Düzce Üniversitesi Bilim ve Teknoloji Dergisi

Araştırma Makalesi

Çok Katmanlı Algılayıcı ile Ağ Trafiği Sınıflandırma Analizi

Kurban KOTAN [a], Bayram KOTAN [a], Serdar KIRIŞOĞLU [b]

[a] *Elekt.-Elektro. ve Bil. Mühendisliği ABD, Lisansüstü Eğitim Enstitüsü, Düzce Üniversitesi, Düzce, TÜRKİYE*
[b] *Bilgisayar Mühendisliği, Mühendislik Fakültesi, Düzce Üniversitesi, Düzce, TÜRKİYE*
** Sorumlu yazarın e-posta adresi: serdarkirisoglu@duzce.edu.tr*
DOI: 10.29130/dubited.980594

ÖZ

Çevrimiçi ağ trafiği sınıflandırması, uzun vadeli ilginin odak noktası olmaya devam ediyor. Ağ trafiğini izleme ve ağ trafiği analizi birçok farklı yoldan yapılabilir. Ağ trafiğini izleme, hizmet kalitesi (QoS) için ham veri girişi sağlar ve bu da ağ analistine ağ kaynaklarını nasıl kullandığını anlama ve ağ performansını belirleme olanağı sağlar. Bu bilgi ile ağ analisti, ağ kaynaklarını kontrol etmek ve yönetmek için QoS politikalarını belirleyebilir. Ağ trafiğinin izlenmesi akademik araştırma için modeller oluşturmak için de kullanılabilir. Bu makalede derin öğrenme algoritması kullanılarak ağ trafiğini doğru şekilde sınıflandıran bir makine öğrenme yaklaşımı sunulmuştur. Aynı zamanda bu çalışmada diğer makine öğrenme algoritmaları ile karşılaştırmalar yapılmıştır. Çok Katmanlı Algılayıcı (MLP), ağın sınıflandırıcısını oluşturmak için kullanılmıştır. Deney sonuçları derin öğrenme algoritmasının diğer algoritmalardan daha iyi sonuç verdiğini ve sınıflandırmada %99,0233 Detection Rate (DR) değerine, %78,3941 doğruluğa (ACC) sahip olduğunu göstermiştir.

Anahtar Kelimeler: *Sinirsel Ağ, Makine Öğrenmesi, Yapay Zekâ, Trafik Sınıflandırması, KDD CUP99 veri seti.*

Network Traffic Classification Analysis with Multi-Layer Sensor

ABSTRACT

Online network traffic classification remains the focus of long-term interest. Network traffic monitoring and network traffic analysis can be done in many different ways. Monitoring network traffic provides raw data input for quality of service (QoS), which gives the network analyst the ability to understand how it uses network resources and determine network performance. With this information, the network analyst can set QoS policies to control and manage network resources. Network traffic monitoring can also be used to build models for academic research. In this article, a machine learning approach that correctly classifies network traffic using a deep learning algorithm is presented. At the same time, comparisons were made with other machine learning algorithms in this study. Multi-Layer Perceptron (MLP) was used to construct the classifier of the network. Experiment results showed that the deep learning algorithm gave better results than other algorithms and had a Detection Rate (DR) value of 99.0233% and an accuracy of 78.3941% (ACC) in classification.

Keywords: *Neural Network, Machine learning, Artificial Intelligence, Traffic Classification, KDD CUP99 dataset*

Geliş: 09/08/2021, Düzeltme: 04/10/2021, Kabul: 02/11/2021

I. GİRİŞ

İnternet, giderek daha büyük miktarda veri ve dijital medya iletişimi içeren ve her gün dünya çapındaki tüm işletmeler için muazzam gelirler sağlayan, büyüyen ve her yerde bulunan ağlara dönüşmektedir. Veri iletimi basit protokollerle yönetilir; İletim Kontrol Protokolü (TCP) ve Kullanıcı Datagram Protokolü (UDP); bunlar işlevselliği olmayan trafik üzerinde izleme, inceleme ve akıllı kontrol olmadan işlem görür [1]. İşletmeler ve hükümetler ağ trafiğini sınıflandırmak ve izlemek, kaynaklarını yönetmek, yatırımlarını ve çıkarlarını korumak ister. Bundan dolayı olası anormallikleri tespit etmek için uygulamalara ihtiyaç duyarlar. Genel olarak, İnternet trafiği, çeşitli ağları, ana bilgisayarları, uygulamaları ve birbirleriyle etkileşime giren farklı istemcileri içeren karmaşık bir sistemin ürünüdür.

Ağ trafiği sınıflandırması (izleme) günümüzde akademik çalışmalarda büyük ilgi görmektedir [2-6]. Trafik akışlarının üretim uygulamalarına göre sınıflandırılması, QoS kontrolü, izinsiz giriş tespiti ve yasal müdahale gibi güvenlik ve ağ yönetiminde çok önemli bir yere sahiptir [7]. Günümüzde milyarlarca cihaz İnternet kaynaklarını kullanmaktadır [8]. Her cihaz, diğer cihazlara bağlantı için talepler gönderir ve internet üzerinden veri alışverişi yapar. Sonuç olarak, büyük miktarda ağ trafiği üretirler, bu nedenle sadece QoS için veya kaynakların kullanılabilirliğini sağlamak için değil, aynı zamanda bilgilerin verimli bir şekilde işlenmesi için de sınıflandırma gereklidir.

Veri örneklerinin elle etiketlenmesi çoğunlukla yorucu, zaman kaybı ve maliyetlidir. Bu karmaşıklık, her gün çok çeşitli ağ uygulamaları üretilmesiyle sürekli artmaktadır. Bu nedenle öğrenebilecek ve uygulayabilecek bir sisteme ihtiyacımız vardır. Bu bağlamda, makine öğrenimini uygulamak daha yararlı olacaktır [9].

Ağ izleme, port tabanlı trafik sınıflandırma yöntemleri, yüke dayalı sınıflandırma yöntemleri (Derin paket incelemesi) ve akış özelliklerine göre sınıflandırma yöntemleri (Makine öğrenimi ve istatistiksel özellik) ile başarılabilir [10]. Trafik sınıflandırmasına olan ilgi arttıkça, birçok sınıflandırma yöntemi bu alana uygulanmıştır [11, 12]. Port tabanlı yöntem, ağ trafiği sınıflandırması için en iyi tekniklerden biri olarak bilinmektedir [13]. Bu yöntem ilk önce İnternet Atanmış Numaralar Kurumu'na (IANA) kayıtlı olan ağ bağlantı noktalarını kullanır. Bununla birlikte, bu yöntem kayıtsız port numarası kullanan ve dinamik bağlantı noktası numaralandırması kullanan Noktadan Noktaya (P2P) uygulamaları yüzünden ağ trafiğini doğru şekilde sınıflandıramamaktadır [2]. Yüke dayalı yöntemler daha iyi sınıflandırma sonuçları verir. Ancak, bu yöntem şifreli yük yani şifreli veriyi barındıran paketler yüzünden ağ trafiğini sınıflandırmamaktadır. Birçok ağ uygulaması verileri korumak için şifreleme kullanmaktadır [14, 15]. Ağ trafiğini izlemek için makine öğrenmesi kullanılarak birçok ağ sınıflandırma yöntemi önerilmiştir. Bu çalışmada Ağ trafiği Derin Öğrenme algoritması ile sınıflandırılmış ve bu alanda çok kullanılan diğer algoritmalar ile performans karşılaştırmaları yapılmıştır. Derin Öğrenme, trafik sınıflandırmasında çok kesin sonuçlar vermektedir [13]. Bu yöntem, bilinmeyen trafik sınıflarını sınıflandırmak için eğitim ve test veri kümelerini kullanır.

Bir ağdaki tüm paketlerin trafiğini aynı anda izlemek kolay değildir [16]. Üst üste gelen protokoller veya protokol katmanlaması, özelliklerin hızlı izlenmesini ve çıkarılmasını zorlaştırır. Bu çalışma, bu gibi zorlukların üstesinden gelmek için, derin öğrenme algoritmasını kullanmak en iyi çözümlerden biri olduğunu ortaya koymuştur.

Bu çalışmanın temel amacı, derin öğrenme algoritmasını ağ trafiği sınıflamasına uygulamak ve sonuçları değerlendirmektir. Bu hedefe ulaşmak için aşağıdaki hedefler göz önünde bulundurulmalıdır:

- Derin öğrenme kullanarak ağ trafiğini sınıflandırmak için mevcut yöntemleri incelemek.
- Tanımlanan taksonomiyi belirlemek ve avantajı ve dezavantajları sağlamak.
- Tanımlanan metodun performansını değerlendirmek ve diğer metotlarla karşılaştırmak.

II. METOT

A. SINIFLANDIRMA

Ağ İzleme aşağıdaki yöntemlerle sağlanabilir [10, 17]:

- Porta dayalı trafik sınıflandırması
- Yüke dayalı sınıflandırma (Derin paket incelemesi)
- Akış özelliklerine dayalı (Makine öğrenimi ve istatistiksel özellik)

Tablo 1 ağ sınıflarını göstermektedir.

Tablo 1. *Ağ Sınıfları [18]*

Ağ Sınıfları	**Örnek Uygulamalar**
BULK	ftp, ftp_data
DATABASE	postgres,sqlnet oracle, ingres
INTERACTIVE	ssh, klogin, rlogin, telnet
MAIL	imap, pop3, smtp
SERVICES	X11, dns, ident, Idap, ntp
WWW	http
P2P	BitTorrent
ATTACK	DoS, Probe
GAMES	Half-Life
MULTIMEDIA	Windows Media Player, Real Time

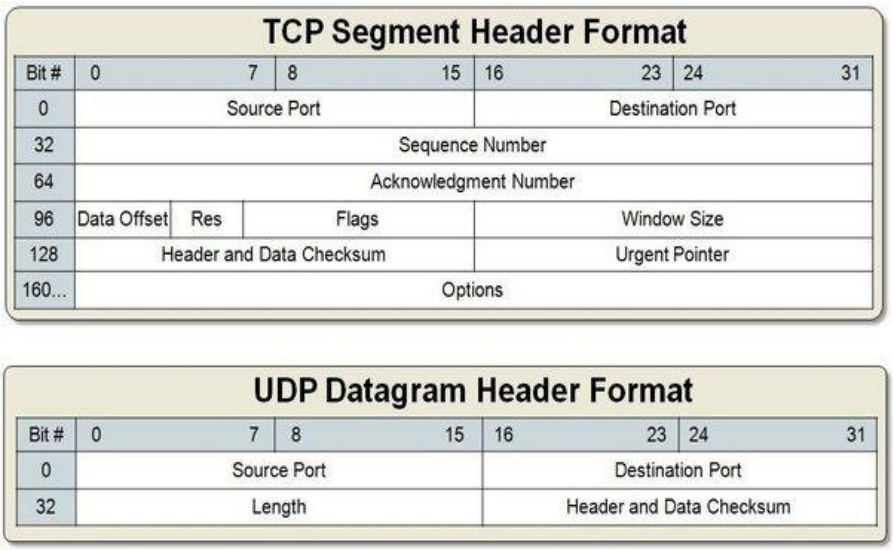
A. 1. Porta Dayalı Sınıflandırma

Bu, trafik sınıflandırmasını gerçekleştirmenin en eski yoludur. Bunun varsayımı, uygulama sunucularının istemcilerin iletişimi başlatması için iyi bilinen bağlantı noktaları kullanmasıdır. Bu tür portlar, IANA'nın kayıtlı portlar listesine kaydedilmiştir [19]:

80: HTTP
22: SSH
20, 21: FTP
25: SMTP
53: DNS
143: IMAP
161, 162: SNMP

Sunucu, uygulamayı anlamak için TCP/UDP paket başlığını okuması yeterlidir. Mesela TCP akışları için, SYN (senkronize) paketi yeterlidir.
Porta dayalı sınıflandırma uygulaması çok basit ve hızlıdır. Yükü denetlemeye gerek yoktur, paket başlıklarını kontrol etmek yeterli olacaktır. Genellikle güvenlik duvarlarında ve erişim kontrol listelerinde kullanılır. Bununla birlikte, çoğu uygulamanın IANA'ya kayıtlı port numaları yoktur. İyi bilinen port noktalarına sahip olsalar bile, başka port numaraları kullanabilirler, örneğin port numarası 80'in arkasına saklanabilirler. Bazı durumlarda port numarası rastgele / dinamik olarak tahsis edilir ve porta dayalı sınıflandırma yöntemleri NAT (Ağ Adresi Çeviricisi) ve IP (İnternet Protokolü) tünellerinde başarısız olur.

TCP Segmenti ve UDP Datagram Başlık Formatı [20] Şekil 1'de gösterilmiştir.



Şekil 1. *TCP Segmenti ve UDP Datagramı Başlık Formatı [20]*

A. 2. Yüke Dayalı Sınıflandırma

Bu, paketlerin TCP veya UDP yüklerini (data) denetleyen yöntemlerdir ve aşağıdaki özellikleri arar:

- Bilinen protokol davranışları (protokol kod çözme)
- Uygulama özel verileri (desen eşleme)

Ayrıca yükün içeriğini incelediklerinde Derin Paket Denetimi (DPI) yöntemleri olarak da adlandırılırlar [21]. Yüke dayalı sınıflandırma, port tabanlı sınıflamanın yapamadığı birçok protokolü tanımlayabilir ve daha yüksek doğruluk (ACC) oranına sahiptir. Yüke dayalı sınıflandırmada işlem için ilk sekiz paket yeterli olacaktır. Trafiği kısa sürede sınıflandırdığı için gerçek zamanlı uygulama da mümkündür. Bu yöntem yükü denetlediğinden şifreli iletişimi sınıflandırmaz. Bu yöntem CPU (Merkezi İşlem Birimi) üzerinde yüksek işlem yükleri oluşturur [22]. Protokol kodu çözme, tüm protokoller hakkında derinlemesine bilgi gerektirdiğinden çok karmaşık bir işlemdir. Sadece çok kullanılan protokol türleri için kullanılır ve bu kod çözücüleri güncel tutmak zordur. Yüke Dayalı yöntemler, sınıflandırmada kesin sonuçlar verir. Bununla birlikte, şifreli veri ağı uygulamaları adı verilen birçok ağ uygulaması, verileri korumak için şifreleme kullanır, bu sebepten yüke dayalı yöntemler şifreli ağ paketlerini sınıflandırmada başarısız olmaktadırlar [2].

A. 3. Akış Özelliğine Dayalı Sınıflandırma

Akış özelliğine dayalı sınıflandırma yöntemleri, yüke dayalı ve porta dayalı sınıflandırma yöntemlerin sorunlarını aşabilir. Protokol/uygulama türünü değerlendirmek için her akışın özelliklerinin istatistiksel özelliklerini kullanır. Bu yöntemler aynı zamanda istatistiksel yöntemler veya makine öğrenme yöntemleri olarak da bilinir. Genel olarak, iki makine öğrenme yöntemi vardır.

A. 3.1. Denetimli Sınıflandırma

Denetimli yöntemlerde, makine "etiketli" veriler kullanılarak eğitilir. Denetimli bir öğrenme algoritması, etiketli eğitim verilerinden yararlanır ve beklenmeyen sonuçların tahmin edilmesine yardımcı olur.

A. 3.2. Denetimsiz Sınıflandırma

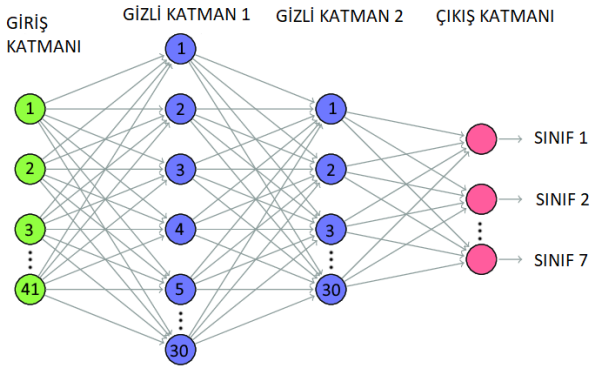
Kümeleme gibi denetlenmeyen yöntemler doğal olarak farklı sınıfları veya hatta yeni uygulamaları ortaya çıkarabilir. Kümelerin etiketlenmesi gerekir, örneğin doğrudan insan tarafından etiketlenebilirler [23].

B. ÖNERİLEN MLP MODELİ

B. 1. KDD CUP 99 Veri Seti:

Askeri bir ağ ortamında simüle edilen ve çok çeşitli müdahaleleri içeren bu veri kümesinde denetlenecek standart veri kümesi vardır. 1999'dan beri, anomali tespit yöntemleri için çılgınca kullanılmaktadır. Bu veri seti, DARPA'98 değerlendirme programında yakalanan verilere dayanılarak oluşturulmuştur. DARPA'98 7 haftalık ağ trafiğinden elde edilmiş olup yaklaşık 5 milyon bağlantı kaydı içerir. Her bağlantı yaklaşık 100 bayttır. Veri kümesi ham ikili TCP dökümü verisidir ve yaklaşık 4GB'dir. KDD Cup99 Beşinci Uluslararası Bilgi Keşfi ve Veri Madenciliği Konferansında Üçüncü Uluslararası Bilgi Keşfi ve Veri Madenciliği Araçları Yarışması düzenlenmiş ve ağ saldırı detektörü için bir veri seti kullanılmıştır. Bu veri seti KDD Cup99'dur. Ağ izinsiz giriş dedektörü izinsiz girişleri (veya saldırıları) öngören ve kötü bağlantıları etiketleyerek normal bağlantıları öngörüp normal olarak etiketledikleri şekilde yapılmıştır. Askeri bir ağ ortamında simüle edilen ve çok çeşitli izinsiz girişleri içeren bu veri setinde denetlenecek standart veri kümeleri vardır. 1999'dan beri, anomalilerin tespit yöntemleri için çılgınca kullanıldı. Bu veri seti DARPA'98 değerlendirme programında elde edilen verilere dayanarak oluşturulmuştur. DARPA'98 7 haftalık ağ trafiğinden elde edildi. Ve yaklaşık 5 milyon bağlantı kaydı içeriyor. Her bağlantı yaklaşık 100 bayttır. Veri kümesi ham ikili tcp döküm verisidir ve yaklaşık 4 GB'dir. KDD eğitim veri seti yaklaşık 5 milyon bağlantı vektörü içerir. Normal etiketli veya saldırı etiketli her vektör 41 özellik içerir [24].

KDD CUP99 verisinde bulunan 41 tane özellikten dolayı MLP topolojisi 41 düğümlü giriş katmanı, 30'ar düğümlü 2 gizli katman ve sınıf sayısı (7) kadar düğüm barındıran çıkış katmanından oluşur ve Şekil 2'de gösterilmiştir.

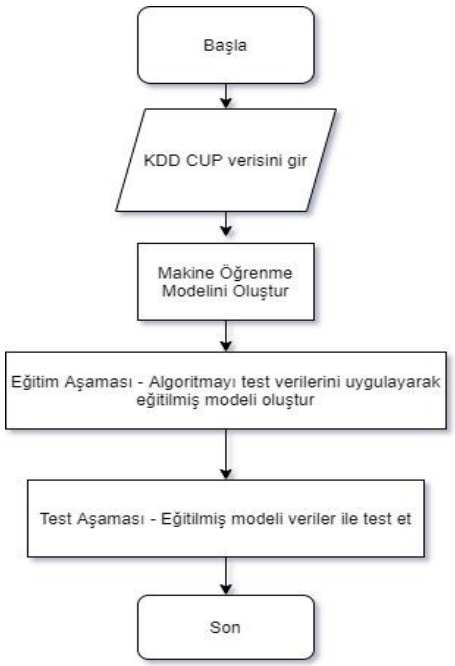


Şekil 2. *Önerilen MLP Modeli*

MLP modeli seri bir model, aktivasyon fonksiyonu softmax, her iki gizli katmanının aktivasyon fonksiyonu relu, loss fonksiyonu categorical_crossentropy, optimizer'ı adam, metrics'i accuracy, epochs 10 ve batch size'ı 5 seçilmiştir.

III. AKIŞ ŞEMASI

Algoritmaların akış şeması Şekil 3'te gösterilmiştir.



Şekil 3. *Akış Aşaması*

C. 1. Ölçütler

İki sınıflı karışıklık matrisi Tablo 2'deki gibidir.

Tablo 2*. İki Sınıflı Karışıklık Matrisi*

		TAHMİN VERİLERİ	
		Anomaly	**Normal**
GERÇEK VERİLER	**Anomaly**	TP	FP
	Normal	FN	TN

C. 1.1. Sensitivity-Detection Rate (DR) – True Positive Rate (TPR)

Duyarlılık, doğru sınıflandırılmış toplam pozitif örnek sayısının toplam pozitif örnek sayısına bölünme oranıdır. Yüksek DR sınıfın doğru tanındığını gösterir [25].

$$DR = TPR = \frac{TP}{P} = \frac{TP}{TP + FN} \quad (1)$$

C. 1.2. Accuracy

Doğruluk, doğru sınıflandırılmış pozitif ve negatif örneklerin toplam sayısının toplam örnek sayısına bölünmesidir [25].

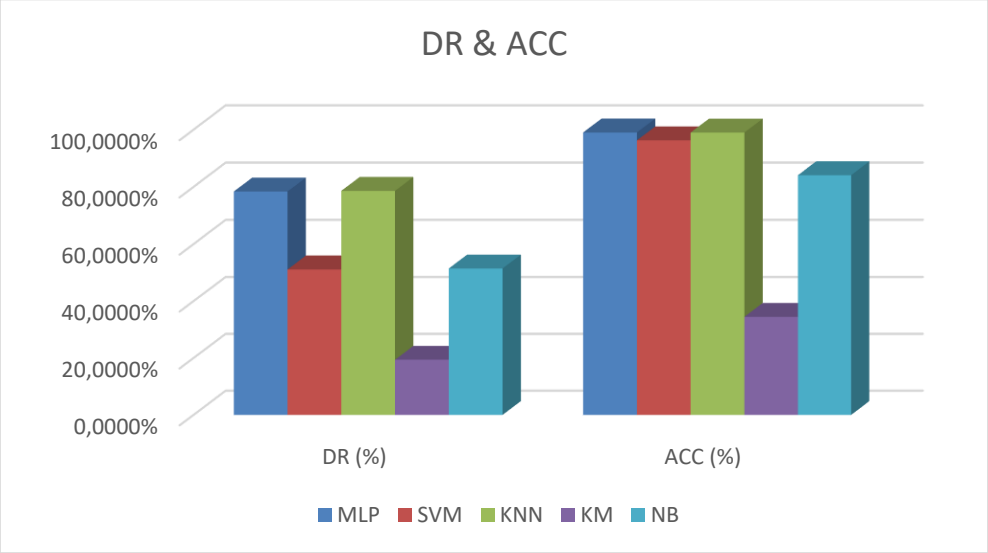
$$ACC = \frac{TP + TN}{Total} \tag{2}$$

D. 1. Deney

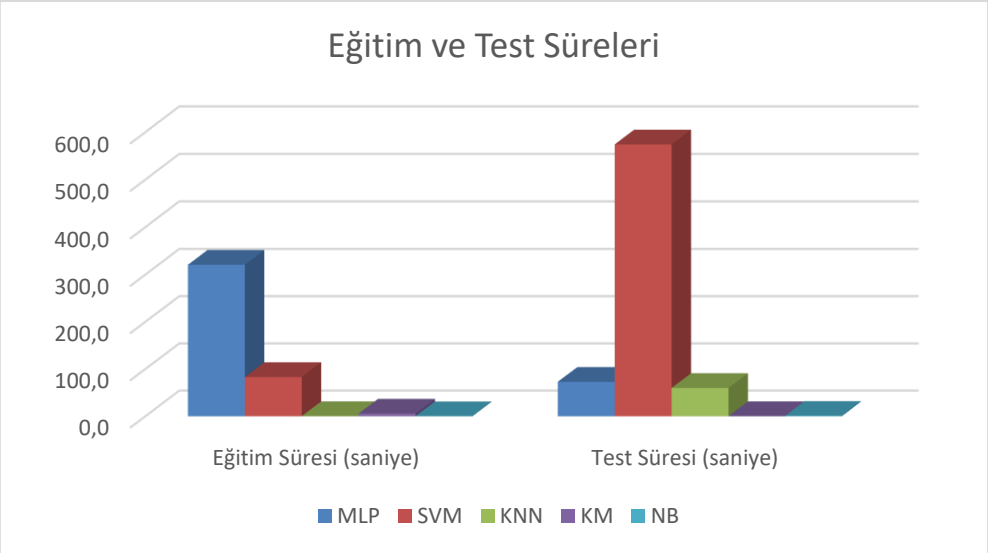
Bu bölümde, MLP, SVM, KNN, KM ve NB algoritmalarını kullanarak ağ trafiğini KDDCUP99 verileri üzerinden sınıflandırmak için deneyler yapılacaktır.

***Tablo 3**. Deney Sonuçları*

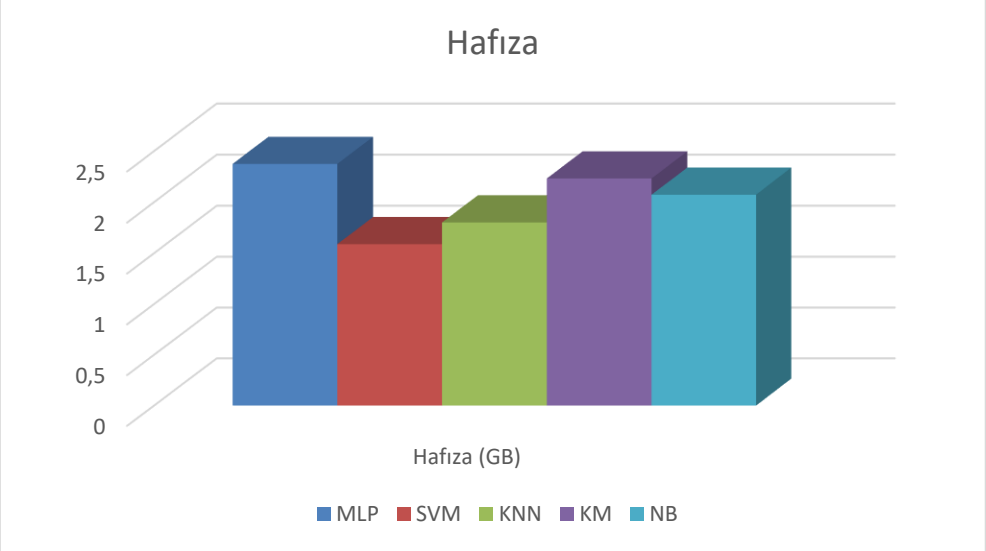
	DENEYLER				
	MLP	**SVM**	**KNN**	**KM**	**NB**
DR (%)	78,3941	51,1479	78,6201	19,4189	51,4787
ACC (%)	99,0233	96,2793	98,9976	36,4961	84,0907
ÇAPRAZ DOĞRULAMANIN ORTALAMASI (%)	98,9200	96,0248	98,1968	4,8519	75,5658
EĞİTİMİ SÜRESİ (SANİYE)	320,818215	83,159853	0,404694	5,799661	0,224237
TEST SÜRESİ (SANİYE)	73,024681	574,283786	60,123494	0,361247	1,064350
HAFIZA(GB)	2,374596	1,588810	1,801605	2,231792	2.073017



***Şekil 4.** DR & ACC Karşılaştırmaları*



Şekil 5. *Eğitim ve Test Sürelerinin Karşılaştırmaları*



Şekil 6. *Bellek Tüketim Karşılaştırmaları*

IV. SONUÇ

Ağ izleme, temel bağlanabilirlikten uygulama çıkışına kadar her türlü ağ işlemine yöneliktir. Bu çalışmanın amacı, ağ paketlerini sınıflandırmaya yardımcı olan MLP kullanan bir Ağ İzleme Sistemi önermektir.

Deney sonuçları önerilen MLP modelinin diğer algoritmalardan daha iyi sonuç verdiğini ve sınıflandırmada %99,0233 DR değerine, %78,3941 doğruluğa (ACC) sahip olduğunu göstermiştir. Diğer tüm algoritmalara kıyasla en iyi DR oranını ve KNN algoritmasından sonra en iyi ACC değerini verir.
Yalnız MLP eğitim süresi olarak en yüksek değere sahip ve test süresi olarak ikinci en büyük değere sahiptir. Hafıza olarak da en yüksek hafıza isteyen algoritmadır.

V. KAYNAKLAR

[1] V. Cerf and R. Kahn, "A Protocol for Packet Network Intercommunication," in *IEEE Transactions on Communications*, vol. 22, no. 5, pp. 637-648, 1974, doi: 10.1109/TCOM.1974.1092259.

[2] T. Karagiannis, A. Broido, M. Faloutsos, and K. Claffy, "Transport layer identification of P2P traffic," in *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, 2004, pp. 121-134.

[3] Y. Wu, G. Min, K. Li, and B. Javadi, "Performance analysis of communication networks in multi-cluster systems under bursty traffic with communication locality," in *GLOBECOM 2009-2009 IEEE Global Telecommunications Conference*, 2009, pp. 1-6.

[4] H. Kim, K. C. Claffy, M. Fomenkov, D. Barman, M. Faloutsos, and K. Lee, "Internet traffic classification demystified: myths, caveats, and the best practices," in *Proceedings of the 2008 ACM CoNEXT conference*, 2008, pp. 1-12.

[5] Y.-s. Lim, H.-c. Kim, J. Jeong, C.-k. Kim, T. T. Kwon, and Y. Choi, "Internet traffic classification demystified: on the sources of the discriminative power," in *Proceedings of the 6th International COnference*, 2010, pp. 1-12.

[6] T. T. Nguyen, G. Armitage, and tutorials, "A survey of techniques for internet traffic classification using machine learning," vol. 10, no. 4, pp. 56-76, 2008.

[7] Y. Xiang, W. Zhou, and M. Guo, "Flexible deterministic packet marking: An IP traceback system to find the real source of attacks," vol. 20, no. 4, pp. 567-580, 2009.

[8] J. Johnson, "Worldwide digital population as of January 2021," 2021.

[9] J. Korteling, G. van de Boer-Visschedijk, R. A. M. Blankendaal, R. C. Boonekamp, and A. R.. Eikelboom, "Human-versus artificial intelligence," *Front. Artif. Intell.*, vol. 4, 2021.

[10] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, 2015, pp. 448-456: PMLR.

[11] T. Auld, A. W. Moore, and S. F. Gull, "Bayesian neural networks for internet traffic classification," *IEEE Transactions on Neural Networks*, vol. 18, no. 1, pp. 223-239, 2007.

[12] M. Crotti, F. Gringoli, P. Pelosato, and L. Salgarelli, "A statistical approach to IP-level classification of network traffic," in *2006 IEEE International Conference on Communications*, 2006, vol. 1, pp. 170-176.

[13] N. Namdev, S. Agrawal, and S. Silkari, "Recent advancement in machine learning based internet traffic classification," *Procedia Computer Science*, vol. 60, pp. 784-791, 2015.

[14] P. Haffner, S. Sen, O. Spatscheck, and D. Wang, "ACAS: automated construction of application signatures," in *Proceedings of the 2005 ACM SIGCOMM workshop on Mining network data*, 2005, pp. 197-202.

[15] S. Sen, O. Spatscheck, and D. Wang, "Accurate, scalable in-network identification of p2p traffic using application signatures," in *Proceedings of the 13th international conference on World Wide Web*, 2004, pp. 512-521.

[16] A. Moore, J. Hall, C. Kreibich, E. Harris, and I. Pratt, "Architecture of a network monitor," in *Passive & Active Measurement Workshop*, 2003, vol. 2003.

[17] W. Li and A. W. Moore, "A machine learning approach for efficient traffic classification," in *2007 15th International symposium on modeling, analysis, and simulation of computer and telecommunication systems*, 2007, pp. 310-317.

[18] P. Schneider, "Tcp/ip traffic classification based on port numbers," 1997.

[19] M. Degermark, B. Nordgren, and S. Pink, *RFC2507: IP header compression*, RFC Editor, 1999.

[20] T. Porter, "The perils of deep packet inspection," 2005.

[21] M. Finsterbusch, C. Richter, E. Rocha, J.-A. Muller, K. J. I. C. S. Hanssgen, and Tutorials, "A survey of payload-based traffic classification approaches," vol. 16, no. 2, pp. 1135-1156, 2013.

[22] G. Hinton and T. J. Sejnowski, *Unsupervised learning: foundations of neural computation*, MIT Press, 1999.

[23] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *2009 IEEE symposium on computational intelligence for security and defense applications*, 2009, pp. 1-6: IEEE.

[24] W. Stallings, *Network security essentials: Applications and standards, 4/e*, Pearson Education India, 2003.