Girdap Optimizasyon Algoritması Tabanlı Destek Vektör Makineleri ile Diyabet Tespiti

Diabetes Determination via Vortex Optimization Algorithm Based Support Vector Machines

Utku Köse [1], *Gür Emre Güraksın* [2], *Ömer Deperlioğlu* [2]

[1] Bilgisayar Bilimleri Uygulama ve Araştırma Merkezi
Uşak Üniversitesi
utku.kose@usak.edu.tr

[2] Bilgisayar Mühendisliği Bölümü
Afyon Kocatepe Üniversitesi
emreguraksin@aku.edu.tr

[3] Bilgisayar Teknolojileri Bölümü
Afyon Kocatepe Üniversitesi
odeper@aku.edu.tr

Özetçe

Bilgisayarlı destek sistemlere dayalı olarak medikal alanda gerçekleştirilen yaklaşımlar her geçen gün daha fazla ilgi görmektedir. Bu tür sistemlerde Yapay Zekâ teknikleri çeşitli hastalık teşhislerinde sıklıkla kullanılmaktadır. Diyabet de bunlardan birisidir. Bu çalışmada Destek Vektör Makineleri tabanlı bir diyabet teşhis sistemi önerilmiştir. DVM eğitimi sırasında, Gauss (RBF) çekirdek fonksiyonunun sigma parametresinin belirlenmesi için Girdap Optimizasyon Algoritması kullanılmış ve Pima Yerlileri'ne ait diyabet veri seti üzerinde sınıflandırma işlemi gerçekleştirilmiştir.
Anahtar Kelimeler: *diyabet tespiti, destek vektör makineleri, girdap optimizasyon algoritması, yapay zekâ*

Abstract

Approaches performed based on computer supported systems within the medical field gain more popularity day by day. In such systems, Artificial Intelligence techniques are often used for several disease diagnostics. Diabetes is one of these diseases. In this study, a diabetes diagnosis system based on Support Vector Machines has been proposed. Along training of SVM, Vortex Optimization Algorithm was used for determining the sigma parameter of the Gauss (RBF) kernel function, and a classification process has been done over the diabetes data set related to Pima Indians.
Keywords: *diabetes determination, support vector machines, vortex optimization algorithm, artificial intelligence*

1. Giriş

Bilgisayar teknolojisi, günlük hayatımızı doğrudan veya dolaylı yönden etkileyen bütün faaliyet alanlarının gelişmesinde ve ilerlemesinde oldukça büyük bir pay sahibidir. Günümüz koşullarını dikkate aldığımızda, bilgisayar sistemlerinin uygulanmadığı neredeyse hiçbir faaliyet alanı yok gibidir. Diğer yandan, bilgisayar teknolojisinin getirdiği avantajlar, ilgili alanların gelişimiyle birlikte yenilikçi fikirlerin ve uygulamaların ortaya çıkmasına, hatta –geri beslemeli bir şekilde– bilgisayar teknolojisinin de gelişmeye devam etmesine katkı sağlamaktadır. Kısacası, bilgisayar teknolojisinin etkinliği çok yönlü olmaktadır. Yine ifade etmek gerekir ki, teknolojik anlamdaki bu ilerlemelerde bilgisayar teknolojisini destekleyen diğer teknolojiler de (örneğin elektronik teknolojisi) aktif rol oynamaktadır.

Bilgisayar teknolojisinin farklı alanlarda kullanılması akla hemen bilgisayar destekli sistem kavramını getirmektedir. Bilgisayar destekli sistemler, bilgisayarların çözüm yollarında etkinliği, verimliliği ve başarımı artırma yönünde devrimsel nitelikte faktörler olarak dikkat çekmektedir. Bu noktada, bilgisayar destekli sistemlerin birçok alanda yaygın kullanım imkânı bulduğunu ifade etmek mümkündür.

Bilgisayar destekli sistemlerin yaygın kullanım imkânı bulduğu söz konusu alanlardan birisi de medikaldir. Buna göre, bilgisayarlı destek sistemlere dayalı olarak medikal alanda gerçekleştirilen yaklaşımlar, elde edilen yüksek başarı oranları sonucunda her geçen gün daha fazla ilgi görmektedir. Bu ilginin sebeplerinden birisi de kuşkusuz ki bilgisayar destekli sistemlerde farklı bilim ve araştırma alanlarından işe koşulan çeşitli unsurların da yer almasıdır. Yapay Zekâ araştırma alanı bu alan içerisinde inceleyebileceğimiz yaklaşımlar, yöntemler ve teknikler bu unsurlardan en dikkat çekici olanıdır. Günümüz uygulamaları incelendiğinde, söz konusu bilgisayar destekli medikal sistemlerinde Yapay Zekâ tekniklerinin özellikle çeşitli hastalık teşhislerinde sıklıkla kullanıldığı görülebilmektedir [1-9].

978-1-4673-7765-2/15/$31.00 ©2015 IEEE

İfade edilen açıklamalar bağlamında bu çalışmanın amacı, Destek Vektör Makineleri tabanlı bir diyabet teşhis sistemi önermektir. Önerilen sistemde esasında iki Yapay Zekâ tekniğinin kullanıldığı bir hibrit yaklaşım oluşturulmuştur. Bu yaklaşım göre, Destek Vektör Makineleri'nin eğitimi sırasında, Gauss (RBF) çekirdek fonksiyonunun parametresi olan, sigma (σ) parametresinin belirlenmesi için Girdap Optimizasyon Algoritması kullanılmaktadır. Önerilen sistem Pima Yerlileri'ne ait diyabet veri seti üzerinde uygulanmıştır.

2. Materyal ve Yöntem

Önerilen sistem ve uygulanan yaklaşım konusunda daha iyi bilgi sahibi olmak adına, öncelikli olarak, seçilen tekniklere ve çalışmaya konu olan diğer faktörlere değinmek gerekmektedir.

2.1. Diyabet Tespitine Konu Olan Veri Seti

Destek Çalışmada Kaliforniya Üniversitesi Irvine Makine öğrenmesi Veri Havuzu Web sitesi üzerinden Pima Yerlileri diyabet veri seti ile gerçekleştirilmiştir [10]. Bu veritabanındaki veriler Arizona civarında yaşayan, en küçüğü 21 yaş olmak üzere Pima Kızılderilileri'ne ait kadınlardan elde edilmiştir. Toplamda 768 örnek ve 8 nitelik mevcuttur. Çıkış olarak ise sağlıklı (diyabet negatif) ve diyabetli (diyabet pozitif) olmak üzere 2 sonuç mevcuttur. Veri setindeki veri sırasına ait nitelikler Tablo 1'de verilmiştir.

Tablo 1: Pima Yerlileri diyabet veri setinde nitelikler.

Nitelik No.	Niteliğin Açıklaması
1	Hamile Kalma Sayısı
2	Plazma Glukoz Konsantrasyonu
3	Diastolik Kan Basıncı
4	Kol Kası Cilt Kıvrım Kalınlığı
5	2-h serum insülin
6	Vücut Kütle İndeksi
7	Diyabet Soyağacı Fonksiyonu
8	Yaş

İlerleyen alt-bölümler çalışmada kullanılan tekniklerden ve uygulanan yaklaşımdan kısaca bahsedilmiştir.

2.2. Destek Vektör Makineleri

Destek Vektör Makineleri (DVM) Vapnik tarafından 1979 yılında tasarlanmıştır. 1995 yılında yine Vapnik tarafından sınıflandırma ve regresyon için önerilmiştir [11, 12]. DVM, sınıflandırma ve doğrusal olmayan fonksiyon yaklaşımı için önerilen bir öğrenme algoritmasıdır. DVM özellikle yazı tanıma, nesne tanıma, ses tanıma, yüz tanıma gibi örüntü tanıma uygulamalarında sıklıkla kullanılmaktadır [13].

DVM temelde, Lagrange çarpanları denkleminin formasyonuna dayanan [14] ve veri noktalarını mümkün olduğu kadar iyi sınıflandıran ve yine mümkün olduğu kadar iki sınıf noktaya ayıran optimum ayırıcı düzlemin bulunmasına dayalı bir Yapay Zekâ tekniğidir. Yani DVM'de amaç iki sınıf arasındaki uzaklığın maksimum olduğu durumun bulunmasıdır. Bu mantığının temel taşları ise eğitim seti içerisinden seçilen destek vektörler adı verilen ve her iki sınıfın uç noktalarında bulunan örneklerdir [15].

DVM'ni, doğrusal DVM ve doğrusal olmayan DVM olmak üzere iki sınıfa ayırmak mümkündür. Doğrusal DVM sadece ayırt edilebilir doğrusal verilere uygulanabilen en basit DVM modelidir. $(x_1,\ldots\ldots,x_n)$ veri seti, $y_i \in (-1,+1)$ ise sınıf etiketleri olmak üzere ve w normal vektörü, b ise eşik değerini göstermek üzere, söz konusu veri, $g(x) = w^T x + b = 0$ hiper düzlemi ile ayrılmaktadır. Dolayısıyla bu durum Eşitlik 1 ve Eşitlik 2'deki formüller ile açıklanabilmektedir [16]:

$$w^T x_t + b \geq +1, \quad \text{eğer } y_t = +1 \text{ (sınıf 2)} \tag{1}$$

$$w^T x_t + b \leq -1, \quad \text{eğer } y_t = -1 \text{ (sınıf 1)} \tag{2}$$

$g(x) = w^T x + b = 0$ hiper düzleminin üst tarafında kalan noktalar Eşitlik 1, alt tarafında kalan noktalar ise Eşitlik 2 ile gösterilmektedir [17].

Pratik uygulamaların çoğunda iki sınıfa ait veriler doğrusal olarak ayırt edilemez ve dolayısıyla verinin daha yüksek boyutlu bir uzayda haritalanması ile sonuca ulaşılabilir. Doğrusal olmayan destek vektörlerindeki ana fikir, orijinal giriş uzayının, eğitim verilerinin ayrılabilir olduğu daha yüksek boyutlu bir özellik uzayında haritalandırılabileceğidir [18]. Böyle durumlarda DVM'nde çekirdek fonksiyonları devreye girer ve n boyutlu bir veri kümesi, $m > n$ olmak üzere m boyutlu yeni bir veri kümesine dönüştürülerek, yüksek boyutta doğrusal sınıflandırma yapılır.

Çekirdek fonksiyonları, DVM algoritmasında önemli bir konumdadır. Doğru çekirdek fonksiyonun seçimi sınıflandırma başarımını önemli ölçüde etkilemektedir. DVM'de kullanılan çekirdek fonksiyonlarından bazıları Eşitlik 3, Eşitlik 4 ve Eşitlik 5'de verilmiştir [13].

Doğrusal: $K(x_i, x_j) = x_i^T x_j$ (3)

Polinom: $K(x_i, x_j) = (1 + x_i^T x_j)^p$ (4)

Gauss (RBF): $K(x_i, x_j) = e^{\frac{\|x_i - x_j\|^2}{2\sigma^2}}$ (5)

2.3. Girdap Optimizasyon Algoritması

Köse ve Arslan tarafından geliştirilmiş olan Girdap Optimizasyon Algoritması (GOA), doğal girdaplardan esinlenilerek geliştirilmiş bir Yapay Zekâ tabanlı optimizasyon tekniğidir [19]. Algoritma, çözüm uzayına ilk aşamada rastgele bırakılan yapay işlem parçacıklarının optimum çözümü aramasına dayanmaktadır. Girdap değerine (v) sahip N parçacık, verilen bir matematiksel fonksiyon için belirli ön değerler ve ek parametreler eşliğinde optimum çözüme ulaşmaya çalışmaktadır. İteratif ilerleyen çözüm adımlarında, her parçacığın sahip olduğu yeni çözüm değerleri (fitness) değerlendirilmekte ve v değeri ortalamaya göre daha iyi konumda olan parçacıklar girdap statüsüne terfi etmektedir. Yine her iterasyon sonunda girdap olmayan (normal statüde) parçacıkların sayısı eleme katsayısından (e) küçük veya eşit ise, bu parçacıklar yok olmakta ve yerlerine aynı sayıda yeni parçacıklar eklenmektedir. Bu süreçler sırasında parçacıkların v değerleri, pozisyonları gibi değerler güncellenmektedir. Çözüm arama süreci, sonlanma kriteri (standart olarak iterasyon sayısı) sağlanıncaya kadar devam etmektedir. Basit matematiksel eşitliklerle kurulan ve doğa temelli olduğu kadar evrimsel özellikler de içeren algoritmanın adımları şu şekildedir:

Adım 1. Parçacık sayısını (N) belirle; bu N parçacığı çözüm uzayına rastgele dağıt. Her bir parçacığa başlangıç

girdap değerini (vorticity: *v*) ata. Eleme katsayısı (*e*) ve maksimum vorticity (*max_v*) değerlerini belirle. Ayrıca minimum vorticity (*min_v*) değerini maksimum vorticity (*max_v*) değerinin negatif işaretlisi olarak kabul et.

Adım 2. Başlangıç pozisyonları için çözüm (fitness) hesapla; en iyi fitness sahip olanı bul ve bu parçacığın *v* değer(ler)ini aşağıdaki eşitliğe göre güncelle:

*baş_en_iyi_p_v_(yeni) = baş_en_iyi_p_v_(geçerli) + (r_sayı * baş_en_iyi_p_v_(geçerli))* (6)

Bu parçacığı aynı zamanda girdap statüsüne yükselt; global en iyi parçacık (g_ei_p) say ve değerlerini de global en iyi (g_e_i) değerler olarak kabul et.

Adım 3. Aşağıdaki adımları durma kriterine kadar (örn. iterasyon sayısı) tekrarla:

Adım 3.1. O andaki ortalama fitness değerinden küçük veya eşit fitness değerine sahip olan parçacıkları girdap; diğerlerini ise normal statüde kabul et.

Adım 3.2. Bütün parçacıklar için *v* değerlerini güncelle:

*p_v_(yeni) = p_v_(geçerli) + (r_sayı * (g_ei_p_v / p_v_(geçerli)))* (7)

Adım 3.3. global en iyi parçacık dışında kalan ve girdap statüsünde olan bütün parçacıkların *v* değerlerini güncelle:

*p_v_(yeni) = r_sayı * p_v_(geçerli)* (8)

Adım 3.4. global en iyi parçacık dışındaki bütün parçacıkların pozisyonlarını güncelle:

*p_poz_(yeni) = p_poz_(geçerli) + (r_sayı * (p_v_(geçerli) * (g_ei_p_poz – p_p_(geçerli))))* (9)

Bir parçacığın fitness değeri, global en iyi değerden iyi ise, o parçacığı global en iyi parçacık ve değerlerini de global en iyi değerler kabul et; parçacığı girdap statüsüne yükselt.

Adım 3.5. Eğer girdap olmayan parçacıkların sayısı, *e* değerinden küçük veya eşitse, girdap olmayan parçacıkları yok et; yerlerine (aynı sayıda) yeni parçacıklar dağıt; büyük problemlerde sistem içi optimizasyon yap; ardından, durma kriteri henüz sağlanmıyorsa döngü başına (Adım 3.1.) dön.

Adım 4. Optimum çözüm, global en iyi parçacığın değerleridir (g_e_i).

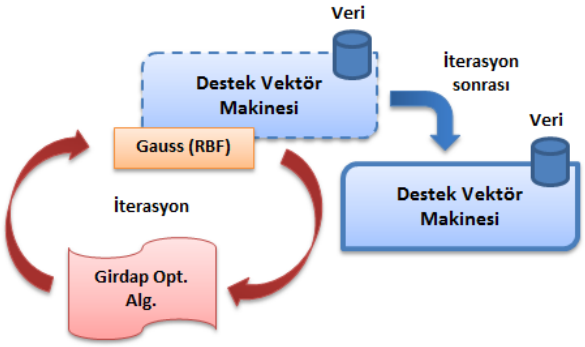
2.4. Diyabet Tespitinde GOA- DVM Yaklaşımı

Bu çalışma kapsamında oluşturulan, GOA-DVM yaklaşımına dair temel çözüm süreçlerini kısaca şu şekilde açıklayabiliriz:

- GOA'da çözüm uzayına dağıtılan her bir parçacık, DVM'deki Gauss (RBF) çekirdek fonksiyonunda kullanılacak sigma (σ) değerini temsil etmektedir.
- GOA belirlenmiş olan iterasyon sayısı kadar çalıştırılmaktadır. Her yeni iterasyonda, her bir parçacığın değerleri DVM'de kullanılmakta ve DVM'nin eğitimi sonrası parçacıkların sınıflandırma doğruluk oran değerlerine (Eşitlik 10) göre, hangi parçacığın o iterasyonda (local) ve genel süreçte (global) en iyi olduğu tespit edilmektedir.

$$\frac{DP+DN}{DP+YP+DN+YN}*100 \quad (10)$$

Eşitlik 10'da, DP: doğru sınıflandırılan diyabet pozitif bireyleri, DN: doğru sınıflandırılan diyabet negatif bireyleri, YP: yanlış sınıflandırılan diyabet pozitif bireyleri ve YN: yanlış sınıflandırılan diyabet negatif bireyleri temsil etmektedir.



Şekil 1: Diyabet tespitinde GOA- DVM yaklaşımı.

- İyi (optimum) parçacıkların tespiti sonrası varsayılan GOA adımları yerine getirilmektedir.
- Iterasyon sonunda, optimum parçacık değeri [sigma (σ) değeri] kullanılarak, DVM'nin sınıflandırma öncesi optimum Gauss (RBF) çekirdek fonksiyonu parametreleri ile eğitilmesi sağlanmış olur.

3. Uygulama ve Bulgular

Geliştirilen GOA-DVM sistemi, ifade edilen Pima Yerlileri diyabet veri seti üzerinde, farklı GOA parçacık sayısı ve iterasyon sayıları ile uygulanmıştır. Söz konusu veri setindeki toplam 768 örneğin 500 tanesi diyabet negatif bireylere, 268 tanesi ise diyabet pozitif bireylere karşılık gelmektedir. Uygulama sürecinde bu örneklerin yarısı eğitim sürecine, yarısı ise test sürecine dâhil edilmiş ve test süreciyle elde edilen doğruluk oranları sistemin etkinliği konusunda dikkate alınmıştır. GOA'da giriş *v* değeri 0.50, max_v değeri 7.0, (min_v değeri -7.0) ve *e* değeri 45 olarak kullanılmıştır.

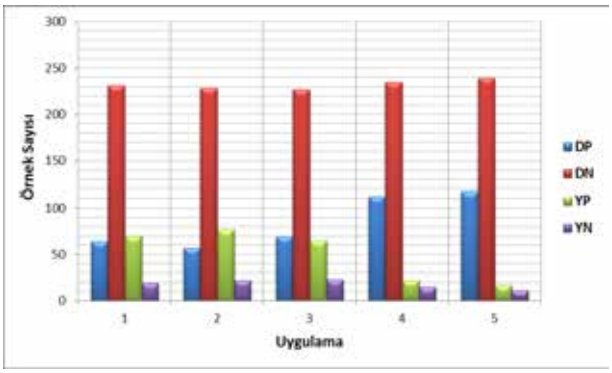
Beş farklı parçacık ve iterasyon sayıları ile gerçekleştirilen uygulama süreçlerine ilişkin elde edilen bulgular Tablo 2'de sunulmuştur.

Tablo 2: Uygulama süreçlerine ilişkin elde edilen bulgular.

N	İterasyon	sigma (σ)	Doğruluk (%)
25	2000	8.6144	76,82
40	2500	4.4587	74,22
50	4000	0.9028	77,08
75	5000	0.7017	90,36
90	5000	0.6618	92,71

Tablo 2'de gerçekleştirilen beş farklı uygulama sonucunda sigma (σ) için bulunan optimum değerler ve bunlara karşılık doğruluk oranları gösterilmiştir. Elde edilen bulgulara göre, oluşturulan sistemin yeter düzeyde bir sınıflandırma ortaya koyabilmektedir.

Uygulamalarda elde edilen DP (doğru sınıflandırılan diyabet pozitif birey), DN (doğru sınıflandırılan diyabet negatif birey), YP (yanlış sınıflandırılan diyabet pozitif birey) ve YN (yanlış sınıflandırılan diyabet negatif birey) sayılarına ilişkin bir grafik Şekil 2'de gösterilmiştir.



Şekil 2*: Uygulamalarda elde edilen DP, DN, YP ve YN.*

4. Sonuçlar ve Gelecek Çalışmalar

Bu çalışmada, GOA tabanlı bir DVM ile oluşturulan, hibrit bir Yapay Zekâ sistemi kullanılarak diyabet tespiti gerçekleştirilmiştir. Kurulan sistem yapısına göre, DVM'nin eğitimi esnasında, Gauss (RBF) çekirdek fonksiyonun parametresi olan, sigma (σ) parametresinin belirlenmesi için GOA işe koşulmuştur. Bu çerçevede kurulan yaklaşımın etkinliği Pima Yerlileri diyabet veri seti üzerinde yapılan sınıflandırma süreci ile değerlendirilmiştir. Süreç ile elde edilen bulgular, önerilen GOA-DVM sisteminin diyabet tespitinde yeter düzeyde etkinliğe sahip olduğunu göstermiştir. Önerilen sistem, ilgili literatürde yer alan yaklaşımlara yeni bir alternatif sunmakta ve yine Yapay Zekâ tabanlı hastalık teşhisi çalışmalarına katkı sağlamaktadır.

Bu metin ile rapor edilenler dışında, yazarlar bazı yeni çalışmaları da gerçekleştirmeyi düşünmektedir. Buna göre, çalışmada yerine getirilen parametre optimizasyon ve diyabet tespit süreçlerinin verimliliğini artırıcı çeşitli deneysel çalışmalarla birlikte, aynı sistemin farklı hastalıkların teşhisi yönünde düzenlenip uygulanması doğrultusundaki yaklaşımlar gelecek süreçler için planlanmaktadır.

5. Kaynakça

[1] Szolovits, P., Patil, R. S., & Schwartz, W. B. "Artificial intelligence in medical diagnosis." *Annals of Internal Medicine*, *108*(1), 80-87, 1988.

[2] Kononenko, I. "Machine learning for medical diagnosis: history, state of the art and perspective." *Artificial Intelligence in Medicine*, *23*(1), 89-109, 2001.

[3] Abbass, H. A. "An evolutionary artificial neural networks approach for breast cancer diagnosis." *Artificial Intelligence in Medicine*, *25*(3), 265-281, 2002.

[4] Al-Shayea, Q. K. "Artificial neural networks in medical diagnosis." *Int Journal of Computer Science Issues*, *8*(2), 150-154, 2011.

[5] Dhar, J., ve Ranganathan, A. "Machine learning capabilities in medical diagnosis applications: Computational results for hepatitis disease." *Int Journal of Biomedical Eng and Tech*, *17*(4), 330-340, 2015.

[6] Botero-Rosas, D., Leon-A, J., Reina, J. M., Obando, A., ve Bastidas, A. R. "Use of artificial ıntelligence in the diagnosis of chronic obstructive pulmonary disease (COPD)". *Am J Respir Crit Care Med*, *191*, 2015.

[7] Thong, N. T. "HIFCF: An effective hybrid model between picture fuzzy clustering and intuitionistic fuzzy recommender systems for medical diagnosis." *Expert Systems with Applications*, *42*(7), 3682-3701, 2015.

[8] Gullapalli, V. K., Brungi, R., ve Gopichand, G. "Application of perceptron networks in recommending medical diagnosis." *Int Journal of Computer Applications*, *113*(4), 2015.

[9] Muralidaran, C., Dey, P., Nijhawan, R., ve Kakkar, N. "Artificial neural network in diagnosis of urothelial cell carcinoma in urine cytology." *Diagnostic Cytopathology*, *43*(6), 443-449, 2015.

[10] Frank, A. ve Asuncion, A., "UCI Machine Learning Repository" [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2010.

[11] Çomak, E., Arslan, A., Türkoğlu, İ., "A decision support system based on support machines for diagnosis of the heart valve diseases", *Computers in Biology and Medicine*, *37*, 21-27, 2007.

[12] Vapnik, V., Golowich, S., Smola, A., "Support vector method for function approximation, regression estimation, and signal processing." *Advances in Neural Information Processing Systems 9*, 281–287, 1997.

[13] Özkaya, A. U., Kaya, M. E., Gürgen, F., "Destek vektör makineleri kullanılarak aritmi sınıflandırılması", *XI. Biyomedikal Mühendisliği Ulusal Toplantısı*, 12-16, 2005.

[14] Tamura, H., Tanno, K., "Midpoint validation method for support vector machines with margin adjustment technique", *Int Journal of Innovative Computing, Information and Control*, 5, 4025-4032, 2009.

[15] Çomak, E., "Destek Vektör Makineleri Çoklu Sınıf Problemleri için Çözüm Önerileri", Yüksek Lisans Tezi, *Selçuk Üniversitesi Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği ABD.*, 2004.

[16] Karaç, E. I., "Model selection for multi-class support vector machines", Yüksek Lisans Tezi, *Boğaziçi Üniversitesi Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği ABD.*, 2005.

[17] Özkan, Y., *Veri madenciliği yöntemleri*, Papatya Yayıncılık, İstanbul, Türkiye, 1. Basım, 2008.

[18] Özkaya, A. U., "Intelligent Arrhythmia Classification Based on Support Vector Machines", Yüksek Lisans Tezi, *İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü*, Bilgisayar Mühendisliği Bölümü, 2003.

[19] Köse, U., ve Arslan, A. "On the idea of a new artificial intelligence based optimization algorithm inspired from the nature of vortex." *Broad Research in Artificial Intelligence and Neuroscience*, *5*(1-4), 60-66, 2015.