1. Uluslararası Bilim, Mühendislik ve Robotik Teknolojisindeki Gelişmeler Konferansı 2019 (ICASERT 2019)

Rastgele Ormanlar ile Küme Tabanlı Alt Örnekleme Kullanarak Dengesiz Ağ İzinsiz Giriş Sınıflandırması için Algılama Doğruluğunun İyileştirilmesi

Md. Ochiuddin Miah, Sakib Shahriar Khan, Swakkhar Shatabda ve Dewan Md. Farid*
Bilgisayar Bilimi ve Mühendisliği Bölümü, Uluslararası Birleşik Üniversite
United City, Madani Avenue, Badda, Dhaka 1212, Bangladeş
E-posta: mmiah131145@bscse.uiu.ac.bd, skbshariar@gmail.com, swakkhar@cse.uiu.ac.bd, dewanfarid@cse.uiu.ac.bd

Özet: Dengesiz büyük veri ortamında *ağ* saldırılarının sınıflandırılması, içinde bulunduğumuz dijital çağda bilgi ve iletişim teknolojilerinde (ICT) önemli bir konu haline gelmiştir. Şu anda, saldırı tespit sistemleri (IDS'ler) iç ve dış ağ saldırılarını / izinsiz girişleri tespit etmek ve önlemek için yaygın olarak kullanılan bir araçtır. IDS'ler temel olarak ana bilgisayar tabanlı ve ağ tabanlı sistemler olarak ikiye ayrılır ve kötüye kullanım tabanlı saldırı tespit sistemi olarak bilinen saldırıları tespit etmek için desen eşleştirme tekniklerini kullanır. Makine öğrenimi (ML) ve veri madenciliği (DM) algoritmaları, son birkaç on yılda IDS'lerde izinsiz girişleri sınıflandırmak için yaygın olarak kullanılmaktadır. Makine öğrenimi ve veri madenciliği algoritmalarını kullanan IDS oluşturmanın en büyük zorluklarından biri, izinsiz giriş sınıflandırma doğruluğunu artırmak ve aynı zamanda yanlış pozitif oranını azaltmaktır. Bu makalede, Rastgele Orman sınıflandırıcısı ile küme tabanlı alt örnekleme kullanarak azınlık sınıfı ağ saldırılarını / izinsiz girişleri sınıflandırmak için algılama oranını artırmak için yeni bir yöntem sunduk. Önerilen yöntem, azınlık/nadir sınıf izinsiz girişleri doğru bir şekilde tanımlamak için oldukça dengesiz büyük verileri işleyebilen çok katmanlı bir sınıflandırma yaklaşımıdır. Başlangıçta, önerilen yöntem bir veri noktasını / gelen veriyi saldırı / izinsiz giriş olup olmadığını (normal davranış gibi) sınıflandırır, eğer bir saldırı ise, önerilen yöntem saldırı türünü ve daha sonra alt saldırı türünü sınıflandırmaya çalışır. Sınıf dengesizliği sorunuyla başa çıkmak için küme tabanlı düşük örnekleme tekniğini ve aşırı uyum sorununu ele almak için popüler topluluk sınıflandırıcısı Random Forest'ı kullandık. Deneysel analiz için KDD99 saldırı tespit ölçüt veri setini kullandık ve önerilen yöntemin performansını mevcut makine öğrenimi algoritmaları ile test ettik: Yapay Sinir Ağı (YSA), na¨ıve Bayes (NB) sınıflandırıcı, Random Forest ve Bagging teknikleri.

Anahtar Kelimeler-Veri Örnekleme; Topluluk Öğrenmesi; İzinsiz Giriş Sınıflandırması; Dengesiz B ig D ata; Rastgele Orman;

I. GIRIŞ

Son zamanlarda bilgi ve iletişim teknolojilerinin yaygınlaşmasıyla birlikte bilgisayar ağları solucanlar, virüsler, reklam yazılımları, casus yazılımlar, rootkitler, truva atları vb. tehditlerle karşı karşıya kalmaktadır. [1]. Bu tür tehditler/izinsiz girişler büyük bir endişeye dönüşmüştür ve bilgisayar ağlarını tahrip etmeden önce algılanmaları gerekmektedir. Bu izinsiz girişler bilgisayar ağlarının, kurumsal elektronik varlıkların ve diğer kaynakların verimliliğini etkilemektedir. İzinsiz giriş tespit sistemi, ağları ve sistemleri inceleyerek ana bilgisayar veya ağdan gelen kötü niyetli faaliyetleri/izinsiz girişleri tespit eden bir ağ güvenliği teknolojisidir [2]. Ağ tabanlı IDS (NIDS) ve ana bilgisayar tabanlı IDS (HIDS), dahili veya harici saldırıları tanımlamak için sıklıkla kullanılır.

978-1-7281-3445-1/19/$31.00 ○c 2019 IEEE

kötü niyetli faaliyetler / izinsiz girişler. NIDS, ağ izinsiz girişlerini tespit etmek için gelen ağ trafiğini inceler ve izler [3]. HIDS, gelen ve giden paketleri inceleyerek izinsiz girişleri tespit etmek için tek bir ana bilgisayara veya bilgisayar cihazına ayrılmıştır [4]. Tespit yaklaşımında, IDS'ler yanlış kullanım tabanlı veya anomali tabanlı IDS olarak kategorize edilir [5]. Yanlış kullanım tabanlı IDS, desen tabanlı veya imza tabanlı IDS olarak da kabul edilir [6]. Anonim saldırıları tespit etmekte yetersizdir ve saldırganlar sürekli olarak farklı imzalarla saldırmaya çalıştıkları için tespit oranı nispeten düşüktür [7]. Aksi takdirde, anomali tabanlı IDS, bilinen veya bilinmeyen ağ izinsiz girişlerini tespit etmek için yaygın olarak kullanılır [3]. Ağ izinsiz girişleri, büyük hacimli ağ izinsiz giriş tespit veri kümesinden gizli izinsiz giriş modellerinin tespit edilmesiyle belirlenir [8]. YSA, DVM, karar ağacı (DT), genetik algoritma, na¨ıve Bayesian, bulanık mantık vb. gibi birçok makine öğrenimi (ML) ve veri madenciliği (DM) algoritması. [9] gibi algoritmalar son birkaç on yıldır IDS'lerde izinsiz girişleri sınıflandırmak için kullanılmaktadır. Günümüzün elde edilebilir ağ saldırı tespit veri kümeleri, dengesiz, çok yönlü, dinamik ve farklı tutarsız verilerden oluşmakta ve bu da tüm verinin işlenmesinde sorunlara neden olmaktadır. Birçok makine öğrenimi algoritması, veri kümesindeki bazı saldırıların diğer saldırılara kıyasla çok düşük olması nedeniyle her türlü saldırıyı yüksek tespit doğruluğu ile tanımlamakta başarısız olmuştur. Günümüzün erişilebilir saldırı tespit sistemleri, bilinen veya bilinmeyen her türlü ağ saldırısı/izinsiz giriş için düşük tespit doğruluğuna ve yüksek yanlış pozitif oranına sahiptir [10].

Bu makalede, Rastgele Orman sınıflandırıcısı ile küme tabanlı alt örnekleme kullanarak azınlık sınıfı ağ saldırılarını / izinsiz girişleri sınıflandırmak için tespit oranını artırmak için yeni bir yöntem sunduk. Başlangıçta, önerilen yöntem bir veri noktasını / gelen veriyi saldırı / izinsiz giriş olup olmadığını sınıflandırır (normal davranış gibi), eğer bir saldırı ise, önerilen yöntem saldırı türünü ve daha sonra alt saldırı türünü sınıflandırmaya çalışır. Sınıf dengesizliği sorunuyla başa çıkmak için küme tabanlı düşük örnekleme tekniğini ve aşırı uyum sorununu ele almak için popüler topluluk sınıflandırıcısı Random Forest'ı kullandık. Bu makalenin en önemli amacı, dengesiz ağ saldırı sınıflandırmasında düşük frekanslı saldırıların tespit doğruluğunu artırmaktır. YSA [9], NB sınıflandırıcı [11], Rastgele Orman [12], Bagging [13] kullanan çeşitli sınıflandırma modelleri oluşturduk ve bu mevcut öğrenme yöntemlerinin performansını önerilen yöntemle karşılaştırdık. Önerilen yöntem, mevcut sınıflandırıcılara kıyasla KDD99 kıyaslama veri kümesinde ortalama %98 tespit oranı elde etmiştir.

Çalışmanın geri kalanı aşağıdaki şekilde düzenlenmiştir. Bölüm II'de ilgili çalışmalar sunulmaktadır. Bölüm III'te etkili bir çok katmanlı hibrit yöntem sunulmaktadır. Bölüm IV, KDD99 kıyaslama veri kümesine dayalı deneysel sonuçlar sunmaktadır. Son olarak, Bölüm V'te sonuçlar ve gelecek çalışmalar sunulmaktadır.

II. İLGILI ÇALIŞMA

Singh ve diğerleri [1], izinsiz giriş tespiti için sinir ağının yavaş öğrenme dezavantajını azaltmak için bir Çevrimiçi Sıralı Ekstrem Öğrenme tekniği sunmuştur. Sırasıyla beta ve alfa profilleme uygulayarak eğitim örneklerinin hacmini ve zaman karmaşıklığını azaltmışlardır. Ayrıca tutarlı özellik seçim teknikleri ve filtrelenmiş topluluk kullanarak alakasız özellikleri atmışlardır. Önerilen tekniğin performansını ölçmek için KDD 2009 veri kümesi kullanılmıştır. Önerilen IDS, ikili sınıf ve çoklu sınıf için sırasıyla %98,66 ve %97,67 doğruluk ve %1,74 yanlış pozitif oranına ulaşmıştır. Önerilen IDS, doğruluk ve tespit süresine göre yayınlanmış diğer tekniklerden daha iyi performans göstermiştir. Farid ve diğerleri [14] uyarlanabilir izinsiz giriş tespiti için NB sınıflandırıcı ve bir topluluk yaklaşımı kullanarak yeni bir öğrenme yöntemi sunmuştur. Önerilen yöntem, bir dizi sınıflandırıcının oylarını dikkate alarak bilinen veya bilinmeyen bir örneği sınıflandırabilir. Bu yöntem, sınıflandırma hata oranlarının sonucuna göre eğitim örneklerinin ağırlıklarını güncelleyebilir, bu da yanlış pozitif (FP) oranlarını azaltır ve tespit oranlarını (DR) iyileştirir. Mevcut veri madenciliği algoritmalarını ve önerilen yöntemi KDD99 veri kümesi üzerinde uygulamışlardır. Önerilen yöntem, yanlış pozitif oranlarını azaltabilir ve farklı türdeki ağ izinsiz girişlerinde yüksek tespit oranlarına ulaşabilir. Farid ve diğerleri [15] karar ağacı kullanarak bilinen veya bilinmeyen izinsiz girişleri tespit etmek için anomali tabanlı IDS için yeni bir öğrenme algoritması tanıtmıştır. Tüm alt veri kümesi aynı sınıfa ait olana kadar her bir örneğin ağırlığını olasılıklara göre böler ve ayarlar. Geleneksel karar ağacı algoritmalarında her örneğin ağırlığı aynıdır, ancak bu yöntemde her örneğin ağırlığı sonsal olasılığa göre değişir. Önerilen öğrenme prosedürü, mevcut diğer algoritmalarla karşılaştırıldığında KDD99 veri kümesi üzerinde %98 tespit oranı elde etmiştir.

Li ve diğerleri [9] NSL-KDD veri kümesini kullanarak k-NN tekniği ve ikili sınıflandırma üzerine inşa edilmiş iki aşamalı bir hibrit saldırı tespit yöntemi önermiştir. İlk olarak, önerilen yöntem belirsiz sınıfları tanımlamak üzere ikili sınıflandırıcıyı eğitmek için C4.5 algoritmasını kullanmaktadır. İkinci olarak, önerilen yöntem belirsiz sınıfları sınıflandırmak için k-NN algoritmasını kullanır. Bu adımların bir araya getirilmesiyle, önerilen teknik Random Forest, na¨ıve Bayes, back- ward propagation neural network ve k-NN gibi temel algoritmalardan daha iyi performans göstermektedir. Önerilen bu tekniğin F1 skorları, U2R ve R2L düşük frekanslı saldırılar için temel algoritmalardan çok daha yüksektir. Zarpela˜o ve diğerleri [16] IOT için saldırı tespit sistemi araştırması üzerine bir analiz sunmuştur. IDS için IoT araştırmasının hala devam etmekte olduğunu ve amaçlarının açık sorunları, önde gelen eğilimleri ve gelecekteki araştırma olanaklarını tanımak olduğunu denemişlerdir. IDS'leri IDS yerleştirme stratejisi, doğrulama stratejisi, tespit şeması ve güvenlik korkutmasına göre kategorize etmişlerdir. IoT için IDS geliştirmek için imza tabanlı, spesifikasyon tabanlı, anomali tabanlı ve hy- brid tespit mekanizmalarını tartıştılar. Ayrıca merkezi IDS yerleştirme, dağıtık IDS yerleştirme ve hibrit IDS yerleştirme teknikleri hakkında tartıştılar.

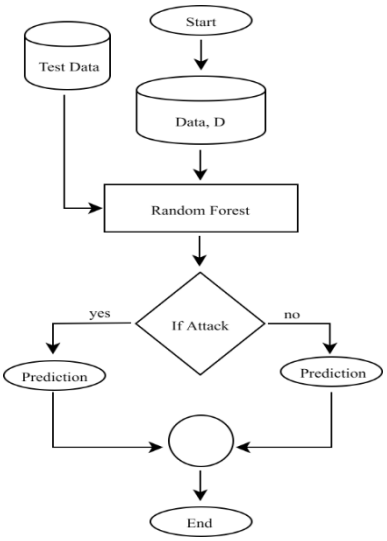
III. ÖĞRENME ALGORITMALARI

A. Öğrenme Programı

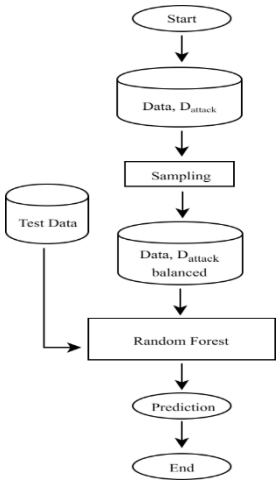
Önerilen yöntemimizde öğrenme şeması olarak Rastgele Orman kullandık. Rastgele Orman bir topluluk öğrenme yöntemidir ve denetimli öğrenmede sınıflandırma ve regresyon için kullanılır[17]. Büyük verileri daha yüksek doğrulukla sınıflandırma yeteneğine sahiptir [18]. Eğitim zamanında keyfi olarak seçilen öznitelik grubuna bağlı olarak birden fazla karar ağacı oluşturur ve ağaç topluluğunun oylarından seçilen sınıfı en popüler sınıf olarak çıkarır [11]. Doğruluk ve ağaç sayısı arasında bir ilişki vardır, daha fazla sayıda ağacın daha yüksek doğruluk oranı üretmesi beklenir. Makine öğreniminin en büyük sorunlarından biri aşırı uyumdur, model eğitim verilerine aşırı uyum sağladığında ortaya çıkar, modele eğitim verilerinin dışından yeni bir test örneği geldiğinde, onu doğru bir şekilde sınıflandıramaz, yalnızca eğitim verileri test verileri olarak kullanıldığında daha yüksek performans verir. Rastgele Orman algoritması, ağaç topluluğundan oy aldığı için bu sorunu çözmüştür, bu nedenle modele aşırı uymayacaktır. Oluşturduğu her karar ağacının oylarını dikkate aldığından, tek bir karar ağacından çok daha yüksek doğruluk elde etmesi beklenir [11].

B. Önerilen Yöntem

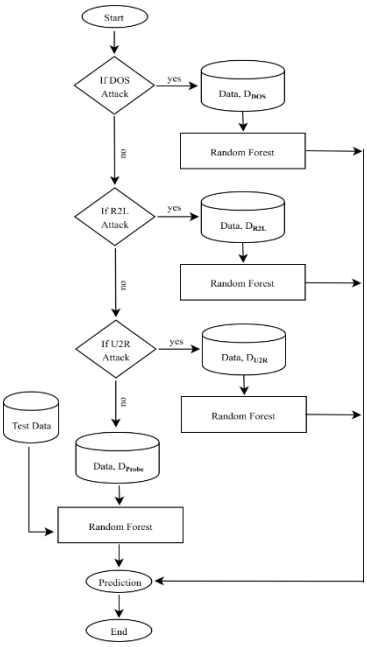
Önerilen yöntemimizin ilk katmanında, *D* veri kümesini ve Rastgele Orman tekniğini kullanarak bir *M* modeli oluşturduk. Yeni bir x_{new} örneği geldiğinde, model *M* bunun bir saldırı olup olmadığını tahmin eder. Bunu takiben, eğer x_{new} normal olarak tanımlanmışsa burada bırakacağız, ancak bir saldırı olarak tahmin edilirse, *D* veri kümesini sırasıyla $_{Dattack}$ ve $_{Dnormal\ olmak\ üzere}$ iki farklı veri kümesine ayırdık. Sadece $Dattack$ veri kümesi ile çalıştık ve 4 farklı saldırı kategorisi içeriyor, $_{Dattack} = D_{DOS} \cup D_{U2R} \cup D_{R2L} \cup D_{Probe}$. Arasında
KDD99 veri setindeki 22 saldırı arasında büyük bir dengesizlik vardır
miktar bağlamında veri. Dengeyi sağlamak için yeniden örnekleme yapmamız gerekti. Bunu yapmak için, ikinci katmanda $_{Dattack}$ veri kümesine eksik örnekleme ve aşırı örnekleme tekniklerini uyguladık. Alt örnekleme yapmak için küme centroid tekniği uygulandı, *X* örneklerini bir dizi küme C_1 , C_2 , ..., Cn ile böler ve yalnızca küme $centroidlerini\ Centroid(C_i)$ alır, böylece bilgilendirici örnekler kayboldu. Bu işlemden sonra, etiketlerini değiştirdik
saldırı kategorisi ile $_{Dattack}$ veri kümesi. Daha sonra $_{Dattack}$ ve Random Forest veri kümesini kullanarak başka bir model $_{Mattack}$ oluşturduk. $_{Mattack}$ tarafından x_{new} veri kümesinin saldırı kategorisi tahmin edilmiştir. Şimdi saldırıların etki alanı kısaltıldı, bu nedenle D_{new} verilerini üçüncü katmandaki $_{Dattack'}$ten ayırdık. D_{new} , yalnızca X_{new} 'in tahmin edilen kategorisinin saldırılarını içeren *X* örneklerini içerir. Nihai modelimiz M_{new} *'yi* D_{new} veri kümesi ve Random Forest ile oluşturduk. Bu model $x'in$ gerçek sınıfını veya saldırısını tahmin edecektir$_{new}$. Önerilen hibrit yöntemin mimarisi Şekil 1'de gösterilmiş ve algoritma Algoritma 1'de özetlenmiştir. Şekil 1'de, (a) önerilen yöntem gelen bir veriyi saldırı/izinsiz giriş olup olmadığını sınıflandırır (normal davranış gibi), eğer bu bir saldırı ise, önerilen yöntem Şekil 1 (b)'de gösterilen ana saldırı türlerini sınıflandırmaya çalışır ve daha sonra Şekil 1 (c)'de gösterilen nihai saldırı/izinsiz girişi sınıflandırır.



(a)



(b)



(c)

Şekil 1: Önerilen yöntem (a) normal veya saldırı tespiti; (b) ana saldırı türlerinin tespiti; (c) nihai saldırı/saldırı tespiti.

IV. DENEYSEL ANALIZ

Sınıflandırma algoritmasının performansı literatürde çeşitli değerlendirme metrikleri ile ölçülmektedir. Bu makalede, tespit oranlarını (DR) [17], yanlış pozitifleri (FP) kullandık [17] standart değerlendirme ölçütleri olarak kullanılmaktadır. Tespit oranı ve yanlış pozitif, Eşitlik 1 ve 2'de gösterilen IDS'nin performansını tahmin etmek için yaygın olarak kullanılmaktadır.

$$DR = \frac{Toplam\ tespit\ edilen\ saldırılar}{Toplam\ saldırılar} \times 100 \quad (1)$$

$$FP = \frac{Toplam\ yanlış\ sınıflandırılmış\ süreç}{Toplam\ normal\ süreç} \times 100 \quad (2)$$

A. İzinsiz Giriş Tespit Veri Kümesi

Saldırı tespit veri setinin kullanılabilirliği çok sınırlıdır çünkü ana bilgisayar, topoloji ve diğer gizli bilgiler hakkında bilgi içerir. KDD99 standart saldırı tespit veri kümesi, ağ saldırı dedektörünü oluşturmak için kullanılmıştır

normal bağlantılar ve izinsiz girişler arasında ayrım yapmak için

Algoritma 1 Önerilen Hibrit Yöntem

Girdi: Veri D ve bir öğrenme şeması.
Çıktı: Tahmin, C
Yöntem:

D'nin Sınıf etiketlerini Normal ve Saldırı ile değiştirin;
D ve Rastgele Orman kullanarak M modelini türetin;
Sınıflandırın, x_{new} M **kullanarak:**
$c = M(x_{new})$; // M tarafından ikili sınıf tahmini
$c ==$ *normal* **ise o zaman**
  dönüş c;
başka
  $D_{attack} = \square$;
  her x **için**$_i \in D$ **do**
    eğer $x_i \in$ *attack* **ise o zaman**
      $D_{attack} = D_{attack} \cup x_i$;
    end if
  için son
  D_{attack} **ve Random Forest kullanarak** M_{attack} **modelini türetin**;
  M_{attack} **kullanarak sınıflandırın**, x_{new} :
  $c = M_{attack}(x_{new})$; // M_{attack} tarafından tahmin
  C = {DOS, *U* 2R, R2L, Prob}
  her k **için**$_i \in C$ **do**
    if $k_i == c$ **then**
      her x **için**$_i \in D_{attack}$ **do**
        $x_i \in k_i$ **ise**
          $D_{new} = D_{new} \cup x_i$;
        end if
      için son
    end if
  için son
  D_{new} **ve Random For- est kullanarak** M_{new} **modelini türetin**;
  $c = M_{new}(x_{new})$; // M tarafından tahmin$_{new}$
  dönüş c
end if

3. Uluslararası Bilgi Keşfi ve Veri Madenciliği Araçları Yarışması [18], [17]. Bu veri kümesinde, her veri noktası bir sınıfın özellik değerlerini gösterir ve her sınıf normal veya saldırı olarak etiketlenir ve tüm örnekler aşağıda belirtilen beş türden biriyle etiketlenir:

Normal: Normal bağlantılar genellikle günlük normal kullanıcı davranışlarıdır ve web sayfalarını ziyaret ederek, dosya yükleyerek, dosya indirerek vb. oluşturulur [17].

Hizmet Reddi: DoS saldırısı, bir cihazın bellek kaynaklarını veya gücünü uygun ağ taleplerini karşılayamayacak kadar dolu veya meşgul hale getirir. Hizmet reddi, sistemleri aşırı yüklemek amacıyla hedeflenen kaynağın gereksiz taleplerle doldurulmasıyla gerçekleştirilir [11].

Uzaktan Kullanıcıya: R2L saldırısında, izole kullanıcı internet üzerinden bir cihaza paketler aktararak yerel bir hesaba erişim sağlar, xlock, xnsnoop, guest, sendmail dictionary, phf vb. içerir [17].

Kullanıcıdan Köke: U2R saldırısında bilgisayar korsanı normal bir kullanıcı hesabıyla sisteme erişir ve kök kullanıcı erişimi elde etmek için güvenlik açıkları bulmaya çalışır. Yaygın örnek Kullanıcıdan köke saldırılar perl, tampon taşmaları, fd-format, load-module, ffb-config ve xterm'dir [18].

Problama: Prob saldırısında bilgisayar korsanı, bilinen güvenlik açıklarını bulmak veya bilgi toplamak için bir ağ cihazını veya bir makineyi tarar [13].

KDD99 veri seti, dört ana kategoriye ayrılan 22 farklı saldırı türüne ve toplam 41 girdi özniteliğine sahiptir ve ağ bağlantıları üç gruba ayrılmıştır ve ayrık veya sürekli değerlere sahiptir [18]. Her TCP bağlantısının temel öznitelikleri ilk grupta yer alır, prototip, süre, kaynak IP adreslerinin bayt miktarı veya hedef IP adresleri, TCP bağlantılarındaki bazı bayraklar ve hizmeti içerir. Alan bilgisi tarafından önerilen ağ bağlantılarının özellikleri ikinci grupta yer alırken, üçüncü grupta iki saniyelik bir zaman dilimi kullanılarak hesaplanan istatistiksel öznitelikler bulunmaktadır. Sınıflandırıcıları eğitmek için eğitim verilerini kullandık ve UCI makine öğrenimi havuzundan test verileri sağlayarak sınıflandırıcı modellerini değerlendirdik. Test verilerinde, eğitim verilerinde bulunmayan bazı bilinmeyen saldırı örnekleri bulunmaktadır. Eğitim ve test örneklerinin sayısı Tablo I'de gösterilmektedir.

TABLO I: KDD99 veri kümesindeki örnek sayısı.

Saldırı Türleri	Eğitim	Test
Normal	972781	60593
R2L	1126	7015
U2R	52	39
DoS	3883370	223298
Sondalama	41102	2377
Toplam	4898431	293322

B. Sonuç

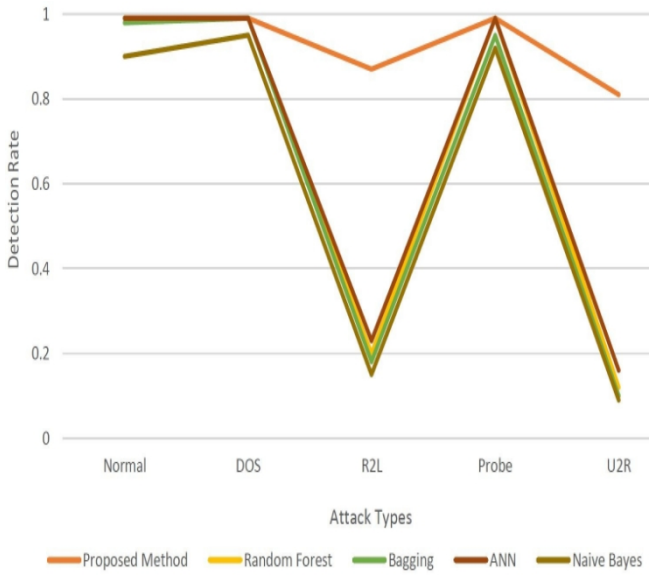
Önerilen yöntemi Python 3.7'de uyguladık ve bilimsel Python geliştirme ortamını (Spyder) kullandık 3.3.1 (https://www.spyder-ide.org). Random Forest, Bagging, na¨ıve Bayes ve ANN gibi bazı popüler makine öğrenimi algoritmalarının performansını KDD99 veri kümesi üzerinde deneyimledik. Önerilen yöntemimizi tespit oranı ve yanlış pozitif oranı kullanarak değerlendirmek için test verileri sağladık. DR, çoğunlukla R2L ve U2R saldırı türlerinde olan azınlık sınıfı örnekleri için çok düşüktür. Random Forest, ANN, Bagging, na¨ıve Bayes gibi bazı güçlü sınıflandırıcılar azınlık sınıfı örneklerini sınıflandırmada başarısız olmuştur. Bizim önerdiğimiz yöntemde ise önemli bir gelişme kaydedilmiştir ve azınlık sınıfı örnekleri için DR %81'in üzerindedir. Önerilen yöntem, R2L ve U2R saldırı türleri için tespit oranlarını artırmakta ve bazı popüler algoritmaların %30'un altında tespit oranına ulaşabildiği durumlarda %87 ve %81 tespit oranına ulaşmaktadır. Yanlış pozitif oranındaki iyileşme de önerilen hibrit yöntemimiz için dikkat çekicidir. Sonuçların karşılaştırılması Tablo II'de ve Önerilen Yöntem, Rastgele Orman, Torbalama, YSA ve na¨ıve Bayes'in tespit oranları Şekil 2'de gösterilmiştir.

V. SONUÇLAR VE GELECEK ÇALIŞMALAR

Son zamanlarda, bilgi güvenliği önemli ölçüde gelişti ve bilgi teknolojisinde hayati bir endişe haline geldi. IDS'ler, hesaplama zekasını kullanarak bilgileri ağ veya ana bilgisayar tabanlı saldırılardan korumak için kullanılmıştır. Ancak, günümüzün erişilebilir IDS'leri kötüye kullanım olarak da bilinen örüntü tabanlıdır.

TABLO II: Farklı güçlü sınıflandırıcıların önerilen yöntemle performans karşılaştırması.

Yöntem	Normal	Sonda	DoS	U2R	R2L
Önerilen Yöntem (DR %)	**0.99**	**0.99**	**0.99**	**0.81**	**0.87**
Önerilen Yöntem (FR %)	**0.03**	0.02	0.02	**0.01**	**0.01**
Rastgele Orman (DR %)	0.99	0.99	0.99	0.12	0.20
Rastgele Orman (FR %)	0.10	0.02	**0.01**	0.01	0.01
Torbalama (DR %)	0.98	0.95	0.99	0.10	0.18
Torbalama (FR %)	0.12	**0.01**	0.02	0.02	0.03
NB Sınıflandırıcı (DR %)	0.90	0.92	0.95	0.09	0.15
NB Sınıflandırıcı (FR %)	0.11	0.03	0.01	0.02	0.01
YSA (DR %)	0.99	0.99	0.99	0.16	0.23
YSA (FR %)	0.98	0.95	0.99	0.84	0.77



Şekil 2: Önerilen yöntem, Rastgele Orman, Torbalama, YSA ve NB sınıflandırıcılarının tespit oranları.

bilinmeyen izinsiz girişleri tespit edememektedir. Bu makalede, Rastgele Orman sınıflandırıcısı ile küme tabanlı alt örnekleme kullanarak azınlık sınıfı ağ saldırılarını / izinsiz girişleri sınıflandırmak için algılama oranını artırmak için yeni bir yöntem sunduk. Önerilen yöntem, bilinen veya bilinmeyen ağ izinsiz girişlerini doğru bir şekilde tanımlamak için oldukça dengesiz büyük verileri işleyebilen çok katmanlı bir sınıflandırma yaklaşımıdır. Başlangıçta, önerilen yöntem bir veri noktasını / gelen veriyi saldırı / izinsiz giriş olup olmadığını (normal davranış gibi) sınıflandırır, eğer bir saldırı ise, önerilen yöntem saldırı türünü ve daha sonra alt saldırı türünü sınıflandırmaya çalışır. Sınıf dengesizliği sorunuyla başa çıkmak için küme tabanlı eksik örnekleme tekniğini ve aşırı uyum sorununu ele almak için popüler topluluk sınıflandırıcısı Random Forest'ı kullandık. Bu makalenin temel amacı, dengesiz ağ saldırı tespit sınıflandırmasında düşük frekanslı saldırıların tespit doğruluğunu artırmaktır. Önerilen yöntemin performansını standart veri madenciliği algoritmaları ile karşılaştırdık. KDD99 benchmark veri kümesi üzerindeki deneysel sonuçlar, önerilen hibrit yöntemin tespit oranlarını artırdığını ve yanlış pozitif oranlarını azalttığını göstermektedir. Gelecekteki çalışmalar, dengesiz veri kümesinde düşük frekanslı saldırıların tespit oranlarını (DR) artırmaya ve bu hibrit yöntemi gerçek dünya NIDS'lerine uygulamaya odaklanacaktır.

REFERANSLAR

[1] R. Singh, H. Kumar ve R. K. Singla, "An intrusion detection system using network traffic profiling and online sequential extreme learning machine," *Expert Systems with Applications,* vol. 42, no. 22, pp. 8609-8624, 2015.

[2] W. Wang, J. Liu, G. Pitsilis, and X. Zhang, "Abstracting massive data for lightweight intrusion detection in computer networks," *Information Sciences,* vol. 433, pp. 417-430, 2018.

[3] D. M. Farid, L. Zhang, C. M. Rahman, M. Hossain, and R. Strachan, "Hybrid decision tree and nave bayes classifiers for multi-class classification tasks," *Expert Systems with Applications*, vol. 41, pp. 1937-1946, March 2014.

[4] P. Garcia-Teodoro, J. Diaz-Verdejo, G. Macia´-Ferna´ndez ve E. Va´zquez, "Anomali tabanlı ağ saldırı tespiti: Techniques, systems and challenges," *Computers & Security,* vol. 28, no. 1-2, pp. 18-28, 2009.

[5] M. S. Pervez ve D. M. Farid, "Feature selection and intrusion classification in nsl kdd cup 99 dataset employing svms," *8th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), Dhaka, Bangladesh,* pp. 1-6, December 2014.

[6] D. M. Farid, N. H. Hoa, J. Darmont, N. Harbi ve M. Z. Rahman, "Scaling up detection rates and reducing false positives in intrusion detection using nbtree," *International Conference on Data Mining and Knowledge Engineering (ICDMKE), Roma, İtalya,* s. 186-190, Nisan 2010.

[7] D. M. Farid ve M. Z. Rahman, "Learning intrusion detection based *on* adaptive bayesian algorithm," *11th International Conference on Computer and Information Technology (ICCIT), Khulna, Bangladesh*, pp. 652-656, Aralık 2008.

[8] D. M. Farid, M. Z. Rahman, ve C. M. Rahman, "Bölüm başlığı: Mining complex network data for adaptive intrusion detection," in *Advances in Data Mining Knowledge Discovery and Applications,* Eylül 2012, ch. 15, pp. 327-348.

[9] S. B. Y. H. Longjie Li, Yang Yu ve X. Chen, "An effective two-step intrusion detection approach based on binary classification and k-nn," *IEEE Access,* vol. 6, pp. 12 060-12 073, 2018.

[10] D. M. Farid ve M. Z. Rahman, "Anomaly detection model for network intrusion detection using conditional probabilities," *6th International Conference on Information Technology in Asia (CITA), Temmuz, 2009, Kuching, Sarawak, Malezya,* s. 104-110.

[11] D. M. Farid, L. Zhang, A. Hossain, C. M. Rahman, R. Strachan, G. Sex- ton, and K. Dahal, "An adaptive ensemble classifier for mining concept drifting data streams," *Expert Systems with Applications*, vol. 40, no. 15,
pp. 5895-5906, Kasım 2013.

[12] D. M. Farid, M. A. Al-Mamun, B. Manderick ve A. Nowe, "An adaptive rule-based classifier for mining big biological data," *Expert Systems with Applications,* cilt. 64, s. 305-316, Aralık 2016.

[13] D. M. Farid, A. Nowe ve B. Manderick, "Ensemble of trees for classifying high-dimensional imbalanced genomic data," in *Proceedings of SAI Intelligent Systems Conference*. Springer, Eylül 2016, s. 172-187.

[14] D. M. Farid, C. M. Rahman, and M. Z. Rahman, "Adaptive intrusion detection based on boosting and na¨ıve bayesian classifier," *International Journal of Computer Applications,* vol. 24, no. 3, pp. 12-19, 2011.

[15] D. M. Farid, C. M. Rahman, N. Harbi, E. Bahri, ve M. Z. Rahman, "Attacks classification in adaptive intrusion detection using decision tree," *International Conference on Computer Science (ICCS), Rio De Janeiro, Brazil,* pp. 86-90, March 2010.

[16] B. B. Zarpela˜o, R. S. Miani, C. T. Kawakani, ve S. C. de Alvarenga, "A survey of intrusion detection in internet of things," *Journal of Network and Computer Applications,* vol. 84, pp. 25-37, 2017.

[17] D. Farid, J. Darmont, N. Harbi, H. H. Nguyen, ve M. Z. Rahman, "Adaptive network intrusion detection learning: attribute selection and classification," *International Conference on Computer Systems Engineering (ICCSE), Bangkok, Thailand,* pp. 82-86, December 2009.

[18] D. M. Farid, N. Harbi, and M. Z. Rahman, "Combining naive bayes and decision tree for adaptive intrusion detection," *International Journal of Network Security & Its Applications (IJNSA),* vol. 2, no. 2, pp. 12-25, April 2010.