, J. A. (2017). Symbolic pattern recognition for sequential data. *Sequential Analysis*, *36*(4), 528-540.

Akbilgic, O., Kamaleswaran, R., Mohammad, A., Ross, W., Masaki, K., Petrovithc, H., Tanner, C., Davis, R., & Goldman, S. (2020). Electrocardiographic Changes Predate Parkinson Disease Onset. *Scientific Reports*, *(Forthcoming)*.

Akbilgic, O., Kamaleswaran, R., Mohammed, A., Ross, G. W., Masaki, K., Petrovitch, H., Tanner, C. M., Davis, R. L., & Goldman, S. M. (2020). Electrocardiographic changes predate Parkinson's disease onset. *Scientific Reports*, *10*(1), 11319. https://doi.org/10.1038/s41598-020-68241-6

Akselrod, S., Gordon, D., Madwed, J. B., Snidman, N. C., Shannon, D. C., & Cohen, R. J. (1985). Hemodynamic regulation: investigation by spectral analysis. *Am J Physiol*, *249*(4 Pt 2), H867-875. https://doi.org/10.1152/ajpheart.1985.249.4.H867

Akselrod, S., Gordon, D., Ubel, F. A., Shannon, D. C., Berger, A. C., & Cohen, R. J. (1981). Power spectrum analysis of heart rate fluctuation: a quantitative probe of beat-to-beat cardiovascular control. *Science*, *213*(4504), 220-222. https://doi.org/10.1126/science.6166045

Alchin, J. E., Signorelli, C., McLoone, J. K., Wakefield, C. E., Fardell, J. E., Johnston, K., & Cohn, R. J. (2022). Childhood Cancer Survivors' Adherence to Healthcare Recommendations Made Through a Distance-Delivered Survivorship Program. *J Multidiscip Healthc*, *15*, 1719-1734. https://doi.org/10.2147/jmdh.S363653

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, *215*(3), 403-410. https://doi.org/10.1016/s0022-2836(05)80360-2

Armenian, S., & Bhatia, S. (2018). Predicting and Preventing Anthracycline-Related Cardiotoxicity. *Am Soc Clin Oncol Educ Book*, *38*, 3-12. https://doi.org/10.1200/edbk_100015

Armenian, S. H., Hudson, M. M., Mulder, R. L., Chen, M. H., Constine, L. S., Dwyer, M., Nathan, P. C., Tissing, W. J., Shankar, S., Sieswerda, E., Skinner, R., Steinberger, J., van Dalen, E. C., van der Pal, H., Wallace, W. H., Levitt, G., & Kremer, L. C. (2015). Recommendations for cardiomyopathy surveillance for survivors of childhood cancer: a report from the International Late Effects of Childhood Cancer Guideline Harmonization Group. *Lancet Oncol*, *16*(3), e123-136. https://doi.org/10.1016/s1470-2045(14)70409-7

Asyali, M. (2003). Discrimination power of long-term heart rate variability measures. Proceedings of the 25th annual international conference of the IEEE engineering in medicine and biology society (IEEE Cat. No. 03CH37439),

Begg, G., Willan, K., Tyndall, K., Pepper, C., & Tayebjee, M. (2016). Electrocardiogram interpretation and arrhythmia management: a primary and secondary care survey. *Br J Gen Pract*, *66*(646), e291-296. https://doi.org/10.3399/bjgp16X684781

Berntson, G. G., Bigger, J. T., Jr., Eckberg, D. L., Grossman, P., Kaufmann, P. G., Malik, M., Nagaraja, H. N., Porges, S. W., Saul, J. P., Stone, P. H., & van der Molen, M. W. (1997). Heart rate variability: origins, methods, and interpretive caveats. *Psychophysiology*, *34*(6), 623-648. https://doi.org/10.1111/j.1469-8986.1997.tb02140.x

Bilt, V. (2022). *Basics ECGpedia*. Retrieved 11.19.2022 from https://en.ecgpedia.org/index.php?title=Main_Page

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Chang, C. C., Spitzer, E., Chichareon, P., Takahashi, K., Modolo, R., Kogame, N., Tomaniak, M., Komiyama, H., Yap, S. C., Hoole, S. P., Gori, T., Zaman, A., Frey, B., Ferreira, R. C., Bertrand, O. F., Koh, T. H., Sousa, A., Moschovitis, A., van Geuns, R. J., . . . Onuma, Y. (2019). Ascertainment of Silent Myocardial Infarction in Patients Undergoing Percutaneous Coronary Intervention (from the GLOBAL LEADERS Trial). *Am J Cardiol*, *124*(12), 1833-1840. https://doi.org/10.1016/j.amjcard.2019.08.049

Cook, D. A., Oh, S. Y., & Pusic, M. V. (2020). Accuracy of Physicians' Electrocardiogram Interpretations: A Systematic Review and Meta-analysis. *JAMA Intern Med*, *180*(11), 1-11. https://doi.org/10.1001/jamainternmed.2020.3989

Costa, M., & Healey, J. A. (2003, 21-24 Sept. 2003). Multiscale entropy analysis of complex heart rate dynamics: discrimination of age and heart failure effects. Computers in Cardiology, 2003,

Coumel, P., Maison-Blanche, P., & Catuli, D. (1994). Heart rate and heart rate variability in normal young adults. *J Cardiovasc Electrophysiol*, *5*(11), 899-911. https://doi.org/10.1111/j.1540-8167.1994.tb01130.x

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Eiben, A. E., & Smith, J. E. (2015). *Introduction to Evolutionary Computing*. Springer Berlin. https://doi.org/https://doi.org/10.1007/978-3-662-44874-8

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C. K., & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, *101*(23), E215-220. https://doi.org/10.1161/01.cir.101.23.e215

Gunturkun, F., Akbilgic, O., Davis, R. L., Armstrong, G. T., Howell, R. M., Jefferies, J. L., Ness, K. K., Karabayir, I., Lucas, J. T., Jr., Srivastava, D. K., Hudson, M. M., Robison, L. L., Soliman, E. Z., & Mulrooney, D. A. (2021). Artificial Intelligence-Assisted Prediction of Late-Onset Cardiomyopathy Among Childhood Cancer Survivors. *JCO Clin Cancer Inform*, *5*, 459-468. https://doi.org/10.1200/CCI.20.00176

Gunturkun, F., Davis, R. L., Armstrong, G. T., Jefferies, J. L., Ness, K. K., Green, D. M., Lucas, J. T., Srivastava, D., Hudson, M. M., Robison, L. L., Mulrooney, D. A., Soliman, E. Z., Karabayir, I., & Akbilgic, O. (2020). Deep learning for improved prediction of late-onset cardiomyopathy among childhood cancer survivors: A report from the St. Jude Lifetime Cohort (SJLIFE). *Journal of Clinical Oncology*, *38*(15_suppl), 10545-10545. https://doi.org/10.1200/JCO.2020.38.15_suppl.10545

Guzzetti, S., Mezzetti, S., Magatelli, R., Porta, A., De Angelis, G., Rovelli, G., & Malliani, A. (2000). Linear and non-linear 24 h heart rate variability in chronic heart failure. *Auton Neurosci*, *86*(1-2), 114-119. https://doi.org/10.1016/s1566-0702(00)00239-3

Hamada, M., Martz, H. F., Reese, C. S., & Wilson, A. G. (2001). Finding Near-Optimal Bayesian Experimental Designs via Genetic Algorithms. *The American Statistician*, *55*(3), 175-181. http://www.jstor.org/stable/2685795

Harris, F. J. (1978). On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE*, *66*(1), 51-83.

Heckbert, S. R., Austin, T. R., Jensen, P. N., Floyd, J. S., Psaty, B. M., Soliman, E. Z., & Kronmal, R. A. (2018). Yield and consistency of arrhythmia detection with patch

electrocardiographic monitoring: The Multi-Ethnic Study of Atherosclerosis. *J Electrocardiol*, *51*(6), 997-1002. https://doi.org/10.1016/j.jelectrocard.2018.07.027

Ho, K. K., Moody, G. B., Peng, C. K., Mietus, J. E., Larson, M. G., Levy, D., & Goldberger, A. L. (1997). Predicting survival in heart failure case and control subjects by use of fully automated methods for deriving nonlinear and conventional indices of heart rate dynamics. *Circulation*, *96*(3), 842-848. https://doi.org/10.1161/01.cir.96.3.842

Hossen, A., & Al-Ghunaimi, B. (2007). A wavelet-based soft decision technique for screening of patients with congestive heart failure. *Biomedical Signal Processing and Control*, *2*(2), 135-143. https://doi.org/https://doi.org/10.1016/j.bspc.2007.05.008

Howell, R. M., Smith, S. A., Weathers, R. E., Kry, S. F., & Stovall, M. (2019). Adaptations to a Generalized Radiation Dose Reconstruction Methodology for Use in Epidemiologic Studies: An Update from the MD Anderson Late Effect Group. *Radiat Res*, *192*(2), 169-188. https://doi.org/10.1667/rr15201.1

Howlader, N., Noone, A. M., Krapcho, M., Miller, D., Brest, A., Yu, M., Ruhl, J., Tatalovich, Z., Mariotto, A., Lewis, D. R., Chen, H. S., Feuer, E. J., & Cronin, K. A. (2020). *SEER Cancer Statistics Review, 1975-2017*. National Cancer Institute. Retrieved 8/4/2020 from https://seer.cancer.gov/csr/1975_2017/

Hu, J., Gao, J., Tung, W. W., & Cao, Y. (2010). Multiscale analysis of heart rate variability: a comparison of different complexity measures. *Ann Biomed Eng*, *38*(3), 854-864. https://doi.org/10.1007/s10439-009-9863-2

InformedHealth.org. (2006). *What is an electrocardiogram (ECG)?* InformedHealth.org. Institute for Quality and Efficiency in Health Care (IQWiG); . Retrieved 11.13.2022 from https://www.ncbi.nlm.nih.gov/books/NBK536878/

Institute, H. H. M. (2009). *Comparing DNA Sequences to Determine Evolutionary Relationships amongMollusks*. http://media.hhmi.org/biointeractive/activities/shells/shell-dna.pdf

Işler, Y., & Kuntalp, M. (2007). Combining classical HRV indices with wavelet entropy measures improves to performance in diagnosing congestive heart failure. *Comput Biol Med*, *37*(10), 1502-1510. https://doi.org/10.1016/j.compbiomed.2007.01.012

Kamaleswaran, R., Akbilgic, O., Hallman, M. A., West, A. N., Davis, R. L., & Shah, S. H. (2018). Applying Artificial Intelligence to Identify Physiomarkers Predicting Severe Sepsis in the PICU. *Pediatr Crit Care Med*, *19*(10), e495-e503. https://doi.org/10.1097/PCC.0000000000001666

Kamaleswaran, R., Mahajan, R., & Akbilgic, O. (2018). A robust deep convolutional neural network for the classification of abnormal cardiac rhythm using single lead electrocardiograms of variable length. *Physiol Meas*, *39*(3), 035006. https://doi.org/10.1088/1361-6579/aaaa9d

Kamen, P. W., & Tonkin, A. M. (1995). Application of the Poincaré plot to heart rate variability: a new measure of functional status in heart failure. *Aust N Z J Med*, *25*(1), 18-26. https://doi.org/10.1111/j.1445-5994.1995.tb00573.x

Landier, W., Wallace, W. H., & Hudson, M. M. (2006). Long-term follow-up of pediatric cancer survivors: education, surveillance, and screening. *Pediatr Blood Cancer*, *46*(2), 149-158. https://doi.org/10.1002/pbc.20612

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE,

Li, Q., & Clifford, G. D. (2012). Dynamic time warping and machine learning for signal quality assessment of pulsatile signals. *Physiological measurement*, *33*(9), 1491.

Lin, J., & Qu, L. (2000). Feature extraction based on Morlet wavelet and its application for mechanical fault diagnosis. *Journal of sound and vibration*, *234*(1), 135-148.

Linge, H. M., & Follin, C. (2021). Mixed methods assessment of impact on health awareness in adult childhood cancer survivors after viewing their personalized digital treatment summary and follow-up recommendations. *BMC Cancer*, *21*(1), 347. https://doi.org/10.1186/s12885-021-08051-9

Lkhagva, B., Suzuki, Y., & Kawagoe, K. (2006a). Extended SAX: Extension of symbolic aggregate approximation for financial time series data representation. *DEWS2006 4A-i8*, *7*.

Lkhagva, B., Suzuki, Y., & Kawagoe, K. (2006b). New time series data representation ESAX for financial applications. 22nd International Conference on Data Engineering Workshops (ICDEW'06),

Lüderitz, B., & de Luna, A. B. (2017). The history of electrocardiography. *J Electrocardiol*, *50*(5), 539. https://doi.org/10.1016/j.jelectrocard.2017.07.014

Mahajan, R., Kamaleswaran, R., & Akbilgic, O. (2017). Effects of varying sampling frequency on the analysis of continuous ECG data streams. VLDB Workshop on Data Management and Analytics for Medicine and Healthcare,

Mahajan, R., Kamaleswaran, R., Howe, J. A., & Akbilgic, O. (2017). Cardiac rhythm classification from a short single lead ECG recording via random forest. 2017 Computing in Cardiology (CinC),

Maheshwari, A., Norby, F. L., Roetker, N. S., Soliman, E. Z., Koene, R. J., Rooney, M. R., O'Neal, W. T., Shah, A. M., Claggett, B. L., Solomon, S. D., Alonso, A., Gottesman, R. F., Heckbert, S. R., & Chen, L. Y. (2019). Refining Prediction of Atrial Fibrillation-Related Stroke Using the P(2)-CHA(2)DS(2)-VASc Score. *Circulation*, *139*(2), 180-191. https://doi.org/10.1161/circulationaha.118.035411

Malliani, A., Pagani, M., Lombardi, F., & Cerutti, S. (1991). Cardiovascular neural regulation explored in the frequency domain. *Circulation*, *84*(2), 482-492. https://doi.org/10.1161/01.cir.84.2.482

Maron, B. J., Friedman, R. A., Kligfield, P., Levine, B. D., Viskin, S., Chaitman, B. R., Okin, P. M., Saul, J. P., Salberg, L., Van Hare, G. F., Soliman, E. Z., Chen, J., Matherne, G. P., Bolling, S. F., Mitten, M. J., Caplan, A., Balady, G. J., & Thompson, P. D. (2014). Assessment of the 12-lead ECG as a screening test for detection of cardiovascular disease in healthy general populations of young people (12-25 Years of Age): a scientific statement from the American Heart Association and the American College of Cardiology. *Circulation*, *130*(15), 1303-1334. https://doi.org/10.1161/cir.0000000000000025

Masoudi, F. A., Magid, D. J., Vinson, D. R., Tricomi, A. J., Lyons, E. E., Crounse, L., Ho, P. M., Peterson, P. N., & Rumsfeld, J. S. (2006). Implications of the failure to identify high-risk electrocardiogram findings for the quality of care of patients with acute myocardial infarction: results of the Emergency Department Quality in Myocardial Infarction (EDQMI) study. *Circulation*, *114*(15), 1565-1571. https://doi.org/10.1161/circulationaha.106.623652

MATLAB. (2020). *Signal Processing Toolbox User' Guide*. Retrieved June 12 from https://www.mathworks.com/help/pdf_doc/signal/signal.pdf

Melillo, P., De Luca, N., Bracale, M., & Pecchia, L. (2013). Classification tree for risk assessment in patients suffering from congestive heart failure via long-term heart rate variability. *IEEE J Biomed Health Inform*, *17*(3), 727-733. https://doi.org/10.1109/jbhi.2013.2244902

Melillo, P., Fusco, R., Sansone, M., Bracale, M., & Pecchia, L. (2011). Discrimination power of long-term heart rate variability measures for chronic heart failure detection. *Med Biol Eng Comput*, *49*(1), 67-74. https://doi.org/10.1007/s11517-010-0728-5

Mietus, J. E., Peng, C. K., Henry, I., Goldsmith, R. L., & Goldberger, A. L. (2002). The pNNx files: re-examining a widely used heart rate variability measure. *Heart*, *88*(4), 378-380. https://doi.org/10.1136/heart.88.4.378

Montano, N., Ruscone, T. G., Porta, A., Lombardi, F., Pagani, M., & Malliani, A. (1994). Power spectrum analysis of heart rate variability to assess the changes in sympathovagal balance during graded orthostatic tilt. *Circulation*, *90*(4), 1826-1831. https://doi.org/10.1161/01.cir.90.4.1826

Mozaffarian, D., Benjamin, E. J., Go, A. S., Arnett, D. K., Blaha, M. J., Cushman, M., Das, S. R., de Ferranti, S., Després, J. P., Fullerton, H. J., Howard, V. J., Huffman, M. D., Isasi, C. R., Jiménez, M. C., Judd, S. E., Kissela, B. M., Lichtman, J. H., Lisabeth, L. D., Liu, S., . . . Turner, M. B. (2016). Executive Summary: Heart Disease and Stroke Statistics--2016 Update: A Report From the American Heart Association. *Circulation*, *133*(4), 447-454. https://doi.org/10.1161/cir.0000000000000366

Mudd, J. O., & Kass, D. A. (2008). Tackling heart failure in the twenty-first century. *Nature*, *451*(7181), 919-928. https://doi.org/10.1038/nature06798

Mulrooney, D. A., Armstrong, G. T., Huang, S., Ness, K. K., Ehrhardt, M. J., Joshi, V. M., Plana, J. C., Soliman, E. Z., Green, D. M., Srivastava, D., Santucci, A., Krasin, M.

J., Robison, L. L., & Hudson, M. M. (2016). Cardiac Outcomes in Adult Survivors of Childhood Cancer Exposed to Cardiotoxic Therapy: A Cross-sectional Study. *Ann Intern Med*, *164*(2), 93-101. https://doi.org/10.7326/m15-0424

Peltola, M. A. (2012). Role of editing of R-R intervals in the analysis of heart rate variability. *Front Physiol*, *3*, 148. https://doi.org/10.3389/fphys.2012.00148

Perez, M. V., Mahaffey, K. W., Hedlin, H., Rumsfeld, J. S., Garcia, A., Ferris, T., Balasubramanian, V., Russo, A. M., Rajmane, A., Cheung, L., Hung, G., Lee, J., Kowey, P., Talati, N., Nag, D., Gummidipundi, S. E., Beatty, A., Hills, M. T., Desai, S., . . . Turakhia, M. P. (2019). Large-Scale Assessment of a Smartwatch to Identify Atrial Fibrillation. *N Engl J Med*, *381*(20), 1909-1917. https://doi.org/10.1056/NEJMoa1901183

PhysionNet. (2022a). *The MIT-BIH Normal Sinus Rhythm database*. https://physionet.org/

physiobank/database/nsrdb/

PhysionNet. (2022b). *Normal Sinus Rhythm RR Interval database*. https://physionet.org/physiobank/database/nsr2db/

Pincus, S. (1995). Approximate entropy (ApEn) as a complexity measure. *Chaos*, *5*(1), 110-117. https://doi.org/10.1063/1.166092

Pincus, S., & Singer, B. H. (1996). Randomness and degrees of irregularity. *Proc Natl Acad Sci U S A*, *93*(5), 2083-2088. https://doi.org/10.1073/pnas.93.5.2083

Pincus, S. M. (1991). Approximate entropy as a measure of system complexity. *Proc Natl Acad Sci U S A*, *88*(6), 2297-2301. https://doi.org/10.1073/pnas.88.6.2297

Poon, C. S., & Merrill, C. K. (1997). Decrease of cardiac chaos in congestive heart failure. *Nature*, *389*(6650), 492-495. https://doi.org/10.1038/39043

Prineas, J. R., Crow, R. S., & Zhang, Z. (2010). *The Minnesota Code Manual of Electrocardiographic Findings*. Springer. https://doi.org/https://doi.org/10.1007/978-1-84882-778-3

Rajendra Acharya, U., Paul Joseph, K., Kannathal, N., Lim, C. M., & Suri, J. S. (2006). Heart rate variability: a review. *Med Biol Eng Comput*, *44*(12), 1031-1051. https://doi.org/10.1007/s11517-006-0119-0

Ream, N. (1977). Discrete-Time Signal Processing. *Electronics and Power*, *23*(2), 157.

Ren, H., Liao, X., Li, Z., & Abdulrahman, A.-A. (2018). Anomaly detection using piecewise aggregate approximation in the amplitude domain. *Applied Intelligence*, *48*(5), 1097-1110.

Richman, J. S., & Moorman, J. R. (2000). Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology*, *278*(6), H2039-H2049. https://doi.org/10.1152/ajpheart.2000.278.6.H2039

Roger, V. L. (2013). Epidemiology of heart failure. *Circ Res*, *113*(6), 646-659. https://doi.org/10.1161/circresaha.113.300268

Sauer, W. H., Olshansky, B., & Yeon, S. B. (2022). *Normal sinus rhythm and sinus arrhythmia*. UpToDate Wolters Kluwer. Retrieved 11.18.2022 from https://www.uptodate.com/contents/normal-sinus-rhythm-and-sinus-arrhythmia/print#:~:text=The%20range%20(defined%20by%201st,)%20%5B1%2D3%5D.

Saykrs, B. M. (1973). Analysis of Heart Rate Variability. *Ergonomics*, *16*(1), 17-32. https://doi.org/10.1080/00140137308924479

Schiff, S. J., Aldroubi, A., Unser, M., & Sato, S. (1994). Fast wavelet transformation of EEG. *Electroencephalography and clinical neurophysiology*, *91*(6), 442-455.

Shima, H., & Nakayama, T. (2009). Wavelet Transformation. In *Higher Mathematics for Physics and Engineering* (pp. 449-480). Springer.

Silverman, M. E. (1992). Willem Einthoven--the father of electrocardiography. *Clin Cardiol*, *15*(10), 785-787. https://doi.org/10.1002/clc.4960151020

Soliman, E. Z., Zhang, Z. M., Chen, L. Y., Tereshchenko, L. G., Arking, D., & Alonso, A. (2017). Usefulness of Maintaining a Normal Electrocardiogram Over Time for Predicting Cardiovascular Health. *Am J Cardiol*, *119*(2), 249-255. https://doi.org/10.1016/j.amjcard.2016.09.051

Sutton, J. R., Mahajan, R., Akbilgic, O., & Kamaleswaran, R. (2019). PhysOnline: An Open Source Machine Learning Pipeline for Real-Time Analysis of Streaming Physiological Waveform. *IEEE J Biomed Health Inform*, *23*(1), 59-65. https://doi.org/10.1109/JBHI.2018.2832610

Thuraisingham, R. A. (2009). A Classification System to Detect Congestive Heart Failure Using Second-Order Difference Plot of RR Intervals. *Cardiol Res Pract*, *2009*, 807379. https://doi.org/10.4061/2009/807379

Todd, K. H., Hoffman, J. R., & Morgan, M. T. (1996). Effect of cardiologist ECG review on emergency department practice. *Ann Emerg Med*, *27*(1), 16-21. https://doi.org/10.1016/s0196-0644(96)70290-1

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *2011*, *45*(3), 67. https://doi.org/10.18637/jss.v045.i03

van Wyk, F., Khojandi, A., Kamaleswaran, R., Akbilgic, O., Nemati, S., & Davis, R. L. (2017). How much data should we collect? A case study in sepsis detection using deep learning. 2017 IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT),

von Tscharner, V., & Zandiyeh, P. (2017). Multi-scale transitions of fuzzy sample entropy of RR-intervals and their phase-randomized surrogates: A possibility to diagnose congestive heart failure. *Biomedical Signal Processing and Control*, *31*, 350-356. https://doi.org/https://doi.org/10.1016/j.bspc.2016.08.014

Westdrop, E. J., Gratton, M. C., & Watson, W. A. (1992). Emergency department interpretation of electrocardiograms. *Ann Emerg Med*, *21*(5), 541-544. https://doi.org/10.1016/s0196-0644(05)82521-1

Yee, K. M., Pringle, S. D., & Struthers, A. D. (2001). Circadian variation in the effects of aldosterone blockade on heart rate variability and QT dispersion in congestive heart failure. *J Am Coll Cardiol*, *37*(7), 1800-1807. https://doi.org/10.1016/s0735-1097(01)01243-8

Yu, S. N., & Lee, M. Y. (2012). Bispectral analysis and genetic algorithm for congestive heart failure recognition based on heart rate variability. *Comput Biol Med*, *42*(8), 816-825. https://doi.org/10.1016/j.compbiomed.2012.06.005

Zhang, Z., & Moore, J. (2014). *Mathematics and Physical Fundamentals of Climate Change* (1st ed.). Elsevier.

7. EKLER

7.1. Python Kodları

```
#!/usr/bin/env python
# coding: utf-8

# In[ ]:

## Değişken Çıkarma
import numpy as np
import pandas as pd
import pandas
from scipy import stats
from scipy import signal
from biosppy.signals import tools as st
from tsfresh.feature_extraction import extract_features
import itertools
import statistics
import itertools
import statistics

from sklearn.preprocessing import StandardScaler
from math import floor

from xgboost import XGBClassifier
from sklearn.model_selection import RandomizedSearchCV
```

```
from sklearn.model_selection import KFold
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import roc_auc_score
#df = pandas.read_csv('ECG_10Sec_ECG.csv')
#df['id'] = df['sjlid'] + df['ageecg'].map(str)
#df['id'] = df['id'].map(str)
#df = df[['id','ix','I']]
#df = df.loc[df['id'] == 'SJL025330143.82']


# In[ ]:


data=pd.read_csv('dfecg_cnn1d.csv')
df= pd.read_csv('data3_stjude_all_full.csv')
cm_pos = df.index[df['CM'] == 1].tolist()
df = df.drop(df.index[cm_pos])
data = data.drop(data.index[cm_pos])
df.iloc[:, np.r_[2,13,15,18,50:56,60:62]] = StandardScaler().fit_transform(df.iloc[:, np.r_[2,13,15,18,50:56,60:62]])
# X = değişken değerleri values
fea = df.iloc[:, np.r_[2,5,12,13,15,18,51:56,60:62,283:288]]
idn =  df.iloc[:, np.r_[0]]
#X.to_csv('X.csv')


# In[ ]:
```

```
data=data.drop(data.index[[410]])

df=df.drop(df.index[[410]])


# In[ ]:


ecg = data.iloc[:,1:60001]

xdata = np.array(ecg)

print(np.argwhere(np.isnan(xdata)))

fea = np.asarray(fea)

idn = np.asarray(idn)


targ=data.iloc[:,60002]

ydata=np.array(targ)


print(xdata.shape)


# In[ ]:


inp_xdata=[]


for j in range(12):

    inp_xdata.append(xdata[0:,j*5000:(j+1)*5000])


# In[ ]:
```

```
#sinyal işleme
nop=80#user defined
sf=500#user defined
filsig=[]
def resample_filter(xdata):
    filtered, _, _ = st.filter_signal(signal=xdata,
                                      ftype='FIR',
                                      band='bandpass',
                                      order=60,#user defined
                                      frequency=[3, 45],#user defined
                                      sampling_rate=sf)
    xdata_resampled = signal.resample(filtered, nop)
    return xdata_resampled
for j in range(12):
    xdata_last = inp_xdata[j]
    xdata_resampled = np.apply_along_axis(resample_filter, 1, xdata_last)
    filsig.append(xdata_resampled)


# In[ ]:


filsig=np.array(filsig)
filsig=filsig.reshape(1217,960)
df1=pd.DataFrame(filsig)
df1['output'] = df['output']
df1['output'] = df['output'].values
```

```
df1['sjlid'] = df['sjlid']

df1['sjlid'] = df['sjlid'].values

cols = list(df1.columns)

cols = [cols[-1]] + cols[:-1]

df1 = df1[cols]

df1.to_csv('stjude_ecg_200_3_45_60.csv', index=False)


# In[ ]:


ecgf = df1.iloc[:,1:961]

xdataf = np.array(ecgf)

targf=df1.iloc[:,961]

ydataf=np.array(targf)

refdf=df1[(df1['output'] == 1)]

refecgf = refdf.iloc[:,1:961]

refxdataf = np.array(refecgf)

reftargf=df1.iloc[:,961]

refydataf=np.array(reftargf)

inp_xdataf=[]

inp_refxdataf=[]

for j in range(12):

    inp_xdataf.append(xdataf[0:,j*80:(j+1)*80])

for j in range(12):

    inp_refxdataf.append(refxdataf[0:,j*80:(j+1)*80])
```

```
# In[ ]:


# Sayısal Değerli Verilerin Semboller ile İfadesi
# Tüm EKG voltaj değerlerinin semboler ile ifadesi.
def fnum2seq2(df):
    t=0
    for j in alphabet:
        for i in range(df.shape[0]):
            if (type(df.value[i]) != str):
                if(df.value[i] <= myquants.iloc[t]):
                    df.value[i] = j
                elif(j==alphabet[len(alphabet)-1]):
                    df.value[i]= j
        if not (t>(len(alphabet)-3)):
            t=t+1
    return df


# In[ ]:


#her bir ECG ID için, sayısal serileri yazıya dönüştür
def df2str(df):
    l=[]
    b=df.id.max()
    for j in range(1,b+1):
        s="
```

```
        a=df.loc[df['id'] == j]
        for i in range(a.shape[0]):
            s=s+a.iloc[i,0]
        l.append(s)
    return(l)

# In[ ]:


#geçil olasılık matrislerinin hesaplanması
def ftp(s,alphabet,n,d):
    x=[''.join(x) for x in itertools.product(alphabet, repeat=n)]
    l=[]
    pl=[]
    for j in range(len(x)):
        t=0
        for i in range(len(s)-n+1):
            if (x[j]==s[i:i+n]):
                t=t+1
        l.insert(j,t)
    if d=='m':
        if not sum(l)==0:
            pl = [x / sum(l) for x in l]
        else:
            pl=l
    elif d=='r':
```

```
        for i in range(len(l)):
            k=[]
            r = i % len(alphabet)
            for j in range(len(alphabet)):
                k.insert(j,l[i-r+j])
            if not sum(k)==0:
                pl.append(l[i]/sum(k))
            else:
                pl.append(l[i])
    else:
        print ('Please enter m for sum of matrix probability, r for sum of row probability')
    #print(pl)
    return pl


# In[ ]:


#geçiş benzerlik matrislerinin hesaplanması
def fsim(tpdf,tprefdf):
    a= [[] for _ in range(len(tpdf))]
    b= [[] for _ in range(len(tpdf))]
    for h in range(len(tpdf)):
        for k in range(len(tprefdf)):
            a[h].append(sum([abs(i - j) for i, j in zip(tpdf[h],tprefdf[k])]))
        b[h]=statistics.mean(a[h])
    #print("i th order similarity to each refid:", a)
```

```
    return b


# In[ ]:


# temel OSMT fonksiyonu
#n sembol geöişlere kadar benzerlikler
sim12=[]
def fspr(df, refdf, myquants, alphabet,n,d):
    sdf=df2str(fnum2seq2(df))
    #print(sdf)
    srefdf=df2str(fnum2seq2(refdf))


    sim = [[] for _ in range(n)]
    for j in range(2,n+2):
        tpdf= [[] for _ in range(df.id.max())]
        for i in range(df.id.max()):
            tpdf[i]=ftp(sdf[i],alphabet,j,d)
        tprefdf= [[] for _ in range(refdf.id.max())]
        for i in range(refdf.id.max()):
            tprefdf[i]=ftp(srefdf[i],alphabet,j,d)
        sim[j-2]=fsim(tpdf,tprefdf)
    #print("i th order similarity:", sim)
    sim12.append(sim)
    #print(sim12)
    #simdataset = pd.DataFrame(sim,columns =['1stsim', '2ndsim','3thsim'])
```

```
    #print("similarity:",[ sum(row[i] for row in sim) for i in range(len(sim[0]))])
    ws = list()
    for j in range(0, len(sim[0])):
        tmp = 0
        for i in range(0, len(sim)):
            tmp = tmp + sim[i][j]/(5**(i+1))
            #print(sim[i][j])
            #print(sim[i][j]/5**(i+1))
            #print('tmp',tmp)
        ws.append(tmp)
    print("weighted similarity:", ws)
    return


# In[ ]:
alphabet='abcde' #user defined
for j in range(12):
    print('j',j)
    xdataf_last = inp_xdataf[j]
    xdataf_last = np.array(xdataf_last)
    #print(xdataf_last.shape)
    xdataf_last = xdataf_last.reshape(97360,1)
    df2=pd.DataFrame({"value": xdataf_last[:,0]})
    df2['id'] = [1 + floor(i / 80) for i in range(df2.shape[0])]
    refxdataf_last = inp_refxdataf[j]
    refxdataf_last = np.array(refxdataf_last)
```

```
    refxdataf_last = refxdataf_last.reshape(9120,1)
    refdf2=pd.DataFrame({"value": refxdataf_last[:,0]})
    refdf2['id'] = [1 + floor(i / 80) for i in range(refdf2.shape[0])]
    madf = df2.value.rolling(window=2).mean()
    myquants=madf.quantile([0.2,0.4,0.6,0.8]) #user defined
    d='r' #user defined 'm' or'r'
    print(myquants)
    df2=df2[72960:97360]
    df2['id'] = [1 + floor(i / 80) for i in range(df2.shape[0])]
    df2=df2.reset_index(drop=True)

    fspr(df2,refdf2,myquants,alphabet,6,d)

# In[ ]:

simdataset = np.reshape(sim12, (72,305)).T
simdataset = pd.DataFrame(simdataset)
#print(sim12)
#print(simdataset)

simdataset.columns = ['s1_1stsim', 's1_2ndsim','s1_3rdsim','s1_4thsim','s1_5thsim',
's1_6thsim','s2_1stsim', 's2_2ndsim','s2_3rdsim','s2_4thsim','s2_5thsim',
's2_6thsim','s3_1stsim', 's3_2ndsim','s3_3rdsim',
's3_4thsim','s3_5thsim', 's3_6thsim',
's4_1stsim', 's4_2ndsim','s4_3rdsim','s4_4thsim','s4_5thsim', 's4_6thsim','s5_1stsim',
's5_2ndsim','s5_3rdsim','s5_4thsim','s5_5thsim',
```

```
's5_6thsim','s6_1stsim', 's6_2ndsim','s6_3rdsim',
's6_4thsim','s6_5thsim', 's6_6thsim',
's7_1stsim', 's7_2ndsim','s7_3rdsim','s7_4thsim','s7_5thsim', 's7_6thsim','s8_1stsim',
's8_2ndsim','s8_3rdsim','s8_4thsim','s8_5thsim',
's8_6thsim','s9_1stsim', 's9_2ndsim','s9_3rdsim',
's9_4thsim','s9_5thsim', 's9_6thsim',
's10_1stsim', 's10_2ndsim','s10_3rdsim','s10_4thsim','s10_5thsim',
's10_6thsim','s11_1stsim', 's11_2ndsim','s11_3rdsim','s11_4thsim','s11_5thsim',
's11_6thsim','s12_1stsim', 's12_2ndsim','s12_3rdsim',
's12_4thsim','s12_5thsim', 's12_6thsim']

#print('simdataset',simdataset)

# In[ ]:

simdataset.to_csv('stjude_simdataset_part4.csv')
X = np.asarray(simdataset)

# In[ ]:

#XGBosst model paramterelerinin optimizasyonu
def random_opt_xg(X,y):
    n_estimators = [int(x) for x in np.linspace(start = 200, stop = 2000, num = 100)]
    gamma = [i/10.0 for i in range(0,5)]
    subsample = [i/10.0 for i in range(1,10)]
    learning_rate = [i/10.0 for i in range(0,10)]
```

```
    colsample_bytree = [i/10.0 for i in range(1,10)]
    reg_alpha = [1e-5, 1e-2, 0.1, 1, 10,100]
    max_depth = [int(x) for x in np.linspace(3, 10, num = 2)]
    min_child_weight = [1, 3, 5, 7]
    random_grid = {'gamma': gamma, 'subsample': subsample,'max_depth': max_depth, 'reg_alpha': reg_alpha,
                   'colsample_bytre': colsample_bytree,'min_child_weight ': min_child_weight ,'n_estimators': n_estimators}
    xg_classifier = XGBClassifier(random_state=42)
    xg_random = RandomizedSearchCV(estimator = xg_classifier, param_distributions = random_grid, n_iter = 100, cv = 5, random_state=42)
    clf_xg = xg_random.fit(X, y)
    print(clf_xg.best_params_)
    return clf_xg.best_params_
random_opt_xg(X,ydata)



# In[ ]:


#XGBoost parametrelerinin Bayesci Yaklaşım ile optimizasyonu
def xgb_evaluate(max_depth, gamma, colsample_bytree, n_estimators, min_child_weight, learning_rate,
                 subsample, reg_alpha):

    params = {'max_depth': int(max_depth),
              'subsample': subsample,
```

```
        'random_state': 42,
        'gamma': gamma,
        'colsample_bytree': colsample_bytree,
        'n_estimators' :int(n_estimators),
        'min_child_weight' : min_child_weight,
        'learning_rate' :learning_rate,
        'reg_alpha':reg_alpha
       }

kf = KFold(n_splits=5, shuffle=True, random_state=42)#n split user defined
x_boost_accuracy=[]
x_boost_recall=[]
x_boost_precision=[]
x_boost_f1=[]
x_boost_auc=[]
x_boost_specificity=[]
x_boost_accuracy_tr=[]
x_boost_recall_tr=[]
x_boost_precision_tr=[]
x_boost_f1_tr=[]
x_boost_auc_tr=[]
x_boost_specificity_tr=[]

for train_index, test_index in kf.split(X):
```

```
# capraz doğrulama için verinin bölünmesi
X_train, X_test ,y_train, y_test = X[train_index], X[test_index], ydata[train_index], ydata[test_index]


# XGBoosting Sınıflayıcısı


modelxgb = XGBClassifier()
modelxgb.set_params(**params)
clf_xg= modelxgb.fit(X_train,y_train)
p_xg=clf_xg.predict_proba(X_test)[:,1]
p_xg_tr=clf_xg.predict_proba(X_train)[:,1]
xg_auc = roc_auc_score(y_test, p_xg)
xg_auc_tr = roc_auc_score(y_train, p_xg_tr)
for i in range(len(p_xg)):
    if p_xg[i] < 0.06:
        p_xg[i]=0
    else:
        p_xg[i]=1
xg_accuracy=accuracy_score(y_test, p_xg)
xg_recall=recall_score(y_test, p_xg)
xg_precision=precision_score(y_test, p_xg)
xg_f1=f1_score(y_test, p_xg)
conf_mat = confusion_matrix(y_test, p_xg)
xg_specificity=conf_mat[0,0]/(conf_mat[0,1]+conf_mat[0,0])
for i in range(len(p_xg_tr)):
    if p_xg_tr[i] < 0.06:
```

```
        p_xg_tr[i]=0
    else:
        p_xg_tr[i]=1
xg_accuracy_tr=accuracy_score(y_train, p_xg_tr)
xg_recall_tr=recall_score(y_train, p_xg_tr)
xg_precision_tr=precision_score(y_train, p_xg_tr)
xg_f1_tr=f1_score(y_train, p_xg_tr)
conf_mat_tr = confusion_matrix(y_train, p_xg_tr)
xg_specificity_tr=conf_mat_tr[0,0]/(conf_mat_tr[0,1]+conf_mat_tr[0,0])
x_boost_accuracy.append(xg_accuracy)
x_boost_recall.append(xg_recall)
x_boost_precision.append(xg_precision)
x_boost_f1.append(xg_f1)
x_boost_auc.append(xg_auc)
x_boost_specificity.append(xg_specificity)
x_boost_accuracy_tr.append(xg_accuracy_tr)
x_boost_recall_tr.append(xg_recall_tr)
x_boost_precision_tr.append(xg_precision_tr)
x_boost_f1_tr.append(xg_f1_tr)
x_boost_auc_tr.append(xg_auc_tr)
x_boost_specificity_tr.append(xg_specificity_tr)
xgb_fea_imp=pd.DataFrame(list(clf_xg.get_booster().get_fscore().items()),
columns=['feature','importance']).sort_values('importance', ascending=False)
#print('',xgb_fea_imp)
#xgb_fea_imp.to_csv('xgb_fea_imp.csv')
```

```
        #print(clf_xg.feature_importances_)
        #from xgboost import plot_importance
        #plot_importance(clf_xg, )

    xg_accuracy=np.mean(x_boost_accuracy, axis=0)
    xg_recall=np.mean(x_boost_recall, axis=0)
    xg_precision=np.mean(x_boost_precision, axis=0)
    xg_f1=np.mean(x_boost_f1, axis=0)
    xg_auc=np.mean(x_boost_auc, axis=0)
    xg_specificity=np.mean(x_boost_specificity, axis=0)
    xg_accuracy_tr=np.mean(x_boost_accuracy_tr, axis=0)
    xg_recall_tr=np.mean(x_boost_recall_tr, axis=0)
    xg_precision_tr=np.mean(x_boost_precision_tr, axis=0)
    xg_f1_tr=np.mean(x_boost_f1_tr, axis=0)
    xg_auc_tr=np.mean(x_boost_auc_tr, axis=0)
    xg_specificity_tr=np.mean(x_boost_specificity_tr, axis=0)

    print (xg_accuracy,xg_recall,xg_precision,xg_f1,xg_auc,xg_specificity)
    print
(xg_accuracy_tr,xg_recall_tr,xg_precision_tr,xg_f1_tr,xg_auc_tr,xg_specificity_tr)
    # Used around 1000 boosting rounds in the full model
    #modelxgb = XGBClassifier()
    #modelxgb.set_params(**params)
    #xgb = modelxgb.fit(egi2, np.argmax(tr_y[i],axis=1))
    #test_predict = modelxgb.predict(tnew2)
```

```
    #print (f1_custom(np.argmax(te_y[i],axis=1), test_predict)[-1])
    print (modelxgb.get_params())

    # Bayesci yaklaşım maksimizasyon için dizayn edilmiş, minimizasyon için değil. Bu nedenle -RMSE değeri ile devam
    return xg_f1

# In[ ]:

# xg_boost paramterlerinin Bayesci yaklaşım ile optimizasyonu
gp_params = {"alpha": 1e-4}

xgb_bo = BayesianOptimization(xgb_evaluate, {'max_depth': (2, 5),
                                             'subsample': (0.6, 1.),
                                             'gamma': (0.0, 0.5),
                                             'colsample_bytree': (0.6, 1.),
                                             'n_estimators': (500,2000),
                                             'min_child_weight':(3, 7),
                                             'learning_rate':(0.05, 0.15),
                                             'reg_alpha':(8,13)
                                            })
# Negatif sayılarla başa çıkabilmek içil 'expected improvement acquisition function' kullan
# optimal sonuç için hesaba katıllması gereken birkaç parameter daha var
xgb_bo.maximize(init_points=200, n_iter=50, acq='ei',  **gp_params)
```