DEÜ FMD 23(67), 55-69, 2021

Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi
Dokuz Eylul University Faculty of Engineering Journal of Science and Engineering

Basılı/Printed ISSN: 1302-9304. Elektronik/Online ISSN: 2547-958X

Sigara Kullanma Durumunun Çoklu Fizyolojik Ölçümler Ve Makine Öğrenmesi Teknikleri Kullanılarak Tahmini

Prediction of smoking status by using multi-physiological measures and machine learning techniques

Aykut Eken [1*], Şevket Çalışkan [1], Soner Çivilibal[1], Pınar Deniz Tosun[1]

[1] Düzce Üniversitesi Mühendislik Fakültesi Biyomedikal Mühendisliği Bölümü, Düzce, TÜRKİYE

Sorumlu Yazar / Corresponding Author *: ekenaykut@gmail.com

Geliş Tarihi / Received:05.02.2020 *Araştırma Makalesi/Research Article*
Kabul Tarihi / Accepted: 13.08.2020 *DOI:10.21205/deufmd.2021236705*
Atıf şekli/ How to cite: EKEN, A., ÇALIŞKAN, Ş., ÇIVILIBAL, S., TOSUN, P.D.(2021). Sigara Kullanma Durumunun Çoklu Fizyolojik Ölçümler Ve Makine Öğrenmesi Teknikleri Kullanılarak Tahmini. DEUFMD, 23(67), 55-69.

Öz

Sigara kullanımı toplumlarda gerek sağlık gerek ekonomik açıdan ciddi kayıplara sebep olmaktadır. Kullanım seviyesinin ölçümünde bir altın standart bulunmamasına rağmen, Fagerstörm Nikotin Bağımlılık Testi (Fagerstörm Test for Nicotine Dependency – FTND) ve HONC (Hooked on Nicotine Checklist) gibi geleneksel testler ve çeşitli nörogörüntüleme yaklaşımları kişinin sigara içme durumunun seviyesi hakkında bir bilgi vermektedir. Bu çalışmada, objektif bir veri olan fizyolojik parametrelerin subjektif bir veri olan bağımlılık testlerinin yerine kullanım seviye tespitinde yeni bir yaklaşım olarak kullanılabileceğini göstermek amaçlanmıştır. Bu amaçla çeşitli seviyelerdeki sigara kullanıcılarından fizyolojik sinyaller (elektrokardiyogram (EKG), Solunum ve Fotopletismografi) toplanmıştır. Bu sinyallerden elde edilen çeşitli öz niteliklerden makine öğrenmesi yaklaşımları kullanılarak katılımcılar düşük seviye veya yüksek seviye olarak tahmin edilmeye çalışılmıştır. Çalışma için önceden FTND bağımlılık testine giren değişik kullanım seviyelerinde 95 katılımcı alınıp bu kişilerden sırasıyla 50 saniyelik EKG, solunum ve fotopletismografi sinyalleri alınmıştır. Öznitelik çıkarımından sonra, parametre optimizasyonu ve sınıflandırma içeren 10 kat içiçe çapraz geçerlilik gerçekleştirilmiştir. Yapılan sınıflandırma sonucunda destek vektör makinesi kullanılarak %93, diskriminant analizi kullanılarak ise %91 doğruluk başarımı elde edilmiştir. Bu sonuçlar, yukarıda belirtilen fizyolojik parametrelerin makine öğrenmesi algoritmaları aracılığı ile sigara kullanım durumunun tespitinde kullanılabileceğini göstermektedir.

Anahtar Kelimeler: *Sigara içme durumu, Fotopletismografi, EKG, Solunum, Makine Öğrenmesi, Sınıflandırma*

Abstract

Smoking causes severe economic and health losses in communities. Despite the lack of a gold standard for the measurement of usage level, conventional tests such as Fagerstörm Test for Nicotine Dependency (FTND), Hooked on Nicotine Checklist (HONC) and various neuroimaging approaches provide information about the level of smoking status. In this study, usage of objective physiological parameters was proposed as a new approach to detect level of status instead of subjective status tests. In order to achieve this physiological signals (i.e.., electrocardiogram (ECG), respiration and photoplestimography) were acquired from participants from different smoking status levels.

Participants' smoking status levels were predicted as high dependent and low dependent from features extracted from these physiological signals using machine learning approaches. For this study, 95 university students with different levels of smoking status were recruited according to FTND test results and ECG, respiration and photopletismography signals were acquired respectively for 50 seconds to provide data for machine learning models. After feature extraction, a 10 fold nested- cross validation that includes hyperparameter optimization and classification was performed. According to the classification results, 93 % accuracy and 91 % accuracy were found by using Support Vector Machine and Discriminant Analysis respectively. These results revealed that physiological parameters might be used to predict smoking status via machine learning algorithms.

Keywords: *Smoking status, Photopletsmography, ECG, Respiration, Machine Learning, Classification*

1. Giriş

Ülkemizde ve dünyada önemli bir halk sorunu olan sigara kullanımı, kardiyovasküler ve metabolik hastalıklar, pulmoner hastalıklar, hamilelik ve doğum sorunları, kanser gibi birçok hastalık kaynaklı ölümlerle direkt olarak ilişkilidir [1]. Dünya Sağlık Örgütü tarafından 2 senede bir yayınlanan, tütün kullanımın etkilerinin detaylandırıldığı bir raporda, sigara tüketiminin dünya genelinde yaklaşık 6 milyon prematür ölüme sebep olduğu bildirilmiştir [2]. Bununla birlikte inme, körlük, sağırlık, ağrılı hastalıklar, kemik erimesi gibi rahatsızlıklar için de risk faktörü oluşturmaktadır [3]. Bu etkilere sebep olan sigara kullanımının temelinde yatan asıl olay ise sigaranın beyine kontrollü bir dozda sağladığı nikotin maddesidir [4] ve bağımlılığın altında yatan sebepler halen tam olarak anlaşılamamıştır [1].

Sigara kullanım seviyesinin tespiti Fagerstörm Nikotin Bağımlılık Testi (Fagerstörm Test for Nicotine Dependence - FTND) [5] ve Hooked Nikotin Kontrol Listesi [6] gibi testler aracılığı ile yapılmakla birlikte, sigara içen ve içmeyen kişiler arasında yapısal manyetik rezonans (MR) [7, 8], Difüzyon Tensör Görüntüleme [9], manyetik rezonans spektroskopi (MRS) verileri arasında da farklar bulunmuştur [10]. Bununla birlikte gri madde değişikliklerini ortaya çıkarmak amacıyla kullanılan bir yapısal manyetik rezonans (MR) analiz tekniği olan voksel bazlı morfometri (VBM) verileri kullanılarak yapılan makine öğrenmesi çalışmalarında, bireysel vakalar için doğruluk değerinin düşük olması (%64) klinik tanı ve tedavide uygulanabilirliğini sınırlamaktadır [11]. Nikotin bağımlılığı olan ve olmayan sağlıklı bireylerin fonksiyonel MR verilerinden dinlenme durumu fonksiyonel bağlanırlık (DDFB) kullanan destek vektör makineleri (DVM) tabanlı bir sınıflandırılması yapıldığında VBM sonuçlarından daha yüksek bir doğruluk yüzdesi (%83) elde edildiği ortaya konmuştur [12]. Yine başka bir makine öğrenmesi çalışmasında sigara içen ve içmeyen kişilerin DDFB verilerinden faydalanılarak gerçekleştirilen bir sınıflandırma çalışmasında %88 doğruluk bulunmuştur [13]. Bununla birlikte kan biyokimyası ve hücre sayımı ile bir derin sinir ağı algoritması kullanılarak yapılan analizde sigara içme seviyesi %83 oranında doğru tespit edilmiştir [14]. Başka bir kan testi tabanlı sınıflandırma çalışmasında ise lojistik regresyon sınıflandırıcı kullanılarak %83,4 oranında başarım elde edilmiştir [15]. Amerika'daki Mayo Clinic tarafından geliştirilen doğal dil işleme bazlı bir sistem, hastalardan elde edilen yazılı veriler sonucu %92 duyarlılık ve %92 hassasiyet ile sigara bağımlılık durumunun tespitinde bulunmuştur [16]. Yine semantik öznitelikler kullanılarak gerçekleştirilen başka bir çalışmada da %90 duyarlılık, %89 hassasiyet bulunmuştur. [17]. Ancak MR ve kan testi klinik olarak uygulanılması pahalı ve çalışmalarda kullanılan verilerin elde edinimi hem bilgisayarlara hem de doktorlara ciddi bir iş yükü getirmektedir. Bununla birlikte, sigara kullanımının seviyesini belirlemek amaçlı öz-raporlama (self-reporting) testleri sübjektif yani testin üzerinde uygulandığı kişilerden kaynaklanan sebeplerden dolayı değişkenlikler gösterebileceğinden dolayı güvenilirlikleri tartışmalıdır.

Bu çalışmada amaç, sübjektif bir sigara kullanım seviyesi ölçüm kriteri olan testlerin yerine daha objektif ve aynı zamanda uygulanması ve analizi kolay veriler olan fizyolojik verilerin sigara kullanım seviyesinin belirlenmesinde

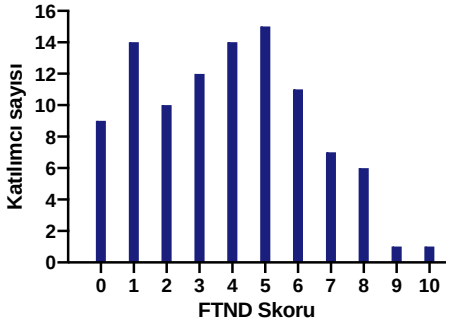
kullanılmasıdır. Sigara kullanımının arteryoskleroz [18] ve kronik obstrüktif pulmoner hastalığı riskini arttırmakla birlikte [19], sigara kullanımının etkileri Elektrokardiyografi (EKG) [20-22], Fotopletismografi [23] ve solunum [24] gibi fizyolojik ölçümler yardımıyla gözlemlenmiştir. Akut sigaranın kullanımının, EKG çalışmalarında, QT aralığı [25], kalp atım değişimi ve ST aralığı [21] üzerinde etkileri ile kronik sigara bağımlılığın QRS ve P dalgası üzerinde etkileri [26] literatürde bildirilmiştir. Buna ek olarak fotopletismografi ölçümlerinde elde edilen oksijen saturasyonun (SpO_2) sigara içenlerde içmeyenlere nazaran daha düşük seviyede olduğu da gözlemlenmiştir [27]. Sigaranın, solunum testi için kullanılan spirometre ölçümlerinde elde edilen parametrelerden biri olan zorlanmış soluk kapasitesinin (forced vital capacity) ise, sigara içenlerde içmeyenlere nazaran daha yüksek olduğunu göstermiştir [28]. Bu temel fizyolojik verilerden elde edilecek öz nitelikler birleştirilip farklı sınıflandırma teknikleri kullanılarak kişilerdeki sigara ölçüm seviyeleri daha objektif verilerle tahmin edilmeye çalışılmıştır. Bu çalışmayı gerçekleştirmekte iki temel amaç bulunmaktadır. Birincisi klinikte yapılan çoklu klasik ölçümlerden elde edilen fizyolojik parametreleri kullanarak, makine öğrenmesi teknikleri aracılığı ile bağımlılığı, tamamen sübjektif testlerden bağımsız olarak sınıfsal düzeyde tahmin edebilmek, ikincisi bu tekniklere sağlanan veriyi daha az maliyetli ve daha kolay analiz edilebilen ölçüm tekniklerinden elde ederek sigara ölçüm seviyesini tahmin etmek.

Yapılan literatür araştırmasına göre şu zamana kadar, EKG, Solunum ve Fotopletismografi gibi fizyolojik sinyallerden elde edilen öznitelikler ile makine öğrenmesi algoritmaları kullanılarak sigara kullanımını seviyesini tahmin etmeye odaklı bir çalışma gerçekleştirilmemiştir. Çalışma bu bahsedilen özgünlüğüne ek olarak kattığı diğer bir önemli yenilik de farklı üç fizyolojik ölçümden elde edilen (EKG, Solunum ve Puls Oksimetre) öznitelikleri bir arada kullanmasıdır.

2. Materyal ve Metot

2.1. Katılımcılar

Bu çalışmaya 95 sigara kullanan lisans öğrencisi (Erkek/Kadın : 79/16, yaş : 21,95 ± 2,17) katılmış olup analizlerde kendilerinden toplanan EKG, solunum ve fotopletismografi işaretleri kullanılmıştır. Katılımcılar deneyden bir saat öncesine kadar sigara tüketilmemesi istenmiş olup yine deneyden önce **** Etik Araştırmalar Kurulu tarafından onaylanan (Proje Numarası : 2018-512) ve Helsinki deklarasyonuna uygun olan protokol hakkında bilgilendirilmiş gönüllü olur formunu imzalatılmıştır. Katılımcıların sigara kullanım seviyesini ölçmek için FTND ölçeği kullanılmıştır. FTND ölçeği, standardize edilmiş bir nikotin bağımlılık testidir. Anket şeklinde sigara kullanıcısına yöneltilen 6 sorudan meydana gelmektedir ve skor diye belirlenen anket sonucu sorulara verilen yanıtların toplamıdır (Evet/Hayır, 1/0 ve çoklu yanıtlar 0-3 puan şeklinde skorlanır.) Elde edilen FTND skorları, 0-4 arası az bağımlı (sınıf 1, 57 kişi) ve 5-10 arası çok bağımlı (sınıf 2, 38 kişi) olacak şekilde 2 sınıfa indirgenmiştir. Katılımcıların FTND skorlarına göre dağılımları Şekil 1. de gösterilmektedir.

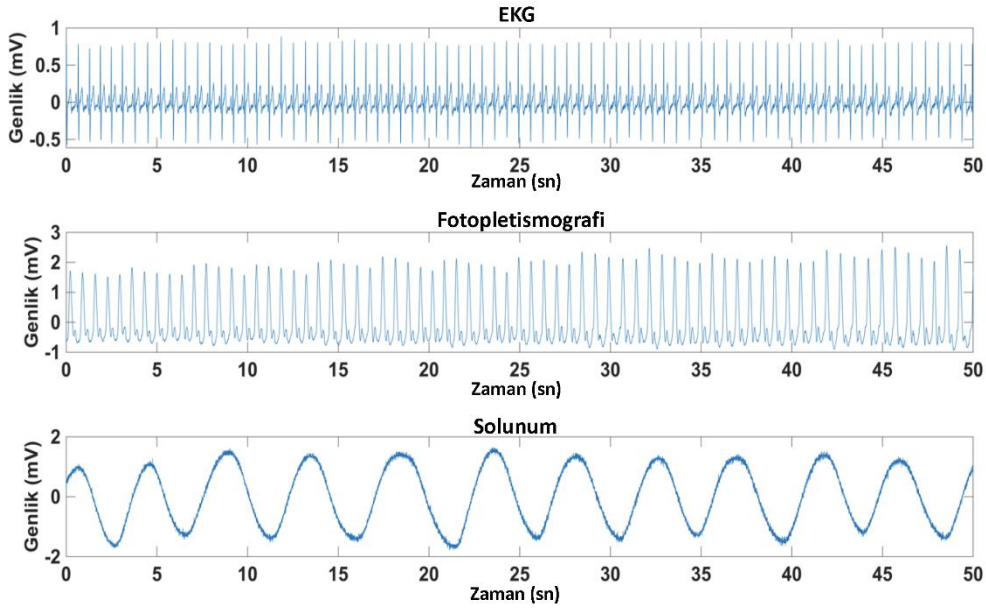


Şekil 1. Katılımcıların FTND skorlarına göre dağılımı

2.2. Deneysel Protokol ve Ölçüm Sistemi

Katılımcılardan dinlenim durumunda iken sırasıyla 50 saniyelik EKG, Solunum ve Fotopletismografi sinyalleri kaydedilmiştir. Ölçüm için KL-730 Biyomedikal Eğitim Seti ve her ölçüm için ilgili modülü (K&H Products, Co.) kullanılmıştır. EKG ölçümü yapılırken elektrotlar sol kol (left arm-LA), sağ kol (right arm-RA), sağ bacak (right leg-RL) ve sol bacak (left leg-LL) bölgesine bağlanmış olup, ön yükselteçten geçirilen sinyallere, 0.1-100 Hz kesim frekansı arasında bir bant geçiren filtre uygulanmıştır. Filtre çıkışından alınan sinyaller bir yükselteç aracılığı ile 10 kat daha yükseltildikten sonra,

sinyalleri 50 Hz şehir şebeke gürültüsünden arındırmak için analog filtre kullanılarak çıkış elde edilmiştir. Solunum sinyallerini toplarken, katılımcılara içinde bir sıcaklık sensörü bulunan bir maske takılarak bir dakika boyunca normal bir ritimde soluk alıp vermeleri istenmiştir. Elde edilen sinyaller, önce farksal bir yükselteçten geçirildikten sonra, 50 Hz bant durduran filtre ile şehir şebeke gürültüsünden arındırıldıktan sonra bir yükselteç işlemine tabi tutulmuştur. Fotopletismografi sinyalleri toplanırken, katılımcılardan 880 nm lik ışık kaynağı ve dedektörü içeren bir fotokuplör aracılığı toplanan sinyaller, öncelikle 2. dereceden bir yüksek geçiren filtreye daha sonra 1 Hz lik kesim frekansı olan bir yüksek geçiren filtre ile filtrelenmiştir. Filtrenin çıkışından alınan sinyaller, kazancı 100 olan bir yükselteç aracılığı ile yükseltildikten sonra, 4. derece kesim frekansı 10 Hz olan bir alçak geçiren filtre tarafından filtrelenmiştir. Elde edilen ham sinyallere bir örnek Şekil 2' de gösterilmektedir.



Şekil 2. Ölçümler sonucu elde edilen ham EKG, Fotopletismografi ve Solunum sinyalleri

2.3. Veri Ön-İşlemleme ve Öznitelik Çıkarımı

Modüllerden elde edilen sinyaller, toplanıp kaydedildikten sonra, öncelikle solunum, EKG ve fotopletismografi sinyallerindeki fizyolojik ve enstrümental gürültüleri yok etmek için 20 Hz kesim frekansı olan 4. derece bir alçak geçiren filtre uygulandı. Gürültüden arındırılan EKG, solunum ve fotopletismografi ve sinyallerinden öncelikle Yule-Walker yöntemi kullanılarak 20. Derece otoregresif (Autoregressive-AR) model katsayıları çıkartıldı. Yule-Walker otoregresif yöntemi, önceden belirlen bir pencere uzunluğundaki verilere otoregresif bir model uydurarak (fitting) ileri tahmin hatasını en küçük kareler (least squares) yöntemi ile çözme yöntemidir[29, 30]. AR modeli, bir zaman serisinin geçmişteki belirli sayıda değerini kullanarak bulunduğu noktadaki değerini kestirmeye yönelik kullanılan bir yöntemdir. En önemli avantajlarından birisi EKG, solunum ve fotopletismografi gibi durağan olmayan (non-stationary) sinyaller için iyi bir kestirim yöntemi olmasıdır. Eğer, elimizdeki zaman serisine x(n), genellikle sıfır ortalama ve beyaz gaussian olduğu varsayılan hata terimine e(n), AR katsayılarına a_p ve AR derecesine p der isek, x(n) Eş. 1'de gösterildiği üzere,

$$e(n) = x(n) + \sum_{k=1}^{p} a_p(k)x(n-k), \qquad (1)$$

$$a_p(0) = 1$$

Eş. 1'e göre elde edeceğimiz $\boldsymbol{a_k}$ katsayılarını bulmak için l gecikme (lag) olmak üzere her iki tarafı Eş. 2'deki gibi $x(n-l)$ ile çarpıp beklenen değerini (Expected value) buluruz.

$$E\{e(n)x(n-l)\} = \sum_{k=1}^{p} a_p(k)E\{x(n - k)x(n-l)\} + E\{x(n)x(n-l)\} \quad (2)$$

Eş 2'den toplam içerisindeki $E\{x(n-k)x(n-l)\}$ $\boldsymbol{r_{xx}}$ otokorelasyon değerini Eş 3. de gösterildiği gibi yazabiliriz.

$$E\{x(n-k)x(n-l)\} = r_{xx}(l-k) \quad (3)$$

Diğer toplamında Eş. 4 deki gibi yazarız.

$$E\{x(n)x(n-l)\} = r_{xx}(l) \quad (4)$$

$e(n)$ ile $x(n-l)$ değişkeninin beklenen değeri $E\{e(n)x(n-l)\}$ ise $l > 0$ iken 0 ve $l = 0$ iken $e(n)$ varyansına σ_e^2 eşit olacaktır. Dolayısı ile karşımıza Eş.5'deki gibi bir bağıntı çıkacaktır.

$$r_{xx}(l) + \sum_{k=1}^{p} a_p(k)r_{xx}(l-k) = \begin{cases} 0, & l > 0 \\ \sigma_e^2, & l = 0 \end{cases} \quad (5)$$

l > 0 olması durumunda ise Eş. 6'da gösterildiği üzere,

$$\sum_{k=1}^{p} a_p(k)r_{xx}(l-k) = -r_{xx}(k) \quad (6)$$

eşitliği ortaya çıkacaktır. Bu noktadan sonra otokorelasyon matrisi oluşturulur. Bu matrise R der isek Eş. 7'de gösterildiği üzere,

$$\underbrace{\begin{bmatrix} r_{xx}(0) & \cdots & r_{xx}(p-1) \\ \vdots & \ddots & \vdots \\ r_{xx}(p-1) & \cdots & r_{xx}(0) \end{bmatrix}}_{R} \underbrace{\begin{bmatrix} a_p(1) \\ . \\ . \\ a_p(p) \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} r_{xx}(1) \\ . \\ . \\ r_{xx}(p) \end{bmatrix}}_{r} \quad (7)$$

RA = r, doğrusal denklemine ulaşırız. Burada, A vektörünündeki $\boldsymbol{a_p}$ katsayıları, Levinson-Durbin özyinelemesi (Levinson-Durbin Recursion) kullanılarak elde edilir [31, 32]. Levinson-Durbin özyinelemesi doğrusal simetrik Toeplitz denklemlerini çözmek için kullanılan bir yöntemdir. Klasik bir matris çözümünde R matrisinin tersini almanın veya Gauss eleme yöntemi ile çözümün $O(p^3)$ kadar bir karmaşıklığı varken, Levinson-Durbin özyinelemesi ile bu karmaşıklık $O(p^2)$ ye kadar düşer[33]. Bu yöntemde, ilk olarak 0. derecedeki AR katsayısı $a_p(0) = 1$, olarak belirlenip başlangıç hata vektörü ϵ_0 $r_{xx}(0)$ olarak belirlendikten sonra algoritma Tablo 1'deki gibidir.

Tablo 1. Levinson-Durbin Özyinelemesi Algoritması

$\boldsymbol{a_p(0) = 1, \epsilon_0 = r_{xx}(0)}$

For $\boldsymbol{j = 0, 1, \ldots\ldots p-1}$

$\boldsymbol{\gamma_j = r_{xx}(j+1) + \sum_{i=1}^{j} a_j(i) r_{xx}}$

$\boldsymbol{\Gamma_{j+1} = -\gamma_j / \epsilon_j}$

For $\boldsymbol{i = 1, 2, \ldots\ldots, j}$

$\boldsymbol{a_{j+1}(i) = a_j(i) + \Gamma_{j+1}\, a_j^*(j - i + 1)}$

$\boldsymbol{a_{j+1}(j+1) = \Gamma_{j+1}}$

$\boldsymbol{\epsilon_{j+1} = \epsilon_j[1 - |\Gamma_{j+1}|^2]}$

Burada Γ_{j+1} j+1. yansıtma katsayısı, ϵ_j j. hata vektörünü temsil etmektedir. Yule-Walker denklemlerinden,Levinson-Durbin özyinelemesi yöntemi kullanılarak her katılımcı için üç ölçümden de (EKG, solunum, fotopletismografi) öznitelikleri çıkartıldıktan sonra, tüm öznitelikler bir araya getirilerek toplamda her bir fizyolojik verinin AR modelinden a_0 hariç 19 katsayı ile birlikte 57 tane öznitelik elde edilmiştir. Böylelikle 95 x 57 boyutlarında bir öznitelik vektörü elde ettik.

2.4. Hampel Filtresi Kullanarak Aykırılık Tespiti

Öz nitelik vektöründe aykırı değerleri (outliers) yok etmek için sıklıkla kullanılan yöntemlerden birisi olan Hampel filtresi kullanılmıştır [34, 35]. Hampel filtresi temelinde medyan filtresi tabanlı bir yöntem olup, bir $\boldsymbol{X} = x_1, x_2, x_3 \ldots\ldots x_N$ vektöründe l uzunluğunda kayan bir pencere üzerinde, belirlenen bir standart sapma değeri σ olmak üzere, i. indeksteki medyan değerini m_i ve

standart sapma σ_i değerlerini Eş. 8 ve 9'da gösterildiği şekilde bulunabilir.

$$\boldsymbol{m_i = median(x_{i-l}, x_{i-l+1}, \dots . x_i, \dots x_{i+l-1}, x_{i+l})} \quad (8)$$

$$\boldsymbol{\sigma_i = h\ median(|x_{i-l} - m_i|, .., |x_i - m_i|, \dots, |x_{i+l} - m_i|)} \quad (9)$$

Burada h için sabit bir değer olmak üzere, medyan standart sapma hesabını yansız (unbiased) elde etmek için kullanılmıştır ve $1/(\sqrt{2}\ erfc^{-1}(0{,}5)) = 1{,}4826$ e eşittir. Filtrenin i. eleman için sonuç d'_i Eş. 10'daki şekliyle bulunur.

$$d'_i = \begin{cases} m_i, & |d_i - m_i| > t\sigma_i \\ d_i, & |d_i - m_i| \leq t\sigma_i \end{cases} \quad (10)$$

Bu denklemde, t eşik değerini göstermekte olup eğer 0 olursa standart bir medyan filtreye karşılık gelmektedir. Bu çalışma için pencere uzunluğu $l = 5$ ve eşik değerini $t = 1$ olarak belirlenmiştir.

2.5. Öznitelik Seçimi

Öznitelik vektörünün içinden, en anlamlı öznitelikleri çıkatmak için, 5-katlı en küçük mutlak daralma ve seçme (Least Absolute Shrinkage and Selection Operator – LASSO) uygulanmıştır [36]. LASSO özellikle küçük boyutlu veri setlerinde çok etkili bir öznitelik seçim yöntemidir [37]. LASSO, cezalandırmalı bir doğrusal regresyon yöntemi olup regresyon katsayılarının hesaplanmasında L1 regülarizayonu uygulayarak bir ceza faktörü eklemesi temeline dayanır.. Eğer β ye regresyon katsayılarının olduğu d boyutlu vektör, y_i binomial değerlerin olduğu etiketli değerlerin (az bağımlı =0, çok bağımlı = 1) olduğu bir vektör, x_i i. katılımcının özniteliklerinin olduğu vektör $x_i : (x_1, x_2, x_3 \dots \dots x_d)^T$, N örneklem sayısı, λ pozitif L1 regülarizasyon parametresi olmak üzere, hedef fonksiyonu ve ceza faktörünü şu Eş. 11'deki gibi tanımlanabilir.

$$\min_{\hat{\beta}} \left\{ \frac{1}{N} \sum_{i=1}^{N} (y_i - x_i^T \hat{\beta})^2 + \lambda \sum_{j=1}^{d} |\beta_j| \right\} \quad (11)$$

Burada öznitelik seçerken uyguladığımız 10-katlı çapraz geçerlilik yöntemi farklı optimal λ değerini minimum ortalama kare hatasını elde etmek için kullanılmaktadır.. Bu optimal λ değerine göre, β katsayıları 0 olmayan katsayıların karşılık geldiği öznitelikler seçilmiştir. LASSO sonucunda seçilen öznitelikler Tablo 2'de gösterilmektedir. LASSO sonucu 9 tane (1 EKG, 2 solunum ve 5 fotopletismografi ile ilişkin) öznitelik seçilmiştir.

Tablo 2. LASSO sonucu seçilen öznitelikler ve karşılık geldiği modalite

Seçilen Otoregresif Katsayısı	İlgili Modalite
a_2	EKG
a_3	Fotopletismografi
a_6	Fotopletismografi
a_7	Fotopletismografi
a_9	Fotopletismografi
a_{13}	Fotopletismografi
a_{18}	Fotopletismografi
a_{18}	Solunum
a_{19}	Solunum

2.6. Destek Vektör Makinesi

Sınıflandırma işlemi için MATLAB Statistics and Machine Learning Toolbox (The MathWorks, Inc., Natick, MA, USA) kullanılmıştır ve her sınıflandırıcıya 10-katlı çapraz geçerlilik (10-fold cross validation) uygulanarak, aşırı uyum (overfitting) problemini en aza indirilmiştir.

Destek Vektör Makinesi (Support Vector Machine – SVM), sınıflandırma çalışmalarında sıklıkla kullanılan popüler bir tekniktir [38]. DVM, girdi olarak verilen verilerin önceden belirlenen sınıflarına göre maksimum mesafede (margin) ayrılması için bir hiperdüzlem oluşturur. Bu hiper düzlemi oluştururken kullanılan mesafenin belirlenmesinde kullanılan veri noktalarına, destek vektörler denir. Bu hiperdüzlem doğrusal bir kernel olabileceği gibi, doğrusal olmayan bir kernel de olabilir. Eğer girdi olarak verilen n gözlem sayılı veriyi tek bir gözlem için x_i ($x_i \in R$) ve çıktı olarak belirlenen sınıf y_i ($y_i \in \{-1, +1\}$) olarak yazılırsa, ideal bir karar hiperdüzlemi doğrusal bir kernel fonksiyonu için $w^T x + b = 0$ olarak temsil edilir. -1 ve +1 burada ikili sınıflandırmadaki sınıfları temsil etmektedir. Burada w hiperdüzleme dik olan ağırlık vektörü x girdi verisi b yanlılık değeri (bias) olarak tanımlanır. Her iki sınıfın en

ayrılabilir koşulda olması için optimal marjin uzunluğu $\frac{2}{\|\boldsymbol{w}\|}$ olmalıdır. Bu işlemler kuadratik programlama teknikleri ile marjin'i maksimize etmeyi sağlamaktadır. Bu, $\|\boldsymbol{w}\|$ vektörünü minimize eden b değeri ve w vektörü bulunarak yapılmaktadır. Dolayısı ile hedef fonksiyonu (objective function) $y_i(w^T x + b) \geq 1 - \varphi_i$ ve $\varphi_i \geq 0$ olma koşulu ile Eş. 12'deki gibi yazabilir.

$$\min_{w,b,\varphi_i} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \varphi_i \qquad (12)$$

Bu denklemde $\boldsymbol{w}$ ağırlık matrisi, C regülarizasyon parametresidir. Regülarizasyon parametresi aşırı uyum (overfitting) olmaması için kullanılan bir parametredir. Eğer C değeri büyük olursa, SVM sınıflandırıcısı daha az destek vektörü atar. Ancak bu da sınıflandırıcının eğitim süresinin uzamasına sebep olur. Burada φ_i gevşek değişkenleri temsil etmekte olup DVM algoritması bu değişkenleri, marjin sınırını geçen veri noktaları için hedef fonksiyonu cezalandırmak için kullanır. Burada eğer φ_i=0 a eşitse bu i. değişkenin o sınırı geçmediğini gösterir. Geçtiği takdirde $\varphi_i \geq 0$ olacaktır. Algoritma bu fonksiyonda optimum değerleri bulmak için Lagrange çarpanlarını kullanır. Eğer bu Lagrange katsayılarına $\alpha_1, \ldots \ldots \alpha_n$ dersek ayrılabilir sınıflar için Karush- Kuhn Tucker kuralı olan $\sum \alpha_i y_i = 0\,, \alpha_i \geq 0$, ve ayrılamayan sınıflar için $\sum \alpha_i y_i = 0\,, 0 \leq \alpha_i \leq C$ koşulu ile doğrusal DVM için Eş 13.'deki fonksiyonu bu katsayılara göre minimize etmek gerekmektedir.

$$\min_{\alpha_1,\ldots\ldots\alpha_n} \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j x_j{}' x_k \qquad (13)$$

Buradan elde edilen, skor fonksiyonu Eş. 14'te verilen denkleme göre hesaplanabilir.

$$\hat{y} = \sum_{i=1}^{n} \hat{a}_i\, y_i x' x_k + \hat{b} \qquad (14)$$

$\hat{b}$ yanlılık kestirimi, $\hat{a}_i$ $\hat{a}$ vektörünün i. kestirimidir. Burada Doğrusal olmayan (non-linear) DVM için ise bir $x_j{}' x_k$ çarpanı yerine bu veri noktalarının bir doğrusal olmayan fonksiyon K aracılığı ile dönüşümünün bu denkleme eklenmesinden sonra yine Lagrange katsayılarına $\alpha_1, \ldots \ldots \alpha_n$ dersek Karush- Kuhn Tucker kuralı gereği $\sum \alpha_i y_i = 0\,, 0 \leq \alpha_i \leq C\,,$ koşulu ile DVM fonksiyonu Eş. 15'teki gibi minimize edilir.

$$\min_{\alpha_1,\ldots\ldots\alpha_n} \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j\, K(x_i, x_j) \qquad (15)$$

Buradan elde edilen skor fonksiyonu Eş. 16'daki gibidir.

$$\hat{y} = \sum_{i=1}^{n} \hat{a}_i\, y_i K(x_j, x_k) + \hat{b} \qquad (16)$$

Bu denklemde x_j ve x_k veri noktaları olmak üzere, polinom kerneli $K(x_j, x_k) = (a + x_j{}' x_k)^d$, olmak üzere a sabit değeri ve d polinom derecesidir. Gauss fonksiyonu $K(x_j, x_k) = e^{\left(\frac{-\|x_j - x_k\|^2}{2\sigma^2}\right)}$ olmak üzere, σ standard sapma değerini temsil etmektedir.

2.7. Diskriminant Analizi

Doğrusal diskriminant analizinde her sınıftaki verilerin normal dağıldığını ve Σ kovaryans matrisi olarak ortak olduğunu varsayarsak; k sınıf olmak üzere doğrusal diskriminant fonksiyonu Eş. 17'deki gibi hesaplanabilir.

$$D_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2}\,\mu_k^T \Sigma^{-1} \mu_k + log p_k \qquad (17)$$

Bu denklemde μ_k her sınıfın ortalaması, p_k her sınıfın önsel olasılık (prior probability) değeri olarak nitelendirilmiştir. Bu denkleme göre doğrusal diskriminant sınıflandırma fonksiyonu $D_k(x)$ Eş. 18'deki gibidir.

$$S_k(x) = \arg\max_k D_k(x) \qquad (18)$$

Kuadratik Diskriminant Analizinde ise her sınıftaki verilerin normal dağıldığını ancak kovaryans matrisi Σ_k her sınıf için farklı olduğunu varsayarsak, kuadratik diskriminant fonksiyonu Eş. 19'daki gibi bulunabilir.

$$D_Q(x) = -\frac{1}{2} log|\Sigma_k| - \frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) + log p_k \qquad (19)$$

Bu durumda sınıflandırılma kuralı da; kuadratik diskriminant fonksiyonunu maksimize eden sınıfı bulmaktır. Bu duruma göre kuadratik diskriminant sınıflandırma fonksiyonu $S_Q(x)$ Eş. 20'deki gibi olmaktadır.

$$S_Q(x) = \arg\max_k D_Q(x) \quad (20)$$

Kullanılan sınıflandırıcı parametrelerinden gamma (γ) kovaryans matrisini regülarizasyonu Eş'21 ve 22'deki gibi kullanılır.

$$M = diag(X^T X) \quad (21)$$

$$\tilde{\Sigma} = (1-\gamma)\Sigma + \gamma M \quad (22)$$

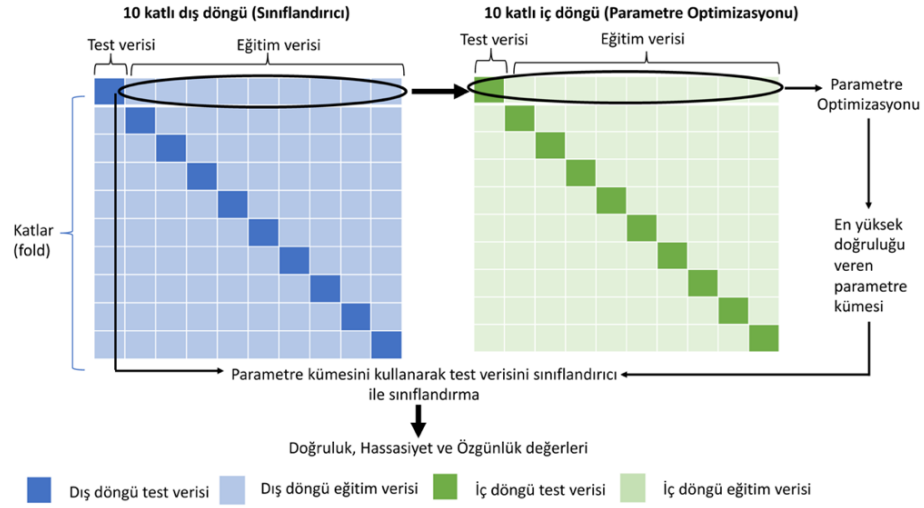
Bu denklemde $\tilde{\Sigma}$ varolan kovaryans matrisinin γ katsayısı aracılığı ile Eş. 22' de gösterildiği gibi yeni elde edilmiş halidir. Bu çalışmada kullanılan diskriminant şekillerinde, doğrusal için tüm sınıflarda aynı kovaryans matrisi, diagonal doğrusal için tüm sınıflarda aynı diagonal kovaryans matrisi, sözde doğrusal için tüm sınıflarda aynı kovaryans matrisinin sözde tersi kullanılmıştır. Kuadratik tip diskriminant için ise tüm sınıflarda farklı kovaryans matrisi, diagonal kuadratik için tüm sınıflarda farklı diagonal kovaryans matrisi ve sözde kuadratik için tüm sınıflarda farklı kovaryans matrisinin sözde tersi kullanılmıştır.

2.7.Sınıflandırıcıların Parametre Optimizasyonu

Öznitelik çıkarımını ve seçimini yaptıktan sonra veri setine maksimum hem DVM hem de DA için parametre optimizasyonu uygulanmıştır. Bu işlem veri setinin en yüksek doğruluk veren parametre kümesini elde etmeyi sağlamaktadır ve en iyi parametreyi elde etmek için Bayes optimizasyonu kullanılmıştır. DVM için, kernel fonksiyonu (Doğrusal, Polinom ve Gaussian), tüm kernel fonksiyonları için regülarizasyon parametresi (C), polinom kernel için polinom derecesi (d) ve Gaussian kerneli için standard sapma (σ) kullanılmıştır. Optimizasyon için kullanılan C değeri $10^{-6} - 10^{6}$, σ değeri $10^{-6} - 10^{6}$ ve d değeri 2-12 arasındadır. DA için, diskriminant şekli (doğrusal, kuadratik, diagonal doğrusal, diagonal kuadratik, sözde doğrusal, sözde kuadratik), regularizasyon katsayısı γ için 0-1 arasında değerler kullanılmıştır.

2.8. İç içe Çapraz Geçerlilik

Elde edilen öznitelik vektörü üzerinde sınıflandırma işlemini gerçekleştirilrken, 10-katlı içiçe çapraz geçerlilik yöntemini uygulanmıştır. İçiçe çapraz geçerlilik yöntemini veri sayımızın azlığından dolayı, öz nitelik çıkarımını dışında tutulup sadece sınıflandırma ve parametre optimizasyonu içerecek şekilde gerçekleştirilmiştir. Bu yöntemi sınıflandırıcı ve parametre optimizasyonunda yanlılık etkisini (bias effect) ve varyansı azalttığı için tercih edilmiştir. İç içe çapraz geçerlilik yönteminde iç ve dış olmak üzere iki döngü bulunmaktadır. Dış döngüde veri seti eğitim verisi ve test verisi olmak üzere k sayıda kata ayrıldıktan sonra, her bir katın eğitim verisi içteki döngüdeki parametre optimizasyonu için eğitim ve test verisi olarak j sayıda kata ayrılarak en ideal sınıflandırıcı parametrelerini bulmak için kullanılır. Parametre optimizasyonunda, eğitim verisi ile eğitilen sınıflandırıcı, değişik parametreler ile eğitildikten sonra en düşük sınıflandırma hatasını vermesini sağlayan parametre kümesi bulunur. Bu parametreler dıştaki döngüde test verisinin sınıflandırmasında kullanır. Bu iki döngü iç içe dönerek sonunda k tane sınıflandırıcı sonucu verecektir. Son olarak bu sınıflandırıcıdan elde edilen doğruluk, hassasiyet ve duyarlılık değerlerinin ortalaması ve standart sapması ile elde edilir. İç içe çapraz geçerlilik Şekil 3'de detaylandırılmıştır.



Şekil 3. İç içe çapraz geçerlilik

2.9. Sınıflandırma Performansını Değerlendirmek İçin Kullanılan Metrikler

Sınıflandırma performansının değerlendirilmesi için altı parametre kullanılmıştır. Bunlar, hassasiyet (sensitivity), özgünlük (specificity), doğruluk (accuracy), F1 skoru ve duyarlılıktır (Precision). Az ve çok bağımlı olarak belirlenen sınıfların parametreleri hesaplarken Tablo 3'de gösterilen karışıklık matrisi'nden (confusion matrix) faydalanılmıştır. Karışıklık matrisinde dört faktör bulunmaktadır. Bunlar doğru pozitif (DP), doğru negatif (DN), yanlış pozitif (YP) ve yanlış negatif (YN) değerleridir.

- DP, algoritmanın gerçekte çok bağımlı katılımcıları yine çok bağımlı olarak bulduğu toplam sayıya,
- DN, algoritmanın gerçekte az bağımlı katılımcıları yine az bağımlı olarak bulduğu toplam sayıya,
- YP, algoritmanın gerçekte az bağımlı katılımcıları çok bağımlı olarak bulduğu toplam sayıya,
- YN, algoritmanın gerçekte az bağımlı katılımcıları çok bağımlı olarak bulduğu toplam sayıya denilmektedir.

Tablo 3. Karışıklık Matrisi

Gerçek sınıf / Tahmini Sınıf	Çok bağımlı	Az bağımlı
Çok bağımlı	DP	YP
Az bağımlı	YN	DN

Bu faktörlerden faydalanarak sınıflandırma performansı için kullanılan duyarlılık, hassasiyet, F1 skoru, ve doğruluk faktörlerini ise Eş. 23-27 arasında verildiği şekilde hesaplanabilir.

$$Hassasiyet = \frac{DP}{DP + YN} \quad (23)$$

$$Özgünlük = \frac{DN}{DN + YP} \quad (24)$$

$$Doğruluk = \frac{DP + DN}{DP + DN + YP + YN} \quad (25)$$

$$Duyarlılık = \frac{DP}{DP + YP} \quad (26)$$

$$F1\ skoru = 2\ x\ \frac{Hassasiyet\ \ x\ Duyarlılık}{Hassasiyet + Duyarlılık} \quad (27)$$

3. Bulgular ve Tartışma

Bu çalışmada, öncelikle üç ölçümde de klinik olarak hastalıklara işaret edebilecek, EKG'de; Q-T aralığı P-Q aralığı, R-R aralığı ve kalp atımı, solunumda; solunum frekansı ve solunum sinyalinin tepe genliği, fotopletismografide ise; ortalama değer ve dikrotik çentik (dicrotic notch) latansı gibi öznitelikler kullanılmıştır.. EKG [20-22, 25, 26], fotopletismografi [23, 27] ve solunum [24, 28] gibi fizyolojik ölçümler kullanılarak sigara kullanan kişilerde başka klinik biyobelirteçler kullanılan çalışmalar literatürde mevcutken, bu çalışmada bu biyobelirteçler hem iki grup arasında istatistiksel olarak fark göstermedikleri hem de sınıflandırma başarımını da düşürdükleri için kullanılmamıştır. Bu sebeple sinyalin durağan olmaması (non-stationarity) özelliğinden faydalanarak AR model katsayıları kullanılmıştır. AR katsayıları EKG sinyallerinde kardiyak aritmi sınıflandırılmasında [39-41], elektroensefalografi (EEG) – beyin-bilgisayar arayüzü uygulamalarında [42] ve EEG mental görevlerin sınıflandırılmasında [43] kullanılmaktadır. Özniteliklerin seçimi için kullandığımız LASSO istatistiksel olarak güçlü bir öznitelik seçim yöntemidir. Klasik, varyans analizi (ANOVA), t-test ve temel bileşen analizi (Principal Component Analysis-PCA) gibi öznitelik azaltmaya yönelik tekniklerin beklenen performansı göstermediği durumlarda LASSO iyi bir alternatiftir. LASSO yöntemi elde ettiği katsayıları küçülterek ve yok ederek, varyansı bir yanlılık artırımı yapmaksızın azaltır. Bu da az sayıda gözlem ve çok sayıda özniteliğimiz olduğu koşullarda kullanılabilmesine olanak sağlar ve bununla birlikte LASSO, sınıflarla ilişkisiz öznitelikleri silerek, aşırı uyumu (overfitting) engellemektedir [37]. Çalışmamızda 95 tane gözlem ve 57 tane özniteliliğimiz olduğundan LASSO öznitelik seçiminde kullanılmıştır .Bununla birlikte çeşitli sebeplerden oluşabilen uç değerleri, Hampel filtresi gibi etkili bir yöntem ile elimine ederek veri setinde oluşabilecek aşırı uyum (overfitting) sorunu daha öznitelik vektörünü oluşturma seviyesinde engellenmiştir. Çalışmadaki en önemli yeniliklerden birisi de üç ölçüm tekniğinin (EKG, Fotopletismografi ve Solunum) birlikte kullanılmasıdır. Bunun sebebi, her teknik tek başına yanıltıcı sonuçlar verebilmekte olup farklı tekniklerden elde edilen özniteliklerin, tek bir teknikten elde edilen özniteliklere nazaran daha geniş bir öznitelik seçeneği sunmasıdır. Öznitelik birleşimi (Feature Fusion) özellikle biyometrik çalışmalarda başarımı arttırmak için sıklıkla kullanılan bir tekniktir [44]. Sınıflandırma algoritmalarının parametre optimizasyonu sonucu her bir kat sonrası elde edilen en iyi parametreler ve eğitim doğruluk (training accuracy) değerleri Tablo 4'de, bu modellere verilen test verisi aracılığı ile elde edilen doğruluk, hassasiyet, özgünlük, F1 skoru ve duyarlılık değerleri standard sapma değerleri ile birlikte DVM için Tablo 5'de DA için Tablo 6'da verilmiştir.

Tablo 4. Sınıflandırıcıların parametre optimizasyonu sonucunda her bir kat sonucu elde edilen en iyi parametreler ve eğitim doğruluk değerleri (D: Doğrusal (Linear), DD : Diagonal Doğrusal, DK : Diagonal Kuadratik, SD : Sözde Doğrusal, G: Gauss, P: Polinom, C : Destek Vektör Makinesi Regülarizasyon parametresi d : Polinom derecesi, σ : Destek Vektör makinesi Gauss kernel'i için standard sapma değeri, γ : Diskriminant Analizi regülarizasyon parametresi)

	Diskriminant Analizi			*Destek Vektör Makinesi*				
Kat (Fold)	Kernel Fonksiyonu	γ	Eğitim Doğruluğu	Kernel Fonksiyonu	σ	C	d	Eğitim Doğruluğu
1	D	0,9698	0,93	G	2,3133	51,1006	-	0,96
2	DD	0,0124	0,93	P	1	70729	3	0,94
3	DD	0,9981	0,94	P	1	963,0503	2	0,94
4	D	0,0001	0,93	P	1	57120	3	0,93

5	DK	0	0,94	P	1	8,1247	2	0,96
6	K	0	0,94	D	1	2,4463	-	0,94
7	SD	0,7790	0,92	P	1	18958	2	0,95
8	SD	0,0975	0,92	P	1	89,85	2	0,95
9	DK	0	0,90	P	1	35331	3	0,96
10	DK	0	0,94	P	1	291,7837	3	0,93

Doğruluk değerleri DVM için 0,93±0,10 ve DA için 0,91±0,08 bulunmuştur. İki sınıflandırıcının doğruluk sonuçları arasında istatistiksel olarak bir fark bulunmamaktadır (t (9) = 0,72, p =0,48, % 95 Güven Aralığı (GA) = [-0,04 0,09]). Hassasiyet değerleri DVM için 0,94±0,11 ve DA için 0,96±0,10 olarak bulunmuştur ve hassasiyet değerleri arasında istatistiksel olarak bir fark bulunmamaktadır (t(9)=1,00, p=0,34, % 95 GA=[-0,06 0,02]). Özgünlük değerleri DVM için 0,92±0,16 ve DA için 0,83±0,21 bulunmuş olup aralarında istatistiksel olarak bir fark bulunmamaktadır (t(9)=1.06, p=0.31, % 95 GA=[-0.10 – 0.28]). F1 skorları DVM için 0,94±0,09 ve DA için 0,93±0,06 bulunmuş olup aralarında istatistiksel olarak bir fark bulunmamaktadır (t(9)=0.65, p=0.53, % 95 GA=[-0.03 0.06]). Duyarlılık değerleri DVM için 0,95±0,09 ve DA için 0,91±0,09 olup aralarında istatistiksel olarak bir fark bulunmamaktadır (t(9)=0.91, p=0.38, %95 GA =[-0.05 – 0.13]). Sonuçlarla birlikte test verisinin alıcı işletim karakteristiği (Reciever Operating Characteristic – ROC) eğrisi Şekil 4'te gösterilmektedir.

Tablo 5. Destek Vektör Makinesi Sınıflandırma sonuçları

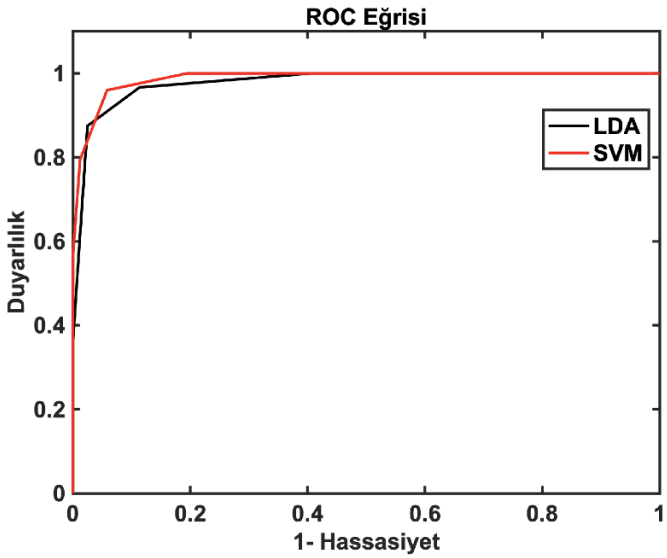
Kat (Fold)	*Doğruluk*	*Hassasiyet*	*Özgünlük*	*F1 Skoru*	*Duyarlılık*
1	1,00	1,00	1,00	1,00	1,00
2	0,80	1,00	0,50	0,86	0,75
3	1,00	1,00	1,00	1,00	1,00
4	1,00	1,00	1,00	1,00	1,00
5	0,70	0,67	0,75	0,73	0,80
6	1,00	1,00	1,00	1,00	1,00
7	1,00	1,00	1,00	1,00	1,00
8	1,00	1,00	1,00	1,00	1,00
9	0,89	0,80	1,00	0,89	1,00
10	1,00	1,00	1,00	1,00	1,00
Ortalama / Standard Sapma	0,93±0,10	0,94±0,11	0,92±0,16	0,94±0,09	0,95±0,09

Tablo 6. Diskriminant Analizi Sınıflandırma Sonuçları

Kat (Fold)	*Doğruluk*	*Hassasiyet*	*Özgünlük*	*F1 Skoru*	*Duyarlılık*
1	1,00	1,00	1,00	1,00	1,00
2	0,90	1,00	0,75	0,92	0,86
3	0,90	1,00	0,75	0,92	0,86
4	0,90	1,00	0,75	0,92	0,86
5	0,80	0,67	1,00	0,80	1,00
6	1,00	1,00	1,00	1,00	1,00
7	0,78	1,00	0,33	0,86	0,75
8	1,00	1,00	1,00	1,00	1,00
9	0,89	1,00	0,75	0,91	0,83
10	1,00	1,00	1,00	1,00	1,00
Ortalama / Standard Sapma	0,91±0,08	0,96±0,10	0,83±0,21	0,93±0,06	0,91±0,09

Önceki sigara bağımlılığının makine öğrenmesi teknikleri ile sınıflandırılması çalışmalarına nazaran daha yüksek başarım elde edilmesinin temel sebebi klinik veriler (kortikal kalınlık, gri madde hacmi, DDFB) gibi parametreler yerine istatistiksel yaklaşımlarla çoklu fizyolojik ölçümlerden elde edilmesinden kaynaklandığı düşünülmektedir. Yapısal MR verilerden elde edilen volümetrik verilerle yapılan bir çalışma sonucu 56 katılımcılı (28 sigara içen, 28 sigara içmeyen) bir çalışmada doğruluk oranı % 64 bulunmuştur [11]. Bir diğer DDFB tabanlı çalışmada ise 42 katılımcıdan (21 sigara içen, 21 sigara içmeyen) %83,3 lük bir doğruluk elde edilmiştir [12]. Son dönemde yapılan, 126 kişinin katıldığı (63 sigara içen 63, sigara içmeyen) ve yine DDFB kullanılarak yapılan bir sınıflandırma çalışmasında ise %88 oranında başarım bulunmuştur [13]. Kan biyokimyası ve hücre sayımı bazlı çalışmalarda ise derin sinir ağı kullanarak gerçekleştirilen bir sınıflandırma çalışmasında sigara içme seviyesi %83 [14] diğer bir 149.000 katılımcı (39000 sigara içen, 110000 sigara içmeyen) kan testi tabanlı sınıflandırma çalışmasında ise lojistik regresyon sınıflandırıcı kullanılarak %83.4 oranında başarım elde edilmiştir [15]. Bu çalışmada yukarıda bahsedilen çalışmalarla uyumlu olarak sigara aiçenlerde sınıflandırmada yüksek bir doğruluk ve hassasiyetle gerçekleştirilmiştir. Buna ek olarak bu çalışmada ilk defa, sigara bağımlıları ile sigara bağımlısı olmayan katılımcılar arasında fark olduğu bulunan fizyolojik veriler kullanarak %90 ve üstü bir başarım ile elde edilmiştir. Bu yüksek doğruluk sayılarının parametre optimizasyonu kullanmasının önemli bir etkisi olduğu düşünülmektedir. Bununla birlikte içiçe çapraz geçerlilik uygulanaması da elde edilen sonuçların yanlılığı ve varyansını azaltmıştır. Öz nitelik seçimi bu içiçe döngüye katılarak hangi modalitelerden elde edilen özniteliklerin nihai sonucun elde edilmesinde etkinliğinin olduğu gösterilebilmiştir. Mamoshina ve arkadaşlarının [14], Wetherhill ve arkadaşlarının [13] ve Savova ve arkadaşlarının [16] çalışması hariç genel olarak literatürdeki diğer çalışmalarla karılaştırıldığında da bu çalışmanın sonuçlarının hem daha yüksek katılımcıya sahip olduğu hem de bu daha yüksek bir doğruluk değerlerine (DA ile %91 ve SVM ile %93) ulaştığı görülmüştür. Bununla birlikte çalışmada kullanılan fizyolojik verilerin, sigara bağımlılığının tespitinde birlikte kullanılmasının avantajlarının yanı sıra FTND ve HOND gibi anketi uygulayan ve soruları yanıtlayan tütün bağımlılarının yanıtlarına ilişkilerinden ötürü

sübjektif testlere iyi bir alternatif olabileceği ve bununla birlikte MR uygulamaları ile karşılaştırıldığında daha az maliyetli ve analizi kolay ölçümler olduğu da gösterilmiştir.



Şekil 4. Destek Vektör Makinesi (SVM) ve Diskriminant Analizi (DA) için ROC eğrisi

4. Sonuçlar

Bu çalışmada sigara bağımlılığının tahmini için klinik kullanımda olan FTND, HOND gibi güvenirliği düşük öz raporlama testleri ve MR gibi pahalı bir ölçüm yöntemleri yerine elde edilmesi daha kolay ve ucuz fizyolojik ölçümler olan EKG, solunum ve fotopletismografi sinyalleri kullanılmıştır. Bu sinyallerden oluşturulan veri setinde makine öğrenmesi algoritmaları kullanılarak sigara bağımlılığı tahmini yapılmaya çalışılmıştır. Her iki sınıflandırıcıda da (SVM, DA) yüksek oranda doğruluk hassasiyet ve özgünlük bulunmuştur. İki sınıflandırıcının performans ölçütleri arasında bir fark bulunamamıştır.

Gerçekleştirilen çalışmanın üç önemli sınırlaması vardır. Bunlardan ilki katılımcıların yaş aralığının kısıtlığıdır (21.95 ± 2.17). Literatürde genç yaş diye tanımlanan bu aralık sigara içmenin etkilerinin vücutta belirli düzeylerde tolere edilebilmesi dolayısıyla, başka yaş guruplarının bağımlılık seviyelerinin incelenmesi esnasında bu çalışmada kullanılan yöntemlerin doğruluk ve hassasiyet ölçülerinde ne seviyelerde bir sınıflandırma gerçekleştirebileceği bilinmemektedir. Literatürdeki sınıflandırma çalışmalarında genel olarak gözlemlediğimiz bu çalışmaların erken yaşta yapıldığıdır. Pariyadath ve arkadaşlarının gerçekleştirdiği çalışmada iki grubun (sigara içenler ve kontroller) yaş ortalaması sırası ile 38 ve 39 dur [12]. Yine benzer bir çalışmada sınıflandırılan katılımcı gruplarının yaş ortalamaları 34 ve 35 dir [13]. Bir başka çalışmada ise 26 yaş ortalamasına sahip bir kadın popülasyonunda sınıflandırma çalışması gerçekleştirilmiştir [45]. Erken yaşta, bu tahminin gerçekleştirilmesi ilerleyen yaşlarda sigaranın bağımlılığının önlenmesini kolaylaştırabilmektedir [46]. İlerleyen yaşlarda, sigara bağımlılığından haricinde başka sebeplerden dolayı kaynaklabilecek rahatsızlıklar ve fizyolojik sinyallere yansımaları bu tahminin yapılmasını zorlaştırabilir.

Çalışmanın sınırlamalarından ikincisi ise katılımcıların cinsiyet sayılarındaki orantısızlıktır. Bu çalışmada hem geniş bir yaş aralığında katılımcı bulmak konusunda zorluklar yaşanmış hem de ağırlıkla erkek öğrenciler çalışmaya katılım konusunda istek göstermişlerdir. Bu sebeple de, cinsiyetin bu fizyolojik veriler üzerindeki etkileri ve kullanılan yöntemin sigara bağımlılığının tespitinde cinsiyete göre sınıflandırma yapılırken ne

seviyede çalışacağı konusu araştırılacak bir konu olarak kalmaktadır. Çalışmanın geçerliliğini güçlendirmek için daha fazla katılımcı, denk bir cinsiyet dağılımı ve daha geniş bir yaş aralığında çalışılması gerekmektedir. Son dönemde yapılan bir çalışmada, sigara bağımlılığının cinsiyetler arasında DDFB ağları arasında olan bir farkla ilişkili olabileceğini ortaya koymuştur [47]. Çalışmanın diğer önemli sınırlaması ise parametre optimizasyonu ve sınıflandırmayı iç içe bir çapraz geçerlilik ile gerçekleştirerek sonuçların yanlılık ve varyans etkisinin minimize edilmesine rağmen, çalışmanın tamamen ayrı ve bağımsız bir veri seti ile geçerliliğini gerçekleştirilememesidir. İlerleyen çalışmalarda daha büyük fizyolojik veri setleri ile çalışarak kullanılan öznitelik ve yöntemlerin klinik uygulamalarda kullanabilirliği konusuna güvenilirliği arttırılacaktır.

Kaynakça

[1] West, R. 2017. Tobacco smoking: Health impact, prevalence, correlates and interventions, Cilt. 32 s. 1018-1036. 10.1080/08870446.2017.1325890

[2] WHO, WHO report on the global tobacco epidemic, 2013. Enforcing bans on tobacco advertising, promotion and sponsorship. Geneva: World Health Organization (in English), 2013, p. 202 pp.

[3] Services, U. D. o. H. a. H., in The Health Consequences of Smoking: A Report of the Surgeon General, (Reports of the Surgeon General. Atlanta (GA), 2004, p. 62.

[4] West, R. 2009. The multiple facets of cigarette addiction and what they mean for encouraging and helping smokers to stop, Cilt. 6 s. 277-83.

[5] Heatherton, T. F., Kozlowski, L. T., Frecker, R. C. and Fagerstrom, K.-O. 1991. The Fagerström Test for Nicotine Dependence: a revision of the Fagerstrom Tolerance Questionnaire, Cilt. 86 s. 1119-1127. 10.1111/j.1360-0443.1991.tb01879.x

[6] DiFranza, J. R., Savageau, J. A., Fletcher, K., Ockene, J. K., Rigotti, N. A., McNeill, A. D., Coleman, M. and Wood, C. 2002. Measuring the loss of autonomy over nicotine use in adolescents: the DANDY (Development and Assessment of Nicotine Dependence in Youths) study, Cilt. 156 s. 397-403.

[7] Brody, A. L., Mandelkern, M. A., Jarvik, M. E., Lee, G. S., Smith, E. C., Huang, J. C., Bota, R. G., Bartzokis, G. and London, E. D. 2004. Differences between smokers and nonsmokers in regional gray matter volumes and densities, Cilt. 55 s. 77-84. 10.1016/s0006-3223(03)00610-3

[8] Gallinat, J., Meisenzahl, E., Jacobsen, L. K., Kalus, P., Bierbrauer, J., Kienast, T., Witthaus, H., Leopold, K., Seifert, F., Schubert, F. and Staedtgen, M. 2006. Smoking and structural brain deficits: a volumetric MR investigation, Cilt. 24 s. 1744-50. 10.1111/j.1460-9568.2006.05050.x

[9] Paul, R. H., Grieve, S. M., Niaura, R., David, S. P., Laidlaw, D. H., Cohen, R., Sweet, L., Taylor, G., Clark, R. C., Pogun, S. and Gordon, E. 2008. Chronic cigarette smoking and the microstructural integrity of white matter in healthy adults: a diffusion tensor imaging study, Cilt. 10 s. 137-47. 10.1080/14622200701767829

[10] Domino, E. F. 2008. Tobacco smoking and MRI/MRS brain abnormalities compared to nonsmokers, Cilt. 32 s. 1778-81. 10.1016/j.pnpbp.2008.09.004

[11] Ding, X., Yang, Y., Stein, E. A. and Ross, T. J. 2015. Multivariate classification of smokers and nonsmokers using SVM-RFE on structural MRI images, Cilt. 36 s. 4869-4879. 10.1002/hbm.22956

[12] Pariyadath, V., Stein, E. A. and Ross, T. J. 2014. Machine learning classification of resting state functional connectivity predicts smoking status, Cilt. 8 s. 425. 10.3389/fnhum.2014.00425

[13] Wetherill, R. R., Rao, H., Hager, N., Wang, J., Franklin, T. R. and Fan, Y. 2019. Classifying and characterizing nicotine use disorder with high accuracy using machine learning and resting-state fMRI, Cilt. 24 s. 811-821. 10.1111/adb.12644

[14] Mamoshina, P., Kochetov, K., Cortese, F., Kovalchuk, A., Aliper, A., Putin, E., Scheibye-Knudsen, M., Cantor, C. R., Skjodt, N. M., Kovalchuk, O. and Zhavoronkov, A. 2019. Blood Biochemistry Analysis to Detect Smoking Status and Quantify Accelerated Aging in Smokers, Cilt. 9 s. 142. 10.1038/s41598-018-35704-w

[15] Frank, C., Habach, A., Seetan, R. and Wahbeh, A., Predicting Smoking Status Using Machine Learning Algorithms and Statistical Analysis. 2018, pp. 184-189.

[16] Savova, G. K., Ogren, P. V., Duffy, P. H., Buntrock, J. D. and Chute, C. G. 2008. Mayo clinic NLP system for patient smoking status identification, Cilt. 15 s. 25-8. 10.1197/jamia.M2437

[17] McCormick, P. J., Elhadad, N. and Stetson, P. D. 2008. Use of semantic features to classify patient smoking status, Cilt. 2008 s. 450-454.

[18] Poredos, P., Orehek, M. and Tratnik, E. 1999. Smoking is associated with dose-related increase of intima-media thickness and endothelial dysfunction, Cilt. 50 s. 201-8. 10.1177/000331979905000304

[19] Rabe, K. F., Hurd, S., Anzueto, A., Barnes, P. J., Buist, S. A., Calverley, P., Fukuchi, Y., Jenkins, C., Rodriguez-Roisin, R., van Weel, C., Zielinski, J. and Global Initiative for Chronic Obstructive Lung, D. 2007. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary, Cilt. 176 s. 532-55. 10.1164/rccm.200703-456SO

[20] Devi, M. R., Arvind, T. and Kumar, P. S. 2013. ECG Changes in Smokers and Non Smokers-A Comparative Study, Cilt. 7 s. 824-6. 10.7860/JCDR/2013/5180.2950

[21] Ramakrishnan, S., Bhatt, K., Dubey, A. K., Roy, A., Singh, S., Naik, N., Seth, S. and Bhargava, B. 2013. Acute electrocardiographic changes during smoking: an observational study, Cilt. 3 s. 10.1136/bmjopen-2012-002486

[22] Bodin, F., McIntyre, K. M., Schwartz, J. E., McKinley, P. S., Cardetti, C., Shapiro, P. A., Gorenstein, E. and Sloan, R. P. 2017. The Association of Cigarette Smoking With High-Frequency Heart Rate Variability: An Ecological Momentary Assessment Study, Cilt. 79 s. 1045-1050. 10.1097/PSY.0000000000000507

[23] Glass, K. L., Dillard, T. A., Phillips, Y. Y., Torrington, K. G. and Thompson, J. C. 1996. Pulse oximetry correction for smoking exposure, Cilt. 161 s. 273-6.

[24] Irizar-Aramburu, M. I., Martinez-Eizaguirre, J. M., Pacheco-Bravo, P., Diaz-Atienza, M., Aguirre-Arratibel, I., Pena-Pena, M. I., Alba-Latorre, M. and Galparsoro-Goikoetxea, M. 2013. Effectiveness of spirometry as a motivational tool for smoking cessation: a clinical trial, the ESPIMOAT study, Cilt. 14 s. 185. 10.1186/1471-2296-14-185

[25] Akbarzadeh, M. A., Yazdani, S., Ghaidari, M. E., Asadpour-Piranfar, M., Bahrololoumi-Bafruee, N., Golabchi, A. and Azhari, A. 2014. Acute effects of smoking on QT dispersion in healthy males, Cilt. 10 s. 89-93.

[26] Chatterjee, S., Kumar, S., Dey, S. K. and Chatterjee, P. 1989. Chronic effect of smoking on the electrocardiogram, Cilt. 30 s. 827-39.

[27] Özdal, M., Pancar, Z., Çınar, V., Bilgiç, M., 2017. Effect of Smoking on Oxygen Saturation in Healthy Sedentary Men and Women, Cilt. 4 s. 178-182.

[28] Tantisuwat, A. and Thaveeratitham, P. 2014. Effects of smoking on chest expansion, lung function, and respiratory muscle strength of youths, Cilt. 26 s. 167-70. 10.1589/jpts.26.167

[29] Walker, G. T. 1931. On periodicity in series of related terms, Cilt. 131 s. 518-532. 10.1098/rspa.1931.0069

[30] Yule, G. U. 1927. VII. On a method of investigating periodicities disturbed series, with special reference to Wolfer's sunspot numbers, Cilt. 226 s. 267-298. 10.1098/rsta.1927.0007

[31] Durbin, J. 1960. The Fitting of Time-Series Models, Cilt. 28 s. 233-244. 10.2307/1401322

[32] Levinson, N. 1946. The Wiener (Root Mean Square) Error Criterion in Filter Design and Prediction, Cilt. 25 s. 261-278. 10.1002/sapm1946251261

[33] Hayes, M. H., Statistical Digital Signal Processing and Modeling. John Wiley & Sons, Inc., 1996.

[34] Hampel, F. R. 1971. A general qualitative definition of robustness, Cilt. s. 1887-1896.

[35] Hampel, F. R. 1974. The influence curve and its role in robust estimation, Cilt. 69 s. 383-393.

[36] Tibshirani, R. 1996. Regression Shrinkage and Selection via the Lasso, Cilt. 58 s. 267-288.

[37] Zou, H. and Hastie, T. 2005. Regularization and Variable Selection via the Elastic Net, Cilt. 67 s. 301-320.

[38] Vapnik, V. N. 1995. The Nature of Statistical Learning, Cilt. s.

[39] Ge, D., Srinivasan, N. and Krishnan, S. M. 2002. Cardiac arrhythmia classification using autoregressive modeling, Cilt. 1 s. 5-5. 10.1186/1475-925X-1-5

[40] Padmavathi, K. and Ramakrishna, K. S. 2015. Classification of ECG Signal during Atrial Fibrillation Using Autoregressive Modeling, Cilt. 46 s. 53-59. https://doi.org/10.1016/j.procs.2015.01.053

[41] Xi, Q., Sahakian, A. V. and Swiryn, S. 2003. The effect of QRS cancellation on atrial fibrillatory wave signal characteristics in the surface electrocardiogram, Cilt. 36 s. 243-9. 10.1016/s0022-0736(03)00046-3

[42] Vidaurre, D., Bielza, C. and Larrañaga, P. 2013. Classification of neural signals from sparse autoregressive features, Cilt. 111 s. 21-26. https://doi.org/10.1016/j.neucom.2012.12.013

[43] Anderson, C. W., Stolz, E. A. and Shamsunder, S. 1998. Multivariate autoregressive models for classification of spontaneous electroencephalographic signals during mental tasks, Cilt. 45 s. 277-86. 10.1109/10.661153

[44] Xin, Y., Kong, L., Liu, Z., Wang, C., Zhu, H., Gao, M., Zhao, C. and Xu, X. 2018. Multimodal Feature-Level Fusion for Biometrics Identification System on IoMT Platform, Cilt. 6 s. 21418-21426. 10.1109/ACCESS.2018.2815540

[45] Kharabsheh, M. K., Meqdadi, O., Al-Abed, M. A., Veeranki, S. P., Abbadi, A. and Alzyoud, S. 2019. A Machine Learning Approach for Predicting Nicotine Dependence, Cilt. 10 s.

[46] Riggs, N. R., Chou, C. P., Li, C. and Pentz, M. A. 2007. Adolescent to emerging adulthood smoking trajectories: when do smoking trajectories diverge, and do they predict early adulthood nicotine dependence?, Cilt. 9 s. 1147-54. 10.1080/14622200701648359

[47] Beltz, A. M., Berenbaum, S. A. and Wilson, S. J. 2015. Sex differences in resting state brain function of cigarette smokers and links to nicotine dependence, Cilt. 23 s. 247-254. 10.1037/pha0000033