Saldırı Tespit Sistemlerinde K-Means Algoritması ve Silhouette Metriği ile Optimum Küme Sayısının Belirlenmesi

Araştırma Makalesi/Research Article

Fatih TOPALOĞLU

Bilgisayar Mühendisliği, Malatya Turgut Özal Üniversitesi, Malatya, Türkiye
fattih.topaloglu@ozal.edu.tr
(Geliş/Received:31.12.2023; Kabul/Accepted:12.02.2024)
DOI: 10.17671/gazibtd.1412641

***Özet*—** Günümüz internetleri neredeyse yarım milyon farklı ağdan oluşmaktadır. Bir ağ bağlantısında, saldırıları türlerine göre tanımlamak zordur. Çünkü farklı saldırılar çeşitli bağlantılara sahip olabilir ve sayıları birkaç ağ bağlantısından yüzlerce ağ bağlantısına kadar değişebilmektedir. Bu nedenden dolayı saldırı tespiti için kullanılan veri setlerinin doğru sınıflandırılması zorlaşmaktadır. Geçmişte pek çok araştırmacı, farklı yöntemler kullanarak davetsiz misafirleri tespit etmek için saldırı tespit sistemleri geliştirmiştir. Ancak mevcut yöntemlerin tespit doğruluğu ve zaman kaybı açısından bazı dezavantajları bulunmaktadır. Çalışmanın temel motivasyonu, saldırı tespit sistemlerinde yüksek boyutluluğun getirdiği zorlukların üstesinden gelmek ve sınıflandırma performansını geliştirmek, sonuçta izinsiz girişlerin daha doğru ve verimli tespitini sağlamaktır. Çalışmada KDD Cup'99 saldırı tespiti veri setinin k-means kümeleme algoritması ile farklı k değerlerine göre analiz edilmesi ve silhouette metriği ile optimum küme sayısının belirlenmesi amaçlanmıştır. Çalışmada farklı k değerleri için yapılan analizlerde, k=10'a kadar olası her konfigürasyon için silhouette skoru hesaplanmıştır. Bu metriğe göre en iyi küme sayısı 4 ve silhouette skoru 0.83 olarak bulunmuştur. Ayrıca silhouette grafiği kalınlıkları ile küme boyutları görselleştirilmiştir.

***Anahtar Kelimeler*—** saldırı tespit sistemleri, k-means, silhouette metriği

Determining the Optimum Number of Clusters with K-Means Algorithm and Silhouette Metric in Intrusion Detection Systems

***Abstract*—** Today's internet consists of almost half a million different networks. In a network connection, it is difficult to identify attacks by type. Because different attacks can have various connections and their number can vary from a few network connections to hundreds of network connections. For this reason, it becomes difficult to correctly classify the data sets used for attack detection. The main motivation of the study is to overcome the challenges of high dimensionality in intrusion detection systems and improve classification performance, ultimately providing more accurate and efficient detection of intrusions. In the past, many researchers have developed intrusion detection systems to detect intruders using different methods. However, existing methods have some disadvantages in terms of detection accuracy and time loss. In the study, it was aimed to analyze the KDD Cup'99 attack detection data set according to different k values with the k-means clustering algorithm and to determine the optimum number of clusters with the silhouette metric. In the analysis carried out for different k values in the study, the silhouette score was calculated for each possible configuration up to k = 10. According to this metric, the best number of clusters was found to be 4 and the silhouette score was 0.83. Additionally, silhouette graphic thicknesses and cluster sizes are visualized.

***Keywords*—** intrusion detection system, k-means, silhouette metric

1. GİRİŞ (INTRODUCTION)

Bilgisayar ağ sistemlerinde kurumsal veya kişisel bilgi güvenliğini sağlamak amacıyla birçok araç veya yazılım kullanılmaktadır [1]. Bu amaçla kullanılan en önemli araçlardan biri saldırı tespit sistemleridir. Saldırı tespit sistemleri, izinsiz girişleri önlemek ve bilgisayar sistemlerine yasa dışı erişimi engellemek için tasarlanmıştır. Saldırı tespit sistemleri, bir kuruluşun bilgisayar ağındaki içsel ve dışsal saldırıları sınıflandırabilir ve güvenlik ihlali varsa alarmı tetikleyebilir [2]. İzinsiz giriş tespit sistemlerinin temel amacı, şimdiye kadar tanımlanmış veya tanımlanamayan saldırılar olmak üzere izinsiz girişleri tanımak, bu saldırıları keşfetmek, bunlara uyum sağlamak ve izinsiz girişleri hızlı bir şekilde tespit etmektir [3].

Saldırı tespit sistemleri için kesinlik ve kararlılık iki önemli metriktir [4] ve son yıllarda, bu önlemleri geliştirmek için birçok çalışma yapılmıştır [5]. Başlangıçtaki çalışmaların çoğu kural tabanlı uzman sistem veya istatistiksel yaklaşıma odaklanmıştır. Ancak çeşitli performans sonuçları, bu yaklaşımların büyük veri kümelerine uygulandığında doğru ve kesin olmadığını göstermektedir [6].

Bu sorunu çözmek için veri madenciliği yaklaşımları [7,8] ve makine öğrenme teknikleri tanıtıldı [9]. Grafik tabanlı yöntemler [10], Doğrusal Genetik Programlama [11], Bayes Ağı [12], k- NN [13], K- Means kümeleme [14], Gizli Markov Modeli [15] vb. saldırı tespit sistemi mimarisi için araştırılmıştır. Makine öğrenimi [16], veri setinde bulunan özellikler ve sınıflar arasındaki korelasyonu tespit edebilir ve özellik seçimi ve boyut azaltma yoluyla alt kümeleri tanımlayabilir, ardından tahminleri gerçekleştirmek üzere bir model oluşturmak için verileri kullanabilir.

Denetimli öğrenme sınıflandırma ve regresyon ile ilişkilendirilen en yaygın öğrenme türüyken, denetimsiz öğrenmede bu ilişkinin önemli bir kısmı kümelemedir. Bu iki öğrenme yöntemi arasındaki temel fark kullanılan veri türüdür. Veriler sınıflandırmada kategorik bir etiketle veya regresyonda sayısal bir değerle etiketlenirken kümelemede etiketlenmez. Bu ise gerçekleştirmeyi ve değerlendirmeyi zor ve karmaşık bir görev yapmaktadır. Bunun için kümeleme algoritmalarının performansının doğru şekilde ölçülmesi çok önemlidir.

Denetimli algoritmalar, doğruluk, R^2 değeri, duyarlılık, özgüllük vb. gibi birçok performans değerlendirme metriğine sahiptir. Bu noktada kümeleme tekniğinin doğruluğunu veya performansını ölçmek için en önemli metrik Silhouette katsayısı veya Silhouette puanıdır. Özellikle yüksek boyutlu veriler analiz edilirken Silhouette puanına ek olarak kümeleme algoritmasının çalışmasını doğrulamak için görselleştirmenin kullanılmasına imkan verir. Yapılan çalışma ile KDD Cup'99 saldırı tespiti veri seti üzerinde k-means kümeleme algoritması ile farklı k değerlerine göre Silhouette metriği ile optimum küme sayısının belirlenmesi amaçlanmıştır.

Bu çalışmanın ana katkıları şu şekilde özetlenebilir:

1) Saldırı tespit sistemleri için K-means algoritması temelli yaklaşım önerilmiştir.
2) Saldırı tespit sistemlerindeki gibi büyük boyutlu veri setleri için Silhouette Puanı performans metriği kullanılmış ve test edilmiştir.
3) Önerilen metrik optimum küme sayısının belirlenmesi için daha kesin puan ve k sayısını vermektedir.

Bu çalışmanın organizasyonu şu şekildedir: Bölüm 2, ilgili çalışmaları anlatmaktadır. Bölüm 3, veri setini ve Siluet Puanı yöntemini tanıtmaktadır. Bölüm 4, önerilen metot ve deney sonuçlarını açıklamaktadır. Bölüm 5, çalışma özetlenmiştir.

2. LİTERATÜR TARAMASI (LITERATURE REVIEW)

Saldırı tespit sistemleri için:

Arif ve ark. [17] hibrit bir yaklaşım tanıtmıştır. Bu yaklaşımda, düğümün budanması parçacık sürü optimizasyonu tarafından gerçekleştirilir ve ağ tabanlı saldırı tespit sistemlerinde sınıflandırma için budanmış karar ağaçları kullanılmıştır. Ahmed ve ark. [18], boyutluluğun azaltılması için naive bayes özellik alt küme seçici tekniğinin uygulandığı hibrit bir saldırı tespit sistemi oluşturmak için üçlü bir strateji geliştirmiştir. Aykırı değerlerin reddi için optimize edilmiş destek vektör makinesi uygulanırken, sınıflandırıcı olarak öncelikli k-nn uygulanmıştır. Dash ve ark. [19], yerçekimsel arama ile parçacık sürüsü optimizasyon algoritmalarının birleşimi olan (GSPSO) dizisi olan iki yeni hibrit saldırı tespit yöntemi önermiştir. Yao ve ark. [20] saldırı tespit sistemleri için hibrit bir model önermiştir. K-means kümeleme algoritması ve sınıflandırma aşamasında tamamı denetimli öğrenme algoritmaları olan destek vektör makinesi, yapay sinir ağları, karar ağaçları ve rastgele orman farklı parametreler üzerinden karşılaştırılmıştır.

Suad ve Fadl [21], makine öğrenimi algoritmasını büyük veri ortamına uygulayan bir saldırı tespit sistemi modeli tanıttı. Bu çalışmada Spark-Chi-SVM modeli kullanılmaktadır. Ijaz ve ark. [22] vektörlere dayalı bir genetik algoritma tanıtmışlardır. Bu teknikte vektör kromozomları uygulanmıştır. Önerilen yaklaşımın benzersizliği, kromozomları bir vektör olarak ve eğitim verilerini metrik olarak göstermesidir. Alauthaman ve ark. [23], karar ağaçlarına yardımcı olmak üzere ileri beslemeli bir sinir ağı üzerine kurulu, eşler arası bot tespitine yönelik bir yaklaşım önerdi. Venkataraman ve Selvaraj [24] verilerin sınıflandırılması için etkili bir hibrit özellik seçim yapısı önermiştir. Sınıflandırma amacıyla ilgili özellikleri bulmak için simetrik belirsizlik uygulanır. Çalışmada

genetik algoritma alt kümeleri daha yüksek doğrulukla aramak için kullanılmıştır.

Kumar ve Kumar [25] akıllı tabanlı bir hibrit model önermiştir. Bu model daha sonra çok katmanlı algıyı, bulanık mantık denetleyiciyi, uyarlanabilir nöro-bulanık girişim sistemini ve bir nöro-bulanık genetiği entegre etmiştir. Çavuşoğlu ve diğerleri[26] makine öğrenimi teknikleri temelli hibrit bir yaklaşım önermişlerdir. Sınıflandırma amacıyla k- nn ve naive bayes algoritmaları kullanılırken, sınıflandırıcı olarak rastgele orman algoritması kullanılmıştır. Saxena ve ark. [27], yüksek kaliteli özellik alt kümelerini elde etmek amacıyla DBSCAN tabanlı bir hibrit teknik önermişlerdir. Verilerdeki gürültüyü ortadan kaldırmak için DBSCAN, verilerin gruplandırılması için k-means kümeleme kullanılmıştır. Kar ve ark. [28], verilerin sınıflandırılması için uygulamada ID3 adı verilen karar ağaçları algoritmasını kullanır. Sınıf etiketlerini, keşfedilmemiş veri noktasına en yakın noktaya atamak için k- nn yaklaşımı uygulanır. Baykara ve Daş [29] saldırı tespit sistemleri için honeypot tabanlı bir yaklaşım önermektedir. Geliştirilen honeypot sunucu uygulaması sunuculardaki ağ trafiğini gerçek zamanlı animasyonla görsel olarak gösterebilmektedir.

Dutta ve ark. [30] saldırı tespit sistemlerindeki sınıflandırma ölçümlerini geliştirmek için hibrit bir model önermiştir ve sınıflandırma doğruluğunu arttırmak için derin bir sinir ağı uygulanmıştır. Latah ve Toker [31] akış tabanlı çok seviyeli hibrit saldırı tespit sistemi tanıtmıştır. Yazar, sınıflandırma amacıyla k-nn ve aşırı öğrenme makinelerini uygulamış ve özellik seçme yöntemi olarak yazılım tanımlı ağ denetleyicisi kullanılmıştır. Sumaiya Thaseen ve ark. [32] korelasyon tabanlı özellik seçimi, en iyi özellik alt kümelerini seçmek için bir özellik seçme yaklaşımı olarak uygulanırken, yapay sinir ağı bir sınıflandırıcı olarak kullanılır. Safaldin ve ark. [33], saldırı tespit sınıflandırması için bir özellik seçme yöntemi ve destek vektör makinesi olarak geliştirilmiş ikili gri kurt optimizerini uygulamıştır. Vallathan ve ark. [34] IoT ortamında derin öğrenme yaklaşımını temel alan şüpheci eylem tespit sistemini önermiştir. Baykara ve Daş [35] web uygulamalarının güvenliği için hibrit bir gerçek zamanlı saldırı ve önleme sistemi yaklaşımı önermişlerdir. Önerilen sistem, kural tabanlı suistimal tespiti ve anormallik tespitini kullanmakta ve veri kaynağı olarak ağ paketlerini kullanmaktadır.

Ishaque ve ark. [36] saldırı tespit sistemi için bulanık mantık, sinir ağları ve genetik algoritma kullanan yeni bir hibrit teknik önermişlerdir. Çalışmada belirsizliğin giderilmesi sürecinde bulanık mantıktan, tahmin amacıyla ise sinir ağlarından yararlanılmıştır. Tahmin sonuçlarının doğruluğunda iyileştirmeler sağlamak amacıyla genetik algoritma kullanılmıştır. Nabi ve Zhou [37] saldırı tespit sistemlerinin geliştirilmesinde makine öğrenimi tekniklerinin karşılaştırıldığı bir çalışma sunulmuştur. Test edilen sınıflandırıcılar arasında J48 ağacı en yüksek doğruluğu sağlamıştır. Sınıflandırıcı performansını geliştirmek için Rastgele Projeksiyon ve PCA projeksiyon yaklaşımları karşılaştırılmıştır. Aljehane ve ark. [38] ağ güvenliği için derin öğrenme destekli saldırı tespit sistemi (GJOADL-IDSNS) tekniği ile yeni bir altın çakal optimizasyon algoritması önermişlerdir. GJOADL-IDSNS sisteminin ana amacı, ağ güvenliğini sağlamak için izinsiz girişlerin etkili bir şekilde tanınması ve sınıflandırılmasıdır. Fraihat ve ark. [39] büyük ölçekli IoT NetFlow tabanlı ağlar için bir güvenlik önlemi olarak bir ağ saldırı tespit sistemi önermişlerdir. Önerilen NIDS, en uygun özellik kümesini belirlemek için aritmetik optimizasyon algoritmasının değiştirilmiş bir versiyonuyla desteklenen makine öğrenimini kullanmıştır. Seçilen yedi özellik, Rastgele Orman ve Ekstra Ağaçlar da dahil olmak üzere çeşitli ML modellerini eğitmek için kullanılmıştır.

Pramilarani ve Kumari [40] etkili bir saldırı tespit sistemi geliştirmek için maliyet tabanlı rastgele orman sınıflandırıcısı (CRFC) önermişlerdir. CRFC tabanlı sınıflandırma, özellik dengesizliği olsa bile özellikleri bölme sürecini iyileştirmeye yardımcı olan, özelliğin önemine göre hesaplanan maliyet matrisinin dahil edilmesiyle doğaçlama yapmışlardır. Al Nuaimi ve ark. [41] Edge-IIoT-2022 veri kümesini kullanarak veri odaklı saldırı tespit sistemlerini IoT ve endüstriyel IoT ortamlarında 6 adet makine öğrenme algoritması ile değerlendirmişlerdir. Sun ve ark. [42] STS sınıflandırmak ve tespit etmek için parçacık sürüsü optimizasyonunu ve AdaBoost algoritmalarını birleştiren bir sistem önermektedir. Deneysel sonuçlar, PSO-AdaBoost yaklaşımının izinsiz giriş tespitinde üstün doğruluk, kesinlik ve hatırlama sağladığını göstermektedir. Korium ve ark. [43] makine öğrenimine dayalı bir saldırı tespit sistemi önermişlerdir. Modelde veri ön işleme için veri dağıtımını koruyan Z-puanı normalleştirmesi, model karmaşıklığını basitleştiren ve yürütme süresini azaltan bir regresyon modeli, model seçimi ve eğitimi için rastgele orman, ekstremgradyan arttırma, kategorik arttırma, hafif gradyan arttırma makinesi ve eğitim aşamasındaki davranışı kontrol etmek ve aşırı uyumu önlemek için hiperparametre optimizasyonu kullanılmıştır.

3. MATERYAL VE METOT (MATERİAL AND METHOD)

Silhouette metriği, kümeleme performansının değerlendirilmesine yardımcı olan bir ölçümdür. Kümeleme kalitesinin değerlendirilmesi, kümeleme algoritmalarının etkinliğini ve güvenilirliğini belirlemek için önemlidir. Kümeleme denetimsiz bir öğrenme görevi olduğundan kümeleri doğrulamak için net etiketler yoktur. Bu nedenle kümeleme sonuçlarının değerlendirilmesi, Silhouette skoru gibi dahili doğrulama metriklerinin kullanılmasını gerektirir.

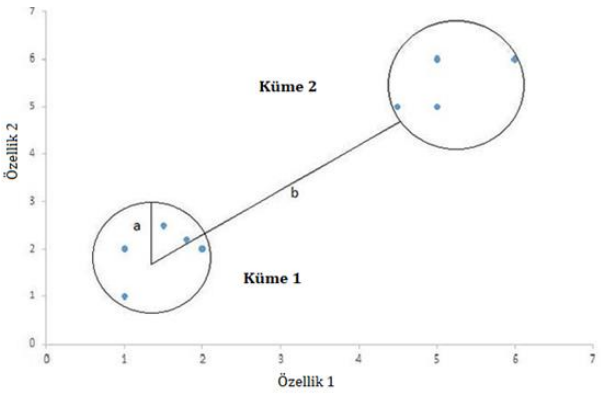
3.1. Veri Seti (Data Set)

Uygulamada KDD Cup'99 veri kümesi kullanılmıştır. KDD Cup'99 veri seti, DARPA98 veri setinden birkaç niteliğin çıkartılmasıyla elde edilmiştir. Veri kümesi 34 sürekli, 7 sembolik olmak üzere 41 öznitelik değerli ve 4.898.430 adet veri içermektedir. KDD99 DARPA, saldırı

tespiti için makine öğrenmesi ve veri madenciliği tabanlı çalışmalarda en çok kullanılan veri setidir. Tavallaee ve ark. [44], çoğunluğunu index dergilerden oluşan tarama çalışması kapsamında incelediği 276 makalenin % 75'inde KDD99 DARPA veri setinin kullanıldığını tespit edilmiştir.

3.2. Silhouette Metriği (Silhouette Metric)

Silhouette puanı ve silhouette grafiği, kümeler arasındaki ayırma mesafesini ölçmek için kullanılır. Bir kümedeki her bir noktanın komşu kümelerdeki noktalara ne kadar yakın olduğunun bir ölçüsünü görüntüler. Bu ölçüm [-1, 1] aralığına sahiptir ve kümeler içindeki benzerlikleri ve kümeler arasındaki farklılıkları görsel olarak da sunmaktadır. Şekil 1'de silhouette küme gösterimi sunulmuştur.



Şekil 1. Silhouette küme gösterimi
(Silhouette cluster representation)

a: ortalama küme içi mesafe, yani bir küme içindeki her nokta arasındaki ortalama mesafe.

b: kümeler arası ortalama mesafe, yani tüm kümeler arasındaki ortalama mesafe.

Bir veri noktası **i** için silhouette puanı şu şekilde verilir:

$$S_i = \frac{b_i - a_i}{\max(b_i, a_i)} \quad (1)$$

bi : bir parçası olması dışında, i veri noktasının en yakın kümesine olan ortalama mesafe olarak tanımlanan kümeler arası mesafedir.

$$b_i = \min \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j) \quad (2)$$

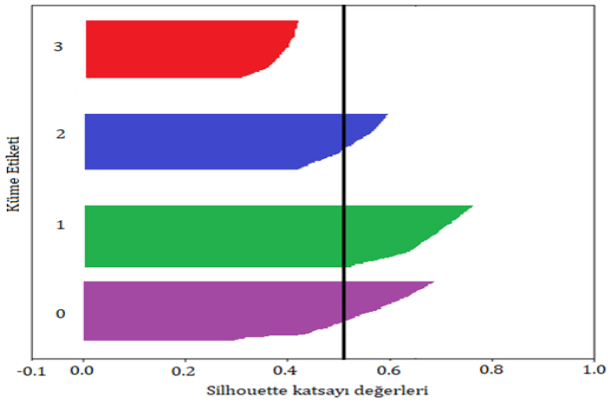
ai : parçası olduğu kümedeki diğer tüm noktalara olan ortalama mesafe olarak tanımlanan küme içi mesafedir.

$$a_i = \frac{1}{|C_i|-1} \sum_{j \in C_i, i \neq j} d(i,j) \quad (3)$$

3.3. Grafiksel Özet (Grafical Abstract)

Tüm veri kümesinin genel silhouette puanı, veri kümesindeki tüm veri noktalarının siluet puanının ortalaması olarak hesaplanabilir. Formülden görülebileceği gibi silhouette puanı her zaman [-1, 1] arasında olacaktır.

Şekil 2'de görüldüğü gibi silhouette grafikleri, y ekseninde küme etiketini temsil ederken, x ekseninde gerçek silhouette puanını temsil eder. Silhouette'lerin boyutu/kalınlığı da o küme içindeki örneklerin sayısıyla orantılıdır.

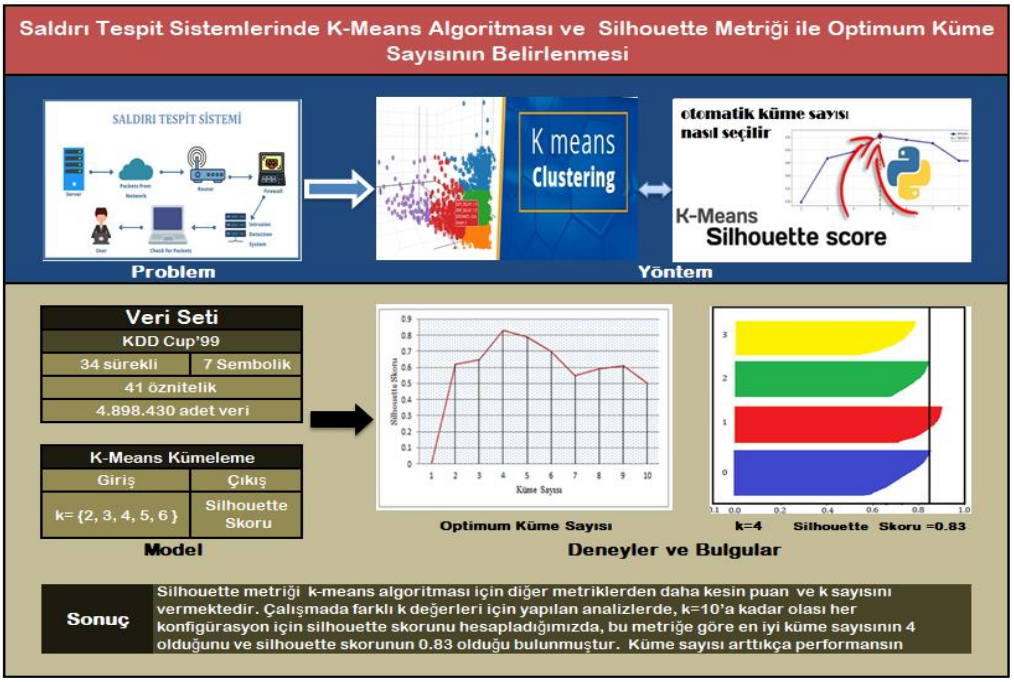


Şekil 2. Silhouette grafik gösterimi
(Silhouette graphic representation)

Silhouette katsayıları +1'e ne kadar yakınsa, kümenin örnekleri komşu küme örneklerinden o kadar uzaktadır. 0 değeri, örneklerin iki komşu küme arasındaki karar sınırında veya çok yakınında olduğunu gösterir. Negatif değerler ise bu örneklerin yanlış kümeye atanmış olabileceğini gösterir. Silhouette katsayılarının ortalamasının alınması tüm kümenin performansınım tek bir silhouette puanı ile hesaplanabilmesini sağlar. Tablo 1'de silhouette metriği algoritması sunulmuştur.

Tablo 1. Silhouette metriği algoritması
(Silhouette metric algorithm)

```
FUNCTION [index] = SILHOUETTE [D]
1     N=size of D;
2     sum_All=0;
3     foreachclass j=1 to c;
4        M= size of Dj;
5        sum_Class=0;
6        for i=1 to M
7           b=b(i);
8           a=a(i);
9           s=(b-a) / max(b,a);
10          sum_Class+=s;
11       endfor
12       sum_All+=sum_Class / M;
13    endfor
14    index=sum_All / c;
```

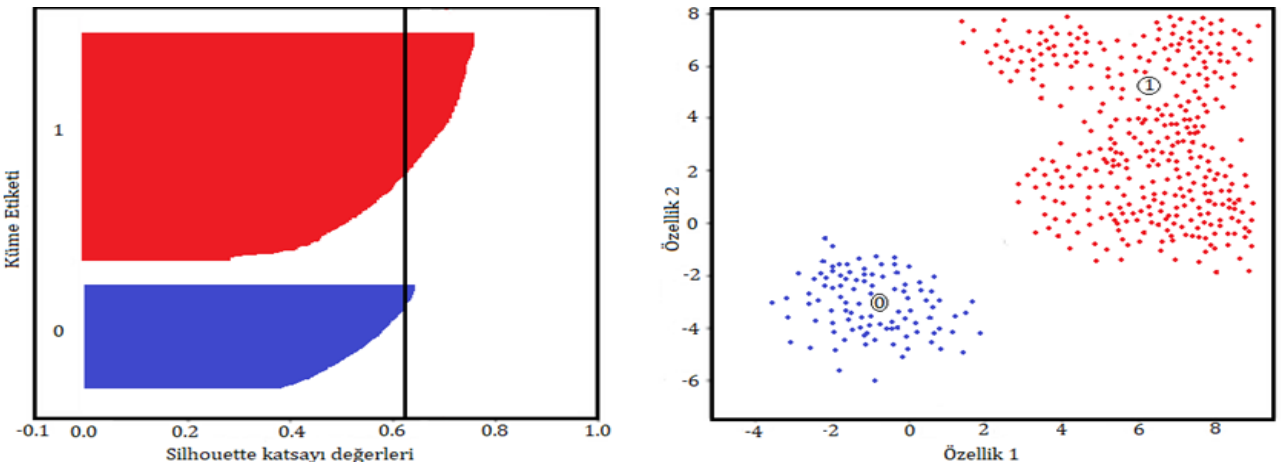


Şekil 3. Grafiksel Özet
(Grafical Abstract)

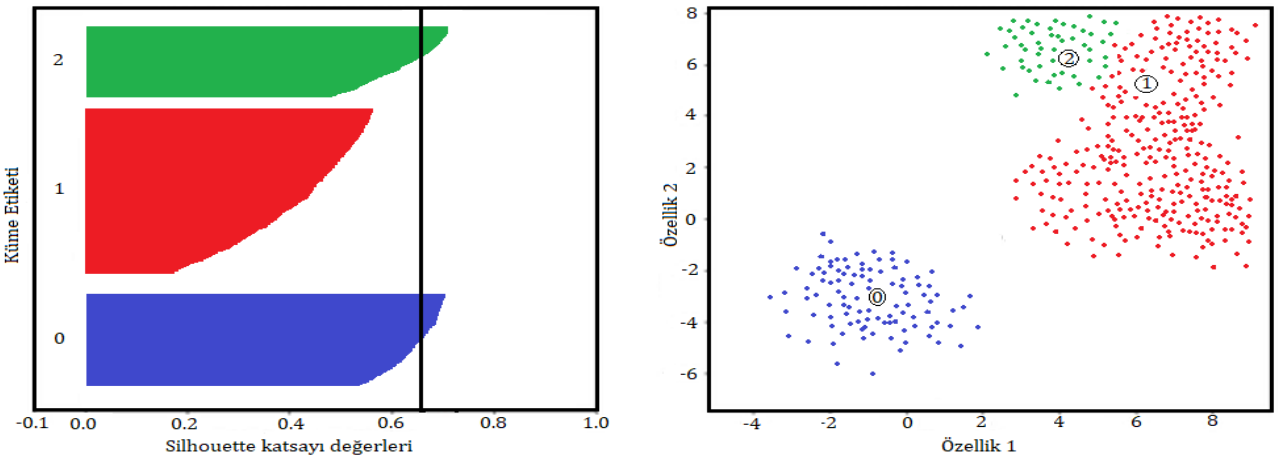
4. UYGULAMA VE SONUÇLARI (APPLİCATİON AND RESULTS)

Uygulama kapsamında ,KDD Cup'99 saldırı tespiti veri seti üzerinde k-means kümeleme algoritması ile farklı k değerlerine göre silhouette metriği ile optimum küme sayısının belirlenmesi amaçlanmıştır. Bunun için 5 farklı k parametresi sırasıyla 2, 3, 4, 5, 6 değerleri analiz edilmiştir. Her bir konfigürasyonun performansına bakılarak silhouette skoru ve silhouette grafiği ile optimum küme sayısını bulması sağlanmıştır.

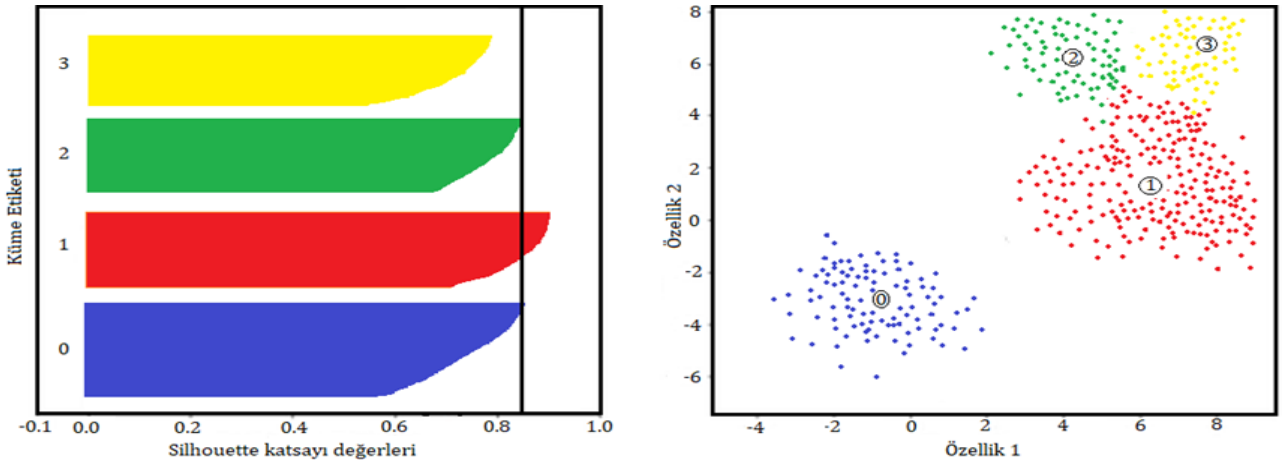
Veriler üzerinde k-means kümelemesi k= {2, 3, 4, 5, 6} değerleri için elde edilen silhouette skorları ve silhouette grafikleri Şekil 4, Şekil 5, Şekil 6, Şekil 7 ve Şekil 8'daki gibidir. Her konfigürasyonun performansına bakılarak silhouette grafiği en iyi sayıda kümenin bulunmasını sağlamaktadır. Şekil 4'de popülasyonu ayırmak için iki küme anlamına gelen k=2'yi kullanarak ortalama 0,62 silhouette puanı elde edilmiştir. Küme sayısı üçe çıkarıldığında ortalama silhouette puanı bir miktar artmıştır. Ayrıca her kümede daha az örnek olması nedeniyle küme sayısı arttıkça silhouette grafiğinin kalınlığının azalmaya devam ettiği de görülmektedir.



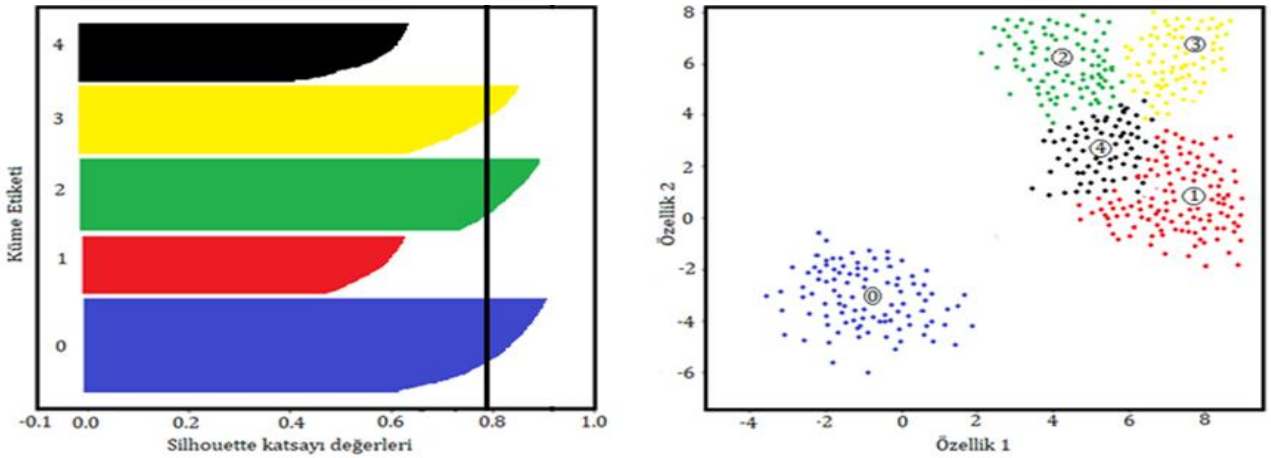
Şekil 4. k=2 için silhouette skoru ve grafiği
(Silhouette score and graph for k=2)



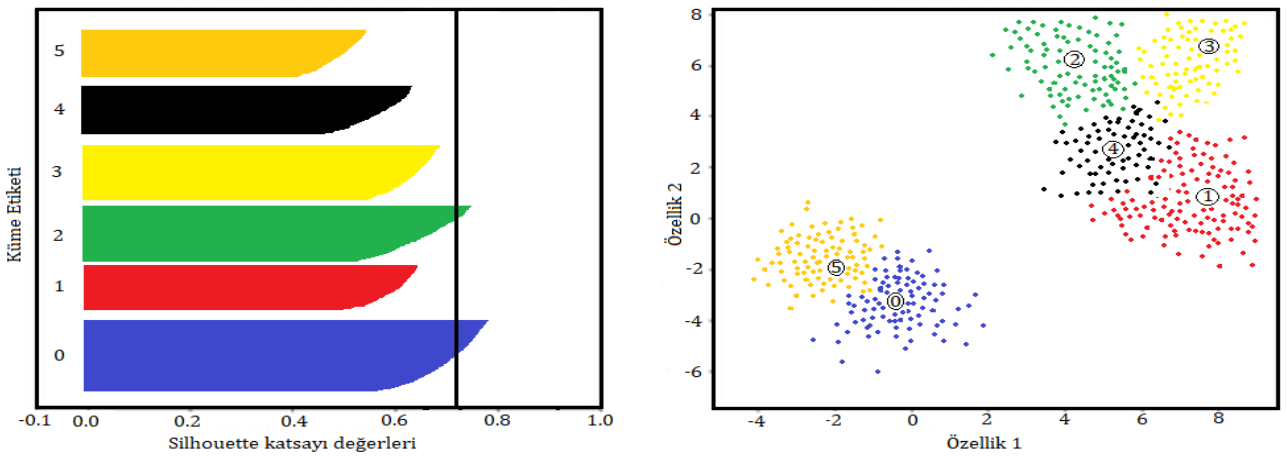
Şekil 5. k=3 için silhouette skoru ve grafiği
(Silhouette score and graph for k=3)



Şekil 6. k=4 için silhouette skoru ve grafiği
(Silhouette score and graph for k=4)



Şekil 7. k=5 için silhouette skoru ve grafiği
(Silhouette score and graph for k=5)



Şekil 8. k=6 için silhouette skoru ve grafiği
(Silhouette score and graph for k=6)

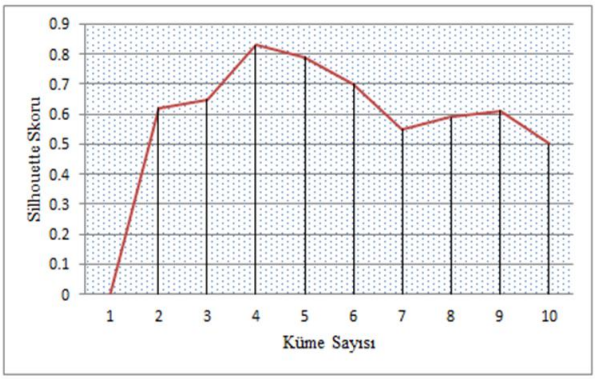
Çalışmada k küme sayısına göre bulunan silhouette puanları Tablo 2'de verilmiştir. Silhouette metriği, en uygun değeri seçmek için kullanılır. Siluet grafiği, ortalamanın altında siluet puanlarına sahip kümelerin varlığı ve ayrıca siluet çizimlerinin boyutunda geniş dalgalanmalar nedeniyle 2, 3 ve 6 değerinin verilen veriler için kötü bir seçim olduğunu 4 ile 5 'in karar vermede daha iyi bir seçim olduğu, en optimum seçimin ise 4 değerinin olduğunu göstermektedir.

Tablo 2. k- means için silhouette analiz sonuçları
(Silhouette analysis results for k- means)

k	Silhouette Skoru
2	**0.62**23615745161018
3	**0.66**28246153550858
4	**0.83**44216225076757
5	**0.79**38500642344114
6	**0.70**37404360466380

5. SONUÇLAR (RESULTS)

Silhouette, veri kümeleri içindeki tutarlılığın yorumlanması ve doğrulanması yöntemini ifade eder. Bu teknik, her bir nesnenin ne kadar iyi sınıflandırıldığına ilişkin kısa ve öz bir grafiksel gösterim sağlar. Çalışmada saldırı tespit sistemleri için en yaygın kullanılan KDD Cup'99 veri seti k-means kümeleme algoritması ile farklı k değerlerine göre analiz edilip silhouette puanlarının ve grafiklerinin bulunması böylelikle optimum küme sayısının belirlenmesi sağlanmıştır. Böylelikle saldırı tespiti için kullanılan veri setlerinin doğru küme değerleri ile sınıflandırılması sağlanmıştır.



Şekil 9. Optimal küme sayısı
(Optimal number of clusters)

Çalışmada farklı k değerleri için yapılan analizlerde, k=10'a kadar olası her konfigürasyon için Silhouette Skorunu hesapladığımızda, Şekil 9'da görüldüğü gibi bu metriğe göre en iyi küme sayısının 4 olduğunu ve küme sayısı arttıkça performansın kötüleştiğini görebiliriz. Ayrıca siluet grafiğinin kalınlığından küme boyutu görselleştirilebilir. Küme sayısı arttıkça her kümede daha az örnek olması nedeniyle silhouette kalınlığının azalmaya devam ettiği de fark edilebilir.

K-means kümelemesi, basit ve popüler bir denetimsiz makine öğrenme algoritmasıdır. Silhouette metriği bize k-means algoritması için diğer metriklerden daha kesin puan ve k sayısını vermektedir. Çalışmada farklı k değerleri için yapılan analizlerde, k=10'a kadar olası her konfigürasyon için silhouette skorunu hesapladığımızda, bu metriğe göre en iyi küme sayısının 4 olduğunu ve silhouette skorunun 0.83 olduğu bulunmuştur. Küme sayısı arttıkça performansın düştüğü ve silhouette kalınlığının azaldığı görülmüştür.

Kümeleme performans metrikleri olarak, silhouette puanı, rand endeksi, düzeltilmiş rand endeksi, karşılıklı bilgi, Calinski-Harabasz endeksi ve Davies-Bouldin endeksi kullanılmaktadır. Çalışmada silhouette performans metriğine odaklanılmıştır. Gelecekte, kümeleme yönteminden daha doğru sonuçlar elde

edilmesi ve daha geniş bir uygulama kapsamına sahip olması için diğer metriklerin performans analizleri gerçekleştirilebilir. Özellikle yüksek boyutlu veri setlerinin sınıflandırılmasındaki zorluklar nedeni ile çalışma KDD Cup'99 veri kümesi üzerinde sınanmıştır. Ancak meteoroloji, enerji ve sağlık gibi büyük boyutlu veri kümelerinde önerilen yöntemin daha kesin ve başarılı sınıflandırma için kullanılabileceği düşünülmektedir.

KAYNAKLAR (REFERENCES)

[1] M. Baykara, R. Daş, "SoftSwitch: a centralized honeypot-based security approach using software-defined switching for secure management of VLAN networks," *Turkish Journal of Electrical Engineering and Computer Sciences*, Vol. 27, no. 5, pp. 3309-3325, 2019.

[2] L. Hung-Jen, C.-h. R. Lin, "Intrusion detection system a comprehensive review", *Journal of network and applications, vol*. 36, no. 1, pp. 16–24, 2013.

[3] H. L. Motoda, H. Motoda, Feature Selection for Knowledge Discovery and Data Mining, vol. 454, Springer, 1998.

[4] L. D. S. Silva, A. C. Santos, T. D. Mancilha, J. D. Silva, A. Montes, "Detecting attack signatures in the real network traffic with ANNIDA", Expert Systems with Applications, vol. 34, no. 4, pp. 2326–2333, 2008.

[5] A. Patcha, J. M. Park, "An overview of anomaly detection techniques: existing solutions and latest technological trends", Computer Networks, vol. 51, no. 12, pp. 3448–3470, 2007.

[6] C. Manikopoulos, S. Papavassiliou, "Network intrusion and fault detection. A statistical anomaly approach," *IEEE Communications Magazine*, vol. 40, no. 10, pp. 76–82, 2002.

[7] P. Fournier-Viger, C. W. Lin, A. Gomariz et al., "The SPMF open-source data mining library version 2", Joint European conference on machine learning and knowledge discovery in databases, pp. 36–40, Riva del Garda, Italy, 2016.

[8] P. Fournier-Viger, J. C.-W. Lin, R. U. Kiran, Y. S. Koh, R. Thomas, "A survey of sequential pattern mining", Data Science and Pattern Recognition, vol. 1, no. 1, pp. 54–77, 2017.

[9] A. Smola, S.V.N. Vishwanathan, Introduction to Machine Learning, Cambridge University Press, ISBN-10: 0521825830, 2008.

[10] Z. Xiaojin, Semi-Supervised Learning Literature Survey, vol. 2, *Computer Science, University of Wisconsin, Madison*, 2008.

[11] S. Mukkamala, A. H. Sung, A. Abraham, "Modeling intrusion detection systems using linear genetic programming approach," in The 17th international conference on industrial & engineering applications of artificial intelligence and expert systems, innovations in applied artificial intelligence, pp. 633–642, Berlin, Heidelberg, 2004.

[12] J. Pearl, "Bayesian networks. A model of self-activated memory for evidential reasoning," in Proceedings of the 7th Conference of the Cognitive Science Society, University of California, pp. 329–334, Irvine, CA, 2009.

[13] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression (PDF)," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

[14] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297, University of California Press, 1967.

[15] L. E. Baum, T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," The annals of mathematical statistics, vol. 37, no. 6, pp. 1554–1563, 1966.

[16] M. Mohammed, M. B. Khan, E. B. Bashier, **Machine Learning Algorithms and Applications**, CRC press Taylor and Francis Group, ISBN-10: 1498705383, 2016.

[17] J. Arif, F. Malik, K. Aslam, "A hybrid technique using binary particle swarm optimization and decision tree pruning for network intrusion detection", *Cluster Computing*, vol. 21, pp. 667–680, 2017.

[18] I. Ahmed, L. Saleh, M. Fatma, L. Talaat, "A hybrid intrusion detection system (HIDS) based on prioritized k-nearest neighbors and optimized SVM classifiers", *Artificial Intelligence Review*, vol. 51, pp. 403–443, 2017.

[19] D. Tirtharaj, "A study on intrusion detection using neural networks trained with evolutionary algorithms", *Soft Computing, vol*. 21, pp. 2687–2700, 2017.

[20] Y. Haipeng, W. Qiyi, "An intrusion detection framework based on hybrid multi-level data mining," *International Journal of Parallel Programming*, vol. 47, pp. 740–758, 2017.

[21] M. Suad, M. Fadl, "Intrusion detection model using machine learning algorithm on Big Data environment", *Journal of big data, vol*. 5, pp. 1–12, 2018.

[22] S. Ijaz, F. A. Hashmi, S. Asghar, M. Alam, "Vector based genetic algorithm to optimize predictive analysis in network security", *Applied intelligence, vol*. 48, no. 5, pp. 1086–1096, 2018.

[23] A. Mohammad, A. Nauman, "A P2P Botnet detection scheme based on decision tree and adaptive multilayer neural networks", *Neural Computing & Applications, vol*. 29, pp. 991–1004, 2018.

[24] V. Sivakumar, S. Rajalakshmi, "Optimal and novel hybrid feature selection framework for effective data classification," *in Advances in Systems, Control and Automation*, pp. 499–514, Springer, Singapore, 2018.

[25] K. Neeraj, K. Upendra, "Knowledge computational intelligence in network intrusion detection systems", *Knowledge Computing and Its Applications, pp*. 161–176, Springer, Singapore, 2018.

[26] C. Unal, "A new hybrid approach for intrusion detection using machine learning methods", *Applied Intelligence*, vol. 49, pp. 2735–2761, 2019.

[27] S. Akash, S. Khushboo, "Hybrid technique based on DBSCAN for selection of improved features for intrusion detection system", *in Emerging Trends in Expert Applications and Security, pp*. 365–377, Springer, Singapore, 2019.

[28] P. Kar, S. Banerjee, K. C. Mondal, G. Mahapatra, S. Chattopadhyay, "A hybrid intrusion detection system for hierarchical filtration of anomalies", *Information and Communication Technology for Intelligent Systems*, vol. 106,

pp. 417–426, Springer, Singapore, 2019.

[29] M. Baykara, R. Daş, " A novel honeypot based security approach for real-time intrusion detection and prevention systems," *Journal of Information Security and Applications* (JISA), Vol.41, pp. 103-116, 2018.

[30] V. Dutta, M. Choras, R. Kozik, M. Pawlicki, "Hybrid model for improving the classification effectiveness on network intrusion detection system**", in Conference on Complex, Intelligent, and Software Intensive Systems**, Cham, 2020.

[31] M. Latah, L. Toker, "An efficient flow-based multi-level hybrid intrusion detection system for software-defined networks", *CCF Transactions on Networking*, vol. 3, pp. 26–271, 2020.

[32] I. Sumaiya Thaseen, J. Saira Banu, K. Lavanya, M. Rukunuddin Ghalib, K. Abhishek, "An integrated intrusion detection system using correlation-based attribute selection and artificial neural network", *Transactions on Emerging Telecommunications Technologies, vol*. 32, no. 2, article e4014, 2021.

[33] M. Safaldin, M. Qtair, L. Abualigah, "Improved binary gray wolf optimizer and SVM for intrusion detection system in wireless sensor networks", *Journal of Ambient Intelligence and Humanized Computing, vol*. 12, no. 2, pp. 1559–1576, 2021.

[34] G. Vallathan, A. John, C. Thirumalai, "Suspicious activity detection using deep learning in secure assisted living IoT environments*", The Journal of Supercomputing*, vol. 77, pp. 3242–3260, 2021.

[35] M. Baykara, R. Daş, " A Novel Hybrid Approach for Detection of WebBased Attacks in Intrusion Detection Systems," *International Journal of Computer Networks and Applications* (IJCNA,) Vol.4, no. 2, pp. 62-76, 2017.

[36] M. Ishaque, Md G. Md Johar, A. Khatibi, M. Yamin, "A novel hybrid technique using fuzzy logic, neural networks and genetic algorithm for intrusion detection system," *Measurement: Sensors*, Vol.30, pp. 1-12 ,2023.

[37] F. Nabi, X. Zhou, "Enhancing intrusion detection systems through dimensionality reduction: A comparative study of machine learning techniques for cyber security", Cyber Security and Applications, Vol.2, pp. 1-8, 2024.

[38] N. O. Aljehane, H. A. Mengash, M. M. Eltahir, F. A. Alotaibi, S. S. Aljameel, A. Yafoz, R. Alsini, M. Assiri, "Golden jackal optimization algorithm with deep learning assisted intrusion detection system for network security", *Alexandria Engineering Journal*, Vol.86, pp. 415-424, 2024.

[39] S. Fraihat, S. Makhadmeh, M. Awad, M. A. Al-Betar, A. Al-Redhaei, "Intrusion detection system for large-scale IoT NetFlow networks using machine learning with modified Arithmetic Optimization Algorithm", *Internet of Things*, Vol. 22, pp. 1-22, 2023.

[40] K. Pramilarani, P. V. Kumari, "Cost based Random Forest Classifier for Intrusion Detection System in Internet of Things", *Applied Soft Computing*, Vol. 151, pp. 1-8, 2024.

[41] T. Al Nuaimi, S. Al Zaabi, M. Alyilieli, M. AlMaskari, S. Alblooshi, F. Alhabsi, M. F. Bin Yusof, A. Al Badawi, "A comparative evaluation of intrusion detection systems on the edge-IIoT-2022 dataset", *Intelligent Systems with Applications*, Vol.20, pp. 1-10, 2023.

[42] Z. Sun, G. An, Y. Yang, Y. Liu, "Optimized machine learning enabled intrusion detection 2 system for internet of medical things*", Franklin Open*, Vol.6, pp. 1-11, 2024.

[43] M. S. Korium, M. Saber, A. Beattie, A. Narayanan, S. Sahoo, P. H.J. Nardelli, "Intrusion detection system for cyberattacks in the Internet of Vehicles environment", Ad Hoc Networks, Vol. 153, pp. 1-16, 2024.

[44] M. Tavallaee, N. Stakhanova, A. A. Ghorbani, "Toward Credible Evaluation of Anomaly-Based Intrusion-Detection Methods", Systems, Man, and Cybernetics, Part C: *Applications and Reviews, IEEE Transactions on*, vol. 40, no. 5, pp. 516-524, 2010.