Dumlupınar Üniversitesi
Fen Bilimleri Enstitüsü Dergisi
E-ISSN: 2651-2769
Sayı 43, Aralık 2019

Journal of Science and Technology
of Dumlupınar University
E-ISSN: 2651-2769
Number 43, December 2019

ARAŞTIRMA MAKALESİ (Research Article)

ROTASYON ORMAN SINIFLANDIRMA ALGORİTMASI KULLANARAK KRONİK BÖBREK RAHATSIZLIĞININ TAHMİNİ

Serhat KILIÇARSLAN[1*] , Mete ÇELİK[2]

[1]Tokat Gaziosmanpaşa Üniversitesi, Enformatik Bölümü,Tokat, serhat.kilicarslan@gop.edu.tr, ORCID: 0000-0001-9483-4425
[2] Erciyes Üniversitesi, Bilgisayar Mühendisliği Bölümü, Kayseri, mcelik@erciyes.edu.tr, ORCID: 0000-0002-1488-1502

Geliş Tarihi:24.09.2018 *Kabul Tarihi: 14.02.2019*

ÖZ

Kronik böbrek rahatsızlığı (KBR) son günlerde artarak insanların yaşamını olumsuz etkileyen ve böbreklere zarar vererek normal görevlerini uzun süre yapmalarını engelleyen bir rahatsızlıktır. KBR'nin erken tanı ve tedavisi yapılmaz ise yüksek tansiyon, kalp rahatsızlığı, şeker rahatsızlığı, böbrek yetmezliği gibi hastalıkları da tetikleyebilmekte ve rahatsızlığa bağlı ölümler artabilmektedir. Bu nedenle kronik böbrek rahatsızlığının teşhis ve tahmininin erken yapılması önemlidir. Literatürde KBR tahmini için sezgisel ve sezgisel olmayan veri madenciliği teknikleri uygulanmıştır. Bu çalışmada KBR'nin tahmini için sezgisel olmayan kolektif veri madenciliği yöntemlerinden olan rotasyon orman algoritmasınının kullanılması önerilmiştir. Deneysel sonuçlar önerilen yaklaşımın, kronik böbrek rahatsızlığını tahmin etmede, diğer algoritmalarından daha iyi performans sergilediğini göstermiştir.

Anahtar kelimeler: *Kronik Böbrek Rahatsızlığı, Rotasyon Orman Algoritması, Veri Madenciliği.*

PREDICTION OF CHRONIC KIDNEY DISEASE USING ROTATION FOREST CLASSIFICATION ALGORITHM

ABSTRACT

Chronic kidney disease (CBR) has increased in recent years by affecting the lives of people adversely. It affects kidneys and prevents them doing their normal duties properly. Without early diagnosis and treatment of CBR, it can trigger diseases such as high blood pressure, heart disease, diabetes mellitus and kidney failure and it can even cause deaths. For this reason, it is important to diagnose and predict CBR early. In the literature, various heuristic and non-heuristic data mining classification techniques have been applied on predicting CBR. In this study, it is proposed to use the rotation forest algorithm as a non-heuristic collective data mining method for predicting CBR. Experimental evaluations show that the proposed approach performs better than other algorithms on predicting CBR.

Keywords: *Chronic Kidney Disease, Rotation Forest Algorithm, Data Mining.*

1. GİRİŞ

Kronik böbrek rahatsızlığı (KBR) son zamanlarda yaşamı olumsuz etkileyen ve böbreklere zarar veren bir rahatsızlıktır. Bu rahatsızlığın teşhisi böbrek fonksiyonlarının azalmasına bağlı olarak idrar yoğunluğunun artmasıyla konulabilmektedir. KBR'nin tedavisi tansiyon yüksekliği, kemik rahatsızlığı, anemi, kan damar rahatsızlıkları, böbrek yetmezliği gibi hastalıkları da tetikleyebilmekte ve rahatsızlığa bağlı ölümler artabilmektedir [46]. Bilgisayar destekli karar sistemi uzman doktorlara yardımcı olma açısından önemi her geçen gün artmaktadır [45-46]. Bu çalışmada veri madenciliği yöntemleriyle KBR'yi tespit etme ve bu sayede uzman doktorlara uygulayacakları tedavi öncesinde tam teşhis koymalarına destek verici yapısı geliştirilmek istenmektedir. Doğru teşhis sağlanması, önlenebilir hastalıklar için zamanında tarama yapılması veya hatalı ilaç etkileşimlerinin önlenmesi gibi klinik gereksinimlere hitap etmesi tıbbi karar destek sistemlerinin en önemli faydalarındandır.

KBR son yıllarda artarak insanın yaşamını olumsuz olarak etkileyen bir rahatsızlıktır [1]. Nisan 2017'de Mexico City'deki Uluslararası Nefroloji Derneği (ISN) Dünya Kongresi 2005-2015 yılları arasında KBR'nin sebep olduğu ölüm sayısı %32 artarak yaklaşık 1 milyon olarak açıklanmıştır [2,29]. Ülkemizde 2012 yılı sonu verilerine göre diyaliz uygulanan veya böbrek nakli yapılmış yaklaşık 62.000 hasta bulunmakta ve devlete olan maliyetinin 2,5-3 milyar doları bulacağını öngörülmektedir [3,28]. Ayrıca, hastalığın ilerlemesiyle birlikte, hastaların bir sağlık kuruluşuna bağlı olarak yaşaması sosyal problemlere sebep olabilmektedir ve hastalığın ilerlemesi diğer organları etkileyebilmekte ve dolayısı ile tıbbi problemlere de sebebiyet vermektedir [3].

Bilişim teknolojilerinin gelişimi ile birlikte elde edilen çok parametreli veriler veri madenciliği teknikleriyle işlenerek anlamlı bilgiler keşfedilmekte ve karar-destek sistemlerinde kullanılabilmektedir [47,48]. Bu teknikler hastalığı belirlemek, ileriye dönük hastalığı veya hastalık durumlarını tahmin etmek amacıyla kullanarak tanı ve tedavi sürecine yardımcı olabilmektedir. Sürekli gelişen teknoloji ile birlikte yeni ve gelişmiş algoritmalar ortaya çıkmakta ve bu algoritmalar tıp gibi çeşitli uygulama alanlarında kullanılmaktadır [48].

Literatürdeki KBR verilerinin sınıflandırılması ile ilgili teknikleri sezgisel ve sezgisel olmayan teknikler olarak ikiye ayırabiliriz. KBR verilerinin sınıflandırılması için kullanılan sezgisel teknikler yapay sinir ağları [6], naïve bayes [10], radial tabanlı sinir ağları [5], destek vektör makinesi [15] olarak ve sezgisel olmayan teknikler de rastgele orman [12], karar ağacı [14], kstar [14], k-NN [16] algoritmaları olarak listelenebilir.

Bu çalışmada kolektif sınıflama algoritmalarından rotasyon orman algoritmasını [20] KBR verilerini sınıflandırmak için kullanmayı öneriyoruz. Önerilen model sınıflandırma performansını arttıran yeni nesil kolektif sınıflandırma algoritmalarından biridir. Diğer kolektif algoritmalardan farklı olarak, rotasyon orman algoritması temel bileşen analizi (TBA) uygulayarak ayırt ediciliği yüksek olan özelliklerin seçilmesiyle başarılı bir sınıflandırma işlemi gerçekleştirmektedir.

Çalışma, doktorların karar verme sürecini basitleştirmeyi ve önerilen modelle kronik böbrek hastalığı teşhisi süresini kısaltmayı amaçlamaktadır.

Bu çalışmanın ikinci bölümünde literatür taraması sunulmuş, üçüncü bölümünde KBR verilerinin sınıflandırılması için önerilen modelin detayları tartışılmış, dördüncü bölümde KBR veri kümesinin detayları verilmiş ve deneysel sonuçlar sunulmuş ve son bölümde ise elde edilen sonuçlar tartışılmıştır.

2. LİTERATÜR TARAMASI

Literatürde kronik böbrek rahatsızlığı verilerinin sınıflandırılması için yapılan çalışmalarda kullanılan teknikleri, sezgisel ve sezgisel olmayan teknikler olarak ikiye ayırmak mümkündür. KBR verilerinin sınıflandırılması için kullanılan sezgisel olmayan sınıflandırma algoritmaları ise istatistiksel tabanlı ve kolektif sınıflandırma algoritmaları olarak ikiye ayrılabilir. Bu çalışmada sezgisel olmayan kolektif sınıflandırma algoritması olan rotasyon orman algoritması KBR verilerinin sınıflandırılması için önerilmiştir. Literatürdeki çalışmaların çoğunda UCI ML veri deposundan alınan kronik böbrek hastalığı verisi kullanılmıştır [4]. Bu çalışmada da aynı veri kümesi kullanılmıştır. UCI ML kronik böbrek hastalığı verisi 400 hastanın laboratuvar sonuçlarından elde edilen 24 niteliğe sahip veri bulunmaktadır ve veri kümesinde hastanın böbrek hastası olup olmadığını gösterir iki sınıf vardır. Veri ile ilgili detaylar Bölüm 4.1'de verilmiştir.

Sezgisel sınıflandırma algoritmaları arasında yapay sinir ağları (YSA), radial tabanlı sinir ağı (RBF) gibi sınıflama algoritmaları sayılabilir [5-9]. İlkuçar'ın yapmış olduğu çalışmada UCI ML kronik böbrek hastalığı veri kümesi üzerinde YSA ve RBF sınıflandırma algoritmaları kullanılarak kronik böbrek hastalarının tahmini yapılmış ve RBF'nin tahmin performansının daha iyi olduğunu ortaya konulmuştur [5]. Jena ve Kamila aynı veri kümesi üzerinde [4] naïve bayes (NB), çok katmanlı sinir ağı (MLP), destek vektör makineleri (SVM), karar ağaçları ve conjunctive rule sınıflama algoritmalarını kullanılmış olup bunlar arasından sezgisel algoritma olan MLP'nin daha iyi sonuç verdiğini ortaya koymuşlardır [6]. Rubini ve Eswaran aynı veri kümesi üzerinde [4] RBF, MLP ve lojistik regrasyon algoritmalarını kullanmışlardır ve sezgisel algoritma olan MLP'nin daha iyi performans sergilediğini ortaya koymuşlardır [7]. Yıldırım'ın yapmış olduğu çalışmada kronik böbrek hastalığı veri kümesi üzerinde [4] sezgisel ve sezgisel olmayan algoritmalar kullanılarak kronik böbrek rahatsızlığı tahmin edilmeye çalışılmış ve en başarılı sonucu MLP ile elde edilmiştir [8]. Ramya ve Radha'nın yapmış oldukları çalışmada 1000 veri ve 15 farklı öznitelikli veri kümesi üzerinde MLP, RBF ve rastgele orman sınıflama algoritmaları kronik böbrek hastalığının teşhisinde kullanılmış ve çalışmada en iyi başarıyı RBF sınıflandırma algoritmasının verdiği ortaya konulmuştur [9]. Ancak literatürde kullanılan sezgisel yöntemler her zaman optimal sonucu verememektedirler ve bu nedenle bu yöntemlerin sağlık verilerinde ve sektöründe kullanımları kısıtlıdır.

Literatürde, KBR tahmin probleminde kullanılan sezgisel olmayan algoritmalar arasında NB ve SVM sayılabilir. Kunwar ve arkadaşlarının yapmış oldukları çalışmada 180 hastanın 25 öznitelikli veri kümesinde %60'lık kısmı hastaları %30'luk kısmı hasta olmayanları verilerinin analizi için NB ve YSA sınıflama algoritmalarını kullanmış olup NB algoritması ile iyi bir başarı elde edilmiştir [10]. Vijayarani ve Dhayanand'ın yapmış olduğu çalışmada sistematik böbrek foksiyonu testinden elde edilen 584 veri, 6 öznitelik veri kümesi üzerine SVM ve NB algoritmalarını kullanarak tahmin işlemi gerçekleştirilmiş ve SVM ile daha yüksek başarı elde edildiği ortaya konulmuştur [11]. Kumar'ın gerçekleştirmiş olduğu çalışmada UCI ML kronik böbrek hastalığı veri kümesi [4] üzerinde rastgele orman (RO), SVM, NB, RBF, MLP ve basit lojistik regrasyon (SLG) sınıflama algoritmaları kullanılmıştır ve rastgele orman algoritmasının daha iyi sonuç verdiği ortaya konulmuştur [12]. Balakrishna ve arkadaşlarının gerçekleştirdiği çalışmada UCI ML kronik böbre hastalığı veri kümesi [4] kullanılarak SVM, NB, hata budama azaltma (REPTree) ve RO algoritmalarının performansları kronik böbrek teşhisi üzerinde incelenmiş ve en iyi başarı sonucunu RO ile elde edilmiştir [13]. Baby ve Vital'in yapmış oldukları çalışmada Visakhapatnam bölgesinde 2014 ve 2015 yılları arasında toplanan 690 veri, 49 öznitelikli veri kümesine alternatif karar ağacı, C4.5, KStar, naïve bayes ve RO algoritmaları böbrek hastalığı tahmininde kullanılmış ve en iyi performansı KStar ve RO sınıflandırma algoritmalarında elde edilmiştir [14]. Polat ve arkadaşlarının yapmış oldukları çalışmada

UCI ML kronik böbrek hastalığı veri kümesi [4] kullanılarak tahmin yapılmıştır ve çalışmada öznitelik indirgeme filtreleri ve SVM algoritması kullanarak %98.5 başarı elde edilmiştir [15]. Sinha'nın yapmış olduğu çalışmada kronik böbrek rahatsızlığı verisi [4] üzerinde k-en yakın komşu (k-NN) ve SVM sınıflandırma algoritmalarının performansları incelenmiş ve en iyi sonucu KNN algoritmasının verdiği ortaya konulmuştur [16].

Sezgisel olmayan kollektif sınıflandırma algoritmaları da kullanılara KBR verileri analiz edilmiştir. Eroğlu ve Palabaş yapmış oldukları çalışmada UCI ML kronik böbrek hastalığı veri kümesinin [4] sınıflandırılması için NB, KNN, SVM, C4.5, rastgele ağaç, karar tablosu ve çeşitli kolektif algoritmaları (adaboost, bagging ve rastgele alt uzaylar) kullanılmıştır. Çalışmada bagging kolektif algoritması C4.5 ve rastgele ağaç algoritmalarıyla yapılan sınıflandırma işleminin daha başarılı olduğu ortaya konulmuştur [17]. Cheng ve arkadaşlarının böbrek rahatsızlığını tahmin etmek üzerine yaptığı çalışmada Tayvan'daki bir büyük şehir hastanesindeki böbrek sağlığı merkezinden 2004 ve 2013 yılları arasında toplanan 2066 hastanın veri kümesi üzerinde sezgisel olmayan istatistiksel ve kolektif algoritmalar kullanılmıştır. En iyi başarının AdaBoost + CART algoritmasının birlikte kullanılmasıyla elde edildiği görülmüştür [18]. Başar ve arkadaşları UCI ML kronik böbrek hastalığı veri kümesi [4] üzerinde yapmış oldukları çalışmada en iyi başarı Adaboost+BFTree kollektif algoritmasıyla elde etmişlerdir [19]. Ancak kolektif algoritmalarla yapılan bu çalışmalarda, kullanılan veri kümelerinde herhangi bir ayırt ediciliği yüksek özniteliklerin seçilme işlemi gerçekleşmediğinden dolayı düşük başarılar elde edilmiştir.

Bu çalışmada, kronik böbrek hastalığının tahmini için, literatürdeki çalışmalardan farklı olarak, sezgisel olmayan kolektif algoritmalarından olan rotasyon orman kolektif sınıflandırma algoritmasının kullanılması önerilmektedir. Bu algoritmanın diğer algoritmalara göre üstünlüğü veri kümesi içerisinden ayırt ediciliği yüksek özniteliklerin seçilmesiyle daha başarılı sınıflandırma işlemi gerçekleştirmesidir.

3. MATERYAL VE METOT

Bu çalışmada KBR verilerinin sınıflandırılması için rotasyon orman ve rastgele ağaç yaklaşımlarının kombinasyonu olan bir yöntem önerilmiştir. Her iki yöntemin detayları aşağıda verilmiştir ve daha sonra önerilen modelin detayları verilmiştir.

3.1. Rotasyon Orman Sınıflandırma Algoritması

Rotasyon orman algoritması yeni nesil güçlü ağaç tabanlı kolektif algoritmalarından biridir. Rotasyon orman algoritması daha az sayıda ağaçla benzer veya daha iyi başarılar elde edebilmektedir. Rotasyon orman algoritması çalışma yapısı olarak rastgele orman algoritma yapısına benzemektedir ve bu algoritmada temel olarak bootstrap algoritma mantığı kullanılmaktadır [20]. Algoritmada eğitim kümesi rastgele olarak alt gruplara ayrılmakta ve her bir alt gruba temel bileşen analizi (principal component analysis-TBA) özellik çıkarım işlemi uygulanmaktadır. Daha sonra temel sınıflandırıcı olarak rastgele seçilen ağaç sonuçların geliştirilmesi için rotasyon orman topluluk algoritması ile değerlendirmeye tutulmaktadır. Rotasyon orman algoritmasının kaba kodu Algoritma 1'de verilmiştir.

Eğitim Aşaması

X	-	Eğitim Veri Kümesi
F	-	Özellik Kümesi
C	-	Sınıf Etiketleri
L	-	Sınıf Sayısı

K - Alt Küme Sayısı
RM - Rotasyon Matrisi
For i = 1:L
1. Rotasyon matrisi hazırlanır R_i
   a. *K* alt kümeler için *F* bölünür
   b. For *j*=1:*K*
      i. $X_{i,j}$ deki sınıfların rastgele bir alt kümesini ele
      ii. $X_{i,j}^{new}$ için bootstrap oranı ile $X_{i,j}$'nin bootstrap örneği ile üretilir.
      iii. $X_{i,j}^{new}$ ye TBA uygulandıktan sonra $C_{i,j}$ elde edilir.
   c. R_i'deki $C_{i,j}$ düzenlenir
   d. R_i'deki sütunları, *F*'deki özelliklerin sırasına uyacak şekilde yeniden düzenleyerek *RM* 'yi oluştur.
2. Eğitim kümesi olarak (*X, RM, Y)* kullanılacak

Sınıflandırma Aşaması

$$\mu_j(x) = \frac{1}{L}\sum_{i=1}^{L} P_{i,j}(xRM), \qquad j = 1,2,\dots,c$$

Algoritma 1. Rotasyon Orman Algoritması Kaba Kodu [20]

Algoritma iki aşamadan oluşmaktadır. Bunlar eğitim ve sınıflandırma aşamalarıdır. Başlangıçta *X* eğitim kümesi, *F* özellik kümesi, *C* sınıf etiketleri, *L* sınıf sayısı ve *K* alt küme sayısı verilmektedir. Daha sonra *F* özellik kümesi *K* alt küme sayısına bölünmektedir. Bölünen *K* alt küme sayısı kadar yeni eğitim veri kümesi bootstrap örneği ile eğitim ve test olacak şekilde veri kümesi üretilir. Üretilen veri kümesine TBA uygulanarak yeni kovaryans değeri hesaplanır. Oluşturulan kovaryans değerine göre yeni rotasyon matrisi oluşturulur. Son olarak da ortalama kombinasyon yöntemi kullanılarak her sınıfa ait güven oyu alınır ve sınıfı bilinmeyenlerden en yüksek güveni alanın sınıfına atama gerçekleştirilir. [20-22].

3.2. Rastgele Ağaç Sınıflandırma Algoritması

Çalışmada eğitim amacıyla verilmiş veri kümesine temel bileşen analizi uygulayarak özellik çıkarım işlemi uygulanmıştır ve daha sonra temel sınıflandırıcı olarak rastgele ağaç algoritması ile alınan sonuçların geliştirilmesi için rotasyon orman algoritması ile değerlendirme yapılmıştır.

Rastgele ağaç algoritması Brieman tarafından geliştirilen sınıflandırma algoritmasıdır [24]. Rastgele ağaç algoritması karar ağacı içerisinde regresyon problemlerinin çözmeye yardımcı olan sınıflandırma algoritmasıdır. Ayrıca oluşturulan ağaç içerisinde rastgele bir şekilde seçilen özellikler ile alt ağaç oluşturmayı sağlamaktadır. Algoritma içerisindeki ağaçlarda budama işlemi mevcut değildir. Rastgele ağaç algoritması, oluşturulan çok sayıdaki rastgele ağaç ile değerlendirmelerde yüksek doğruluk oranına sahip olabilmektedir [23,24]. Algoritmanın en önemli avantajlarından birisi karar ağaçlarındaki aşırı uyum problemini çözmüş olmasıdır. Rastgele ağaç yönteminde Denklem 1' deki şekilde ağaç tipi sınıflandırıcı kullanılır.

$$\{h(x, \theta_k)k = 1, \dots\} \tag{1}$$

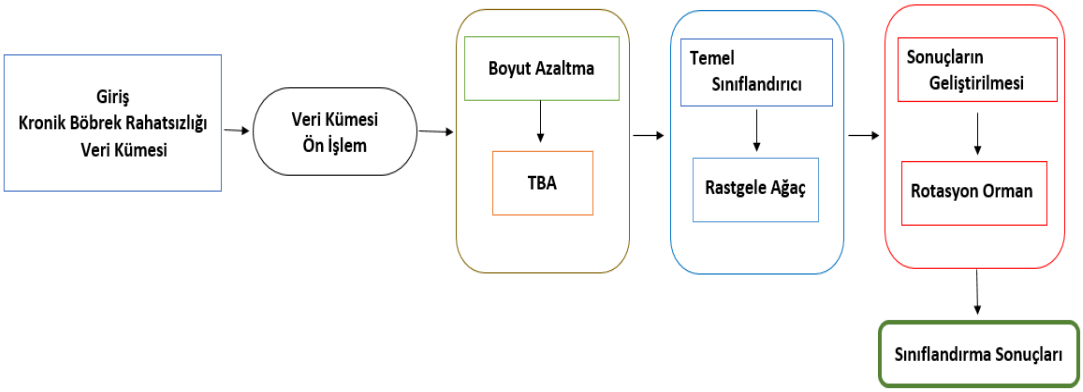
Denklem 1'de girdi verisi olarak *x* ve rastgele vektörü ise θ_k ile temsil etmektedir [24]. Rastgele ağaç algoritması mevcut dallar içinde en iyi dalı belirlemek için Denklem 2'de verilen GINI indeksini kullanır [44].

$$\sum\sum j \neq i(f(C_i,T)/|T|)(f(C_j,T)/|T|) \quad (2)$$

Denklem 2'de T eğitim veri kümesini, C_i verinin ait olduğu sınıfı, $f(C_i, T)/|T|$ seçilen verinin C_i sınıfına ait olma olasılığını gösterir [43-44]. Rastgele ağaç algoritması rastgele örneklemeyi ve topluluk algoritmaları tekniklerinin iyileştirilmiş bir yapıyı içermesi nedeniyle diğer sınıflandırma algoritmalarına göre daha iyi genelleme ve doğru tahminlemede bulunur [42].

3.3. Önerilen Sınıflandırma Yöntemi

Bu çalışmada KBR verilerinin sınıflandırılması için rotasyon orman ve rastgele ağaç yaklaşımlarının kombinasyonu olan bir yöntem önerilmiştir. Deneyler, çalışmada kullanılan veri kümesine ön işlem uygulanarak, veri kümesi içindeki gereksiz veriler temizlenip kullanıma hazır hale getirilmiştir. Ön işlemin ardından veri kümesine TBA yöntemi uygulanarak veri kümesi öznitelik sayısında azaltma işlemi gerçekleştirilmiştir. Boyutu azaltılmış veri kümesine sırayla temel sınıflandırıcı algoritmalarından rastgele ağaç uygulanmıştır. Sonuçların geliştirilmesi için rotasyon orman algoritması uygulanarak model oluşturulmuştur. Önerilen model Şekil 1'de gösterildiği gibi gerçekleştirilmiştir.



Şekil 1. Önerilen model.

3.4. Diğer Sınıflandırma Algoritmaları

Çalışmada önerilen modelin dışında performansları karşılaştırmak için temel sınıflandırıcılardan (destek vektör makinesi, yapay sinir ağı, naive bayes ve rasgrele orman) ve topluluk algoritmaları (decorate, grading ve filteredclassifier) kullanılmıştır.

Biyolojik sinir hücrelerinin işleyişinden esinlenerek geliştirilen yapay sinir ağları, sinir hücreleri (nöronlar) olarak adlandırılan çok sayıda bağlantılı işlem elemanlarının problem çözümünde beraberce çalışması prensibine dayanmaktadır [30]. Destek vektör makinesi, iki veya çok boyutlu veri kümelerini sınıflandırmak amacı için doğrusal ve hiperdüzlem ayırma mekanizmalarını kullanır [31]. Naive Bayes sınıflandırma algoritması olasılık hesabına göre sınıflandırma işlemini gerçekleştirir [32]. Rastgele orman algoritması, sınıflandırma işlemi sırasında birden fazla karar ağacı üreterek sınıflandırma başarısını yükseltmeyi amaçlamaktadır [32]. Oluşturulan karar ağacı yapısı her sınıflandırma işleminin sonucunda bir oy almakta ve mevcut olan tüm ağaçlardan en yüksek oy alana göre sınıflandırmaya uygun olan ağaç yapısı tespit edilir. Veri setinin test aşamasında sınıf etiketi

bilinmeyen örnekler için tüm ağaç tahmininde en fazla oy alan ilgili sınıfa atanmasıyla sınıflandırma yapılmış olmaktadır [33].

Topluluk algoritmalardan, Adaboost algoritması yanlış olarak sınıflandırılmış örneklere daha fazla odaklanarak yüksek başarının elde edilmesi amaçlanmaktadır [34]. Decorate topluluk algoritması özel olarak oluşturulmuş yapay örnekleri kullanarak, çeşitli sınıflandırıcı topluluklarının oluşturulması ile sınıflandırma başarısının artırılması üzerine kurulmuştur[35]. Filtered Classifier topluluk algoritmasında, veriler keyfi bir süzgeçten geçirilerek, elde edilen verilere rastgele sınıflandırıcıların çalıştırıldığı algoritmadır [36]. Grading, temel sınıflandırıcıların çıktısının doğru ve hatalı etiketler ile derecelendiren ve derecelendirilmiş sonuçları daha sonra birleştiren topluluk algoritmasıdır [37].

4. DENEYSEL DEĞERLENDİRME

Bu bölümde öncelikle çalışmada kullanılan veri kümesinin detayları verilmiş ve daha sonra ise önerilen modelin deneysel değerlendirilmesi sunulmuştur.

Çalışmadaki sınıflama algoritmalarının değerlendirilmesi WEKA (Waikato Environment for Knowledge Analysis) yazılımı ile gerçekleştirilmiştir. Çalışmada Intel Core i5 7200U 2.5 GHZ işlemciye sahip, 4GB DDR3 hafızası bulunan sistem üzerinde uygulama gerçekleştirilmiştir.

4.1. Kronik Böbrek Rahatsızlığı Veri Kümesi

Çalışmada kullanılan kronik böbrek rahatsızlığı verileri 400 kayıttan oluşmaktadır ve '*UC Irvine Machine Learning Repository'* veri tabanından elde edilmiştir [4]. Veri kümesinde 24 özellik (11 özellik sayısal, 14 özellik kategorik) bulunmaktadır. Veri kümesinde 400 kayıtın 250 tanesi hastalıklı kişilere ait verilerden 150 tanesi ise sağlıklı kişilere ait verilerden oluşmaktadır. Çalışmada kullanılan verinin özellikleri Çizelge 1'de verilmektedir. Yapılan çalışmada Çizelge 1'de verilen özellikler sistemde giriş olarak kullanılmış çıkış olarak ise bireyin kronik böbrek rahatsızlığı olup olmadığı tahmin edilmeye çalışılmıştır.

Çizelge 1. Kronik Böbrek Rahatsızlığı Veri Kümesi Özellikleri ve Açıklamaları.

No	Özellik	Özellik Açıklama	Tip
1	age	Age (yaş)	Sayısal
2	Bp	Blood Pressure (kan basıncı-tansiyon)	Sayısal
3	Sg	Specific Gravity(idrara yoğunluğu- densite)	Kategorik
4	Al	Albumin(Albumin)	Kategorik
5	Su	Sugar (Şeker)	Kategorik
6	Rbc	Red Blood Cells(Kırmızı kan hücresi-eritrosit)	Kategorik
7	Pc	Pus Cell (iltihap)	Kategorik
8	Pcc	Pus Cell clumps(iltihap hücresi kümeleşmesi)	Kategorik
9	Ba	Bacteria (bakteri)	Kategorik
10	Bgr	Blood Glucose Random (Kan şekeri)	Sayısal
11	Bu	Blood Urea (Kan üre)	Sayısal
12	Sc	Serum Creatinine (Serum keratin)	Sayısal
13	Sod	Sodium (sodyum)	Sayısal
14	Pot	Potassium (Potasyum)	Sayısal
15	Hemo	Hemoglobin	Sayısal
16	Pcv	Packed Cell Volume (Hücre hacmi-Herotroksit)	Sayısal
17	Wc	White Blood Cell Count (Beyaz kan hücresi -Lokosit sayımı)	Sayısal
18	Rc	Red Blood Cell Count (Kırmızı kan hücresi sayımı)	Sayısal
19	Htm	Hypertension(Hipertansiyon)	Kategorik
20	Dm	Diabetes Mellitus(Diyabet melitis)	Kategorik
21	Cad	Coronary Artery Disease(Kroner arter hastalığı)	Kategorik
22	appet	Appetite(İştah durumu)	Kategorik
23	Pe	Pedal Edema (Ayak ödemi)	Kategorik
24	Ane	Anemia(Anemi-Kansızlık)	Kategorik

4.2. Algoritma Performanslarının Değerlendirilmesi

Çalışmada kullanılan veri kümesindeki veriler KBR rahatsızlığı olma veya olmama (ckd ve nockd) şeklinde iki farklı sınıf ile ifade edilmektedir. Çalışma 10 katlı çapraz doğrulama yöntemi kullanılarak veri kümesi eğitim ve test aşamasında geçirilmiştir. Çalışmada çeşitli algoritmaların performansları karşılaştırılmıştır ve bu algoritmaların test doğrulukları, aldıkları süre ve ağaç boyutları Çizelge 2'de verilmiştir. [27]. Algoritma performansları algoritmaların doğruluk keskinlik, duyarlılık, f-ölçütü ve MCC değerlerine göre karşılaştırılmıştır (sırasıyla Denklem 3, 4, 5, 6, ve 7). Formüllerdeki *DP* gerçek pozitifi, *YP* gerçek negatifi, *DN* yanlış pozitifi ve *YN* ise yanlış negatifi ifade etmektedir [41].

$$\boldsymbol{Doğruluk} = \frac{DP+DN}{DP+DN+YP+YN} \tag{3}$$

$$\boldsymbol{Keskinlik} = \frac{DP}{DP+YP} \tag{4}$$

$$\boldsymbol{Duyarlılık} = \frac{DP}{DP+YN} \tag{5}$$

$$\mathbf{F-Ölçütü} = \frac{2*\mathbf{Keskinlik}*\mathbf{Duyarlılık}}{\mathbf{Keskinlik}+\mathbf{Duyarlılık}} \tag{6}$$

$$\boldsymbol{MCC} = \frac{DP*DN-YP*YN}{\sqrt{(DP+YP)(DP+YN)(DN+YP)(DN+YN)}} \tag{7}$$

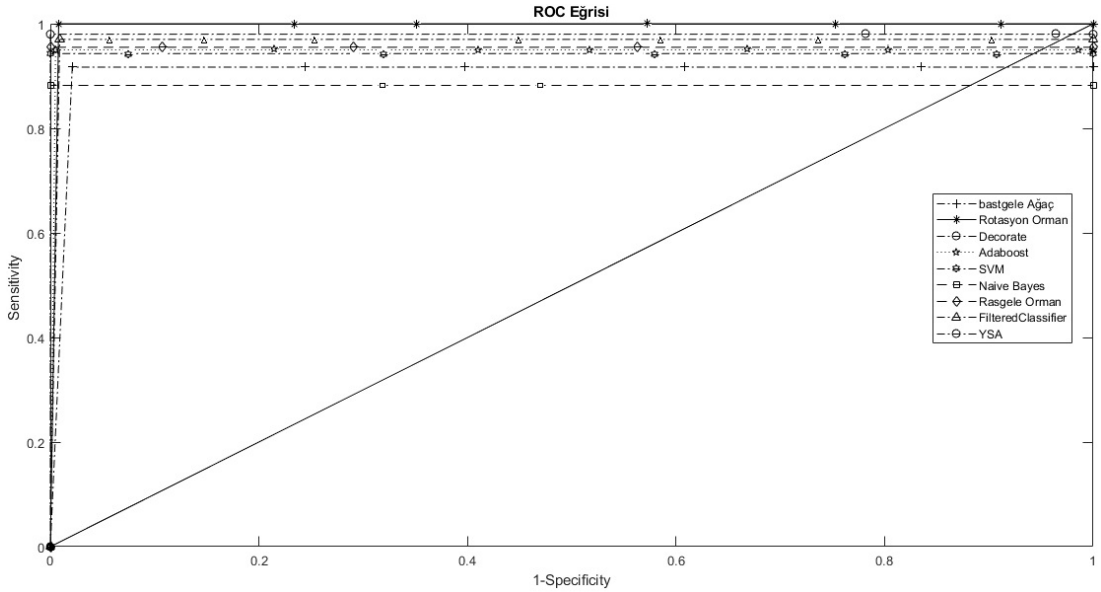
Çalışmada algoritmaların performanslarını karşılaştırmanın yanında, istatistiksel anlamlılık testi olan t-istatistiği kullanılarak p değeri hesaplanmıştır.

Çizelge 2. Topluluk Algoritmaları ve Temel Sınıflandırıcıların Deneysel Sonuçları.

Algoritma	Test Doğruluk %	Keskinlik	Duyarlık	F-Score	MCC	Süre (saniye)	Ağaç Boyutu
Rotasyon Orman+ Rastgele Ağaç	99,5	0,995	0,995	0,995	0,989	0,19	33
Decorate+ Rastgele Ağaç	99,25	0,993	0,993	0,993	0,984	0,24	959
Grading+ Rastgele Ağaç	62,5	0,391	0,625	0,481	0	0,01	102
FilteredClassifier + Rastgele Ağaç	98,5	0,986	0,985	0,985	0,969	0,03	87
Adaboost+ Rastgele Ağaç	98	0,981	0,980	0,980	0,958	0,01	76
Rastgele Ağaç	95,5	0,956	0,955	0,955	0,906	0.05	67
Destek Vektör Makinesi	97,75	0,979	0,978	0,978	0,954	0,06	-
Naive Bayes	95	0,956	0,950	0,951	0,901	0	-
Rastgele Orman	98,25	0,983	0,983	0,983	0,964	0,01	-
Yapay Sinir Ağı	99,25	0,993	0,993	0,993	0,984	0,75	-

Bu çalışmada rastgele ağaç algoritması sezgisel olmayan kolektif algoritmalarından adaboost, rotasyon orman, decorate, grading, filtered classfication algoritmaları ayrıca destek vektör makinesi, naive bayes, rastgele orman ve yapay sinir ağı ile çalıştırılmıştır. Algoritmaların uygulanması sonrasında, kronik böbrek rahatsızlığını tespite yönelik uygunluklarını kıyaslamak amacıyla doğruluk, ksinlik, duyarlılık, f-ölçütü ve mcc alanı ölçütleri kullanılmıştır. Çalışmada kullanılan rotasyon orman algoritması için sınıflandırıcı sayısı (*L*) 10, alt küme sayısı (*K*) 3 olarak belirlenmiştir [39]. Ayrıca

grup sayısı 3 ile 10 arasından sayıyla (varsayılan 3), yineleme sayısını varsayılan 10 ve örneklerin kaldırılma yüzdesini %50 varsayılan değerlerini ayarladık [40]. Bununla birlikte, algoritmaların veri kümesi üzerinde işletim sürelerine göre performans karşılaştırmaları gerçekleştirilmiştir. Elde edilen sonuçlar Çizelge 2'de verilmiştir. Çizelge 2 incelendiğinde topluluk algoritmalarının temel sınıflandırıcı algoritmalarına göre performansının arttığı görülmektedir. Algoritma sonuçları incelendiğinde rotasyon orman algoritması test Doğruluk oranı %99.5, Kesinlik 0,995, Duyarlılık 0,995, F-ölçütü 0,995, MCC 0,989, çalışma zamanı 0.19 saniye ve oluşturulan ağaç boyutu 33 olarak görülmektedir. Çalışmada en kötü başarı Grading algoritması ile Doğruluk oranı %62.5, Kesinlik 0,391, Duyarlılık 0,625, F-ölçütü 0,481, MCC 0, çalışma zamanı 0.01 saniye ve oluşturulan ağaç boyutu 102 olarak görülmektedir.



Şekil 2. Rotasyon Orman+ Rastgele Ağaç ROC Eğrisi.

Şekil 2'deki ROC grafiği sensitivity ve specificity oranlarından elde edilen eğrinin altında kalan alanı 1 değerine ne kadar yakın ise grafiğin daha iyi sonuç verdiği söylenebilmektedir. Şekil 2'de de ROC alanı değeri 1 olduğundan başarı oranı yüksek olduğu gözlemlenmektedir.

Çizelge 3. Topluluk Algoritmaları Ve Temel Sınıflandırıcı Algoritmalarının Performans Analizi İçin Hesaplanan İstatistik Değerle.

Algoritma	P Değeri
Rotasyon Orman+ Rastgele Ağaç	0.470
Decorate+ Rastgele Ağaç	0.248
Grading+ Rastgele Ağaç	<2.2e
FilteredClassifier + Rastgele Ağaç	0.412
Adaboost+ Rastgele Ağaç	0.077
Rastgele Ağaç	0.098
Destek Vektör Makinesi	0.007
Naive Bayes	2.152e-05
Rastgele Orman	0.023
Yapay Sinir Ağı	0.248

Çizelge 2'deki sonuçların istatistiksel olarak anlamlılığı ölçek için %95 güven ve p-değeri (<0.05) aralığında olduğu sonucuna Çizelge 3 incelenerek ulaşılmıştır. Böylece rotasyon ormanı algoritması kullanarak yüksek doğruluk düzeyi elde edildiği görülmüştür. Algoritma sonucunda oluşturulan ağacın boyutunun küçük olması ve derinliğinin minimum düzeyde olması beklenmektedir [25,26]. Karar ağacı algoritmalarında meydana gelen genelleştirme hatalarını minimuma indirebilmek için ağaç derinliğinin en aza indirgenmesi önemlidir. Bundan dolayı seçmiş olduğumuz algoritmanın ağaç derinliğinin küçük olması başarı sonucunun yüksek olmasını sağlamaktadır. Sonuç olarak elde edilen sonuçlar karşılaştırıldığında en iyi performansı önerilen rotasyon orman algoritmasının verdiği görülmektedir.

5. SONUÇ

Kronik böbrek rahatsızlığı (KBR) dünya genelinde yaygın görülen hastalıklardan ve hastalığın tahmini tıbbi tanıdaki temel konulardan biridir. Kronik böbrek rahatsızlığı dünya çapında önde gelen ölüm nedenlerinden biridir. Literatürde hastalığı tahmin etmek için çok sayıda veri madenciliği algoritmaları kullanılmış olduğu görülmüştür. Çalışmada hastalığı tahmin etmek amacıyla sınıflandırma algoritmalarının performansları incelenmiştir. Bu çalışmada rotasyon orman algoritması ile rastgele ağaç algoritmasının kombinasyonu KBR verilerinin sınıflandırılması için önerilmiştir. Sonuçlar değerlendirildiğinde önerilen modelin diğer sınıflandırma algoritmalarına göre daha iyi performans sergilediği görülmüştür. Bu sayede uzman doktorlara uygulayacakları tedavi öncesinde doğru teşhis sağlanması, önlenebilir hastalıklar için zamanında tarama yapılması gibi klinik gereksinimlere destek vermesine katkı sağlayacağı düşünülmektedir.

İleriki çalışmalarda, kronik böbrek rahatsızlığını tahmini için derin öğrenme yöntemleri gibi farklı algoritmaların kullanımı mümkün olabilir. Ayrıca, veriye uygun algoritma parametrelerinin tespiti için çalışmalar yapılabilir.

KAYNAKÇA

[1] Erdem, B. K., & Akbas, H., (2017), Kronik Böbrek Hastalığı ve Vasküler Kalsifikasyon, Türk Klinik Biyokimya Derg. , 152: 89-98.

[2] Singh, P., Chandola, V., & Fox, C., (2017), Automatic Extraction of Deep Phenotypes for Precision Medicine in Chronic Kidney Disease, Proceedings of the 2017 International Conference on Digital Health - DH '17, 195–199. https://doi.org/10.1145/3079452.3079489

[3] Topbaş, E., (2015), Kronik Böbrek Hastalığının Önemi , Evreleri Ve Evrelere Özgü Bakımı, Nefroloji Hemşireliği Dergisi, 53-59.

[4] Dua, D. and Karra Taniskidou, E., (2017), UCI Machine Learning Repository [http://archive.ics.uci.edu/ml. Irvine, CA: University of California, School of Information and Computer Science.

[5] İlkuçar, M., (2015), Kronik Böbrek Hastalarının Yapay Sinir Ağı ve Radyal Temelli Fonksiyon Ağı ile Teşhisi. Mehmet Akif Ersoy Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 6(2), 82-88.

[6] Jena, L., & Kamila K. N., (2015), Distributed Data Mining Classification Algorithms for Prediction of Chronic-Kidney-Disease, International Journal of Emerging Research in Management &Technology, 93594, 2278–9359.

[7] Rubini, L. J., & Eswaran, P., (2015), Generating comparative analysis of early stage prediction of Chronic Kidney Disease. International Journal of Modern Engineering Research (IJMER), 5(7), 49-55.

[8] Yildirim, P., (2017), Chronic Kidney Disease Prediction on Imbalanced Data by Multilayer Perceptron: Chronic Kidney Disease Prediction, 2017 IEEE 41st Annual Computer Software and Applications Conference COMPSAC, 193–198.

[9] Ramya, S., & Radha, N., (2016), Diagnosis of chronic kidney disease using machine learning algorithms, International Journal of Innovative Research in Computer and Communication Engineering, 4(1), 812-820.

[10] Kunwar, V., Chandel, K., Sabitha, A. S., & Bansal, A., (2016), Chronic Kidney Disease analysis using data mining classification techniques, In IEEE 6th International Conference Cloud System and Big Data Engineering (Confluence), 300-305.

[11] Vijayarani, S., & Dhayanand, S., (2015), Data Mining Classification Algorithms for Kidney Disease Prediction, International Journal on Cybernetics & Informatics, 44, 13–25.

[12] Kumar, M., (2016), Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm Running Title: Prediction of Chronic Kidney Disease, International Journal of Computer Science and Mobile Computing, 52522, 24–33.

[13] Balakrishna, T., Narendra, B., Reddy, M. H., & Jayasri, D., (2017), Diagnosis of Chronic Kidney Disease Using Random Forest Classification Technique, HELIX, 7(1), 873-877.

[14] Baby, P. S., & Vital, P., (2015), Statistical Analysis and Predicting Kidney Diseases using Machine Learning Algorithms, International Journal of Engineering Research & Technology, 407, 206–210.

[15] Polat, H., Mehr, H. D., & Cetin, A., (2017), Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods, Journal of Medical Systems, 41(4), 55.

[16] Sinha, P., (2015), Comparative study of chronic kidney disease prediction using KNN and SVM, International Journal of Engineering Research and Technology, 4(12), 608-12.

[17] Eroğlu K.ve Palabaş T., (2016), Kronik Böbrek Hastalığı Tespitinde Farklı Sınıflandırma Yöntemleri ve Farklı Topluluk Algoritmalarının Birlikte Kullanımının Sınıflandırma Performansına Etkisi, Elektrik-Elektronik Mühendisliği Odası, 512–516.

[18] Cheng, L. C., Hu, Y. H., & Chiou, S. H., (2017), Applying the Temporal Abstraction Technique to the Prediction of Chronic Kidney Disease Progression, Journal of medical systems, 41(5), 85.

[19] Başar, M. D., Sarı, P., Kılıç, N., & Akan, A., (2016), Detection of chronic kidney disease by

using Adaboost ensemble learning approach, In IEEE Signal Processing and Communication Application Conference (SIU), 2016 24th . 773-776.

[20] Rodríguez, J. J., Kuncheva, L. I., & Alonso, C. J., (2006), Rotation forest: A New classifier ensemble method, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2810, 1619–1630.

[21] Akçetin, E., & Çelik, U., (2014), İstenmeyen Elektronik Posta (Spam) Tespitinde Karar Ağacı Algoritmalarının Performans Kıyaslaması, Journal of Internet Applications & Management/İnternet Uygulamaları ve Yönetimi Dergisi, 5(2).

[22] Namlı, Ö. H., & Özcan, T., (2017), Makine Öğrenmesi Algoritmaları Kullanarak Gişe Hasılatının Tahmini, Yönetim Bilişim Sistemleri Dergisi, 3(2), 130-143.

[23] Ali, J., Khan, R., Ahmad, N., & Maqsood, I., (2012), Random forests and decision trees, IJCSI International Journal of Computer Science Issues, 9(5), 272-278.

[24] Breiman, L., (2001), Random forests. Machine Learning, 451, 5–32.

[25] Onan, A., (2015), Şirket İflaslarının Tahminlenmesinde Karar Ağacı Algoritmalarının Karşılaştırmalı Başarım Analizi, Bilişim Teknolojileri Dergisi, 81, 9–19.

[26] Sebban, M., Nock, R., Chauchat, J. H., & Rakotomalala, R., (2000), Impact of learning set quality and size on decision tree performances. International Journal of Computers, Systems and Signals, 11, 85–105.

[27] Fawcett, T., (2006), An introduction to ROC analysis, Pattern recognition letters, 27(8), 861

[28] Türkiye Halk Sağlığı Kurumu, Türkiye Böbrek Hastalıkları Önleme ve Kontrol Programı Eylem Planı, (2014-2017). Sağlık Bakanlığı, Yayın No: 946, Ankara, 2014, ss. 1. http://www.tsn.org.tr/pdf/Turkiye_Bobrek_Hastaliklari_Onleme_ve_Kontrol_Programi.pdf (E.T. 08.06.2018).

[29] Horspool S., (2016), Global Burden of Disease Study 2015 outlines chronic kidney disease as a cause of death worldwide. https://www.theisn.org/news/item/2969-global-burden-of-disease-study-2015-outlines-chronic-kidney-disease-as-a-cause-of-death-worldwide (Erişim Tarihi: 20.06.2018).

[30] Kunwar, V., Chandel, K., Sabitha, A. S., & Bansal, A., (2016), Chronic Kidney Disease analysis using data mining classification techniques. In Cloud System and Big Data Engineering (Confluence), 2016 6th International Conference(pp. 300-305). IEEE.

[31] Vapnik, V., (1995), "The nature of statistical learning theory," Springer-Verlag: New York, pp. 75-100.

[32] Orhan, U., Adem, K., & Comert, O., (2012), Least squares approach to locally weighted naive Bayes method. Journal of New Results in Science, 1(1).

[33] Breiman, L., (2001), Random forests. Machine learning, 45(1), 5-32.

[34] Hu, W., Hu, W., & Maybank, S., (2008), Adaboost-based algorithm for network intrusion detection. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 38(2), 577-583.

[35] Melville, P., & Mooney, R. J., (2003), Constructing diverse classifier ensembles using artificial training examples. In IJCAI, Vol. 3, pp. 505-510.

[36] Witten, I. H., Frank, E., & Hall, M. A., (2005), Data mining: Practical machine learning tools and techniques, (morgan kaufmann series in data management systems). Morgan Kaufmann, June, 104, 113.

[37] Pasha, M., & Fatima, M., (2017), Comparative Analysis of Meta Learning Algorithms for Liver Disease Detection. Journal of Software, 12(12), 923-934.

[38] Bagnall, A., Bostrom, A., Cawley, G., Flynn, M., Large, J., & Lines, J., (2018), Is rotation forest the best classifier for problems with continuous features?. arXiv preprint arXiv:1809.06705.

[39] Hosseinzadeh, M., & Eftekhari, M., (2015), Improving rotation forest performance for imbalanced data classification through fuzzy clustering. In Artificial Intelligence and Signal Processing (AISP), 2015 International Symposium on (pp. 35-40). IEEE.

[40] Mauša, G. O. R. A. N., Bogunović, N., Grbac, T. G., & Bašić, B. D., (2015), Rotation forest in software defect prediction. SQAMIA, 1375, 35-43.

[41] Coşkun, C., & Baykal, A., (2011), Veri Madenciliğinde Sınıflandırma Algoritmalarının Bir Örnek Üzerinde Karşılaştırılması. Akademik Bilişim, 2011, 1-8.

[42] Qi, Y., (2012), Random forest for bioinformatics. In Ensemble machine learning (pp. 307-323). Springer, Boston, MA.

[43] Gislason, P. O., Benediktsson, J. A., & Sveinsson, J. R., (2006), Random forests for land cover classification. Pattern Recognition Letters, 27(4), 294-300.

[44] Pal, M., (2005), Random Forest Classifier For Remote Sensing Classification, International Journal Of Remote Sensing, 26(1) , 217-222.

[45] Veerappan, I., & Abraham, G., (2013), Chronic kidney disease: Current status, challenges and management in India. Ch, 130, 593-7.

[46] Go, A. S., Chertow, G. M., Fan, D., McCulloch, C. E., & Hsu, C. Y., (2004), Chronic kidney disease and the risks of death, cardiovascular events, and hospitalization. New England Journal of Medicine, 351(13), 1296-1305.

[47] Tan, P. N., (2007), Introduction to data mining. Pearson Education India.

[48] Garg, A. X., Adhikari, N. K., McDonald, H., Rosas-Arellano, M. P., Devereaux, P. J., Beyene, J. & Haynes, R. B., (2005), Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. Jama, 293(10), 1223-1238.