

EMBER 2024 Üzerinde 5-Fold ve Temporal Değerlendirme ile Malware Sınıflandırma: Hibrit SAE–Top-K–LightGBM Modeli

¹Sultan Tazefidan, ²Heya Meylem ve ³Gülay Çiçek

^{1,2} Yazılım Mühendisliği Departmanı, Mühendislik-Mimarlık Fakültesi
İstanbul Beykent Üniversitesi, Sarıyer, İstanbul, Türkiye

¹sultantazefidan.1@gmail.com ²heyameylem96@gmail.com ³gulaycicek@beykent.edu.tr

I. DENEYSEL ANALİZ

A. Veri Seti ve Metodolojik Özeti

1) Veri Seti Tanımı ve Bölme Stratejisi

Bu çalışmada kullanılan veri seti, EMBER2024 (Win32) veri setidir; veri kaynağı olarak GitHub üzerindeki [EMBER 2024 GitHub Repostosu](#) adresi referans alınmıştır. EMBER 2024 veri seti, zararlı yazılım tespiti alanında global güncel ve kapsamlı bir benchmark olarak öne çıkmaktadır. Yüksek boyutlu, dengeli ve zengin özellik yapısıyla, makine öğrenmesi ve derin öğrenme tabanlı modellerin gerçekçi koşullarda değerlendirilmesine olanak tanımaktadır. GitHub Veri setinin orijinal boyutu, Win32 formatında yaklaşık 1.56 milyon eğitim örneği ve 360.000 test örneği olarak tanımlanmıştır.

Bu çalışmada, 5-Fold çapraz doğrulama deneyi ve model kararlılığı analizi için veri setinin ilk 52 haftasından her hafta 3.850 benign (0) ve 3.850 malicious (1) örnek seçilerek dengeli bir alt küme oluşturulmuştur. Böylece toplam 400.400 örnekten oluşan bir çalışma seti elde edilmiştir. Şekil 1'de görüldüğü üzere, bu alt küme içerisinde 3 yinelenen kayıt tespit edilip temizlenmiş ve nihai veri seti 400.397 örneğe (Benign: 200.198, Malicious: 200.199) düşmüştür. Her haftadan eşit miktarda örnek alınması, veriyi kronolojik olarak dengede tutarak zaman temsili sağlamış; modelin farklı dönemlerdeki davranış örüntülerini görebilmesine ve temporal bias'in azaltılmasına önemli katkı sunmuştur. Bu yaklaşım sayesinde model, belirli haftaların baskınılığinden kaynaklanabilecek yanlışlıklardan arındırılarak daha tutarlı bir genellemeye yapabilir hale gelmiştir.

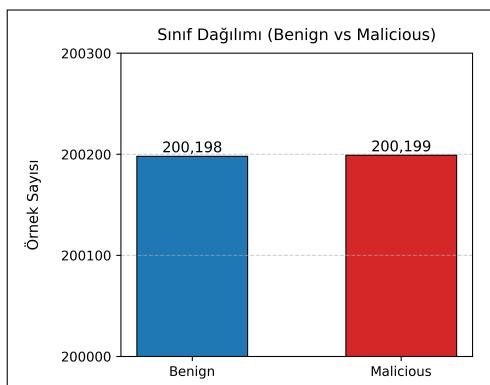


Fig. 1: Sınıf dağılımının görsel temsili (Benign vs Malicious).

Eğitim ve doğrulama süreçlerinde bu dengeli alt küme üzerinde Stratified K-Fold ($k = 5$) yöntemi uygulanmıştır. Bu metodoloji, yalnızca overfitting riskini azaltmakla kalmamış, aynı zamanda literatürdeki çalışmalara kıyasla daha kapsamlı, sistematik ve metodolojik açıdan güçlü bir değerlendirme çerçevesi sunmuştur.

Ayrıca, zaman temelli (temporal) analizi desteklemek amacıyla, modelin hiç görmediği verilerden oluşan kilitli nihai test seti hazırlanmıştır. Bu test seti, ana deneyde kullanılan stratejiye benzer biçimde oluşturulmuş ve veri setinin son 12 haftasından, benign ve malicious sınıflardan eşit sayıda örnek seçilerek dengeli hale getirilmiştir. Zaman kronolojisinin bozulmaması için bu aşamada Stratified K-Fold yerine tek seferlik (hold-out) bölme stratejisi uygulanmış ve yaklaşık 200.000 eğitim ile 200.000 test örneği olacak şekilde ayırmaya çalışılmıştır.

Bu bölüm stratejisi, modelin yalnızca rastgele bölünmüş veriler üzerindeki performansını değil, aynı zamanda zaman ilerledikçe karşılaşacağı örneklerle karşı sereceği genellemeye yeteneğini değerlendirmeyi amaçlamaktadır. Bu kapsamlı analiz, ilerleyen bölgelerde ayrıntılı olarak sunulmuştur.

2) Veri Ön İşleme ve Öznitelik Çıkarma

Veri ön işleme aşaması bu çalışmada büyük bir titizlik ve sistematiğin yanı sıra yürütmüştür. İlk olarak EMBER 2024 veri setinin ham .jsonl yapısı incelenmiş; etiket dağılımı, duplicate kayıtlar, SHA-256 bütünlük doğrulaması, satır eşleşme kontrolleri ve alan yapısı detaylı biçimde analiz edilmiştir. İlk vektörleştirme adımda 3.12 milyon kayıt tespit edilmiş, tekilleştirme sonunda eğitim seti 1.56 milyon, test seti ise 359,994 örnekten oluşan şekilde temizlenmiştir. Test setindeki 6 eksik satırın hatalı veya eksik etiketli olduğu belirlenmiş ve filtreleme mekanizması tarafından otomatik olarak çıkarılmıştır.

Bu aşamadan sonra, veri setine hafta (week) bilgisi enjekte edilerek kronolojik bütünlük sağlanmış; haftalar ilk haftadan son haftaya kadar dengeli biçimde düzenlenmiş ve her hafta için örnek sayıları ilgili log dosyalarıyla birlikte kaydedilmiştir. Böylece ilk 52 haftayı kapsayan, yinelenen kayıtları giderilmiş ve kronolojik olarak düzenlenmiş "unique" 5-Fold deney altkümesi oluşturulmuştur. Aynı süreç, temporal analizde kullanılan test seti için de birebir uygulanmıştır.

Ardından ham veriler doğrudan kullanılmadığı için .jsonl dosyaları vektörleştirme aşamasına tabi tutulmuştur. Bu süreçte yalnızca ilk 52 haftayı içeren eğitim periyodu işlenmiş, hata önleme amacıyla win32_train.jsonl ve win32_challenge.jsonl için placeholder kayıtlar eklenmiştir. thremler aracılıyla elde edilen tüm feature vektörleri ve mapping dosyaları çıktı klasörüne

taşınmış, ardından thremler.read_vectorized_features() ile float32 girdi özellikleri ve int8 etiketler okunarak .parquet formatında nihai vektörleştirilmiş alt küme oluşturulmuştur.

Vektörleştirme sonrasında, kapsamlı ön işleme adımları uygulanmıştır. İlk olarak tüm metadata + PE tabanlı mühendislik özellikleri sayısal forma dönüştürülmüş; ardından düşük varyans kontrolü uygulanmış ve 32 sabit/nereye sabit özellik kaldırılarak boyut 2.536'ya düşürülmüştür. Daha sonra yüksek korelasyon analizi yapılmış; Pearson $|\rho| \geq 0.95$ olan 81 özellik çifti tespit edilmiş ve tekrarlayıcı nitelikteki 31 özellik çıkarılarak nihai özellik sayısı 2.505 olarak belirlenmiştir.

Hem 5 fold çapraz doğrulama deney hem de temporal subset üzerinde NaN/Inf kontrolleri, duplicate taraması, bellek kullanımı denetimi, şekil/dtype kontrolü, özellik indeksi-etiket eşleşmesi doğrulamaları ve temporal set için orijinal JSON dosyalarıyla tutarlılık kontrolleri uygulanmıştır. Ayrıca tüm modelleme adımlarında, VarianceThreshold (threshold = 0.0) ile sabit sütunlar ek güvenlik önlemi olarak yeniden temizlenmiş; her işlem sonrası bulletproof_clean() fonksiyonu ile veri yapısı sağlamlaştırılmıştır. Veri kümelerinde sınıf dengesi-zligi bulunmadığından SMOTE/ADASYN gibi sentetik yöntemler kullanılmamış; bunun yerine alt kümeler doğal olarak dengeli olacak şekilde oluşturulmuş ve modeller yalnızca gerçek gözlemler üzerinden eğitilmiştir.

Bu çalışmada uygulanan tüm temel veri işleme adımları Şekil 2'de bütüncül bir iş akışı olarak özetlenmiştir.

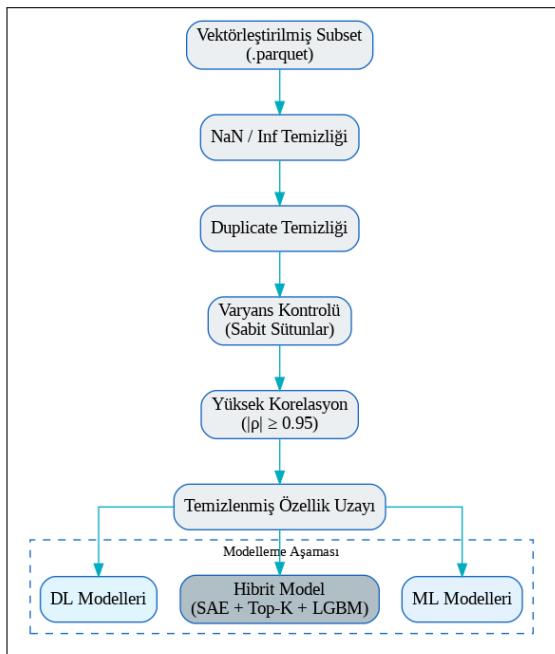


Fig. 2: Veri ön işleme ve genel modelleme iş akışı

3) Normalizasyon

Normalizasyon, farklı ölçeklerdeki özelliklerin model üzerindeki etkisini dengeleyerek öğrenme sürecinin daha kararlı ilerlemesini sağlar. Bu işlem sayesinde, model parametreleri daha dengeli güncellenir ve özellikle derin öğrenme yapılarında gradyan kararsızlıklar önemli ölçüde azaltılır.

Makine Öğrenimi Modellerinde Normalizasyon Stratejisi: Makine öğrenmesi tabanlı modellerde, özellikler arasında üç değerlerin yoğunlukta bulunması nedeniyle RobustScaler yöntemi tercih edilmiştir. Bu ölçekleyici, medyan ve IQR (Interquartile

Range) temelli yapısı sayesinde üç değerlerin model performansını olumsuz etkilemesini önlemekte ve değişkenlerin ölçek farklarından kaynaklanan bozulmaları azaltmaktadır. Özellikle EMBER gibi yüksek varyansa ve geniş değer aralıklarına sahip veri setlerinde, RobustScaler'in medyan ve IQR tabanlı yaklaşımı, üç değerlerin baskın etkisini minimize ederek sıralama ve bölünme noktalarının daha kararlı biçimde belirlenmesini sağlar. Böylece ağaç tabanlı modellerin daha istikrarlı kararlar alması, genellemeye gücünün artması ve üç değerlerden etkilenmeden doğru bölünmeler gerçekleştirmesi mümkün hale gelmiştir.

Derin Öğrenme Modellerinde Normalizasyon Stratejisi: Derin öğrenme ve lojistik regresyon modellerinde, ağ ağırlıklarının dengeli biçimde öğrenebilmesi ve gradyan güncellemelerinin kararlılığının sağlanması amacıyla StandardScaler yöntemi tercih edilmiştir. Bu ölçekleyici, tüm özellikler ortalaması sıfır ve varyansı bir olacak şekilde dönüştürerek veriyi z-skor normalizasyonuna tabi tutar. Bu dönüşüm, özellikle sınır ağı tabanlı modellerde aktivasyon fonksiyonlarının giriş değerlerini ideal aralıktaki (ortalama 0, varyansı 1 civarında) tutarak eğitim sürecinin daha stabil ilerlemesini sağlar. Böylece gradyan kaybı (vanishing gradient) riski azalır, Batch Normalization ve Dropout katmanlarının etkinliği artar, ve modelin daha hızlı ve kararlı yakınsaması mümkün hale gelir.

4) Boyut İndirgeme Yöntemleri

EMBER 2024 veri setinin öznitelik dağılımı sağa çarpık, uzun kuyruklu ve normal (Gaussian) olmayan bir yapıya sahiptir. Önceki sürümde 2.381 öznitelik bulunurken, son sürüm (EMBER24-v3) ile bu sayı 2.568'e yükselmiştir. Bu artış, DOS header, RVH header ve PE parse warning gibi yeni kategorilerin eklenmesinden kaynaklanmaktadır, antivirüs yazılımlarını atlatabilen karmaşık kötü amaçlı yazılımların daha doğru tespit edilmesini sağlayarak modellerin gerçek dünya senaryolarında daha güvenilir biçimde eğitilmesine katkı sağlamaktadır.

Tüm 2.568 öznitelinin doğrudan kullanımı veri zenginliği sağlasa da, yüksek korelasyon ve gürültü nedeniyle bazı modellerde performans dalgalanmalarına ve gereksiz karmaşıklığa yol açabilmektedir. Bu nedenle, çalışmada hem ham öznitelikli (OFF) yapı korunmuş hem de boyut indirgeme aşamasında bilgi açısından daha zengin temsiller elde etmek amacıyla PLS-DA (Partial Least Squares Discriminant Analysis) ve VAE (Variational Autoencoder) yöntemleri uygulanmıştır. Bu iki yöntemle elde edilen indirgenmiş temsiller üzerinde modeller ayrı ayrı eğitilmiş; böylece hem orijinal özelliklerle doğrudan performans karşılaştırması yapılmış hem de boyut indirgeme sonrasında modelin genellemeye kabiliyeti ve işlem verimliliği analiz edilmiştir.

PLS-DA: PLS-DA, sınıf etiketleri ile özellikler arasındaki kovaryansi maksimize eden gözetimli bir boyut indirgeme yöntemidir. Veri boyutunu azaltırken sınıflar arasındaki ayrimı daha belirgin bir alt uzaya projekte eder [1]. Bu çalışmada sınıflandırma problemlerine uygun olarak PLS-DA yaklaşımı uygulanmıştır. Kodlama aşamasında, sklearn kütüphanesindeki PLSRegression sınıfı kullanılmış; hedef değişkenin ikili (0–1) yapıda olması nedeniyle yöntem filen PLS-DA prensipleri doğrultusunda çalışmıştır. Çalışmada kullanılan temel PLS-DA hiperparametreleri Tablo I'de sunulmuştur.

TABLE I: Çalışmada kullanılan PLS-DA temel hiperparametreleri

| Parametre | Değer |
|--|--------------------|
| Bileşen sayısı (<i>n_components</i>) | 64 |
| Ölçekleme (<i>scale</i>) | False |
| Maksimum iterasyon (<i>max_iter</i>) | 1000 |
| Eşik değeri (<i>tol</i>) | 1×10^{-6} |

VAE: VAE, olasılıksal temelli bir derin öğrenme yöntemidir ve boyut indirmeye ile veri temsili amacıyla kullanılmaktadır. Klasik otoencoderlerden farklı olarak, VAE giriş verilerini sabit bir latent vektöre değil, olasılık dağılımına (mean ve variance) dönüştürür. Bu sayede model, verinin altında yatan istatistiksel yapıyı öğrenir ve benzer özellikte yeni örnekler üretebilir [2]. Bu çalışmada kullanılan veri setinin yapısına uygun olarak, VAE modeli; doğrusal ilişkileri latent uzayda etkin biçimde öğrenebilmesi ve gürültülü, karmaşık özniteliklerden bilgi açısından zengin temsil vektörleri üretebilmesi nedeniyle tercih edilmiştir. Elde edilen bu temsil uzayı kullanılarak hibrit modelin performans değerlendirmesi gerçekleştirılmıştır. Çalışmada kullanılan temel β -VAE hiperparametreleri Tablo II'de özetlenmiştir.

TABLE II: Temel β -VAE Hiperparametreleri

| Parametre | Değer |
|--------------------------|------------------------------------|
| Latent boyutu | 256 |
| Encoder katmanları | 512 → 256 (ReLU) |
| Düzenleştirme | Dropout(0.10), GaussianNoise(0.01) |
| β değeri (warm-up) | 0.0 → 1.0 (10 epoch) |
| Optimizer | Adam ($lr = 1e-3$) |
| Batch size | 256 |

5) Eşik Optimizasyonu (Youden's J ile Threshold Tuning)

Sınıflandırıcıdan elde edilen olasılık skorları üzerinde ROC eğrisi tabanlı bir eşik optimizasyonu uygulanmıştır. Her fold'da, model yalnızca eğitim verisiyle eğitildikten sonra, tutulan held-out (fold'un test parçası) üzerinde ROC eğrisi hesaplanmış ve Youden'in J istatistiğini

$$J = TPR - FPR$$

maksimize eden karar eşigi belirlenmiştir. Böylece sabit 0.5 yerine veri sete özgü optimal eşik seçilmiş ve eşik değeri fold'daki tüm değerlendirmelerde tutarlı olarak uygulanmıştır. Başka bir deyişle, her fold'da ROC eğrisi üzerinden Youden J istatistiği ile optimal karar eşigi hesaplanmış ve aynı fold'daki eğitim/test değerlendirmelerinde kullanılmıştır. Bu yaklaşım, IDS senaryolarında kritik olan yanlış negatifleri azaltmayı ve Recall/F1 değerlerini iyileştirmeyi hedeflemektedir; pratikte sabit eşik kulanımına kıyasla daha dengeli bir hata profili sağlamaktadır.

Kısa not: En katı değerlendirme için eşigin, train verisi içinden ayrılan bağımsız bir validation diliminde seçilmiş test setinde yalnızca uygulanması önerilmektedir. Ancak bu çalışmada, katmanlı çapraz doğrulamanın tutarlılığını korumak amacıyla eşik her fold'un held-out parçası üzerinde seçilmiş ve fold içinde hem train hem test değerlendirmelerinde aynı eşik değeri τ^* kullanılmıştır. Bu yöntem, veri sizintisini önlerken eşik-bağımlı metriklerin karşılaştırılabilirliğini arttırmıştır.

B. Deneysel Kurulum

1) Donanım/Yazılım Ortamı

Deneysel çalışmalar, Windows 11 Pro 64-bit işletim sistemine sahip bir masaüstü bilgisayarda gerçekleştirılmıştır. Kullanılan sisteme AMD Ryzen 7 9700X işlemci, NVIDIA GeForce RTX 5060 Ti (16 GB) GPU ve 32 GB DDR5 RAM bulunmakta olup, tüm deneyler Python 3.12 ortamında PyCharm IDE üzerinde yürütülmüştür. Kullanılan temel yazılım kütüphaneleri sırasıyla TensorFlow 2.16, scikit-learn 1.5, NumPy 1.26 ve Matplotlib 3.9 sürümleridir. Tüm modeller, Stratified KFold ($n_{splits} = 5$, $random_state = 42$) yapısı altında tekrarlanabilir biçimde değerlendirilmiştir.

Bu çalışmada tüm modeller, tutarlı karşılaştırma sağlamak amacıyla yalnızca CPU üzerinde çalıştırılmış; GPU hızlandırması hiçbir aşamada kullanılmamıştır.

C. Model Grupları ve Karşılaştırma Seti

1) ML Modelleri

Bu grupta, klasik tabanlı makine öğrenmesi algoritmaları yer almaktadır. Her bir modelin temel çalışma prensibi aşağıda özetlenmiştir.

- **Gradient Boosting (GB):** Art arda zayıf karar ağaçları oluşturarak hata oranını azaltan topluluk tabanlı bir yöntemdir.
- **Histogram-based Gradient Boosting (HistGB):** GB algoritmasının daha hızlı ve bellek açısından verimli bir versiyonudur. Büyük veri setlerinde histogram yaklaşımıyla hesaplama yükünü azaltır.
- **CatBoost:** Kategorik verileri doğrudan işleyebilen, veri dengesizliğini azaltan ve aşırı uyum (overfitting) riskini düşüren bir boosting algoritmasıdır.
- **ExtraTrees (Extremely Randomized Trees):** Rastgele seçilen özellik ve eşik değerleriyle çalışan, yüksek çeşitlilik sağlayan bir topluluk öğrenme yöntemidir.
- **SGDClassifier:** Stokastik gradyan inişiyle optimize edilen, doğrusal sınıflandırıcıların (ör. lojistik regresyon, SVM) hızlı ve basit bir versiyonudur.

Bu çalışmada seçilen makine öğrenmesi algoritmaları, EMBER24 veri setinin yüksek boyutlu, dengesiz ve doğrusal olmayan yapısını farklı açılardan ele alabilecek tamamlayıcı modellere dayanmaktadır. Bu beş algoritmanın birlikte seçilme amacı, farklı öğrenme stratejilerinin (boosting, bagging, linear optimization) aynı veri üzerinde nasıl genelleme sağladığını bütüncül biçimde analiz etmektedir. Böylece model karmaşaklı ve doğruluk arasındaki denge EMBER24 veri setinin yapısına uygun biçimde karşılaştırmalı olarak değerlendirilmiştir.

2) DL Modelleri

Derin öğrenme grubu, karmaşık veri örüntülerini ve temsil gücünü analiz etmek üzere oluşturulmuştur.

- **Wide & Deep Model:** Geniş (memorization) ve derin (generalization) bileşenleri birleştirerek hem doğrusal hem doğrusal olmayan ilişkileri öğrenir.
- **DNN (Deep Neural Network):** Çok katmanlı ileri beslemeli sinir ağı yapısıdır ve temel derin öğrenme modeli olarak görev yapar.
- **MLP-Mixer** MLP-Mixer, evrişim veya dikkat (attention) katmanları yerine yalnızca MLP blokları kullanarak hem uzamsal (patch) hem de kanal boyutlarında özellik karışımı yapan saf MLP tabanlı bir modeldir.
- **ResNet+MLP:** Derin artık öğrenme (residual learning) tabanlı evrişimsel katmanları çok katmanlı algılayıcı (MLP) ile birleştirerek karmaşık temsilleri yakalar.
- **gMLP-Tabular:** gMLP-Tabular, klasik MLP yapısına "gated" (kapılı) lineer katmanlar ekleyerek öznitelikler arası etkileşimi dinamik biçimde öğrenen hafif bir derin ağ mimarisidir. Özellikle tabular verilerde düşük parametre sayısıyla verimli genelleme sağlar.

Bu çalışmada seçilen derin öğrenme mimarileri, farklı öğrenme stratejilerini temsil eden tamamlayıcı yapılarından oluşmaktadır. DNN saf ve klasik MLP yapısını temsil ederken, ResNetMLP, gMLP ve MLP-Mixer modelleri MLP'nin geliştirilmiş varyantlarıdır; her biri farklı bir yapısal yenilikle klasik MLP'nin sınırlarını genişletmektedir. Özellikle MLP ailesine ait üç modelin (DNN,

gMLP-Tabular ve MLP-Mixer) çalışmada yer almasının temel gerekçesi, bu yapıların evrişimsel veya dönüşüm tabanlı karmaşık yapılara kıyasla daha düşük parametre maliyetiyle, tabular ve yüksek boyutlu veri setlerinde hızlı, hafif ve genelleme gücü yüksek sonuçlar verebilmesidir. Böylece farklı derin mimari tiplerinin karşılaştırılması hem model karmaşıklığına hem de temsil gücüne göre bütüncül biçimde değerlendirilmiştir.

3) Hiperparametre Uyarlaması (ML, DL ve Hibrit İçin Ortak Çerçeve))

Tablo III’de, önerilen hibrit mimaride kullanılan temel hiperparametreler, Tablo IV’de Makine öğrenmesi modellerinde kullanılan temel hiperparametreler ve derin öğrenme modellerine ait ayarlar ise Tablo V’de sunulmuştur. Bu tablolar, modellerin temel yapılandırma ayarlarını karşılaştırmalı olarak sunmaktadır. Belirlenen parametre kombinasyonları, modellerin dengeli bir öğrenme süreci gerçekleştirmesine ve yüksek genelleme başarımı elde etmesine katkı sağlamaktadır. Genel olarak, her üç model grubu (ML, DL ve Hibrit) için yapılandırma değerleri, deneysel bütünlüğün korunmasını ve adil bir performans karşılaştırmasının yapılabilmesini mümkün kılmıştır.

TABLE III: Önerilen hibrit modelin temel hiperparametreleri

| Parametre | Bileşen | Değer |
|----------------------|---------------|----------------|
| Latent Boyutu | SAE (Encoder) | 256 |
| Dropout Oranı | SAE (Encoder) | 0.10 |
| Epoch | SAE (Encoder) | 40 |
| Batch Size | SAE (Encoder) | 512 |
| Optimizer | SAE (Encoder) | Adam($1e-3$) |
| Öznitelik Sayısı (K) | Top-K | 220 |
| Önem Sıralaması | Top-K | LightGBM |
| n_estimators | LGBM | 4000 |
| learning_rate | LGBM | 0.02 |
| num_leaves | LGBM | 72 |
| max_depth | LGBM | 10 |
| min_child_samples | LGBM | 140 |

TABLE IV: Makine öğrenmesi modellerinde kullanılan temel hiperparametreler.

| Model | Parametre | Değer |
|---------------|-------------------|------------|
| GradientBoost | n_estimators | 90 |
| GradientBoost | learning_rate | 0.1 |
| GradientBoost | max_depth | 6 |
| HistGB | max_iter | 100 |
| HistGB | learning_rate | 0.1 |
| HistGB | l2_regularization | 1.0 |
| CatBoost | iterations | 300 |
| CatBoost | learning_rate | 0.1 |
| CatBoost | depth | 8 |
| ExtraTrees | n_estimators | 500 |
| ExtraTrees | max_features | ”sqrt” |
| ExtraTrees | bootstrap | True |
| SGDClassifier | loss | log_loss |
| SGDClassifier | penalty | elasticnet |
| SGDClassifier | learning_rate | adaptive |

TABLE V: Derin öğrenme modellerinde kullanılan temel hiperparametreler.

| Model | Parametre | Değer |
|---------------------|--------------------|----------------|
| WideDeep | deep_units | (512, 256, 64) |
| WideDeep | wide_units | 1 |
| WideDeep | dropout | 0.0 |
| WideDeep | activation | ReLU |
| DNN | units | (512, 256, 64) |
| DNN | dropout | 0.2 |
| DNN | activation | ReLU |
| MLP_Mixer | blocks | 2 |
| MLP_Mixer | mix | 128 |
| MLP_Mixer | hidden | 256 |
| MLP_Mixer | dropout | 0.1 |
| ResNetMLP | u1, u2, bottleneck | 512, 256, 64 |
| ResNetMLP | dropout | 0.3 |
| ResNetMLP | activation | ReLU |
| gMLP-Tabular (Tiny) | blocks | 2 |
| gMLP-Tabular (Tiny) | dim_ff | 256 |
| gMLP-Tabular (Tiny) | dropout | 0.1 |
| gMLP-Tabular (Tiny) | activation (gate) | Sigmoid |

D. Değerlendirme Metrikleri

Bu çalışmada modellerin değerlendirilmesi ve karşılaştırılmasında kullanılan temel metrikler; Doğruluk (Accuracy), Hassasiyet (Precision), Duyarlılık (Recall), F1-Skoru, Ağırlıklı Ortalama F1-Skoru (F1-W), AUC-ROC ve Özgüllük (Specificity) olarak belirlenmiştir.

Tüm modeller, 5-katlı Stratified K-Fold çapraz doğrulama yöntemi kullanılarak değerlendirilmiştir. Bu yaklaşım, her bir modelin veri kümesinin farklı bölgelerinde tutarlı performans sergilemesini sağlayarak genelleme gücünün istatistiksel olarak güvenilir biçimde ölçülmesine olanak tanımaktadır. Ayrıca, her katmandan elde edilen sonuçların ortalamasının alınmasıyla varyans etkisi azaltılmış ve modellerin tekrarlanabilirlik düzeyi güçlendirilmiştir. Bu değerlendirme sürecinin genel akışını gösteren sadeleştirilmiş işlem hattı Şekil 3’de sunulmuştur.

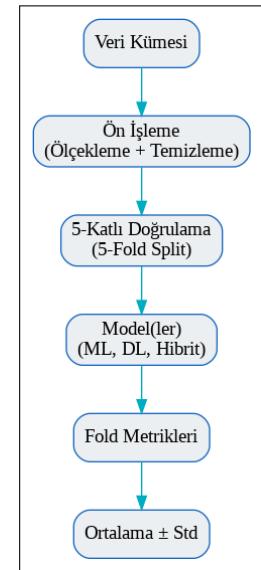


Fig. 3: Tüm modeller (ML, DL ve Hibrit) için uygulanan 5-Katlı çapraz doğrulama işlem hattı.

E. Hibrit Modelle Dair Bilgiler

Bu çalışmada önerilen hibrit model; derin öğrenme tabanlı SAE yapısı, kazanç temelli öznitelik seçim yaklaşımı olan Top-K yöntemi ve ağaç tabanlı gradyan artırmalı LightGBM sınıflandırıcısından oluşan çok aşamalı bir mimariye sahiptir. Tüm modellerde olduğu gibi, hibrit modelde de analiz öncesiinde üç değerlerin giderilmesi, NaN/Inf değerlerinin temizlenmesi ve varyans eşigi kontrolü gibi temel ön işlemler uygulanarak veri tutarlılığı ve bütünlüğü sağlanmıştır. Ayrıca, her fold'da yalnızca eğitim alt kümesi üzerinde fit edilen *StandardScaler* yöntemi ile veri standartlaştırılmış; böylece değişkenler arasındaki ölçek farklılıklarının model performansı üzerindeki olası etkileri azaltılmıştır.

1) SAE

SAE, verinin gizli temsillerini katmanlı biçimde öğrenen, birden fazla autoencoder'in ardışık olarak birleştirildiği derin öğrenme mimarisidir [3]. Bu çalışmada kullanılan SAE mimarisi, yüksek boyutlu öznitelik uzayını daha kompakt ve anlamlı temsillere dönüştürmek amacıyla tasarlanmış yarı simetrik (semi-symmetric) bir derin öğrenme yapısıdır. Encoder kısmı sırasıyla 512, 256 ve 256 nöronlu tam bağlantılı (Dense) katmanlardan oluşmaktadır; her katmanda ReLU aktivasyon fonksiyonu uygulanmıştır. Öğrenme sürecinin kararlılığını artırmak ve iç dağılım kaymasını azaltmak için her yoğun katmanın ardından Batch Normalization katmanı eklenmiştir. Ayrıca Dropout (0.10) oranı uygulanarak aşırı öğrenme (overfitting) riski azaltılmıştır.

Encoder'in çıkış katmanında elde edilen 256-boyutlu latent temsil, verinin daha az gürültüyle, bilgi açısından zengin bir biçimde özetlenmesini sağlamıştır. Decoder kısmında ise bu latent vektör, simetrik biçimde 256 nöronlu bir katmandan geçirilerek yeniden yapılandırılmış ve çıkışta orijinal boyuta sahip linear aktivasyonlu bir katman kullanılmıştır. Modelin eğitimi, Adam optimizasyon algoritması (learning rate = 1e-3) ve ortalama karesel hata (MSE) kayıp fonksiyonu ile gerçekleştirilmiştir. Bu yapı, verideki karmaşık örüntülerin doğrusal olmayan biçimlerde yakalanmasını ve daha anlamlı öznitelik temsillerinin öğrenilmesini sağlamıştır.

2) Top-K Öznitelik Seçimi

Top-K öznitelik seçimi, modelin performansına en çok katkı sağlayan en iyi K adet özelliğin istatistiksel önem veya ağırlık skorlarına göre seçilmesi yöntemidir. Bu sayede gereksiz veya düşük etkili değişkenler elenerek daha hızlı, genelleme kabiliyeti yüksek ve sade modeller elde edilir [4]. Bu çalışmada, modelin karmaşıklığını azaltmak ve yalnızca en bilgilendirici değişkenleri korumak amacıyla Top-K tabanlı özellik seçimi uygulanmıştır. İlk olarak, eğitim verisi üzerinde yalnızca bir "probe" LGBM modeli eğitilmiş ve modelin ürettiği feature_importances_değerleri kullanılarak her bir özniteligin sınıflandırma performansına katkısı nicel olarak hesaplanmıştır. Elde edilen önem puanları azalan sırayla sıralanmış ve en yüksek katkı sağlayan K = 220 öznitelik seçilmiştir. Bu süreçte, herhangi bir önem değeri bulunmaması durumuna karşı hata önleyici bir yapı eklenmiş ve tüm öznitelikler güvenli şekilde değerlendirilebilir hale getirilmiştir. Seçilen Top-K öznitelikler, X_train_final32 ve X_test_final32 matrislerinden indeksleme yoluyla çıkarılmış (Xtr_topK, Xte_topK) ve ardından SAE'den elde edilen latent temsillerle yataş olarak birleştirilmiştir (np.hstack). Böylece nihai öznitelik uzayı hem derin öğrenme temsillerini hem de en yüksek bilgi kazancı sağlayan özelliklerini içerecek biçimde oluşturulmuştur. Bu yaklaşım, gürültülü veya düşük kataklı öznitelikleri eleverek modelin hesaplama verimliliğini ve genelleme kabiliyetini artırmıştır.

3) LGBM Sınıflandırıcısı

Karar ağaçlarını ardışık biçimde eğiterek hızlı ve yüksek doğruluklu tahminler üreten bir makine öğrenmesi algoritmasıdır [5]. Veri seti büyük ve öznitelik sayısı fazla olduğunda hafiza verimliliği ve işlem hızı açısından oldukça etkilidir. Bu çalışmada Hibrit yapının son aşamasında, sınıflandırma işlemi LGBM algoritmasıyla gerçekleştirilmiştir. LGBM, karar ağaçları üzerine inşa edilmiş gradyan artırmalı (gradient boosting) bir yöntem olup, büyük ölçekli veriler üzerinde hem yüksek doğruluk hem de işlem verimliliği sağlamaktadır. Model, SAE'den elde edilen latent temsiller ile Top-K yöntemiyle seçilen özniteliklerin birleştirilmesiyle oluşturulan genişletilmiş öznitelik uzayında eğitilmiştir. Eğitim sürecinde, modelin karmaşıklığı ve genelleme kabiliyeti arasında denge sağlamak amacıyla optimum düzeyde ayarlanmış parametreler kullanılmıştır. Bu parametre ayarları, modelin aşırı öğrenmeden kaçınarak daha sağlam bir karar sınırı oluşturmasını sağlamıştır. Eğitim süreci sırasında erken durdurma (early stopping) mekanizması etkinleştirilmiş ve AUC ile binary_logloss metrikleri üzerinden değerlendirme yapılmıştır.

F. Önerilen Hibrit Mimari Üzerine Derin Analiz

1) Ablasyon Analizi

Tablo VI'de, önerilen hibrit mimarinin temel bileşenlerinden biri olan SAE entegrasyonunun, LGBM, LGBM+SAE ve LGBM+Top-K modelleri üzerindeki etkisini gösteren karşılaştırmalı performans sonuçları sunulmuştur. Bu tablo, söz konusu bileşenlerin modelin doğruluk düzeyi ve hesaplama süresi üzerindeki katkısını nicel olarak ortaya koyarak, hibrit yapının performans optimizasyonundaki rolünü açık biçimde göstermektedir.

TABLE VI: LGBM, LGBM+SAE ve LGBM+Top-K modellerinin karşılaştırmalı performans sonuçları.

| Model | Metrik | Sonuç (Test Ortalaması ± Std) |
|------------|--------------------|-------------------------------|
| LGBM | Doğruluk | 0.9788 ± 0.0004 |
| | Precision | 0.9831 ± 0.0002 |
| | Recall | 0.9744 ± 0.0008 |
| | Specificity | 0.9832 ± 0.0002 |
| | F1 | 0.9787 ± 0.0004 |
| | F1-W | 0.9788 ± 0.0004 |
| | AUC-ROC | 0.9977 ± 0.0001 |
| | Eğitim Süresi (sn) | ≈ 500 |
| LGBM+SAE | Doğruluk | 0.9678 ± 0.0004 |
| | Precision | 0.9717 ± 0.0008 |
| | Recall | 0.9637 ± 0.0009 |
| | Specificity | 0.9720 ± 0.0009 |
| | F1 | 0.9677 ± 0.0004 |
| | F1-W | 0.9678 ± 0.0004 |
| | AUC-ROC | 0.9950 ± 0.0001 |
| | Eğitim Süresi (sn) | ≈ 313 |
| LGBM+Top-K | Doğruluk | 0.9780 ± 0.0005 |
| | Precision | 0.9823 ± 0.0017 |
| | Recall | 0.9736 ± 0.0025 |
| | Specificity | 0.9825 ± 0.0017 |
| | F1 | 0.9779 ± 0.0005 |
| | F1-W | 0.9780 ± 0.0005 |
| | AUC-ROC | 0.9974 ± 0.0001 |
| | Eğitim Süresi (sn) | ≈ 210 |

Her üç model de yüksek doğruluk oranlarına ulaşmıştır; ancak SAE'nin eklenmesiyle bazı metriklerde hafif düşüşler gözlemlenmiştir. Bu durum, SAE'nin özellik temsili sırasında

sınırlı bilgi kaybına yol açabilmesi ve LGBM'in ham özniteliklerle zaten güçlü bir genelleme yeteneğine sahip olmasıyla açıklanabilir. Buna karşın, SAE'nin eklenmesi hesaplama süresini 510 saniyeden 313 saniyeye düşürerek yaklaşık %38 oranında hız kazanımı sağlamıştır. Bu sonuç, SAE'nin boyut indirgeme yoluyla hesaplama maliyetini azaltan, ancak doğrulukta yalnızca marginal bir kayıp yaratır verimli bir bileşen olduğunu göstermektedir.

LGBM, karar ağacı temelli yapısı sayesinde karmaşık ilişkileri doğrudan yakalayabildiğinden, SAE eklemesi her zaman ek bir doğruluk artışı sağlayamaz. Ancak SAE, özellikle büyük veya yüksek boyutlu veri setlerinde modelin genel verimliliğini artırır, aşırı öğrenmeyi azaltır ve hesaplama süresini önemli ölçüde kısaltır. Böylece doğrulukta küçük bir azalma karşılığında anlamlı bir işlem verimliliği ve metodolojik katkı sunmaktadır.

Ayrıca LGBM + Top-K modeli, temel LGBM ile neredeyse aynı doğruluk ve AUC-ROC değerlerini üretmiş; metrikler arasındaki farklar istatistiksel olarak önemsiz düzeyde kalmıştır (ör. 97.88 → 97.80 ACC, 99.77 → 99.74 AUC). Bununla birlikte, Top-K seçimi modelin eğitim süresini 510 saniyeden yaklaşık 210 saniyeye düşürerek önemli bir hız kazanımı sağlamıştır. Bu durum, Top-K'ın gereksiz öznitelikleri başarıyla eleverek modeli daha hafif ve optimize hale getirdiğini, buna rağmen genelleme gücünü koruduğunu göstermektedir. Her üç modelde de AUC-ROC değerlerinin %99.5'in üzerinde olması, sınıflar arası ayırm gücünün korunduğunu ortaya koymaktadır.

Elde edilen yüksek doğruluk, AUC-ROC ve tutarlı F1 değerleri, yalnızca mimari tasarımın değil, aynı zamanda hiperparametre optimizasyon sürecinin titizlikle yürütülmesinin bir sonucudur. Her üç modelde de kritik parametrelerin özenle ayarlanması, sistemin kararlı ve dengeli performans sergilemesini sağlamıştır. Sonuç olarak, modelin istikrarlı başarısını otomatik optimizasyondan ziyade bilinçli hiperparametre tasarımlıyla elde edilmiş; böylece deneysel başarısının ardından tasarım odaklı, mühendislik temelli bir katkı olduğu ortaya konmuştur.

2) Öznitelik Seçiminin Nihai Hibrit Model Üzerindeki Etkisi

Tablo VII, Top-K özellik seçiminin hibrit model performansı üzerindeki etkisini açık bir biçimde göstermektedir. Top-K uygulandığında OFF veri temsili üzerinde doğruluk %98.30, precision %98.53, recall %98.07, specificity %98.54, F1-score %98.30, F1-weighted %98.30 ve AUC-ROC %99.83 elde edilmiştir. Bu sonuçlar, modelin hem yüksek başarıya hem de karmaşık veri dağılımlarını etkin bir şekilde temsil edebilen güçlü öğrenme kapasitesine sahip olduğunu kanıtlamaktadır. Özellikle Top-K bileşeni, SAE tarafından oluşturulan derin temsillerin yalnızca en bilgilendirici kısımlarını seçerek modelin gürültüye karşı dayanıklılığını artırmış ve genelleme kabiliyetini güçlendirmiştir. Böylece hibrit model, geniş ölçekli ağ trafiği verilerinde hem istatistiksel hem yapısal örüntüler etkin biçimde yakalayabilmştir.

TABLE VII: Hibrit Modelin 5-Kat Çapraz Doğrulama Sonuçları (OFF)

| Metrik | Sonuç (Test Ortalaması ± Std) |
|-------------|-------------------------------|
| Accuracy | 0.983 ± 0.0004 |
| Precision | 0.9853 ± 0.0006 |
| Recall | 0.9807 ± 0.0008 |
| Specificity | 0.9854 ± 0.0006 |
| F1-Score | 0.983 ± 0.0004 |
| F1-Weight | 0.983 ± 0.0004 |
| AUC-ROC | 0.9983 ± 0.0001 |

3) Hibrit Mimari İçin Kararlılık ve Güvenilir Sınıflandırma Analizi

K-Fold sonuçlarında elde edilen düşük standart sapma değerleri, modellerin farklı eğitim bölmelerinde tutarlı performans sergilediğini göstermektedir. Özellikle hibrit modelde gözlenen son derece düşük varyans, modelin tüm katmanlarda benzer doğruluk seviyelerine ulaştığını ve veri bölünmesinden kaynaklanan değişkenliğe karşı yüksek dayanıklılık gösterdiğini kanıtlamaktadır. Bu bulgu, hibrit mimarinin yalnızca yüksek başarı sağladığını, aynı zamanda kararlı ve güvenilir bir sınıflandırma performansı sunduğunu da güçlü biçimde desteklemektedir.

Hibrit model, literatürde çoğu çalışmada yer verilmeyen ancak saldırı tespit sistemleri açısından kritik öneme sahip FPR (False Positive Rate), FNR (False Negative Rate), NPV (Negative Predictive Value) ve FDR (False Discovery Rate) gibi hata oranlarında da güçlü performans sergilemektedir (bkz. Tablo VIII). Hibrit modelin FPR ve FNR değerlerinin sırasıyla %1.46 ve %1.93 seviyesinde olması, saldırısı sınıfı örneklerinde hem yanlış alarm oranını hem de kaçırılan saldırısı olasılığını oldukça düşük tuttuğunu göstermektedir. Bu bulgular, hibrit mimarinin ağ trafiğinde saldırısı ve normal örnekleri yüksek doğrulukla ayırt edebildiğini; hata oranlarının istikrarlı biçimde minimize edilerek güvenilir bir sınıflandırma performansı sergilediğini ortaya koymaktadır.

TABLE VIII: Hibrit Modelin Ek Hata Oranları ve Doğruluk Metrikleri (OFF-Ham Veri)

| Metrik | Açıklama | Değer (%) |
|--------|-----------------------|-----------|
| FPR | Yanlış Pozitif Oranı | 1.465 |
| FNR | Yanlış Negatif Oranı | 1.935 |
| NPV | Negatif Öngörü Değeri | 98.07 |
| FDR | Yanlış Keşif Oranı | 1.466 |

4) Hibrit Mimarının Model Boyutu ve Hesaplama Performansı Metrikleri

Çalışmada, karmaşık ve yüksek boyutlu bir veri kümesi üzerinde geliştirilen hibrit mimarinin değerlendirilmesinde yalnızca doğruluk metrikleri değil, aynı zamanda gerçek zamanlı sistem performansını doğrudan etkileyen ölçütler de dikkate alınmıştır. Tablo IX'da hibrit modelin model boyutu (model size), çıkışım süresi (inference time), gecikme süresi (latency), işleme hızı (throughput) ve eğitim süresine ilişkin hesaplama performansı gösterilmektedir. Tüm değerler float32 formatında ölçülmüş ve hesaplamalar CPU üzerinde gerçekleştirilmiştir.

TABLE IX: Hibrit Modelin Ham Veri Üzerinde Hesaplama Performans Metrikleri

| Metrik | Değer | Açıklama |
|----------------|---------------------------------------|------------------------|
| Model Boyutu | 28.67 MB (\approx 1.48M parametre) | Toplam model boyutu |
| Inference Time | 1.236 s (n = 80 079) | Toplam çıkışım süresi |
| Latency | 0.015 ms/ornek | Ortalama gecikme |
| Throughput | 64 780 örnek/s | İşleme hızı |
| Eğitim Süresi | 493.039 s | Ortalama eğitim süresi |

Tablodan gösterildiği üzere; hibrit modelin ortalama gecikmesi (latency) 0.015 ms/ornek, inference time süresi 1.236 s (n = 80 079) ve işleme hızı (throughput) 64 780 örnek/s olarak hesaplanmıştır. Model boyutu 28.67 MB olup, bu değer yaklaşık 1.48 milyon eğitilebilir parametreye karşılık gelmektedir. Bu sonuç, önerilen hibrit yapının hem hafif hem de dağıtıma uygun bir mimarı sunduğunu göstermektedir.

Elde edilen işlem performansı, hibrit mimarinin gerçek zamanlı saldırısı tespiti için oldukça verimli çalıştığını ortaya koymakta; düşük gecikme, yüksek throughput ve kompakt model

boyutu sayesinde gerçek zamanlı uygulamalarda gecikmesiz ve ölçeklenebilir analiz imkânı sağlamaktadır. Eğitim süresi her fold için ortalama 493.039 saniye olmasına rağmen, mimarinin karmaşılığına rağmen dengeli ve yüksek performans üretmesi bu süreyle hesaplama açısından kabul edilebilir bir maliyet hâline getirmektedir. Dolayısıyla, önerilen hibrit model yalnızca doğruluk bakımından değil, hesaplama verimliliği, pratik uygulanabilirliği ve gerçek zamanlı kullanım potansiyeli ile literatürdeki benzer yaklaşımardan belirgin şekilde ayırmaktadır.

5) Hibrit Modelin OFF Temsilinde Sınıflandırma Performansı (Confusion Matrix)

Şekil 4'de sunulan hibrit model confusion matrix sonucu, OFF veri temsili üzerinde oldukça dengeli ve yüksek bir sınıflandırma performansı sergilemiştir. Confusion matrix sonucuna göre model, 197.278 benign örneği doğru şekilde Benign, 196.326 malicious örneği doğru şekilde Malicious olarak sınıflandırmıştır. Buna karşın, yalnızca 2.920 benign örnek hatalı şekilde saldırı (False Positive) olarak işaretlenmiş; 3.873 malicious örnek ise benign (False Negative) şeklinde sınıflandırılmıştır. Bu dağılım, modelin hem yanlış alarm oranını hem de saldırı atlama oranını oldukça düşük seviyede tuttuğunu göstermektedir. Özellikle True Positive ve True Negative hücrelerinde yer alan yüksek değerler, hibrit mimarinin hem benign hem de malicious trafiği ayırt etmede güçlü bir genellemeye yeteneğine sahip olduğunu ortaya koymaktadır. Sonuç olarak, OFF temsili üzerinde hibrit modelin ürettiği confusion matrix, modelin pratik IDS senaryolarında güvenilir bir ayrımlı yapısını, yanlış pozitif/negatif oranlarını minimumda tuttuğunu ve yüksek doğrulukla çalıştığını göstermektedir.

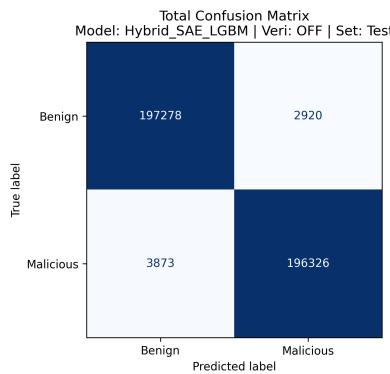


Fig. 4: Hibrit Modelin confusion Matrix Sonucu

6) Hibrit Model İçin İstatistiksel Kararlılık Analizi (Bootstrap %95 CI)

Bu bölümde, elde edilen nihai Hibrit modelin istatistiksel kararlılığı bootstrap yöntemi kullanılarak değerlendirilmiştir. 5-Fold çapraz doğrulama protokolü kapsamında, her fold'un test kümelerinde uygulanan 1000 tekrarlı bootstrap örnekleme sonucunda her performans metriğine ait %95 güven aralıkları hesaplanmıştır; tüm fold'ların birleşik sonuçları Tablo X'da raporlanmıştır. Bootstrap analizine göre tüm metriklerin güven aralıkları oldukça dardır (genellikle ± 0.001 – 0.0015 aralığında), bu da modelin sonuçlarının rastgele örnekleme varyansından etkilenmediğini ve yüksek bir istikrarla performans gösterdiğini doğrulamaktadır. Özellikle AUC-ROC değeri (0.9983; CI: [0.9979, 0.9986]), modelin ayrımlı gücünün neredeyse mükemmel düzeyde olduğunu göstermektedir. Genel olarak, önerilen SAE+LGBM hibrit mimarisi doğruluk, hassasiyet ve genel güvenilirlik açısından oldukça tutarlı ve kararlı bir performans sergilemiştir.

TABLE X: Hibrit Modelin Temporal Test Performansı ve %95 Güven Aralıkları (Bootstrap, $B = 1000$).

| Metric | Mean | CI_low | CI_high |
|-------------|--------|--------|---------|
| Accuracy | 0.9830 | 0.9817 | 0.9842 |
| Precision | 0.9853 | 0.9840 | 0.9866 |
| Recall | 0.9807 | 0.9789 | 0.9818 |
| Specificity | 0.9854 | 0.9842 | 0.9869 |
| F1 Score | 0.9830 | 0.9815 | 0.9841 |
| F1 Weighted | 0.9830 | 0.9817 | 0.9843 |
| AUC-ROC | 0.9983 | 0.9979 | 0.9986 |

G. Önerilen Hibrit Modelin Literatür Hibrit Model ile Karşılaştırılması

Bu çalışmada önerilen hibrit model, SAE tabanlı derin temsil öğrenimi, Top-K öznitelik seçimi ve LightGBM sınıflandırıcısının ardışık kullanımından oluşmaktadır. SAE, veriyi daha anlamlı bir latent uzaya indirgerken; Top-K yöntemi, en yüksek bilgi kazancına sahip öznitelikleri seçerek modeli sadeleştirir ve verimliliği artırır. Son aşamada LightGBM, bu derin + seçilmiş temsilleri kullanarak yüksek doğruluk, düşük gecikme ve güçlü genellemeye performansı sağlamaktadır. Tablo XI'de, literatürde yer alan BVR-SFO-AEDL hibrit modeli ile bu çalışmada önerilen SAE + Top-K + LGBM tabanlı hibrit modelin kötü amaçlı yazılım sınıflandırmasındaki temel metrikler açısından karşılaştırması sunulmuştur. Literatürdeki BVR-SFO-AEDL modeli, Autoencoder-1D-CNN-BiLSTM bileşenlerinden oluşan derin hibrit bir mimari olup, Windows tabanlı kötü amaçlı yazılım tespitinde %96.47 doğruluk seviyesinde yüksek bir başarı elde ettiği rapor edilmiştir [6]. Buna karşılık, bu çalışmada önerilen mimarı; SAE katmanının temsil öğrenme kapasitesini, Top-K öznitelik seçiminin optimizasyon etkisini ve LightGBM'in güçlü karar mekanizmasını bir araya getirerek daha dengeli, genellemeye kabiliyeti yüksek ve hesaplama açısından verimli bir yapı sunmaktadır. Elde edilen sonuçlar, yalnızca yüksek doğruluk değeri (%98.30) ile değil; saldırı tespit duyarlığını doğrudan temsil eden recall metriğinde elde edilen %98.07'lik performansla da literatürdeki hibrit modele kıyasla belirgin bir üstünlük ortaya koyduğunu göstermektedir. Dolayısıyla önerilen hibrit model, hem genel sınıflandırma başarısı hem de saldırı tespit hassasiyeti açısından daha istikrarlı, güvenilir ve pratik olarak uygulanabilir bir çözüm sunmaktadır.

TABLE XI: Temel Metriklerde Literatür Hibrit Model ile Karşılaştırma

| Model | Accuracy | Recall | Specificity | AUC-ROC |
|----------------------|----------|--------|-------------|---------|
| Önerilen Hibrit | 0.9830 | 0.9807 | 0.9854 | 0.9983 |
| Literatür Hibrit [6] | 0.9647 | 0.9648 | 0.9632 | — |

Ayrıca, literatürdeki BVR-SFO-AEDL hibrit modelinde AUC-ROC metriği açık biçimde raporlanmamıştır. Oysa malware sınıflandırma problemlerinde AUC-ROC, modelin genel ayrı edicilik gücünü ve sınıflar arasındaki ayrımlı performansını ölçmek açısından kritik öneme sahiptir. Bu eksiklik, ilgili literatür çalışmasının sınıflandırma performansını tam olarak karşılaştırılabilir hale getirmemektedir. Bu çalışmada ise AUC-ROC metriği açık biçimde hesaplanmış ve modelin AUC-ROC değeri %99.83 olarak elde edilmiştir. Dolayısıyla, önerilen hibrit model; performans metriklerinin şeffaf, eksiksiz ve karşılaştırılabilir biçimde raporlanması bakımından literatürdeki mevcut hibrit yaklaşımardan ayırmaktadır.

1) Hibrit Modelin Sınıf Bazlı ROC Eğrileri ve AUC Değerlerinin Analizi

Şekil 5'de, önerilen hibrit modelin toplam çapraz doğrulama (5-fold CV) sonuçlarına göre elde edilen ROC eğrileri sunulmaktadır. Tablo XII'de gösterildiği üzere, model hem Benign (0) hem de Malicious (1) sınıflarında 0.9983 AUC değeri elde etmiş, ortalama AUC skoru ise 0.99827 (± 0.00007) olarak hesaplanmıştır. Bu sonuçlar, önerilen hibrit yapının sınıflar arasında son derece yüksek ayrımla sahip olduğunu, yanlış sınıflandırmaları neredeyse tamamen ortadan kaldırdığını ve genelleme başarısında istisnai bir kararlılık sergilediğini göstermektedir. Dolayısıyla, hibrit model yalnızca yüksek doğrulukta değil, aynı zamanda istatistiksel olarak mükemmel yakını ayırt edicilik performansıyla gerçek zamanlı saldıri teşpiti için üst düzey güvenilirlik sunmaktadır.

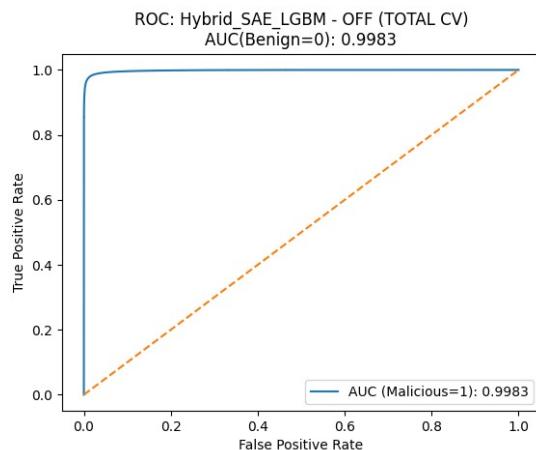


Fig. 5: Önerilen hibrit modelin OFF veri temsili üzerinde 5-katlı çapraz doğrulama ile elde edilen toplam ROC eğrisi

TABLE XII: Hibrit Modelin Sınıf Bazlı AUC Performans Sonuçları

| Sınıf | AUC Değeri | Ortalama AUC (\pm Std) |
|---------------|------------|-----------------------------|
| Benign (0) | 0.998271 | 0.998272 (± 0.000065) |
| Malicious (1) | 0.998271 | |

H. Karşılaştırmalı Sınıflandırma Sonuçları ve İstatistiksel Analiz

1) ML Modelleri: Karşılaştırmalı Performans Analizi

Tablo XIII'te sunulan karşılaştırmalı sonuçlar, farklı makine öğrenmesi algoritmalarının test kümesi üzerindeki genel sınıflandırma performanslarını ortaya koymaktadır. Elde edilen bulgular, ağaç tabanlı yöntemlerin (CatBoost, ExtraTrees, GradientBoost, HistGB) doğrusal tabanlı SGDClassifier modeline kıyasla belirgin biçimde daha yüksek performans sergilediğini göstermektedir.

CatBoost modeli, %97.24 doğruluk ve %99.61 AUC-ROC degeriyle dikkat çekmiş; dengeli precision-recall değerleri (%97.7-%96.7) sayesinde istikrarlı bir sınıflandırma davranışını sergilemiştir. ExtraTrees ise tüm modeller arasında en yüksek doğruluk (%97.54) ve AUC-ROC (%99.72) değerlerini elde ederek istatistiksel olarak en başarılı model olmuştur. Bu sonuç, çok sayıda karar ağacının rastgele öznitelik alt kümeleri üzerinde eğitilmesiyle modelin genelleme kabiliyetinin güçlendiğini göstermektedir.

GradientBoost modeli, %95.27 doğruluk ve %99.10 AUC-ROC değerleriyle göreceli olarak daha düşük performans göstermiştir;

bu farkın, özellikle öğrenme oranı ve veri dengesizliği gibi hiperparametre etkilerinden kaynaklandığı değerlendirilmektedir. Benzer şekilde, HistGB algoritması %96.64 doğruluk ve %99.49 AUC-ROC ile düşük varyanslı, kararlı sonuçlar üretmiştir.

Veri kümesinin lineer ayrılabilirliğini sınamak amacıyla ayrıca SGDClassifier modeli de test edilmiştir. Beklendiği üzere, doğrusal karar sınırlarına dayalı yapısı nedeniyle düşük performans (%60.47 doğruluk, %60.47 AUC-ROC) sergilemiştir. Bu bulgu, EMBER veri kümesinin karmaşık ve non-lineer ilişkiler içerdigini empirik olarak doğrulamaktadır. Buna karşın, doğrusal olmayan örüntülerin yakalayabilen yöntemler — örneğin KNN veya ağaç tabanlı modeller — çok daha yüksek doğruluk ve genelleme başarısı göstermiştir. Ancak KNN modeli, yüksek hesaplama maliyeti ve bellek gereksinimleri nedeniyle büyük ölçekli senaryolarda pratik sınırlamalara sahiptir.

TABLE XIII: Ham Veri Üzerinde Makine Öğrenmesi Modellerinin Test Performans Sonuçları (Ortalama \pm Std)

| Model | Metrik | Sonuç (Test Ortalama \pm Std) |
|---------------|-------------|---------------------------------|
| CatBoost | Doğruluk | 0.9724 \pm 0.0004 |
| | Precision | 0.9772 \pm 0.0009 |
| | Recall | 0.9674 \pm 0.0011 |
| | Specificity | 0.9774 \pm 0.0009 |
| | F1 | 0.9723 \pm 0.0004 |
| | F1-W | 0.9724 \pm 0.0004 |
| | AUC-ROC | 0.9961 \pm 0.0001 |
| ExtraTrees | Doğruluk | 0.9754 \pm 0.0007 |
| | Precision | 0.9826 \pm 0.0007 |
| | Recall | 0.9681 \pm 0.001 |
| | Specificity | 0.9828 \pm 0.0006 |
| | F1 | 0.9753 \pm 0.0007 |
| | F1-W | 0.9754 \pm 0.0007 |
| | AUC-ROC | 0.9972 \pm 0.0001 |
| GradientBoost | Doğruluk | 0.9527 \pm 0.0007 |
| | Precision | 0.9561 \pm 0.0009 |
| | Recall | 0.9490 \pm 0.0014 |
| | Specificity | 0.9564 \pm 0.0009 |
| | F1 | 0.9525 \pm 0.0007 |
| | F1-W | 0.9527 \pm 0.0007 |
| | AUC-ROC | 0.9910 \pm 0.0001 |
| HistGB | Doğruluk | 0.9664 \pm 0.0008 |
| | Precision | 0.9717 \pm 0.0005 |
| | Recall | 0.9607 \pm 0.0016 |
| | Specificity | 0.9721 \pm 0.0007 |
| | F1 | 0.9662 \pm 0.0008 |
| | F1-W | 0.9664 \pm 0.0008 |
| | AUC-ROC | 0.9949 \pm 0.0002 |
| SGDClassifier | Doğruluk | 0.6047 \pm 0.0176 |
| | Precision | 0.5927 \pm 0.0136 |
| | Recall | 0.6682 \pm 0.0506 |
| | Specificity | 0.5411 \pm 0.0317 |
| | F1 | 0.6276 \pm 0.0266 |
| | F1-W | 0.6026 \pm 0.0165 |
| | AUC-ROC | 0.6047 \pm 0.0176 |

Sonuç olarak, elde edilen bulgular ağaç tabanlı ve derin öğrenme temelli modellerin tercih edilme gereklisini güçlü biçimde desteklemektedir. Özellikle CatBoost, ExtraTrees, HistGB ve GradientBoost gibi yöntemlerin non-lineer öznitelik etkileşimlerini modelleyebilme kabiliyeti ve yüksek AUC-ROC skorları, bu yaklaşımın karmaşık siber güvenlik verilerinde güçlü genelleme sağladığını göstermektedir.

Bununla birlikte, önerilen hibrit mimari, hem derin öğrenmenin temsil gücünü hem de LGBM'in karar ağacı temelli ayrımla kapsatesini bütünlüğe getirerek, aynı veri üzerinde doğruluk ve AUC ölçütlerinde tüm ML tabanlı yaklaşımları aşmıştır. Bu durum, hibrit yapının yalnızca öznitelikler arasındaki doğrusal olmayan ilişkileri yakalamakla kalmayıp, aynı zamanda SAE'nin derin temsilleriyle

bilgi yoğunluğunu artırması ve LGBM'in optimizasyon gücüyle bu temsilleri etkili biçimde ayırtılması sayesinde gerçekleşmiştir.

Eğitim maliyeti açısından bakıldığından, ağaç tabanlı ML modellerinin yaklaşık 290–580 s aralığında eğitildiği gözlenmiştir; 400,397 örnekten oluşan veri kümesi üzerinde, tamamen CPU (GPU'suz) koşullarda bu süreler beklenen ve normal düzeydedir. Ayrıca bu modellerin —ve hibrit yapıda dağıtımında aktif kalan LGBM modülünün— çıkışım gecikmesi düştür. Dolayısıyla hibrit mimari, yalnızca en yüksek doğruluk ve AUC performansını değil, aynı zamanda maliyet/performans dengesi açısından da gerçek zamanlı kullanım senaryoları için en uygun yapıyı sunmaktadır.

2) DL Modelleri: Karşılaştırmalı Performans Analizi

OFF veri tipi üzerinde gerçekleştirilen deneylerde, DNN, MLP-Mixer, ResNet-MLP, Wide&Deep ve gMLP gibi farklı derin öğrenme mimarileri test edilmiş ve sonuçlar Tablo XIV'de sunulmuştur. EMBER24 veri kümesi, yüksek boyutlu, uzun kuyruklu, dağınık ve non-lineer bir özellik dağılımına sahip olduğu için bu tür modellerin genellemeye kapasitesini sınayan zorlu bir yapıya sahiptir. Buna rağmen, oluşturulan kod altyapısı ve dikkatle belirlenen hiperparametre ayarları sayesinde, literatürde genellikle daha düşük doğruluk ve AUC değerleri rapor edilen bu tür tabular veri problemlerinde oldukça yüksek sonuçlar elde edilmiştir.

TABLE XIV: Ham Veri Üzerinde Derin Öğrenme Modellerinin Test Performans Sonuçları (Ortalama \pm Std)

| Model | Metrik | Sonuç (Test Ortalama \pm Std) |
|-----------|-------------|---------------------------------|
| Wide&Deep | Doğruluk | 0.9706 \pm 0.0013 |
| | Precision | 0.9726 \pm 0.0028 |
| | Recall | 0.9684 \pm 0.0021 |
| | Specificity | 0.9727 \pm 0.0029 |
| | F1 | 0.9705 \pm 0.0012 |
| | F1-W | 0.9706 \pm 0.0013 |
| | AUC-ROC | 0.9944 \pm 0.0003 |
| DNN | Doğruluk | 0.9707 \pm 0.0011 |
| | Precision | 0.9742 \pm 0.0012 |
| | Recall | 0.9669 \pm 0.0024 |
| | Specificity | 0.9744 \pm 0.0013 |
| | F1 | 0.9706 \pm 0.0012 |
| | F1-W | 0.9707 \pm 0.0011 |
| | AUC-ROC | 0.9945 \pm 0.0003 |
| MLP-Mixer | Doğruluk | 0.9700 \pm 0.0007 |
| | Precision | 0.9740 \pm 0.0022 |
| | Recall | 0.9657 \pm 0.0033 |
| | Specificity | 0.9742 \pm 0.0023 |
| | F1 | 0.9698 \pm 0.0008 |
| | F1-W | 0.9700 \pm 0.0007 |
| | AUC-ROC | 0.9949 \pm 0.0003 |
| ResNetMLP | Doğruluk | 0.9708 \pm 0.0006 |
| | Precision | 0.9755 \pm 0.0023 |
| | Recall | 0.9658 \pm 0.0017 |
| | Specificity | 0.9758 \pm 0.0024 |
| | F1 | 0.9707 \pm 0.0005 |
| | F1-W | 0.9708 \pm 0.0006 |
| | AUC-ROC | 0.9949 \pm 0.0002 |
| gMLP | Doğruluk | 0.9719 \pm 0.0006 |
| | Precision | 0.9756 \pm 0.0033 |
| | Recall | 0.9681 \pm 0.0039 |
| | Specificity | 0.9757 \pm 0.0035 |
| | F1 | 0.9718 \pm 0.0007 |
| | F1-W | 0.9719 \pm 0.0006 |
| | AUC-ROC | 0.9950 \pm 0.0003 |

DL modellerinin test doğrulukları yaklaşık 97.0–97.2%, AUC değerleri ise 0.994–0.995 (± 0.0003) bandında gerçekleşmiştir. Bu seviyedeki başarı, yalnızca model mimarisinden değil, özellikle veri ön işleme, ölçekleme, regularization ve erken durdurma stratejilerinin optimize edilmesiyle mümkün olmuştur. Başarı, modelin

türünden ziye yapının bütünsel olarak tasarlanmasıından kaynaklanmaktadır.

Derin öğrenme tabanlı modeller eğitim süresi açısından ilk bakışta daha hızlı görünse de, CPU ortamında 200–435 s aralığında tamamlanmıştır; özellikle MLP-Mixer gibi çok katmanlı yapılar bu aralığın üst sınırına ulaşmıştır. Ayrıca bu modellerin inference süresi, matris çarpımına dayalı derin yapı nedeniyle hibrit modele göre belirgin biçimde yüksektir. Buna karşın, önerilen hibrit mimari yaklaşık 493 s'lik eğitim süresine rağmen, inference aşamasında yalnızca LGBM modülünü kullandığından çok daha düşük latency üretmektedir. Bu durum, özellikle gerçek zamanlı IDS veya malware tespitinde önemli bir avantajdır; zira eğitim yalnızca bir kez yapılmakta, inference ise sürekli olarak çalışmaktadır. Böylece hibrit model, hem eğitim hem test performansında üstünlik (AUC = 0.9983 ± 0.0001) sağlarken, gerçek zamanlı uygulamalara uygun bir çalışma verimliliği de sunmuştur.

Ayrıca bu sonuçlar, önerilen hibrit mimari ile kıyaslandığında anlamlı bir fark ortaya koymaktadır. Hibrit model, aynı veri üzerinde %98.3 doğruluk, %99.83 AUC ve %98.07 recall değeriyle DL tabanlı yaklaşımlarından daha yüksek performans sergilemiştir. DL modellerinde recall değerleri %96.57–%96.81 aralığında kalmıştır. Bu fark, hibrit yapının hem SAE'nin derin temsillerinden hem de LGBM'in karar ağacı tabanlı öğrenme kapasitesinden yararlanmasından ileri gelmektedir. Sonuç olarak, DL tabanlı modeller iyi optimize edildiğinde tabular verilerde güçlü sonuçlar üretse de, hibrit mimari bu başarıyı yaklaşık +%1.2 doğruluk, +0.003 AUC ve +%1.2–%1.5 puan recall farkıyla gerçek genel performans açısından üstünlik sağlamıştır.

Şekil 6'da, ML, DL ve hibrit modellere ait malware sınıflandırmasındaki temel metrik sonuçları karşılaştırılmış olarak sunulmuştur. Bu görsel, modellerin doğruluk, duyarlılık ve AUC-ROC performanslarını bir arada göstererek, her yaklaşımın genel sınıflandırma başarısını açık biçimde ortaya koymaktadır. Genel olarak değerlendirildiğinde, önerilen hibrit model, hem klasik makine öğrenmesi hem de derin öğrenme tabanlı yaklaşımara kıyasla daha yüksek doğruluk, istikrar ve genellemeye başarısı sergilemiştir. Makine öğrenmesi modelleri (örneğin CatBoost, ExtraTrees) yüksek doğruluk ve AUC değerleri üretmiş olsa da, veri setindeki karmaşık öznitelik etkileşimlerini derin düzeye temsil etme kapasitesi hibrit mimariye göre daha sınırlı kalmıştır. Derin öğrenme modelleri ise yüksek temsil gücüne rağmen, hesaplama maliyeti nedeniyle hibrit yapının altında performans göstermiştir. Böylece model, karmaşık siber güvenlik verilerinde yüksek doğruluk, düşük hata oranı ve pratik uygulanabilirliğiyle literatürdeki mevcut hibrit yaklaşımından anlamlı biçimde ayırtmıştır.

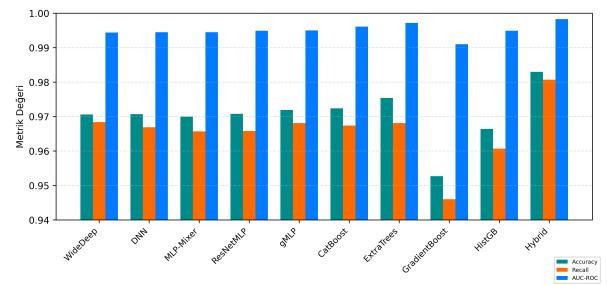


Fig. 6: DL, ML ve Hibrit modellerin Accuracy, Recall ve AUC-ROC metriklerinin karşılaştırılmalı sunumu.

3) Ek Analiz: Ölçekleyicinin Derin Öğrenme Modelleri Üzerindeki Etkisi

Bu çalışmada ölçekleyici seçiminin DL performansına etkisi incelenmiş ve sonuçlar Tablo XV'de sunulmuştur. StandardScaler ile elde edilen sonuçlar, RobustScaler kullanıldığında anlamlı biçimde düşmüştür. Düşüş özellikle DNN, ResNetMLP ve WideDeep modellerinde belirgindir; gMLP ve MLP-Mixer'da ise daha sınırlı kalmıştır. Bunun olası nedeni, veri temizliği sonrasında aykırı değer etkisinin zaten azaltılmış olması ve StandardScaler'ın ortalaması-standart sapma temelli normalize etmesinin, DL eğitiminde gradyan ölçeklerini ve Batch/LayerNorm ile etkileşimi daha iyi hizalamasıdır. Buna karşılık RobustScaler (medyan-IQR) bazı öznitelikleri fazla sıkıştırarak sinyal-gürültü oranını düşürebilir ve aktivasyonların doygunluğa girmesine yol açabilir.

Fold std'ları düşük kaldığından, gözlenen farklar rastgelelikten değil, ölçekleyici seçime duyarlılıktan kaynaklanmaktadır.

Öneri: Bu veri/kurulum için DL tarafında varsayılan olarak StandardScaler'ı koruyun. Aykırı patlamaları olan durumlarda StandardScaler öncesi winsorization/clipping veya gerekirse Quantile/Yeo-Johnson gibi dönüştürmeler ekleyin; RobustScaler kullanılıcaksa öğrenme oranı/weight decay gibi hiperparametreleri yeniden ayarlayın.

TABLE XV: Ham Veri Üzerinde Derin Öğrenme Modellerinin Test Performans Sonuçları (Ortalama \pm Std) — RobustScaler

| Model | Metrik | Sonuç (Test Ortalama \pm Std) |
|-----------|-------------|---------------------------------|
| DNN | Doğruluk | 0.6455 \pm 0.0024 |
| | Precision | 0.5865 \pm 0.0018 |
| | Recall | 0.9873 \pm 0.0016 |
| | Specificity | 0.3038 \pm 0.0058 |
| | F1 | 0.7358 \pm 0.0012 |
| | F1-W | 0.5987 \pm 0.0038 |
| | AUC-ROC | 0.7407 \pm 0.0112 |
| MLP-Mixer | Doğruluk | 0.9159 \pm 0.0013 |
| | Precision | 0.9368 \pm 0.0082 |
| | Recall | 0.8920 \pm 0.0092 |
| | Specificity | 0.9397 \pm 0.0091 |
| | F1 | 0.9138 \pm 0.0016 |
| | F1-W | 0.9158 \pm 0.0013 |
| | AUC-ROC | 0.9726 \pm 0.0007 |
| ResNetMLP | Doğruluk | 0.6511 \pm 0.0046 |
| | Precision | 0.7427 \pm 0.2107 |
| | Recall | 0.7176 \pm 0.0364 |
| | Specificity | 0.5846 \pm 0.3714 |
| | F1 | 0.6342 \pm 0.1403 |
| | F1-W | 0.6068 \pm 0.0062 |
| | AUC-ROC | 0.7483 \pm 0.0103 |
| Wide&Deep | Doğruluk | 0.6746 \pm 0.0832 |
| | Precision | 0.6601 \pm 0.0918 |
| | Recall | 0.7463 \pm 0.0717 |
| | Specificity | 0.6030 \pm 0.1491 |
| | F1 | 0.6980 \pm 0.0688 |
| | F1-W | 0.6704 \pm 0.0873 |
| | AUC-ROC | 0.7354 \pm 0.0731 |
| gMLP | Doğruluk | 0.9216 \pm 0.0014 |
| | Precision | 0.9355 \pm 0.0055 |
| | Recall | 0.9057 \pm 0.0069 |
| | Specificity | 0.9375 \pm 0.0061 |
| | F1 | 0.9203 \pm 0.0016 |
| | F1-W | 0.9215 \pm 0.0014 |
| | AUC-ROC | 0.9750 \pm 0.0007 |

4) Overfitting ve Underfitting Durumlarının Değerlendirilmesi

Model geliştirme sürecinde overfitting ve underfitting olasılıkları dikkatle değerlendirilmiş, ancak hiçbir modelde bu tür durumlar gözlemlenmemiştir. Eğitim ve test seti sonuçlarının birbirine oldukça yakın seyretmesi, modellerin yüksek genellemeye kabiliyetine sahip olduğunu açık biçimde ortaya koymuştur. Özellikle

underfitting riskine karşı, model karmaşaklılığı ve öğrenme parametreleri dengeli biçimde ayarlanmış; bu sayede modellerin verideki örüntüler yeterli düzeyde öğrenmesi sağlanmıştır. Ayrıca, 5-kat çapraz doğrulama stratejisinin kullanılması, her bir modelin farklı veri bölgümlerinde tutarlı performans göstermesini sağlamış ve genelleme başarısının istatistiksel olarak doğrulanmasına olanak tanımıştır. Böylece, tüm modellerin hem eğitim verisine aşırı uyum göstermediği hem de yetersiz öğrenme sergilemeden gerçek dünya senaryolarında güvenilir biçimde genellenebildiği tescillemiştir.

5) En İyi Modelin İstatistiksel Anlamlılık ve Karşılaştırması

Bu çalışmada, Tablo XVI'de sonuçları verilen en iyi model olan Hibrit ile bir sonraki en iyi model olan ExtraTrees arasındaki performans farklarının istatistiksel olarak anlamlı olup olmadığı tablo XVII'de değerlendirilmiştir. OFF temsili üzerinde yapılan karşılaştırında, Hibrit modelin elde ettiği tüm metrik farklarının istatistiksel olarak anlamlı olduğu görülmektedir. Standart sapmaların ± 0.0004 – 0.0008 aralığında kalması, her iki modelin de son derece kararlı ve tekrarlanabilir sonuçlar ürettiğini; beş temel metrikte $p < 0.001$ düzeyindeki değerler ise farkların rastlantusal olasılıkla açıklanamayacağını göstermektedir. Yapılan istatistiksel anlamlılık testlerinde, Accuracy, Recall, AUC-ROC ve F1-Score metriklerinde farkların $p < 0.001$ düzeyinde; Precision metriğinde ise $p < 0.01$ düzeyinde anlamlı olduğu belirlenmiştir.

Özellikle Recall metriği (Hibrit: 0.9807 ± 0.0008 ; ExtraTrees: 0.9681 ± 0.0010 ; $\Delta = 0.0123$, $p \approx 0.0004$) en belirgin farkı ortaya koymaktadır. Bu bulgu, modelin saldırısı örneklerini yakalama kapasitesinde ve yanlış negatifleri azaltmadı anlamlı bir iyileşmeye işaret etmektedir. Siber güvenlik alanında yanlış negatiflerin kritik önemi göz önüne alındığında, Recall'daki bu üstünlük, Hibrit modelin pratik kullanım açısından da tercih edilir olduğunu desteklemektedir.

Sonuç olarak, Hibrit modelin yüksek doğruluğu (Accuracy = 0.983), güçlü ayırm gücü (AUC-ROC = 0.9983) ve düşük varyansı ile gözlenen performans farkı rassal değildir; tamamen istatistiksel olarak anlamlıdır ($p < 0.001$) ve yüksek tekrarlanabilirlik düzeyinde elde edilmiştir. Bu bulgular, EMBER-IDS saldırısı analizlerinde güvenilirlik ve tespit tutarlılığı açısından Hibrit modelin referans niteliğinde olduğunu göstermektedir.

TABLE XVI: Hibrit ve ExtraTrees Modellerinin Performans Karşılaştırması (OFF Veri Temsili, Ortalama \pm Std)

| Metrik | Hibrit | ExtraTrees |
|-----------|---------------------|---------------------|
| Accuracy | 0.9830 ± 0.0004 | 0.9754 ± 0.0007 |
| Precision | 0.9853 ± 0.0006 | 0.9826 ± 0.0007 |
| Recall | 0.9807 ± 0.0008 | 0.9681 ± 0.0010 |
| AUC-ROC | 0.9983 ± 0.0001 | 0.9972 ± 0.0001 |
| F1-Score | 0.9830 ± 0.0004 | 0.9753 ± 0.0007 |

TABLE XVII: Hibrit ve ExtraTrees Modelleri Arasındaki Fark (Δ) ve p-Degerleri

| Metrik | Δ (Fark) | p-değeri |
|-----------|-----------------|------------------|
| Accuracy | 0.0076 | <0.001 |
| Precision | 0.0027 | <0.001 |
| Recall | 0.0126 | ≈ 0.0004 |
| AUC-ROC | 0.0011 | <0.001 |
| F1-Score | 0.0077 | <0.001 |

I. Temsillerin Model Performansına Etkisi Analizi

1) PLS-DA Temsilinin Model Performansına Etkisi

Bu çalışmada hibrit model, ML ve DL modellerinin, boyut indirgeme amacıyla kullanılan PLS-DA temsili üzerindeki

sınıflandırma performanslarına etkisi analiz edilmiştir.

Hibrit Modelin PLS-DA Temsili Üzerindeki Performansı

Tablo XVIII'de hibrit modelin PLS-DA temsili üzerindeki sınıflandırma sonuçları, Tablo XIX'da ise modelin eğitim süresi, gecikme ve throughput gibi hesaplama performansı göstergeleri sunulmuştur. Hibrit model, PLS-DA temsili üzerinde test doğruluğu %96.9, F1-score %96.9 ve özellikle AUC-ROC %99.53 ile oldukça güçlü sonuçlar elde etmiştir. Bu bulgular, modelin hem benign hem de malicious örnekleri yüksek doğrulukla ayırt ettiğini göstermekte; siber güvenlik uygulamalarında kritik kabul edilen AUC-ROC'un %99.5'in üzerinde olması, zararlı yazılımların tespitinde yüksek duyarlılığın korunduğunu teyit etmektedir. Modelin hesaplama performansı da oldukça verimlidir: SAE boyutu 0.88 MB, LGBM 30.37 MB, toplam 31.26 MB'lık kompakt bir yapı ile yalnızca 195.3 s içinde eğitilmiş ve 0.021 ms/ornek gecikme ile 48 140 örnek/saniye hızında çıkışım (inference) yapmıştır. Buna karşın, OFF (ham özellik uzayı) temsili üzerinde hibrit modelin eğitimi daha uzun sürmüştür (\approx 493 s), ancak çıkışım süresi 0.015 ms/ornek ve throughput değeri 64 779 örnek/saniye ile oldukça yüksek bir hızda gerçekleşmiştir. Bu durum, OFF temsili verinin doğrudan modele sunulması sayesinde ağır daha verimli optimize edilemeye açıklanabilir. Ayrıca, OFF varyantı $AUC \approx 0.998$ ile en yüksek genel doğruluğu sağlamış, böylece çalışmanın asıl (referans) sonucu olarak kabul edilmiştir. Sonuç olarak, PLS-DA temsili hesaplama açısından daha kompakt ve hızlı olsa da, OFF temsili hem doğruluk hem de genelleme başarımı bakımından daha üstün bir performans sergileyerek hibrit modelin en verimli varyantı olarak öne çıkmıştır.

TABLE XVIII: PLS-DA Üzerinde Hibrit Modelin Test Performans Sonuçları

| Metrik | Sonuç (Test Ortalaması \pm Std) |
|-------------|-----------------------------------|
| Accuracy | 0.9693 ± 0.0006 |
| Precision | 0.9738 ± 0.0022 |
| Recall | 0.9646 ± 0.0018 |
| Specificity | 0.9740 ± 0.0023 |
| F1-Score | 0.9692 ± 0.0005 |
| F1-Weighted | 0.9693 ± 0.0006 |
| AUC-ROC | 0.9953 ± 0.0001 |

TABLE XIX: Hibrit Modelin Hesaplama Performans Metrikleri (PLS-DA Temsili)

| Metrik | Değer (Ortalama) | Açıklama |
|----------------|---------------------------------------|------------------------|
| Model Boyutu | 31.26 MB (\approx 0.23M parametre) | Toplam model boyutu |
| Inference Time | 1.663 s ($n = 80\,079$) | Test çıkışım süresi |
| Latency | 0.021 ms/ornek | Ortalama gecikme |
| Throughput | 48 140 örnek/s | İşleme hızı |
| Eğitim Süresi | 195.355 s | Ortalama eğitim süresi |

2) Hibrit Modelin PLS-DA Temsilinde Sınıflandırma Performansı (Confusion Matrix)

PLS-DA veri temsiline ait hibrit model confusion matrix'i Şekil 7'de sunulmuştur. Sonuçlar incelendiğinde, modelin Benign sınıfında 194.996 doğru / 5.202 yanlış pozitif, Malicious sınıfında ise 193.108 doğru / 7.091 yanlış negatif ürettiği görülmektedir. Bu değerler, hibrit modelin PLS-DA temsili üzerinde genel olarak kabul edilebilir bir doğruluk sunduğunu ancak OFF temsiline kıyasla belirgin şekilde daha yüksek hata oranları ürettiğini göstermektedir.

Özellikle yanlış negatif sayısının (7.091) OFF temsiline göre belirgin ölçüde artması, saldıri trafığının ayrıstırılmasında PLS-DA'nın hibrit mimariye beklenen seviyede katkı sağlayamadığını ortaya koymaktadır. Benzer şekilde yanlış pozitif hatalar da OFF

temsiline göre yükselmiştir. Bu durum, PLS-DA'nın veri yapısını lineer bir şekilde küçültüğü için özellikle karmaşık ve yüksek varyanslı ağ trafığında sınıflandırma gücünü bir miktar zayıflatlığını işaret etmektedir.

Bununla birlikte, PLS-DA gibi lineer projeksiyon temelli bir boyut indirgeme yönteminin oldukça agresif bir şekilde bilgi kaybı oluşturmasına rağmen hibrit mimarinin hâlâ yüksek sayıda doğru sınıflandırma yapabilmesi, modelin genel dayanıklılık kapasitesinin güçlü olduğunu göstermektedir. Dolayısıyla PLS-DA temsilinin OFF kadar başarılı olmasa da, hibrit yapının bu temsilde dahi tatmin edici ve istikrarlı bir performans sergilediği söylenebilir.

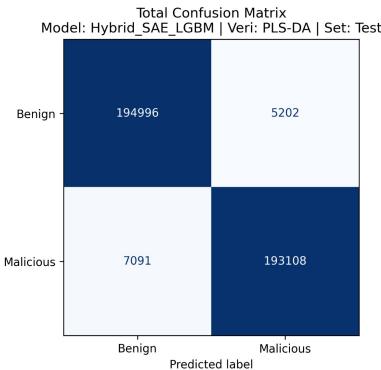


Fig. 7: Hibrit Modelin PLS-DA temsilindeki confusion Matrix Sonucusu

ML Modellerinin PLS-DA Temsili Üzerindeki Performansı

ML modellerine ait karşılaştırmalı performans değerleri Tablo XX'de sunulmuştur.

PLS-DA temsili kullanıldığında, makine öğrenmesi modellerinin doğruluk oranlarının OFF sonuçlara kıyasla yaklaşık %3 ila %9 aralığında azaldığı gözlemlenmiştir. Bu düşüşün temel nedeni, PLS-DA'nın lineer bir boyut indirgeme yöntemi olması ve yüksek boyutlu, karmaşık ve non-lineer dağılıma sahip EMBER24 veri setindeki ilişkileri tam olarak temsil edememesidir. Bununla birlikte, uygun hiperparametre ayarları ve dikkatli model optimizasyonu sonucunda PLS-DA temsiliyle de yüksek ve tutarlı performans değerleri elde edilmiştir. Özellikle CatBoost ve ExtraTrees modelleri, PLS-DA dönüşümü sonrasında bile genel doğruluk, F1-score ve AUC-ROC metriklerinde güçlü sonuçlar üretmiş; bu durum PLS-

DA'nın, lineer yapısına rağmen, doğru yapılandırıldığından karmaşık veri setleri üzerinde dahi etkili bir öznitelik temsili sağlayabildiğini göstermektedir.

DL Modellerinin PLS-DA Temsili Üzerindeki Performansı

Tablo XXI'de sunulan PLS-DA temsili altında derin öğrenme modellerinin performans sonuçları incelendiğinde, OFF durumuna kıyasla doğruluk değerlerinde yalnızca %1–1.5 civarında bir düşüş gözlemlenmiştir. Bu farkın sınırlı kalmasının temel nedeni, derin öğrenme modellerinin optimum hiperparametrelerle yeniden yapılandırılması ve temsili dönüşümünün bilgi kaybını telafi edecek biçimde öğrenme kapasitesinin artırılmasıdır. Normal koşullarda PLS-DA gibi lineer bir temsili yöntemi, karmaşık ve non-lineer veri ilişkilerini tam olarak yakalayamadığından performansın daha belirgin biçimde azalması beklenir. Ancak bu çalışmada, özellikle MLP-Mixer ve gMLP gibi modellerin dikkatli optimizasyonu sayesinde PLS-DA temsiliyle de yüksek genelleme ve tutarlı doğruluk değerleri elde edilmiştir.

Ote yandan, hibrit mimari, PLS-DA temsili altında derin öğrenme modellerine göre yaklaşık 0.5 puanlık bir farkla daha

TABLE XX: PLS-DA Temsili Üzerinde ML Öğrenmesi Modelerin Test Performans Sonuçları (Ortalama \pm Std)

| Model | Metrik | Sonuç (Test Ortalama \pm Std) |
|---------------|-------------|---------------------------------|
| CatBoost | Doğruluk | 0.8936 \pm 0.0033 |
| | Precision | 0.9045 \pm 0.0053 |
| | Recall | 0.8801 \pm 0.0013 |
| | Specificity | 0.9071 \pm 0.0056 |
| | F1 | 0.8922 \pm 0.0031 |
| | F1-W | 0.8936 \pm 0.0033 |
| | AUC-ROC | 0.9598 \pm 0.0017 |
| ExtraTrees | Doğruluk | 0.9313 \pm 0.0014 |
| | Precision | 0.9393 \pm 0.0022 |
| | Recall | 0.9223 \pm 0.0007 |
| | Specificity | 0.9403 \pm 0.0033 |
| | F1 | 0.9307 \pm 0.0013 |
| | F1-W | 0.9313 \pm 0.0014 |
| | AUC-ROC | 0.9763 \pm 0.0005 |
| GradientBoost | Doğruluk | 0.8795 \pm 0.0042 |
| | Precision | 0.8901 \pm 0.0041 |
| | Recall | 0.8658 \pm 0.0019 |
| | Specificity | 0.8931 \pm 0.0041 |
| | F1 | 0.8788 \pm 0.0022 |
| | F1-W | 0.8795 \pm 0.0042 |
| | AUC-ROC | 0.9500 \pm 0.0013 |
| HistGB | Doğruluk | 0.8756 \pm 0.0028 |
| | Precision | 0.8836 \pm 0.0052 |
| | Recall | 0.8651 \pm 0.0017 |
| | Specificity | 0.8860 \pm 0.0059 |
| | F1 | 0.8743 \pm 0.0024 |
| | F1-W | 0.8756 \pm 0.0028 |
| | AUC-ROC | 0.9492 \pm 0.0016 |
| SGDClassifier | Doğruluk | 0.5775 \pm 0.0281 |
| | Precision | 0.6246 \pm 0.1048 |
| | Recall | 0.5013 \pm 0.1687 |
| | Specificity | 0.6536 \pm 0.2064 |
| | F1 | 0.5297 \pm 0.0835 |
| | F1-W | 0.5631 \pm 0.0275 |
| | AUC-ROC | 0.5775 \pm 0.0281 |

düşük doğruluk elde etse de, hesaplama maliyeti açısından belirgin biçimde daha verimli çalışmış ve benzer doğruluk düzeyini daha kısa süre ve daha düşük kaynak kullanımıyla sağlamıştır. Bu durum, hibrit yapının genel verimliliğini ve uygulama ölçüğinde maliyet-performans dengesi açısından üstünlüğünü ortaya koymaktadır.

Genel olarak değerlendirildiğinde, PLS-DA lineer bir boyut indirgeme yöntemi olduğundan, özellikle saldırı örtüntülerindeki non-lineer ilişkileri temsil etmede sınırlı kalmaktadır. Bu nedenle ML modellerinde doğruluk oranlarında yaklaşık %2–6 aralığında bir düşüş gözlemlenmiştir. Buna karşın DL modelleri, yüksek genelleme kapasiteleri sayesinde bu bilgi kaybını büyük ölçüde telafi etmiş ve doğruluk kaybı yalnızca

Önerilen Hibrit model, PLS-DA temsili altında dahi %96.9 doğruluk ve %99.5 AUC-ROC değerleriyle en yüksek genel başarıyı göstermiştir. Bu performans, yalnızca yüksek doğruluğu değil, aynı zamanda düşük varyansı (Accuracy std \pm 0.0006, AUC-ROC std \pm 0.0001) ile modelin son derece kararlı bir yapıda olduğunu ortaya koymaktadır.

PLS-DA temsili altında doğruluk ve AUC-ROC değerlerindeki düşüş, ham varyanta kıyasla yalnızca %1–3 düzeyinde kalmıştır; bu durum, özellikle DL modellerinin lineer temsil kaynaklı bilgi kaybını etkili biçimde telafi edebildiğini göstermektedir. ML modellerindeki azalma daha belirgin olsa da, CatBoost ve ExtraTrees hâlâ istatistiksel olarak anlamlı şekilde yüksek performans üretmiştir ($p < 0.01$).

Sonuç olarak, Hibrit model, PLS-DA temsili altında bile en düşük varyansla en kararlı sonuçları üretmiş; elde edilen başarı

TABLE XXI: PLS-DA Temsili Üzerinde DL Öğrenme Modelerin Test Performans Sonuçları (Ortalama \pm Std)

| Model | Metrik | Sonuç (Test Ortalama \pm Std) |
|-----------|-------------|---------------------------------|
| DNN | Doğruluk | 0.9630 \pm 0.0013 |
| | Precision | 0.9702 \pm 0.0018 |
| | Recall | 0.9553 \pm 0.0031 |
| | Specificity | 0.9707 \pm 0.0021 |
| | F1 | 0.9627 \pm 0.0013 |
| | F1-W | 0.9630 \pm 0.0013 |
| | AUC-ROC | 0.9935 \pm 0.0003 |
| MLP-Mixer | Doğruluk | 0.9625 \pm 0.0025 |
| | Precision | 0.9694 \pm 0.0025 |
| | Recall | 0.9552 \pm 0.0036 |
| | Specificity | 0.9699 \pm 0.0026 |
| | F1 | 0.9618 \pm 0.0006 |
| | F1-W | 0.9625 \pm 0.0025 |
| | AUC-ROC | 0.9931 \pm 0.0003 |
| ResNetMLP | Doğruluk | 0.9629 \pm 0.0019 |
| | Precision | 0.9707 \pm 0.0020 |
| | Recall | 0.9547 \pm 0.0023 |
| | Specificity | 0.9712 \pm 0.0019 |
| | F1 | 0.9626 \pm 0.0011 |
| | F1-W | 0.9629 \pm 0.0019 |
| | AUC-ROC | 0.9934 \pm 0.0004 |
| Wide&Deep | Doğruluk | 0.9623 \pm 0.0015 |
| | Precision | 0.9686 \pm 0.0026 |
| | Recall | 0.9585 \pm 0.0020 |
| | Specificity | 0.9661 \pm 0.0027 |
| | F1 | 0.9621 \pm 0.0015 |
| | F1-W | 0.9623 \pm 0.0015 |
| | AUC-ROC | 0.9915 \pm 0.0005 |
| gMLP | Doğruluk | 0.9636 \pm 0.0016 |
| | Precision | 0.9699 \pm 0.0028 |
| | Recall | 0.9569 \pm 0.0028 |
| | Specificity | 0.9703 \pm 0.0028 |
| | F1 | 0.9628 \pm 0.0014 |
| | F1-W | 0.9636 \pm 0.0016 |
| | AUC-ROC | 0.9936 \pm 0.0004 |

rastlanışal değil, istatistiksel olarak güçlü ($p < 0.001$) ve tekrarlanabilir bir performans ortaya koymuştur.

3) VAE Temsilinin Model Performansına Etkisi Hibrit Modelin VAE Temsili Üzerindeki Performansı

Tablo XXII'de hibrit modelin VAE temsili üzerindeki performansına göre, modelin en yüksek genel başarıyı koruduğu gözlemlenmiştir. Ham veri durumuna kıyasla doğruluk değerinde yalnızca yaklaşık 0.0016'lık çok küçük bir düşüş meydana gelmiş olup, bu fark istatistiksel olarak ihmali edilebilir düzeydedir. Bu sonuç, hibrit modelin VAE temsili altında karmaşık ve non-lineer veri ilişkilerini etkin biçimde temsil edebilmesi sayesinde, bilgi kaybı yaşamadan genelleme gücünü koruduğunu göstermektedir. Sonuç olarak, hibrit mimari (SAE + LGBM) hem VAE hem de OFF temsillerinde yüksek doğruluk (\approx 0.98), güçlü AUC-ROC (\approx 0.998) ve düşük varyans değerleriyle tutarlı ve kararlı performans sergilemiştir. Bu bulgu, hibrit yaklaşımın veri temsiline bağımlı olmaksızın yüksek genelleme kapasitesi ve hesaplama verimliliği sunduğunu metodolojik olarak doğrulamaktadır.

TABLE XXII: VAE Temsili Üzerinde Hibrit Modelin 5-Kat Çapraz Doğrulama Sonuçları

| Metrik | Sonuç (Test Ortalama \pm Std) |
|-------------|---------------------------------|
| Accuracy | 0.9814 \pm 0.0005 |
| Precision | 0.9853 \pm 0.0012 |
| Recall | 0.9775 \pm 0.0016 |
| Specificity | 0.9854 \pm 0.0013 |
| F1-Score | 0.9814 \pm 0.0005 |
| F1-Weight | 0.9814 \pm 0.0005 |
| AUC-ROC | 0.9981 \pm 0.0001 |

4) Hibrit Modelin VAE Temsilinde Sınıflandırma Performansı (Confusion Matrix)

VAE veri temsiline ait hibrit modelin confusion matrix’i şekil 8’de sunulmuştur. Matriç incelendiğinde, modelin Benign sınıfında 197.279 doğru / 2.919 yanlış pozitif, Malicious sınıfında ise 195.687 doğru / 4.512 yanlış negatif ürettiği görülmektedir. Bu dağılım genel olarak dengeli olsa da, OFF temsili altında elde edilen sonuçlarla karşılaşıldığında hibrit modelin VAE temsilinde daha yüksek hata oranları ürettiği ve saldırı tespit性能ansında belirgin bir düşüş yaşadığını anlaşılmaktadır.

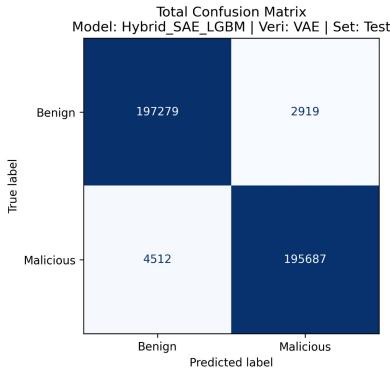


Fig. 8: Hibrit Modelin VAE temsilindeki confusion Matrix Sonucu

OFF temsilinde hibrit model, yanlış pozitif ve yanlış negatif değerlerini oldukça düşük seviyede tutarak en güçlü performansını ortaya koymuştur. VAE temsili altında ise hata sayılarının artması, özellikle saldırı trafisinin ayrıştırılmasında OFF temsiline kıyasla daha zayıf bir performansa işaret etmektedir. Bununla birlikte, VAE’nin çok daha sıkıştırılmış ve düşük boyutlu bir temsil sunduğu göz önüne alındığında, hibrit mimarının bu temsil altında dahi yüksek doğruluk ve dengeli bir sınıflandırma başarısı üretmiş olması dikkat çekicidir.

Sonuç olarak, VAE temsili OFF’ya göre daha zorlu bir senaryo sunmasına rağmen hibrit modelinin bu temsilde bile başarılı, istikrarlı ve pratik olarak kullanılabilir bir performans ortaya koyduğu söylenebilir.

ML Modellerinin VAE Temsili Üzerindeki Performansı

Tablo XXIII’de makine öğrenmesi modellerinin VAE temsili üzerindeki sonuçları sunulmuştur. OFF durumuna kıyasla doğruluk değerlerinde belirgin ancak kabul edilebilir düzeyde düşüller gözlemlenmiştir. Bu farkın temel nedeni, VAE’nin öznitelikleri yeniden olasılıksal biçimde temsil etmesi nedeniyle bazı ayrıştırıcı bilgilerin kısmen yumoşatılmıştır. Buna rağmen, CatBoost ve ExtraTrees modelleri, doğruluk ($\approx 0.88\text{--}0.93$) ve AUC-ROC ($\approx 0.95\text{--}0.97$) değerleriyle yüksek genelleme başarısını korumuştur. Bu durum, VAE temsillerinin doğru yapılandırıldığından ağaç tabanlı algoritmalar için yeterli bilgi yoğunluğu sunduğunu göstermektedir.

Öte yandan, SGDClassifier modeli hem OFF hem de VAE temsillerinde düşük sonuçlar üretmiş; doğruluk (≈ 0.52) ve F1 (≈ 0.22) değerleri, bu modelin söz konusu veri seti ve temsil yapısı için uygun olmadığını açık biçimde ortaya koymuştur. Sonuç olarak, VAE temsili altında makine öğrenmesi modelleri genel olarak istikrarlı ancak ham veriye göre sınırlı bir performans sergilemiş; buna karşın hesaplama verimliliği ve genelleme dengesi açısından tammin edici sonuçlar elde edilmiştir.

TABLE XXIII: VAE Temsili Üzerinde Makine Öğrenmesi Modelerinin Test Performans Sonuçları (Ortalama \pm Std)

| Model | Metrik | Sonuç (Test Ortalama \pm Std) |
|---------------|-------------|---------------------------------|
| CatBoost | Doğruluk | 0.8810 \pm 0.0014 |
| | Precision | 0.8969 \pm 0.0021 |
| | Recall | 0.8610 \pm 0.0018 |
| | Specificity | 0.9010 \pm 0.0023 |
| | F1 | 0.8786 \pm 0.0014 |
| | F1-W | 0.8810 \pm 0.0014 |
| | AUC-ROC | 0.9525 \pm 0.0008 |
| ExtraTrees | Doğruluk | 0.9265 \pm 0.0015 |
| | Precision | 0.9418 \pm 0.0027 |
| | Recall | 0.9093 \pm 0.0015 |
| | Specificity | 0.9438 \pm 0.0028 |
| | F1 | 0.9253 \pm 0.0014 |
| | F1-W | 0.9265 \pm 0.0015 |
| | AUC-ROC | 0.9771 \pm 0.0006 |
| GradientBoost | Doğruluk | 0.8614 \pm 0.0019 |
| | Precision | 0.8693 \pm 0.0030 |
| | Recall | 0.8507 \pm 0.0042 |
| | Specificity | 0.8721 \pm 0.0037 |
| | F1 | 0.8599 \pm 0.0020 |
| | F1-W | 0.8614 \pm 0.0019 |
| | AUC-ROC | 0.9387 \pm 0.0015 |
| HistGB | Doğruluk | 0.8604 \pm 0.0015 |
| | Precision | 0.8692 \pm 0.0038 |
| | Recall | 0.8486 \pm 0.0023 |
| | Specificity | 0.8723 \pm 0.0045 |
| | F1 | 0.8588 \pm 0.0011 |
| | F1-W | 0.8604 \pm 0.0015 |
| | AUC-ROC | 0.9391 \pm 0.0008 |
| SGDClassifier | Doğruluk | 0.5209 \pm 0.0186 |
| | Precision | 0.6043 \pm 0.1016 |
| | Recall | 0.1382 \pm 0.0249 |
| | Specificity | 0.9036 \pm 0.0418 |
| | F1 | 0.2231 \pm 0.0328 |
| | F1-W | 0.4382 \pm 0.0187 |
| | AUC-ROC | 0.6071 \pm 0.0758 |

DL Modelerinin VAE Temsili Üzerindeki Performansı

Tablo XXIV’da, derin öğrenme modellerinin VAE temsili üzerindeki performans sonuçları sunulmuştur. OFF (ham veri) durumuna kıyasla doğruluk değerlerinde yaklaşık %2–2.5 oranında bir düşüş gözlemlenmiş, ancak bu fark genelleme performansını zayıflatacak düzeyde olmamıştır. Bu düşüşün temel nedeni, VAE’nin veriyi olasılıksal bir latent uzaya dönüştürmesi sonucu öznitelikler arasındaki deterministik ilişkilerin kısmen yumuşamasıdır.

Buna karşın, özellikle Wide&Deep, MLP-Mixer ve gMLP modelleri, doğruluk ($\approx 0.94\text{--}0.95$) ve AUC-ROC ($\approx 0.988\text{--}0.989$) değerleriyle hâlen iteratürde raporlanan birçok çalışmaya kıyasla yüksek performans sergilemiştir. Bu sonuç, VAE’nin non-lineer veri dağılımlarını yakalayabilme yeteneği sayesinde modellerin bilgi kaybına rağmen güçlü genelleme kabiliyeti koruduğunu göstermektedir.

Genel olarak, derin öğrenme modelleri VAE temsili altında hafif doğruluk kaybı yaşamış ancak istikrarlı, kararlı ve yüksek AUC-ROC performansı ile sonuçlarını sürdürmüştür. Bu bulgular, VAE’nin uygun hiperparametrelerle optimize edildiğinde derin modellerin temsil öğrenme kapasitesini destekleyici bir mekanizma sunduğunu metodolojik olarak doğrulamaktadır.

TABLE XXIV: VAE Temsili Üzerinde Derin Öğrenme Modelerinin Test Performans Sonuçları (Ortalama \pm Std)

| Model | Metrik | Sonuç (Test Ortalama \pm Std) |
|-----------|-------------|---------------------------------|
| DNN | Doğruluk | 0.9462 \pm 0.0099 |
| | Precision | 0.9566 \pm 0.0106 |
| | Recall | 0.9349 \pm 0.0114 |
| | Specificity | 0.9575 \pm 0.0105 |
| | F1 | 0.9456 \pm 0.0101 |
| | F1-W | 0.9462 \pm 0.0099 |
| | AUC-ROC | 0.9886 \pm 0.0035 |
| MLP-Mixer | Doğruluk | 0.9486 \pm 0.0080 |
| | Precision | 0.9597 \pm 0.0021 |
| | Recall | 0.9365 \pm 0.0161 |
| | Specificity | 0.9607 \pm 0.0021 |
| | F1 | 0.9479 \pm 0.0086 |
| | F1-W | 0.9485 \pm 0.0080 |
| | AUC-ROC | 0.9890 \pm 0.0029 |
| ResNetMLP | Doğruluk | 0.9469 \pm 0.0084 |
| | Precision | 0.9587 \pm 0.0062 |
| | Recall | 0.9341 \pm 0.0141 |
| | Specificity | 0.9597 \pm 0.0061 |
| | F1 | 0.9462 \pm 0.0088 |
| | F1-W | 0.9469 \pm 0.0084 |
| | AUC-ROC | 0.9889 \pm 0.0033 |
| Wide&Deep | Doğruluk | 0.9495 \pm 0.0080 |
| | Precision | 0.9586 \pm 0.0037 |
| | Recall | 0.9395 \pm 0.0133 |
| | Specificity | 0.9595 \pm 0.0033 |
| | F1 | 0.9490 \pm 0.0084 |
| | F1-W | 0.9495 \pm 0.0080 |
| | AUC-ROC | 0.9884 \pm 0.0028 |
| gMLP | Doğruluk | 0.9493 \pm 0.0077 |
| | Precision | 0.9590 \pm 0.0055 |
| | Recall | 0.9387 \pm 0.0118 |
| | Specificity | 0.9599 \pm 0.0053 |
| | F1 | 0.9487 \pm 0.0080 |
| | F1-W | 0.9493 \pm 0.0077 |
| | AUC-ROC | 0.9892 \pm 0.0028 |

Zamana Dayalı Model Performans Analizi

(Temporal Evaluation Experiments)

Zaman-Temelli Deney Tasarımı, Model Seçimi ve Temporal Drift Analizi

Temporal deneyin temel amacı, geçmiş verilerden yararlanarak gelecekte karşılaşılabilen örnekleri tahmin edebilmektir. Bu yaklaşımada model, yalnızca geçmişe ait gözlemler üzerinden öğrenerek gerçekçi bir zaman serisi genellemesi sağlar. Rastgele bölmeye yöntemleri veri sırasını bozduğundan, temporal senaryolarda ciddi bir veri sizıntısı (data leakage) riski oluşturabilir; buna karşılık temporal bölmeye yöntemi kronolojik yapıyı koruyarak bu riski tamamen ortadan kaldırır. Dosya veya hafta bazlı zaman sırasının korunması, özellikle tehditlerin sürekli evrildiği malware tespiti gibi alanlarda gerçek dünya koşullarının daha doğru temsil edilmesi açısından kritik öneme sahiptir. Bu sayede modelin daha önce hiç karşılaşmadığı gelecekteki saldırı örneklerine karşı genellemeye kapasitesi güvener bir biçimde değerlendirilebilir.

Bu çalışmada, modelin zamana bağlı olarak değişen veri dağılımlarına karşı dayanıklılığını değerlendirmek amacıyla temporal drift analizi gerçekleştirilmiştir. 5-Fold çapraz doğrulama aşamasında tüm makine öğrenmesi (ML) ve derin öğrenme (DL) modelleri üç farklı veri temsili (OFF, PLS ve VAE) üzerinde test edilmiş ve kapsamlı bir performans karşılaştırması yapılmıştır. Temporal analiz aşamasında ise, tüm modelleri yeniden eğitmek yerine yalnızca en yüksek başarı gösteren iki ML modeli (Extra-Trees ve CatBoost), iki DL modeli (DNN ve gMLP) ile önerilen Hibrit model seçilmiştir. Bu seçim, hem hesaplama verimliliğini

artırmak hem de zaman temelli senaryolarda pratik olarak uygulanabilir ve genellemeye kapasitesi yüksek modellerin performansını değerlendirmek amacıyla yapılmıştır.

Temporal analizde, zaman düzenini bozabilecek rastgele çapraz doğrulama (K-Fold) yöntemleri yerine, zaman bağımlılığını koruyan tek seferlik (hold-out) bir bölmeye stratejisi uygulanmıştır. Veri kümesi haftalık kronolojik sıraya göre dizilmiş ve bu yapı JSONL dosyaları üzerinden doğrulanmıştır.

Eğitim seti, ilk 53 haftayı kapsamaktadır. Bu dönem aşağıdaki JSONL kayıtlarıyla doğrulanmıştır:

- İlk hafta (Train): 2023-09-24_2023-09-30_Win32_train.jsonl
- Son hafta (Train): 2024-09-15_2024-09-21_Win32_train.jsonl

Bu aralık toplam 200.000 örnek (Benign: 100.000, Attack: 100.000) içermektedir.

Temporal test seti ise, eğitim döneminin hemen ardından gelen 12 haftayı kapsamaktadır. Bu dönemde aşağıdaki JSONL kayıtlarıyla doğrulanmıştır:

- İlk hafta (Test): 2024-09-22_2024-09-28_Win32_test.jsonl
- Son hafta (Test): 2024-12-08_2024-12-14_Win32_test.jsonl

Bu test seti toplam 199.990 örnekten oluşmaktadır (Benign: 99.994, Attack: 99.996).

Bu yapı sayesinde eğitim ve test kümeleri arasında hiçbir zaman sizıntıları olmamış, gerçekçi bir gelecek tahmini senaryosunu elde edilmiştir. Eğitim ve temporal test kümelerine ait sınıf dağılımı Şekil 9'de sunulmuştur.

5-Fold çapraz doğrulama deneyinde kullanılan tüm veri ön işleme ve öznitelik çıkarma adımları temporal alt kümelere de aynı şekilde uygulanmıştır. Eğitim ve test setleri birbirinden bağımsız olarak temizlenmiş; varyans ve korelasyon analizleri her bir kümeye ayrı ayrı uygulanmıştır. Temporal test setinde düşük varyanslı 37 sütun ve yüksek korelasyon göstergen 110 özellik çifti belirlenmiş; bu analiz sonucunda 42 özellik yüksek korelasyon nedeniyle veri kümelerinden çıkarılmıştır.

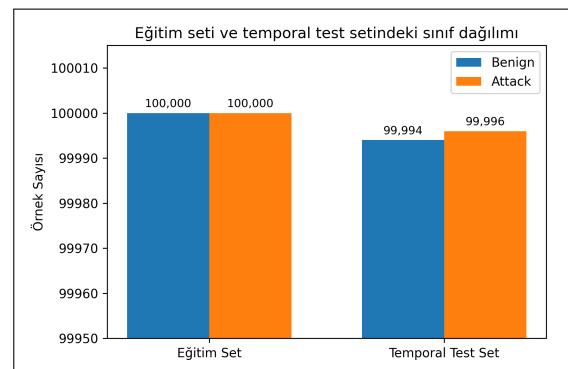


Fig. 9: Temporal deneyde kullanılan eğitim ve temporal test setlerindeki sınıf dağılımı

Temporal senaryoda kullanılan hibrit modelin işlem hattı Şekil 10'da sunulmuştur. Bu akış, StandardScaler ile ölçekleme ve varyans eşigi kontrolü sonrasında elde edilen özelliklerin Stacked Autoencoder tabanlı gizil gösterime dönüştürülmesi ve eşzamanlı olarak kazanç temelli Top-K öznitelik seçimiyle zenginleştirilmesi adımlarını içermektedir. Her iki temsilin birleştirilmesiyle oluşturulmuş nihai özellik seti LightGBM sınıflandırıcısına aktarılmakta ve temporal test kümesi üzerinde değerlendirme gerçekleştirilmektedir. Bu yapı, temporal veri akışında ortaya çıkan özellik kaymaları karşısında temsil gücünü artırarak

modelin zaman bağımlı genelleme performansını iyileştirmeyi amaçlamaktadır.

Temporal senaryoda tüm modeller varsayılan 0.5 eşik değeriyle değerlendirilirken yalnızca hibrit model için ek bir karar eşiği optimizasyonu uygulanmıştır. Bu eşinin belirlenmesinde önce çoklu-metrik hedef dikkate alınmış; bu hedef karşılanamadığında Youden J ölçüyü (TPR–FPR) kullanılarak en yüksek performans veren eşik seçilmiştir. Böylece karar sınırı iyileştirmesi yalnızca hibrit modele özgü bir adım olarak tasarılmıştır.”

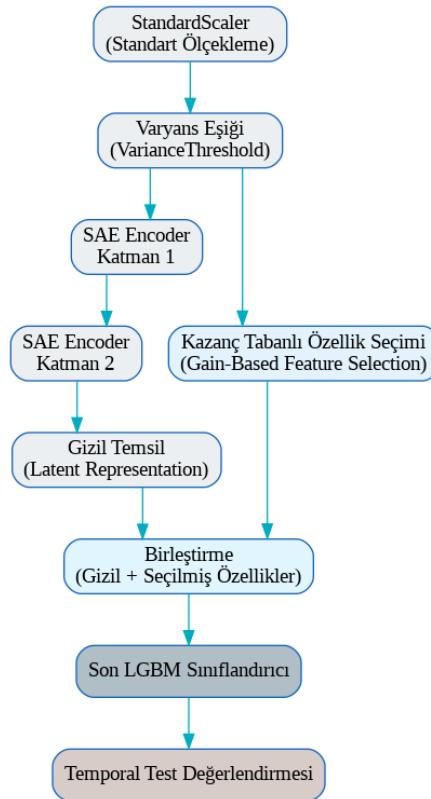


Fig. 10: Temporal senaryoda hibrit SAE–Top-K–LGBM modelinin işlem hattı.

Temporal Drift Analizinde Eğitim ve Test Öz niteliklerinin Hizalanması (Feature Alignment)

Temporal senaryoda eğitim ve test kümeleri farklı haftalardan geldiği için, özellik uzayında doğal bir zaman kayması (feature drift) ortaya çıkmıştır. Nitekim eğitim setinde 2.504 özellik bulunurken, temporal test setinde temizlik sonrasında yalnızca 2.493 özellik kalmıştır; bu fark, bazı öz niteliklerin zamanla bilgi taşıma gücünü yitirdiğini veya daha yüksek korelasyon göstermeye başlayarak veri kümelerinden elendiğini göstermektedir. Bu nedenle kodlama aşamasında eğitim ve test veri kümeleri parquet formatından okunmuş, tüm öz nitelikler float32 tipine dönüştürülmüş ve kolon adları karşılaştırılarak eğitimde olup teste eksik kalan özellikler nötr bir başlangıç değeri olan 0.0 ile test setine eklenmiş, test setine özgü fazla kolonlar ise çıkarılmıştır. Ardından test kümelerinin sütun sıralaması eğitim kümlesiyle birebir aynı hâle getirilmiş, kolon sayısı ve veri tipi denetimlerle doğrulanmış ve hizalanmış test kümlesi kayıt altına alınmıştır. Bu süreç, temporal deneye de eğitim–test ayrimını bozmadan öz nitelik uzayının tutarlı tutulmasını sağlamış ve gelecekteki veriler üzerinde modelin yeniden eğitilmesi veya adaptasyon mekanizmalarının gereklileceğine işaret etmiştir.

Temporal Drift Senaryosunda Hiperparametre Uyarlaması (ML, DL ve Hibrit Modeller)

Temporal senaryo, 5-fold deney üzerinden dağılımsal olarak belirgin şekilde farklılık gösterdiği için modeller bu aşamada yeniden optimize edilmiştir. Zaman bağımlı veri setlerinde aynı hiperparametrelerin kullanılması metodolojik olarak doğru kabul edilmemektedir; çünkü en uygun ayarlar, veri dağılımindaki kaymaların (drift) şiddetine ve yapısına göre değişiklik göstermektedir. Bu nedenle temporal deneylerde kullanılan temel hiperparametreler, ilgili modeller için yeniden belirlenmiş ve özet bir biçimde Tablo XXV’da sunulmuştur. Tablo, ML, DL ve hibrit modellerin temporal koşullara uyarlanmış çekirdek ayarlarını sistematik şekilde göstermektedir.

TABLE XXV: Temporal test senaryosunda kullanılan Hibrit, ML ve DL modellerinin temel hiperparametreleri.

| Model | Parametre | Değer |
|--|-------------------|---------------------|
| Hibrit Model (SAE + Top-K + LGBM) | | |
| Hybrid | SAE latent_dim | 192 |
| Hybrid | Top-K features | 256 |
| Hybrid | LGBM n_estimators | 6000 |
| Makine Öğrenmesi Modelleri | | |
| CatBoost | iterations | 150 |
| CatBoost | learning_rate | 0.3 |
| ExtraTrees | n_estimators | 200 |
| ExtraTrees | max_depth | 12 |
| Derin Öğrenme Modelleri | | |
| DNN | hidden_units | (512, 256, 128, 64) |
| DNN | dropout | 0.3 |
| gMLP-Tabular | blocks | 2 |
| gMLP-Tabular | dim_ff | 256 |

Temporal Drift Senaryosunda Modellerin Karşılaştırmalı Performans Analizi

Tablo XXVI ve performansı ilk üç sırada yer alan modellerin sonuçlarını gösteren Şekil 11 incelendiğinde, hibrit modelin klasik makine öğrenmesi tabanlı modellere (CatBoost ve ExtraTrees) kıyasla daha dengeli ve yüksek bir performans ortaya koyduğu açık biçimde görülmektedir. Özellikle Accuracy (%96.70), Recall (%96.10), F1 (%96.68), F1-Weighted (%96.70) ve AUC-ROC (%99.53) değerlerinin diğer modellere kıyasla belirgin biçimde yüksek olması, hibrit mimarının hem genel doğrulukta hem de pozitif (saldıri) sınıfına yönelik duyarlılıkta güçlü bir genelleme yeteneğine sahip olduğunu göstermektedir. Ayrıca, hibrit modele ait ek hata oranları ve tamamlayıcı doğrulama metriklerinin sunulduğu Tablo XXVII, modelin düşük FPR ve FNR değerleriyle birlikte istikrarlı bir hata profiline sahip olduğunu ortaya koyarak bu değerlendirmeyi desteklemektedir.

TABLE XXVI: Temporal Test Senaryosunda Tüm Modellerin Performans Sonuçları

| Model | Acc | Rec | Spec | Prec | F1 | FIW | AUC |
|------------|--------|--------|--------|--------|--------|--------|--------|
| Hibrit | 0.9670 | 0.9610 | 0.9730 | 0.9726 | 0.9668 | 0.9670 | 0.9953 |
| CatBoost | 0.9574 | 0.9450 | 0.9698 | 0.9691 | 0.9569 | 0.9574 | 0.9933 |
| ExtraTrees | 0.9391 | 0.9255 | 0.9528 | 0.9514 | 0.9383 | 0.9391 | 0.9889 |
| DNN | 0.9210 | 0.9677 | 0.8744 | 0.8851 | 0.9246 | 0.9209 | 0.9818 |
| gMLP | 0.9143 | 0.8859 | 0.9426 | 0.9391 | 0.9118 | 0.9142 | 0.9762 |

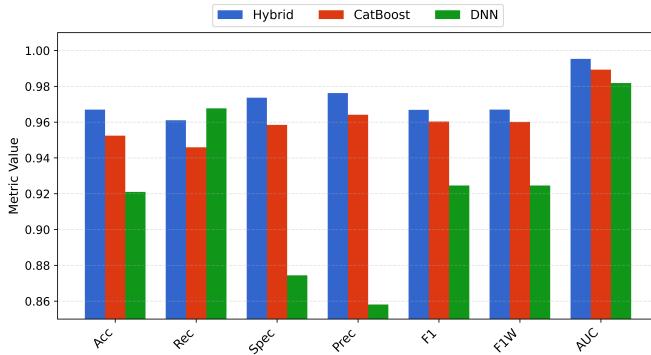


Fig. 11: Temporal test senaryosunda Hibrit, CatBoost ve DNN modellerine ait çubuk grafik karşılaştırılması.

TABLE XXVII: Hibrit Modelin Ek Hata Oranları ve Doğruluk Metrikleri (Temporal Test)

| Metrik | Açıklama | Değer (%) |
|--------|-----------------------|-----------|
| FPR | Yanlış Pozitif Oranı | 2.698 |
| FNR | Yanlış Negatif Oranı | 3.870 |
| NPV | Negatif Öngörü Değeri | 96.14 |
| FDR | Yanlış Keşif Oranı | 2.71 |

CatBoost ve ExtraTrees modelleri Precision ve Specificity açısından sınırlı avantaj göstermiştir; ancak bu farklar istatistiksel olarak ihmali edilebilir düzeydedir ($<1\%$). Daha önemlisi, Recall ve AUC-ROC gibi saldırı tespitinde kritik metriklerde hibrit modelin üstünlüğü, gerçek dünyada hatalı negatiflerin (yani tespit edilemeyen saldırıların) azaltılması açısından daha anlamlıdır. Bu nedenle, modelin öncelikli amacı olan saldırıların olabildiğince eksiksiz yakalanması bağlamında hibrit yapı daha avantajlıdır.

Klasik ağaç tabanlı modeller bazı alt metriklerde (ör. precision, specificity) sınırlı avantajlar gösterse de, hibrit mimari genel başarım, dengesizlik toleransı ve uzun dönemli kararlılık açısından daha tutarlı ve üstün sonuçlar üretmiştir. Bu durum, “derin özellik çıkarımıyla güçlendirilmiş hibrit yapıların, zaman içinde değişen saldırı örüntülerine karşı klasik yöntemlere kıyasla daha yüksek genelleme kabiliyeti sunduğu” yönündeki temel argümanı desteklemektedir.

Temporal analiz kapsamında derin öğrenme modelleri de değerlendirilmiş ve DNN ile gMLP'nin hibrit modele kıyasla belirgin şekilde daha düşük performans sergilediği gözlemlenmiştir. DNN modeli temporal teste %92.10 doğruluk ve %92.46 F1-score üretirken, benign sınıflarını ayırt etme gücünü temsil eden Specificity değerinde belirgin bir düşüş yaşamıştır (%87.44). Benzer şekilde gMLP modeli de %91.43 doğruluk ve %88.59 recall ile daha zayıf sonuçlar elde etmiştir. Bu bulgular, DL modellerinin zaman içinde değişen veri dağılımlarına karşı daha duyarlı olduğunu göstermektedir.

Hibrit modelde yer alan Stacked Autoencoder tabanlı özellik çıkarımı ve bunu takip eden LGBM sınıflandırıcı yapısı, verinin karmaşık dağılımlarını daha iyi temsil edebilmekte ve zamana bağlı dağılım kaymalarına (concept drift) karşı esnek bir adaptasyon sağlamaktadır. Böylece hibrit mimari, yalnızca anlık doğruluk başarımı değil, aynı zamanda uzun vadeli temporal dayanıklılık açısından da üstün bir performans sergilemektedir.

Elde edilen bulgular, hibrit yaklaşımın saldırı tespitinde sürdürülebilir performans sağlama potansiyeline sahip olduğunu ve geleneksel yöntemlere kıyasla daha ileri bir çözüm olarak konumlandırılmasını bilimsel açıdan haklı kıldığını göstermektedir.

Temporal Test Kümesinde Modellerin Confusion Matrix Karşılaştırması

Temporal test sonuçlarına ait confusion matrix görselleri incelediğinde (şekil 12 - şekil 16), hibrit SAE+LGBM mimarisinin diğer tüm modellere kıyasla çok daha dengeli ve tutarlı bir sınıflandırma performansı sergilediği açıkça görülmektedir. Hibrit model, Benign sınıfında 97.296 doğru sınıflandırma ve yalnızca 2.698 yanlış pozitif hatası üretmiş; Malicious sınıfında ise 96.100 doğru tespit ve 3.896 yanlış negatif sonucu elde etmiştir. Bu dağılım, hem benign hem saldırı sınıflarında hataların düşük ve dengeli olduğunu, yani modelin hem precision hem recall açısından güçlü bir denge kurduğunu göstermektedir. Özellikle yanlış negatif sayısının düşük olması, saldırı tespiti açısından kritik bir avantajdır.

CatBoost modelinin confusion matrix sonuçları incelediğinde, Benign = 96.982 doğru / 3.012 yanlış ve Malicious = 94.499 doğru / 5.497 yanlış tespiti yaptığı görülmektedir. Model genel olarak yüksek doğruluk sunsa da, hibrit modele göre hem false positive hem de false negative sayıları daha yüksektir. Özellikle saldırı sınıfındaki hata oranının (TP: 94.499 → hibritten 1.600 daha düşük) artması, temporal drift altında CatBoost'un genelleme kabiliyetinin sınırlı kaldığını göstermektedir.

ExtraTrees modelinin performansı daha da belirgin şekilde düşmektedir: Benign = 95.275 doğru / 4.719 yanlış, Malicious = 92.549 doğru / 7.447 yanlış değerleri, saldırı sınıfında çok daha fazla hatalı negatif ürettiğini göstermektedir. Bu durum, ExtraTrees'in temporal senaryoda dağılım kaymasına (drift) karşı daha zayıf tepki verdiği ve kötü amaçlı trafiği tespit etmede ciddi performans kayipları yaşadığını ortaya koymaktadır.

Derin öğrenme tabanlı gMLP modelinde drift etkisi daha da belirgindir. Model, Benign = 94.252 doğru / 5.742 yanlış ve Malicious = 88.594 doğru / 11.402 yanlış sonuçlarıyla hem false positive hem false negative oranlarında ciddi artış sergilemiştir. Özellikle saldırı sınıfında 11.402 yanlış negatif üretmesi, gMLP'nin temporal koşullarda ciddi şekilde zorlandığını göstermektedir. Bu değer, hybrid modele kıyasla neredeyse üç kat daha fazla saldırıyı kaçırdığı anlamına gelir.

DNN modelinin confusion matrix sonuçları, gMLP'den daha da olumsuzdur. Benign = 87.440 doğru / 12.554 yanlış ve Malicious = 96.767 doğru / 3.229 yanlış dağılımı, benign sınıfında hatalı pozitiflerin çok ciddi düzeyde arttığını göstermektedir. DNN, temporal teste benign trafiği ayırt etmede güçlük yaşamış ve yüksek false positive üretmiştir. Bu durum, modelin zaman-temelli veri kaymasına karşı oldukça duyarlı olduğunu doğrulamaktadır.

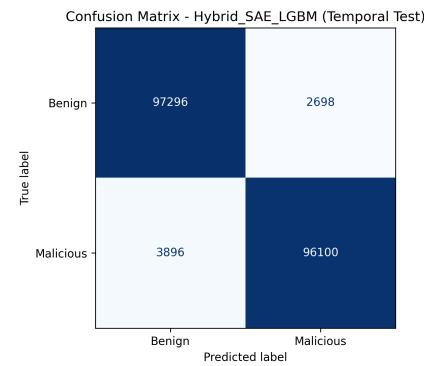


Fig. 12: Temporal Test kümesi üzerinde Hibrit Modelin confusion Matrix Sonucu

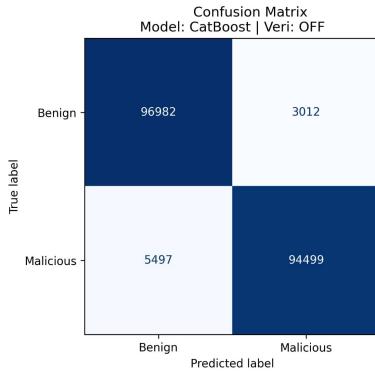


Fig. 13: Temporal Test kümesi üzerinde CatBoost Modelinin confusion Matrix Sonucu

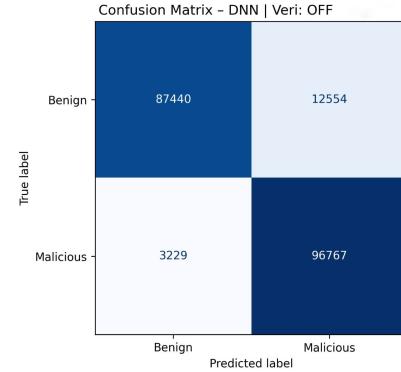


Fig. 16: Temporal Test kümesi üzerinde DNN Modelinin confusion Matrix Sonucu

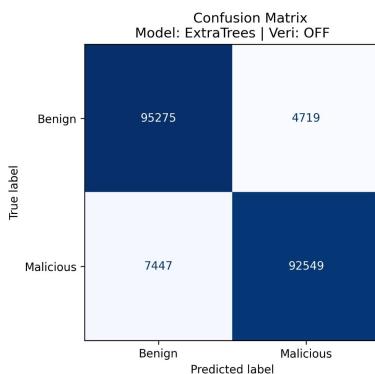


Fig. 14: Temporal Test kümesi üzerinde ExtraTrees Modelinin confusion Matrix Sonucu

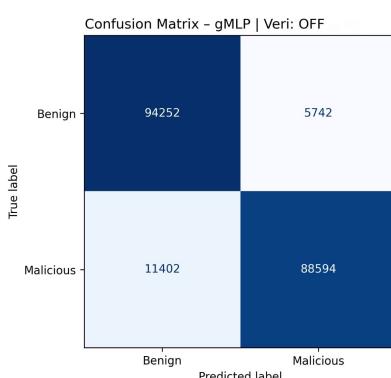


Fig. 15: Temporal Test kümesi üzerinde gMLP Modelinin confusion Matrix Sonucu

Temporal Analizde Hibrit Mimarının Model Boyutu ve Hesaplama Performansı Metrikleri

Temporal test senaryosunda hibrit modelin hesaplama performansı incelendiğinde, önerilen mimarının yalnızca 21.30 MB gibi oldukça kompakt bir model boyutuna ve yaklaşık 3.03 milyon eğitilebilir parametreye sahip olduğu görülmektedir. Model, 1.664 saniyelik toplam çıkışım süresi boyunca 199.990 örneği işleyebilmiş ve bunun sonucunda 0.008 ms/ornek gibi son derece düşük bir gecikme üretmiştir. Ayrıca 120.180 örnek/s throughput değeri, mimarının saniyede yüz binden fazla akış kaydını işleyebildiğini göstererek hibrit modelin gerçek zamanlı IDS sistemlerinde gerekli olan yüksek veri işleme kapasitesini fazlasıyla karşıladığı kanıtlamaktadır. Bu hesaplama metrikleri Tablo XXVIII’de özetlenmiştir.

Bu göstergeler bir arada değerlendirildiğinde, hibrit mimarının yalnızca yüksek doğruluk sağlayamadığı, aynı zamanda optimizasyona son derece uygun, hafif ve verimli bir model yapısı sunduğu açıkça görülmektedir. Düşük gecikme süresi, yüksek throughput değeri ve kompakt model boyutu sayesinde mimari, gerçek zamanlı ağ trafiği izleme, uç cihaz entegrasyonları ve kurumsal IDS dağıtımları gibi hesaplama maliyetinin kritik olduğu pratik senaryolarda minimum kaynak kullanımıyla maksimum performans sunabilmektedir. Düşük gecikme, yüksek throughput ve kompakt model boyutunun aynı anda sağlanması, önerilen yaklaşımın yalnızca doğruluk açısından değil; hesaplama maliyeti, hız ve gerçek zamanlı kullanılabilirlik bakımından da güçlü ve sürdürülebilir bir çözüm olduğunu açık biçimde ortaya koymaktadır. Ayrıca hibrit yapının yoğun hesaplama gerektirmeyen, optimize edilebilir ve donanım dostu mimarisini; GPU gereksinimi olmadan dahi yüksek performans sunabilmesi; ve düşük parametre yükü sayesinde kaynak kısıtlı cihazlarda bile kararlı çalışılabilmesi, modelin gerçek dünyadaki operasyonel verimliliğini daha da artırmaktadır. Bu yönleriyle hibrit model, hem akademik hem de pratik uygulamalarda ölçülebilir, hafif ve yüksek performanslı bir tehdit tespit çözümü sunmaktadır.

TABLE XXVIII: Hibrit Modelin Temporal Test Üzerindeki Hesaplama Performansı Metrikleri

| Metrik | Değer (Ortalama) | Açıklama |
|----------------|---------------------------------------|-------------------------|
| Model Boyutu | 21.30 MB ($\approx 3.03M$ parametre) | Toplam model boyutu |
| Inference Time | 1.664 s (n = 199,990) | Test çıkışım süresi |
| Latency | 0.008 ms/ornek | Ortalama gecikme süresi |
| Throughput | 120 180 örnek/s | İşleme hızı |
| Eğitim Süresi | 71.151 s | Toplam eğitim süresi |

Temporal Analizde Hibrit Modelin Sınıf Bazlı ROC Eğrisi

Şekil 17'de sunulan ROC eğrisi, hibrit SAE+LGBM modelinin temporal test kümesi üzerindeki ayırmayı açık biçimde ortaya koymaktadır. Eğri, düşük yanlış pozitif oranlarında dahi yüksek doğrulama oranı üreterek ideale yakın (sol üst köşeye yakın) bir davranış sergilemektedir. Elde edilen AUC-ROC değeri = 0.9954, hibrit modelin saldırısı (Malicious) sınıfını benign trafikten ayırma konusunda neredeyse mükemmel bir performans gösterdiğini doğrulamaktadır. Bu yüksek AUC değeri, modelin yalnızca belirli bir eşikte başarılı olmasının ötesinde, tüm olası karar eşikleri boyunca istikrarlı ve güvenilir bir ayırmayı yeteneğine sahip olduğunu göstermektedir. Temporal drift altında dahi ROC eğrisinin ideal şekilde son derece yakın olması, hibrit mimarinin zamanla değişen saldırısı dağılımlarına karşı güçlü genelleme kabiliyeti sunduğunu bir kez daha ortaya koymaktadır.

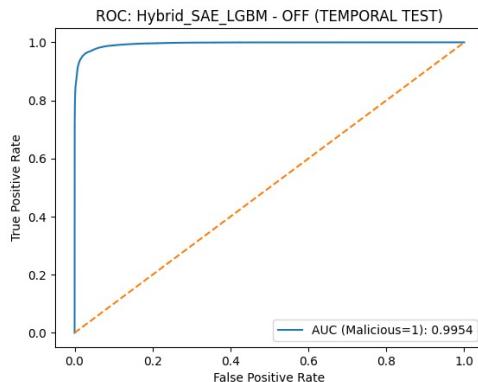


Fig. 17: Hibrit modelin Temporal Analizdeki ROC eğrisi

REFERENCES

- [1] Svante Wold, Michael Sjöström, and Lennart Eriksson. PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2):109–130, 2001.
- [2] Stephen Odaibo. Tutorial: Deriving the standard variational autoencoder (vae) loss function. *arXiv preprint arXiv:1907.08956*, 2019.
- [3] Tim Silhan, Stefan Oehmcke, and Oliver Kramer. Evolution of stacked autoencoders. In *2019 IEEE Congress on Evolutionary Computation (CEC)*, pages 823–830. IEEE, 2019.
- [4] David Durfee and Ryan M Rogers. Practical differentially private top-k selection with pay-what-you-get composition. *Advances in Neural Information Processing Systems*, 32, 2019.
- [5] Hualong Liao, Xinyuan Zhang, Can Zhao, Yu Chen, Xiaoxi Zeng, and Huafeng Li. Lightgbm: an efficient and accurate method for predicting pregnancy diseases. *Journal of Obstetrics and Gynaecology*, 42(4):620–629, 2022.
- [6] Payal Awwal and Smita Naval. Development of heuristic adapted serial-based deep learning for efficient adversarial malware detection framework in windows. 326:114032.