Black Sea Journal of Engineering and Science
doi: 10.34248/bsengineering.1839347

Open Access Journal
e-ISSN: 2619 – 8991

Araştırma Makalesi (Research Article)
Cilt 9 - Sayı 4: 1725-1736 / Temmuz 2026
(Volume 9 - Issue 4: 1725-1736 / July 2026)

TF-IDF VE YAPISAL ÖZELLİKLERE DAYALI URL TABANLI KİMLİK AVI TESPİTİNDE MAKİNE ÖĞRENİMİ VE ANSAMBL YÖNTEMLERİNİN KARŞILAŞTIRMALI ANALİZİ

Ülkü ÖZMEN[1]*, Esra ODABAŞ YILDIRIM[1]*

[1]Atatürk University, Faculty of Engineering, Department of Software Engineering, 25240, Erzurum, Türkiye

Özet: Kimlik avı (phishing) saldırıları, kullanıcıların kimlik ve finansal bilgilerini hedef alan, hızla yaygınlaşan bir siber güvenlik tehdidi dir; bu nedenle saldırıların erken ve düşük maliyetle tespiti kritik önem taşımaktadır. Web sayfası içeriğinin analizine dayalı geleneksel yöntemler yüksek hesaplama maliyeti ve gecikme oluştururken, statik kara liste tabanlı yaklaşımlar yeni oluşturulan veya kısa ömürlü URL'ler karşısında yetersiz kalmaktadır. Bu çalışma, yalnızca URL bilgisinden yararlanan hibrit bir tespit hattı önermektedir: karakter düzeyindeki TF-IDF n-gram temsilleri; URL uzunluğu, sembol yoğunluğu, alt alan adı derinliği ve alan adı/TLD özellikleri gibi 30 yapısal göstergeyle birleştirilmektedir. Deneyler, 549.346 URL içeren Phishing Site URLs veri seti üzerinde yürütülmüş; tekrar eden ve geçersiz etiketli kayıtların temizlenmesiyle 507.191 benzersiz URL elde edilmiş, sınıf oranı korunarak %30'luk katmanlı bir alt küme (152.157 URL) %80 eğitim / %20 test biçiminde ayrılmıştır. Logistic Regression, SGDClassifier, RidgeClassifier, Complement Naive Bayes, Random Forest, LightGBM ve XGBoost modelleri ile bunların olasılık tabanlı soft-voting birleşimi değerlendirilmiştir. Sınıf dengesizliği doğrusal modellerde RandomOverSampler, ağaç tabanlı modellerde sınıf ağırlıklandırmasıyla ele alınmış; hiperparametreler katmanlı çapraz doğrulamayla (GridSearchCV ve RandomizedSearchCV) ayarlanmış ve karar eşikleri out-of-fold tahminler üzerinden F1 skorunu maksimize edecek biçimde optimize edilmiştir. Sonuçlar, Logistic Regression'ın %97,33 doğruluk ve 0,94 F1 ile en iyi sonucu verdiğini; Ridge ve soft-voting modellerinin de Bad sınıfı için 0,93–0,94 F1 aralığında, 0,99'a ulaşan ROC-AUC değerleriyle güçlü performans sergilediğini göstermektedir. Bulgular, yüksek boyutlu URL özellik uzayında doğrusal modellerin etkili bir ayrım sağladığını ve optimize edilmiş, yalnızca URL tabanlı bir hattın güvenilir ve düşük maliyetli kimlik avı tespiti sunabileceğini ortaya koymaktadır.

Anahtar Kelimeler: Kimlik avı, URL, TF-IDF, Makine öğrenmesi, Ansambl öğrenme, Sınıf dengesizliği

Comparative Analysis of Machine Learning and Ensemble Methods for URL-Based Phishing Detection Using TF-IDF and Structural Features

Abstract: Phishing attacks, which target users' credentials and financial data, are a rapidly growing cybersecurity threat, making their early and low-cost detection critical. Traditional approaches that analyze web page content introduce high computational cost and latency, while static blacklist-based methods fail against newly created or short-lived URLs. This study proposes a hybrid detection pipeline that relies solely on URL information: character-level TF-IDF n-gram representations are combined with 30 structural indicators such as URL length, symbol density, subdomain depth, and domain/TLD features. Experiments were conducted on the Phishing Site URLs dataset of 549,346 URLs; after removing duplicate and invalid-label records, 507,191 unique URLs were obtained, and a 30% stratified subset (152,157 URLs) was split into 80% training and 20% testing while preserving the class ratio. Logistic Regression, SGDClassifier, RidgeClassifier, Complement Naive Bayes, Random Forest, LightGBM, and XGBoost, together with a probability-based soft-voting ensemble, were evaluated. Class imbalance was addressed using RandomOverSampler for the linear models and class weighting for the tree-based models; hyperparameters were tuned via stratified cross-validation (GridSearchCV and RandomizedSearchCV), and decision thresholds were optimized on out-of-fold predictions to maximize the F1 score. The results show that Logistic Regression achieved the best performance with 97.33% accuracy and a 0.94 F1 score, while the Ridge and soft-voting models also performed strongly, with F1 scores in the 0.93–0.94 range for the Bad class and ROC-AUC values reaching 0.99. The findings demonstrate that linear models provide effective discrimination in the high-dimensional URL feature space and that an optimized, URL-only pipeline can deliver reliable and low-cost phishing detection.

Keywords: Phishing, URL, TF-IDF, Machine learning, Ensemble learning, Class imbalance

***Corresponding author:** Ataturk University, Faculty of Engineering, Department of Software Engineering, 25240, Erzurum, Türkiye
E mail: esra.odabas@atauni.edu.tr
Ülkü ÖZMEN https://orcid.org/0009-0005-1988-391X
Esra ODABAŞ YILDIRIM https://orcid.org/0000-0002-0936-1342

Gönderi: 10 Aralık 2025 **Received:** December 10, 2025
Kabul: 16 Haziran 2026 **Accepted:** June 16, 2026
Yayınlanma: 15 Temmuz 2026 **Published:** July 15, 2026

Cite as: Özmen, Ü., & Odabaş Yıldırım, E., &. (2026). Comparative analysis of machine learning and ensemble methods for URL-based phishing detection using TF-IDF and structural features. *Black Sea Journal Engineering and Science*, *9*(4), 1725-1736.

This work is licensed (CC BY-NC 4.0) under Creative Commons Attribution 4.0 International License

1. Giriş

Kimlik avı saldırıları, günümüzde siber güvenliğin karşı karşıya olduğu en yaygın ve en tehlikeli tehditler arasında yer almaktadır. Bu saldırılar, kullanıcıları meşru kurumlar veya güvenilir hizmetler gibi davranan sahte dijital varlıklar aracılığıyla kandırmayı hedefler. Özellikle banka işlemleri, çevrim içi ödeme sistemleri, kurumsal e-posta servisleri ve sosyal medya platformları, kimlik avı saldırılarının en sık hedef aldığı alanlardır. Saldırganlar, kullanıcıların güvenini kazanmak için çoğunlukla sahte web siteleri ve yanıltıcı URL'ler kullanarak kişisel bilgileri, kimlik doğrulama verilerini ve finansal bilgileri ele geçirmeye çalışmaktadır. Bu tehdit, Türkiye özelinde de hızla büyümektedir. Ulusal Siber Olaylara Müdahale Merkezi (USOM) verilerine göre oltalama saldırısı sayısı 2022 yılında 72.209 iken 2023 yılında 105 binin üzerine çıkmıştır (USOM, 2024). Bilgi Teknolojileri ve İletişim Kurumu (BTK), 2024 yılında oltalama saldırıları dâhil toplam 123 binin üzerinde zararlı bağlantı tespit edilerek altyapı seviyesinde erişim engeli getirildiğini açıklamıştır (BTK, 2025). Konunun URL ve alan adı boyutuyla doğrudan ilişkili olarak, USOM'un yapay zekâ tabanlı tespit sistemiyle 61.293 alan adının vatandaşları dolandırmaya yönelik oltalama amacıyla kullanıldığı belirlenmiştir (USOM, 2024). Bu artış eğilimi, 19 Mart 2025 tarihli ve 32846 sayılı Resmî Gazete'de yayımlanan 7545 sayılı Siber Güvenlik Kanunu ile ulusal düzeyde kurumsal çerçevenin güçlendirilmesini de beraberinde getirmiştir (Siber Güvenlik Kanunu, 2025). Söz konusu tablo, özellikle alan adı ve URL düzeyinde çalışan, hızlı ve ölçeklenebilir tespit yöntemlerine duyulan ihtiyacı Türkiye bağlamında da somut biçimde ortaya koymaktadır.

Kimlik avı saldırılarının temel özelliği, kullanıcı etkileşimini tetikleyen bağlantılar üzerinden gerçekleştirilmesidir. Bu nedenle URL yapısı, saldırının doğası hakkında önemli ipuçları barındırmaktadır. Alan adı benzerlikleri, aşırı uzun URL'ler, olağandışı karakter kullanımı, derin alt alan adları ve aldatıcı anahtar kelimeler kimlik avı saldırılarında sıklıkla kullanılan teknikler arasında yer almaktadır. Bu tür sözel ve yapısal manipülasyonlar, URL tabanlı analiz yöntemlerini kimlik avı tespiti için cazip hale getirmektedir.

Geleneksel kimlik avı tespit yaklaşımları çoğunlukla web sayfası içeriğinin analizine, HTML ve JavaScript tabanlı incelemelere veya statik kara liste (blacklist) mekanizmalarına dayanmaktadır. Bu yaklaşımlar, saldırıların belirli özelliklerini ayrıntılı biçimde ortaya koyabilse de web sayfasının indirilmesi ve işlenmesini gerektirdiğinden yüksek hesaplama maliyetleri ve gecikme süreleri oluşturabilmektedir. Ayrıca içerik tabanlı yöntemler, dinamik olarak değişen web yapıları ve istemci taraflı betikler nedeniyle kararsız sonuçlar üretebilmektedir. Benzer biçimde, e-posta tabanlı kimlik avı tespitinde de metinsel içerik temsilleri (TF-IDF, Word2Vec ve BERT) yüksek başarı sağlamakta; ancak bu yaklaşımlar e-posta gövdesi gibi içeriğe doğrudan erişimi gerektirmektedir (Al Tawil vd., 2024).

Statik kara liste tabanlı yöntemler ise yalnızca daha önce tespit edilmiş zararlı alan adlarını kapsadığından, yeni oluşturulan veya kısa ömürlü kimlik avı URL'leri karşısında yetersiz kalabilmektedir. Saldırganların alan adı yapısını ve URL bileşenlerini hızlı bir şekilde değiştirebilmesi, bu tür yöntemlerin etkinliğini önemli ölçüde azaltmaktadır. Bu nedenlerle, geleneksel kimlik avı tespit yaklaşımlarının ölçeklenebilirlik, gerçek zamanlı uygulama ve daha önce görülmemiş saldırılara karşı dayanıklılık açısından çeşitli sınırlamalar içerdiği literatürde yaygın olarak vurgulanmaktadır (Aljofey vd., 2022).

Bu bağlamda, yalnızca URL bilgisine dayalı kimlik avı tespit yöntemlerine olan ilgi son yıllarda artmıştır. Bu yöntemler, web sayfasının içeriğini indirmeye veya karmaşık analizler yapmaya gerek kalmadan saldırıların erken bir aşamada tespit edilmesine olanak tanımaktadır. Bu sayede hesaplama maliyetleri azalmakta ve söz konusu çözümler gerçek zamanlı çalışan güvenlik sistemlerine daha kolay entegre edilebilmektedir. Ayrıca URL tabanlı yaklaşımlar, saldırganların sıklıkla kullandığı yanıltıcı kelime yapıları ile URL üzerindeki yapısal düzensizlikleri yakalayabilme potansiyeline sahiptir (Jalil vd., 2023; Karim vd., 2023). Öte yandan, derin öğrenme tabanlı kimlik avı tespit yöntemleri yüksek doğruluk sağlasa da hesaplama maliyeti, gerçek zamanlı çalışma ve mobil/edge cihazlarda dağıtım gibi açılardan önemli zorluklar barındırmaktadır (Sangeetha ve Nair, 2025). Bu durum, hafif ve yalnızca URL tabanlı yaklaşımların önemini artırmaktadır.

Bu çalışma, kimlik avı saldırılarının tespiti için URL-Only Hybrid Detection Pipeline adı verilen bir yaklaşım önermektedir. Önerilen yöntem, URL'nin karakter düzeyindeki sözel yapısını temsil eden TF-IDF tabanlı n-gram özellikleri ile URL uzunluğu, sembol yoğunluğu, parametre sayısı, alt alan adı derinliği ve domain/TLD özellikleri gibi mühendislik ürünü yapısal göstergeleri bir araya getirmektedir. Bu hibrit yaklaşım sayesinde, kimlik avı URL'lerinin hem sözel manipülasyonları hem de yapısal düzensizlikleri aynı anda modellenebilmektedir. Çalışmada, Logistic Regression, SGDClassifier, RidgeClassifier, Naïve Bayes, Random Forest, LightGBM ve XGBoost gibi birden fazla makine öğrenimi algoritması sistematik bir şekilde değerlendirilmiştir. Veri kümesinde bulunan belirgin sınıf dengesizliği, doğrusal modellerde çapraz doğrulama aşamalarında uygulanan RandomOverSampler (ROS), ağaç tabanlı modellerde ise sınıf ağırlıklandırması ile ele alınmış; ek olarak, kimlik avı sınıfında en yüksek F1 performansının elde edilmesi için out-of-fold eşik optimizasyonu yapılmıştır. Ağaç tabanlı modellerin yüksek boyutlu TF-IDF özellik matrisiyle çalışırken karşılaşabileceği hesaplama ve bellek maliyetlerini azaltmak amacıyla TruncatedSVD tabanlı boyut indirgeme opsiyonel olarak uygulanmıştır. Bu sayede, verimlilik artırılırken performans kayıplarının önüne geçilmiştir.

Deneysel sonuçlar, hibrit özellik entegrasyonunun tüm değerlendirilmiş modellerde yüksek doğruluk,

hassasiyet, geri çağırma ve F1 skorları ortaya koyduğunu göstermektedir. Özellikle Logistic Regression ve benzeri lineer modeller, hibrit özellik uzayında dikkat çekici bir performans sergilemiş ve yüksek boyutlu URL özelliklerinin doğrusal karar sınırlarıyla etkili bir şekilde temsil edilebildiğini göstermiştir. Ayrıca, en başarılı modellerin soft-voting tabanlı bir VotingClassifier ile bir araya getirilmesi, farklı kimlik avı saldırılarına karşı daha kararlı ve hata toleransı yüksek bir yapı sağlamıştır.

Sonuç olarak, bu çalışma; URL'nin yapısal ve sözel özelliklerini tek ve optimize bir pipeline'da birleştiren bir yaklaşım sunmaktadır. Elde edilen sonuçlar, bu yöntemin hem akademik literatürde hem de siber güvenlik uygulamaları açısından değerli katkılar sunduğunu açıkça göstermektedir. Özellikle hibrit özellik kümesinin lineer ve ansambl modeller üzerinde sağladığı güçlü performans, URL verilerini kullanarak kimlik avı saldırılarının hızlı ve yüksek doğrulukla tespit edilebileceğini ortaya koymuştur. Bu yaklaşım, içerik tabanlı tekniklere göre hem daha az işlem gerektirmekte hem de farklı sistemlere daha kolay entegre edilebilmektedir.

2. Materyal ve Yöntem

Bu çalışma, URL tabanlı kimlik avı saldırılarını tespit etmek amacıyla, yapısal özellikleri ve karakter bazlı TF-IDF analizini bir araya getiren hibrit bir yaklaşım sunmaktadır. Hedefimiz, URL'lerin hem yapısal formundan hem de metinsel diziliminden gelen ipuçlarını birleştirerek, gerçek zamanlı çalışabilen ve yüksek doğruluk oranına sahip bir model geliştirmekti. Süreç boyunca veri temizliği, özellik mühendisliği, model eğitimi, hiperparametre optimizasyonu ve karar eşiği ayarlaması gibi adımlar titizlikle uygulanmış; farklı makine öğrenimi modelleri kıyaslanarak en iyi sonucu veren mimari seçilmiştir. Bu bölümde, kullanılan veri setinden değerlendirme kriterlerine kadar izlenen tüm metodolojik süreç detaylarıyla ele alınmaktadır.

2.1. Veri Seti

Çalışmanın deneysel aşamasında, toplam 549.346 adet URL içeren Phishing Site URLs veri seti kullanılmıştır (Kaggle, n.d.). İkili sınıflandırma yapısına sahip olan bu veri setindeki her bir örnek, meşru siteleri temsil eden "Good" ve oltalama girişimlerini ifade eden "Bad" etiketiyle tanımlanmaktadır. Veri setinin temel özellikleri ve uygulanan ön işleme adımlarının özeti Tablo 1'de sunulmaktadır. Tablo 1'de görüldüğü üzere, ham veri kümesinde toplam 42.154 yinelenen URL kaydı ve 3 çelişkili etiketli örnek tespit edilmiştir. Bu kayıtların temizlenmesi sonucunda toplam 42.155 kayıt çıkarılmış ve çalışma 507.191 benzersiz ve tutarlı URL üzerinden sürdürülmüştür. Temizlenmiş veri seti incelendiğinde, sınıflar arasında yaklaşık %77,5 (Good) ve %22,5 (Bad) oranında belirgin bir dengesizlik bulunduğu görülmektedir. Model eğitiminin hesaplama yükünü azaltmak ve sınıf dengesini korumak amacıyla veri setinin %30'u Stratified Shuffle Sampling yöntemi ile alt örneklemeye tabi tutulmuştur. Bu işlemle nihai çalışılan veri kümesi 152.157 URL içermektedir. Ayrılan bu alt küme daha sonra %80 eğitim ve %20 test olacak şekilde bölünerek deneysel süreç devam ettirilmiştir.

2.2. Veri Ön İşleme

Bu çalışmada kullanılan "Phishing Site URLs" veri seti üzerinde modellemeden önce kapsamlı bir veri ön işleme süreci uygulandı. İlk aşamada, veri setindeki tüm URL'ler standart hâle getirilip, gereksiz boşluklar temizlendi ve etiketler ikili sınıflandırma yapısına uygun olarak "Good = 0" ve "Bad = 1" biçiminde dönüştürüldü. Veri setinde yer alan tekrarlayan URL kayıtları ve özellikle aynı URL'nin birden fazla farklı etiketle işaretlendiği çelişkili örnekler tespit edilerek veri setinden çıkarılmıştır. Bu işlem sonucunda hatalı veya tutarsız örneklerin modele etki etmesi engellenmiştir.

Veri setindeki tüm URL'ler arasında önemli bir sınıf dengesizliği bulunduğundan, modelleme aşamasında veri setinin %30'luk bir alt kümesi oluşturulmuş ve katmanlı örnekleme (Stratified Sampling) yöntemi ile sınıf oranlarının korunması sağlanmıştır. Böylece hem hesaplama maliyeti azaltılmış hem de orijinal veri setinin istatistiksel dağılımı bozulmadan temsil edilmiştir. Alt örneklemeden sonra veri seti %80 eğitim ve %20 test olacak şekilde ayrılmıştır. Ham veri kümesinde toplam 42,154 yinelenen (duplicate) URL tespit edilmiş, ayrıca 3 adet çelişkili etiketli URL belirlenmiştir. Bu çelişkili örnekler veri setinden çıkarılmış ve tekrarlayan kayıtlar temizlenmiştir.

Tablo 1. Kimlik avı URL tespiti için veri kümesi özeti

Özellik	Değer	Açıklama
Toplam kayıt (ham)	549,346	Tüm veri seti boyutu
Benzersiz URL sayısı	507,191	Ham verinin ≈ %92,3'ü
Silinen tekrarlar	42,154	Ham verinin ≈ %7,7'si
Sütunlar	URL, Label	Yalnızca iki sütun
Sınıf dağılımı	Good: % 77,5, Bad: %22,5	Dengesiz veri seti
Etiket eşlemesi	Good → 0, Bad → 1	Binary Encoding
Eğitim/Test bölünmesi	%80 / %20	Stratified Sampling
Kullanılan alt küme	152,157	Katmanlı alt örneklem

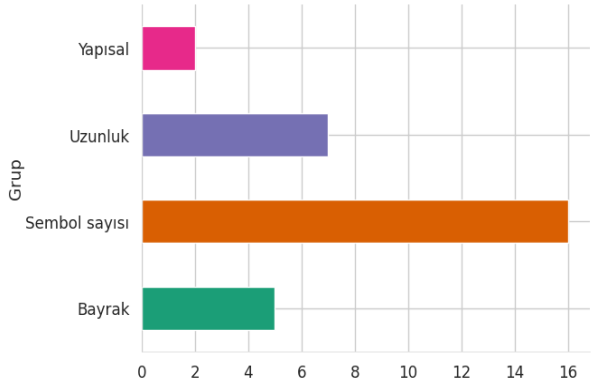
2.3. Özellik Çıkarımı

Kimlik avı URL tespit performansını artırmak için iki tür özellik çıkarımı yapılmıştır. Bunlar yapısal özellikler ve

karakter düzeyinde TF–IDF özellikleridir. Bu özelliklerden oluşan hibrit bir özellik uzayı oluşturulmuştur. URL'lerden çıkarılan yapısal özellikler toplam 30 adet olup Şekil 1'de bu özelliklerin dört ana gruba (Uzunluk, Sembol Sayısı, Bayrak Özellikleri ve Diğer Yapısal Özellikler) göre dağılımı gösterilmektedir. Tablo 2 ise metinsel özellik çıkarımı sürecinde kullanılan TfidfVectorizer algoritmasının hiperparametrelerini özetlemektedir.

2.3.1. Yapısal özellik mühendisliği

URL'lerin kimlik avı davranışlarını daha etkili biçimde analiz edebilmek için çalışmada hibrit özellik çıkarımı yaklaşımı benimsenmiştir. İlk olarak URL'lerden toplam 30 adet yapısal özellik elde edilmiştir. Bu özellikler; URL, hostname, path ve query bölümlerinin uzunluklarını içeren uzunluk tabanlı özellikler, çeşitli karakterlerin (., -, /, ?, =, @, %, _, !, # vb.) frekansını ölçen sembol sayımı özellikleri, URL'nin "https" ile başlayıp başlamadığı, "@" içerip içermediği, path içerisinde çift eğik çizgi bulunup bulunmadığı gibi bayrak (binary) özellikleri, ayrıca IP adresi kullanımı, parametre sayısı ve subdomain sayısı gibi diğer yapısal bileşenlerden oluşmaktadır. Bu özellikler URL'nin biçimsel ve yapısal karmaşıklığını temsil ederek sınıflandırma modellerine güçlü ayrıştırıcı sinyaller sağlamaktadır.



Şekil 1. URL'den çıkarılan 30 yapısal özelliğin dört ana kategoriye göre dağılımı.

2.3.2. TF-IDF tabanlı metinsel özellikler

Yapısal özelliklere ek olarak, URL içindeki karakter örüntülerini yakalayabilmek için karakter tabanlı TF–IDF n-gram özellikleri çıkarılmıştır. URL'lerin metinsel içeriği, kimlik avı tespitinde yüksek ayrım gücü sağlayan değerli özellikler barındırır. Bu kapsamda, hostname–path–query birleşiminden oluşturulan metin üzerinde 3–5 uzunluğunda karakter n-gram'ları üretilmiş ve düşük frekanslı diziler elenmiştir. Kimlik avı URL'lerinde sık görülen karakter dizileri (ör. /login/verify, secure-update, paypal-confirm, cgi-bin, account-reset, rastgele karmaşık diziler vb.) belirli bir örüntü oluşturur. TF-IDF, bu örüntüleri nadiren görülen ancak ayrıştırıcı karakter dizilerine daha yüksek ağırlık vererek yakalar. Yüksek boyutlu metinsel temsili yönetilebilir kılmak amacıyla TF–IDF vektörü 60,000 özellik ile sınırlandırılmıştır. Böylece URL'lere özgü istatistiksel karakter kalıpları modele aktarılmıştır. Bu iki özellik grubunun birleşimiyle oluşturulan hibrit özellik uzayı hem URL'nin yapısal karakteristiklerini hem de karakter düzeyindeki olası kimlik avı izlerini tek bir bütün içinde temsil ederek model performansının anlamlı şekilde artmasını sağlamıştır. Metinsel özellik mühendisliği kapsamında kullanılan TF-IDF vektörleştiricisine ait hiperparametreler Tablo 2'de görüldüğü üzere belirlenmiştir (eşitlik 1).

$$idf(t) = log\left(\frac{N+1}{df(t)+1}\right) + 1, \quad (1)$$

Burada,

- t: İlgili terimi (kelime veya n-gram),
- d: İncelenen dokümanı (burada URL metni),
- $tf(t,d)$: Terimin $d$dokümanı içindeki terim frekansını (kaç kez geçtiğini),
- $idf(t)$: Terimin tüm doküman kümesindeki ayırt ediciliğini gösteren ters doküman frekansını temsil etmektedir.

Ters doküman frekansı ise (eşitlik 2),

$$tfidf(t,d) = tf(t,d) \cdot idf(t) \quad (2)$$

şeklinde hesaplanmakta olup:

- N: Toplam doküman (URL) sayısını,
- $df(t)$: Terimin geçtiği doküman sayısını,
- log: Logaritmik ölçeklendirmeyi

İfade etmektedir. Bu tanımlama sayesinde, çok sayıda dokümanda geçen yaygın terimlerin ağırlığı azaltılırken, yalnızca belirli URL'lerde görülen ayırt edici karakter dizilerine daha yüksek ağırlık verilmektedir.

Tablo 2. Metinsel özellik mühendisliği için tfidfvectorizer hiperparametreleri

Parametre	Değer
analyzer	"char" (Karakter)
ngram_range	(3, 5)
max_features	60,000
min_df	2
sublinear_tf	True

2.3.3. Hibrit özellik mühendisliği

Bu çalışmada kimlik avı URL tespit performansını artırmak amacıyla hibrit bir özellik çıkarım yaklaşımı uygulanmıştır. Yöntem, URL'lerin hem yapısal karakteristiklerini hem de TF-IDF tabanlı karakter-seviyesi metinsel temsilini bir araya getirir. Böylece model, URL'nin hem semantik hem de biçimsel özelliklerine aynı anda erişmiş olur. Bu birleştirme, hafıza dostu ve optimize bir yapıda scipy.sparse.hstack() fonksiyonu kullanılarak yapılmıştır (eşitlik 3):

$$X_{hybrid} = \left[X_{struct} \parallel X_{tfidf}\right] \quad (3)$$

Bu sayede model hem URL'nin yapısal karakteristiklerini hem de metinsel örüntülerini aynı anda değerlendirebilmektedir. Bu hibrit yaklaşım, çalışmadaki tüm makine öğrenimi algoritmaları için ortak giriş temsilini oluşturmuş ve özellikle Logistic Regression,

RidgeClassifier ve hibrit Soft-Voting modellerinde yüksek F1-skoru elde edilmesini sağlamıştır.

2.4. Model ve Mimari

Bu çalışmada kimlik avı URL'lerinin tespiti için hibrit özellik uzayı üzerinde çeşitli makine öğrenimi modelleri değerlendirilmiştir. Eğitim süreci, önceki aşamalarda elde edilen 30 yapısal manuel özellik ile karakter düzeyinde TF–IDF n-gramlarından oluşan yüksek boyutlu metinsel özelliklerin birleştirilmesiyle oluşturulan veri temsili üzerinde yürütülmüştür. Veri kümesi, sınıf dağılımını korumak amacıyla katmanlı örnekleme yöntemiyle %80 eğitim ve %20 test olacak şekilde ayrılmıştır. Alt örnekleme işlemlerinde StratifiedShuffleSplit kullanılmıştır. Hiperparametre optimizasyonu ve model seçimi aşamalarında katmanlı k-katlı çapraz doğrulama uygulanmış; doğrusal ve hafif modeller için k=5, hesaplama maliyeti daha yüksek olan ağaç tabanlı modeller için ise k=3 olarak belirlenmiştir. Çapraz doğrulama işlemlerinde StratifiedKFold yapısı shuffle=True seçeneğiyle kullanılmış ve deneylerin tekrarlanabilirliğini sağlamak amacıyla tüm rastgelelik kaynakları random_state=42 ile sabitlenmiştir.

Modelleme aşamasında Logistic Regression, Stochastic Gradient Descent Classifier, Ridge Classifier, Complement Naive Bayes, Random Forest, LightGBM ve XGBoost gibi hem doğrusal hem de doğrusal olmayan farklı sınıflandırıcılar kullanılmıştır. Hiperparametre optimizasyonunda model ailesinin karmaşıklığına göre farklı arama stratejileri tercih edilmiştir. Arama uzayı görece dar olan doğrusal ve hafif modellerde GridSearchCV kullanılırken, parametre uzayı daha geniş ve hesaplama maliyeti daha yüksek olan LightGBM ve XGBoost modellerinde RandomizedSearchCV uygulanmıştır. Böylece hem sistematik bir hiperparametre araması yapılmış hem de yüksek maliyetli modellerde hesaplama yükü kontrol altında tutulmuştur.

Veri setindeki sınıf dengesizliği model ailesine göre farklı biçimde ele alınmıştır: doğrusal modeller (LR, SGD, Ridge) ile Complement Naive Bayes'te çapraz doğrulama katmanları içinde RandomOverSampler (ROS) uygulanırken, ağaç tabanlı modellerde (RF, LightGBM, XGBoost) ROS yerine sınıf ağırlıklandırması (sırasıyla balanced_subsample, is_unbalance ve scale_pos_weight) tercih edilmiştir. RandomOverSampler, veri sızıntısını önlemek amacıyla yalnızca her çapraz doğrulama katmanının eğitim kısmına uygulanmış; doğrulama ve test verileri hiçbir şekilde örnekleme işlemine dahil edilmemiştir.

Dengesizliği gidermek için SMOTE veya ADASYN gibi sentetik örnekleme yöntemleri yerine ROS'un seçilmesinin temel nedeni özellik uzayının niteliğidir: SMOTE ve ADASYN, azınlık sınıfı örnekleri arasında komşuluk (k-NN) temelli doğrusal ara değerlemeyle sentetik örnekler üretir. Ancak karakter düzeyindeki TF-IDF temsili 60.000 boyutlu ve oldukça seyrek olduğundan, bu uzayda mesafe ve komşuluk hesapları anlamını büyük ölçüde yitirir; ara değerlemeyle üretilen sentetik vektörler gerçek URL'lerin karakter n-gram örüntülerine karşılık gelmeyebilir ve bu işlemler yüksek boyutlu seyrek matrislerde yüksek bellek/zaman maliyeti getirir. ROS ise mevcut azınlık örneklerini birebir çoğalttığından gerçek n-gram örüntülerini bozmadan korur, seyrek yapıyla uyumludur ve veri sızıntısını önlemek için yalnızca her çapraz doğrulama katmanının eğitim kısmında uygulanmıştır.

Buna ek olarak, bireysel modellerin güçlü yönlerini bir araya getirerek daha dengeli bir performans elde etmek amacıyla olasılık tabanlı bir Soft Voting ensemble modeli de oluşturulmuştur; bu yapıda yalnızca predict_proba çıktısı üretebilen sınıflandırıcılar yer almıştır. Bu seçim süreci sonucunda Soft Voting yapısı Logistic Regression ve SGDClassifier (logistic loss) modellerinin birleşiminden oluşturulmuştur.

Her modelin hiperparametreleri çapraz doğrulama temelli arama süreciyle optimize edilmiş; değerlendirme sırasında yalnızca tahminlerin doğruluğu değil, aynı zamanda dengesiz veri setlerinde kritik öneme sahip olan "Bad" sınıfının yakalanma başarısı da göz önünde bulundurulmuştur. Bu doğrultuda, modellerin eğitim seti üzerindeki out-of-fold (OOF) tahminleri kullanılarak her bir model için F1(Bad) skorunu maksimize eden optimal karar eşiği belirlenmiştir. Performans değerlendirmesi Accuracy, Precision (Bad), Recall(Bad), F1(Bad), ROC-AUC ve PR-AUC metriklerine ek olarak karmaşıklık matrisleri üzerinden yapılmıştır. Böylece modellerin özellikle kimlik avı URL'lerini doğru şekilde tespit etme kabiliyeti detaylı olarak incelenmiştir.

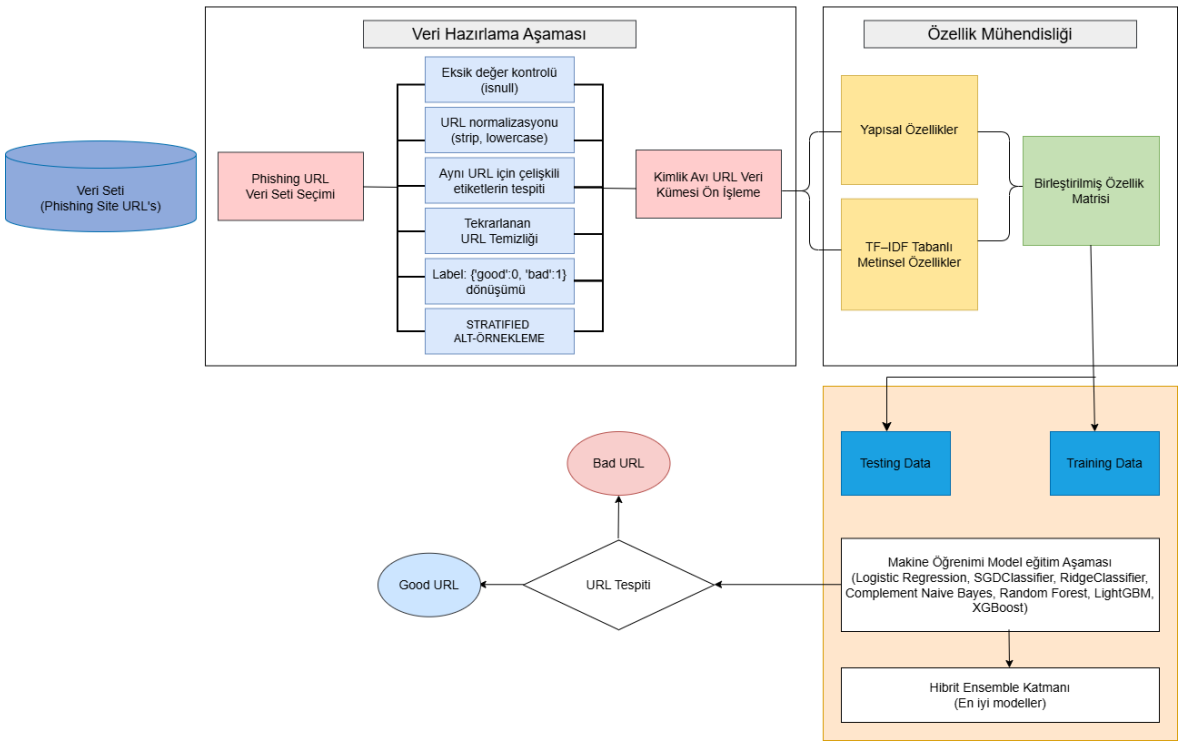
Şekil 2, çalışmada kullanılan URL kimlik avı tespit sisteminin genel mimarisini göstermektedir. Mimaride veri temizleme, ön işleme, yapısal ve TF-IDF tabanlı özellik çıkarımı, eğitim-test ayrımı ve makine öğrenimi modellerinin eğitilmesi gibi aşamalar ayrıntılı olarak sunulmaktadır. Bu yapı, çalışmanın uçtan uca işleyişini ve kullanılan yöntemlerin birbirleriyle olan etkileşimini açıkça ortaya koymaktadır.

Logistic Regression, doğrusal sınırlarla ayrılabilen verilerde etkili ve yorumlanabilir bir ikili sınıflayıcıdır. Kimlik avı URL detection görevlerinde, özellikle özelliklerin (örneğin URL uzunluğu, özel karakter sayısı) doğrusal etkiler gösterdiği durumlarda yüksek başarı sunar. Ayrıca hızlı çalışması ve düşük hesaplama maliyeti nedeniyle gerçek zamanlı sistemlerde tercih edilir (Hosmer vd., 2013).

SGD Classifier, özellikle büyük URL veri kümelerinde hızlı model eğitimi sunar ve sınırlı bellek kullanımıyla öne çıkar. Kimlik avı gibi yoğun örnekli alanlarda sık güncellemelerle verimli öğrenme sağladığı için uygundur. LR'ye benzerdir ancak veri yığınları üzerinde artımlı olarak eğitilir (Jalil vd., 2023).

Ridge Sınıflandırıcı L2 düzenlileştirmeli doğrusal bir karar fonksiyonudur. Model, ceza terimi ile karesel kaybı minimize eder. Kimlik avı URL veri kümelerinde, birbirine çok benzeyen veya korelasyonlu URL özelliklerinin yer alması yaygın olduğundan, L2 regülerizasyonuyla daha stabil ve genellenebilir sonuçlar

sağlar (Hoerl ve Kennard, 1970). ComplementNB Özellikle dengesiz metin veri kümeleri (bir sınıfın diğerinden çok az olduğu) için geliştirilmiş, olasılığa dayalı bir modeldir. Standart Naive Bayes'in aksine, bir sınıfın olasılığını doğrudan o sınıftaki örneklerle değil, o sınıfa ait olmayan örneklerin özelliklerini analiz ederek hesaplar. XGBoost, gradyan artırma algoritmasının optimize edilmiş yüksek performanslı versiyonudur. Model, zayıf öğreniciler olarak karar ağaçlarını kullanır ve her yeni ağaç, önceki ağaçların yaptığı hataları azaltacak şekilde ardışık olarak eğitilir. Kimlik avı URL'lerdeki karmaşık karar sınırlarını modelleyebilir, overfitting'i kontrol eder ve gelişmiş regülerizasyon teknikleri içerir. (Das Guptta vd., 2024).



Şekil 2. Çalışmada önerilen kimlik avı URL tespit sisteminin genel mimarisi.

LightGBM, büyük ölçekli veri kümelerinde hızlı ve yüksek performanslı öğrenme sağlayan, histogram tabanlı Gradient Boosting algoritmasına dayalı bir makine öğrenmesi yöntemidir. Dengesiz veri setleri için uygun loss fonksiyonları ve class-weight desteği de sağlar (Karim vd., 2023).

Soft-voting yaklaşımı, her bir temel modelin tahmin olasılıklarını birleştirerek daha kararlı ve genellenebilir bir çıktı üretmeyi amaçlayan bir topluluk yöntemidir. Her model, giriş örneği için "Bad (1)" sınıfı olasılığını üretir; ardından tüm modellerin bu olasılıklarının ortalaması alınır. Son karar, ortalama olasılığı en yüksek sınıfın seçilmesiyle verilir. Bu yöntem, farklı modellerin güçlü yönlerini bir araya getirerek hataya karşı dayanıklılığı artırır (Mathew vd., 2025) (eşitlik 4).

$$\bar{p}(y=1|x) = \frac{1}{M}\sum_{m=1}^{M} p_m(y=1|x)\ ,$$

$$\hat{y} = \arg\max_{c\backslash in\,\{0,1\}} \bar{p}\,(y=c|x) \qquad (4)$$

Bu yapı, tüm modellerin olasılık tahminlerini birleştirerek nihai sınıf kararının verilmesini sağlar.

2.5. Değerlendirme Metrikleri

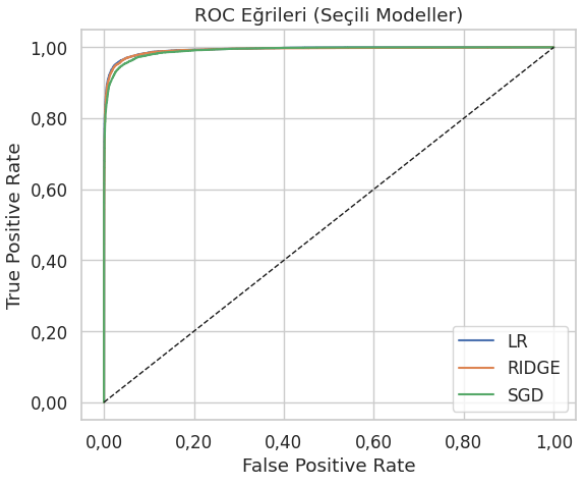
İki sınıflı (Bad ve Good) bir sınıflandırma problemi olması nedeniyle, model performansını kapsamlı şekilde değerlendirmek için Tablo 3'te özetlenen metrikler kullanılmıştır. Özellikle siber güvenlik gibi kritik uygulama alanlarında, yalnızca genel doğruluk değil sınıflar arasındaki dengesizlikleri de göz önüne alan metriklerin raporlaması büyük önem taşımaktadır. Bu kapsamda, pozitif sınıf "Bad", negatif sınıf ise "Good" şeklinde tanımlanmıştır. Formülde yer alan TP, FP, TN ve FN kısaltmaları; sırasıyla Gerçek Pozitif, Yanlış Pozitif, Gerçek Negatif ve Yanlış Negatif değerlerine karşılık gelmektedir.

Tablo 3. Kimlik avı URL tespiti için kullanılan performans metriklerinin ve açıklamaları

Metric	Formül
Doğruluk (Accuracy)	(TP + TN) / (TP + TN + FP + FN)
Kesinlik (Precision)	TP / (TP + FP)
Duyarlılık (Recall)	TP / (TP + FN)
Özgüllük (Specificity)	TN / (TN + FP)
F1-Skoru (F1-Score)	2 × (Kesinlik × Duyarlılık) / (Kesinlik + Duyarlılık)

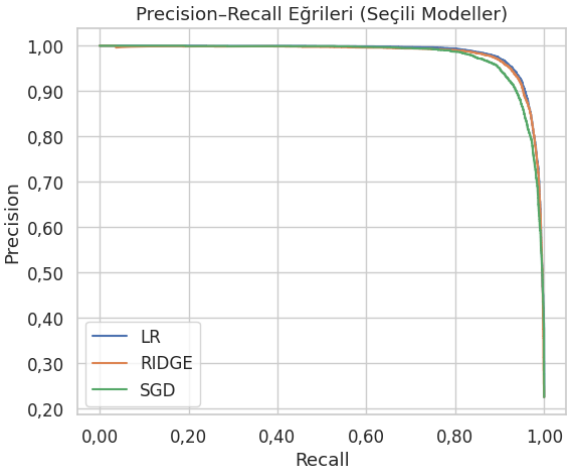
3. Bulgular

Bulgular ve tartışma kapsamında, önerilen TF-IDF tabanlı özellik çıkarımı yöntemiyle eğitilen farklı makine öğrenimi modellerinin performansı; ROC eğrileri, Precision-Recall (Kesinlik-Duyarlılık) eğrileri ve temel değerlendirme metrikleri üzerinden değerlendirilmiştir. En başarılı performansı sergileyen lineer modellere (LR, Ridge, SGD) ait ROC eğrileri Şekil 3'te yer almaktadır. Şekil 3 incelendiğinde, modellerin Sol-Üst köşeye yakınsayan eğrileri, genel ayrıştırma yeteneklerinin üst düzeyde olduğunu ortaya koymaktadır. AUC bakımından en başarılı model 0,9935 değeriyle Lojistik Regresyon (LR) olurken, bu modeli 0,9924 AUC ile Ridge ve 0,9913 AUC ile SGD izlemiştir. Bu sonuç, özellikle LR ve Ridge modellerinin, yasal ve kimlik avı URL'lerini ayırt etme konusunda son derece kararlı bir karar sınırına sahip olduğunu göstermektedir.



Şekil 3. En başarili modellerin (LR, Ridge, SGD) ROC eğrileri.

Modellerin sınıflandırma başarısını daha detaylı incelemek adına, dengesiz veri setlerinde kritik öneme sahip olan Precision-Recall eğrileri Şekil 4'te sunulmaktadır. Grafikte görüldüğü üzere LR ve Ridge modelleri, yüksek duyarlılık (Recall) seviyelerinde dahi kesinlik (Precision) değerlerini korumayı başarmıştır. Bu durum, modellerin yanlış alarm (False Positive) üretmeden saldırıları tespit edebildiğini kanıtlamaktadır. Diğer taraftan Random Forest (RF) gibi ağaç tabanlı modellerde (Tablo 4'te görüldüğü üzere), yanlış negatif oranlarının artmasıyla performansın lineer modellere kıyasla daha düşük kaldığı gözlemlenmiştir.



Şekil 4. En başarılı modellerin (LR, Ridge, SGD) precision–recall eğrileri.

Test sonuçlarının doğruluk (accuracy), kesinlik (precision), duyarlılık (recall) ve F1-skoru açısından karşılaştırıldığı Tablo 4 incelendiğinde, LR modeli en yüksek doğruluğa (%97,33) ve en dengeli performansa (F1: 0,9400) sahiptir. Voting modeli, %97,21 doğruluk ile LR'ye çok yakın bir başarı göstererek ikinci sırada yer almaktadır. Ridge modelinin doğruluk değeri %97,16'ya ulaşmakta, RF ise %93,52 ile en geride kalmaktadır. Kesinlik ve duyarlılık bakımından da LR modeli, hem pozitif tahminlerin doğruluğunu hem de saldırıların atlanmamasını aynı anda sağlayarak sınıflandırmadaki dengesini kanıtlamaktadır.

Şekil 5'te sunulan F1-skoru ve doğruluk grafikleri de bu tabloyu desteklemekte; LR, Voting ve Ridge modellerinin üst sıralarda sıkıştığı; LGBM, XGB ve NB modellerinin orta düzeyde performans sergilediği; RF modelinin ise hem doğruluk hem de F1-skorunda en düşük değeri aldığı açıkça görülmektedir. Özellikle F1-skoru grafiğinde LR, Voting ve Ridge modelleri birbirine oldukça yakın değerler alırken, RF modelinin keskin bir şekilde düşük performans göstermesi hibrit özellik kümesi ile lineer modellerin daha etkili çalıştığını ortaya koymaktadır.

Tablo 4. Modellerin doğruluk, kesinlik, duyarlılık ve F1-skora dayalı test sonuçları

Model	Doğruluk	Kesinlik	Duyarlılık	F-1 Skoru
LR	0,973285	0,952267	0,927967	0,939960
VOTING	0,972102	0,950382	0,924468	0,937246
RIDGE	0,971576	0,945709	0,927092	0,936308
SGD	0,966187	0,938865	0,909157	0,923772
LGBM	0,957742	0,915821	0,894721	0,905148
XGB	0,952583	0,904890	0,882327	0,893466
NB	0,950217	0,903124	0,872703	0,887653
RF	0,935200	0,863325	0,846457	0,854808

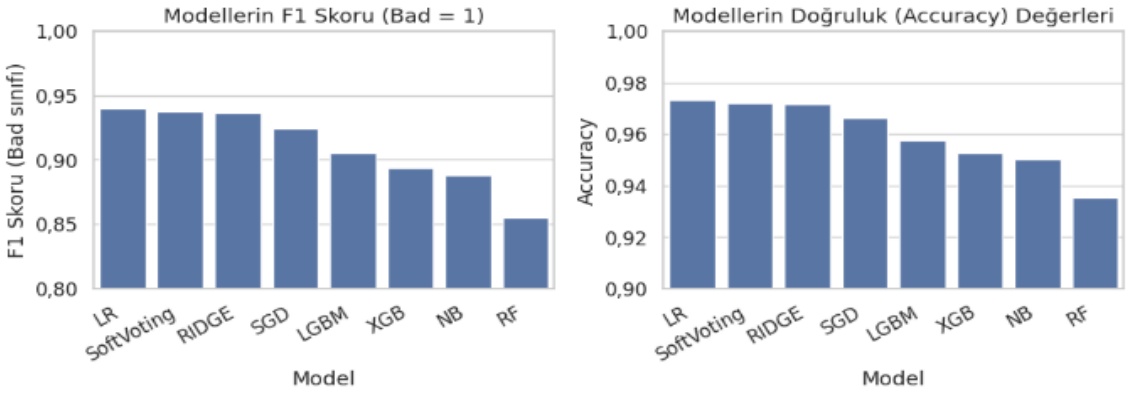
Bu sonuçlar, kullanılan özellik uzayının yapısıyla doğrudan açıklanabilir. Karakter düzeyindeki TF-IDF temsili, 60.000 boyutlu, son derece seyrek (sparse) ve yüksek boyutlu bir uzay oluşturur. Bu tür uzaylarda sınıflar çoğunlukla doğrusal olarak ayrılabilir hâle geldiğinden, L2 düzenlileştirmeli Lojistik Regresyon ve Ridge gibi modeller aşırı öğrenmeye karşı kararlı kalarak güçlü bir karar sınırı öğrenebilmekte; ayrıca seyrek matrislerle verimli çalışıp çok sayıda zayıf ayırt edici n-gram özelliğinin katkısını ağırlıklandırarak birleştirebilmektedir. Buna karşılık Random Forest gibi ağaç tabanlı modeller, her bölünmede tek bir özelliği eşik değerine göre değerlendirdiğinden, on binlerce seyrek ve düşük yoğunluklu n-gram özelliği karşısında bilgilendirici bölünmeler bulmakta zorlanır; bu da yüksek boyutlu seyrek metin verisinde görece daha düşük genelleme performansına ve daha yüksek yanlış negatif oranına yol açar. Nitekim ağaç tabanlı modellerin başarımını korumak için TF-IDF uzayı TruncatedSVD ile yoğun bir alt uzaya indirgenmiş, buna rağmen bu modeller doğrusal modellerin gerisinde kalmıştır.

Tablo 5. Ağaç tabanlı modellerde TruncatedSVD'nin doğruluk, F1 ve eğitim süresine etkisi

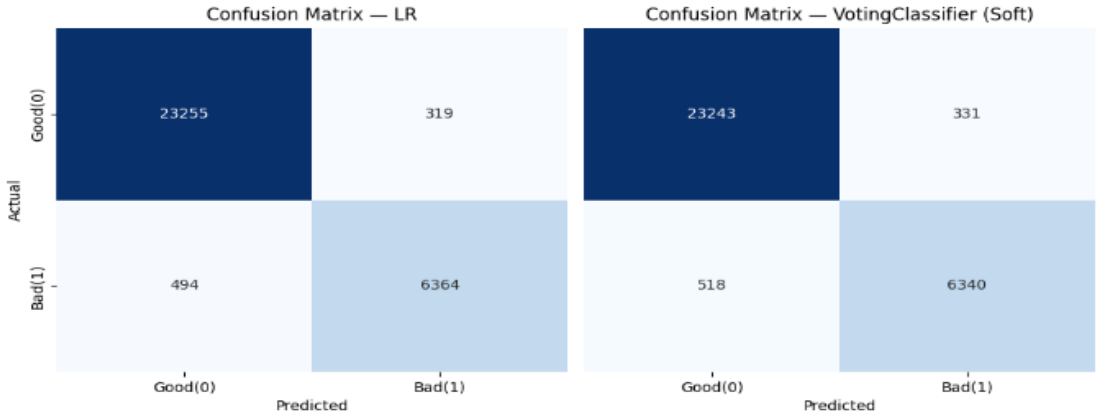
Model	Özellik uzayı	Boyut	Doğruluk	F1 (Bad)	Eğitim süresi (sn)
Random Forest	Tam	60,030	0,9166	0,8211	194,6
Random Forest	SVD	200	0,9352	0,8548	218,3
LightGBM	Tam	60,030	0,9781	0,9514	395,8
LightGBM	SVD	200	0,9577	0,9051	39,0
XGBoost	TAM	60,030	0,9640	0,9195	427,8
XGBoost	SVD	200	0,9526	0,8935	41,0

Tablo 5, TruncatedSVD'nin ağaç tabanlı modeller üzerindeki etkisini özetlemektedir. Gradient boosting modellerinde (LightGBM ve XGBoost) 60,030 boyutlu seyrek TF-IDF uzayının 200 bileşene indirgenmesi eğitim süresini yaklaşık 10 kat azaltmıştır (LightGBM: 395,8 sn → 39,0 sn; XGBoost: 427,8 sn → 41,0 sn). Bu hızlanma, doğrulukta sınırlı bir düşüşle birlikte gelmiştir (F1(Bad) LightGBM'de 0,9514 → 0,9051; XGBoost'ta 0,9195 → 0,8935). Random Forest'ta ise SVD, eğitim süresini benzer düzeyde tutarken (≈195 sn → 218 sn) doğruluğu ve F1 skorunu artırmıştır (F1 0,8211 → 0,8548). Bu bulgular, TruncatedSVD'nin yüksek boyutlu seyrek metin uzayında ağaç tabanlı modeller için önemli bir hesaplama ve bellek tasarrufu sağladığını; bu tasarrufun gradient boosting modellerinde küçük bir doğruluk bedeli karşılığında, Random Forest'ta ise doğrudan bir kazanç olarak elde edildiğini göstermektedir. Çalışmanın düşük hesaplama maliyetini koruma hedefi doğrultusunda ağaç tabanlı modellerde SVD'li temsil tercih edilmiştir.

Modellerin doğru ve hatalı tahminlerinin dağılımını detaylandırmak amacıyla, en yüksek performansı gösteren Lojistik Regresyon (LR) ve Voting (Soft) modellerine ait karmaşıklık matrisleri Şekil 6'da analiz edilmiştir.



Şekil 5. Modellerin F1-skoru ve doğruluk performanslarının karşılaştırılması.



Şekil 6. Lojistik regresyon (LR) ve voting modellerine ait karmaşiklik matrisleri.

Matrisler incelendiğinde, LR modelinin test setindeki toplam 6,858 adet zararlı URL'nin 6,364 tanesini doğru tespit ettiği (Doğru Pozitif), sadece 494 tanesini kaçırdığı (Yanlış Negatif) görülmektedir. Bir siber güvenlik sisteminde saldırının gözden kaçırılmasının maliyeti yüksek olduğundan, LR modelinin sunduğu bu düşük yanlış negatif oranı kritik bir avantajdır. Voting modeli ise 518 adet yanlış negatif ile LR'ye göre biraz daha fazla saldırıyı kaçırmıştır. Ayrıca, LR modelinde 23,574 yasal URL içerisinden sadece 319 tanesi hatalı olarak zararlı (Yanlış Pozitif) etiketlenmişken, Voting modelinde bu sayı 331'dir. Bu durum, LR modelinin hem saldırı tespitinde hem de yanlış alarmları minimize etmede Voting modeline kıyasla daha kararlı olduğunu kanıtlamaktadır.

3.1. Mevcut Çalışmanın Literatür ile Karşılaştırılması

Bu çalışmada geliştirilen TF-IDF ve yapısal özelliklere dayalı hibrit makine öğrenimi yaklaşımı, URL verilerini kullanarak hafif, hızlı ve rekabetçi doğrulukta kimlik avı tespiti yapabilen bir sistem ortaya koymuştur. Model, 549.346 URL içeren Kaggle veri setinden temizleme ve katmanlı alt örnekleme sonrası elde edilen 152.157 URL'lik küme üzerinde eğitilmiş ve Lojistik Regresyon (LR) algoritmasıyla %97,33 doğruluk ve 0,94 F1-skoru elde etmiştir. Ayrıca Soft-Voting ansambl modeli %97,21 doğruluk, Ridge modeli ise %97,16 doğruluk oranına ulaşmıştır. Bu sonuçlar, önerilen hibrit özellik kümesinin doğrusal sınıflandırıcılarla birlikte, herhangi bir boyut indirgemesine gerek kalmadan ve düşük hesaplama maliyetiyle yüksek genellenebilirlik ve kararlılık sunduğunu göstermektedir.

Bu noktada, ağaç tabanlı modeller için yapılan SVD karşılaştırması, Tablo 5'te önemli bir bağlam sağlamaktadır. Söz konusu karşılaştırma, gradient boosting modellerinin SVD'siz tam özellik uzayında doğrusal modellere yakın, hatta LightGBM örneğinde bir miktar daha yüksek doğruluğa ulaşabildiğini; ancak bunun yaklaşık 10 kat daha yüksek eğitim maliyeti gerektirdiğini ortaya koymaktadır. Dolayısıyla bu çalışmada vurgu, salt en yüksek doğruluk yerine doğruluk-maliyet dengesine verilmiştir. Doğrusal modeller, çok daha düşük hesaplama yüküyle rekabetçi bir doğruluk sağladığından, gerçek zamanlı ve ölçeklenebilir uygulamalar için daha uygun bir denge sunmaktadır.

Literatürde yer alan benzer çalışmalar incelendiğinde, önerilen yöntemin performans açısından oldukça rekabetçi olduğu görülmektedir. Jalil vd. (2023), aynı Kaggle veri setini kullanarak gerçekleştirdikleri çalışmada Random Forest (RF) modeliyle %96,25 doğruluk bildirmiştir. Önerilen LR modeli, aynı veri seti üzerinde bu değeri yaklaşık %1,1 oranında aşmıştır. Benzer şekilde, Lakshmanarao, Babu ve Bala Krishna (2021a), aynı Kaggle veri seti üzerinde Hashing Vectorizer ve Random Forest ile %97,5 doğruluk elde etmiştir. Önerilen çalışma aynı veri setinde benzer doğruluğa ulaşmakla birlikte, yalnızca metinsel vektörleştirmeyle sınırlı kalmayıp karakter düzeyi TF-IDF özelliklerini yapısal URL özellikleriyle birleştiren ve ansambl ile eşik optimizasyonunu da değerlendiren daha kapsamlı bir model ve özellik karşılaştırması sunmaktadır.

Karim vd. (2023), 11.054 URL içeren bir Kaggle veri seti üzerinde geliştirdikleri LSD hibrit modelinde %98,12 doğruluk, Kumar (2023) ise Word2Vec + PCA + Random Forest yaklaşımıyla %99,75 doğruluk elde etmiştir. Lakshmanarao vd. (2021b) ise UCI Phishing Websites veri seti üzerinde ANOVA ve Mutual Information tabanlı özellik seçimiyle PA1 ve PA2 öncelik algoritmalarına dayalı bir fusion sınıflandırıcı önermiş ve %97 doğruluk elde etmiştir.

Ancak bu çalışmaların veri setleri daha küçük ve dengeli olduğundan, elde edilen yüksek oranların büyük ölçekli ve dengesiz veri kümelerinde genellenebilirliği sınırlıdır. Bu bağlamda, önerilen çalışmanın yarım milyondan fazla URL içeren dengesiz bir veri setinde %97'nin üzerinde doğruluğu düşük maliyetle sağlaması, modelin sağlamlığını ve genellenebilirliğini desteklemektedir. Mathew vd. (2025) tarafından geliştirilen PhishGuard Ensemble çerçevesi %99 doğruluk elde etmiş olsa da, bu model CNN ve LSTM bileşenleri içermesi nedeniyle yüksek hesaplama maliyeti taşımaktadır. Benzer biçimde, Maneriker vd. (2021) tarafından önerilen transformatör (BERT/RoBERTa) tabanlı URLTran modeli, yaklaşık bir

milyon URL içeren büyük ölçekli bir tarama telemetrisi üzerinde %99,67 doğruluk elde etmiştir. Ancak bu başarım, ağır transformatör mimarileri ve kamuya açık olmayan geniş özel veri kümeleri gerektirmesi nedeniyle yüksek hesaplama maliyeti ve sınırlı yeniden üretilebilirlik taşımaktadır. Buna karşılık mevcut çalışma, yalnızca TF-IDF + yapısal özelliklerle çalışan doğrusal modeller kullanarak çok daha düşük maliyetle benzer bir sınıflandırma başarısı sunmuştur.

Buna ek olarak, Ok ve Kara (2021) tarafından USOM veri seti ile yapılan çalışmada %95 doğruluk, Rashid vd. (2020) tarafından UCI veri seti üzerinde gerçekleştirilen çalışmada ise %95,66 doğruluk rapor edilmiştir. Bu sonuçlar, önerilen modelin hem ulusal hem de uluslararası veri setlerine göre rekabetçi bir başarı sağladığını ortaya koymaktadır. Das Guptta vd. (2024)'ün XGBoost tabanlı hibrit modelinde %99,17 doğruluk elde edilmiştir. Anandaraj vd. (2024) ise URL ve köprü (hyperlink) temelli hibrit özellikleri birleştiren bir kimlik avı tespit sistemi (PDS-HML) önermiştir. Benzer biçimde Alam vd. (2020), 32 yapısal web/URL özelliği üzerinde özellik seçimi ve Temel Bileşen Analizi (PCA) ile boyut indirgeme uygulayarak Karar Ağacı ve Random Forest modelleriyle kimlik avı web sitelerini sınıflandırmıştır. Ancak bu modeller hem HTML içerik hem de URL yapısı özelliklerini birlikte kullanmakta olup, yalnızca URL tabanlı hafif sistemlerle doğrudan karşılaştırılabilir değildir.

Önerilen yaklaşımın ayırt edici yönü, yakın zamanlı benzer çalışmalarla karşılaştırıldığında daha açık biçimde görülmektedir. Asha Shajilal (2024), kimlik avı tespitinde e-posta metni, sayfa içeriği ve URL özelliklerini bir arada kullanan ansambl tabanlı melez bir yaklaşım önermiştir. Ancak bu yöntem e-posta gövdesi ve içerik erişimine bağımlı olduğundan, içerik indirmeden çalışan hafif senaryolara doğrudan uygulanamaz. Ogunleye vd. (2024), URL tabanlı kimlik avı tespitinde çeşitli makine öğrenimi algoritmalarını karşılaştırmalı olarak değerlendirmiştir; ancak karakter düzeyi TF-IDF ile el yapımı yapısal özelliklerin füzyonu, dengesizliğin çapraz doğrulama içinde ele alınması ve out-of-fold eşik optimizasyonu gibi bileşenleri bir arada içermemektedir. Palavali ve Pothireddy (2025), açıklanabilir bir ansambl modelini hafif siber tehdit istihbaratı (CTI) özellikleriyle birlikte kullanmış; bu yaklaşım ise dış istihbarat beslemelerine bağımlıdır. Buna karşılık önerilen çalışma; yalnızca URL bilgisine dayanan, içerik veya dış istihbarat gerektirmeyen, karakter düzeyi TF-IDF ile yapısal özellik füzyonunu dengesizlik yönetimi ve eşik optimizasyonuyla bütünleştiren ve büyük ölçekli, dengesiz bir veri kümesinde değerlendirilen kendi kendine yeten bir hat sunmaktadır.

Genel olarak, literatürdeki çalışmalar ve Tablo 6'daki karşılaştırmalı bulgular birlikte değerlendirildiğinde, önerilen modelin doğruluk-maliyet dengesi açısından verimli yaklaşımlardan biri olduğu görülmektedir. Doğrusal modeller rekabetçi doğruluğu düşük hesaplama maliyetiyle sağlarken, Voting ve Ridge modelleriyle sistemin kararlılığı artırılmış; özellikle LR modelinin düşük yanlış negatif oranı sayesinde kimlik avı URL'lerinin kaçırılma riski azaltılmıştır.

Öte yandan literatürde, kimlik avı tespit modellerinin yalnızca doğruluk veya F1 skoruyla değerlendirilmesinin yeterli olmadığı kullanım senaryosunun güvenlik gereksinimine göre yanlış pozitif ve yanlış negatif oranları arasındaki dengenin karar eşiği ve eğitim verisi sınıf dağılımı ayarlanarak düzenlenmesi gerektiği vurgulanmıştır (Dalvi vd., 2019). Önerilen çalışmada benimsenen sınıf dengesizliği yönetimi ve out-of-fold eşik optimizasyonu da tam bu dengeyi gözeten tasarım tercihleridir.

Literatürdeki çalışmalar topluca değerlendirildiğinde, yüksek doğruluk bildiren yaklaşımların önemli bir kısmının ya görece küçük ve dengeli veri kümelerine, ya web sayfası içeriği/HTML gibi ek veri kaynaklarına, ya da CNN-LSTM türü yüksek maliyetli derin mimarilere dayandığı görülmektedir. Bu durum, gerçek dünyadaki dengesiz ve büyük ölçekli URL akışlarında düşük maliyetle çalışabilen, içerik indirmeden yalnızca URL bilgisiyle yüksek başarı sağlayan hafif yöntemlere yönelik belirgin bir boşluk bırakmaktadır. Önerilen çalışma tam da bu boşluğu hedefleyerek; karakter düzeyi TF-IDF ile yapısal özellikleri tek bir hatta birleştiren, sınıf dengesizliğini çapraz doğrulama içinde ele alan ve karar eşiğini out-of-fold tahminlerle F1 odaklı optimize eden bütünleşik bir yaklaşımı, yarım milyondan fazla URL içeren dengesiz bir veri kümesi üzerinde değerlendirmektedir. Böylece çalışma; içerik bağımsız, gerçek zamanlı sistemlere kolayca entegre edilebilen ve büyük ölçekli dengesiz veride genellenebilirliğini koruyan düşük maliyetli bir alternatif sunarak literatüre katkı sağlamaktadır.

Sonuç olarak, önerilen çalışma; yalnızca URL verisine dayalı hafif bir sistem sunmakta, büyük ölçekli ve dengesiz veri kümelerinde rekabetçi doğruluğu düşük maliyetle elde etmekte ve doğrusal modellerle düşük hesaplama maliyeti ile yüksek genellenebilirlik sağlamaktadır.

Tablo 6. Mevcut çalışma ve literatür karşılaştırması

Çalışma	Model	Veri Seti	Doğruluk
Jalil vd. (2023)	Hybrid TF-IDF + Structural (LR), URL-Based (Random Forest)	Kaggle Phishing URL (507K), Kaggle Phishing URL (507K)	97,33, 96,25
Lakshmanarao vd. (2021a)	Hybrid URL (RF + Hashing Vectorizer)	Kaggle Phishing URL (507K)	97,50
Lakshmanarao vd. (2021b)	Fusion Model (RF + NB)	UCI Phishing Websites (11K)	97,00
Karim vd. (2023), Kumar (2023), Das Guptta vd. (2024)	Hybrid LSD (LightGBM + SVM + DNN), NLP-Based Hybrid (RF + PCA + Word2Vec), Hybrid Feature-Based (XGBoost)	PhishTank + UCI+ Kaggle, PhishTank + Alexa + Kaggle (15K), Özel URL + HTML (10K)	98,12, 99,75, 99,17
Ok ve Kara (2021), Rashid vd. (2020), Alam vd. (2020),Dalvi vd. (2019)	URL-Based (Random Forest), URL-Based (SVM), Traditional ML (Hybrid Features), Optimized ML (Tuned Model)	USOM Dataset (12K, Türkiye), UCI Phishing Websites (11K), Smart Systems Dataset (ICSSIT 2020), UCI Phishing Websites (11K)	95,00, 95,66, 94,50, 95,60

4. Sonuç

Bu çalışma, URL tabanlı kimlik avı saldırılarının tespitinde, yapısal özellikler ile karakter düzeyinde TF-IDF temsillerini birleştiren hibrit bir makine öğrenimi yaklaşımının etkinliğini kapsamlı bir şekilde ele almış ve farklı mimarilerin (Lojistik Regresyon, Ridge, SGD, Topluluk ve Ağaç tabanlı modeller) performanslarını deneysel sonuçlarla ortaya koymuştur. Bulgular, özellikle Lojistik Regresyon (LR) ve Ridge Sınıflandırıcı modellerinin yüksek doğruluk, ROC-AUC ve dengeli sınıflandırma ölçütleri (F1-Skoru) ile zararlı URL'leri güvenilir biçimde saptayabildiğini göstermiştir. Bununla birlikte, Random Forest (RF) gibi ağaç tabanlı modeller, yüksek boyutlu seyrek verilerde göreceli olarak daha düşük sonuçlar üretmiş ve eğitim maliyeti açısından en verimsiz modeller olarak konumlanmıştır.

Çalışmanın sonuçları, karakter tabanlı n-gram TF-IDF yönteminin sağladığı zengin metinsel temsillerin, kimlik avı gibi metin manipülasyonuna dayalı siber tehditlerin tespitinde kritik önem taşıdığını doğrulamaktadır. Özellikle Lojistik Regresyon modelinin %97,33 doğruluk ve 0,9400 F1-skoru ile gösterdiği üstün başarı, yüksek boyutlu ve seyrek veri uzaylarında lineer modellerin genelleme kapasitesinin, karmaşık ağaç yapılarından daha etkili olabildiğini göstermektedir.

Bununla birlikte, çalışmanın bazı sınırlılıkları bulunmaktadır. Önerilen yöntem yalnızca URL bilgisine dayandığından, dışarıdan meşru görünen ancak içerik düzeyinde zararlı davranış sergileyen sayfaların veya ele geçirilmiş alan adlarının tespitinde sınırlı kalabilir. Ayrıca modelin tek bir veri kümesi üzerinde değerlendirilmiş olması, farklı zamanlarda, farklı kaynaklardan veya farklı dağılımlardan gelen URL'ler karşısındaki genelleme performansının ayrıca incelenmesini gerekli kılmaktadır. Bu doğrultuda gelecek çalışmalarda yöntemin farklı ve güncel veri kümeleri üzerinde çapraz olarak değerlendirilmesi, hafif karakter düzeyi derin öğrenme temsilleriyle karşılaştırılması, kavram kaymasına karşı artımlı öğrenme stratejilerinin incelenmesi ve modelin gerçek zamanlı bir servis olarak konuşlandırılması hedeflenmektedir.

Genel olarak bu çalışma, Lojistik Regresyon gibi lineer modellerin, uygun özellik temsili ve eşik optimizasyonu ile birlikte kullanıldığında siber tehdit analizinde hem yüksek başarı hem de operasyonel verimlilik sunabileceğini göstermektedir.

Katkı Oranı Beyanı

Yazarların katkı yüzdeleri aşağıda verilmiştir. Yazarlar makaleyi incelemiş ve onaylamıştır.

	Ü.Ö.	E.O.Y
K	50	50
T	50	50
Y	50	50
VTI	50	50
VAY	50	50
KT	50	50
YZ	60	40
KI	50	50
GR	50	50
PY	50	50

K= kavram, T= tasarım, Y= yönetim, VTI= veri toplama ve/veya işleme, VAY= veri analizi ve/veya yorumlama, KT= kaynak tarama, YZ= yazım, KI= kritik inceleme, GR= gönderim ve revizyon, PY = proje yönetim.

Çatışma Beyanı

Bu çalışmaya ilişkin olarak yazarlar herhangi bir çıkar çatışması bulunmadığını beyan etmektedir.

Etik Onay Beyan

Bu çalışma insan veya hayvan denekler üzerinde herhangi bir deney içermediğinden etik kurul onayı gerektirmemektedir.

Kaynaklar

Al Tawil, A., Almazaydeh, L., Qawasmeh, D., Qawasmeh, B., Al Shinwan, M., & Elleithy, K. (2024). Comparative analysis of machine learning algorithms for email phishing detection using TF-IDF, Word2Vec, and BERT. *Computers, Materials & Continua*, *81*(2), 3395–3412. https://doi.org/10.32604/cmc.2024.057279

Alam, M. N., Saha, I., Sarma, D., Ulfath, R. E., Lima, F. F., & Hossain, S. (2020). Phishing attacks detection using machine learning approach. In *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)* (pp. 1173–1179). IEEE. https://doi.org/10.1109/ICSSIT48917.2020.9214225

Aljofey, A., Jiang, Q., Qu, Q., Huang, M., & Niyigena, J. P. (2022). An effective detection approach for phishing websites using URL and HTML features. *Scientific Reports*, *12*(1), Article 8842. https://doi.org/10.1038/s41598-022-10841-5

Anandaraj, A. P. S., Ramesh, P. S., Arifulla, S., Madhuri, C. G., & Malepati, N. Y. (2024). Phishing detection system through hybrid machine learning based on URL. In *2024 Third International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)* (pp. 1–7). IEEE. https://doi.org/10.1109/ICEEICT61591.2024.10718484

Asha Shajilal, B. (2024). A hybrid approach for detecting phishing mails using textual, content, and URL analysis with ensemble learning [Master's thesis, National College of Ireland]. NORMA@NCI Library.

BTK. (2025). *Oltalama saldırıları ve zararlı bağlantı istatistikleri* [Phishing attacks and malicious link statistics]. Bilgi Teknolojileri ve İletişim Kurumu. https://www.btk.gov.tr

Dalvi, S., Gressel, G., & Achuthan, K. (2019). Tuning the false positive rate / false negative rate with phishing detection models. *International Journal of Engineering and Advanced Technology, 9*(1S5), 7–13. https://doi.org/10.35940/ijeat.A1002.1291S519

Das Guptta, S., Shahriar, K. T., Alqahtani, H., Alsalman, D., & Sarker, I. H. (2024). Modeling hybrid feature-based phishing websites detection using machine learning techniques. *Annals of Data Science, 11*(1), 217–242. https://doi.org/10.1007/s40745-022-00379-8

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Hosmer, D. W., Jr., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). John Wiley & Sons. https://doi.org/10.1002/9781118548387

Jalil, S., Usman, M., & Fong, A. (2023). Highly accurate phishing URL detection based on machine learning. *Journal of Ambient Intelligence and Humanized Computing*, *14*(7), 9233–9251. https://doi.org/10.1007/s12652-022-04426-3

Kaggle. (n.d.). *Phishing site URLs dataset*. https://www.kaggle.com/datasets/taruntiwarihp/phishing-site-urls (erişim tarihi: 08 Kasım 2025).

Karim, A., Shahroz, M., Mustofa, K., Belhaouari, S. B., & Joga, S. R. K. (2023). Phishing detection system through hybrid machine learning based on URL. *IEEE Access*, *11*, 36805–36822. https://doi.org/10.1109/ACCESS.2023.3252366

Kumar, J. (2023). Hybrid feature-based machine learning method for phishing URL detection. In *2023 Third International Conference on Secure Cyber Computing and Communication (ICSCCC)* (pp. 222–227). IEEE. https://doi.org/10.1109/ICSCCC58608.2023.10176901

Lakshmanarao, A., Babu, M. R., & Bala Krishna, M. M. (2021a). Malicious URL detection using NLP, machine learning and FLASK. In *2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)* (pp. 1–4). IEEE. https://doi.org/10.1109/ICSES52305.2021.9633889

Lakshmanarao, A., Rao, P. S. P., & Krishna, M. M. B. (2021b). Phishing website detection using novel machine learning fusion approach. In *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)* (pp. 1164–1169). IEEE. https://doi.org/10.1109/ICAIS50930.2021.9395810

Maneriker, P., Stokes, J. W., Lazo, E. G., Carutasu, D., Tajaddodianfar, F., & Gururajan, A. (2021). URLTran: Improving phishing URL detection using transformers. In *MILCOM 2021 – 2021 IEEE Military Communications Conference (MILCOM)* (pp. 197–204). IEEE. https://doi.org/10.1109/MILCOM52596.2021.9653028

Mathew, J. T., Nair, L. S., Menon, H. P., & A. P., A. (2025). PhishGuard: A hybrid ensemble framework for detecting malicious URLs. In *2025 8th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 1901–1908). IEEE. https://doi.org/10.1109/ICCMC65190.2025.11140768

Ogunleye, G. O., Ayogu, B. A., Tolulope, O., & Ogunrekun, P. O. (2024). Comparative analysis of some machine learning techniques for uniform resource locator-based phishing detection. *Journal of Engineering, Technology and Innovation*, *3*(2), 1–12.

Ok, M., & Kara, İ. (2021). İnternet sayfası verileri kullanarak kimlik avı web sitelerinin makine öğrenme tekniğiyle tespiti [Detection of phishing web sites using web page data with machine learning technique]. *Avrupa Bilim ve Teknoloji Dergisi*, (31), 182–187. https://doi.org/10.31590/ejosat.1008335

Palavali, D. R., & Pothireddy, S. (2025). Explainable ensemble learning for detecting phishing URLs using lightweight cyber threat intelligence. In *2025 9th International Conference on Electronics, Communication and Aerospace Technology (ICECA)* (pp. 207–214). IEEE. https://doi.org/10.1109/ICECA66444.2025.11382884

Rashid, J., Mahmood, T., Nisar, M. W., & Nazir, T. (2020). Phishing detection using machine learning technique. In *2020 First International Conference of Smart Systems and Emerging Technologies (SMARTTECH)* (pp. 43–46). IEEE. https://doi.org/10.1109/SMART-TECH49988.2020.00026

Sangeetha, T. S., & Nair, P. P. (2025). Phishing URL detection with deep learning: Trends, challenges, and future directions. In *2025 International Conference on Power, Instrumentation, Control, and Computing (PICC)* (pp. 1–8). IEEE. https://doi.org/10.1109/PICC67314.2025.11291322

Siber Güvenlik Kanunu [Cyber Security Law]. (2025). Kanun No. 7545. *T.C. Resmî Gazete* (No. 32846, 19 March 2025). https://www.resmigazete.gov.tr (erişim tarihi: 01 Kasım 2025).

USOM. (2024). *Ulusal siber güvenlik istatistikleri* [National cyber security statistics]. Ulusal Siber Olaylara Müdahale Merkezi. https://www.usom.gov.tr (erişim tarihi: 11 Kasım 2025).