Zaman Serisi Verilerinde Anomali Tespiti için 5 Modelin Eğitilmesi ve Karşılaştırılması

Furkan YILMAZ & Eyup CINAR

Öz: Anomali tespiti, verilerden aykırı bileşenleri tespit etme ve çıkarma için yıllardır kullanılmaktadır. Anomali tespitinde birçok teknik kullanılmaktadır. Önemi her gün artan tekniklerden biri de bu alanda büyük bir rol oynayan makine öğrenmesidir. Bu çalışmada, Airbus Sas tarafından sağlanan ağır donanımlı bir helikopterin farklı bölgelerine yerleştirilen ivme ölçerlerden elde edilen zaman serisi verisi kullanılarak titreşim seviyelerinin analizi yapılmıştır. Veri, 1 boyutlu 1 dakikalık dizilere dönüştürülmüş ve bu formatta hazırlanmıştır. Beş farklı model (Isolation Forest, One-Class SVM, K-means, GMM ve LSTM) kullanılarak eğitim yapılmış ve veri setine üç farklı imputasyon yöntemi (Ortalama, K-NN ve FB-Fill) uygulanmıştır. Elde edilen sonuçlara göre, 1C-SVM ve LSTM modelleri model eğitimi sonucunda en yüksek F1 skoru olan 0.89'u elde etmiştir. Bu bulgular, helikopter titreşim seviyelerinin etkili bir şekilde modelleyebilen ve anomali tespiti konusunda başarılı sonuçlar veren bu iki modelin öne çıktığını göstermektedir. Çalışma, havacılık endüstrisinde güvenlik ve bakım optimizasyonu açısından önemli bir adım olan titreşim analizi konusunda değerli bilgiler sunmaktadır.

Anahtar kelimeler: anomali tespiti, zaman serisi verileri, makine öğrenmesi.

1. GIRIŞ

Anomali tespiti(ayrıca **aykırı değer** ve bazı yerlerde **yenilik tespiti** olarak adlandırılır) yıllardır üzerinde çalışılan bir konudur. Farklı uygulamalar üzerinde anomali tespiti için çeşitli şahsına münhasır metodlar geliştirilmiş ve kullanılmıştır, [1]. Anomali Tespiti "alışılmışın dışında davranan veri modellerini bulma sorunu" olarak tanımlanır , [2],[3].

Anomali tespitinin uygulama alanları şunlardır; izinsiz giriş tespiti, dolandırıcılık tespiti, arıza tespiti, sistem sağlığı izlenmesi, sensör ağlarındaki olay tespiti, ekosistem bozukluklarının tespiti ve görüntülerde kusur tespiti[5]. Bu araştırmada da sensör verileri ve zaman serisi şeklinde olan veriler üzerinde anomali tespiti yapan makaleler incelenmektedir. Zaman serisi verileri, düzenli zaman aralıkları ardışık olarak ölçülen verilerdir.

Bu araştırmanın kullandığı veri seti Airbus SAS tarafından toplanmış ve yayınlanmıştır. Ağır donanımlı hava araçlarının uçuş testlerindeki ana problem toplanan sinyal verilerinin teyidinin, verilerin çokluğu yüzünden zor olmasıdır. Elle teyit çok fazla zaman ve insan gücüne ihtiyaç duyduğu için bu işlemin otonom hale gelmesi kritiktir. Bu durumda, helikopterin tüm çalışma şartlarındaki titreşim seviyelerini ölçmek için, helikoptere farklı yönlerde (enlemesine, boylamasına, yanlamasına) ve farklı konumlarda ivmeölçerler konulmuştur. Araştırmada kullanılan veri seti farklı uçuşlardan toparlanan 1024 Hz frekansında 1 dakikalık dizilere bölünmüş 1B zaman serilerinden oluşmaktadır [4].

Bu araştırmanın amacı bu veri setini kullanarak anomali tespiti için dört adet makine öğrenmesi ve bir adet derin öğrenme modeli eğitmektir. Makine öğrenmesi modelleri şunlardır; **Isolation Forest**, **K-Means Cluster**, **Gaussian Mixture Model, One-Class SVM.** Derin öğrenme modeli ise **LSTM'**dir**.** Eğitilen modellerin sonuçları karşılaştırılıp anomali tespiti için en iyisi hangisi karar verilecektir.

2. SEMBOLLER, BIRIMLER, KISALTMALAR VE TANIMLAR

İsim	Açıklama
Fuzzy C-Mean Clustering (FCM)	Kümeleme problemlerinde kullanılan bir yöntemdir.
Least Squares Support Vector Machine (LS-SVM)	Regresyon veya sınıflandırma problemleri için kullanılan bir öğrenme modelidir.
Recurrent Neural Networks (RNN)	Zaman serileri ve sıralı verilerde etkili olan bir tür yapay sinir ağıdır.
Long Short-	RNN'lerdeki problemleri

Term Memomry (LSTM)	çözmek için tasarlanmış bir özel hücre tipidir ve uzun vadeli bağımlılıkları daha etkili bir şekilde öğrenir.
Gated Recurrent Unit (GRU)	LTSM'nin daha basitleştirilmiş bir versiyonudur ve benzer performansı sağlarken daha az hesaplama maliyetine sahiptir.
Support Vector Data Description (SVDD)	Sağlıklı örnek verilerin işlenerek tespit eşiği oluşturan sınıflandırıcıdır.
Gauss Karışım Modeli (GMM)	Bir sistemin normal davranışını modellemek ve bu normal davranıştan sapmaları belirlemek için kullanılabilen istatistiksel bir yöntemdir.
Holt-Winters	Zaman serileri analizi için kullanılan bir ön görüleme yöntemidir ve zaman içindeki trend, mevsimsellik ve seviye değişimlerini modellemek amacıyla üç temel bileşeni içerir.
Auto Regressive Integrated Moving Average (ARIMA)	Zaman serileri analizi ve ön görüleme için kullanılan istatistiksel bir modeldir.
Gradient Boosted Regression Trees (GBRT)	Ensemble öğrenme yöntemi olan Gradient Boosting'in bir türüdür ve regresyon problemlerinde güçlü tahminler yapmak için karar ağaçlarını birleştirir.
Long Short-Term Memory-Variational Autoencoder (LTSM-VAE)	Zaman serileri üzerinde değişkenlik analizi için kullanılan bir yapay sinir ağı modelidir. LSTM uzun vadeli bağımlılıkları yakalarken, VAE de değişkenlik analizi ve veri üretimi için kullanılır, böylece birleştirilmiş model zaman serileri verilerini etkili bir şekilde öğrenir ve analiz eder.
Z-Score	Bir veri noktasının ortalama değerden ne kadar uzakta olduğunu, ölçümün standart sapma birimleri cinsinden ifade eden istatistiksel bir ölçüdür.
K-NN Imputation	Eksik değerlere sahip bir gözlem biriminin eksik olan değerini, benzer diğer gözlem birimlerinin bilinen değerleri kullanılarak tahmin etmeye dayanan bir imputasyon yöntemidir.
Forward-Backward Fill	Birbirini takip eden zaman serisi gözlemlerindeki eksik değerleri sırasıyla önceki ve sonraki değerlerle dolduran imputasyon yöntemleridir.
Principal Component Analysis (PCA)	Çok boyutlu bir veri setindeki değişkenler arasındaki ilişkileri anlamak ve veri setini daha az sayıda, bağımsız değişkenle ifade etmek için kullanılan bir boyut indirgeme tekniğidir.

3. LİTERATÜR İNCELEMESI

2.1 Makale 1: “Anomaly Detection in Time Series Data using a Fuzzy C-Means Clustering”[5]

Bu makalede araştırmacılar **Fuzzy C-Means Clustering (**bulanık kümeleme**)** kullanarak zaman serisi verisinde anomali tespiti üzerinde çalışma yapmışlardır. Bu algoritma belirli bir veri kümesini belirli özelliklere göre gruplamak için kullanılır.

Araştırmacılar ilk olarak zaman serilerindeki anomalileri genlik ve şekil anomalisi olarak ikiye ayrıştırmaktadır. İki tür anomalininde tespitini sağlayacak bir birleşik çerçeve kullanmışlar. Bulanık kümeleme yöntemi kullanılarak zaman serilerindeki yapı ortaya çıkarılmış ve bir yeniden yapılandırma kriteri kullanılarak her bir alt diziye anomali skoru tayin edilmiştir. Genlik anomalilerin tespitinde zaman serilerinin orjinal durumları kullanılırken, şekil anomalilerinin tespitinde ise zaman serilerinin otokorelasyon gösterimleri kullanılmıştır.

Araştırmacılar sabit uzunlukta bir aralık penceresi ile bir alt dizi seti oluşturmuş ve bu alt diziler içerisindeki normal yapıyı göstermek için **FCM** algoritmasından yararlanmıştır. Her bir alt dizinin bir kümeye benzersizliğini ölçmek için yeniden yapılandırma kriteri kullanmışlar ve yeniden yapılandırma hatasını hesaplayarak anomali skoruna atamışlardır.

2.2 Makale 2: "Large-Scale Unusual Time Series Detection"[6]

Bu makalede araştırmacılar büyük ölçekte ki zaman serileri kullanarak anomali tespiti üzerine çalışma yapmışlardır. Bu çalışmada zaman verileri binlerce sunucunun saat başı toplanan performans ölçümlerinden oluşmaktadır. Ve araştırmacılar da olağan dışı davranan sunucuları tespit etmeyi hedeflemektedirler.

Araştırmacılar, her zaman serisi için karakteristik ölçümleri hesapladıktan sonra öznitelik vektörlerini çıkartmışlardır. Elde edilen öznitelik vektörleri üzerinde ana bileşen ayrışımı uygulayarak, çeşitli iki değişkenli anomali tespiti yöntemlerini kullanmışlardır. Bu yaklaşım, en olağan dışı zaman serilerini öznitelik vektörleri üzerinden belirlemelerini sağlamıştır. Yazarlar, kullandıkları yöntemlerin en yüksek yoğunluklu bölgelere ve α-hull'lara dayandığını ifade etmektedirler.

Araştırmacılar bu yazıda büyük miktardaki zaman serilerinde olağandışı zaman serilerini tespit etmek için Ana Bileşen Analizi (Principal Component Analysis) ile beraber çok-boyutlu anomali tespiti kullanmayı önermektedirler. Bu yaklaşımlarının 80% isabet oranına sahip olmakla beraber, 1000 zaman serisini işlemek için 0.5 saniyeden daha azına ihtiyaç duyduğunu belirterek bu yöntemlerinin güçlü ve isabetli olduğunu vurgulamaktadırlar.

2.3 Makale 3: "Anomaly Detection of Spacecraft Based on Least Squares Support Vector Machine"[7]

Bu yazı, yörüngede bulunan uzay araçlarında ortaya çıkan anomalileri tespit etmek için **Least Squares Support Vector Machine** (**LS-SVM)** algoritmasına dayanan bir yöntem sunmaktadır. Voltaj, sıcaklık, akım, titreşim gibi veriler toplanıp ön işleme tabii tutulduktan sonra öznitelikleri çıkartılıp Ana Bileşen Analizi kullanılarak öznitelik dizileri oluşturulmuş ve LS-SVM kullanılarak uzay aracındaki anormal davranışlar tespit edilmeye çalışılmıştır.

Bu çalışma, uzay aracındaki anormal durumları tespit etmek amacıyla Least Squares Support Vector Machine (LS-SVM) tabanlı bir yöntem önermektedir. PCA tabanlı öznitelik çıkarma ve LS-SVM kullanılarak gerçekleştirilen anormal durum tespiti sonuçları, yüksek doğruluk ve etkinlik göstermektedir. Ancak, yöntemin sınırlamaları ve PCA tabanlı öznitelik çıkarma yöntemlerinin geliştirilmesi ihtiyacı vardır.

2.4 Makale 4: "Probabilistic anomaly detection in natural gas time series data"[8]

Bu makale doğalgaz zaman serisi verileri için bir olasılıksal yaklaşım sunmaktadır. Bir dizi anomalinin sebepleri incelenip kategorize edilmiş ve Bayesian maksimum olabilirlik sınıflandırıcısına bilinen anomalilerin geçici yapısı öğretilmiştir. Bilinmeyen bir zaman serisi verisi sunulduğunda, sistem, hava durumu girdilerini kullanarak bir lineer regresyon modeli ile anormallikleri tespit etmektedir. . Daha sonra, anormallikler yanlış pozitiflerle test edilir ve Bayesian bir sınıflandırıcı kullanılarak sınıflandırılır. Bu yöntem aynı zamanda bilinmeyen kaynaklı anormallikleri tanımlayabilir. Bu nedenle, bir veri noktasının anormal olma olasılığı, hem bilinen hem de bilinmeyen kaynaklı anormallikler için sağlanmaktadır.

Bu yöntem günlük tüketim verilerine uygulandığında, veri temizleme yoluyla kök ortalama kare hatalar (RMSE'ler) için %37.5 ve ortalama mutlak yüzde hataları (MAPE'ler) için %7.84'ün üzerinde ortalama bir iyileşme sağlar. Çalışma, veri imputasyonu için güçlü tahmin modellerinin kullanılması ve Bayesian sınıflandırıcıyı geliştirmek için dış kaynaklı girdilerin eklenmesi gibi daha fazla iyileştirmeler önermektedir. Yöntem, zaman serisi verilerinde bilinen dış kaynaklı faktörlere sahip çeşitli alanlarda uygulanabilirliğini göstermektedir.

2.5 Makale 5: "Generic and Scalable Framework for Automated Time-series Anomaly Detection"[9]

Bu makalede Yahoo'nun kendi kullandığı, büyük ölçekli zaman serili veriler için açık kaynaklı bir otonom anomali tespit kütüphanesi tanıtılmaktadır. Bu kütüphane (**EGADS)** mevcut çoğu anomali tespit yaklaşımlarının büyük sorun yaşadığı ölçeklenebilirlik konusunda birden fazla tespit ve tahmin modelleri kullanarak bir çözüm ürettiğini belirtmektedir. Gerçek ve yapay veri kullanarak

kendi yaklaşımlarının diğer sistemlerden 50-60% oranında daha iyi olduğunu keşfetmişlerdir.

2.6 Makale 6: "ANOMALY DETECTION IN AIRCRAFT DATA USING RECURRENT NEURAL NETWORKS (RNN)"[10]

Bu makale, uçakların Uçuş Veri Kaydedici (FDR) veya Uçuş Operasyonel Kalite Güvence (FOQA) verilerinden toplanan çoklu değişkenli, zaman serisi verilerinde anomalilerin tespiti üzerine odaklanmaktadır. Endüstri standardı "Aşma Algılama" algoritması, belirli parametrelerin ve eşiklerinin listesini kullanarak bilinen sapmaları tanımlar. Buna karşılık, Makine Öğrenimi algoritmaları, verideki bilinmeyen olağandışı desenleri yarı gözetimli veya gözetimsiz öğrenme yoluyla tespit eder.

Makale, **Recurrent Neural Networks** (RNN)'in **Long Term Short Term Memory** (LTSM) ve **Gated Recurrent Units (**GRU**)** mimarileriyle uygulanışını tanımlar. RNN algoritmalarının hassasiyet=1, anımsama=0.818 ve F1 skoru=0.89 değerleriyle 11 anomaliden 9 tanesinin tespit edildiği belirtilmektedir.

2.7 Makale 7: "Hybrid robust convolutional autoencoder for unsupervised anomaly detection of machine tools under noises"[11]

Bu makale gürültülü ortamlarda makinelerden etiketli veri elde etmenin zorluğuna bir çözüm üretmek için gürültülü içerisinde çalışan makinelerin denetimsiz anomali tespiti için hibrit sağlam bir konvolüsyonel otokodlayıcı (HRCAE) geliştirilmesinden bahsetmektedir. Bu bahsedilen method bir Paralel Konvolüsyonel dağılım Uydurma (PCDF) modülü oluşturur ve bir . Birleştirilmiş Yönlü Mesafe (FDD) kayıp fonksiyonu tasarlayarak anomali tespitinin sağlamlığını artırır. Diğer denetimsiz otokodlayıcı yöntemleriyle karşılaştırıldığında bu metodun farklı gürültülerin arasında daha iyi tespit yapabildiği belirtilmiştir.

2.8 Makale 8: "Vibration-based anomaly detection using LSTM/SVM approaches"[12]

Bu çalışma, yarı-üstünlü anomali tespiti için yeni veri odaklı mimariler önermektedir. LSTM regresörleri ve tek sınıf SVM sınıflayıcılarını birleştiren bu mimariler, dişli aşınması ve rulman arızalarının titreşim sinyallerini hedeflemektedir. Standart bir model, dişli arızalarını tespit etmede etkili olsa da, rulman arızalarının neden olduğu zayıf değişiklikleri tespit etmekte zorlanır. Bu zorluğu aşmak için, sağlıklı sinyallerle eğitilmiş iki aşamalı bir model önerilmektedir.

Bu model, sinyalin sağlıklı bileşenlerini kaldıran birinci LSTM regresörü ve ardından artık sinyaldeki belirgin bileşenleri temizleyen ikinci bir LSTM regresörü içerir. Bu ekleme, rulman arızasının ilerlemesiyle sinyalin rastgele bileşenlerindeki bir artışı tanıma yeteneği sağlar. Sonuç olarak, sürekli olmayan veri kümeleri için ise, otoregresif bir LSTM eklemenin avantajı olmadığı ve basit istatistiksel özelliklere dayalı tek sınıf SVM aykırı değer tespitin daha etkili ve güvenilir olduğu gösterilmiştir.

2.9 Makale 9: "A Deep Support Vector Data Description Method for Anomaly Detection in Helicopters "[13]

Bu makalede helikopterlerde anomali tespiti için **Support Vector Data Description** (SVDD) yöntemi baz alınmıştır. Bu çalışmada da otonom bir anomali tespiti kurmak için derin bir SVDD modeli sunulmuştur. SVDD modeline uygun öznitelik çıkarmak içinde bir CNN uygulanmıştır.

Deneyler, bu çalışmanında kullandığı AirbusSAS tarafından sağlanan helikopter titreşim veri seti üzerinde yürütülmüş ve F1 skorunun 94%'e kadar ulaştığını göstererek yöntemin etkili olduğunu göstermiştir. Karşılaştırmalı analiz, önerilen yöntemin diğer giriş tiplerinden daha iyi performans gösterdiğini ortaya koymuştur.

2.10 Makale 10: "ANOMALY DETECTION AND CLASSIFICATION IN TIME SERIES WITH KERVOLUTIONAL NEURAL NETWORKS "[14]

Bu makale, kervolutional sinir ağlarının (KCNN'ler) zaman serisi verilerine uygulanan potansiyelini araştırıyor. KCNN'ler, konvolüsyonel sinir ağlarının (CNN'ler) uzun mesafeli ilişkileri yakalama yeteneklerini, CNN'lerin yerel özelliklerini yakalama yetenekleriyle birleştiren yeni bir tür koniyeldir.

Makale, KCNN'lerin CNN'lerden daha iyi performans gösterdiğini gösteren iki deney rapor ediyor. İlk deneyde, KCNN'ler zaman serilerinde bir sınıflandırma görevinde test edilmiş ve CNN'lerden

daha yüksek doğruluk elde etmişlerdir. İkinci deneyde ise KCNN'ler helikopterlerde ivme ölçerler tarafından kaydedilen zaman serisi verilerinde anomalileri tespit etmek için test edilmiş ve CNN'lerden daha iyi performans göstermişlerdir. Bu sonuçlar, KCNN'lerin zaman serisi verilerinde arıza tespiti ve sınıflandırma için etkili bir araç olabileceğini göstermektedir..

2.11 Makale 11: "Exemplar Learning for Extremely Efficient Anomaly Detection in Real-Valued Time Series "[15]

Bu makale, gerçek değerli zaman serilerindeki anomalileri tespit etmek için yeni bir algoritma sunmaktadır. Algoritma, bir eğitim zaman serisinde bulunan çeşitli alt dizileri temsil etmek için küçük bir örnek seçimine dayanır. Örnekler, benzer alt dizilerde bulunan karakteristik yörüngeyi (düşük frekans bilgisi) ve stokastik özelliklerini (yüksek frekans bilgisi) yakalayan İstatistiksel ve Pürüzsüz Yörünge özelliklerini kullanır. Makale, yeni algoritmanın önceki algoritmalardan çok daha hızlı ve daha doğru olduğunu göstermektedir.

2.12 Makale 12: "Detecting Anomalies in a Time Series Database"[16]

Bu makalede yazarlar yarı denetimli anomali tespit tekniklerini dair yaptıkları değerlendirmeleri sunmuşlardır. Bu değerlendirmenin sonunda şu sonuçlara varmışlardır;

- Sürekli zaman serilerinde çalışan teknikler, genellikle ayrık diziler üzerinde çalışan tekniklerden üstündür.
- Model bağımsız olan çekirdek ve pencere tabanlı teknikler, zaman serisi verileri için bir model oluşturmaya çalışan tahmini ve segmentasyon tabanlı tekniklerden daha iyi performans gösterme eğilimindedir.
- Çekirdek tabanlı teknikler, pencere tabanlı tekniklere göre daha hızlıdır.
- Normal ve anormal verilerin doğası ile farklı tekniklerin performansı arasında birkaç ilişki vardır.

2.13 Makale 13: "Anomaly Detection Method of Aircraft System using Multivariate Time Series Clustering and ClassificationTechniques"[17]

Bu makalede uçak sistemlerinde sapma tespiti için yeni bir yöntem önerilmiştir. Yöntem, uçuş evreleri içindeki değişkenlerin döngüselliğini dikkate alarak normal ve anormal davranışları ayırt eder. Yöntem, DBSCAN ve DTW algoritmalarını kullanarak yüksek doğrulukla çalışır.

Önerilen yöntemde, bu anormal ve normal davranışlar karşılaştırıldığında, yeni bir önem puanı kullanılarak açıklanır. Davranışlar, Zaman Serisi Ormanı (TSF) algoritması ve siluet kriteri kullanılarak belirlenir. Metod, Bombardier'in Uçak Sağlık İzleme Sisteminden bir örnek kullanılarak eğitilmiş ve test edilmiştir. Normal ve anormal davranışları kümelenme siluet puanı 0,95 olacak şekilde ayırt eder ve bilinmeyen davranışları %89 hassasiyetle tespit eder.

2.14 Makale 14: "Anomaly Detection Method of Aircraft System using Multivariate Time Series Clustering and ClassificationTechniques"[18]

Bu makalede gerçek dünyadaki verilerde anomaliler seyrek olarak gerçekleştiğinden, çok değişkenli zaman serilerinde sapma tespitinin zorluğu ele alınmıştır. Yazarlar bir seferde yalnızca bir özellik kümesini dikkate alan bir özellik toplama tekniği önermektedir. Bu, anomalilerin daha küçük bir özellik kümesinde daha olası olduğundan, anomali tespiti için daha etkilidir.

Ek olarak, yaklaşımın etkinliğini ve genelleştirilmesini iyileştirmek için, yazarlar PCA ile hesaplanan iç içe geçmiş rotasyona dayalı bir dönüşüm uygularlar. Bu, anomalileri daha iyi ayırt etmeye yardımcı olur. Son olarak, temel modellerin çıktılarını birleştirmek için, yazarlar bir Lojistik Regresör kullanan yarı denetimli bir yaklaşım önerirler. Bu, tahmin performansını daha da artırır. Genel olarak, önerilen yöntem, çok değişkenli zaman serilerinde sapma tespiti için etkili bir yaklaşımdır.

2.15 Makale 15: "Operational Anomaly Detection in Flight Data Using a Multivariate Gaussian Mixture Model"[19]

Bu makale, helikopter uçuş verilerinde sapmaları tespit etmek için GMM'leri kullanma yöntemini önermektedir. Önerilen yöntem, öncelikle veri kümesi içindeki özelliklerin sayısını azaltmak için bir boyutsallık azaltma tekniği kullanır. GMM daha sonra azaltılmış boyutlu veri kümesi üzerinde eğitilir. Eğitilmiş GMM daha sonra yeni veri noktalarını puanlamak için kullanılır ve düşük puan alan veri noktaları anomali olarak işaretlenir.

Önerilen yöntem, Maryland Üniversitesi'nden helikopter uçuş verisi kümesi üzerinde değerlendirilmiştir. Sonuçlar, yöntemin yüksek doğrulukla anomalileri tespit edebildiğini göstermektedir. Yöntem ayrıca hesaplama açısından verimlidir, bu da onu gerçek zamanlı uygulamalar için uygun hale getirir. Genel olarak, bu metodun helikopter uçuş verilerinde sapma tespiti için umut verici bir yaklaşım olduğu belirtilmektedir. Helikopter operasyonlarının güvenliğini ve güvenilirliğini iyileştirme potansiyeline sahiptir.

2.16 Makale 16: "Hybrid approach for Anomaly Detection in Time Series Data"[20]

Bu makalede, zaman serisi verilerinde anomali tespiti için bir hibrit yaklaşım önerilmektedir. Önerilen yaklaşım, LSTM Autoencoder ve doğrusal sınıflandırıcıyı birleştirir. LSTM Autoencoder, zaman serisinden verimli temsiller öğrenir. Bu temsiller, doğrusal sınıflandırıcı tarafından anomalileri tespit etmek için kullanılır.

Deneyler, gerçek dünya veri kümelerinde yapılan karşılaştırmalara göre, önerilen hibrit yaklaşımın mevcut en iyi yöntemlerden daha iyi performans gösterdiğini göstermektedir.

2.17 Makale 17: "A Method to Handle Unstable Time Series in Anomaly Detection Problem"[21]

Bu makalede, zaman serisi varyasyonunu azaltmak için bir yöntem önerilmektedir. Bu yöntem, zaman serisinden kararsız parçaları otomatik olarak arayan ve hariç tutan bir algoritmadan oluşur. Algoritma, ham veri örneklerinin toplanmasından makine öğrenimi modelleriyle gelecekteki çalışmalar için vektörlemelerine kadar adımları kapsayan bir ön işleme algoritmasının parçası olarak uygulanmaktadır.

Önerilen yöntem, gaz sensörlerinden gelen zaman serilerinde test edilmiştir. Bu zaman serileri, sensörün hassas yüzeyinin çevrede bulunan çeşitli gazlara karşı reaksiyonunu yansıtmaktadır. Deneyler, yöntemin mevcut en iyi yöntemlerden daha iyi performans gösterdiğini göstermektedir.

2.18 Makale 18: "Anomaly Detection for Key Performance Indicators Through Machine Learning"[22]

Bu makale ağ hizmetlerinin stabil olup olmadığını anlamak için çeşitli kilit performans göstergeleri üzerinden anomali tespiti üzerinde çalışma yapmıştır. Sorun iki kısma ayrılmıştır. İlk kısımda zaman serisinde bir sonraki noktayı tahmin etmek için zaman serisi analiz metodu olan **Holt-Winters**, **ARIMA** modeli, regresyon tabanlı **GBRT** ve **LTSM** teknikleri kullanılmıştır. Sonrasında ise anomali kuralı belirlenmektedir. Ve tahmin edilen değer ile asıl değer karşılaştırılarak asıl değerin anomali olup olmadığı tespit edilir.

Bahsedilen metodlar arasında en iyi performansa sahip olan tekniğin, en düşük **MSE**, en yüksek F1 skoru ve en kısa eğitim süresine sahip olan **GBRT** olduğu belirtilmektedir.

2.19 Makale 19: "A Comparative Study of Cluster Based Outlier Detection, Distance Based Outlier Detection and Density Based Outlier Detection Techniques "[23]

Bu makalede yazarlar kümeleme bazlı, uzaklık bazlı ve yoğunluk bazlı anomali tespit teknikleri üzerine bir karşılaştırma araştırması yapmaktadırlar. Araştırmacılar bu tekniklerin performanslarını ölçmek için farklı türde gürültü ve anomali içeren çeşitli veri setleri kullanmaktadırlar. Her bir tekniğin kendine has parametreleri olduğundan direkt bir karşılaştırmanın mümkün olmadığı sonucunu çıkarmışlardır.

2.20 Makale 20: "Hybrid Machine Learning for Anomaly Detection in Industrial Time-Series Measurement Data"[24]

Bu makale, çoklu değişkenli zaman serisi ölçüm verilerinden oluşan büyük veri kümelerinde anomalilerin tespiti için yeni bir yöntem önermektedir. Yöntem, anahtar performans göstergeleri (KPI'lar) ve uzun kısa süreli bellek (LSTM-VAE) varyasyonel otoenkoder adlı iki farklı makine öğrenimi tekniğini birleştirir. KPI'lar, sistem davranışını tahmin etmek için kullanılır. **LSTM-VAE** ise verilerin istatistiksel belirsizliğini dikkate alarak sistemin normal davranışını modeller. Bu iki yöntemin kombinasyonu, verilerde anomalileri daha güvenilir bir şekilde tespit etmeyi sağlar.

Yöntem, bina temelleri için bir zemin iyileştirme işlemiyle ilgili bir vaka çalışması ile doğrulanır. Sonuçlar, yöntemin verilerde anomalileri tespit etmede etkili olduğunu göstermektedir.

3. YÖNTEM

Helikopter verilerinde anomali tespiti için beş adet model eğitilmiştir. Modeller eğitilmeden önce veri ön işlemeye tâbi tutulmuştur.

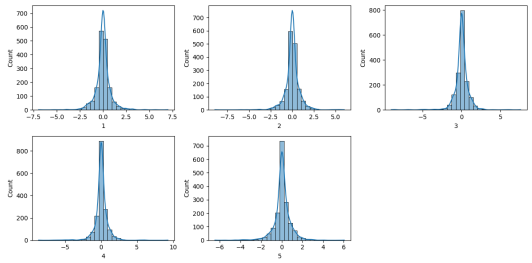
3.1 Ön İşleme

Eğitim veri seti doğası gereği yüksek boyutludur. Eğitim verisi 1677 satır 61440 kolondan, test verisi ise 594 satır 61440 kolondan oluşmaktadır. Ayrıca test verisi için 594 satır 2 kolonluk etiket verisi bulunmaktadır. Etiketler için 0 normali, 1 anomaliyi ifade etmektedir.

Tablo 1. Veri setinin Tanımı

Eğitim Seti	*Test Seti*
RangeIndex: 1677 entries, 0 to 1676	*RangeIndex: 594 entries, 0 to 593*
Columns: 61440 entries, 0 to 61439	*Columns: 61440 entries, 0 to 61439*
dtypes: float64(61440)	*dtypes: float64(61440)*

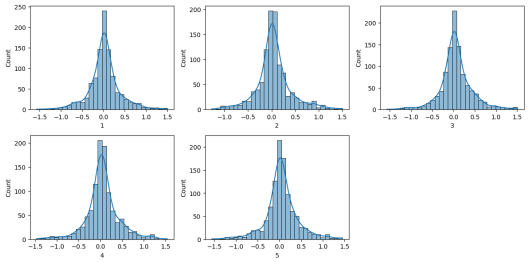
3.1.1 Aykırı Değerleri Silme



Şekil 1. Eğitim verisinin düzensiz hali

Şekil 1'de eğitim verisinin ön işlemeye tâbi tutulmadan önceki hali sunulmuştur. Bu grafikte ilk 5 öğenin dağılımını gösterilmektedir. Grafikte değerlerin dağılımının oldukça düzensiz olduğu görülmektedir. Ve -7.5 ile 7.5 arasında değişmektedir. Ayrıca değerler arasındaki korelasyon da oldukça zayıftır. Bu bilgiler doğrultusunda verinin heterojen olduğu çıkarımı yapılabilir.

Veri setinde mevcut bulunan gürültüyü ve düzensizliği ortadan kaldırmak için verilerin Z-Skoru hesaplanmıştır. Bu skordan büyük olan veriler düzensizliğe sebep olmaktadır ve modellerin iyi bir şekilde eğitilmesi için veri setinden çıkartılmıştır.



Şekil 2. Eğitim verisinin düzenli hali

Z-Skoru 3'ten büyük olan veriler silindikten sonra Şekil 2'de görüldüğü üzere veri seti daha düzenli ve homojen hale gelmiştir. Satır sayısı 1038'e düşmüştür.

3.1.2 Eksik değerlerin doldurulması

Veri önişlemenin önemli adımlarından biri de eksik değerlerin imputasyonudur. Bu çalışma da üç adet imputasyon yöntemi kullanılmıştır. İlk olarak Ortalama İmputasyonu kullanılmıştır. Bu metodla eksik değerler, değişkenin ortalama değeri ile doldurulmuştur. Tablo 2 ve 3'te ilk 5 sütun için ortalama imputasyondan öncesi ve sonrası paylaşılmıştır.

	0	*1*	*2*	*3*	*4*
count	*1038*	*1038*	*1038*	*1038*	*1038*
mean	*0.0034*	*0.0224*	*0.0465*	*0.0577*	*0.0548*
std	*0.3618*	*0.3641*	*0.3676*	*0.3715*	*0.376*
min	*-1.553*	*-1.568*	*-1.273*	*-1.574*	*-1.47*
25%	*-0.147*	*-0.121*	*-0.101*	*-0.11*	*-0.0968*
50%	*0.017*	*0.0187*	*0.0252*	*0.0362*	*0.036*
75%	*0.159*	*0.1645*	*0.1824*	*0.2073*	*0.1969*
max	*1.407*	*1.4994*	*1.5049*	*1.5083*	*1.5432*

Tablo 2. Ortalama İmputasyonu öncesi

	0	*1*	*2*	*3*	*4*
count	*1038*	*1038*	*1038*	*1038*	*1038*
mean	*0.003397*	*0.022433*	*0.046546*	*0.057736*	*0.054794*
std	*0.361824*	*0.364087*	*0.367609*	*0.371529*	*0.375983*
min	*-1.553164*	*-1.567793*	*-1.272794*	*-1.574047*	*-1.470111*
25%	*-0.1479*	*-0.121382*	*-0.100552*	*-0.110418*	*-0.096798*
50%	*0.016996*	*0.018709*	*0.025158*	*0.036214*	*0.035987*
75%	*0.159156*	*0.164466*	*0.182395*	*0.207288*	*0.196914*
max	*1.406972*	*1.499392*	*1.504855*	*1.508307*	*1.54317*

Tablo 3. Ortalama İmputasyonu sonrası

Yukarıdaki tablolarda sütun isimleri şöyle açıklanır:

- **count:** satır sayısı,
- **mean:** her sütundaki değerlerin ortalaması,
- **std:** her sütundaki değerin standart sapması,
- **min:** en küçük değer,
- **25%**: her sütundaki değerlerin alt çeyrek yüzdeliği,
- **50%:** her sütundaki değerlerin medyanı,

- **75:** her sütundaki değerlerin üst çeyrek yüzdeliği,
- **max:** en büyük değer.

Tablo 2 ve 3 incelendiğinde ortalama imputasyonunda kayda değer bir değişiklik olmadığı görülmektedir. Bu da ortalama imputasyonunu veri setine etkisinin sınırlı olduğu anlamına gelmektedir.

İkinci yöntem olarak *n_neighbors = 5* parametresiyle **K-NN** imputasyonu kullanılmıştır. Bu imputasyon sonrası sonuçlar Tablo 4'te paylaşılmıştır.

	0	1	2	3	4
count	1038	1038	1038	1038	1038
mean	0.003397	0.003397	0.046546	0.057736	0.054794
std	0.361824	0.361824	0.367609	0.371529	0.375983
min	-1.553164	-1.553164	-1.272794	-1.574047	-1.470111
25%	-0.1479	-0.1479	-0.100552	-0.110418	-0.096798
50%	0.016996	0.016996	0.025158	0.036214	0.035987
75%	0.159156	0.159156	0.182395	0.207288	0.196914
max	1.406972	1.406972	1.504855	1.508307	1.54317

Tablo 4. K-NN İmputasyonu sonrası

	0	1	2	3	4
count	1038	1038	1038	1038	1038
mean	0.003397	0.022433	0.046546	0.057736	0.054794
std	0.361824	0.364087	0.367609	0.371529	0.375983
min	-1.553164	-1.567793	-1.272794	-1.574047	-1.470111
25%	-0.1479	-0.121382	-0.100552	-0.110418	-0.096798
50%	0.016996	0.018709	0.025158	0.036214	0.035987
75%	0.159156	0.164466	0.182395	0.207288	0.196914
max	1.406972	1.499392	1.504855	1.508307	1.54317

Tablo 5. F-B Fill İmputasyonu sonrası

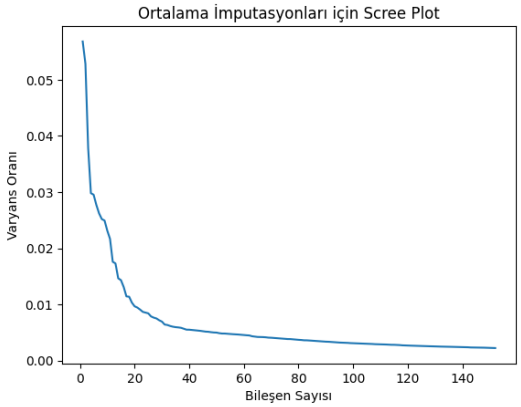
Tablo 2, Tablo 4 ve Tablo 5 karşılaştırıldığında ilk 5 sütun için yine büyük bir değişiklik söz konusu değildir. Bu durum, imputasyon yöntemlerinin belirli bir istatistiksel özet üzerinde benzer etkilere sahip olduğunu düşündürebilir. Ancak, bu değerlerin benzer olması, imputasyon yöntemlerinin tümüyle aynı sonuçları ürettiği anlamına gelmez.

3.1.2 Öznitelik Çıkarımı

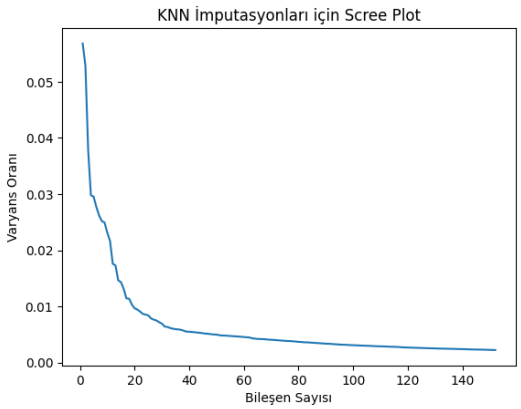
Makine öğrenmesinde öznitelik, bir fenomenin ölçülebilir özelliği ya da karakteristiğidir [25]. Bu veri setinde yapısı gereği yüksek boyutta öznitelik bulunmaktadır. Aşırı öznitelik sayısı, modelin genelleme performansını düşürebilir ve aşırı öğrenmeye neden olabilir. İdeal öznitelik sayısı, hem yeterli bilgi sağlamalı hem de modelin karmaşıklığını kontrol altında tutmalıdır. Buna istinaden tek tek öznitelik çıkarılması bu veri seti için mümkün olmadığından yaygın bir öznitelik çıkarma metodu olan **PCA** kullanılmıştır.

PCA'nın uygulanması için özniteliklerden oransal olarak ne kadar çıkarım yapılacağı belirtilmelidir. Burada dikkat edilmesi gereken nokta az öznitelikle verisetinin çoğunluğunu örnekleyecek şekilde bir oran seçilmesidir. Veri setinin ne durumda olduğunu görmek için ilk olarak standart bir değer olan *n_components=0.7* parametresiyle PCA uygulanmıştır. PCA uygulanmadan önce eğitim ve test setleri *(1038,61440)* ve *(594,61440)* şekline sahiplerdir.

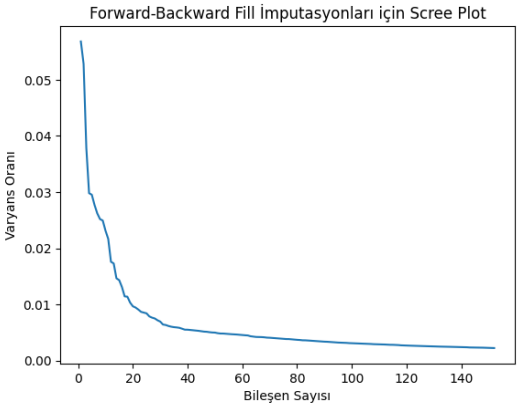
PCA uygulandıktan sonra ise eğitim ve test setleri *(1038,152)* ve *(594,152)* şekillerine sahip olmuşlardır. Ayrıca PCA'nın uygulanması için veri setlerine Min-Max yöntemiyle normalizasyon uygulanmıştır.



Şekil 3.



Şekil 4.

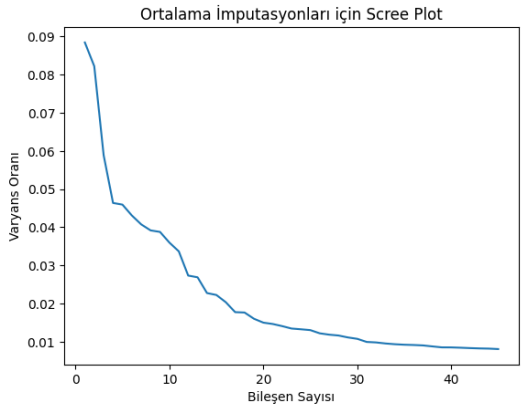


Şekil 5.

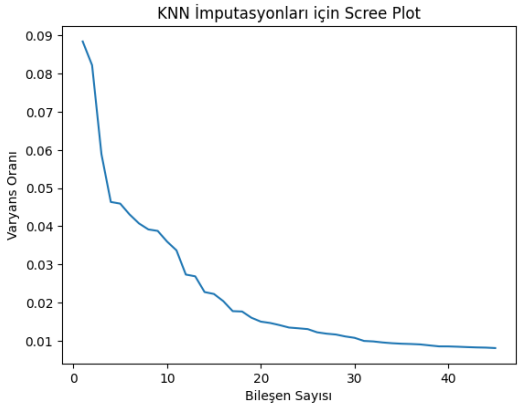
Şekil 3-5'te görülen grafiklerde eksenlerin açıklaması aşağıda verilmiştir;

- **bileşen:** Analiz sonrasında elde edilen bileşen sayısını ifade eder. PCA, veri setindeki değişken sayısını azaltmak için kullanılır, ve bu bileşen sayısı analizin sonucunda belirlenir.
- **varyans:** Her bir bileşenin toplam varyans içindeki oranını gösterir. Yüksek varyans açıklama oranına sahip olan bileşenler, orijinal veri setindeki önemli bilgileri korur.

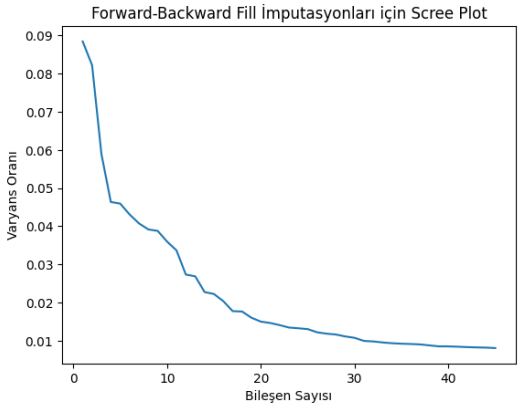
Şekil 3-5'te görüldüğü gibi varyans oranı 40. bileşenden sonra azalmaya başlamıştır. Bu ilk 40 bileşenin sahip olduğu bilgi oranının, veri setinin genelini kapsadığı anlamına gelmektedir. Çok düşük varyans oranına sahip bileşenleri modele beslemek aşırı uyum riskini ortaya çıkarır. Bu yüzden PCA *0.45* bileşen oranı ile tekrar uygulanmıştır.



Şekil 6.



Şekil 7.



Şekil 8.

Şekil 6-8'te görüldüğü üzere bileşen oranı *0.45* yapıldığında Şekil 3-5'te görülen düşük varyanslı bileşenler eğitim ve test setlerinden çıkarılmıştır. Öznitelik çıkarımı yapılarak eğitim seti model eğitmeye uygun hale gelmiştir.

3.2 Model Eğitimi

Bu çalışmada dört adet makine öğrenmesi modeli ve bir adet derin öğrenme modeli, üç farklı şekilde impute edilmiş verilerle eğitilmiştir. Bu modellerin karşılaştırılmasında **Kesinlik (A)**, **Hassasiyet (P), Duyarlılık (R)** ve **F1 Skoru** metrikleri kullanılmıştır. Ayrıca derin öğrenme modeli için **Receiver Operator Characteristics (ROC)** eğrisi de kullanılmıştır.

3.2.1 Isolation Forest

Isolation Forest, ilk olarak 2008'de Fei Tony Liu tarafından anomali tespiti için geliştirilmiştir [26]. Bu algoritma anomali tespitini ikili ağaçlar kullanarak tespit etmektedir [27][28].

Bu model eğitilirken yapılan ince ayar sonucunda en iyi hiper parametreler şu şekilde açıklanmıştır;

- **n_estimators=200:** Toplamda oluşturulan izolasyon ağaçlarının sayısını belirler.
- **max_samples=0.2:** Her bir izolasyon ağacının oluşturulurken kullanılacak örnek sayısını belirler.
- **contamination=0.175:** Veri kümesindeki anomali oranını belirler. Bu oran, izolasyon ağaçları tarafından izole edilmiş örneklerin oranını ifade eder.
- **random_state=42:**Modelin tekrarlanabilir olmasını aynı parametrelerle çalıştırıldığında aynı sonuçların elde edilmesini sağlar.

3.2.2 One-Class SVM

Bu algoritmanın çalışma prensibi, önce eğitim verilerinde yalnızca normal sınıfa ait olan örnekleri kullanarak bir model oluşturmasıdır. Bu model, normal sınıfa ait örnekleri içeren bir bölge veya hacim tanımlar. Daha sonra, bu bölge dışında kalan noktaları aykırı değer olarak kabul eder.

Bu bölgeyi oluşturmak için seçilen yapılan ince ayarda en iyi parametreler şunlar seçilmiştir;

- **nu=0.11:** Modelin eğitimi sırasında hata toleransını kontrol eder.
- **tol=0.008:** Eğitim süreci bu tolerans eşiği altına düştüğünde durur.
- **kernel=rbf:** SVM'in veri kümesindeki örnekleri daha yüksek boyutlu bir özellik uzayına haritalamak için kullanılan bir matematiksel işlemdir. (Radial Basis Function)
- **gamma_value=scale:** RBF çekirdeğinin genişliğini kontrol eden bir parametredir. "scale" seçeneği, gamma'nın 1 / (n_feature*X.var()) şeklinde ölçeklendiği anlamına gelir.

3.2.3 K-Means Cluster

Bu algoritma, veri noktalarını belirli sayıda küme içinde gruplandırmak için kullanılan bir kümeleme algoritmasıdır. Temel olarak, veri noktalarını birbirine benzerliklerine göre gruplandırmaya çalışır. Belirli sayıda küme merkezi rastgele seçilir. Her veri noktası en yakın küme merkezine atanır. Küme merkezleri, kendi içindeki veri noktalarının ortalaması olarak güncellenir. Atama ve güncelleme işlemleri, küme merkezleri değişmeyene veya belirli bir iterasyon sayısına ulaşana kadar tekrarlanır [29].

Bu algoritma ile eğitilen modelin parametreleri şu şekilde açıklanmıştır;

- **n_cluster=6:** Küme sayısını ifade eder.
- **n_init=8:** Bu algoritma rastgele başlangıç merkezleri seçer ve her bir başlangıçta algoritmayı çalıştırır. *n_init* parametresi, bu rastgele başlangıçların kaç kez deneneceğini belirtir ve en iyi sonucu seçer.
- **algorithm=lloyd:** K-Means algoritmasının kullanacağı algoritmayı belirler. 'lloyd', K-Means'in klasik Lloyd's algoritması olduğunu belirtir.
- **threshold=99:** Anomali tespiti için eşik değerini ifade eder.

3.2.4 Gaussian Mixture Model

Gaussian Mixture Model (GMM), istatistiksel bir modelleme tekniğidir ve veri kümesini birden fazla Gauss (normal) dağılımının bir kombinasyonu olarak temsil eder. Her bir Gauss bileşeni, veri noktalarının belirli bir kısmını temsil eder ve ağırlıkları, merkezi konumları ve kovaryans matrisleri kullanılarak tanımlanır. GMM, esnek bir olasılık dağılımı sağlar ve karmaşık veri yapılarını modellemek, veri setindeki farklı grupları tanımlamak için kullanılır [30]. Bu modelin eğitilmesinde kullanılan parametreler şunlardır:

- **n_components=110:** bileşen sayısı,
- **covariance_type=full:** kovaryans tipi,
- **threshold=40:** eşik değeri.

3.2.5 LSTM

Long Short-Term Memory (LSTM), özellikle zaman serisi verileri gibi uzun vadeli bağımlılıkları işlemek için tasarlanmış bir tür rekürrent sinir ağı (RNN) yapısıdır. LSTM, geleneksel RNN'lerin karşılaştığı uzun vadeli bağımlılık sorununu çözmek için özel hücre yapıları kullanır. Bu hücreler, giriş verisinin anlık değeri yanı sıra önceki zaman adımlarından gelen bilgileri koruyarak ve güncelleyerek çalışırlar. Bu özellikleri sayesinde LSTM, uzun vadeli bağımlılıkları etkili bir şekilde modelleyebilir ve bu nedenle özellikle zaman serisi analizi, doğal dil işleme ve benzeri uygulamalarda tercih edilir [31].

Bu model, iki LSTM katmanı, 2 *Dropout* katmanları, *RepeatVector* katmanı ve bir *Dense*

katmanından oluşmaktadır. Model, *Adam* optimizasyonu ve ortalama karesel hata kaybı ile derlenir. LSTM katmanlarının birinci katmanındaki aktivasyon fonksiyonu *sigmoid*'dir, ve *kernel*, *recurrent* ve *bias* regularizasyonları L2 normu kullanılarak *0.003* oranında uygulanır. Öğrenme oranı *0.005* olarak belirlenmiştir. *Dropout* oranları ise sırasıyla *0.2*'dir. *Dropout* katmanıyla rastgele nöronların oluşması engellenip aşırı uyum azaltılmaktadır. *RepeatVector* katmanında , ikinci LSTM katmanının çıkışını, orijinal zaman serisi boyutuna geri döndürmek için kullanılır. Son olarak, *Dense* katmanı, modelin çıkışını oluşturur. Bu katmanın çıkış boyutu, modelin öğrenmeye çalıştığı veri setinin özellik sayısına eşittir.

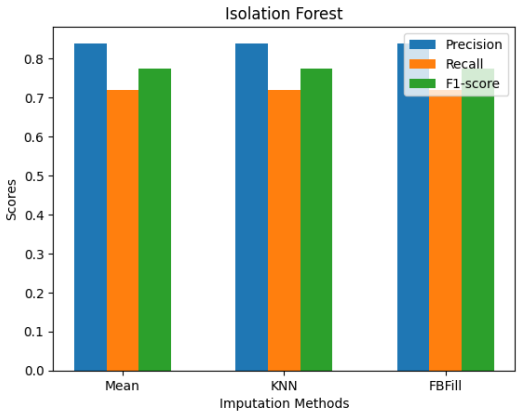
4. BULGULAR VE ANALİZ

Bu başlıkta eğitilen modellerden elde edilen bulgular ve bunların analizi anlatılmıştır. Önce her bir modelin çıktısı verilmiş olup altına da yapılan analiz yazılmıştır.

4.1 Isolation Forest

Tablo 4. IF Sonuçları

	Etiket	*P*	*R*	*F1*	*A*
Ortalama	*0*	*0.76*	*0.86*	*0.81*	*0.79*
	1	*0.84*	*0.72*	*0.78*	
K-NN	*0*	*0.76*	*0.86*	*0.81*	*0.79*
	1	*0.84*	*0.72*	*0.78*	
FB-Fill	*0*	*0.76*	*0.86*	*0.81*	*0.79*
	1	*0.84*	*0.72*	*0.78*	



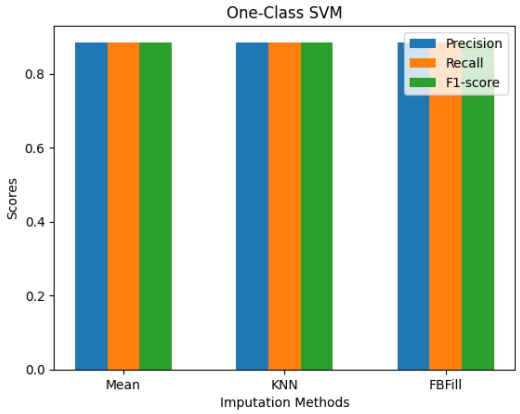
Şekil 9. IF Sonuçları

Tablo 4 ve Şekil 9'dan yola çıkarak bu modelin imputasyon yöntemlerine göre fark göstermeden aynı performansı gösterdiği çıkarılabilir. Model *%84 P* elde ederek, etiketlediği anomali sınıfındaki örneklerin %84'ünün gerçekten anomali olduğunu doğrulamıştır. Ayrıca, *%72 R* elde ederek, gerçek anormal örneklerin %72'sini doğru bir şekilde tespit etmiştir. Bu değerlerden yola çıkarak *%78 F1 Skoru* hesaplanmıştır.

4.2 One-Class SVM

Tablo 5. 1-C SVM Sonuçları

	Etiket	*P*	*R*	*F1*	*A*
Ortalama	*0*	*0.89*	*0.89*	*0.89*	*0.88*
	1	*0.89*	*0.89*	*0.89*	
K-NN	*0*	*0.89*	*0.89*	*0.89*	*0.88*
	1	*0.89*	*0.89*	*0.89*	
FB-Fill	*0*	*0.89*	*0.89*	*0.89*	*0.88*
	1	*0.89*	*0.89*	*0.89*	



Şekil 10. 1-C SVM Sonuçları

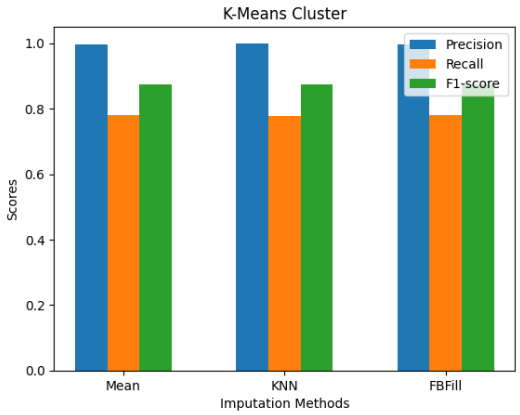
Tablo 5 ve Şekil 10'da görüldüğü üzere bu model anomalileri tespit etmekte hem hassasiyet olarak hemde duyarlılık olarak %89 oranında bir başarı elde etmiştir. P, R ve F_1 oranlarının aynı olması modelin anormal durumları etiketleme yeteneğinin oldukça güçlü olduğunu göstermektedir.

4.3 K-Means Clustering

Tablo 6. K-Means Clustering Sonuçları

	Etiket	*P*	*R*	*F1*	*A*
Ortalama	*0*	*0.82*	*1*	*0.90*	*0.88*
	1	*1*	*0.78*	*0.88*	
K-NN	*0*	*0.82*	*1*	*0.90*	*0.88*
	1	*1*	*0.78*	*0.88*	

FB-Fill	***0***	***0.82***	***1***	***0.90***	***0.88***
	1	***1***	***0.78***	***0.88***	



Şekil 11. K-Means Clustering Sonuçları

Bu modelin sonuçlarından modelin tüm anomalileri tespit edebildiğini, bu anomalilerin %78'inin doğru olduğu çıkartılmaktadır. %88 F_1 skoruyla bu modelinde anormal durumları etiketleme yeteneğinin güçlü olduğu söylenebilir.

4.4 GMM

Diğer modellerin aksine bu modelde imputasyon yöntemine göre sonuçlarda farklılıklar vardır.

Tablo 7. GMM Sonuçları

	Etiket	***P***	***R***	***F_1***	***A***
Ortalama	***0***	***0.81***	***0.97***	***0.88***	***0.87***
	1	***0.96***	***0.77***	***0.86***	
K-NN	***0***	***0.80***	***96***	***0.88***	***0.86***
	1	***0.95***	***0.76***	***0.85***	
FB-Fill	***0***	***0.79***	***0.95***	***0.90***	***0.84***
	1	***0.93***	***0.75***	***0.83***	

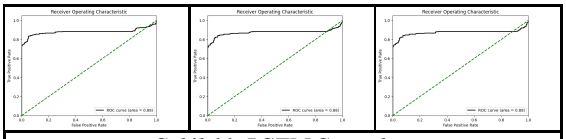
Model ortalama yöntemiyle impute edilmiş veriyle eğitildiğinde *%96* hassasiyetle anomali tespit edebilmekte ve bunun *%77*'sinin doğru tespit edildiği gözükmektedir. K-NN yöntemiyle işlenmiş verinin beslenmesiyle modelin hassasiyet ve duyarlılığı *%1* oranında düşmüştür. FB-Fill yöntemiyle işlenmiş verinin beslenmesi sonucu ise model, anomalilerin *%93*'ünü tespit edebilmekte ve bunların *%75*'ini doğru bilmektedir. F_1 skorları karşılaştırıldığında üç imputasyon yönteminin, modelin eğitiminde kayda değer bir değişim yaptığı söylenemez.

4.5 LSTM

Bu model bir derin öğrenme modeli olduğu için eğitimi diğerlerinden farklıdır. Eğitim verisi, aykırı değerlerinden ayıklanıp PCA uygulanmadan modele verilmiştir. Veri seti önce Min-Max yöntemiyle ölçeklendirilmiştir. Daha sonra eğitime başlanmıştır.

Tablo 8. LSTM Sonuçları

	Etiket	***P***	***R***	***F_1***	***ROC***
Ortalama	***0***	***0.85***	***0.95***	***0.90***	***0.89***
	1	***0.94***	***0.84***	***0.89***	
K-NN	***0***	***0.85***	***0.93***	***0.89***	***0.88***
	1	***0.92***	***0.84***	***0.88***	
FB-Fill	***0***	***0.85***	***0.93***	***0.89***	***0.88***
	1	***0.92***	***0.84***	***0.88***	



Şekil 11. LSTM Sonuçları

Tablo 8 ve Şekil 11'deki bulgulardan yola çıkarak bu modelin kendi özniteliklerini çıkarıp anomali tespitinde başarılı olduğu gözlenmektedir. Ortalama imputasyonuyla işlenmiş veriyle beslenen model, *%94* hassasiyetle anomali tespiti yapıp bunların *%84*'ünü doğru bilirken, diğer iki yöntemle işlenmiş veriyle beslenen model ise sadece %2'lik bir kayıpla hassasiyet bakımından geri kalmıştır. F_1 skorları ve ROC alanları göz önüne alındığında bu model anomali tespitinde oldukça başarılıdır.

5. SONUÇLAR VE ÖNERİLER

Bu çalışma kapsamında kullanılan veri seti için eksik değerlerin yerine konulması için uygulanan ortalama, K-NN ve FB-Fill yöntemlerinin, modellerin performanslarına çok büyük bir katkı sağlamadığı çıkarımı yapılabilir. Her bir modelin imputasyon yöntemi gözetmeksizin en iyi F_1 skorları ele alındığında aşağıdaki tablo ortaya çıkmaktadır.

	IF	***1C-SVM***	***K-Means***	***GMM***	***LSTM***
F_1	***0.78***	***0.89***	***0.88***	***0.86***	***0.89***

Bu sonuçlara göre, One-Class SVM ve LSTM modellerinin, veri setindeki anormal durumları etiketleme konusunda diğer modellere göre daha başarılı bir performans sergilediği görülmektedir.

Model güncelleme ve ayarlama süreçleri üzerinde çalışmak, mevcut modellerin daha da iyileştirilmesine katkı sağlayabilir. Ayrıca model sonuçlarının iş uygulamalarına entegrasyonunu sağlamak ve gerçek dünya senaryolarında nasıl kullanılacağı düşünülebilir. Bu, anomali tespiti modellerinin etkili bir şekilde implemente edilerek, iş süreçlerinde olası anormal durumları önceden belirleme yeteneğini artırabilir.

REFERANSLAR

[1] A. B. Nassif, M. A. Talib, Q. Nasir and F. M. Dakalbab, "Machine learning for anomaly detection: A systematic review", IEEE Access, vol. 9, pp. 78658-78700, 2021.

[2] V. Chandola, A. Banerjee, and V. Kumar, ''Anomaly detection: A survey,'' ACM Comput. Surv., vol. 41, no. 3, pp. 71–97, 2009, doi:10.1145/1541880.1541882.

[3] M. Injadat, F. Salo, A. B. Nassif, A. Essex, and A. Shami, ''Bayesian optimization with machine learning algorithms towards anomaly detection,'' in Proc. IEEE Global Commun. Conf. (GLOBECOM), Dec. 2018, pp. 1–6, doi: 10.1109/GLOCOM.2018.8647714.

[4] Anomaly Detection in Helicopter, https://www.kaggle.com/datasets/nagasai524/anomaly-detection-in-helicopter/data.

[5] H. Izakian and W. Pedrycz. 2013. Anomaly detection in time series data using a fuzzy c-means clustering. In Proceedings of the 2013 Joint IFSA World Congress and NAFIPS Annual Meeting. IEEE, Edmonton, Alberta, Canada, 1513–1518.

[6] R. J. Hyndman, E. Wang, and N. Laptev. 2015. Large-Scale Unusual Time Series Detection. In Proceedings of the 15th IEEE International Conference on Data Mining Workshop (ICDMW '15). IEEE, Atlantic City, NJ, USA, 1616–1619.

[7] L. Xiong, H.-D. Ma, H.-Z. Fang, K.-X. Zou and D.-W. Yi, "Anomaly detection of spacecraft based on least squares support vector machine", Proc. Prognostics Syst. Health Management Conf., pp. 1-6, May 2011..

[8] H. N. Akouemo and R. J. Povinelli. 2016. Probabilistic anomaly detection in natural gas time series data. Int. J. Forecast. 32, 3 (2016), 948–956.

[9] N. Laptev, S. Amizadeh, and I. Flint. 2015. Generic and Scalable Framework for Automated Time-series Anomaly Detection. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15). ACM, Sydney, NSW, Australia, 1939–1947.

[10] A. Nanduri and L. Sherry, "Anomaly Detection in Aircraft Data Using Recurrent Neural Networks (RNN)", Proceedings of Integrated Communications Navigation and Surveillance, pp. 5C2-1-5C2-8, April 2016.

[11] Yan, S., Shao, H., Xiao, Y., Liu, B., & Wan, J. (2023). Hybrid robust convolutional autoencoder for unsupervised anomaly detection of machine tools under noises. Robotics and Computer-Integrated Manufacturing, 79, Article 102441.

[12] Vos K, Peng Z, Jenkins C, Shahriar MR, Borghesani P, Wang W. Vibration-based anomaly detection using LSTM/SVM approaches. Mech Syst Signal Process 2022; 169:108752. https://doi.org/10.1016/j.ymssp.2021.108752.

[13] Liu, C., & Gryllias, K. . (2021). A Deep Support Vector Data Description Method for Anomaly Detection in Helicopters. PHM Society European Conference, 6(1), 9. https://doi.org/10.36001/phme.2021.v6i1.2957.

[14] Ammann, Oliver et al. "Anomaly Detection And Classification In Time Series With Kervolutional Neural Networks." ArXiv abs/2005.07078 (2020).

[15] M. Jones, D. Nikovski, M. Imamura, and T. Hirata. 2016. Exemplar learning for extremely efficient anomaly detection in real-valued time series. Data Min.Knowl. Discov. 30, 6 (2016), 1427–1454.

[16] Chandola, Varun; Cheboli, Deepthi; Kumar, Vipin. (2009). Detecting Anomalies in a Time Series Database. Retrieved from the University of Minnesota Digital Conservancy, https://hdl.handle.net/11299/215791.

[17] Mohamed Ben Slimene, Mohamed-Salah Ouali,Anomaly Detection Method of Aircraft System using Multivariate Time Series Clustering and Classification Techniques, IFAC-PapersOnLine,Volume 55, Issue 10,2022,Pages 1582-1587,ISSN 2405-8963,https://doi.org/10.1016/j.ifacol.2022.09.616.

[18] Iliopoulos, Anastasios, et al. "Detection of Anomalies in Multivariate Time Series Using Ensemble Techniques." 2023 IEEE Ninth International Conference on Big Data Computing Service and Applications (BigDataService). IEEE, 2023.

[19] G. Li, A. Rai, H. Lee, A. Chattopadhyay, "Operational anomaly detection in flight data using a multivariate gaussian mixture model". In proceedings of the 10th Annual Conference of the Prognostics and Health Management Society, 2018. doi: 10.36001/phmconf.2018.v10i1.474.

[20] Z. Ghrib, R. Jaziri and R. Romdhane, "Hybrid approach for Anomaly Detection in Time Series Data," 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 2020, pp. 1-7, doi: 10.1109/IJCNN48605.2020.9207013.

[21] A. Gaev and A. Lantsberg, "A Method to Handle Unstable Time Series in Anomaly Detection Problem," 2021 3rd International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA), Lipetsk, Russian Federation, 2021, pp. 548-552, doi: 10.1109/SUMMA53307.2021.9632243.

[22] J. Shi, G. He and X. Liu, "Anomaly Detection for Key Performance Indicators Through Machine Learning," 2018 International Conference on Network Infrastructure and Digital Content (IC-NIDC), Guiyang, China, 2018, pp. 1-5, doi: 10.1109/ICNIDC.2018.8525714.

[23] H. C. Mandhare and S. R. Idate, "A comparative study of cluster based outlier detection, distance based outlier detection and density based outlier detection techniques," 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2017, pp. 931-935, doi: 10.1109/ICCONS.2017.8250601.
[24] Y. Jiang, N. Le, Y. Zhang, Y. Zheng and Y. Jiao, "Research on the Flight Anomaly Detection During Take-off Phase Based on FOQA Data," 2019 CAA Symposium on Fault Detection, Supervision and Safety for Technical Processes (SAFEPROCESS), Xiamen, China, 2019, pp. 756-760, doi: 10.1109/SAFEPROCESS45799.2019.9213422.
[25] Bishop, Christopher (2006). Pattern recognition and machine learning. Berlin: Springer. ISBN 0-387-31073-8.
[26] Liu, Fei Tony. "First Isolation Forest implementation on Sourceforge".
[27] Liu, Fei Tony; Ting, Kai Ming; Zhou, Zhi-Hua (December 2008). "Isolation Forest". 2008 Eighth IEEE International Conference on Data Mining. pp. 413–422. doi:10.1109/ICDM.2008.17. ISBN 978-0-7695-3502-9. S2CID 6505449.
[28] Liu, Fei Tony; Ting, Kai Ming; Zhou, Zhi-Hua (December 2008). "Isolation-Based Anomaly Detection". ACM Transactions on Knowledge Discovery from Data. 6: 3:1–3:39. doi:10.1145/2133360.2133363. S2CID 207193045.
[29] Hastie, T.; Tibshirani, R.; Friedman, J. (2009). "Unsupervised Learning". The Elements of Statistical Learning (2nd ed.). Springer.
[30] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.
[31] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

github:
https://github.com/frkanyilmaz2/machine_learning/blob/main/main.ipynb