HAVA KALİTESİ TAHMİNLEMESİ İÇİN BİR YÖNTEM GERÇEKLEŞTİRİLMESİ

Derya AKBAYIR
221450103

DOKTORA TEZİ
Bilgisayar Mühendisliği Anabilim Dalı
Bilgisayar Mühendisliği Doktora Programı
Danışman: Prof.Dr. Emin Murat ESİN

İstanbul
T.C. Maltepe Üniversitesi
Lisansüstü Eğitim Enstitüsü
Temmuz, 2025

HAVA KALİTESİ TAHMİNLEMESİ İÇİN BİR YÖNTEM GERÇEKLEŞTİRİLMESİ

Derya AKBAYIR
221450103
ORCID: 0009-0000-6495-3028

DOKTORA TEZİ
Bilgisayar Mühendisliği Anabilim Dalı
Bilgisayar Mühendisliği Doktora Programı
Danışman: Prof. Dr. Emin Murat ESİN

İstanbul
T.C. Maltepe Üniversitesi
Lisansüstü Eğitim Enstitüsü
Temmuz, 2025



JÜRİ VE ENSTİTÜ ONAYI

DERYA AKBAYIR'ın "Hava Kalitesinin Tahminlemesi İçin Bir Yöntem Geliştirilmesi " başlıklı tezi 14.07.2025 tarihinde aşağıdaki jüri tarafından değerlendirilerek "Maltepe Üniversitesi Lisansüstü Eğitim ve Öğretim Yönetmeliği" nin ilgili maddeleri uyarınca Bilgisayar Mühendisliği Anabilim Dalı Yüksek Lisans/Doktora/Sanatta Yeterlik tezi oy birliğiyle/oy çokluğuyla, başarılı/başarısız olarak kabul edilmiştir.

	Unvanı. Adı ve Soyadı	İmza
Üye (Tez Danışmanı)	Prof.Dr. Emin Murat ESİN (Danışman) Maltepe Üniversitesi	.....
Üye	Prof.Dr. Sinan APAK Fenerbahçe Üniversitesi	.....
Üye	Dr.Öğr.Üyesi Erdal GÜVENOĞLU Maltepe Üniversitesi	.....
Üye	Dr.Öğr.Üyesi Ender ŞAHİNARSLAN Mudanya Üniversitesi	.....
Üye	Dr.Öğr.Üyesi Ayşe SALMAN Maltepe Üniversitesi	.....

Prof. Dr. Selva ERSÖZ KARAKULAKOĞLU
Enstitü Müdürü

ETİK İLKE VE KURALLARA UYUM BEYANI

Bu belge, Yükseköğretim Kurulu tarafından 19.01.2021 tarihli "Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge" ile bildirilen 6698 Sayılı Kişisel Verilerin Korunması Kanunu kapsamında gizlenmiştir.

TEŞEKKÜR

Bu doktora tezinin hazırlanması sürecinde, değerli bilgileri ve rehberliği ile her aşamada beni yönlendiren danışman hocam Prof. Dr. Emin Murat ESİN'e en içten teşekkürlerimi sunarım. Ayrıca, bu süreçte bana sabırla destek sağlayan ve her daim yanımda olan kıymetli eşime ve hayatımın anlamı çocuklarıma çok teşekkür ederim. Son olarak, hayatım boyunca her zaman sundukları koşulsuz sevgi ve destekleri için sevgili anneme ve ömrü bu tezin tamamlandığını görmeye yetmeyen canım babama teşekkürlerimi sunuyorum.

Derya AKBAYIR

Temmuz, 2025

ÖZET

HAVA KALİTESİ TAHMİNLEMESİ İÇİN BİR YÖNTEM GERÇEKLEŞTİRİLMESİ

Derya Akbayır
Doktora Tezi
Bilgisayar Mühendisliği Anabilim Dalı
Bilgisayar Mühendisliği Doktora Programı
Danışman: Prof. Dr. Emin Murat Esin
Maltepe Üniversitesi Lisansüstü Eğitim Enstitüsü, 2025

Hava kirliliği, modern şehirleşmenin göz ardı edilen bir sonucu olarak, yaşamımızı etkileyen en önemli çevresel ve halk sağlığı sorunlarından biri olmuştur. Kentsel hava kalitesi izlenme süreçlerinin önemli bir aşaması olan hava kalitesi tahmin modelleri kullanılarak, hava kalitesinin gelecekte nasıl olacağını tahmin etmemizi sağlar. Bu çalışmada, hava kalitesi ve meteoroloji verilerini kullanarak kısa ve orta vadeli PM_{10} tahminleri için özgün bir hibrit derin öğrenme modeli önerilmektedir. Bu model, doğrusal olmayan ve mevsimsel bileşenleri etkili bir şekilde modelleyerek tahminlerin doğruluğunu ve genellenebilirliğini artırmayı amaçlamaktadır. Bu çalışma, Ankara il sınırları içerisindeki 18 hava kalitesi izleme istasyonundan 2019-2023 yılları arasında saatlik olarak kaydedilen PM_{10}, $PM_{2.5}$, SO_2, CO, NO_2, NO_x, NO ve O_3 kirleticileri ile sıcaklık, nem, yağış miktarı, rüzgâr hızı ve basınç gibi meteorolojik değişkenleri içeren kapsamlı bir veri setine dayanmaktadır.

Önerilen model yapısı, zaman serilerindeki uzun mesafeli bağımlılıkları tespit etmede başarılı olan Uzun Kısa Süreli Bellek (Long Short -Term Memory, LSTM) katmanları ile verinin farklı bölümlerine odaklanabilme yeteği sunan dikkat mekanizmaları kullanan Transformer Encoder bloklarını birleştirir. Tek adımlı tahmin yöntemleri yerine özyinelemeli tahmin tekniği uygulanmıştır. Böylelikle model, her tahmin sonucunu bir sonraki adımın girdisi olarak kullanarak üç günlük ardışık öngörüler gerçekleştirmiştir. Bu yöntemle, özellikle geleceğe yönelik tahminlerde toplam hata birikiminin azaltılması amaçlanmıştır.

Modelin hiperparametre ayarlama sürecinde, Optuna kütüphanesinde bulunan TPE (Tree-structured Parzen Estimator) örnekleyici kullanılarak, öğrenme oranı, katman sayısı ve pencere uzunluğu gibi önemli parametreler sistematik bir şekilde optimize edilmiştir. Bununla birlikte, aşırı öğrenmenin (overfitting) engellenmesi ve modelin genelleme yeteneğini yükseltmek amacıyla, Batch Normalization, Dropout ve ağırlıkların güncellenmesinde AdamW optimizasyon algoritması kullanılmıştır.

Deneysel bulgular, önerilen hibrit modelin hem kısa vadeli (1–24 saat) hem de orta vadeli (25–72 saat) tahminlerde literatürdeki geleneksel istatiksel yöntemler ile güncel derin öğrenme yöntemlerine nazaran anlamlı derecede daha düşük ortalama mutlak hata (MAE) ve kök ortalama kare hata (RMSE) sonuçları ortaya koyduğunu göstermektedir. Diğer taraftan, yalnızca 588 bin parametreli yapısıyla, daha büyük grafik-temelli modellere göre dikkate değer bir hesaplama verimliliği elde edilmiştir.

Geliştirilen model ayrıca, Streamlit temelli kullanıcı dostu bir arayüzle desteklenmiştir. Bu arayüz aracılığıyla, istasyon bazında görselleştirmeler, zamansal tahmin grafikleri ve gerçek zamanlı veri akışlarını etkili bir şekilde sunmak mümkündür. Modelin gerçek zamanlı veri akışlarıyla birleşimi, ani hava kalitesi değişimlerine karşı sürekli ve güncel öngörüler sunabilmektedir. Bu bağlamda sistem, sadece tahmin çıktıları değil, aynı zamanda dinamik çevresel koşullara duyarlı ve sürdürülebilir bir öngörü sistemi olarak tasarlanmıştır.

Sonuç olarak, bu tez kapsamında geliştirilen hibrit derin öğrenme modeli ile destekleyici sistem altyapısının, hava kirliliğinin olumsuz etkilerinin azaltılması ve koruyucu önlemlerin zamanında alınabilmesi için yetkili kurumlara önemli katkılar sağlayabileceği öngörülmektedir. Özellikle, kentsel hava kalitesi yönetimi alanında, veri odaklı stratejik karar alma süreçlerini destekleyecek fonksiyonel bir karar destek sistemi altyapısı olarak entegre edilebilme potansiyeline sahip olduğu değerlendirilmektedir.

Anahtar Sözcükler: Hibrit Hava Kirliliği Tahmin Modeli, LSTM, Transformer, Optuna, Gerçek Zamanlı Tahmin Sistemi, Derin Öğrenme, Streamlit, MICE

ABSTRACT

A METHODOLOGICAL FRAMEWORK FOR AIR QUALITY FORECASTING

Derya Akbayır
PhD Thesis
Computer Engineering Department
Computer Engineering Doctoral Programme
Thesis Advisor: Prof. Dr. Emin Murat Esin
Maltepe University Graduate School, 2025

Air quality problems have become a big challenge for the environment and public health. They often show up as a hidden effect of urbanization today. Air quality forecasting models are important for monitoring air in cities. They can predict future air quality, which helps in taking timely actions and planning effectively. This thesis introduces a novel hybrid deep learning model that incorporates meteorological data to improve the accuracy and generalizability of short- and medium-term PM_{10} concentration forecasts, effectively capturing the nonlinear and seasonal components inherent in atmospheric time series data. This study uses a broad dataset that includes hourly air pollutant measurements. The pollutants measured are PM_{10}, $PM_{2.5}$, SO_2, CO, NO_2, NOx, NO and O_3. It also includes meteorological data like temperature, humidity, precipitation, wind speed and atmospheric pressure. The data were gathered from 18 air quality monitoring stations in Ankara, Turkey, from 2019 and 2023.

The proposed architecture combines Long Short Term Memory (LSTM) layers with Transformer Encoder blocks. LSTM layers are effective for modeling long-range dependencies. Meanwhile, Transformer blocks apply attention mechanisms to focus on various parts of the input sequence. The model uses a recursive prediction strategy instead of traditional single-step forecasting. It feeds each predicted output back as input for the next forecast. This allows for sequential predictions over three days. This framework is designed to mitigate cumulative error propagation typically associated with multi-step forecasting.

Hyperparameter optimization was systematically done with Optuna. This automated optimization framework relies on the Tree-structured Parzen Estimator (TPE) algorithm.We systematically and automatically adjusted important factors like learning rate, the number of layers and input window size. To additionally improve generalization and avert overfitting, Batch Normalization, Dropout, and the AdamW optimizer were integrated into the training process.

Test findings indicate that the suggested hybrid model reliably exceeds conventional statistical techniques and modern deep learning approaches regarding Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), for both short term (1–24 hours) and medium term (25–72 hours) predictions. As well, while it demonstrates significant predictive capabilities, the model preserves computational efficiency with a compact architecture containing nearly 588 parameters, offering a competitive alternative to larger graph-based models.

The developed model is further supported by a user-friendly interface based on Streamlit. By means of this interface, location-based visual representations, time prediction charts, and real-time data streams can be effectively demonstrated. The merging of the model with real-time data streams allows for ongoing immediate updates. This enables respond quickly to sudden changes in air quality. Consequently, the system is designed not only to produce reliable forecast outputs but also to operate as a responsive and sustainable prediction system that adapts to changing environmental conditions.

Accordingly, the hybrid deep learning model and the accompanying supportive system infrastructure that this thesis developed are expected to provide meaningful contributions to authorized institutions in mitigating the adverse effects of air pollution and facilitating timely implementation of preventive measures. Specifically, the suggested system has great potential. It can effectively support decisions in urban air quality management. This will improve data-driven strategies and decision-making processes.

Keywords: Hybrid Air Pollution Forecasting Model, LSTM, Transformer , Optuna, Real-time Forecasting System, Deep Learning, Streamlit, MICE

İÇİNDEKİLER

JÜRİ VE ENSTİTÜ ONAYI ii

ETİK İLKE VE KURALLARA UYUM BEYANI iii

TEŞEKKÜR iv

ÖZET v

ABSTRACT vii

İÇİNDEKİLER ix

TABLOLAR LİSTESİ xiii

ŞEKİLLER LİSTESİ xiv

KISALTMALAR LİSTESİ xvii

SİMGELER LİSTESİ xxi

1. GİRİŞ 1

1.1 Problem 1

1.2 Amacı ve Akademik Katkıları 2

1.3 Hava Kirliliği Tahmin Modellerinin Önemi 4

1.4 Hibrit Modellerin Avantajları ve Hava Kalitesi Tahminindeki Rolü 6

1.5 Sınırlılıklar 8

1.6 Tezin Kurgusu 9

2. LİTERATÜR ARAŞTIRMASI 10

3. VERİ HAZIRLAMA VE ÖN İŞLEME 16
3.1 Çalışma Alanı 16
3.2 Veri Kaynakları 17
3.2.1 Hava kalitesi izleme istasyon verileri 17
3.2.1.1 Hava kalitesi ulusal sınır değerleri 22
3.2.2 Meteoroloji istasyon verileri 27
3.2.3. Hava kalitesi ve meteoroloji istasyon verilerinin birleştirilmesi 29
3.3 Veri Temizleme ve Normalizasyon Süreçleri 30
3.4 Eksik Verilerin İşlenmesi ve Tamamlannması 32
3.4.1 Eksik veri dağılımı ve görselleştirilmesi 32
3.4.2 Eksik veri tamamlama yöntemleri 37
3.4.3. Karşılaştırılmalı performans analizi 39
3.4.4 MICE ile Veri Tamamlama 41
3.5 Aykırı Değerlerin Tespiti 42
3.6 Veri Normalizasyonu 47
3.7 Zaman Serisi Analizi (Time Series Analysis) 48
4. VERİ KEŞFİ (EXPLORATORY DATA ANALYSIS) 52
4.1 Hava Kalitesi Verilerinin Genel Durumu 52
4.2 Dendrogram ile Değişkenler Arası Benzerlik Analizi 55
4.3 Hava Kirletici Değerlerinde Korelasyon Analizi 57
4.4 Hava Kirliliği Değerlendirmesi ve Kirlilik Kontrol Stratejilerine Etkisi 59

4.5 Hava Kalitesi İstasyonlarından Ölçülen Değerlerin Histogram Analizi 59

4.6. Hava Kalitesi Verilerinin İlişkisel İncelemesi 61

4.7. Hava Kalitesi Konsantrasyonlarının Yıllık Değişim ve Uzun Dönem Eğilim Analizleri 63

5. MODEL GELİŞTİRME 68

5.1 Kullanılan Algoritmaların Teorik Açıklaması 70

5.1.1. Uzun kısa süreli bellek mimarisi (LSTM) 70

5.1.2. Transformer model 72

5.2. Önerilen Tahmin Model Mimarisi 76

5.2.1. Önerilen modelin mimari yapısı 76

5.2.2. Model mimarisi ve matematiksel gösterim 78

5.3. Özyinelemeli ve Çok Adımlı Tahmin Yaklaşımları 82

5.4. Optimizasyon Yöntemleri 84

5.4.1. Optuna ile hiperparametre optimizasyonu 87

5.5. Model Performans Değerlendirme Metrikleri 88

6. HAVA KİRLİLİĞİ TAHMİN UYGULAMASI GELİŞTİRME 91

6.1. Uygulama Arayüzü 91

6.1.1. Kullanılan teknoloji altyapısı 92

6.1.2. Uygulama mimarisi 93

6.2. Gerçek Zamanlı Hava Kalitesi Tahmin Sistemi 97

6.2.1. Gerçek zamanlı veri sağlayıcıları 97

6.2.2. Gerçek zamanlı uygulama mimarisi 98

7. ARAŞTIRMANIN BULGULARI 102

7.1. Model Eğitim Süreci ve Optimizasyonu 102

7.1.1 Veri işleme 103

7.1.2 Model işlemleri ve optimizasyon 106

7.1.3 Optuna ile hiperparametre analizi 107

7.3 Deneysel Çalışmaların Bulguları 113

7.3.1 Karşılaştırmalı performans analizi 115

7.3.2 Literatürle ilişkilendirme 116

8. SONUÇ VE ÖNERİLER 117

KAYNAKÇA 119

ÖZ GEÇMİŞ 129

TABLOLAR LİSTESİ

Tablo 1. Hava Kalitesi İstasyon Verileri 21

Tablo 2. Hava Kirliliği Tahmin Uygulamasında Kullanılan Teknolojiler 93

Tablo 3. Hiperparametre Ayarları 108

Tablo 4. 72 Saatlik Çok-Adımlı PM_{10} Tahmini İçin Modeller Arası Performans ve Karmaşıklık Karşılaştırmaları 114

ŞEKİLLER LİSTESİ

Şekil 1. Ankara ilinin fiziki haritası 17

Şekil 2. İstasyon Bilgileri Haritası 19

Şekil 3. İstasyonların Yıllara Göre PM10 Günlük Sınır Aşım Sayısı 24

Şekil 4. Yıllara Göre İstasyonların Günlük PM10 Aşım Sayıları 25

Şekil 5. Yıllara Göre İstasyonların Günlük PM10 Aşım Sayıları 3D Gösterimi 27

Şekil 6. Hava Kalitesi ve Meteoroloji İstasyonlarının Konum Bilgisi 29

Şekil 7. Hava Kalite ve Meteoroloji İstasyon Konum Eşleştirme Matrisi 30

Şekil 8. Saatlik Meteoroloji Verilerindeki Eksik Değerlerin Isı Haritası 33

Şekil 9. Saatlik Hava Kirletici Verilerindeki Eksik Veri Değerlerin Isı Haritası 34

Şekil 10. Hava Kirleticilerin Violin Plot Gösterimi 34

Şekil 11. Hava Kirletici Verilerindeki Eksik Veri Miktarı 36

Şekil 12. Hava Kirletici Verilerindeki Eksik Veri Oranları 36

Şekil 13. Meteoroloji Verilerindeki Eksik Veri Oranları 37

Şekil 14. Eksik Veri Tamamlama Yöntemlerinin Performans Karşılaştırması 41

Şekil 15. Eksik ve Tamamlanmış Verilerin Isı Haritası 42

Şekil 16. PM10 Değeri için IQR yöntemi ile Aykırı Deger Tespiti 44

Şekil 17. PM10 Değeri için LOF yöntemi ile Aykırı Deger Tespiti 45

Şekil 18. PM10 Değeri için Isolation Forest ile Aykırı Deger Tespiti 46

Şekil 19. Zaman Serisi Bileşenlerinin Görsel Analizi 51

Şekil 20. Hava Kalitesi Verilerinin İstatiksel Analizi 53

Şekil 21. Hava Kalitesi Verilerinin İstasyonlara Göre Doluluk Oranı 54

Şekil 22. İstasyonlara Göre Yıllık Ortalama PM10 Dağılım Grafiği 55

Şekil 23. Hava Kalitesi Verilerinin Dendrogram Analizi..... 55

Şekil 24. Hava Kirleticilerin Korelasyon Analizi..... 57

Şekil 25. Hava Kirleticiler Arası Korelasyon Analizi Çubuk Grafiği 58

Şekil 26. Hava Kalitesi İstasyon Verilerinin Dağılımı (Histogram Analizi)..... 60

Şekil 27. Hava Kirleticilerin Aykırı Değer Analizi (Boxplot Grafik Analizi) 61

Şekil 28. Hava Kalitesi Verilerinin İlişkisel İncelemesi (Pair Plot Analizi) 62

Şekil 29. Hava Kirleticilerin Yıl Bazında Yüzdelik Değişim Grafiği..... 65

Şekil 30. Kirleticilerin Yıllık Konsantrasyon Grafiği 66

Şekil 31. Hava Kirleticilerin Yüzdelik Değişim Grafiği (Trend Analizi) 67

Şekil 32. Hibrit Hava Kirliliği Tahmin Modeli Mimarisi Akış Şeması 68

Şekil 33. Transformer Model Mimarisi..... 73

Şekil 34. Önerien Modelin Mimari Akış Şeması 78

Şekil 35. Hava Kirliliği Tahmin Uygulaması: Veri ve Model Yükleme Arayüzü 94

Şekil 36. Hava Kirliliği Tahmin Uygulaması: Eğitim Aşaması Arayüzü (2)..... 95

Şekil 37. Hava Kirliliği Tahmin Uygulaması : Harita Seçimi (3)..... 95

Şekil 38. Hava Kirliliği Tahmin Uygulaması: Anlık Tahmin Sonucu 96

Şekil 39. Gerçek Zamanlı PM10 Tahmin Uygulaması Ekran Görüntüsü 99

Şekil 40. Gerçek Zamanlı PM10 Tahmin Uygulaması Ekran Görüntüsü (2) 100

Şekil 41. Zaman Serisi Verisinde Veri Bölümleme 104

Şekil 42. Optuna ile Slice Plot ile Analizi 109

Şekil 43. Hiperparametre Önem Analizi Grafiği 109

Şekil 44. Optuna ile Optimizasyon Geçmişi 110

Şekil 45. PM10 Gerçek ve Tahmin Verisi Gösterimi 111

Şekil 46. PM10 Gerçek ve Tahmin Verisi Gösterimi (İlk 5000 Gözlem) 112

KISALTMALAR LİSTESİ

AI : Artificial Intelligence / Yapay Zeka

AQI : Air Quality Index / Hava Kalitesi İndeksi

ARIMA : AutoRegressive Integrated Moving Average / Otoregresif Entegre Hareketli Ortalama

BiLSTM : Bidirectional Long Short-Term Memory / Çift Yönlü Uzun Kısa Süreli Bellek

CNN : Convolutional Neural Networks / Evrişimli Sinir Ağları

CO : Carbon Monoxide / Karbon Monoksit

COVID-19 : Koronavirüs Hastalığı

CSV : Comma-Separated Values (Virgülle Ayrılmış Değerler Dosyası)

ÇŞİDB : T. C. Çevre, Şehircilik ve İklim Değişikliği Bakanlığı

EPA : Environmental Protection Agency / Çevre Koruma Ajansı

GBM : Gradient Boosting Machines / Gradyan Artırmalı Makineler

GAT : Graph Attention Networks / Grafik Dikkat Ağları

GNN : Graph Neural Networks / Grafik Sinir Ağları

GPU : Graphics Processing Unit / Grafik İşleme Birimi

GRU : Gated Recurrent Unit / Kapılı Yinelemeli Birim

HKDYY : Hava Kalitesi Değerlendirme ve Yönetimi Yönetmeliği

JSON : JavaScript Object Notation / JavaScript Nesne İşaretlemesi

KNN : K-Nearest Neighbors / En Yakın Komşular Algoritması

LEZ : Low Emission Zones / Düşük Emisyon Bölgeleri

LSTM : Long Short-Term Memory / Uzun Kısa Süreli Bellek

MAE : Mean Absolute Error / Ortalama Mutlak Hata

MAPE : Mean Absolute Percentage Error / Ortalama Mutlak Yüzde Hatası

MGM : Meteoroloji Genel Müdürlüğü

MICE : Multiple Imputation by Chained Equations / Zincirli Denklemlerle Çoklu Atama

ML : Machine Learning / Makine Öğrenmesi

MLR : Multiple Linear Regression / Çoklu Lineer Regresyon

MSE : Mean Squared Error / Ortalama Kare Hata

NASA : National Aeronautics and Space Administration / Ulusal Havacılık ve Uzay Dairesi

NN : Neural Network / Sinir Ağı

NO_2 : Nitrogen Dioxide / Azot Dioksit

NOx : Nitrogen Oxides / Azot Oksitler

O_3 : Ozone / Ozon

PM_{10} : Particulate Matter ≤10 microns / 10 mikrondan küçük Partikül Madde

PM25 : Particulate Matter ≤2,5 microns / 2,5 mikrondan küçük Partikül Madde

PMM : Predictive Mean Matching / Tahmin Edilen Ortalama

RF : Random Forest / Rastgele Orman

RMSE : Root Mean Squared Error / Kök Ortalama Kare Hata

RNN : Recurrent Neural Network / Tekrarlayan Sinir Ağları

SARIMA : Seasonal Auto Regressive Integrated Moving Average / Mevsimsel Otoregresif Entegre Hareketli Ortalama

SİM : Sürekli İzleme Merkezi

SO2 : Sulfur Dioxide / Kükürt Dioksit

STL **:** Seasonal-Trend Decomposition using Loess / Loess Kullanarak Mevsimsellik-Trend Ayrıştırma

SVM **:** Support Vector Machine / Destek Vektör Makineleri

UHKİA **:** Ulusal Hava Kalitesi İzleme Ağı

WHO **:** World Health Organization / Dünya Sağlık Örgüt

SİMGELER LİSTESİ

°C : Celcius / Derece

Ct : Cell State / Hücre Durumu

ft : Forget Gate / Unutma Kapısı

İt : İnput gate / Giriş kapısı

K : Key/ Anahtar

mg : miligram

m³ : metre küp

µg : mikrogram

ot : Output Gate / Çıkış Kapısı

Q : Query/Sorgu

V : Value/ Değer

1. GİRİŞ

1.1 Problem

Hava kirliliği, modern şehirleşmenin getirdiği bir çevre sorunu gibi gözüksede, aynı zamanda halk sağlığı, şehir planlaması, ekosistem bütünlüğü ve iklim değişikliğiyle mücadele kapsamında öncelikli çözüm gerektiren küresel ölçekte bir sorundur. Dünya Sağlık Örgütü (World Health Organization, WHO) verilerine göre, her yıl yaklaşık 7 milyon kişi dış ve iç ortam hava kirliliğine bağlı hastalıklar nedeniyle hayatını kaybetmektedir (WHO, 2021). Hava kirliliği; astım, bronşit gibi solunum yolu hastalıklarından, nörolojik bozulmalara, kardiyovasküler rahatsızlıklara, akciğer kanserine kadar ciddi sağlık sorunlarına neden olmaktadır. Bu yüzden, hava kirliliği verilerinin düzenli olarak takip edilmesi, güvenilir tahminler için doğru bir şekilde modellenmesi ve elde edilen bilgilerin veri odaklı ve izlenebilir karar destek sistemlerine entegre edilmesi sadece akademik bir zorunluluk değil, aynı zamanda sosyal bir sorumluluktur. Özellikle, bölgesel hava kalitesi planlamasında model destekli analitik yöntemler, merkezi ve yerel yönetimler için daha işlevsel yönetim stratejileri oluşturmak ve hayata geçirmek için sağlam bir temel sunmaktadır.

İnce partikül maddeler (PM_{10} ve $PM_{2.5}$), azot dioksit (NO_2), kükürt dioksit (SO_2), karbon monoksit (CO) ve ozon (O_3) hava kirliliğinin temel bileşenleridir. Bu maddeler, hem insan sağlığına hem de çevreye zarar verebilir. Bu kirleticilerin havadaki varlığı ve zaman içindeki dağılımları, meteorolojik durumlara, coğrafi özelliklere, insan kaynaklı emisyonlara ve atmosferik kimyasal tepkimelere bağlı olarak büyük ölçüde değişkenlik göstermektedir. Özellikle COVID-19 pandemisi sırasında hava kalitesinde geçici bir iyileşme oldu. Bu süreçte kardiyovasküler ölüm oranları da azaldı. Böylece, hava kalitesinin halk sağlığı üzerindeki etkilerini belirgin hale getirdi (Venter, Aunan, Chowdhury ve Lelieveld, 2020).

Hava kirliliği, günümüzde halk sağlığından başlayarak ekosistem bütünlüğünü ve iklim direncini doğrudan etkileyen karmaşık bir problem alanı olarak öne çıkmaktadır. Mevcut tahmin yöntemleri, bu karmaşık yapıyı zaman serisi odaklı, çok değişkenli ve istasyonel

düzeyde bir araya getirme noktasında yetersiz kalmaktadır. Bu bağlamda, kirletici konsantrasyonlarını ve meteorolojik değişkenleri birlikte değerlendirerek kısa ve orta vadeli hava kirliliği tahmini yapabilen, açıklanabilirliği yüksek, zaman serisi tabanlı hibrit bir derin öğrenme modeline duyulan ihtiyaç kritik bir akademik ve uygulamalı problem olarak öne çıkmaktadır.

1.2 Amacı ve Akademik Katkıları

Bu çalışmanın temel amacı, çok bileşenli hava kirleticileri ile meteorolojik değişkenler arasındaki dinamik ilişkileri dikkate alarak, zaman serisi verisi üzerinde derin öğrenme tabanlı hibrit bir model aracılığıyla üç günlük hava kirliliği tahminleri gerçekleştirmektir. Üç günlük tahmin aralığı, kısa vadeli kirlilik tahminlerinde karar vericilere yeterli müdahale süresi tanırken, aynı zamanda modelin özyinelemeli tahminlerde (recursive prediction) kümülatif hata birikimini minimize edecek şekilde dengeli bir zaman ufku sağlar. Literatürde, 1–3 günlük tahminlerin hava kalitesi yönetimi için kritik olduğu vurgulanıyor. Bu tahminler, halk sağlığı uyarı sistemleri açısından önemli bir karar desteği sağlıyor. Bu nedenle, çalışmada 3 günlük bir tahmin aralığı uzunluğu (prediction horizon) benimsenmiştir.

Bu kapsamda geliştirilen model, her tahmin adımında önceki sonuçları yeni bir girdiye dönüştürerek zamanla hata birikimini azaltmayı amaçlamaktadır. Model, özyinelemeli tahmin yöntemine dayanır. Böylece, kısa vadeli karar destek sistemleri için uygun hale gelir. Ayrıca, hava kalitesi yönetimi, halk sağlığı uyarıları ve sürdürülebilir şehircilik gibi alanlara da özgün katkıda bulunur.

Bu bağlamda çalışma, literatürdeki mevcut yaklaşımlarla karşılaştırıldığında aşağıdaki özgün akademik katkıları sunmaktadır.

A. Hibrit LSTM+Transformer model tasarımı

Bu tez çalışmasında, Uzun Kısa Süreli Bellek (Long Short-Term Memory, LSTM) ve Transformer Encoder yapıları birleştirilmiştir. Amacımız, veri setindeki doğrusal olmayan ve mevsimsel bileşenleri etkili bir şekilde modellemektir. Bu hedefle özgün bir hibrit model önerilmiştir.

LSTM katmanı, zaman bağımlılıklarını uzun vadeli öğrenme yeteneği sayesinde etkili bir şekilde modelleyebilir. Ardından, Transformer Encoder ise dikkat mekanizması kullanarak verideki uzun süreli ilişkileri yakalayabilir. LSTM+Transformer yapısı, zaman serisi tahmininde daha güçlü ve genellenebilir sonuçlar elde etmemizi sağladı. Önerilen modelin başarısı, tam bağlantılı katmanların entegrasyonu ile artırılmıştır. Bu katmanlar, farklı zaman ölçeklerinde öğrenilen temsilleri birleştirir.

B. Özyinelemeli tahmin yaklaşımı

Önerilen model, özyinelemeli yaklaşım kullanarak 3 günlük tahminler yapar. Bu tahminleri, önceki kendi tahminlere dayanarak ardışık olarak yapmaktadır. Bu sayede, uzun vadeli öngörü doğruluğu yükselir. Bu çalışmada, tahmin edilen çıktılar bir sonraki zaman adımına girdi olarak kullanılır. Bu, tek adımlı tahmin yöntemlerinden farklı bir modelleme stratejisidir. Özyinelemeli tahminle, hava kirliliği sistemlerinde daha kararlı ve esnek bir model geliştirildi. Bu yöntem, değişkenliği yüksek olan sistemler için önemlidir.

C. Optuna ile model optimizasyonu

Hiperparametrelerin sistematik olarak optimize edilmesi için Optuna tabanlı bir yöntem uygulanmıştır. Bu süreçte, öğrenme oranı, dropout oranı, gizli katman sayısı ve pencere uzunluğu gibi önemli parametreler en uygun kombinasyonla ayarlandı. Böylece modelin tahmin başarısı yükseltildi. Ayrıca, aşırı öğrenme (overfitting) riskini azaltmak ve modelin genel doğruluğunu artırmak için eğitim sürecinde çeşitli optimizasyon teknikleri kullanılmıştır. Modelin genelleme yeteneği, Batch Normalization, Dropout ve AdamW gibi gelişmiş yöntemlerle artırılmıştır.

D. Hava kirliliği tahmin uygulaması geliştirilmesi

Geliştirilen hava kirliliği tahmin uygulaması ile hava kirliliği tahminlerini istasyonlara göre harita üzerinde görsel olarak sunmaktadır. Ayrıca, tahmin sonuç grafikleri kullanıcı dostu şekilde sunularak, öngörü sonuçlarının anlaşılabilirliği arttırılmıştır. Uygulamanın modüler tasarımı sayesinde, farklı veri kümeleri ve yeni modelleme yaklaşımlarını kolaylıkla entegre etme imkanı sunar. Önerilen tahmin modeli, modelin performans

başarısını artırmanın yanı sıra, bir Streamlit arayüzüyle son kullanıcı için kullanılabilir hale getirilmiştir.

E. Gerçek zamanlı öngörü yeteneği

Önerilen modele, gerçek zamanlı tahmin altyapısı entegre edilmiştir. Bu entegrasyon, sürekli veri akışlarını işleyecek biçimde tasarlanmıştır. Bu sayede, hava kalitesindeki ani değişimlere karşı hızlı tahminler üretebilmektedir. Böylece modelin uygulanabilirliği arttırılmıştır. Önerilen model, anlık karar almayı gerektiren alanlarda etkili bir şekilde kullanılabilir. Bu alanlar, halk sağlığı uyarı sistemleri, çevresel izleme platformları ve yerel yönetim karar sistemlerini kapsamaktadır.

F. Akademik katkısı ve uygulama alanları

Geliştirilen modelin akademik düzeyde başarılı sonuçlar elde edilmesinin yanı sıra, aynı zamanda kamu ve özel sektör paydaşları için uygulanabilir karar destek aracı niteliği taşımasına katkı sağlanmıştır. Modelin katkı sağlayabileceği uygulama alanları

- Yerel yönetimler için bölgesel hava kalitesi planlamalarının veri temelli olarak yürütülmesi,
- Halk sağlığı yönetimi kapsamında, hassas ve risk altındaki grupların önceden bilgilendirilmesine yönelik erken uyarı sistemlerinin desteklenmesi,
- Kentsel sürdürülebilirlik politikaları çerçevesinde uzun vadeli hava kalitesi izleme altyapılarının geliştirilmesi,
- Kentsel hava kalitesi yönetimi süreçlerinde karar alma mekanizmalarını güçlendirecek etkin ve esnek bir karar destek sistemi olarak entegre edilebilmesi.

Bu kapsamda geliştirilen model, ileri düzey tahminleme yöntemlerini yalnızca kuramsal bir çerçevede bırakmayıp, bunları çevresel yönetimin pratik gereksinimleriyle bütünleştirerek; hem bilimsel bilgi üretimine hem de veri odaklı, akıllı ve sürdürülebilir kent yönetimi yaklaşımlarının güçlendirilmesine stratejik düzeyde katkı sağlamaktadır.

1.3 Hava Kirliliği Tahmin Modellerinin Önemi

Hava kirliliği tahmin modelleri, şehir yönetiminden sağlık hizmetlerine kadar birçok alanda kritik bir işlev görmektedir. Ayrıca, çevresel politikaların belirlenmesi ve sanayi

planlamasında da etkilidir. Hava kirliliğinin tahmin edilmesinin bazı önemli katkıları aşağıda açıklanmıştır.

Halk sağlığına olan etkisi açısından değerlendirildiğinde; hava kirliliği, solunum yolu hastalıkları, kardiyovasküler rahatsızlıklar ve erken ölümler gibi ciddi sağlık sorunlarıyla bağlantılıdır (WHO, 2021). Hava kirliliği öngörü sistemleri, bu zararlı etkilerin önceden belirlenmesini sağlayarak sağlık otoritelerine ve yerel yönetimlere koruyucu stratejiler geliştirme imkanı sunar (Manisalidis, Stavropoulou, Stavropoulos ve Bezirtzoglou, 2020). Özellikle toz parçacıkları (PM_{10}, $PM_{2.5}$) ve azot dioksit (NO_2) gibi kirleticiler, astım, KOAH ve kalp hastalıkları gibi kronik durumları olan bireyler için büyük tehlike arz etmektedir. Bu modeller aracılığıyla, hassas bireyler (çocuklar, yaşlılar ve bağışıklık sistemi zayıf kişiler) uyarılabilir ve dış alan maruziyetleri azaltılabilir.

Endüstriyel verimlilik ve ekonomik etkinliğe katkısı bakımından incelendiğinde; öngörü kabiliyeti yüksek hava kirliliği tahminleri, sanayi işletmelerinin emisyon kontrol stratejilerini iyileştirmelerine imkan tanır. Bu durum hem çevresel yükü azaltır hem de enerji verimliliğini artırır. Ayrıca, çalışma günlerinde hava kalitesine bağlı üretim aksaklıkları ve iş gücü kayıpları da en aza indirilebilir (Chang, Graff Zivin, Gross ve Neidell, 2019).

Şehir planlaması ve çevresel sürdürülebilirlik bakımından katkısını inceleyecek olursak; hava kirliliği tahmin sistemleri, sürdürülebilir şehirleşme stratejilerinin belirlenmesinde başlıca veri kaynağıdır. Tahminleme modelleri sayesinde, kirletici kaynakların mekânsal ve zamansal dağılımları analiz edilerek, düşük emisyon bölgeleri (Low Emission Zones - LEZ) planlanabilir ve ulaşım politikaları daha etkin biçimde şekillendirilebilir (Doval-Miñarro, Bueso ve Guillén-Alcaraz, 2025). Ayrıca, iklim değişikliğiyle mücadelede hava kalitesi tahminleri, karbon yönetimi politikalarının etkili biçimde izlenmesini sağlamaktadır.

Her ne kadar üç günlük tahmin aralığı oldukça kısa bir zaman dilimini kapsasa da, bu süre zarfında elde edilen yüksek doğruluklu hava kirliliği tahminleri, çevresel sürdürülebilirlik ve kentsel planlama stratejileri açısından önemli bir rol oynamaktadır. Kısa süreli öngörüler, dinamik ve anlık politika geliştirme süreçlerinde esas girdi olarak değerlendirilmektedir. Mesela, öngörülen kirlilik artışı nedeniyle toplu taşımanın teşvik

edilmesi, motorlu araç kullanımının kısıtlanması ya da inşaat çalışmalarının geçici olarak durdurulması gibi acil durum eylem planları devreye alınabilir (WHO, 2021).

Ayrıca, üç günlük tahmin verileri zamanla birikerek uzun dönemli istatistiksel çalışmaların temelini oluşturur. Bu öngörüler sayesinde, kirlilik örüntülerinin mevsimsel, haftalık ve bölgesel eğilimleri tespit edilebilir ve buna dayanarak şehir odaklı sürdürülebilir ulaşım stratejileri oluşturulabilir. Örneğin, bir bölgede belirli zaman dilimlerinde yükselen PM_{10} düzeylerinin saptanması, düşük emisyon bölgelerinin (LEZ) sınırlarının tekrar gözden geçirilmesine ya da ağır taşıt trafiğinin kontrol altına alınmasına yönelik yapısal tedbirleri gündeme getirebilmektedir (Doval-Miñarro, Bueso ve Guillén-Alcaraz, 2025).

Bununla birlikte, hava kalitesi modelleri iklim değişikliğiyle mücadelede karbon emisyonlarının dolaylı bir belirtisi olarak da önem taşımaktadır. Kısa süreli hava kirliliği verileri, karbon salımıyla ilişkili NOx, CO gibi kirleticilerin zamansal yoğunluğunu açığa çıkararak sera gazı azaltım stratejilerinin etkilerini günlük ölçekte takip etme imkanı sağlamaktadır.

1.4 Hibrit Modellerin Avantajları ve Hava Kalitesi Tahminindeki Rolü

Geleneksel istatistiksel modelleme, hava kalitesi gibi çok değişkenli, zamana bağlı ve istasyonlar arası heterojenliğe sahip karmaşık sistemlerde belirli başarılar sağlamıştır. Fakat, bu yöntemler doğrusal olmayan ilişkileri ve zamana bağlı değişkenlikleri tam olarak yakalayamıyor. Bu durum önemli bir sınırlılık olarak karşımıza çıkmaktadır. Bu nedenle, son yıllarda makine öğrenimi (Machine Learning, ML) ve özellikle derin öğrenme (Deep Learning, DL) temelli yaklaşımlar, hava kalitesi tahmininde artan bir şekilde kullanılmaya başlanmıştır (Liao, Zhu, Wu, Pan, Tang ve Wang ,2020). Fakat, tek başına uygulanan bir algoritmanın tüm veri kalıplarını ve karmaşık etkileşimleri tespit etmesi çoğu zaman yetersiz kalmaktadır. Bu çerçevede, hibrit modelleme yöntemleri, çeşitli yaklaşımların faydalarını birleştirerek daha güvenilir ve yaygın tahmin performansı sağlayan alternatif çözümler sunmaktadır.

a) Derin öğrenme yöntemleri ile zaman serisi tahmini: Zaman bağımlı hava kalitesi verilerinin modellenmesinde, özellikle Uzun Kısa Süreli Bellek (LSTM) ağları gibi

tekrarlayan sinir ağı (RNN) varyantları, zaman serisinin önceki değerlerini uzun vadeli bağımlılıklara göre öğrenebilme kabiliyetleri sayesinde kayda değer başarılar kazanmıştır. LSTM hücreleri, bilgi akışını yöneten giriş, çıkış ve unutma kapıları sayesinde gradyan kaybolması (vanishing gradient) sorununu hafifletmekte ve hem kısa vadeli hem de uzun vadeli desenleri etkili bir şekilde yakalayarak tahmin doğruluğunu yükseltmektedir (Hochreiter ve Schmidhuber, 1997). Sun ve arkadaşları (2022) gerçekleştirdiği çalışma ile**,** $PM_{2.5}$ ve O_3 tahmininde kullanılan istasyon merkezli LSTM modelini yayınlama katmanı ile genişleterek hem lokal hem de bölgesel düzeyde performansı %20–40 oranında yükseltmiştir. Bu yöntem, zaman bağımlılıklarını daha hassas biçimde yakalamasını sağlamıştır (Sun, Fung, Chen, Li, Yuan, Chen ve Lu, 2022).

b) Dikkat mekanizmaları ve transformer temelli yapılar: Klasik Tekrarlayan Sinir Ağları (Recurrent Neural Network, RNN) yapılarına ait kısıtlamaları geçmek için geliştirilen dikkat (attention) temelli yapılar, özellikle Transformer Encoder yapıları, zaman serisi modellemesinde metodolojik bir değişime öncülük etmiştir. Bu yapılar, zaman içindeki dikkate değer kalıpları ağırlıklandırarak öğrenme sürecini kuvvetlendirmektedir (Vaswani, Shazeer, Parmar, Uszkoreit, Jones vd., 2017). Fang ve arkadaşları (2022), bu dikkat mekanizmasını hem zamanla hem de mekânla ilgili kalıpları bir arada modelleyen hibrit bir spatiotemporal yapıya entegre ederek, $PM_{2.5}$ tahmininde dikkat farkındalığına sahip (attention-aware) bir çözüm geliştirmiştir. Bu yapı, çeşitli istasyonlar arasındaki ilişkileri de hesaba katarak modelin genelleme kapasitesini yükseltmiştir (Fang, Li, Karimian ve Mo, 2022).

c) Hibrit yapılar ile performans ve genelleme kapasitesi: Hibrit modelleme yöntemleri, istatistiksel modellerin açıklanabilirliğini ve derin öğrenme modellerinin güçlü tahmin yeteneğini bir araya getirir. Bu sayede daha esnek ve etkili sistemler ortaya çıkar. Bu çerçevede, LSTM ve Transformer Encoder bileşenlerinin bir araya getirildiği hibrit yapılar, zaman serisinin karmaşık yapısını çözümlemede yüksek başarı yakalamaktadır (Liu, Tee, Lu, Zhou ve Lu, 2025). Bu çalışma kapsamında önerilen hibrit yapı, özyinelemeli tahmin (recursive prediction) yaklaşımını benimseyerek çok adımlı tahmin yeteneği kazandırmaktadır. LSTM ve Transformer Encoder bileşenlerini birleştirmek, modelin kısa vadeli çıktısını bir sonraki zaman

diliminin girişi olarak kullanmasını sağlar. Bu sayede çok adımlı tahmin elde edilir. Böylece hava kalitesinin zaman içindeki gelişimini daha gerçekçi şekilde yakalayabilir. Kheder ve arkadaşları tarafından önerilen *AQ-Net* modeli, LSTM ve çok başlıklı dikkat içeren hibrit yapıları hem gözlemlenen hem gözlemlenmeyen istasyonlara yönelik $PM_{2.5}$ yeniden değerlendirmesini gerçekleştirmişlerdir. Bu model, tüm zaman ufuklarında kararlı ve tutarlı bir modelleme sunmuştur (Kheder, Foreback, Wang, Liu ve Boy, 2025).

1.5 Sınırlılıklar

Bu çalışmada kullanılan gerçek zamanlı PM_{10} verileri bazı sorunlar içerebilir. Sensör hataları, iletişim kopmaları ve kalibrasyon uyumsuzlukları yüzünden eksik ve tutarsız gözlemler barındırabilir. MICE yöntemi gibi çoklu veri tamamlama teknikleriyle bu boşluklar doldurulsa da tamamlanan verinin orijinal dağılımı tam yansıtamaması, özellikle ani sıçramalar ve uç değerlerde model doğruluğunu kısmen kısıtlamıştır.

Bu çalışma, Ankara ili ve belirli bir zaman aralığına (1.10.2019-19.12.2023 tarihleri arası) ait veriler kullanıldığı için modelin farklı iklim değişkenlerine veya diğer şehirlerdeki meteorolojik şartlara doğrudan uygulanabilirliği bu veriler ile sınırlıdır. Ayrıca model eğitimi sırasında erken durdurma ve dropout yöntemleri uygulanmış olmasına rağmen, modelin farklı veri setlerine ne oranda genelleme yapabileceği belirsiz kalmıştır.

Hibrit LSTM + Transformer Encoder mimari yapısı, çok başlıklı dikkat mekanizmaları ve derin LSTM blokları sebebiyle yüksek hesaplama ve bellek kaynağı gereksinimi duymaktadır. Sınırlı GPU süreleri ve mevcut donanım nedeniyle kapsamlı hiperparametre taramaları ve alternatif mimari varyant denemelerini kısıtlı bir düzeyde gerçekleştirilmiştir.

Hiperparametre ayarlaması için kullanılan Optuna kütüphanesi ile yalnızca belirli hiperparametre kombinasyonları denenmiştir. Tüm kombinasyonları incelemek pratik olmadığından modelin performansı, seçilen alt küme konfigürasyonuna dayanır. Bu nedenle, olası olarak daha iyi sonuçlar elde edebilecek konfigürasyonların olup olmadığı hala belirsizdir.

1.6 Tezin Kurgusu

Bu tez, toplam sekiz bölümden oluşmaktadır. İlk bölümde çalışmanın amacı, önemi, katkıları ve sınırlılıkları ele alınmıştır. İkinci bölümde, hava kirliliği tahmini üzerine yapılan önceki çalışmalar kapsamlı bir şekilde incelenmiştir. Üçüncü bölümde, veri toplama, ön işleme ve normalizasyon süreçleri detaylandırılmıştır. Dördüncü bölümde, detaylı veri analiz keşfi görselleri yer almıştır. Beşinci bölümde, önerilen hibrit modelin geliştirilmesi, algoritmik yapısı ve optimizasyon süreçleri açıklanmıştır. Altıncı bölümde, hava kirliliği tahmin uygulaması ve gerçek zamanlı tahmin sistemi açıklanmıştır. Yedinci bölümde, modelin eğitim süreçleri ve test bulguları ile performans değerlendirmesi ortaya konulmuştur. Sekizinci bölümde, araştırma bulguları derlenmiş ve gelecek çalışmalara dair önerilerde bulunulmuştur.

2. LİTERATÜR ARAŞTIRMASI

Hava kirliliği tahmini, atmosferdeki kirleticilerin zaman ve mekâna bağlı dağılımının öngörülmesiyle nasıl yayıldığını gösterir. Bu, çevresel riskleri azaltmak ve halk sağlığı stratejileri oluşturmak açısından kritik öneme sahiptir. PM_{10} konsantrasyonlarının öngörülmesine yönelik literatürde, geleneksel makine öğrenmesi algoritmalarının yanısıra derin öğrenme yöntemlerinin de etkin şekilde kullanıldığı görülmektedir. Son yıllarda derin öğrenme ve transformer model gibi yöntemlerin kullanılması yaygınlaşmıştır. Hava kirliliği tahmininde öncelikle geleneksel makine öğrenmesi yöntemlerine, sonrasında derin öğrenme ve hibrit modelleri uygulayan çalışmalar incelenmiştir.

Box ve Jenkins (1976) tarafından geliştirilen Otomatik Regresif Bütünleşik Hareketli Ortalama (AutoRegressive Integrated Moving Average, ARIMA) modeli, hava kirliliği zaman serilerinin geçmiş verilerine dayanarak tahmin yapılmasına olanak tanıyan temel yöntemlerden biridir. Zaman serisi analizinde sıkça kullanılan model, zaman serisinin geçmiş değerleri (AutoRegression, AR), geçmiş hata terimlerini (moving average - MA) ve fark alma işlemlerini (integration , I) kullanarak durağanlaştırılmış verilere dayanır. Ancak çevresel ve meteorolojik olaylar gibi karmaşık ve çok değişkenli süreçlerin modellenmesinde yetersiz kalmaktadır (Fenger, 1999).

SARIMA (Mevsimsel Otoregresif Bütünleşik Hareketli Ortalama) modeli, Box ve Jenkins (1976) tarafından geliştirilen klasik ARIMA modelinin bir uzantısıdır. Bu model, zaman serilerindeki mevsimsel döngüleri modellemek için tasarlandı. ARIMA yapısına mevsimsel parametrelerin (P, D, Q, s) eklenerek oluşturulmuştur. SARIMA modeli, hem doğrusal bağımlılıkları hem de mevsimsel örüntüleri içerir. Bu model, özellikle $PM_{2.5}$ ve PM_{10} gibi hava kirleticilerin tahmininde iyi bir performans göstermektedir. SARIMA, hem doğrusal bağımlılıkları hem de mevsimsel kalıpları olan zaman serilerinde tahmin başarısını büyük ölçüde artırmıştır.

Regresyon (Multiple Linear Regression, MLR) modeli, hava kirliliği ile meteorolojik değişkenler arasındaki doğrusal ilişkileri modellemek için kullanılan istatistiksel bir

yöntemdir. Bu yöntemde, bir bağımlı değişken (PM_{10} vb.) birden fazla bağımsız değişken (örneğin sıcaklık, rüzgar hızı, nem vb.) ile doğrusal ilişki içinde modellenmiştir. Ancak Fenger (1999) belirtmiştir ki, çoklu doğrusal regresyon karmaşık atmosferik süreçleri ve doğrusal olmayan etkileşimleri yakalamakta yetersiz kalmıştır. Bu yüzden, özellikle hava kalitesi gibi çok değişkenli ve karmaşık sistemlerde sınırlı başarısı ile bilinmektedir.

Destek Vektör Makineleri (Support Vector Machine, SVM), Cortes ve Vapnik tarafından 1995'te geliştirilmiş olup, küçük boyutlu veri kümelerinde bile yüksek performans gösterebilmesiyle öne çıkar. Bu algoritma, doğrusal olmayan sınıflandırma ve regresyon problemlerinde etkilidir. Kernel dönüşümleri sayesinde karmaşık ilişkileri modelleyebilme yeteneğine sahiptir.

Moursi ve arkadaşları tarafından 2019'da, bu algoritmaların karşılaştırmalı performansını detaylı biçimde incelemiştir. Bu çalışmada, $PM_{2.5}$ konsantrasyonunun kısa vadeli tahmini için bazı algoritmalar değerlendirildi. Destek Vektör Regresyonu (SVR), Rastgele Orman, LSTM ve Extra Trees algoritmaları karşılaştırıldı. Sonuçlar, Extra Trees algoritmasının diğer yöntemlere göre daha iyi tahmin doğruluğu sağladığını göstermektedir. Bu sonuç, model karmaşıklığı ve genelleme başarısı arasındaki dengeyi inceleyen araştırmalar için dikkat çekicidir. Ayrıca çalışmada SVR'nin düşük örneklem hacimlerinde dahi tutarlı sonuçlar verebildiği bulunmuştur. RF ise çeşitli veri tiplerine karşı dayanıklılığını göstermiştir (Moursi, Shouman, Hemdan ve El-Fishawy , 2019).

Derin öğrenme yöntemlerinden LSTM, Hochreiter ve Schmidhuber (1997) tarafından geliştirilmiş olup, zaman serilerindeki uzun vadeli bağımlılıkları yakalamadaki yeteneği nedeniyle hava kalitesi modellemesinde öne çıkmıştır. Dikkat mekanizmaları ve LSTM katmanlarıyla bir araya geldiğinde, sıralı bilgilerin yerel ve küresel bağlamda işlenmesini sağlamaktadır. Bu yaklaşım, Bahdanau, Cho ve Bengio tarafından önerilen "attention-based encoder-decoder" mimarisiyle ilk kez gündeme gelmiştir (Bahdanau, Cho ve Bengio, 2015).

Seng ve diğerleri (2021) yaptığı çalışmada, Uzun Kısa Süreli Bellek (LSTM) tabanlı geliştirilen, çok çıktılı ve çoklu denetimli öğrenme endeksi (MMSL) yapısını içeren kapsamlı bir hava kalitesi tahmin modeline odaklanmışlardır. Model kapsamında, Pekin'deki mevcut izleme istasyonlarından elde edilen $PM_{2.5}$ partikül madde

konsantrasyon verileri, komşu istasyonlardan sağlanan meteorolojik parametreler ve hava kirletici gaz verileri (çoğunlukla CO, NO olmak üzere NO_2, O_3 ve SO_2) bir araya getirilerek modelin veri kaynağı oluşturulmuş. MMSL yaklaşımıyla geliştirilen modelin tahmin performansı değerlendirme metrikleri ile ölçülmüş; önerilen MMSL modeli, doğrusal regresyon, Rastgele Orman, Destek, SVM ve güncel modeller olan CNN-LSTM ve CNN-Bi-RNN ile karşılaştırılmıştır. Karşılaştırma sonuçları ile önerilen MMSL modelinin üstün tahmin performansı gösterdiği raporlanmıştır (Seng, Zhang, Zhang, Chen ve Chen, 2021).

Bir başka çalışmada ise Yan ve arkadaşları, Pekin kentine yönelik, uzay-zamansal kümeleme temelli derin öğrenme ağlarını kullanarak çok zamanlı ve çok merkezli bir hava kalitesi tahmin modeli geliştirmiştir. Bu kapsamda performansları geri yayılım sinir ağı (BPNN), evrişimli sinir ağı (CNN), LSTM ve CNN-LSTM mimarileri kullanılarak karşılaştırılmıştır. Özellikle bir sonraki saatlik tahmin (t+1) doğruluğu üzerine odaklanan çalışmanın sonuçları, hem CNN-LSTM hem de LSTM modellerinin genel olarak LSTM ve CNN-LSTM modellerinin diğer yöntemlere kıyasla daha iyi performans göstermiştir (Yan, Liao, Yang, Sun, Nong ve Li, 2021).

Muthukumar ve arkadaşları yaptığı çalışmada, meteorolojik, çevresel ve uydu görüntülerinden elde edilen çok kaynaklı verilerin entegrasyonunu hedefleyerek, Graph Convolutional Network (GCN) ve Convolutional LSTM (ConvLSTM) yöntemlerini uygulamışlardır. Bu çalışmada, veri türlerini çeşitlendirmek ve derin ağ mimarilerini birleştirmekle tahmin doğruluğunun anlamlı şekilde arttığı belirtilmiştir.Çok boyutlu veri setlerinin birleştirilmesi ve gelişmiş derin öğrenme modellerinin bir arada kullanılması, tahmin performansını önemli ölçüde artırmıştır (Muthukumar, Pathak, Nagrecha, Hosseini, Comer ve Amini, 2020).

Yang ve Ismail (2022) ise tahmininde transfer öğrenme ile desteklenen hibrit bir derin öğrenme modeli önererek, karmaşık ve yüksek boyutlu veri kümelerinin modellenmesine yönelik olarak LSTM ve Çok Katmanlı Algılayıcı (MLP), CNN ve transfer öğrenmesine dayalı yapılarının entegrasyonunu gerçekleştirmiştir. Modelin, Pekin ve Hengshui hava kalitesi verileriyle doğrulanmış, yüksek doğruluğun yanı sıra güçlü bir kararlılığa sahip olduğu sunulmuştur.

Yang ve arkadaşları çalışmalarında, Pekin'e ait veri eksikliği sorunuyla başa çıkabilmek amacıyla kanal ve mekansal dikkat mekanizmaları ekledikleri transfer öğrenme temelli hibrit derin öğrenme modelini geliştirmiştir. Önerien modelde LSTM ve Çok katmanlı algılayıcı (MLP) birlikte kullandıkları ağ yapısına tahmin doğruluğunu arttırmak için artık sinir ağ birimi eklenmiştir. Bulgulara göre, klasik derin öğrenme modelleri, hibrit derin öğrenme modelleri ve en güncel transfer öğrenme yaklaşımlarıyla karşılaştırıldığında, TL-Modified modelinin RMSE değeri, TL-LSTM ve TL-CNN-LSTM modellerine kıyasla farklı istasyonlarda sırasıyla %38, %16.5 ve %25.6 oranında azaltıldığı gösterilmiştir (Yang, Wang, Liu ve Zhang, 2023).

Zhang, Zheng ve Yu (2023), hava kalitesi tahmininde açıklanabilirliği artırmak amacıyla dikkat mekanizması (attention) ile güçlendirilmiş LSTM (Attention-LSTM) modelini önermiştir. Çalışmada, dikkat mekanizması kullanılarak modelin zaman serisi verilerindeki önemli zaman noktalarına daha fazla ağırlık vermesi sağlanmış ve böylece $PM_{2.5}$ ve PM_{10} seviyelerinin daha isabetli tahmini hedeflenmiştir. Model, yalnızca doğruluk açısından değil, aynı zamanda hangi girdilerin ve zaman dilimlerinin tahmine daha fazla katkı sağladığını açıklayabilme potansiyeli ile de öne çıkmıştır. Önerilen STA-ResCNN modeli, LSTM,GRU,CNN ve TA-ResCNN yöntemleri ile kıyaslandığında, $PM_{2,5\ ve}$ PM_{10} için üç büyük şehirde 1-4 saat öncesi tahminlerin ortalaması için ortalama karekök hatasını sırasıyla %5, 595-%15, 247 ve %6, 827-%16, 906 oranlarında azalmıştır. Yapılan deneyler, Attention-LSTM modelinin klasik LSTM ve diğer derin öğrenme yaklaşımlarına kıyasla hava kalitesi tahmininde daha yüksek performans sergilediğini göstermiştir.

Jesupriya, Mahalakshmi ve Komalavalli (2025) çalışmalarında, hava kalitesi tahmini için çok başlı dikkat mekanizmalı, çift yönlü LSTM'lere ve yoğun katmanlara dayalı bir hibrit derin öğrenme modeli önermiştir. Model, hem farklı kirletici konsantrasyonlarını tahmin etmeyi hemde hava kalitesi indeksini hesaplamayı amaçlamış. Hibrit CNN-LSTM modeli, girdi verilerinin mekansal özelliklerini çıkarırken CNN katmanları önemli rol oynar. LSTM ise zaman serisi verilerinin zamansal dinamiklerini öğrenir. Ayrıca, modelin dikkat mekanizması eklenerek önemli zaman adımlarına ve kritik girdilere daha fazla ağırlık vermesi sağlanmıştır. Bu sayede, hem kirletici konsantrasyonlarının tahmininde hemde AQI sınıflandırmasında yüksek doğruluk elde edilmiştir.

Hava kalitesi tahmininde mekânsal ve zamansal bağımlılıkları etkili bir şekilde modellemek için grafik sinir ağları (Graph Neural Networks, GNN) daha fazla bir ilgi çekiyor. Wang ve arkadaşları, kritik meteorolojik ve coğrafi bilgileri kullanan $PM_{2.5}$-GNN modelini geliştirdi. Bu, zenginleştirilmiş bir grafik sinir ağı mimarisi sunmaktadır. Model, düğüm ve kenar özniteliklerini meteorolojik ve coğrafi verilerle destekler. Böylece, ince çözünürlüklü mekânsal detayları ve uzun vadeli bağımlılıkları etkili bir şekilde yakalamaktadır. Gerçek dünya veri setleri üzerinde yapılan deneyler, $PM_{2.5}$-GNN'nin ince detayları ve uzun dönem korelasyonları iyi bir şekilde modellediğini ortaya koymuştur. Bu özellikler, onun alanındaki en iyi performansı sergilediğini göstermiştir. Ancak, bu modelin uygulaması zordur. Bunun nedeni, $PM_{2.5}$'ye ait birçok parametre ve karmaşık ön işleme adımları gerektirmesidir ve bu durum pratikte kullanımını kısıtlayabilir (Wang, Li, Zhang, Meng, Meng ve Gao , 2020).

Kıyaslanılan diğer bir model olan Group-Aware Graph Neural Network (GAGNN), Chen ve arkadaşları tarafından 2021'de ülke çapında şehirler arası hava kalitesi tahmini için geliştirilmiştir. GAGNN, şehirler arasındaki gizli bağımlılıkları hiyerarşik bir grafik yapısı üzerinden tespit etmektedir. Model, önce bireysel şehirleri düğüm olarak içeren bir şehirler grafiği oluşturur ve ardından benzer şehirleri biraraya getirerek üst düzey bir grup grafiği hazırlar. Kendinden öğrenimli gruplayıcı ağ (self-supervised clustering network) sayesinde benzer özelliklere sahip şehirler dinamik bir şekilde gruplandırılmakta ve bu gruplarlar arasındaki korelasyonlar bir kodlayıcı aracılığıyla modellenmektedir. Mesaj geçişi (message passing) mekanizması sayesinde, coğrafi olarak uzak olan ancak benzer hava kirliliği profiline sahip şehirler etkili bir şekilde ilişkilendirilmektedir. Deneysel sonuçlar, GAGNN'nin mevcut tüm karşılaştırmalı modellere nazaran en yüksek tahmin doğruluğunu sunduğunu göstermiştir. Modelin en büyük dezavantajı, ek gruplayıcı katmanlar ve grup sayısı gibi hiperparametrelerin optimizasyon ihtiyacı nedeniyle mimarinin karmaşık hale gelmesidir (Chen, Xu, Wu, Qian, Du, Li ve Zhang, 2021).

Güncel çalışmalarda kullanılmaya başlayan Transfer öğrenme modeli sayesinde genellikle veri kıtlığı sorunu üzerinde çalışılmaktadır. Dhole ve arkadaşları tarafından yapılan çalışmada, çok kaynaklı transfer öğrenme yaklaşımlarını değerlendirerek, çeşitli kaynak istasyonlardan elde edilen verilerin hedef istasyona aktarılması yoluyla kümülatif bir tahmin oluşturmuştur. Bu yöntemle, yakın istasyonlardan kolayca erişilen verilerden

faydalanılarak tahmin performansının artırılması istenmiştir (Dhole, Ambekar, Gunjan ve Sonawani, 2021).

Liang ve arkadaşları, Çin genelinde ki 1.085 izleme istasyonundan toplanan dört yıllık verileri kullanan ve özellikle $PM_{2.5}$ parametrelerine odaklanarak kentsel hava kalitesini tahmin etmek için transformer tabanlı bir yaklaşım sağlayan AirFormer modelini geliştirdiler. Bu çalışmada, iki aşamalı hibrit bir yapı geliştirilmiştir. İlk aşama, "Dartboard Spatial MSA" ve "Causal Temporal MSA" olarak adlandırılan iki yeni öz-dikkat mekanizması ile uzamsal ve zamansal bağımlılıkları etkili bir şekilde öğrenen deterministik aşamasıdır. İkinci aşamada ise, tahmin belirsizliğini modellemek için gizli rastgele değişkenler (latent değişkenler) eklenerek stokastik bir aşama oluşturulmuştur. Bu yapı, dinamik uzamsal ilişkilerin yanı sıra uzun vadeli belirsizliklerin de etkili bir şekilde ele alınmasına olanak tanımaktadır. Deneysel sonuçlar, AirFormer'ın 72 saatlik tahminlerde mevcut yöntemlere göre hata oranını %5–8 oranında düşürebildiğini ortaya koymaktadır. Literatürdeki diğer çalışmalara göre, uzun vadeli tahminlerde yüksek doğruluk sağlaması, değişken belirsizlikleri modelleyebilmesi ve büyük ölçekli veri setlerinde etkili bir şekilde uygulanabilmesi gibi avantajları bulunmaktadır (Liang, Li, Zhang, Wang, Liu ve Peng, 2023).

Yi ve arkadaşları 2018 yılında, $PM_{2.5}$ tahmini için tasarlanmış hibrit bir CNN-BiLSTM mimarisi olan DeepAir modelini önermiştir. Bu modelde, çevresel sensörlerden elde edilen çok değişkenli verilerden yerel eğilimleri ve mekânsal korelasyonları çıkarmak için tek boyutlu evrişimli sinir ağı katmanları kullanılmıştır. Sonrasında, BiLSTM katmanı, bu özellikler aracılığıyla zaman içindeki bağımlılıkları öğrenerek daha kapsamlı bir zaman serisi analizi sunmuştur. Son katmanlarda elde edilen özellikler birleştirilerek $PM_{2.5}$ konsantrasyonu tahmini yapılmaktadır. Gerçek dünya veri setleri üzerinde gerçekleştirilen deneysel çalışmalar, DeepAir modelinin yüksek doğrulukla tahminler yaptığını ortaya koymuştur. Modelin en önemli avantajı, CNN katmanları ile yerel uzamsal desenleri, BiLSTM ile de zamansal bağımlılıkları eş zamanlı olarak öğrenebilmesi ve bu sayede daha zengin ve anlamlı özellikler elde etmesidir. Ancak, modelin dezavantajı, meteorolojik verilerin ek veri kaynaklarıyla birleştirilmemesi ve tahminlerin çoğunlukla istatistiksel korelasyonlara dayalı olmasıdır (Yi, Zhang, Wang, Li ve Zheng, 2018).

3. VERİ HAZIRLAMA VE ÖN İŞLEME

Veri ön işleme ve veri hazırlama süreçleri, modelin öğrenme performansını artırmakta ve hava kirliliği tahmininin doğruluğunu yükseltmektedir. Yüksek kaliteli bir veri seti oluşturmak, hava kirliliği tahmininin başarısını ve modelin genelleştirme yeteneğini olumlu yönde etkilemektedir. Bu bölümde, veri kaynaklarının özellikleri, veri temizleme ve normalizasyon yöntemleri, ayrıca zaman serisi analizine ilişkin detaylar ele alınmıştır.

3.1 Çalışma Alanı

Çalışma alanı olarak, Türkiye'nin başkenti Ankara seçilmiştir. Ankara, 5.782.285 kişilik nüfusu ile Türkiye'nin en kalabalık ikinci ili olup, Orta Anadolu'nun kuzeybatısında yer almaktadır. Coğrafi olarak Kızılırmak ve Sakarya nehirlerinin kolları arasında bulunan şehir, 39.57°K enlemi ve 32.53°D boylamı koordinatlarında konumlanmaktadır. 26.897 km² yüzölçümü bakımından Türkiye'nin üçüncü büyük ili olan Ankara'nın, 25 ilçesi bulunmaktadır. Bu ilimiz, coğrafi ve iklimsel koşullar bakımından önemli özellikler taşımaktadır. Karasal iklimin hâkim olduğu Ankara'da, yıllık sıcaklık farkları belirgindir ve kışlar soğuk, yazlar ise sıcak geçmektedir.

Ankara, yüzölçümünün büyük bir kısmında bozkır bitki örtüsüne sahipken, il merkezinde yoğun bir yerleşim alanı bulunmaktadır. Nüfus yoğunluğunun büyük çoğunluğu merkez ilçelerde toplanmıştır. Ancak çevre ilçelerdeki yoğun tarımsal faaliyetler ve son yıllarda gelişen termal turizm ve kültür turizmi gibi ekonomik faaliyetler, hava kalitesini etkileyen faktörler arasında yer almaktadır. Şekil 1'de Ankara ilinin fiziki haritası görsel olarak sunulmaktadır (MSB HGM, 2025).



Şekil 1. Ankara ilinin fiziki haritası (MSB HGM, 2025)

Çalışmada kullanılacak veri seti, Ankara ilinin çeşitli bölgelerine ait hava kalitesi, meteoroloji istasyonlarından erişilen verilere dayanacaktır.

3.2 Veri Kaynakları

Bu çalışmada, hava kirliliği tahmin modellerinin oluşturulmasında kullanılacak veri seti, Ankara ilindeki hava kalitesi ve meteorolojik parametrelerin çok değişkenli bir zaman serisini içermektedir. Çalışmanın ana hedefi, özellikle PM_{10} konsantrasyon seviyesini tahmin etmektir. Bu hedef değişkenler, halk sağlığı açısından önemli olan hava kirliliği bileşenlerini temsil eder. Girdi değişkenleri ise hava kirliliği ve meteorolojik faktörlerden oluşmaktadır.

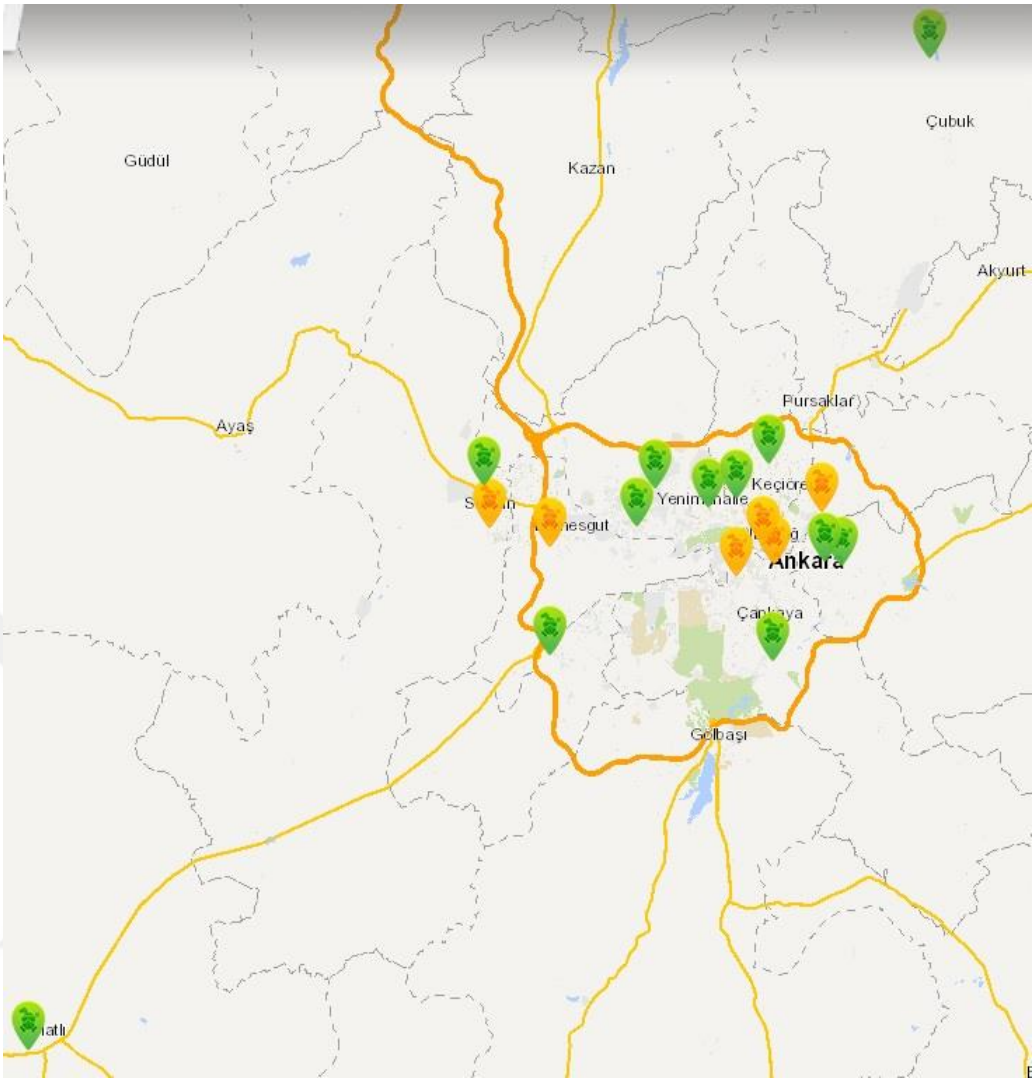
Veri setinin temel bileşenleri, hava kalitesi izleme istasyonları ve meteoroloji izleme istasyonları tarafından sağlanan verilerden oluşmaktadır. Bu veriler, Ankara ilinin hava kalitesine ilişkin kapsamlı bir veri analizi ve hava kirliliği tahminleri için modellerin eğitim sürecine temel oluşturacaktır.

3.2.1 Hava kalitesi izleme istasyon verileri

Türkiye Cumhuriyeti Çevre, Şehircilik ve İklim Değişikliği Bakanlığı'na bağlı hava kalitesi ölçüm istasyonları, hava kirliliği izleme ve değerlendirme süreçlerinde kritik bir işlev görmektedir. Bu tez kapsamında kullanılan hava kirliliğine ait istasyon verileri, T.C. Çevre, Şehircilik ve İklim Değişikliği Bakanlığı Ulusal Hava Kalitesi İzleme Ağı'ndan

(UHKİA) temin edilerek kullanılmıştır (UHKİA, 2023). Bu istasyonlar, hava kirliliğini izlemek ve çevresel yönetmeliklere uygun olarak düzenli aralıklarla hava kalitesini ölçmek için stratejik olarak yerleştirilmiştir. Çalışmada kullanılan veriler, Ankara ilindeki 18 farklı hava kalitesi izleme istasyonundan, 2019-2023 yılları arasında saatlik ölçülen kirleticilerin değerlerini içermektedir.

Hava kalitesi veri seti, şehirdeki hava kirliliği düzeylerini analiz etmek için toplandı. Bu veri seti, kirliliğin mekânsal ve zamansal değişimlerini incelemeye olanak tanır. Ölçüm istasyonlarından toplanan veriler, temel kirletici parametreleri gösterir. Bunlar arasında partikül madde (PM10, PM2.5), azot dioksit (NO_2), kükürt dioksit (SO_2), karbon monoksit (CO) ve ozon (O_3) bulunur. Bu parametreler, hava kirliliğini belirleyen ve insan sağlığına olumsuz etkileri olan bileşenlerdir. Ankara'daki hava kalitesi ölçüm istasyonlarının dağılımları, coğrafi konumları ile birlikte Şekil 2'de gösterilmiştir.



Şekil 2. İstasyon Bilgileri Haritası (SIM, 2024)

Hava kirliliği parametreleri:

PM10 (µg/m³) hava içindeki 10 mikron veya daha küçük partikül maddelerin konsantrasyonunu gösterir. Bu kirleticiler, trafik, sanayi ve inşaat faaliyetlerinden ortaya çıkar. Bu maddeler, solunum yolu hastalıklarına neden olabilir.

Partikül Madde ($PM_{2.5}$ (µg/m³)), 2.5 mikron veya daha küçük parçacıkların yoğunludur. $PM_{2.5}$ partikülleri, derin solunum yollarına kadar inebilir. Bu durum, kalp-damar hastalıkları ve kanser gibi ciddi sağlık sorunlarına neden olabilir.

Sülfür Dioksit (SO_2 (µg/m³)), fosil yakıtların yanmasıyla ortaya çıkmaktadır. Bu gaz, sanayi bölgelerinde sık görülür. Ayrıca, astım gibi solunum hastalıklarına neden olabilir.

Azot Dioksit (NO_2 ($\mu g/m^3$)), motorlu taşıtlardan ve sanayi emisyonlarından oluşur. Bu gaz, hava kalitesini kötüleştirir ve ozon tabakasına zarar verebilir.

Karbon Monoksit (CO ($\mu g/m^3$)), yetersiz yanma süreçlerinden oluşan bir gazdır. Vücutta oksijen taşımayı azaltır. Bu durum, baş ağrısı ve baş dönmesi gibi semptomlara yol açabilir.

Ozon(O_3 ($\mu g/m^3$)), fotokimyasal reaksiyonlar sonucu ortaya çıkan bir ikincil kirleticidir ve atmosferde bulunur. Ozon, solunum yollarını etkileyerek alerjik reaksiyonları tetikleyebilir.

Aşağıdaki Tablo 1'de, UHKİA'dan elde edilen veriler işlenerek Ankara il sınırları içerisinde konumlanan hava kalitesi izleme istasyonlarına ait kirletici parametrelerin dağılımı detaylı ve sistematik bir biçimde sunulmaktadır. İlgili görselleştirme, her bir istasyonda izlenen kirletici türlerinin kapsamını ve çeşitliliğini tabloda sunmakta; "✓" sembolü ile ölçüm yapılan parametreleri, "X" sembolü ise ölçümün gerçekleşmediği durumları göstermektedir.

Bu yaklaşım, istasyon bazında ölçüm kapasitesinin ve parametre çeşitliliğinin karşılaştırmalı analizine olanak tanıyarak, ölçüm altyapısındaki heterojenliği açık bir şekilde ortaya koymaktadır. Böylelikle, kirletici türleri açısından istasyonlar arası farklılıklar görünür kılanarak, özellikle eksik ölçüm alanlarının belirlenmesiyle, gelecekteki modelleme ve izleme stratejilerinde bu tespitlerin dikkate alınması açısından oluşturulmaktadır.

Tablo 1. Hava Kalitesi İstasyon Verileri

İstasyon Adı	**Hava Kirliliği Parametreleri**							
	PM10	***PM2.5***	***SO2***	***CO***	***NO2***	***O3***	***NOx***	***NO***
Bahçelievler	✓	✓	✓	✓	✓	X	✓	✓
Batıkent	✓	X	✓	X	✓	X	✓	✓
Demetevler	✓	✓	✓	X	✓	X	✓	✓
Etimesgut	✓	✓	✓	✓	✓	X	✓	✓
Etlik	✓	X	✓	✓	X	X	X	X
Kayaş	✓	X	✓	X	X	X	X	X
Keçiören	✓	✓	✓	X	✓	✓	✓	✓
Mamak	✓	X	✓	✓	✓	✓	✓	✓
Ostim	✓	✓	✓	✓	✓	✓	✓	✓
Polatlı	✓	X	✓	✓	✓	✓	✓	✓
Sıhhıye	✓	✓	✓	✓	✓	✓	✓	✓
Sincan	✓	✓	✓	X	✓	X	✓	✓
Siteler	✓	✓	✓	✓	✓	✓	✓	✓
Törekent	✓	✓	✓	✓	✓	X	✓	✓
Ulus	✓	✓	✓	X	✓	✓	✓	✓
Yaşamkent	✓	✓	✓	✓	✓	✓	✓	✓
Çankaya	✓	✓	X	X	✓	✓	✓	✓
Çubuk	X	X	✓	X	✓	✓	✓	✓

3.2.1.1 Hava kalitesi ulusal sınır değerleri

Türkiye'de hava kalitesinin izlenmesi ve yönetimi, insan sağlığını ve çevreyi korumak amacıyla geliştirilen yasal düzenlemeler çerçevesinde gerçekleştirilmektedir. Bu bağlamda, Hava Kalitesi Değerlendirme ve Yönetimi Yönetmeliği (HKDYY), ulusal hava kalitesi standartlarını belirleyen temel referans doküman olarak öne çıkmaktadır. Yönetmelik, 21 Mayıs 2008 tarihli ve Avrupa Birliği'nin (AB) 2008/50/EC sayılı "Hava Kalitesi ve Temiz Hava için Daha İyi Bir Avrupa" Direktifi ile uyumlu şekilde hazırlanmıştır. Yönetmelikte, belirli kirletici parametreler için ulusal sınır değerler, hedef değerler ve uzun vadeli hedefleri belirlenmiştir. Bu değerler, toplum sağlığını ve çevresel sürdürülebilirliği koruma amacıyla belirlenmiş olup, hava kirliliği yönetiminin temel dayanaklarından birini oluşturmaktadır.

Yönetmelikte tanımlanan sınır değerler, insan sağlığı ve ekosistemlerin korunması için kritik bir eşik olarak kabul edilmektedir. Bu değerler, Dünya Sağlık Örgütü'nün (WHO) önerdiği kılavuz değerler ve AB uyum süreçleri dikkate alınarak belirlenmiştir. Sınır değerlerin tanımlanmasında özellikle şu hususlar göz önünde bulundurulmuştur:

- İnsan sağlığı üzerindeki kısa ve uzun vadeli etkileri (solunum ve kardiyovasküler hastalık riskleri gibi hastalıklar),
- Su kaynakları, tarım ve ormanlar üzerindeki çevresel etkileri,
- Trafik, sanayi ve enerji sektörü gibi kirletici kaynaklardan gelen emisyonların kontrolü,
- Uluslararası çevre standartlarına uyumun sağlanması.

Bu bölümde, Türkiye'de geçerli olan temel hava kirleticilerine ilişkin sınır değerler ve bu değerlerin belirlenme amaçları ele alınmıştır.

a) Partikül Madde (PM_{10} ve $PM_{2.5}$)

Partikül madde (PM), havada bulunan ve çapı 10 mikrometre (μm) veya daha küçük olan katı ve sıvı parçacıkları ifade eder. PM_{10} ve $PM_{2.5}$, insan sağlığı üzerinde ciddi etkileri olan iki temel partikül madde türüdür. Bu kirleticiler, özellikle solunum sistemi hastalıkları ve kardiyovasküler rahatsızlıklara yol açabilmektedir. Yönetmelikte PM için belirlenen sınır değerler şu şekildedir:

- 24 saatlik ortalama: 50 µg/m³ (günlük aşım sayısı 35)
- Yıllık ortalama: 40 µg/m³

b) Kükürt Dioksit (SO2)

Kükürt dioksit (SO_2), fosil yakıtların yanması ve endüstriyel faaliyetler sonucu ortaya çıkan bir hava kirleticisidir. SO_2, solunum sistemi üzerinde tahriş edici etkiler yaratabilir ve asit yağmurlarına neden olarak çevresel zararlar oluşturabilir. Türkiye'de SO_2 için belirlenen sınır değerler şunlardır:

- 1 saatlik ortalama: 350 µg/m³ (yılda 24 kez aşılabilir) (günlük aşım sayısı 3)
- 24 saatlik ortalama: 125 µg/m³ (yılda 3 kez aşılabilir)

c) Azot Dioksit (NO2)

Azot dioksit (NO_2), motorlu taşıt emisyonları ve endüstriyel faaliyetler gibi kaynaklardan salınan bir kirletici gazdır. NO_2, solunum yolu enfeksiyonlarının artmasına ve akciğer fonksiyonlarının azalmasına neden olabilir. Yönetmelik kapsamında NO_2 için belirlenen sınır değerler şunlardır:

- 1 saatlik ortalama: 250 µg/m³
- Yıllık ortalama: 40 µg/m³

d) Karbonmonoksit (CO)

Karbonmonoksit (CO), genellikle yetersiz yanma süreçleri sonucu oluşan bir gazdır. CO, oksijen taşıyan hemoglobin moleküllerine bağlanarak vücutta oksijen eksikliği nedeniyle ciddi sağlık sorunlarına yol açabilmektedir. CO için belirlenen sınır değer şu şekildedir:

- 8 saatlik ortalama: 10 mg/m³

e) Ozon (O3)

Ozon (O_3), atmosferdeki azot oksitlerin (NOx) ve uçucu organik bileşiklerin (VOC) güneş ışığıyla fotokimyasal reaksiyona girmesi sonucu oluşan bir ikincil kirleticidir. Ozon, yüksek konsantrasyonlarda solunum yollarında tahrişe neden olabilir ve bitki örtüsüne zarar verebilir. Türkiye'de O_3 için belirlenen hedef değer şunlardır:

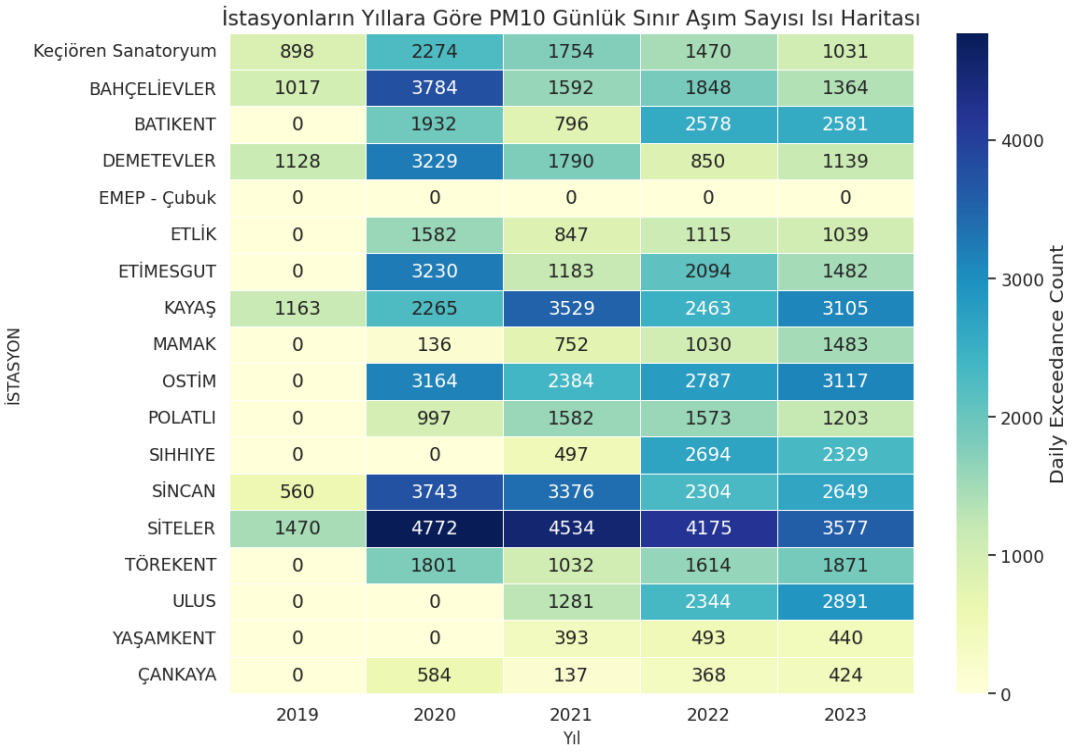
- 8 saatlik ortalama: 120 µg/m³

Bu sınır değerler, insan sağlığını korumak amacıyla belirlenmiştir. Hava kalitesi izleme istasyonları, bu kirleticilerin konsantrasyonlarını sürekli olarak izler ve ölçümler yapar. Yönetmelikte belirlenen sınır değerlerin aşılması durumunda, ilgili yetkililer tarafından gerekli önlemler alınır.

A. Sınır değer aşım analizi

Hava kalitesi istasyonlarındaki PM10 aşım sayılarının incelenmesi, hava kalitesi üzerindeki etkilerin değerlendirilmesine katkı sağlamaktadır. Bu analizler sonucunda elde edilen bulguların, merkezi ve yerel yönetimlerin kirlilik kaynaklarını kontrol etmek ve azaltmak için strateji geliştirilmesine yardımcı olabileceği değerlendirilmektedir.

T.C. Çevre, Şehircilik ve İklim Değişikliği Bakanlığı Ulusal Hava Kalitesi İzleme Ağı'ndan alınan, Ankara iline ait saatlik istasyon verileri işlenerek PM10 için günlük sınır aşım sayılarının analizi Şekil 3'te ısı haritası olarak yer almaktadır.



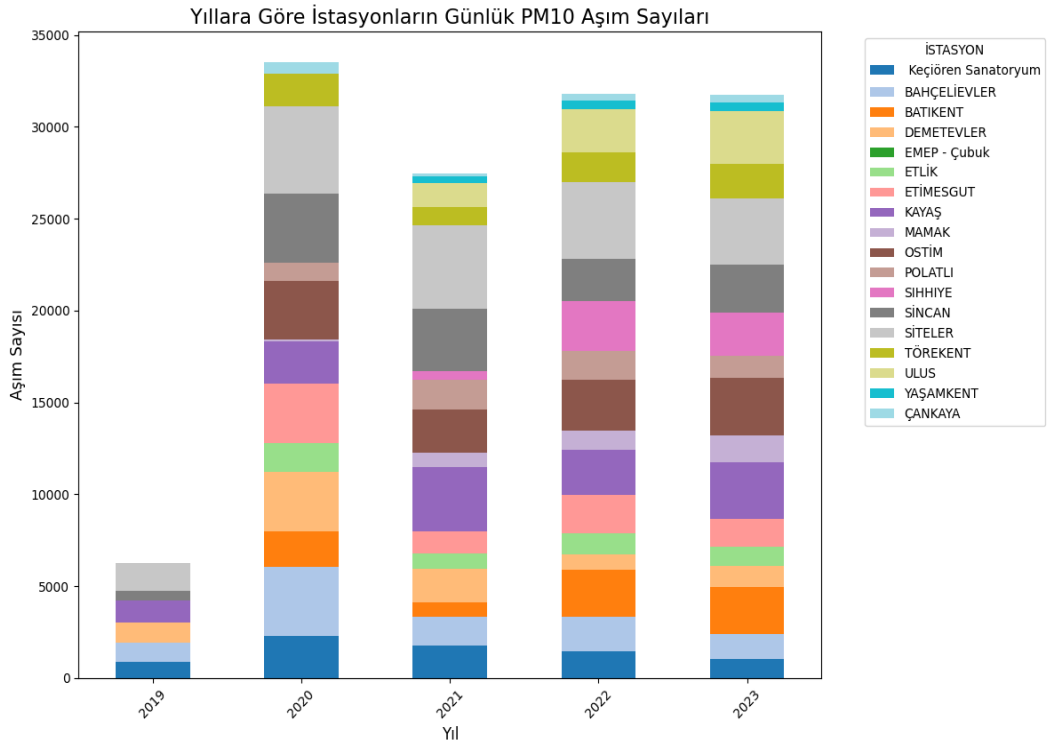
Şekil 3. İstasyonların Yıllara Göre PM10 Günlük Sınır Aşım Sayısı

Şekil 3'de UHKİA'ndan temin edilen ham veriler kullanılarak elde edilen analiz sonucuna göre, 2019-2023 yılları arasında, istasyonlardaki PM10 günlük sınır aşımı sayıları yer alan ısı haritası incelendiğinde aşağıdaki değerlendirmelere varılmıştır:

2020 yılı, COVID-19 pandemisi nedeniyle trafik ve endüstriyel faaliyetlerin geçici olarak azaldığı bir yıl olmasına rağmen, birçok istasyonda en yüksek aşım değerlerinin gözlendiği yıl olması dikkat çekicidir. Bu durum, meteorolojik koşulların veya ev içi yakıt vb. kullanımının artışı gibi alternatif etkenleri işaret edebilir.

2021'den itibaren ise, aşım sayılarında kademeli bir düşüş eğilimi gözlemlenmektedir. Ancak istasyon bazında bu değişimin homojen olmadığı görülmektedir.

Ayrıca Şekil 3'de sunulan PM10 kirleticisine ait veriler kullanılarak yıl ve istasyon bazında günlük aşım sayıları yıl bazında Şekil 4'te sunulmuştur.



Şekil 4. Yıllara Göre İstasyonların Günlük PM10 Aşım Sayıları

Yapılan analiz sonucunda Şekil 4'deki grafikte yer alan istasyonlar arasında farklı yıllarda önemli farklılıklar gözlemlenmiştir. Aşağıda her istasyonun yıllara göre PM10 aşım sayıları değerlendirilmiştir.

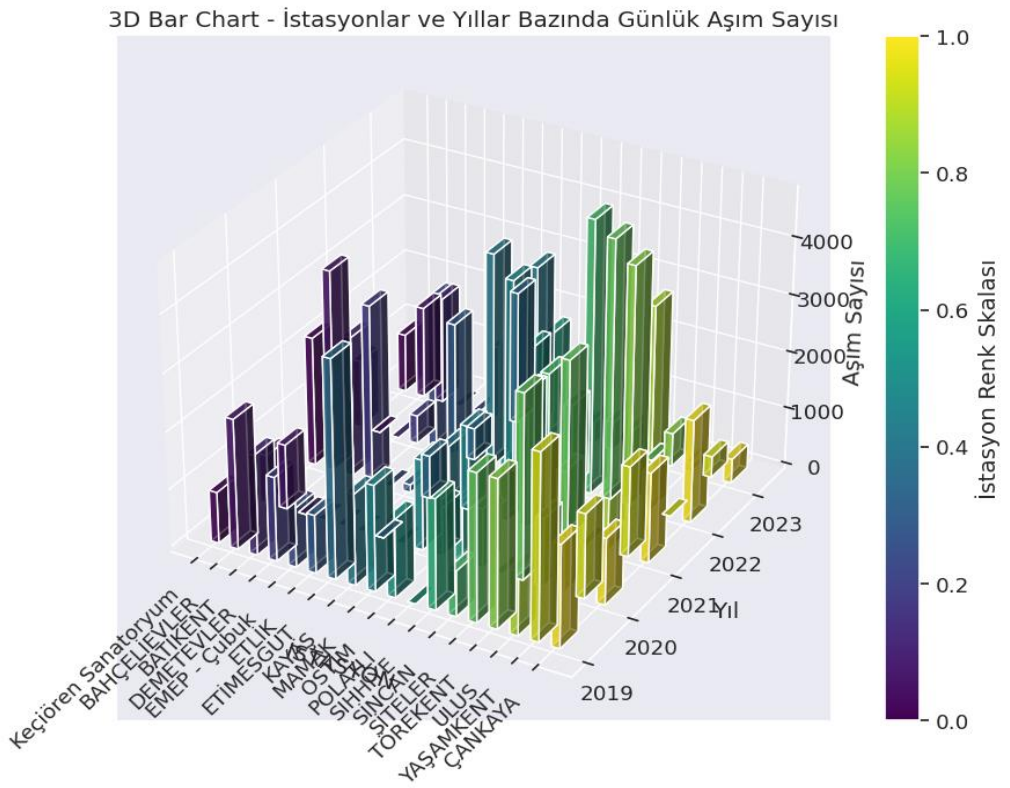
- Keçiören Sanatoryum, 2019 yılında belirgin bir artış göstermektedir.

- Bahçelievler istasyonundaki hava kirletici verileri, stabil bir performans izlenirken yıllar arasında az değişiklik gösterdiği izlenmiştir.

- Batıkent, Demetler, EMEP-Çubuk istasyonlarda ise dalgalanmalar görülmekte ise de bazı yıllarda ciddi aşım sayıları kaydedilmiştir.

- Diğer istasyonlarda ise çeşitli yıl ve dönemlerdeki aşım sayıları, hava kalitesinin bölgesel değişkenliğini göstermektedir.

Şekil 5'de, Şekil 3 ve Şekil 4'de elde edilen sonuçların birleştirilmesiyle farklı istasyonlarda yıllar bazında günlük hava kalitesi sınır aşım sayılarının üç boyutlu dağılımı sunulmuştur. Ankara ilindeki çeşitli hava kalitesi izleme istasyonlarında 2019–2023 yılları arasında ölçülen günlük PM_{10} aşım sayıları görselleştirilmiştir. Üç boyutlu çubuk grafik ile zaman serisi eğilimleri, mekânsal farklar ve genel kirlilik seviyelerindeki değişimler istasyonlar arası karşılaştırmalı olarak verilmiştir. Birçok istasyonda 2019 ve 2020 yıllarında yüksek aşım sayıları gözlemlenirken, 2021 ve 2023 yıllarında genel olarak bir azalma eğilimi gözlenmiştir. Sincan, Törekent, Ulus gibi istasyonlarda tüm yıllar boyunca yüksek aşım değerleri görülmektedir. COVID-19 pandemisinin etkili olduğu 2020 yılında, bazı istasyonlarda ani düşüşler gözlemlenmiştir. Bu sonuçlara göre COVID-19 pandemi döneminde uygulanan kısıtlamalara bağlı azalan trafik ve endüstriyel faaliyetleriyle açıklanabilir. İstasyonlar arası aşım sayıları bakımından önemli mekânsal farklılıklar söz konusudur. Bu durum, kent merkezi, sanayi alanı ve konut bölgesi gibi yerleşim yerlerinin yoğunluğuyla açıklanabilir. Keçiören Sanatoryum ve Çankaya istasyonlarında aşım sayıları belirgin şekilde daha düşüktür.



Şekil 5. Yıllara Göre İstasyonların Günlük PM10 Aşım Sayıları 3D Gösterimi

3.2.2 Meteoroloji istasyon verileri

Meteorolojik parametreler, hava kirliliği tahmin modellerinde temel belirleyiciler arasında yer almakta olup, atmosferik koşulların fiziksel özelliklerini ve dinamik süreçlerini temsil etmektedir. Bu parametreler, atmosferde bulunan hava kirleticilerin taşınımı, yayılımı, seyreltimi ve çökelme süreçlerini doğrudan etkileyerek, hem kısa vadeli hem de uzun vadeli hava kalitesi tahminlerinin doğruluğu üzerinde önemli derecede katkıları olduğu bilinmektedir (Seinfeld ve Pandis, 2016).

Bu kapsamda, çalışmada 2019-2023 yılları arasında Ankara'daki ölçüm istasyonlarından alınan saatlik meteorolojik veriler kullanılmıştır. Bu veriler, T.C. Çevre Şehircilik ve İklim Değişikliği Bakanlığı, Meteoroloji Genel Müdürlüğü, MEVBİS (Meteorolojik Veri Bilgi Satış ve Sunum Sistemi) kayıtlarından temin edilmiştir (MEVBİS, 2023). Elde edilen her bir meteorolojik değere ait veriler ayrı metin dosyaları (Notepad - .txt) biçiminde yapılandırılmış olup, bu dağıtık veri yapısı, makine öğrenme sürecine uygun hale getirilmek için tek bir .csv dosyasında toplanmıştır.

Aşağıda, çalışmada kullanılan temel meteorolojik değerler ve bunların hava kirliliği üzerindeki etkileri açıklanmaktadır:

Sıcaklık (°C): Özellikle yaz aylarında artan sıcaklık, uçucu organik bileşikler (VOC'ler) ile azot oksitler (NOx) arasında fotokimyasal reaksiyonları hızlandırarak troposferik ozon (O_3) oluşumunu artırmaktadır. Bu süreç, insan sağlığına olumsuz etkileri olan seviyelerde ozon konsantrasyonlarının oluşmasını etkilemektedir (Jacob ve Winner, 2009).

Basınç (hPa): Basınç değişimleri, hava kütlelerinin hareketini yönlendirir. Bu da kirleticilerin yatay ve dikey taşınma süreçlerini etkiler. Düşük basınç sistemleri, yağışlı ve hareketli hava kütleleriyle birlikte çalışır. Bu sistemler, hava sirkülasyonunu artırarak hava kirliliğini seyreltimi ve uzaklaştırması açısından iyileştirici etki gösterir. Öte yandan, yüksek basınç sistemleri kirleticilerin birikmesine neden olabilir (Tai, Mickley ve Jacob, 2010).

Rüzgâr Hızı ve Yönü (m/s, derece): Rüzgâr, atmosferdeki hava kirleticilerin taşınımında ve dağılımında belirleyici bir faktördür. Özellikle yüksek rüzgâr hızları, ince partikül maddelerin ($PM_{2.5}$) atmosferde geniş alanlara yayılmasına neden olarak bölgesel hava kalitesi üzerinde önemli etkileri bulunmaktadır. Rüzgar yönü ise hava kirletici konsantrasyonlarının hangi yöne taşınacağını belirleyerek kirlilikteki bölgesel değişimleri etkilemektedir (Elminir, 2005; Seinfeld ve Pandis, 2016).

Bağıl Nem (%): Bağıl nem, partikül maddelerin atmosferdeki yaşam süresini ve higroskopik büyümesini etkileyerek, PM konsantrasyonlarının artmasına veya azalmasına neden olabilmektedir. Yüksek bağıl nem koşullarında, partiküller su buharı ile birleşerek çapça büyür ve yer seviyesine inme eğilimini arttırmaktadır (Elminir, 2005; Tai vd., 2010).

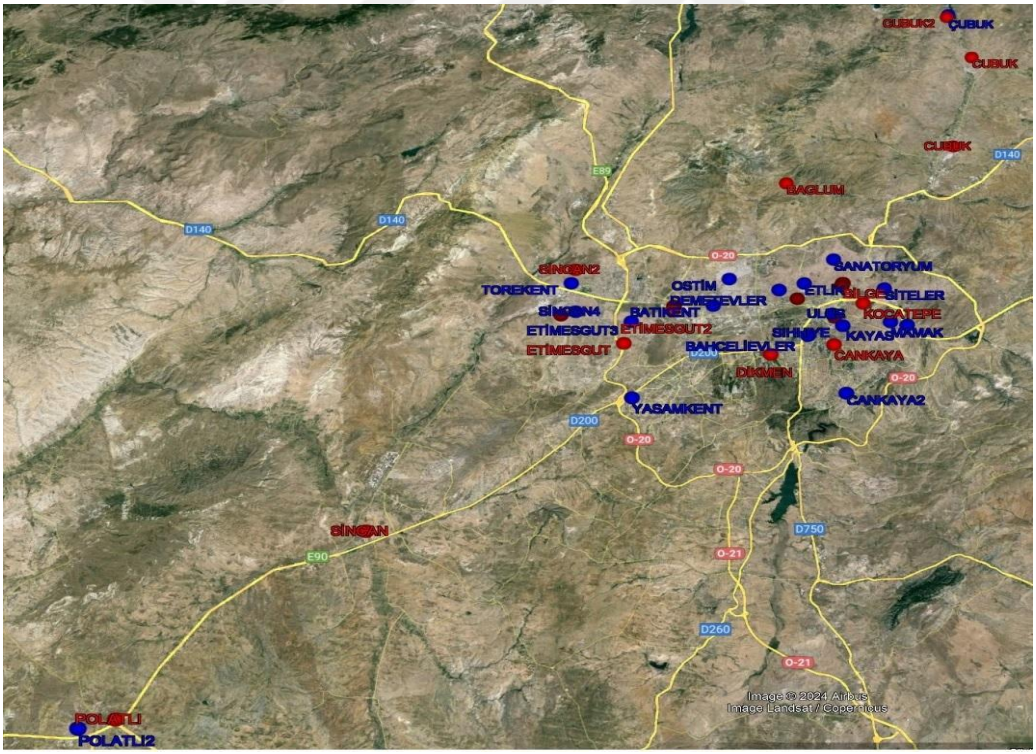
Yağış (mm): Yağış, atmosferde askıda duran partikül maddelerin ve çözünmüş gazların yer yüzeyine taşınmasında etkili olan başlıca doğal temizleme (wet deposition) mekanizmasıdır. Bu nedenle yağışlı günlerde partikül ve gaz kirleticilerin konsantrasyonlarında belirgin azalmalar gözlemlenebilir (Tai vd., 2010).

Bu meteorolojik parametrelerin zamansal çözünürlükte modellenmesi ve hava kalitesi tahminleme sürecine entegre edilmesi, modelin çevresel koşullara duyarlılığını artırmakta

ve tahmin performans başarısını iyileştirmektedir. Bu bağlamda, çalışmada kullanılan meteorolojik veriler, yalnızca yardımcı değişkenler olarak değil, aynı zamanda atmosferik sistemin fiziksel temsilleri olarak modelle dahil edilmiştir.

3.2.3. Hava kalitesi ve meteoroloji istasyon verilerinin birleştirilmesi

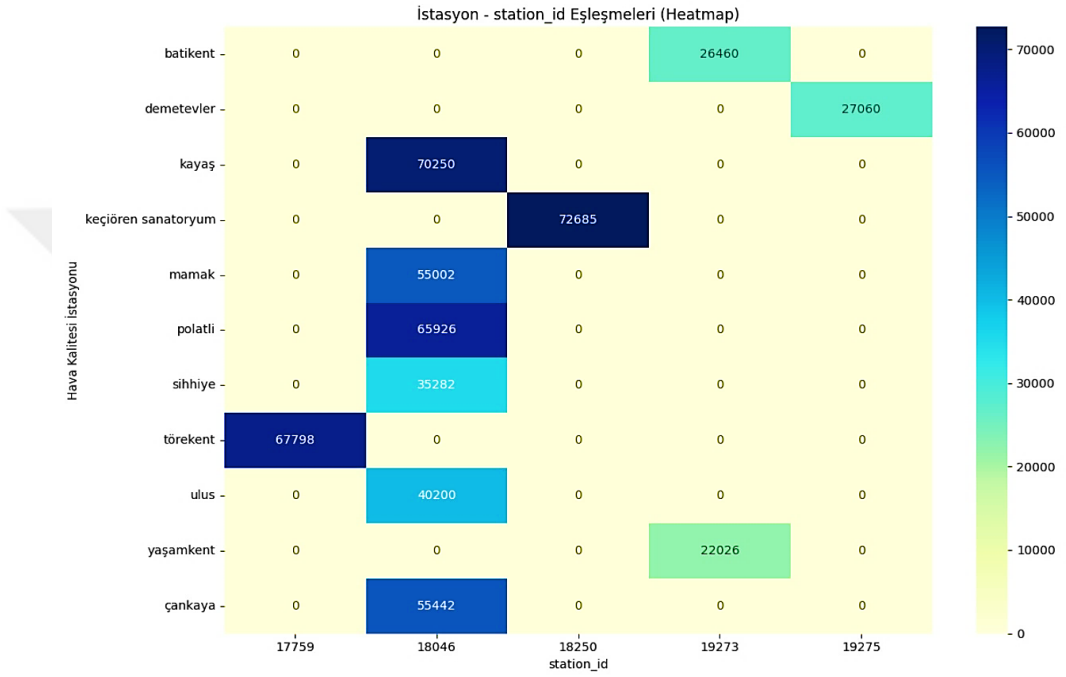
Veri setlerinde kullanılan tüm hava kalitesi izleme ve meteoroloji istasyonlarının coğrafi koordinatları, mekânsal analizlerin gerçekleştirilmesi amacıyla Google Maps Pro'ya aktarılmıştır. Şekil 6'da, hava kalitesi istasyonları mavi işaretçilerle, meteoroloji istasyonları ise kırmızı işaretçilerle gösterilmiştir. Her iki hava kalitesi izleme ve meteoroloji istasyonlarının Ankara ilindeki dağılımı ve birbirleriyle olan ilişkileri görsel olarak sunulmuştur. Bu görselden, hava kirletici ölçümleri ile yerel meteorolojik koşullar arasındaki mekânsal ilişkilerin incelenerek, her iki veri kaynağının birleştirilmesinde faydalanılmıştır.



Şekil 6. Hava Kalitesi ve Meteoroloji İstasyonlarının Konum Bilgisi

Hava kalitesi ve meteoroloji verileri tek bir .csv dosyasında birleştirilmiştir. Bu bütünleştirme sürecinde, meteoroloji verilerinde yer alan istasyon kodları temel alınmış; hava kalitesi verilerinde ise istasyon adları, konum koordinatları ve veri doluluk oranları

dikkate alınarak her hava kalitesi istasyonu için en uygun meteoroloji istasyonu eşleştirilmiştir. Böylece, her hava kalitesi istasyonu, coğrafi yakınlık ve veri dololuk miktarlarına göre bir meteoroloji istasyonu ile ilişkilendirilmiştir. Bu yöntem, hem hava kirliliği tahminlerinin doğruluğunu artırmayı hem de kirliliğin farklı bölgelerdeki etkilerinin daha kapsamlı bir şekilde değerlendirilmesini sağlamayı amaçlamaktadır. Söz konusu eşleştirme matrisi, ısı haritası olarak Şekil 7'de gösterilmektedir.



Şekil 7. Hava Kalite ve Meteoroloji İstasyon Konum Eşleştirme Matrisi

3.3 Veri Temizleme ve Normalizasyon Süreçleri

Veri madenciliği ve makine öğrenimi tabanlı tahmin modellerinde, model başarımı büyük ölçüde ham verilerin işlenme kalitesine de bağlıdır. Bu bağlamda, modelleme sürecinin ilk ve en kritik aşamalarından biri, verilerin temizlenmesi, dönüştürülmesi ve analiz aşamasına hazır hale getirilmesidir. Ham veriler; eksik değerler, ölçüm hataları, aykırı gözlemler ve biçimsel tutarsızlıklar gibi çeşitli sorunları içerebilir. Bu tür sorunların, doğrudan modelin tahmin hatalarına ve genellenebilirlik sorunlarına neden olabileceği bilinmektedir (Han, Kamber ve Pei, 2011; García, 2015).

Bu çalışmada kullanılan T.C. Çevre, Şehircilik ve İklim Değişikliği Bakanlığı ve Meteoroloji Genel Müdürlüğü'nden sağlanan saatlik meteorolojik ve hava kalitesi

verileri, çok kaynaklı (Excel ve Notepad formatlarında) dosyalarda temin edilmiştir. Hava kalitesi istasyon verileri ve meteoroloji verileri arasında zaman damgası ve istasyon bilgileri temel alınarak; en yakın ve veri doluluk oranı göz önünde bulundurularak tek bir yapısal veri kümesinde (csv formatında) birleştirilmiştir. Bu işlemler sırasında aşağıdaki ön işleme adımları dikkatle uygulanmıştır.

Öncelikle farklı formatlarda sunulan tarih, istasyon adı ve parametre adları gibi yapısal bileşenler standardize edilerek veri kümesi içi tutarlılık sağlanmıştır. Zaman serisi modellerinin etkili bir şekilde eğitilebilmesi için tüm sayısal özellikler, RobustScaler yöntemi ile normalize edilmiştir (Géron, 2019). Bu kapsamlı veri hazırlama süreci, modelin girdilerinin hem zamansal hem de mekânsal olarak tutarlı olmasını sağlamıştır. Böylece model performansı açısından iyileştirilmiş bir veri seti hazırlanmıştır.

Zaman serisi modelleme çalışmalarında, verilerin doğru bir şekilde dizinlenmiş ve kronolojik sıralanmış olması, tahmin modellerin performansı açısından temel bir gerekliliktir (Hyndman ve Athanasopoulos, 2018). Bu amaçla, çalışma kapsamında kullanılan hava kalitesi ve meteorolojik veriler, aşağıdaki biçimlendirme adımlarına tabi tutulmuştur:

- Zaman Damgası Oluşturulması: Ham veri setinde bulunan gün, ay, yıl ve saat sütunları birleştirilerek datetime formatında yeni bir zaman damgası sütunu oluşturulmuştur. Bu işlem, zaman serisi indekslemesinin sağlıklı yürütülmesi için kritik öneme sahiptir.
- Zamansal Sıralama ve Dizinleme: Tüm veriler, oluşturulan zaman damgası üzerinden kronolojik olarak sıralanmış ve zaman serisi formatında pandas DatetimeIndex yapısına dönüştürülmüştür.
- Eksik Zaman Dilimlerinin İkmal Edilmesi: Zaman serisi içerisindeki olası eksik zaman noktaları, belirlenen örnekleme frekansına (saatlik) göre tamamlanarak dizinin bütünlüğü korunmuştur. Bu işlem, zaman serisinin düzgün periyotlara sahip olması açısından model eğitimi öncesinde önemli ve gerekli bir aşamadır.

3.4 Eksik Verilerin İşlenmesi ve Tamamlannması

Zaman serisi temelli hava kalitesi ve meteorolojik veri analizlerinde eksik verilerin doğru biçimde ele alınması, elde edilen modellerin hem doğruluk hem de genellenebilirlik açısından güvenilir olmasını sağlar. Eksik gözlemler, zaman serilerinde yapısal bozulmalara neden olabilir ve istatistiksel analizlerin geçerliliğini ciddi ölçüde etkileyebilir. Bu nedenle, eksik veri problemi sadece teknik değil, aynı zamanda bilimsel bir titizlikle ele alınması gereken kritik bir konudur.

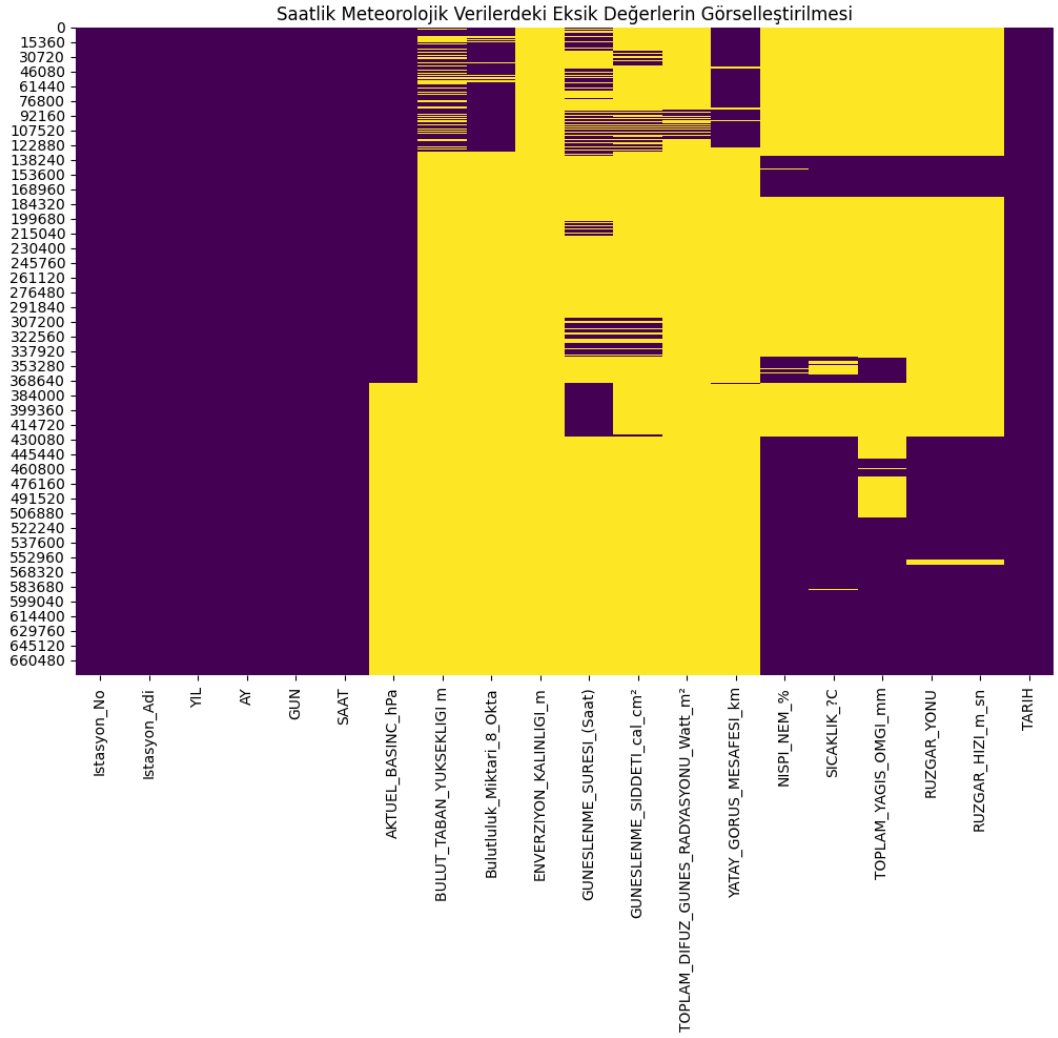
Bu çalışmada, Ankara iline ait 2019–2023 yılları arasındaki PM_{10}, $PM_{2.5}$, NO_2, CO, SO_2, O_3, NO, NOX gibi hava kirletici verileri ile sıcaklık, bağıl nem, rüzgâr hızı gibi meteorolojik değişkenlerin oluşturduğu zaman serileri üzerinde eksik veri analizi gerçekleştirilmiş ve belirlenen veri tamamlama yöntemleri ile iyileştirme stratejileri uygulanmıştır.

3.4.1 Eksik veri dağılımı ve görselleştirilmesi

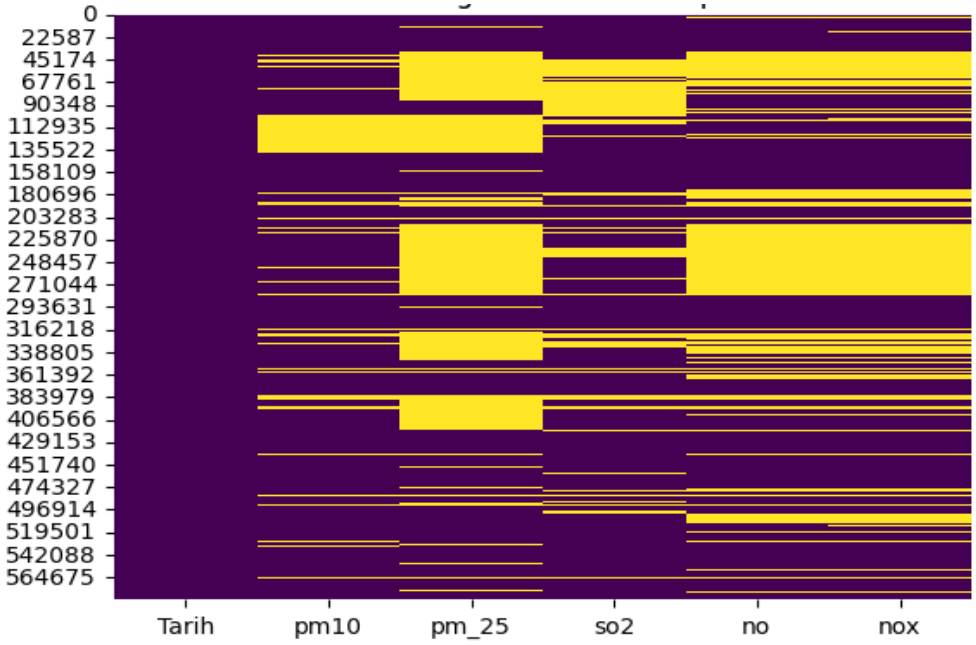
Eksik veri örüntüsünü anlamak, verinin zaman içindeki yapısal eksikliklerini anlamamızı sağlar. Zaman içerisindeki dağılımını görsele dayalı analiz etmek amacıyla, istatistiksel ve grafiksel araçlardan ısı haritası, keman grafiği (violin plot), histogram, pasta grafiği (pie chart) ve çubuk grafik yöntemleri kullanılmıştır.

a) Eksik veri ısı haritası (Heatmap)

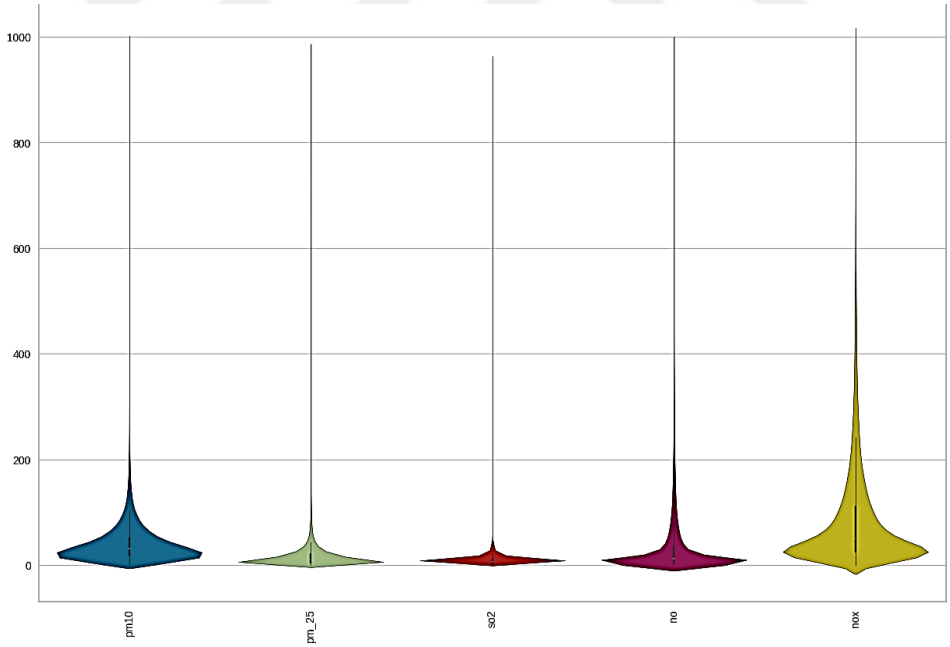
UHKİA'ndan temin edilen ham veriler kullanılarak eksik verilerin zamansal ve değişken bazlı desenleri, ısı haritaları ile görselleştirilmiştir. Şekil 8 ve Şekil 9'da sunulan bu haritalarda, sarı bölgeler eksik gözlemleri, mor bölgeler ise mevcut verileri temsil etmektedir. Her bir satır zaman aralığını, sütunlar ise analiz edilen değişkenleri göstermektedir. Bu görselleştirme, eksik veri probleminin zaman içerisindeki sürekliliğini ve belirli dönemlere yoğunlaşma eğilimini ortaya koymuştur.



Şekil 8. Saatlik Meteoroloji Verilerindeki Eksik Değerlerin Isı Haritası



Şekil 9. Saatlik Hava Kirletici Verilerindeki Eksik Veri Değerlerin Isı Haritası



Şekil 10. Hava Kirleticilerin Violin Plot Gösterimi

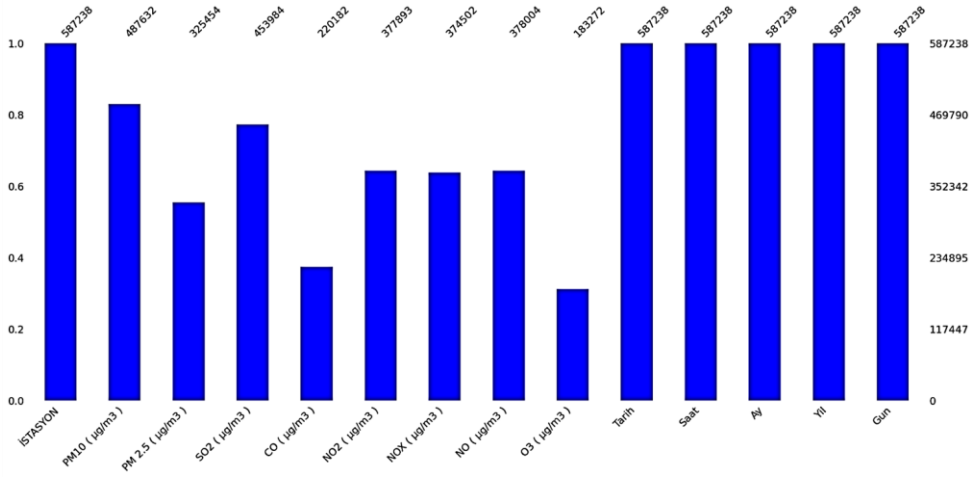
Eksik veriler farklı yöntemlerle tamamlandıktan sonra, elde edilen verilerin istatistiksel dağılımları, eksik veri öncesi durumla karşılaştırmalı olarak analiz edilmiştir. Bu amaçla, PM_{10} ve $PM_{2.5}$ gibi kritik hava kalitesi göstergeleri başta olmak üzere tüm değişkenler

için violin plot görselleştirmeleri kullanılmıştır (Şekil 10). Bu görselleştirmeler, her bir veri tamamlama yönteminin veri dağılımı üzerindeki etkisini açık biçimde ortaya koymakta olup, özellikle uç değerlerin bastırılması, dağılımın daralması veya kayması gibi yapısal değişikliklerin görsel olarak izlenmesine olanak sağlamaktadır. Bu analizler sonucunda bazı yöntemlerin, özellikle PM_{10} ve $PM_{2.5}$ gibi varyansı yüksek değişkenlerde, bazı yöntemlerin uç değerleri bastırarak varyansı azalttığı, dolayısıyla dağılımı yapay biçimde daralttığı gözlemlenmiştir. Bu durum, modelleme sürecinde varyans kaybı yaşanmasına ve potansiyel yanlılıklara neden olabileceğinden dikkatle değerlendirilmiştir.

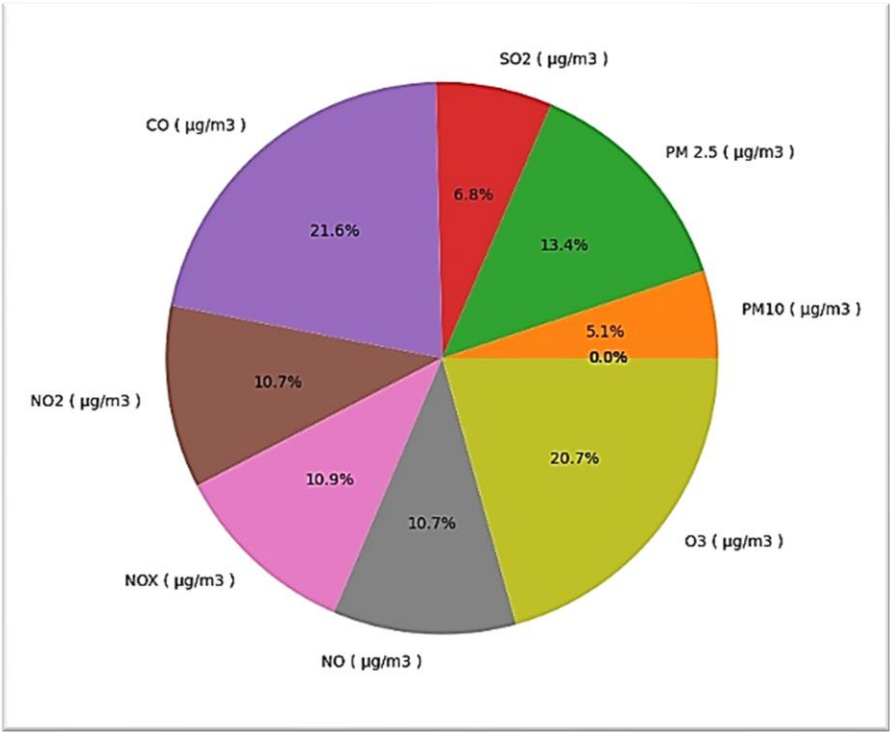
(b) Histogram ve çubuk grafikler ile eksik veri dağılımı analizi

Eksik veri analizi için, önce Şekil 11'de tüm değişkenlerin gözlemlenen veri sayıları histograla gösterilmiştir. Sonrasında, Şekil 12'deki pasta grafik ile her bir hava kalitesi değişkeninin eksik veri yüzdesi hesaplandı ve görselleştirildi. Bu görsel analiz, hangi değişkenlerin tahmin modellerinde daha güvenilir olduğunu belirlemek açısından çok önemlidir.

Analiz sonuçları, CO (%21.6), O_3 (%20.7), $PM_{2.5}$ (%13.4) ve NOx (%10.9) için yüksek veri eksikliklerini ortaya koymuştur. Bu çarpıcı yüzdeler, bu değişkenlerin model performansını düşürme riski taşıdığını göstermektedir. Buna karşın, PM_{10} değişkenindeki eksik veri oranı %5'tir. Bu durum, PM_{10}'un zaman serisi tahmin modellerinde daha güvenilir bir hedef değişken olduğunu göstermektedir.

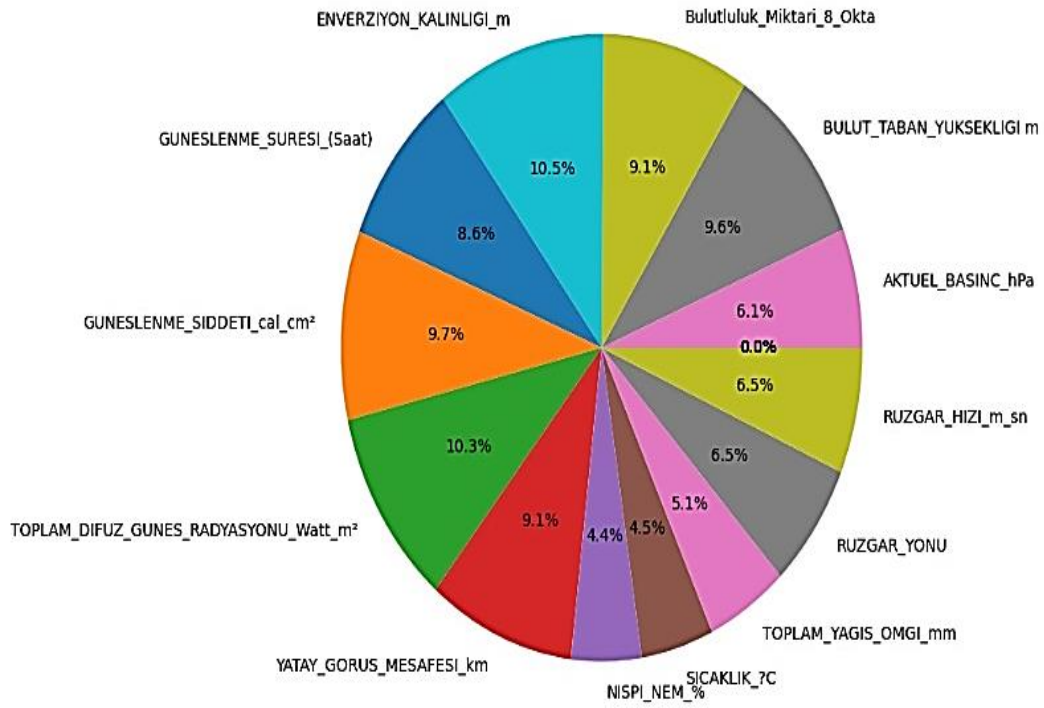


Şekil 11. Hava Kirletici Verilerindeki Eksik Veri Miktarı



Şekil 12. Hava Kirletici Verilerindeki Eksik Veri Oranları

Hava kalitesi değişkenleri ve meteorolojik verilerdeki eksiklik oranları da incelenmiştir. Şekil 13'de, bu değişkenlere ilişkin eksik veri oranları pasta grafik biçiminde sunulmuştur. Bu görselleştirme, özellikle sıcaklık, bağıl nem ve rüzgar hızı gibi değişkenlerin modelleme sürecinde ne kadar güvenilir veri sağladığını değerlendirmek için önemlidir. Pasta grafik, eksik veri oranlarını görsel olarak sunar. Böylece, modellerin ön işleme (preprocessing) aşamasında dikkate alması gereken önemli unsurları ortaya koymaktadır.



Şekil 13. Meteoroloji Verilerindeki Eksik Veri Oranları

3.4.2 Eksik veri tamamlama yöntemleri

Eksik verileri tamamlamak için hem geleneksel hem de istatistiksel yöntemler kullanılmıştır. Bu çalışmada altı farklı yöntem uygulanmış ve performansları karşılaştırılmıştır. Her bir yöntemin teorik altyapısı, varsayımları, avantaj ve dezavantajları detaylı olarak açıklanmış ve Python ortamında sklearn, fancyimpute, statsmodels gibi kütüphaneler aracılığıyla uygulanmıştır.

a) K-En yakın komşu yöntemi

K-en yakın komşu (K-Nearest Neighbors, KNN) algoritması, eksik değerlerin tamamlanmasında komşu gözlemler arasındaki öklidyen mesafeye dayalı benzerlikleri temel alır. Komşu gözlemlerin etkisini dikkate alması nedeniyle yerel yapıyı korur, ancak yüksek boyutlu veri kümelerinde gürültüye duyarlı hale gelebilmektedir. Özellikle veri setinde mekân veya zaman korelasyonunun güçlü olduğu veri setlerinde genelllikle iyi performans göstermektedir (Batista ve Monard, 2010).

b) Zaman serisi tabanlı tahmin yöntemleri

Zaman serisi analizlerinde eksik verilerin tamamlanması için doğrusal, polinom veya spline interpolasyon gibi yöntemler yaygın olarak kullanılmaktadır. Bu teknikler, zaman serisinin sürekliliğini koruyarak eksik değerleri önceki ve sonraki gözlemler üzerinden tahmin etmeye olanak tanımaktadır. Hyndman ve Athanasopoulos (2018), özellikle kısa süreli ve tek değişkenli eksikliklerde bu yöntemlerin etkili olduğunu vurgulamaktadır.

c) Doğrusal interpolasyon (Linear Interpolation)

Doğrusal interpolasyon yöntemi, eksik değeri bir önceki ve bir sonraki gerçek değer arasındaki doğrusal ilişki üzerinden tahmin eden basit ancak etkili bir yöntemdir. Veri yapısının istatistiksel olarak düzenli olduğu ve eksiklik oranının düşük olduğu durumlarda yeterli sonuç verir. Ancak karmaşık zaman serisi desenlerini yakalamada yetersiz kalabilmektedir (Hyndman ve Athanasopoulos, 2018).

d) Zincirli Denklemlerle Çoklu Atama (Multiple Imputation by Chained Equations - MICE)

MICE yöntemi, her değişken için sırasıyla regresyon modelleri kurup, eksik verilerin birden fazla olasılık dağılımına dayanarak zincirleme tekrarlayarak tahmin eder. Tahminlerin istatistiksel olarak birleştirilmesine dayanan güçlü bir yaklaşımdır. Bu yöntem, özellikle yüksek boyutlu ve çok değişkenli veri setlerinde, özellikle trend mevsimsellik gibi yapısal bileşenlerin korunduğu durumlarda güvenilir tahminler elde etmek amacıyla tercih edilmektedir (van Buuren ve Groothuis-Oudshoorn, 2011). Schafer (1997), MICE yönteminin sistematik eksikliklerin modellenmesinde etkili

olduğunu ve belirsizliklerin tahmin sürecine entegre edilerek daha sağlam istatistiksel çıkarımlar yapılmasını sağladığını ifade etmiştir.

e) Rastgele Orman (Random Forest) yöntemi

Rastgele Orman algoritması, karar ağaçları temelinde çalışan ve eksik değerleri sınıflandırma yöntemiyle tahmin edebilen çok yönlü bir makine öğrenmesi yöntemidir. Özellikle doğrusal olmayan ilişkiler içeren, yüksek değişkenlik gösteren ve düzensiz veri yapılarında üstün performans göstermektedir. Breiman (2001), bu yöntemin yüksek boyutlu veri setlerinde esnekliği ve tahmin performans başarısı ile öne çıktığını vurgulamıştır.

f) Doğrusal Regresyon

Eksik değeri, diğer değişkenler üzerinden kurulan doğrusal modelle tahmin edilmesini sağlayan klasik bir istatistiksel yaklaşımdır. Değişkenler arasında güçlü doğrusal ilişkilerin bulunduğunda performanslıdır. Özellikle PM_{10} ve $PM_{2.5}$ gibi kirleticiler arasındaki yüksek korelasyon, bu yöntemin uygulanabilirliğini artırmaktadır. Ancak, veride doğrusal olmayan etkileşimler varsa iyi sonuçlar sağlamayabilir. Eksik veri tamamlama yöntemleri arasında en eski ve temel yaklaşımlardan biridir. Ancak modern veri yapılarında tek başına yeterli olmayabileceği savunulmuştur (Little ve Rubin, 2002).

3.4.3. Karşılaştırılmalı performans analizi

Bu çalışmada uygulanan veri tamamlama yöntemlerinin etkinliği, eksik veriler kullanılarak oluşturulan doğrulama altkümesi üzerinden değerlendirilen Ortalama Kare Hata Karekökü (Root Mean Squared Error, RMSE) ve Ortalama Mutlak Hata (Mean Absolute Error, MAE) metrikleri aracılığıyla incelemiştir. RMSE metriği, büyük hatalara karşı duyarlılığı; MAE metriği ise ortalama sapma düzeyini ortaya koyar. Ayrıca yöntemlerin hata dağılım dengesi MAE/RMSE oranı ile analiz edilmiştir.

a) KNN : Düşük oranda eksiklik içeren kısa süreli boşluklarda, KNN ve zaman serisini bir önceki ve bir sonraki verisiyle doldurma yöntemleri, beklenen aralık içinde düşük sapma göstererek etkin sonuçlar sunmuştur.

b) Doğrusal Interpolasyon: Doğrusal varsayımlara dayanması, karmaşık ve doğrusal olmayan bağıntılar içeren veri yapılarında yüksek RMSE ve MAE değerlerine neden olmuş, bu yöntemlerin özellikle uç değerleri doğru modelleyemediği belirlenmiştir.

c) Basit atama (Simple Imputer): Ortalama ve medyan tabanlı doldurma metodolojisi, MAE bakımından nispeten düşük hata oranları sağlarken, mekânsal ve zamansal örüntüleri korumada sınırlı kalmıştır.

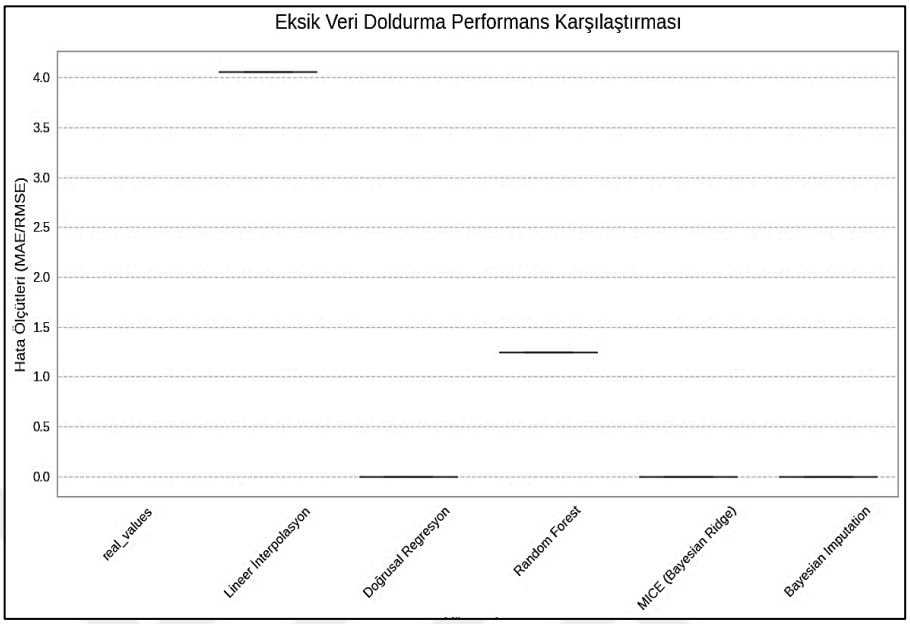
d) Rastgele orman yöntemi: Ağaç tabanlı ensemble yaklaşımı, ortalama MAE bakımından rekabetçi sonuçlar sunsa da, yüksek RMSE değerleri, yöntemin uç değer hassasiyetine karşı yeterince dayanıklı olmadığını ve bazı değişkenlerde büyük sapmalara yol açtığını göstermiştir. Bu da modelin büyük hataları önleyemediğini ve karmaşık bağımlılıkları yansıtmakta yetersiz kaldığını göstermektedir.

e) MICE yöntemi : Hem RMSE hem de MAE metriklerinde dengeli bir performans sergileyen MICE, eksiklik oranının yüksek olduğu PM2.5 ve NO_2 gibi değişkenlerde en düşük hata değerlerini elde etmiştir. MICE'in ürettiği birden fazla tamamlama seti, verinin varyans yapısını korurken, Predictive Mean Matching mekanizması marjinal dağılımların bozulmasını engellemiş, bu da yöntemin genel hata profiline olumlu yansımıştır. Çok değişkenli yapıları dikkate alarak tahmin ürettiği için istatistiksel bütünlüğü koruma kapasitesi yüksektir.

Hata metrikleri arasındaki dengeyi nicelleştirmek amacıyla bugün sıkça kullanılan MAE/RMSE oranı incelenmiştir. Oranın 1'e yakın olması, büyük ve küçük hata büyüklüklerinin benzer düzeyde yönetildiğini gösterirken; yüksek oranlar, büyük hataların baskın olduğunu ve yöntemin uç değerlerde yetersiz kaldığını işaret eder. Bu bağlamda MAE/RMSE oranı, yöntemlerin uç değer hassasiyetini ve verinin temel yapısını koruma kapasitesini değerlendirmek açısından kritik bir ölçüt sunmaktadır.

Şekil 14'te sunulan eksik veri tamamlama yöntemlerinin MAE/RMSE oranlarına dayalı performans karşılaştırması, MICE yönteminin hata dağılımında istatistiksel dengeyi en

iyi şekilde koruduğunu ve bu açıdan diğer yöntemlere kıyasla daha tutarlı sonuçlar ürettiğini ortaya koymuştur.



Şekil 14. Eksik Veri Tamamlama Yöntemlerinin Performans Karşılaştırması

3.4.4 MICE ile Veri Tamamlama

Yukarıda açıklanan performans karşılaştırmaları sonucunda, hem hava kalitesi hem de meteorolojik verilerde eksik verilerin tamamlanmasında Zincirli Denklemlerle Çoklu Atama (Multiple Imputation by Chained Equations, MICE) yöntemi seçilmiştir. Bu yöntem, eksik verilerin çok boyutlu bağımlılık yapılarını dikkate alarak, hem değişkenler arası ilişkileri hem de zaman serisi bağlamındaki bağımlılıkları başarıyla modellemektedir. Böylece hem tahmin doğruluğu hem de modelin açıklanabilirliği açısından üstünlük sağlamıştır. Seçimi yönlendiren temel gerekçeler şunlardır:

a) Çok Boyutlu Korelasyon Yapısının Korunması: MICE, her bir eksik değişken için diğer tüm değişkenleri açıklayıcı olarak kullanarak, veri kümesinin içsel korelasyon matrisine uygun koşullu modeller oluşturur.

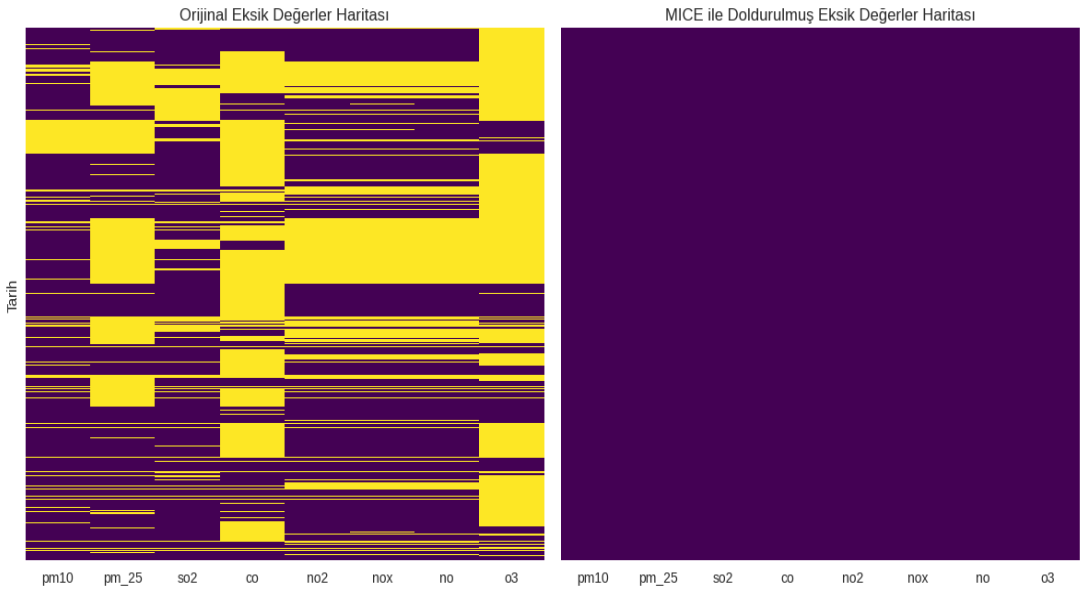
b) Tahmin Edilen Ortalama (Predictive Mean Matching, PMM) tekniği, atanan değerlerin orijinal veri dağılımına yakın olmasını sağlayarak marjinal dağılımların bozulmasının önüne geçer.

c) Çoklu Tamamlama (Multiple Imputation) yöntemi ile, birden çok eksik veri tamamlama üretilerek belirsizlikler modellemede dikkate alınır ve varyans tahmini iyileştirilir.

d) Zaman serisi uyumu için iteratif döngüler sayesinde hem trend hem de mevsimsel bileşenler korunarak eksik değerler güncellenir.

e) Dengeli Hata Profili: Hem RMSE hem de MAE değerlerinde dengeli sonuçlar elde edilerek uç değer duyarlılığı ve ortalama sapma aynı anda minimize edilir.

Sonuç olarak, MICE yöntemi ile tamamlanan hava kalitesi verileri, Şekil 15'de eksik ve tamamlanmış değerlerin ısı haritasıyla görsel olarak desteklenmiştir. Aynı işlem, meteoroloji verilerine de uygulanarak, eksik veriler tamamlanmıştır. Bu yöntem ile, derin öğrenme modellerinin eğitiminde sağlam ve tutarlı bir veri temeli sunarak, model doğruluğu ve genellenebilirliğinin arttırılması amaçlanmıştır.



Şekil 15. Eksik ve Tamamlanmış Verilerin Isı Haritası

3.5 Aykırı Değerlerin Tespiti

Makine öğrenmesi tabanlı tahmin modellerinin güvenilirliği, eğitim verisindeki aykırı değerlerin (outlier) doğru biçimde tespit edilmesi ve uygun stratejilerle yönetilmesine bağlıdır. Aykırı değerler, ölçüm hatalarından, sensör arızalarından veya beklenmeyen

çevresel olaylardan kaynaklanabilir; bu yüzden hem hedef değişkenler (PM_{10}) hem de bağımsız değişkenler üzerinde sistematik bir analiz yapılmıştır. Aykırı değerler, hem ölçüm hatalarından kaynaklı gürültüler hem de beklenmeyen çevresel olayların göstergesi olabileceği için, her bir gözlem alan bağlamında dikkatle değerlendirilmiştir.

Bu çalışmada aykırı değerlerin tespitinde, hem klasik hem de güncel makine öğrenimi tabanlı yöntemler kullanılarak aykırı değerlerin belirlenmiştir. Z- skoru, kutu grafiği, yerel aykırı değer faktörü (Local Outlier Factor, LOF) ve izolasyon ormanı (isolation forest, IF) yöntemleri detaylı incelenerek hava kirletici ve meteoroloji verilerinde uygulanmıştır.

a) Z-Skoru yöntemi: Z-skoru, istatistiksel dağılımın merkezindeki gözlemlerin standart sapma miktarını ne kadar uzaklaştığını ölçerek aykırı değerleri belirlemeye olanak tanır. Gözlem Yi için Z skoru 3.1'deki formülüyle tanımlanır. Denkemdeki Y ortalama, s standart sapmayı temsil etmektedir.

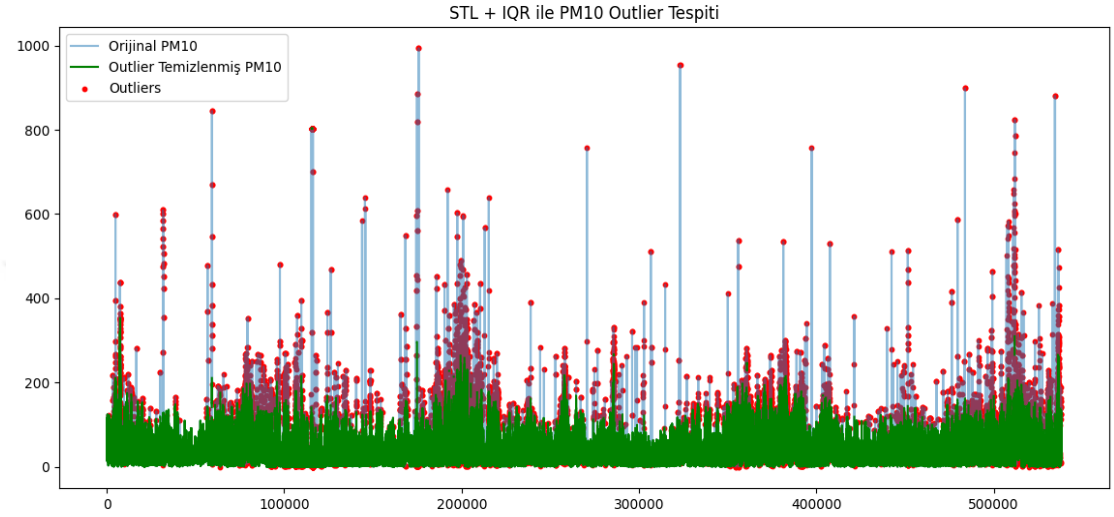
$$Zi = \frac{(Yi - \bar{Y})}{s} \qquad (3.1)$$

Bu çalışmada, $|Z|>3$ koşulunu sağlayan gözlemler potansiyel aykırılar olarak değerlendirilmiştir. Bu yöntem, özellikle normal dağılıma yakın değişkenlerde etkili bir tarama sağlamaktadır (Barnett ve Lewis, 1994).

b) Kutu Grafiği (Boxplot) ve Çeyrekler Arası Aralık (Interquartile, IQR) grafiği

Kutu grafikleri, değişkenlerin dağılımı üzerindeki aykırı gözlemleri görselleştirmek amacıyla kullanılmıştır. Verinin birinci ve üçüncü çeyrek değerleri arasındaki fark, interquartile range (IQR) ile hesaplanır. Aykırı değerler, genellikle $Q1-1.5\times IQR$ altı veya $Q3+1.5\times IQR$ üstü noktalar olarak tanımlanır. Bu yöntem dağılımın şekline bağlı kalmaksızın ara değerler aralığına göre ekstrem noktaları belirler. Kutu grafikleri (boxplot) bu durumu göstermek için kullanılır. Kutu dışındaki noktalar, potansiyel aykırılar olarak belirtilir. IQR bazlı yöntem özellikle simetrik olmayan veya çarpık dağılımlarda Z-skoruna göre daha duyarlı olabilir.

Mevsimsellik ve trend bileşenleri, çalışma veri setinden çıkarıldıktan sonra, veri setindeki kalan residual bileşen üzerinden IQR yöntemi uygulanmıştır. Uç noktalar, Şekil 16'da kırmızı renkle işaretlenerek gösterilmektedir. IQR değeri standart 1.5 * IQR sınırı ile çalıştırıldığında 36113 değer tespit edilmiştir. Yalnızca uç aykırı değerleri yakalaması için k değeri 3.0 incelenmiş ve 9844 adet uç değer belirlenmiştir.



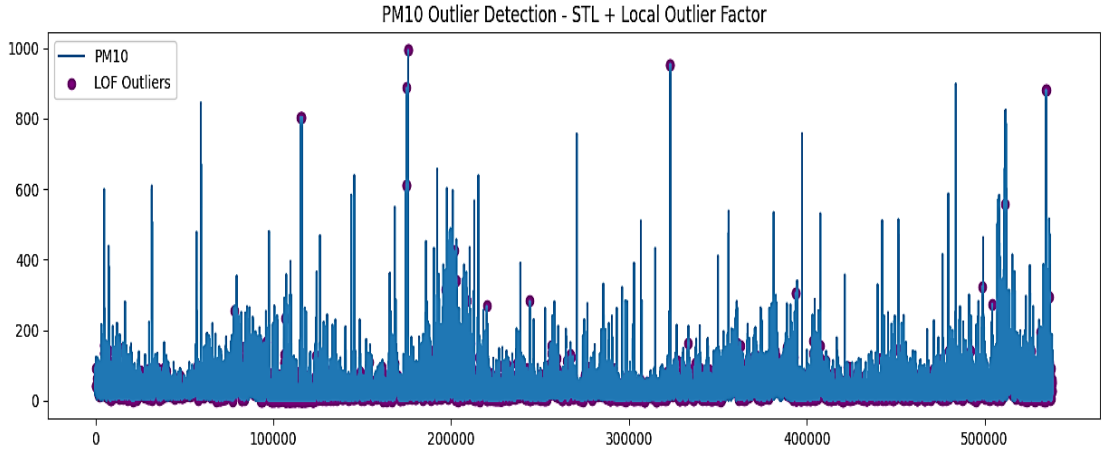
Şekil 16. PM10 Değeri için IQR yöntemi ile Aykırı Deger Tespiti

c) Local Outlier Factor (LOF) yöntemi

LOF algoritması (Breunig, Kriegel, Ng ve Sander, 2000), her gözlemin etrafındaki k-en yakın komşu yoğunluklarını kıyaslayarak aykırı değerleri ölçer. LOF, her nokta için çevresindeki komşu noktaların yoğunluğu ile karşılaştırmalı bir "aykırılık skoru" verir. Eğer bir gözlemin yerel yoğunluğu komşularına kıyasla çok daha düşük yerel yoğunluğa sahipse, bu gözlemyüksek LOF skoru alarak aykırı olarak işaretlenir. LOF yöntemi, özellikle çok boyutlu, yoğunluk değişkenliği gösteren veri setlerinde global eşiklerin yetersiz kaldığı durumlarda etkili olduğu dikkate alınmalıdır (Campos, Zimek, Sander, Campello, Micenková, vd., 2016).

Bu çalışmada bir skor eşiği ya da üst sınır kullanılmadan, doğrudan yerel yoğunluk kıyaslamasıyla anomali belirlenmiştir. Özellikle çok boyutlu ve doğrusal olmayan veri yapılarında, global eşiklere duyarsız kalan bu yöntem, hava kirliliği gibi yerel varyansların yüksek olduğu zaman serisi verileri için oldukça etkilidir. Zaman serisi analizi için trend, mevsimsellik, döngüsellik ve rastgelelik (residual) bileşenlerine

ayrıştırılan veri seti üzerine uygulan LOF yöntemi sonucu, Şekil 17'de LOF ile belirlenen PM_{10} aykırı değerleri kutu grafikleriyle görselleştirilmiştir. LOF yöntemi ile işaretlenen anomaliler grafik üzerinde mor renk ile, kalan PM_{10} değerleri mavi renkleriyle temsil edilerek sunulmuştur. Bu yöntem ile 26906 değer belirlenmiştir.

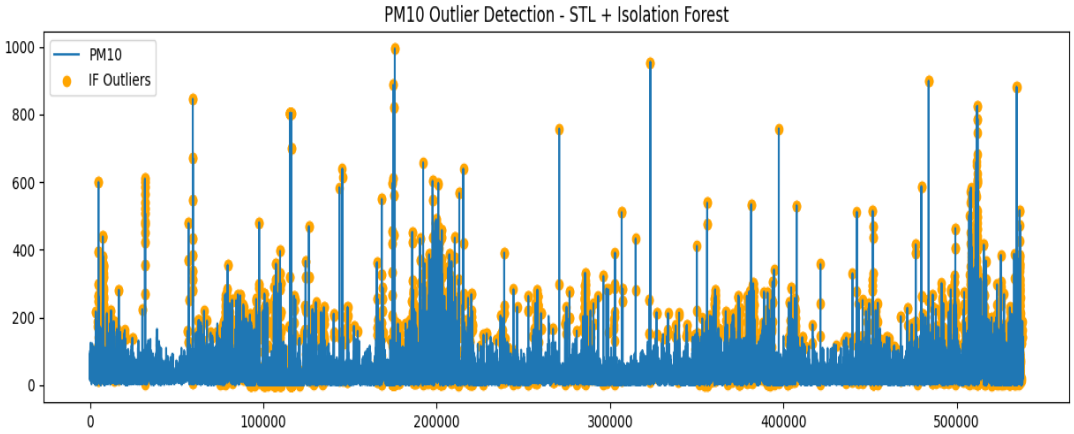


Şekil 17. PM_{10} Değeri için LOF yöntemi ile Aykırı Deger Tespiti

d) Isolation forest (IF)

Isolation Forest (Liu, Ting ve Zhou, 2008), aykırı değerleri diğer veri noktalarından izolasyon yoluyla belirleyen, ağaç tabanlı ve makine öğrenimi tabanlı bir algoritmadır. Temel prebsibi, aykırı değerlerin veri uzayında farklı ve nadir olması nedeniyle rastgele seçilecek daha az bölme ile izole edilebilmesidir. Ortalama ağaç yol uzunluğu kısa olan gözlemler yüksek aykırılık potansiyeline sahiptir. IF, eğitim verisi üzerinde rastgele özellik ve bölme değerleriyle çok sayıda binary ağaç oluşturur. Her gözleme ait ortalama ağaç yol uzunluğu, o noktanın ne kadar çabuk izole edildiğini gösterirken, kısa yol uzunluğu yüksek skoru yani anormal olarak yorumlanmaktadır.

Özellikle yüksek boyutlu veri setlerinde etkinliğiyle öne çıkan IF, bu çalışmada PM_{10} gözlemleri üzerinde uygulanmış ve istatistiksel yöntemlerle belirlenemeyen anomalileri tespit etmede başarı göstermiştir. Zaman serisi analizi için trend, mevsimsellik, döngüsellik ve rastgelelik (residual) bileşenlerine ayrıştırılan veri seti üzerine uygulan Isolation Forest yöntemi sonucu Şekil 18'de, PM_{10} değerine ait aykırı değerler grafik üzerinde sarı renk ile gösterilmiş, kalan PM değerleri grafik üzerinde mavi renkleriyle temsil edilerek sunulmuştur. Bu yöntem ile de 26868 değer belirlenmiştir.



Şekil 18. PM10 Değeri için Isolation Forest ile Aykırı Deger Tespiti

Zaman serisi verilerinde gözlemlene aykırı değerler (outliers), model tahmin sürecinde önemli sapmalara neden olabilecek kritik unsurlardır. Özellikle doğrusal regresyon ve komşuluk tabanlı modeller, uç değerler karşısında yüksek duyarlılık sergileyerek önyargılı tahminlere neden olabilir (Aggarwal, 2017). Bu değerler, ölçüm hatalarından veya olağandışı çevresel koşulların etkisiyle oluşabilir ve modelleme sürecinde dikkate alınmaları gerekli olan bir bileşendir (Zimek, Schubert ve Kriegel, 2012). Aykırı değerlerin modeller üzerindeki etkisi, büyük ölçüde kullanılan yöntemine yapısına bağlıdır.

Buna karşın, ağaç tabanlı modeller ve derin öğrenme mimarileri, parametrik olmayan yapıları sayesinde aykırı değerlere karşı daha dayanıklıdır (Breiman, 2001). Derin öğrenme modelleri, aykırı değerleri veri yapısının doğal bir parçası olarak görebilir. Bu yetenekleri sayesinde bu değerleri çıkarmadan öğrenme sürecine doğrudan entegre edilebilir. Özellikle otomatik özellik çıkarımı yapan mimariler, lokal anomalileri tespit etmede yüksek başarı gösterir (NIST, t.y.).

Veri setinde sistematik bir sensör hatası olmadığı varsayıldığında, tespit edilen aykırılıklar "gürültü" olarak kabul edilmiş ve eğitim verisinden çıkarılmamıştır. Bu yaklaşım, NIST'in veri analizi kılavuzlarında da önerildiği gibi, istatistiksel olarak hatalı olmayan ancak aykırı değer oluşturan gözlemlerin modelden çıkarılmadan analizde tutulmasının daha uygun olacağını önermektedir (Hodge ve Austin (2004); NIST, t.y.).

Sonuç olarak, aykırı veri tespit yöntemleri incelemiş ve her bir yöntem çalışma veri seti üzerinde uygulanmıştır. Ön işleme tabi tutulmuş çalışma veri setinin boyutu 538,131'dir.

Tespit edilen aykırı değer sayısı ise 9,844 ile 36,113 aralığında değişmektedir. Belirlenen değerlerin modelin performansını önemli ölçüde etkileyebileceği dikkate alındığında, gerçek zamanlı verilerde meydana gelen ani değişimlerin (kısa süreli meteorolojik değişiklikler) veya gerçek anomali olaylarının (yangınlar, trafik yoğunluğu artışı, fırtına) modelin öğrenme sürecine dahil edilmesi ve bu değişimlere karşı duyarlı hale gelmesi için veri setinde tutulması önemlidir. Aykırı verilerin modelin kararlılığı üzerindeki olumsuz etkisini azaltmak için, 3.7 bölümünde açıklanan Loess ile mevsimsellik-trend ayrıştırması (Seasonal-Trend Decomposition using Loess, STL) ve uç değerleri tespit edebilen Robust ölçeklendirme yöntemleri kullanılarak bu etkilerin kısmen giderilmesi hedeflenmiştir.

3.6 Veri Normalizasyonu

Makine öğrenmesi ve derin öğrenme algoritmalarının performansı, modelin giriş değişkenlerinin ölçekleriyle doğrudan ilişkilidir. Bu nedenle, farklı değişkenler arasında karşılaştırılabilirlik sağlamak ve gradyan bazlı optimizasyon süreçlerini iyileştirmek amacıyla veri ölçeklendirme adımı uygulanması için aşağıdaki yöntemler incelenmiştir.

a) Min-Max ölçeklendirme

Min-Max normalizasyonu, tüm özellikleri [0, 1] aralığına dönüştürerek modellerin aynı ölçekte giriş almasını sağlar. Özellikle, dağılımı simetrik ve uç değer içermeyen değişkenlerde tercih edilmiştir. Bu yöntem, ağaç tabanlı algoritmalar için uygun olup, performans üzerinde olumlu etki sunmuştur (Han vd., 2011).

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{3.2}$$

b) Z-Skoru Standardizasyonu

Z-skor standardizasyonu, değişkenlerin ortalamasını 0, standart sapmasını 1 olacak şekilde dönüştürür. Verilerin normal dağılıma yakın olduğu durumlarda etkili olmuştur. Bu yöntem, ağın öğrenme sürecini istikrarlı hale getirmiştir (Ioffe ve Szegedy, 2015).

$$Z = \frac{X - \mu}{\sigma} \tag{3.3}$$

c) Robust Ölçeklendirme

Bu çalışmada aykırı değerlerin veri dağılımı üzerindeki etkisini azaltmak amacıyla medyan ve interkuartil aralık (IQR) temelli robust scaler yöntemi tercih edilmiştir. Bu yaklaşım, özellikle uç değerlerin yoğun olduğu değişkenlerde, daha güvenilir bir normalizasyon göstermektedir. Verinin merkezi eğilimi ve IQR temel alarak uç değerleri miznimize etmesi ile makine öğrenmesi modellerinde önyargıyı azaltarak daha kararlı bir performans sağlanmasına katkıda bulunur (Iglewicz ve Hoaglin, 1993; Hodge ve Austin, 2004).

$$X' = \frac{X - Medyan}{IQR} \quad (IQR = Q3 - Q1) \tag{3.4}$$

Ölçekleme yöntemi, hem model mimarisine hem de verinin dağılım özelliklerine göre uygun olan aykırı, uç değerleri yakalayabilen Robust ölçeklendirme uygunlanmıştır.

3.7 Zaman Serisi Analizi (Time Series Analysis)

Zaman serisi analizi, belirli zaman aralıklarında toplanan verilerin yapısal bileşenlerine ayrılarak incelenmesini sağlayan bir yöntemdir. Hava kirliliği gibi çevresel veriler, genellikle eğilim mevsimsellik, döngüsellik ve rastgelelik bileşenlerini içermektedir. Bu

bileşenlerin sistematik şekilde ayrıştırılması, kirlilik seviyelerinin dinamik yapısının anlaşılması ve gelecekteki değerlerin yüksek doğrulukla tahmin edilmesi açısından kritik öneme sahiptir (Hyndman ve Athanasopoulos, 2018). Bu çalışma kapsamında, PM_{10} konsantrasyonuna yönelik zaman serisi analizi gerçekleştirilmiş ve üç ana bileşen detaylı olarak incelenmiştir.

a) Trend Bileşeni (Trend Component)

Trend bileşeni, zaman serisindeki uzun vadeli yapısal değişimleri göstermektedir. Bu değişimler genellikle çevresel faktörler, demografik artış veya ekonomik büyüme gibi etkenlerle şekillenir. Bu çalışmada, PM_{10} konsantrasyonları üzerine bir trend analizi yapıldı. Sonuçlar, özellikle 2022 yılı itibarıyla PM_{10} konsantrasyonlarında belirgin bir artış eğilimi olduğunu göstermiştir. Bu eğilim, kentleşmeye bağlı nüfus artışı, sanayi faaliyetlerinin yoğunlaşması, artan motorlu taşıt kullanımı gibi etkenlerle ilişkilendirilebilir. STL (Seasonal-Trend Decomposition using Loess) yöntemi kullanılarak, trend bileşeni, hareketli ortalamalar, düşük geçişli filtreler ve Loess kullanılarak mevsimsellik ve trend ayrıştırması yapılmıştır. Loess yöntemi, zaman serisini trend, mevsimsellik ve rastgelelik olarak üç bileşene ayıran esnek ve doğrusal olmayan bir ayrışma tekniğidir. Bu yöntemin en büyük avantajı ise, aykırı değerlere karşı dayanıklı olmasıdır (Cleveland, Cleveland, McRae, Terpenning vd., 1990).

b) Mevsimsellik Bileşeni (Seasonal Component)

Mevsimsel bileşen, yılın belirli zamanlarında tekrarlayan ve tahmin edilebilir dalgalanmaları temsil etmektedir. Hava kirliliği mevsimsel faktörlerden etkilenmektedir. Kışın, ısınma nedeniyle emisyon artışları gözlenirken, yazın ise fotokimyasal reaksiyonlar ozon seviyelerini değiştirir. Analiz sonuçlarına göre, özellikle kış aylarında PM_{10} konsantrasyonlarının mevsimsel olarak arttığını göstermektedir. Bu mevsimsel yapı bileşenin çıkarımı STL yöntemiyle sağlanmış olup, mevsimsel desenler net bir şekilde sergilenmiştir (Hyndman ve Athanasopoulos, 2018).

c) Rastgelelik Bileşeni (Residual Component)

Trend ve mevsimsellik etkileri çıkarıldıktan sonra, seride kalan varyasyonlar rastgele bileşen olarak ele alınır. Bu varyasyonlar sistematik olarak açıklanamayan olarak değerlendirilir. Bu bileşen, rüzgar yönündeki ani değişiklikler ve fırtına gibi beklenmedik hava olaylarından etkilenir. Ayrıca, yeni bir inşaat faaliyenin başlaması gibi lokal emisyon kaynaklarındaki artışlar ve ölçüm hataları da etkili olabilir (Jiang, Qiao, Jiang ve Tian, 2021). Bu çalışmada elde edilen residual bileşen, bazı zaman dilimlerinde yüksek varyans göstermektedir. Bu durum, modelin mevcut girdilerle açıklayamadığı dışsal etkilerin olduğunu göstermektedir. Bu durum, tahmin modelinin performansını arttırmak için ek meteorolojik ve çevresel değişkenlerin (rüzgar yönü, yağış miktarı, trafik yoğunluğu gibi değişkenlerin) modele eklenmesi gerektiğini göstermektedir.

A. Zaman serisi bileşenlerinin incelemesi

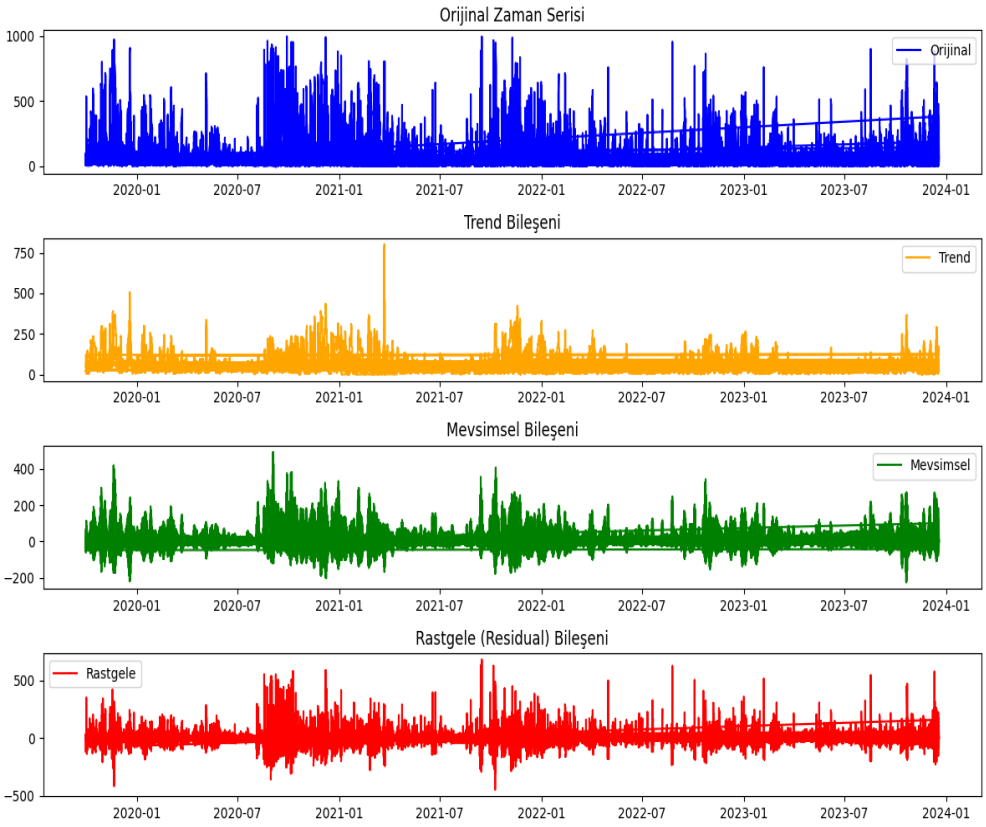
Zaman serisi analizinin görselleştirilmesi, her bileşenin dinamiklerini daha iyi anlamamıza yardımcı olur. Bu çerçevede, Şekil 19'da yer alan grafiksel zaman serisi bileşenlerinin ayrıştırma sonuçları aşağıdaki listelenmiştir.

a) Orijinal Seri (Mavi): PM_{10} seviyeleri zaman içinde dalgalı bir seyir sergilemektedir. 2022–2023 döneminde, özellikle kısa süreli fakat yüksek yoğunlukta pikler dikkat çekmektedir.

b) Trend Bileşeni (Turuncu): Zamanla artan bir yönelim gözlenmektedir. Bu eğilim, 2022 yılının ortalarından itibaren hız kazanmıştır. Bölgesel şehirleşme ve sanayileşme eğilimleriyle korelasyon gösterdiği düşünülmektedir.

c) Mevsimsel Bileşen (Yeşil): Yıl boyunca tekrarlayan döngüsel yapılar gözlemlenmektedir. Bu yapılar, sıcaklık, nem, ısınma ihtiyacı gibi meteorolojik değişkenlerle yüksek korelasyon içerisindedir.

d) Rastgelelik /Residual Bileşen (Kırmızı): Açıklanamayan sapmalar, modelin bazı dışsal faktörlere duyarlı olabileceğini göstermektedir. Bu sapmalar, istasyon bazında lokal emisyon kaynakları ve ani meteorolojik etkilerle ilişkilendirilebilir.



Şekil 19. Zaman Serisi Bileşenlerinin Görsel Analizi

4. VERİ KEŞFİ (EXPLORATORY DATA ANALYSIS)

Bu bölümde, Ankara'daki meteoroloji ve hava kalitesi izleme istasyonlarından elde edilen verilerin keşifsel analizi gerçekleştirilecektir. Hava kirleticileri arasındaki ilişkilerin, zaman içindeki dağılım kalıplarının ve istasyonlar arası benzerliklerin ortaya konulması amaçlanmaktadır. Veri keşfi aşaması, hava kalitesi verilerinin analizinde temel bir adım olup, elde edilen bulgular, hava kirliliği kaynaklarının belirlenmesi ve kontrol stratejilerinin geliştirilmesi açısından önemli içgörüler sunmaktadır. Bu bulgular, ilerleyen bölümlerde model geliştirme süreçlerine ışık tutacak ve hava kalitesi tahminlerinin doğruluğunu arttıracaktır.

4.1 Hava Kalitesi Verilerinin Genel Durumu

Bu bölümde, çalışmada kullanılan hava kalitesi verilerinin temel istatistiksel özellikleri ve dağılım yapısı incelenmiştir. Veri seti, PM_{10}, $PM_{2.5}$, SO_2, CO, NO_2, NO_x, NO ve O_3 gibi hava kirleticilerini içermektedir. Her bir parametre için aykırı değer (outlier) analizleri gerçekleştirilmiş ve ölçüm değerlerinin dağılımları grafikler ile görselleştirilmiştir.

Şekil 20, analiz edilen hava kalitesi verilerinin temel istatistiksel özelliklerini özet olarak sunmaktadır. Bu analiz sonucu, her bir parametre için ortalama (mean), standart sapma (std), minimum (min) ve maksimum (max) yanı sıra çeyrek değerleri (25%, 50%, 75%) içeren tanımlayıcı istatistikleri kapsamaktadır. Elde edilen bu bulgular, hava kirleticilerinin dağılım özelliklerini sayısal olarak ortaya koyarak veri setinin merkezi eğilim ve yayılım ölçütleri üzerinden değerlendirilmesine olanak tanımaktadır. Özellikle bazı kirleticilerin maksimum değerlerinin ortalamadan oldukça uzaklaşması, belirli zaman dilimlerinde ekstrem konsantrasyon artışlarının yaşandığını ve bu durumun hava kalitesinin değerlendirilmesinde dikkate alınması gereken kritik bir unsur olduğunu göstermektedir.

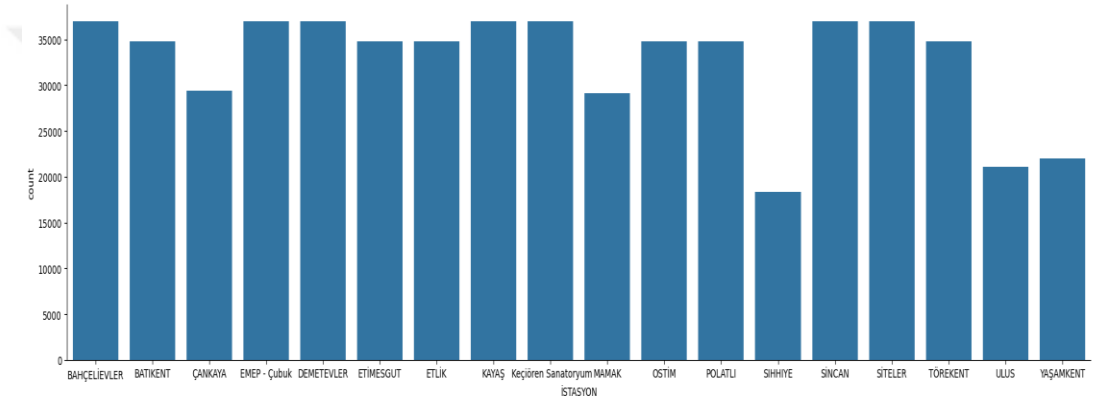
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 587238 entries, 0 to 587237
Data columns (total 11 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   İSTASYON          587238 non-null  object
 1   PM10 ( µg/m3 )    487632 non-null  float64
 2   PM 2.5 ( µg/m3 )  325454 non-null  float64
 3   SO2 ( µg/m3 )     453984 non-null  float64
 4   CO ( µg/m3 )      220182 non-null  float64
 5   NO2 ( µg/m3 )     377893 non-null  float64
 6   NOX ( µg/m3 )     374502 non-null  float64
 7   NO ( µg/m3 )      378004 non-null  float64
 8   O3 ( µg/m3 )      183272 non-null  float64
 9   Tarih             587238 non-null  datetime64[ns]
 10  Saat              587238 non-null  object
dtypes: datetime64[ns](1), float64(8), object(2)
memory usage: 49.3+ MB
None
       PM10 ( µg/m3 )  PM 2.5 ( µg/m3 )  SO2 ( µg/m3 )   CO ( µg/m3 )
count   487632.000000     325454.000000  453984.000000  220182.000000
mean        42.171599         17.252116       7.427162     346.603845
min          0.000000          0.000000       0.000000       0.000000
25%         17.290000          4.980000       2.470000      10.133500
50%         30.820000         10.530000       5.160000     308.545000
75%         52.250000         20.300000      10.160000     571.210000
max        995.110000        982.280000     962.260000     999.990000
std         45.731199         26.355597       8.347012     293.138906

       NO2 ( µg/m3 )  NOX ( µg/m3 )   NO ( µg/m3 )   O3 ( µg/m3 )  \
count  377893.000000  374502.000000  378004.000000  183272.000000
mean       53.761545      90.181689      37.549015      36.076274
min         0.000000       0.000000       0.000000       0.000000
25%        18.880000      25.100000       5.000000       9.870000
50%        38.570000      49.790000      11.520000      26.100000
75%        72.460000     111.090000      35.730000      54.070000
max       983.910000     999.870000     990.060000     828.310000
std        53.908566     110.974301      69.631890      34.528443

                              Tarih
count                        587238
mean   2022-01-15 21:15:04.073645824
min              2019-10-01 00:00:00
25%              2021-01-30 00:00:00
50%              2022-02-07 00:00:00
75%              2023-01-13 00:00:00
max              2023-12-19 00:00:00
std                              NaN
```

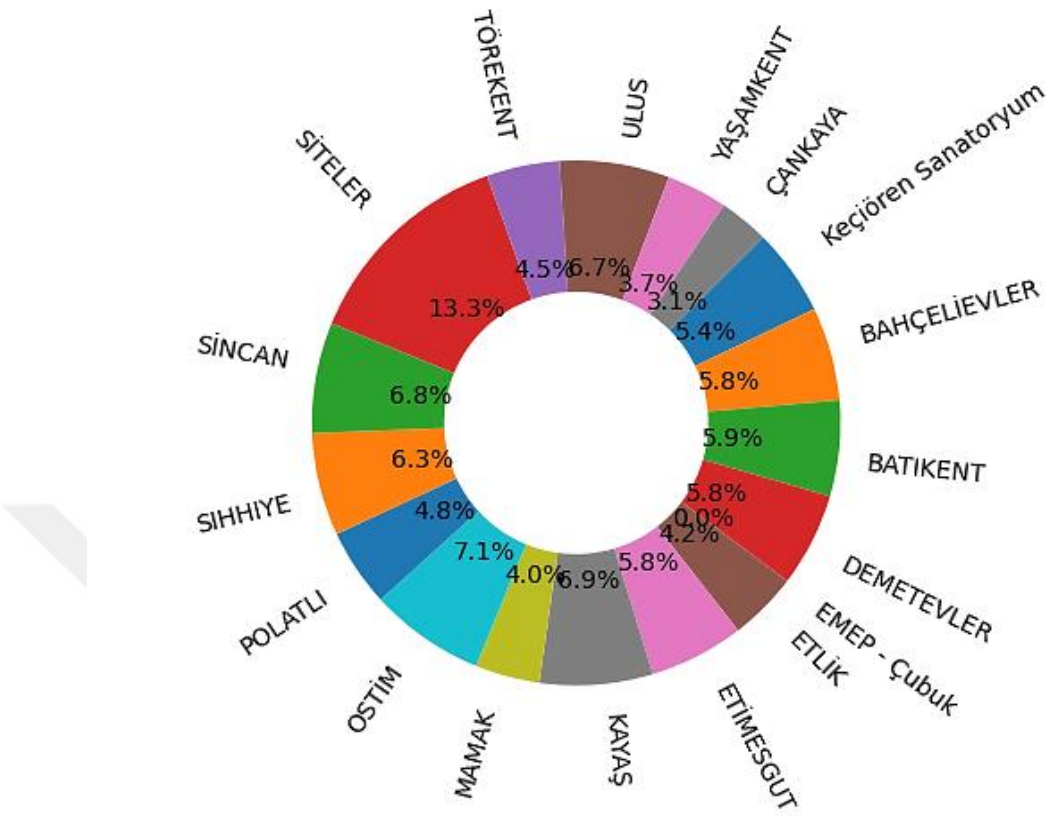
Şekil 20. Hava Kalitesi Verilerinin İstatiksel Analizi

Şekil 21’de, hava kalitesi izleme istasyonlarına ait ölçüm sıklıkları istasyon bazında görselleştirilmiştir. İlgili grafik, her bir istasyonun belirli bir zaman aralığında sağladığı veri miktarını açık biçimde ortaya koyarak, veri toplama süreçlerinin düzenliliği ve sürekliliği hakkında önemli bir gösterge sunmaktadır. İstasyonlar arasında gözlemlenen ölçüm sayısı farklılıkları, hem uzamsal hem de zamansal açıdan homojen ve güvenilir bir izleme sisteminin gerekliliğine işaret etmektedir. Bu bağlamda, eksik veri kaynaklı belirsizliklerin azaltılması amacıyla çeşitli veri tamamlama (data imputation) yöntemleri uygulanmıştır.



Şekil 21. Hava Kalitesi Verilerinin İstasyonlara Göre Doluluk Oranı

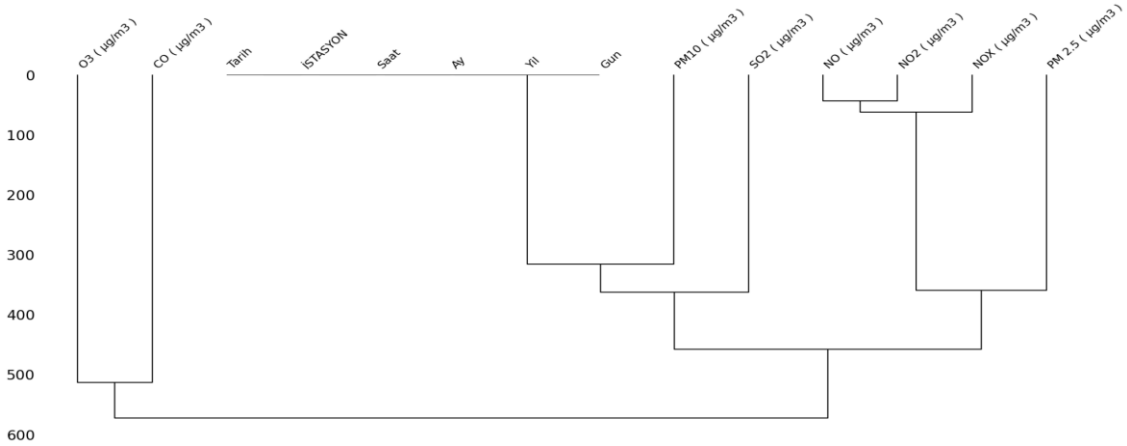
Buna ek olarak, her bir kirleticinin yıllık ortalama değerlerinin istasyonlara göre dağılım oranları da incelenmiştir. Özellikle Şekil 22’de, yıllık ortalama PM10 konsantrasyonlarının istasyon bazlı yüzde (%) dağılımları grafiksel olarak sunularak, mekânsal ölçekte hava kalitesi düzeyleri arasındaki farklılıklar açık biçimde ortaya konulmuştur.



Şekil 22. İstasyonlara Göre Yıllık Ortalama PM10 Dağılım Grafiği

4.2 Dendrogram ile Değişkenler Arası Benzerlik Analizi

Şekil 23'de sunulan dendrogram, hava kirleticileri arasındaki benzerlik ve ilişkileri hiyerarşik kümeleme (hierarchical clustering) analizi yoluyla görselleştirmektedir.



Şekil 23. Hava Kalitesi Verilerinin Dendrogram Analizi

Bu analiz, kirleticilerin birlikte değişim eğilimlerini ve potansiyel ortak kaynaklarını ortaya koyarak, çok değişkenli yapının yapısal desenlerini analiz etmeye olanak tanımaktadır.

Dendrogramda yer alan gruplama, değişkenler arasındaki ilişkilerin derinlemesine analizi için önemli bir araçtır. Örneğin, O_3 ve CO'nun benzer gruplar içinde yer alması, bu kirleticilerin ortak kaynaklardan kaynaklanabileceğine dair bir ipucu sunmaktadır. Zaman ve konum değişkenleri de belirli bir küme oluşturarak, hava kalitesinin mekansal ve zamansal dinamiklerini yansıtmaktadır.

a) Genel gruplama bulguları

O_3 (µg/m³) ve CO (µg/m³) konsantrasyonları, erken aşamada ayrılarak kendi içlerinde bir grup oluşturmuş, bu durum O_3 ve CO'nun benzer özelliklere sahip olduğunu veya ortak kaynaklardan geldiğini düşündürmektedir.

Zaman ve Konum Değişkenleri: Tarih, istasyon, saat, ay, yıl ve gün değişkenleri, birlikte büyük bir küme oluşturarak zamana ve konuma bağlı yapısal benzerlikleri göstermektedir.

PM_{10} (µg/m³) konsantrasyonu başlangıçta zaman ve konum değişkenleriyle gruplandıktan sonra diğer kirleticilere katılarak zamansal etkilerden yoğun şekilde etkilendiğini göstermektedir.

SO_2, NO, NO_2, NOx ve $PM_{2.5}$ kirleticileri benzer yapıda alt gruplar halinde birleşmiş olup, potansiyel olarak ortak emisyon kaynaklarına veya atmosferik süreçlere işaret etmektedir.

b) Dikey eksen yorumu (Benzerlik ölçeği)

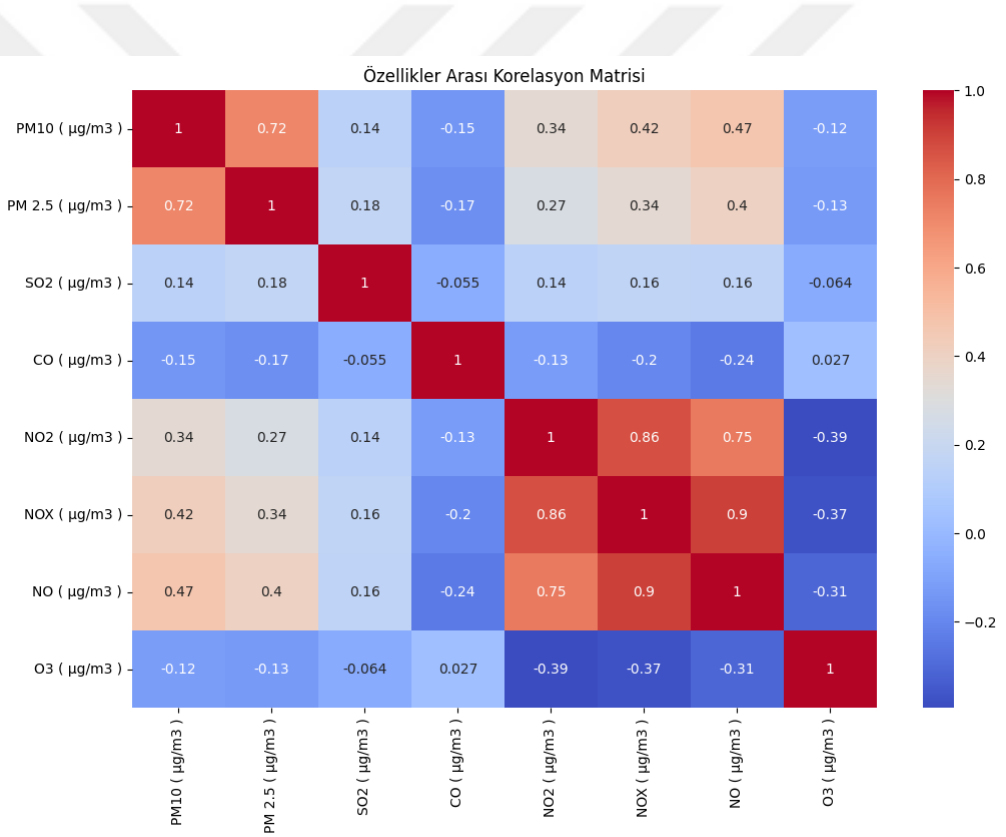
Dikey eksen, kümeler arasındaki benzerlik veya uzaklık ölçüsünü göstermektedir. Değer ne kadar küçükse, kümeler o kadar birbirine benzer.

- 0–200 aralığında: O_3 ve CO kirleticileri birlikte gruplanmıştır.
- 200–400 aralığında: Zaman ve konum değişkenleri birleşmiştir.
- 400–600 aralığında: Kirleticiler genel bir küme oluşturarak bütünleşmiştir.

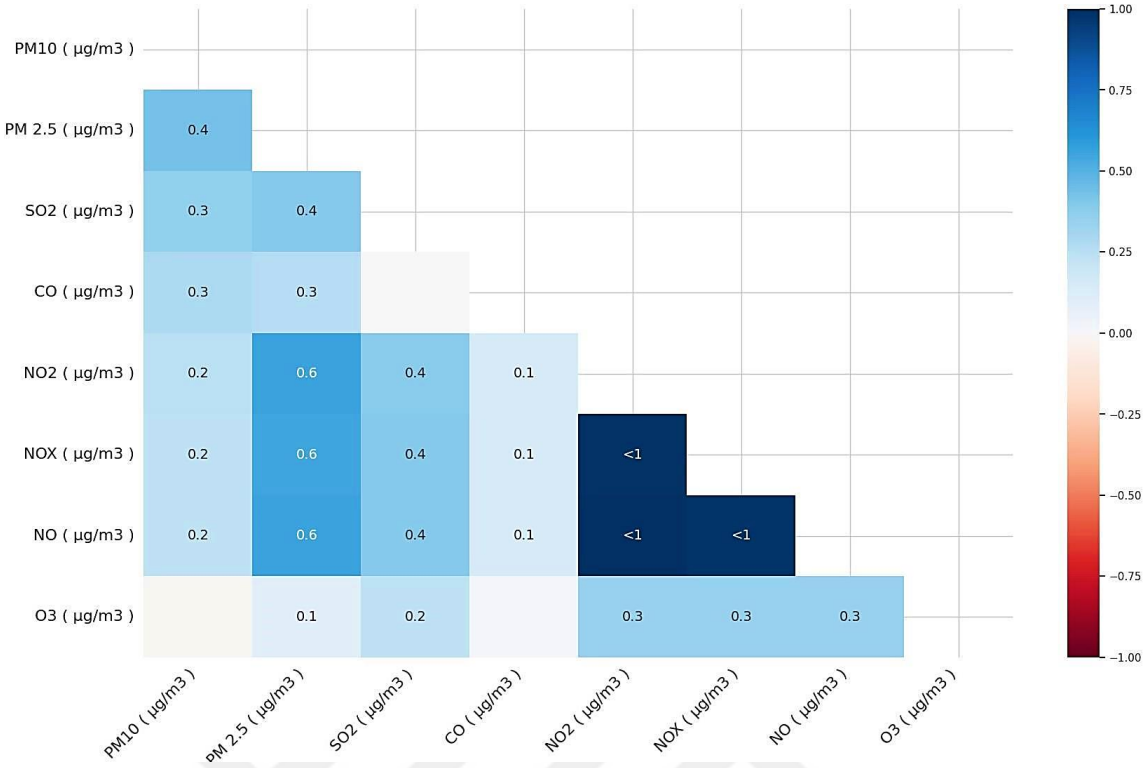
Sonuç olarak, dendrogram analizi, hem zamansal/mekansal değişkenlerin hem de kirleticilerin kendi içlerinde kümelenme eğilimini ortaya koyarak çok değişkenli yapıdaki verilerde önemli içgörüler sunmaktadır.

4.3 Hava Kirletici Değerlerinde Korelasyon Analizi

Hava kirletici arasındaki ilişkiyi analiz ettiğimizde Şekil 24 ve 25'de sunulan korelasyon matrisi ve çubuk grafik, hava kirleticileri arasındaki ilişkilerin gücünü ve yönünü göstermektedir. Korelasyon katsayıları, -1 ile +1 arasında değer alarak iki değişken arasındaki doğrusal ilişkinin derecesini temsil etmektedir.



Şekil 24. Hava Kirleticilerin Korelasyon Analizi



Şekil 25. Hava Kirleticiler Arası Korelasyon Analizi Çubuk Grafiği

Korelasyon analizi sonucundaki genel bulgular incelenmiştir. Verinin büyük kısmında pozitif korelasyonlar gözlemlenmiştir. Bir kirleticinin artışının diğer kirleticiler de artış eğilimi yarattığı görülmektedir. Ayrıca, anlamlı negatif korelasyonlar gözlemlenmemiştir. Korelasyon katsayı değerleri genellikle 0.1 ila 0.6 aralığında değişmekte olup zayıf ile orta düzeyde ilişkiler olduğunu göstermektedir.

Seçili Korelasyonlar:

- PM_{10} – $PM_{2.5}$: 0.4 ile orta derecede güçlü pozitif ilişki
- $PM_{2.5}$ – NO_2, NOx, NO: 0.6 ile kuvvetli pozitif ilişki
- SO_2 – NO_2, NO_x, NO: 0.4 ile orta düzeyde ilişki
- CO – NO_2, NOx, NO: 0.4 ile ortak kaynak ihtimali
- NO_2 – NO_x – NO: 1.0 ile çok güçlü pozitif ilişki
- O_3 – diğer gazlar: 0.1–0.3 aralığında zayıf ilişki

4.4 Hava Kirliliği Değerlendirmesi ve Kirlilik Kontrol Stratejilerine Etkisi

Hava kirliliği bileşenleri arasındaki korelasyonlar, kirlilik kaynaklarının belirlenmesi, sağlık risklerinin değerlendirilmesi ve entegre kontrol politikalarının geliştirilmesi açısından stratejik önem taşımaktadır. Özellikle partikül madde (PM_{10} ve $PM_{2.5}$) ile gaz formundaki kirleticiler (NO_2, SO_2, CO, O_3) arasındaki istatistiksel ilişkiler, kirlilik kontrol stratejilerinin çok yönlü olarak yapılandırılması için önemli bilgiler sağlamaktadır.

A. Ortak kaynakların tanımlanması

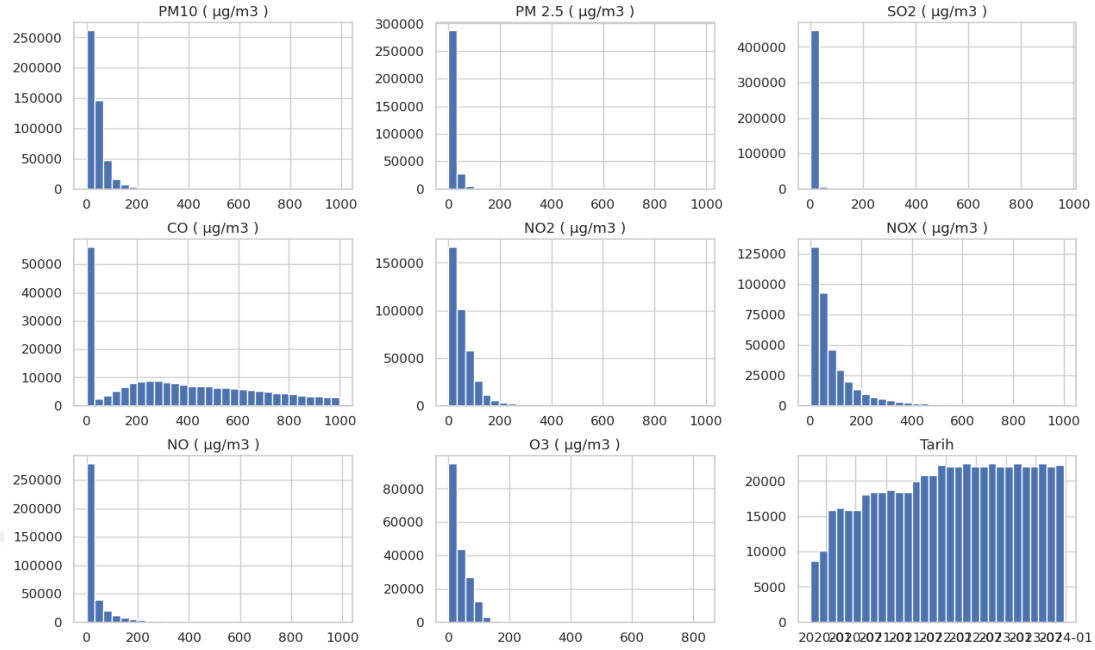
Kirleticiler arasındaki güçlü pozitif korelasyonlar, genellikle bu bileşenlerin ortak bir emisyon kaynağından yayıldığını göstermektedir.

Trafik Emisyonları: NO_2, NOx, CO ve $PM_{2.5}$ arasındaki yüksek korelasyon, bu kirleticilerin büyük ölçüde taşıtlardan kaynaklandığını göstermektedir. Bu durumda, egzoz gazlarının kontrolü ve elektrikli araç kullanımının teşvik edilmesi gibi önlemler etkili olabilmektedir.

Sanayi Emisyonları: SO_2 ile PM_{10} arasındaki pozitif korelasyon, bu iki kirleticinin endüstriyel süreçlerden yayıldığını göstermektedir. Temiz üretim teknolojilerinin teşviki ve emisyon limitlerinin sıkılaştırılması faaliyetlerine öncelik verilmesi gerektiği sonucuna varılmıştır.

4.5 Hava Kalitesi İstasyonlarından Ölçülen Değerlerin Histogram Analizi

Şekil 26, hava kalitesi izleme istasyonlarından elde edilen kirletici verilerinin frekans dağılımlarını histogramlar aracılığıyla sunmaktadır. Histogramlar, belirli zaman dilimlerinde ölçülen kirletici konsantrasyonlarının sıklığını görselleştirerek verilerin merkezi eğilimi, yayılımı, çarpıklığı ve potansiyel aykırı değerleri hakkında çıkarımlar yapmamıza olanak tanımaktadır.



Şekil 26. Hava Kalitesi İstasyon Verilerinin Dağılımı (Histogram Analizi)

- PM_{10} (µg/m³) konsantrasyon sağa çarpık bir dağılım göstermektedir. Bu kirleticinin ölçümleri çoğunlukla 200 µg/m³ altında yoğunlaşmıştır.

- $PM_{2.5}$ (µg/m³) konsantrasyonunda sağa çarpık dağılım görülmektedir. Ölçümler genellikle 100 µg/m³ altındadır. Bununla birlikte, 200 µg/m³ üzerindeki değerler nadir bulunmaktadır.

- SO_2 (µg/m³) konsantrasyonunda belirgin bir sağa çarpık dağılım gösterir. Kükürt dioksit değerlerinin çoğu 100 µg/m³ altında toplandığı görülmüştür.

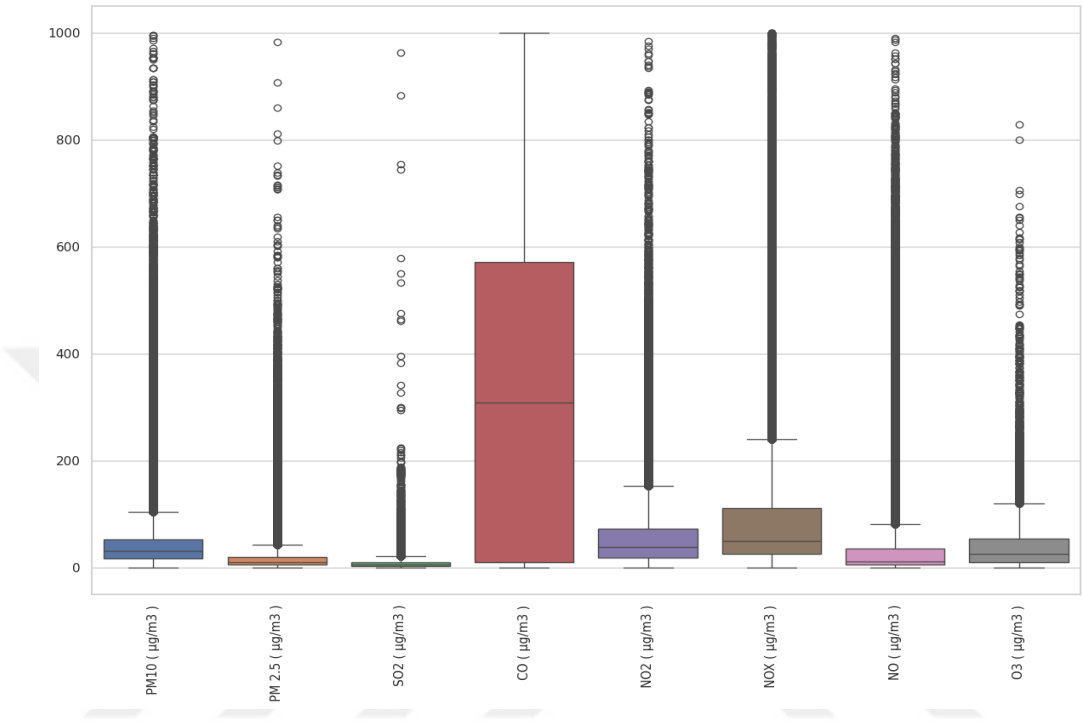
- CO (µg/m³) konsantrasyonunda, daha dengeli ama sağa çarpık bir dağılım görülmüştür. CO konsantrasyon değerleri, 800 µg/m³'e kadar geniş bir aralıkta dağılmaktadır.

- NO_2, NO_x, NO ve O_3 seviyelerinde, sağa çarpık dağılım sergileyerek, genellikle 100 µg/m³ altında yoğunlaşmışlardır.

- Tarih verisi histogramı, zamanla artan bir örneklem sıklığını sergiliyor. Veriler tutarlı görünüyor. Zaman serilerinde belirgin boşlukların olmaması, güvenilir bir izlemenin göstergesidir.

4.6. Hava Kalitesi Verilerinin İlişkisel İncelemesi

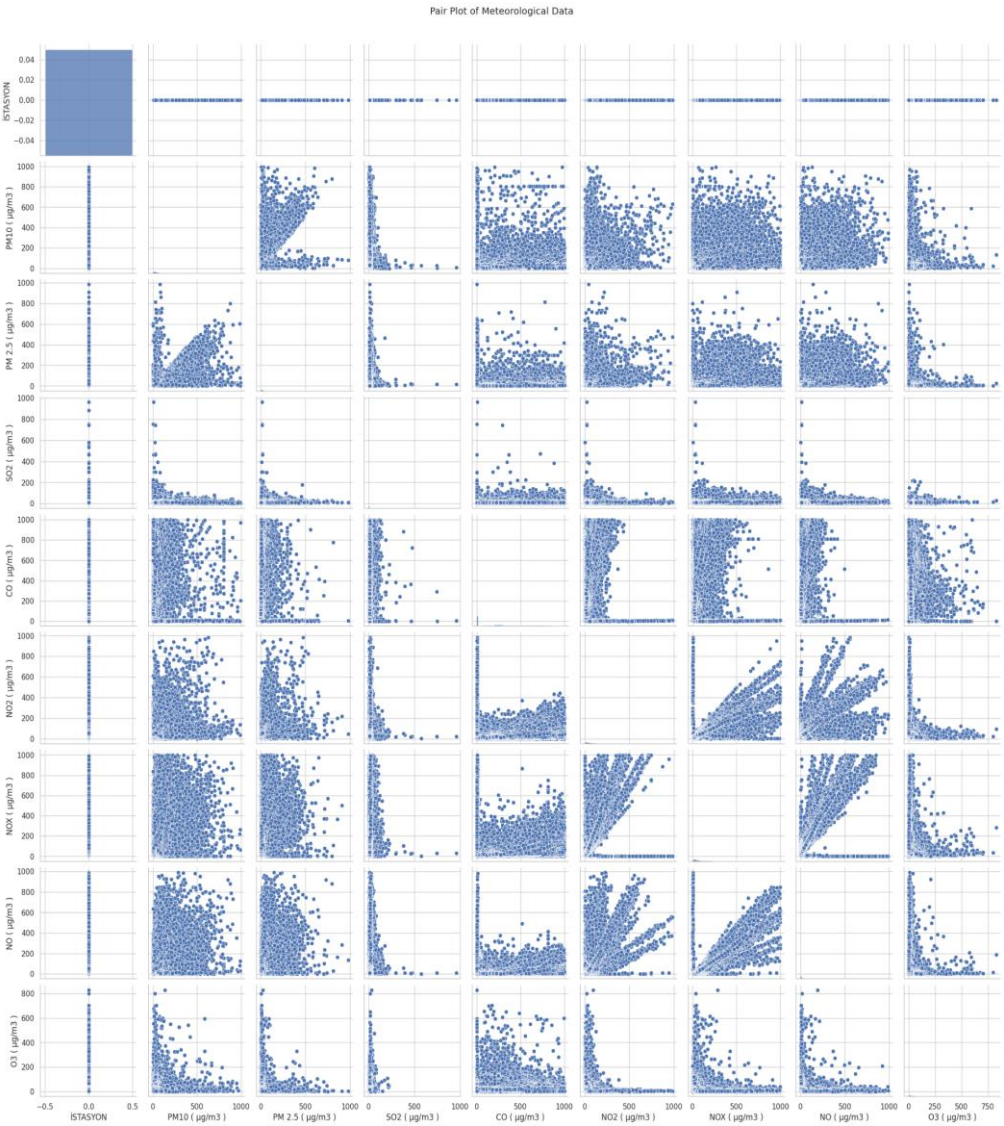
Boxplot görselleştirmeleri, her bir kirletici için minimum, maksimum, medyan, alt ve üst çeyrekleri göstererek veri kümesindeki aykırı değerlerin tespiti için güçlü bir araç sunar.



Şekil 27. Hava Kirleticilerin Aykırı Değer Analizi (Boxplot Grafik Analizi)

Hava kirletici verileri üzerinde boxplot grafiği oluşturulmuştur. Şekil 27'de elde ettiğimiz sonucu incelediğimizde aşağıda öne çıkan bulgular yer almaktadır:

- PM_{10} ve $PM_{2.5}$: Her iki kirletici için de kutu dışına çıkan çok sayıda aykırı değer gözlemlenmiştir. Bu değerler mevsimsel olaylar, trafik yoğunluğu veya sanayi faaliyetlerindeki ani değişimlere işaret edebilir.
- CO: Diğer kirleticilere göre daha uzun kuyruklar ve daha fazla aykırı değer gözlemlenmiştir. Bu durum, karbonmonoksit düzeylerinin farklı kaynaklardan etkilenerek değişkenlik gösterdiğini ortaya koymaktadır.
- SO_2, NO_2, NO_X, NO, O_3: Görece daha kompakt kutulara rağmen birkaç aşırı uç değer mevcuttur. Bu durum, genel olarak istikrarlı bir çevre havasının mevcut olduğunu ancak nadiren lokal kirlenme olaylarının yaşandığını göstermektedir.



Şekil 28. Hava Kalitesi Verilerinin İlişkisel İncelemesi (Pair Plot Analizi)

Şekil 28'de elde edilen pair plot analizi grafiği, kirleticiler arasındaki çift yönlü ilişkileri hem dağılım hem de korelasyon düzeyinde değerlendirerek potansiyel etkileşimleri ve ortak kaynakları incelemek için kullanılmıştır.

- NO_2-NO_X-NO: Bu üç bileşen arasında yüksek korelasyon gözlemlenmiştir. Bu durum, aynı kaynaktan (örneğin taşıt emisyonlarından) türeyen azot türevli kirleticilerin birlikte arttığını göstermektedir.
- PM_{10}-$PM_{2.5}$: Yüksek pozitif korelasyon, bu iki partikül madde türünün büyük oranda ortak kaynaklara sahip olduğunu ve birlikte arttığını ortaya koymaktadır.

- SO_2-CO: Zayıf ancak anlamlı ilişkiler, sanayi ve trafik kaynaklı emisyonların bu kirleticileri birlikte etkileyebileceğini düşündürmektedir.

a) Veri kalitesi ve zaman tutarlılığı

Verilerin zaman boyunca homojen şekilde toplandığı gözlemlenmiştir. Büyük eksikliklerin bulunmaması, istasyonların etkin şekilde çalıştığını ve güvenilir bir veri seti elde edildiğini göstermektedir.

b) İstatistiksel özellikler ve çarpıklık

Çoğu kirletici sağa çarpık dağılım sergilemektedir. Bu çarpıklık, istatistiksel analizler açısından dönüşüm (ör. logaritmik) gerektirebilir. Çarpık yapılar ayrıca, istisnai olaylara (örneğin mevsimsel kirlilik, meteorolojik faktörler) karşı duyarlılığı da ortaya koymaktadır.

c) Çevresel ve Sağlık Açısından Çıkarımlar

Genel olarak düşük kirletici seviyeleri, istasyon çevresinde iyi hava kalitesi profili çizmektedir. Ancak nadiren yaşanan yüksek konsantrasyon olayları, halk sağlığı açısından dikkatle izlenmeli ve kontrol stratejileriyle desteklenmelidir. CO'nun değişkenliği ise trafik veya endüstriyel kaynakların bölgesel etkisini göstermektedir.

4.7. Hava Kalitesi Konsantrasyonlarının Yıllık Değişim ve Uzun Dönem Eğilim Analizleri

Bu bölümde, 2019–2023 arasındaki dört yıllık dönemde PM_{10}, $PM_{2.5}$ ve SO_2 kirleticilerinin yıllık bazlı yüzde değişimleri ile uzun döneme yayılan trend yönelimleri incelenmiştir. Analizler, pandemi kaynaklı kısıtlamalar ve sonrasındaki ekonomik canlanma süreçlerinin hava kalitesi üzerindeki kısa ve uzun dönem etkilerini açığa çıkarmak ve emisyon kontrol politikalarının etkinliğini nesnel olarak değerlendirmek amacıyla tasarlanmıştır.

Çalışmada, her bir kirletici için yıllık konsantrasyon değerlerinin bir önceki yıla göre yüzde değişimi hesaplanmıştır.

$$\Delta_{\%}^{t} = \frac{C^{t} - C^{t-1}}{C^{t-1}} \times 100 \tag{4.1}$$

Burada Ct ve Ct-1 sırasıyla t ve t−1 yıllarındaki ortalama yıllık konsantrasyon değerlerini temsil etmektedir. Yüzde değişim serileri elde edildikten sonra, her bir kirleticiye ilişkin uzun dönem eğilimi belirlemek için basit doğrusal regresyon modeli uygulanmıştır.

$$\Delta_{\%}^{t} = \beta_0 + \beta_1 t + \varepsilon_t \tag{4.2}$$

Burada β1 katsayısı, yıllık yüzde değişimdeki ortalama eğilim eğimini (trend yönelimini) göstermekte, εt ise hata terimini temsil etmektedir.

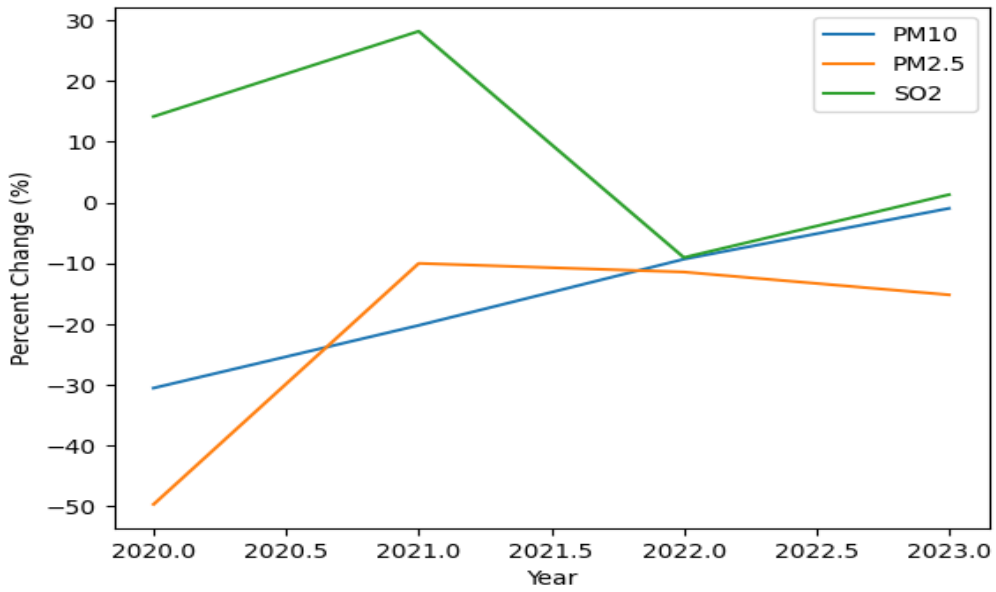
a) Yıllık yüzde değişim analizi

PM_{10}, $PM_{2.5}$, SO_2 verileri üzerinde yıllık yüzde değişim grafiği oluşturularak Şekil 31'de sunulmuştur. Uzun dönem trend analizleri incelendiğinde, doğrusal regresyon sonuçlarına göre elde edilen β1 katsayıları aşağıdaki gibi gözlemlenmiştir.

- PM_{10}: β1≈+1.5% yıllık artış yönelimi
- $PM_{2.5}$: β1≈−2.0% yıllık azalış yönelimi
- SO_2: β1≈+0.5% hafif artış yönelimi (2022'deki anomalinin etkisiyle belirginleşmiştir)

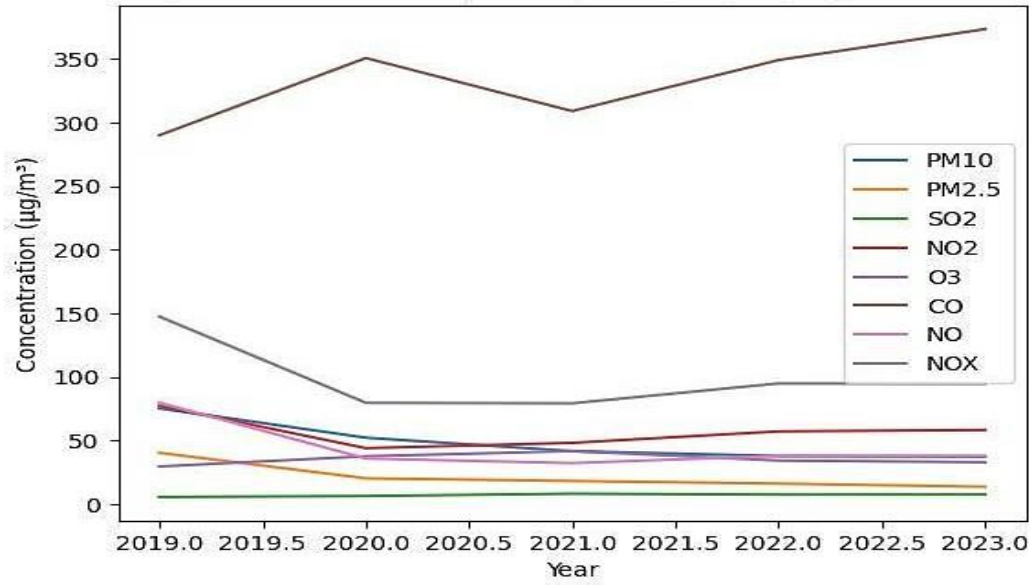
Analiz edilen beş yıllık dönemde, COVID-19 kısıtlamalarının yarattığı ani emisyon azalmaları ve sonrasındaki toparlanma dinamikleri, hava kalitesi yönetiminde hem kısa vadeli kriz müdahalelerinin hem de uzun vadeli sürdürülebilir stratejilerin dengeli bir şekilde yürütülmesi gerekliliğini vurgulamaktadır. $PM_{2.5}$ için yakalanan uzun dönem azalma trendi, devam eden teknolojik yatırımların ve denetim mekanizmalarının

etkinliğini desteklerken; PM_{10} ve SO_2 için güçlendirilecek ilave tedbirler ihtiyacını işaret etmektedir. Bundan sonraki çalışmalarda, meteorolojik değişkenlerin kontrol edildiği çok değişkenli modellerle benzer eğilim analizlerinin tekrarlanması, elde edilen bulguların güvenilirliğini daha da artıracaktır.



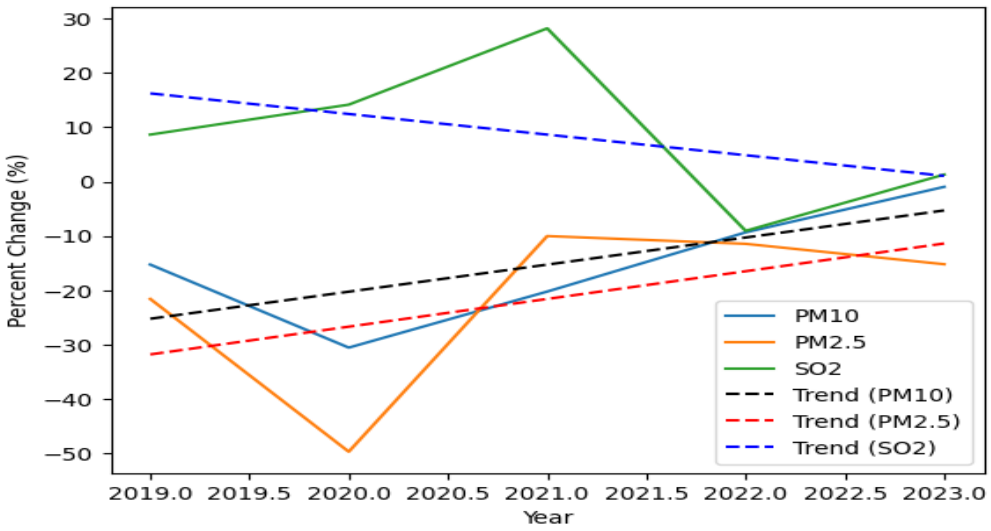
Şekil 29. Hava Kirleticilerin Yıl Bazında Yüzdelik Değişim Grafiği

Şekil 29'da 2020–2023 döneminde PM_{10}, $PM_{2.5}$ ve SO_2 kirletici parametrelerinin bir önceki yıla göre yüzdesel değişimleri sunulmaktadır. PM_{10} değeri, 2020 yılında %-30'dan başlayarak izleyen yıllarda kademeli bir iyileşme göstermiş; 2021'de yaklaşık %-20, 2022'de %-10 ve nihayet 2023'te %0 seviyesine ulaşarak pandemi kaynaklı kısıtlamaların azalması ve endüstriyel faaliyetlerin kısmi toparlanmasının partikül madde düzeylerini göreceli olarak artırdığını işaret etmektedir. $PM_{2.5}$, 2020'de %-50'ye varan dramatik bir düşüş sergiledikten sonra, 2021'de %-10'a kadar toparlanmış; 2022 ve 2023'te yine negatif bölgede kalarak sırasıyla %-12 ve %-15 değerleri ile hava kalitesindeki salınımların daha ince partikül fraksiyonlarında devam ettiğini göstermektedir. SO_2 ise 2020–2021 arasında %15'ten %28'e yükselerek yüksek dalgalanma eğilimi sergilemiş; bunu izleyen yıllarda 2022'de %-10'a kadar gerileme, 2023'te ise hafif bir toparlanma (%+2) ile izleyerek sülfür dioksit emisyonlarının kontrolünde dalgalı ancak genel olarak iyileşme yönünde bir seyir izlediğini ortaya koymaktadır.



Şekil 30. Kirleticilerin Yıllık Konsantrasyon Grafiği

Şekil 30'da tüm kirleticlerin yıllık konsantrasyon grafiği sunulmuştur. Bu analizde, 2019–2023 dönemi boyunca kirletici gaz ve partikül madde konsantrasyonlarında belirgin ayrışmalar gözlenmektedir. Karbonmonoksit (CO) en yüksek seviyeyi koruyarak 2019'da 290 µg/m³'ten 2020'de 350 µg/m³'e yükselmiş, 2021'de pandemi kaynaklı azalmaya bağlı olarak 310 µg/m³'e gerilemiş; ancak 2022–2023'te yeniden artış trendine girerek 375 µg/m³'e ulaşmıştır. NO_x ($NO+NO_2$) bileşeni de 2019'daki 150 µg/m³'den 2020'de 80 µg/m³'e ani düşüş sergileyip sonraki yıllarda kademeli toparlanarak 2023'te 95µg/m³ düzeyine yükselmiştir. PM_{10} ve $PM_{2.5}$ partiküllerinde; pandemi döneminde gözlenen keskin düşüş (PM_{10}: 80→50 µg/m³, $PM_{2.5}$: 40→20 µg/m³) sonrasında toparlanma oranı sınırlı kalmış, 2023'te sırasıyla 65 ve 15 µg/m³ civarında istikrar bulmuştur. SO_2 konsantrasyonu ise tüm dönemde düşük ve dalgalanmasız bir seyir izleyerek 5 µg/m³ altı değerleri korumuştur. Ozon (O_3) ve NO_2 eğrileri pandemi dönemini takiben görece yalın bir toparlanma göstermekle beraber, 2023'te sırasıyla 32 µg/m³ ve 58 µg/m³ düzeyine yerleşmiş; bu durum fotokimyasal reaksiyon dengelerindeki değişimlerin yanı sıra taşıt ve endüstriyel emisyon kontrol politikalarının birleşik etkisine işaret etmektedir. Genel olarak, COVID-19 kısıtlamalarının neden olduğu ani emisyon azalmaları ile takiben gelen ekonomik canlanma ve düzenleyici önlemlerin heterojen etki mekanizmaları, kirleticiler arasında farklı iyileşme hızı ve ölçeklenebilirlik profilleri ortaya koymuştur.

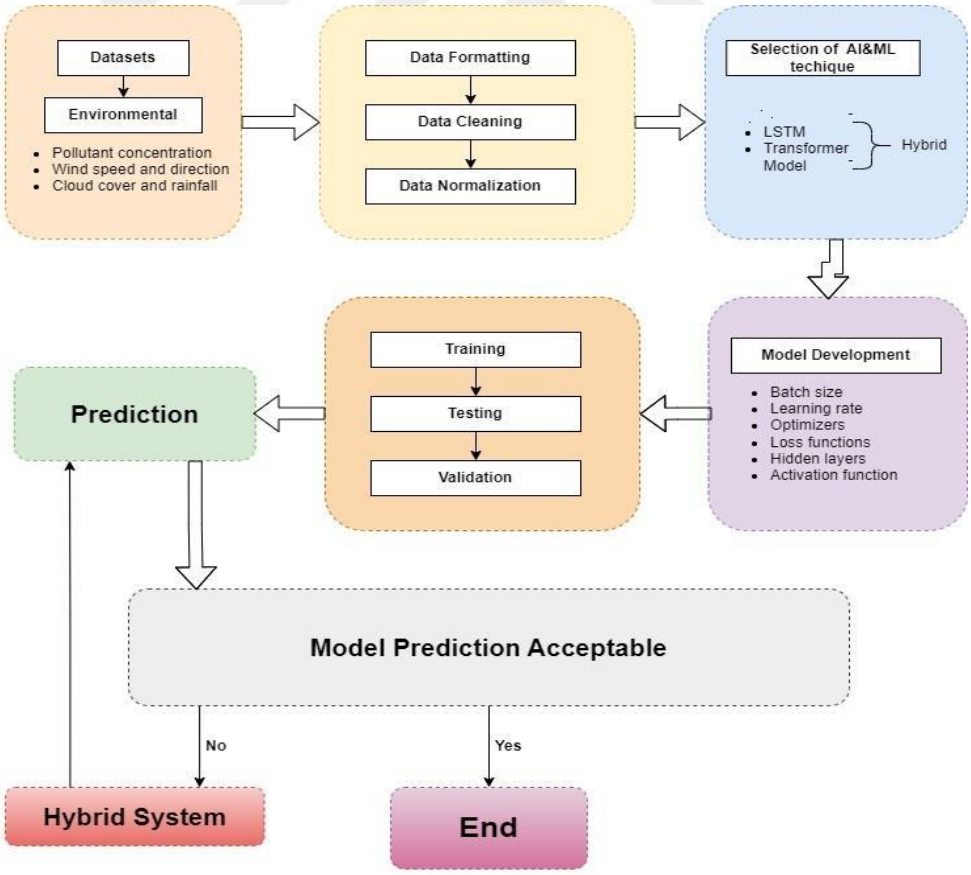


Şekil 31. Hava Kirleticilerin Yüzdelik Değişim Grafiği (Trend Analizi)

Şekil 31'de 2019–2023 dönemi kapsanarak her bir kirletici için doğrusal regresyon trend çizgileri eklenmiştir. PM_{10} trendi, genel olarak hafif pozitif eğimli (%yıllık ≈+1.5) olup, iki uç arasındaki toparlanmanın istikrarlı bir seyir izlediğini göstermektedir. $PM_{2.5}$ için trend çizgisi negatif eğimli (%yıllık ≈-2.0) kalarak ince partiküllerin uzun dönemde de azalmakta olduğuna işaret etmektedir; bu durum, gelişmiş yakıt temizleme teknolojileri ve sıkı denetim politikalarının etkisini yansıtır. SO_2'nin 2019–2023 trendi ise düzleştirilmiş (%yaklaşık +0.5) bir artış eğilimi sunmakla birlikte 2022'deki ani düşüş regression hattını belirgin biçimde etkilemiştir. Bu kapsamlı analiz, kirletici türleri arasındaki uzun vadeli eğilim farklılıklarını ortaya koyarak, hem halk sağlığı açısından öncelikli parametrelerin belirlenmesi hem de geleceğe dönük politika tasarımında hangi kirleticilere ağırlık verilmesi gerektiği konusunda rehberlik sağlamaktadır.

5. MODEL GELİŞTİRME

Bu bölümde, hava kalitesi tahmini amacıyla geliştirilen hibrit derin öğrenme modelinin mimarisi ayrıntılı olarak sunulmuştur. Model, zaman serisi tahminine özgü olarak LSTM (Long Short-Term Memory) ve Transformer mimari yapılarının güçlü yönlerini bir araya getirerek hem kısa vadeli dalgalanmaları hem de uzun vadeli yapısal desenleri öğrenebilen bir sistem olarak tasarlanmıştır. Bu mimari yapı, hava kirleticilerinin (PM_{10}, $PM_{2.5}$ vb.) zamansal örüntülerini yüksek doğrulukla tahmin etmeyi amaçlamaktadır. Hibrit hava kirliliği tahmin modeli geliştirme sürecini, veri işleme aşamasından model eğitimi ve değerlendirmesine kadar adım adım bütün süreci kapsamlı ve sistematik bir biçimde gösteren bir akış diyagramı Şekil 32'de yer almaktadır.



Şekil 32. Hibrit Hava Kirliliği Tahmin Modeli Mimarisi Akış Şeması

Bu çalışmanın akış şeması, veri mühendisliği ve yapay zeka (YZ) modelleme süreçlerini bir araya getiriyor. Bu birleşim, çalışma açısından oldukça değerlidir ve önemli iç görüler sunar.

Bu süreç, veri toplama, temizleme, normalizasyon ve model seçimi gibi veri madenciliğinin temel aşamalarını kapsamaktadır. Ayrıca hibrit yapay zeka modelleri ile doğruluk merkezli bir tahmin sistemine de dikkat çekmektedir.

İlk aşamada, veri toplama süreci çevresel göstergelere dayanmaktadır. Bu aşamada, hava kirleticiler ($PM_{2.5}$, PM_{10} gibi), rüzgar hızı ve yönü, nem ve sıcaklık gibi meteorolojik veriler toplanır. Bu veriler, modelin öngörü kapasitesini artırmak için çoklu çevresel koşulları yansıtacak şekilde entegre edilmiştir. Sonrasında veri biçimlendirme, temizleme ve normalleştirme işlemleri gerçekleştirilir. Bu aşama, zaman serisi analizi için uygun veri bütünlüğünü sağlamakla kalmayıp, aynı zamanda modelin eğitilebilirliğini de artırmayı amaçlamaktadır. Eksik verileri işlemek, aykırı değerleri ayıklamak ve tüm değişkenleri ölçeklendirmek, bu sürecin temel parçalarıdır.

Zaman serisi verileri, özellikle PM_{10} ve $PM_{2.5}$ gibi hava kirletici maddelerin saatlik ölçümlerini içerecek şekilde alınmakta ve STL (Seasonal-Trend decomposition using Loess) yöntemi ile trend, mevsimsellik ve rastgelelik olarak üç bileşene ayrılmıştır. Bu ayrıştırma, modelin sadece verinin genel eğilimini değil, mevsimsel döngüleri ve düzensiz dalgalanmaları da öğrenebilmesini sağlar. Bu bileşenler, veri temsili açısından modelin öğrenme kapasitesini büyük ölçüde artırabilir.

Veri, özel olarak tasarlanmış bir HybridDataset sınıfı üzerinden hazırlanır. Bu sınıfın girdileri trend, mevsimsellik ve rastgelelik bileşenleri olup, hedef değer olarak PM konsantrasyonları kullanılır. Mini-batch yapısında çalışan ve shuffle=True olarak tanımlanan DataLoader nesnesi, verinin eğitim sırasında rastgele dağılımını sağlayarak modelin genelleme yeteneğini güçlendirir.

Sonrasında, modelleme sürecinde kullanılacak yapay zeka ve makine öğrenimi tekniklerinin seçimi yapılır. Burada dikkat çekici olan, yalnızca tekil modellerin (LSTM, Transformer) değil, bunların hibrit yapı halinde entegrasyonunun öngörülmesidir. Bu seçim, hava kirliliği gibi çok değişkenli ve karmaşık sistemleri modellemek için oldukça

uygundur. Devamında gelen model geliştirme aşamasında ise Optuna ile gerçekleştirilen hiperparametre optimizasyonu (batch size, öğrenme oranı, aktivasyon fonksiyonları, katman sayısı vb.) devreye girerek modelin genel başarısını artıracak şekilde yapılandırma yapılmıştır.

Eğitim sürecinde, ModelCheckpoint ve EarlyStopping stratejileri entegre edilmiştir. Bu stratejiler, overfitting'i önlemek ve en iyi model parametrelerini otomatik olarak kaydetmek adına kritik rol oynar. Eğitim sonunda, modelin başarımı hem validation loss değerleriyle hem de tahmin-gözlem karşılaştırmaları ile görselleştirilerek değerlendirilir. Ayrıca, modelin hata performansı Karekök Ortalama Kare Hatası (Root Mean Squared Error-RMSE) ve Ortalama mutlak hata (Mean Absolute Error - MAE) gibi nicel metriklerle analiz edilir.

Model eğitimi, test edilmesi ve doğrulanması süreci tamamlandıktan sonra, elde edilen tahminlerin kabul edilebilirliği değerlendirilir. Eğer modelin tahminleri belirli bir doğruluk eşiğinin altında kalıyorsa, hibrit sisteme geri dönülerek yeni bir modelleme stratejisi veya parametre ayarlamaları yapılır. Bu süreç, modelin sürekli olarak geliştirilmesine olanak tanır. Tahminler yeterli doğrulukta olduğunda ise süreç başarıyla tamamlanır. Bu yapı, yalnızca teknik doğruluğun yanı sıra, aynı zamanda metodolojik bütünlüğü de ön plana çıkarmaktadır. Sonuç olarak, Şekil 32, hem araştırma hem de uygulama açısından hibrit model tabanlı hava kirliliği tahmini için kapsamlı ve gelişmiş bir yol haritası sunmaktadır.

5.1 Kullanılan Algoritmaların Teorik Açıklaması

Zaman serisi tahmini literatüründe sıklıkla karşılaşılan iki temel derin öğrenme mimarisi olan LSTM ve Transformer yapılarına ilişkin teorik bilgiler bu başlık altında açıklanmıştır.

5.1.1. Uzun kısa süreli bellek mimarisi (LSTM)

LSTM, klasik tekrarlayan sinir ağlarının (RNN) uzun dönemli bağımlılıkları öğrenmekte yaşadığı zorlukları (özellikle vanishing gradient problemi) aşmak amacıyla geliştirilmiş bir mimaridir. LSTM, zaman serisi içerisindeki geçici dalgalanmaların ve kısa vadeli örüntülerin başarılı biçimde modellenmesini sağlar.

LSTM'nin temel birimi, hücre durumu (cell state) ve çeşitli kapılardan (gates) oluşmaktadır.

Hücre Durumu (Cell State, C_t): Uzun vadeli bilgiyi saklar. Ct , bir önceki durum $Ct-1$ ile kapılar aracılığıyla güncellenir.

Giriş Kapısı (Input Gate, $İ_t$): Giriş bilgisinin ne kadarının hücre durumuna ekleneceğini kontrol etmektedir.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{5.1a}$$

Unutma Kapısı (Forget Gate, f_t): Önceki hücre durumundan hangi bilginin silineceğini denetlemektedir.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{5.1b}$$

Çıkış Kapısı (o_t, Output Gate): Hücre durumunun hangi kısmının çıktı olarak kullanılacağını belirler.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{5.1c}$$

Aday Hücre Durumu:

$$\tilde{C}t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{5.1d}$$

Hücre Durumunun Güncellenmesi:

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \tag{5.1e}$$

Çıktı

$$h_t = o_t \odot \tanh(C_t) \tag{5.1f}$$

LSTM katmanı uzun vadeli bağımlılıkları etkili şekilde modelleyebilir. Gürültülü verilerde önemli özellikleri ayıklama becerisi yüksektir.

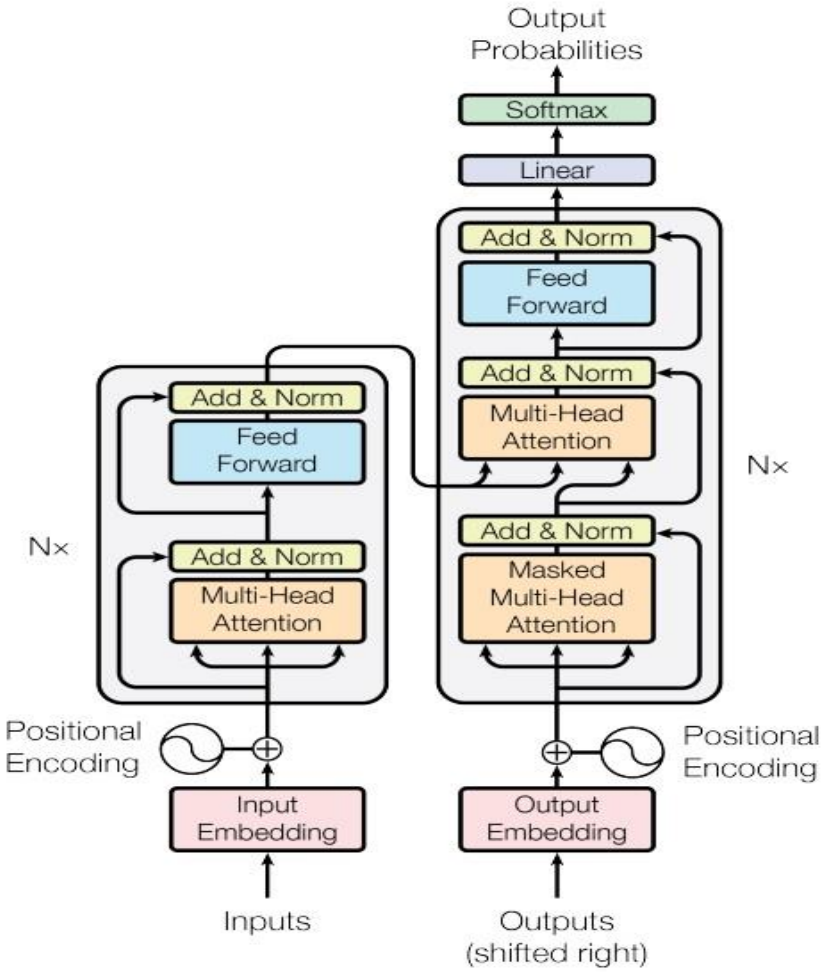
5.1.2. Transformer model

Transformer modeli, sıralı verilerde paralel hesaplama yapabilen, öz-dikkat (self-attention) mekanizması üzerine kurulu bir modeldir. İlk olarak doğal dil işleme (Natural Language Processing) alanında yaygınlaşmakla birlikte zaman serisi analizlerinde de başarıyla kullanılmaya başlanmıştır.

Uzun dönemli bağımlılıkları LSTM gibi sıralı işleme gerekmeden başarılı bir şekilde öğrenebilir. Zaman bağımlılığı içeren veri setlerinde performans ölçüm değerlerini arttırmada başarılı olmuştur. Paralel işlem yapabilme yeteneği sayesinde büyük veri setlerinde hızlı eğitim imkanı sunması büyük bir avantajıdır.

Transformer modeli, özellikle dikkat mekanizmalarına (attention mechanisms) dayanır ve sıralı veriler üzerinde paralel hesaplama yapabilmesi ile öne çıkmaktadır. Dikkat mekanizması sayesinde, farklı zaman adımları arasındaki uzun vadeli ilişkiler etkili bir şekilde öğrenebilmektedir. Model, sabit bir uzunluktaki girişlere ihtiyaç duymadan esnek bir şekilde çalışabilmektedir. Bu model, öz dikkat (self-attention) mekanizması

kullanarak tüm zaman adımları arasında doğrudan ilişki kurabilir (Vaswani vd., 2017). Model genellikle bir kodlayıcı ve kod çözücü (encoder- decoder) bileşenden oluşur (Şekil 33). Çoğunlukla, zaman serisi tahmini çalışmalarında yalnızca kodlayıcı bileşeni bağımsız olarak kullanılır.



Şekil 33. Transformer Model Mimarisi (Vaswani vd., 2017)

Transformer model mimarisinde, çok başlı bir öz-dikkat mekanizması ve mekansal olarak bağımsız (Positional Encoding), tam bağlı bir ileri besleme ağı yer alır.

A. Öz-Dikkat (Self-Attention) Mekanizması

Transformer mimarisinin en yeni ve orjinal bileşeni olan öz-dikkat mekanizması ile modelin, giriş dizisinin tüm öğeleri arasındaki bağlamsal ilişkileri aynı anda ve konumdan bağımsız olarak değerlendirmesini sağlar. RNN ve LSTM mimarilerindeki sıralı modellerin aksine, öz-dikkat mekanizması bileşeni, uzun menzilli bağımlılıkları etkin bir şekilde yakalayabilmek için tüm diziyi paralel olarak işlemektedir. Dizideki iki nesne arasındaki ilişkileri konumlarından bağımsız olarak işaretleyebilmektedir. Böylelikle, her öğenin diğer öğelere olan ilişkisini dinamik olarak değerlendirilmiş olur. Sorgu vektörü, dikkati hesaplanan öğeyi temsil eder; anahtar vektörü, dikkatin yönlendirildiği öğeyi temsil eder; değer vektörü ise aktarılacak bilgiyi taşır [Luong, Pham, Mnnning, 2015;Vaswani vd., 2017]. Bu mekanizma üç temel vektör ile temsil edilmektedir.

Sorgu (Query, Q), Anahtar (Key, K), Değer (Value, V)

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V \tag{5.2}$$

Her bir öğe çifti için dikkat ağırlık puanlarını, sorgu ve anahtar vektörlerinin nokta çarpımı (dot-product) ile hesaplanır. Ardından sayısal kararlılığı arttırmak için, bu ağırlık puanları ile anahtar vektörlerinin boyutunun karekköüne bölünerek ölçeklendirilmiş dikkat (scaled attention) uygulanır. Son olarak, softmax fonksiyonu ile normalleştirilir ve dikkat ağırlıkları belirlenmiş olur. Böylece her öğenin ne kadar önemli olduğunu belirlenmiş olur ve modelde hangi öğelere daha fazla odaklanması gerektiğini öğrenir.

Dikkat ağırlıkları aşağıdaki formül ile tanımlanır:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5.3}$$

Burada X, giriş dizisini (örneğin zaman serisi verisini), d_k ise anahtar vektörlerin boyutunu ifade eder. Bu mekanizma sayesinde, modelin giriş dizisinin farklı kısımlarına aynı anda odaklanmasını sağlar. Zaman serisi tahmininde uzun menzilli bağımlılıkları yakalama yeteneği ile farklı zaman dilimlerindeki olayların birbirini etkileyebileceği durumları ilişkilendirip tespit etmesi ile etkili bir özellik olarak değerlendirilmektedir. Örneğin, hava kirliliği seviyeleri üzerinde etkili olan geçmiş günlerdeki hava durumu koşulları veya trafik yoğunluğu verileri gibi değişkenlerin, mevcut hava kirliliği seviyeleri üzerindeki etkilerini doğru bir şekilde analiz edebilir.

B. Konumsal Kodlama (Positional Encoding)

Veri girişlerini sıralı olarak işlemediğinden, veri setinin sırasını belirlemek için konumsal kodlama mekanizmasını uygular; dizideki konumunu kodlayan her giriş öğesine benzersiz bir vektör ekleyerek sıralar. Bu yöntem ile, modelin dikkat puanlarını hesaplarken zamansal sırasını dikkate almasını sağlar. Böylelikle zaman serisi verilerindeki sıranın korunması amacıyla giriş verilerine pozisyon bilgisi şu şekilde eklenir:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d}}\right), \quad PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d}}\right) \tag{5.4}$$

C. Çok Başlı Dikkat Mekanizması (Multi-Head Attention)

Çok başlı dikkat mekanizmasını, veri setlerindeki ilişkilerin farklı yönlerini yakalamak için kullanır. Bu mekanizma sayesinde, tek bir dikkat fonksiyonuna bağlı kalmadan modeldeki veri setinin farklı bölümlerinin önemini değerlendirerek birden fazla bakış açısını dikkate almış olur. Farklı alt uzaylarda öğrenilen dikkat çıktılarının paralel olarak elde edilip birleştirilmesi şu şekilde tanımlanmıştır:

$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h)\, W^O \tag{5.5}$$

5.2. Önerilen Tahmin Model Mimarisi

LSTM ve Transformer katmanları birleştirilirken paralel yerine sıralı füzyon stratejisi tercih edilmiştir. Böylece hesaplama maliyeti dengelenir ve model hem kısa vadeli tepkileri hem de mevsimsel/trend bileşenlerini etkili şekilde öğrenir.

Paralel mimarilerde LSTM ve Transformer katmanları eş zamanlı işlenirken, bu durum hesaplama maliyetlerinin artmasına ve model karmaşıklığının yükselmesine sebep olmaktadır. Bu çalışmada önerilen sıralı füzyon stratejisi, önce LSTM katmanının kısa vadeli zaman serisi dinamiklerini kodlamasını, ardından Transformer Encoder ile bu temsilin uzun vadeli ve bağlamsal özelliklerinin öğrenilmesini sağlar. Bu sayede, hem hesaplama yükü optimize edilir hem de modelin kısa ve uzun vadeli bağımlılıkları dengeli şekilde öğrenmesi mümkün olur.

$$\begin{aligned} h_t^{\mathrm{LSTM}} &= \mathrm{LSTM}(x_t, h_{t-1}, C_{t-1}) \\ Z &= \mathrm{TransformerEncoder}(h_t^{\mathrm{LSTM}}) \end{aligned} \tag{5.6}$$

5.2.1. Önerilen modelin mimari yapısı

a. LSTM Encoder Katmanı

Zaman serisi girişleri öncelikle LSTM katmanına yönlendirilir. LSTM, kısa vadeli bağımlılıkları analiz eder.

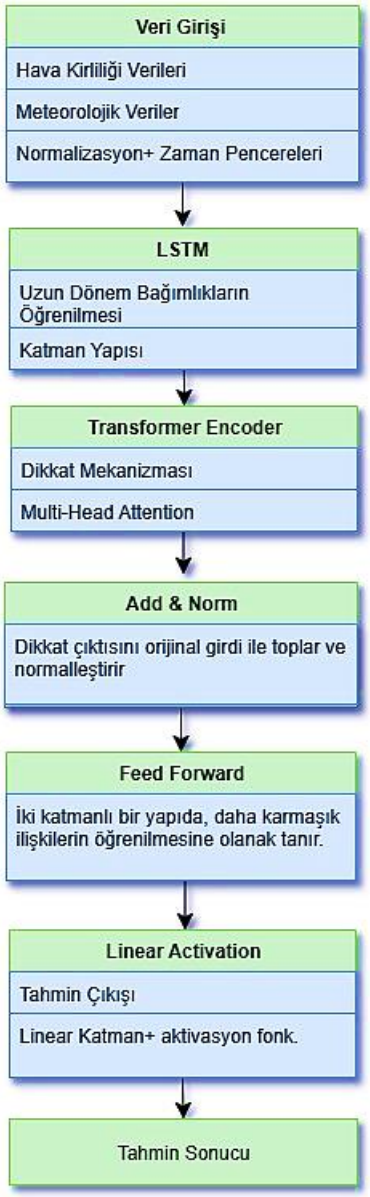
b. Transformer Encoder (Multi-head Attention)

LSTM'den elde edilen temsil vektörleri, Transformer'a aktarılır. LSTM çıktıları, ara özellikler (hidden state, ht) olarak Transformer encoder'a iletilir. Bu yapı, kısa vadeli bağımlılıkların LSTM ile, uzun vadeli bağlamların ise Transformer ile öğrenilmesine olanak tanır.

c. Dense Çıkış Katmanı

Transformer tarafından işlenen özellikler, tam bağlantılı (fully connected) katman üzerinden hedef değişkenlere iletilir. PM_{10} ve $PM_{2.5}$ gibi hedef değişkenlerin tahmini bu katmanda gerçekleştirilir.

Önerilen modelin mimari yapısı, Şekil 34'de detaylandırılmış bir şekilde sunulmuştur.



Şekil 34. Önerien Modelin Mimari Akış Şeması

5.2.2. Model mimarisi ve matematiksel gösterim

Zaman serisi tahmininde, verinin kısa vadeli dinamikleri ile uzun vadeli mevsimsel ve trend bileşenlerini etkili bir şekilde yakalamak, model performansını belirleyen en önemli faktördür. Bu bağlamda, klasik tekrarlayan sinir ağları (RNN) ve gelişmiş versiyonları olan Uzun Kısa Süreli Bellek (Long Short-Term Memory, LSTM) ağları, özellikle kısa ve orta vadeli bağımlılıkların tespitinde oldukça etkilidir. Diğer yandan, Transformer tabanlı modeller, paralel işleme ve çoklu dikkat mekanizmaları sayesinde uzun vadeli ilişkileri öğrenme yeteneğine sahiptir.

Transformer tabanlı modeller, paralel işlem yetenekleri ve uzun menzilli bağımlılıkları modelleme becerileri sayesinde zaman serisi tahminlerinde etkileyeci sonuçlar sunmaktadır. Ancak, bu modeller, girdi verilerin zamansal sırasının (saatlik izleme gibi) önemli olduğu durumlarda sınırlı başarı gösterebilir (Vaswani, 2017). Bu problemi çözmek için, LSTM katmanlarının katılması ile verinin zamansal yapısının öğrenilip Transformer encoder katmanına aktarılması, zaman serisindeki ardışık bağımlılıkları etkin bir şekilde yakalaması sağlanarak, konumsal kodlamadaki sınırlı zaman bilinci problemi çözülmüştür [Bahdanau vd., 2015; Hochreiter, 1997; Qin vd., 2017].

A. Girdi Temsili

Önerilen modele, zaman serisi verisi girişi aşağıdaki gibi gösterilmektedir.

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T], \quad \mathbf{x}_t \in \mathbb{R}^d \tag{5.7}$$

burada xt t-zaman noktasındaki d-boyutlu gözlemi temsil eder. T, gözlem sayısını ifade eder.

B. LSTM Encoder Katmanı

LSTM katmanı, zaman serisindeki kısa ve orta vadeli bağımlılıkları yakalamak için kullanılır. LSTM hücreleri, geçmiş zaman noktalarındaki bilgiyi unutma, hatırlama ve güncelleme mekanizmaları ile kodlayarak zaman içinde taşıma yeteneğine sahiptir.

LSTM encoder çıkışı, her zaman adımı için gizli durum dizisi:

$$\mathbf{H}^{\text{LSTM}} = \text{LSTM}(\mathbf{X}) = [\mathbf{h}_1^{\text{LSTM}}, \mathbf{h}_2^{\text{LSTM}}, \ldots, \mathbf{h}_T^{\text{LSTM}}], \quad \mathbf{h}_t^{\text{LSTM}} \in \mathbb{R}^{d_h} \tag{5.8}$$

C. Transformer Encoder Katmanı

LSTM tarafından oluşturulan ara temsiller, uzun vadeli bağlam ve mevsimsel trend gibi karmaşık zaman bağımlılıklarını öğrenmek üzere Transformer Encoder'a iletilir. Transformer, çok başlı (multi-head) dikkat mekanizması ile hem uzamsal hem de zamansal bağlamı geniş bir perspektiften kavrayabilir.

Transformer Encoder çıkışı:

$$\mathbf{Z} = \text{TransformerEncoder}(\mathbf{H}^{\text{LSTM}}) = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_T], \quad \mathbf{z}_t \in \mathbb{R}^{d_z} \tag{5.9}$$

D. Dikkat Ağırlıklı Birleştirme (Attention Fusion)

Modelin tahmin gücünü artırmak için LSTM'nin kısa vadeli temsilleri ile Transformer'ın uzun vadeli bağlam çıkarımları dikkat ağırlıklı olarak birleştirilir. Çok başlı dikkat (Multi-Head Attention) mekanizması, LSTM çıktılarını self-attention yoluyla zenginleştirir.

$$\mathbf{A} = \text{MultiHead}(\mathbf{H}^{\text{LSTM}}, \mathbf{H}^{\text{LSTM}}, \mathbf{H}^{\text{LSTM}}) \in \mathbb{R}^{T \times d_a} \tag{5.10a}$$

Burada **A**, LSTM temsillerinin farklı alt uzaylarda dikkatle yeniden şekillendirilmiş versiyonudur.

Birleştirilmiş nihai temsili:

$$\mathbf{h}^{\text{final}} = \text{concat}(\mathbf{H}^{\text{LSTM}}, \mathbf{A}) \in \mathbb{R}^{T \times (d_h + d_a)}$$

(5.10b)

dikkat mekanizması, tüm giriş değerlerinin ağırlıklı bir toplamını hesaplar, burada ağırlıklar her giriş değerinin mevcut çıktı için ne kadar ilgili olduğuna göre belirlenir. Bu yaklaşım, Transformer'ın uzak veri noktaları arasındaki ilişkileri etkili bir şekilde modellemesine olanak tanır. Bu, gelecekteki hava kirliliği koşullarını etkileyebilecek çeşitli meteorolojik faktörleri içeren hava kirliliği tahmininde özellikle faydalıdır.

E. Tahmin Katmanı

Son aşamada, birden çok tam bağlantılı katmandan (MLP - Multi-Layer Perceptron) oluşan çıkış katmanı, birleşik temsili kullanarak her bir zaman adımı için tahmin gerçekleştirir:

$$\hat{\mathbf{y}}_t = \text{MLP}(\mathbf{h}_t^{\text{final}}), \quad \hat{\mathbf{y}}_t \in \mathbb{R}^{d_y}$$

(5.10c)

LSTM, sensör bazlı dalgalanmalar gibi kısa süreli değişimleri modellerken, Transformer katmanı mevsimsel döngüler ve meteorolojik koşullar gibi uzun dönemli yapıları ele alır. Bu nedenle, seri ve çok katmanlı hibrit mimari, hava kalitesi gibi karmaşık sistemlerin tahmininde ileri teknik performans sağlayabilir.

Önerilen modelde, zaman serisi verisi önce LSTM tarafından kısa vadeli özellikler açısından kodlanmakta, ardından Transformer Encoder aracılığıyla uzun vadeli ve bağlamsal ilişkiler işlenmektedir. LSTM çıkışları, aynı zamanda çok başlı dikkat mekanizması ile yeniden değerlendirilip orijinal temsillerle birleştiriliyor ve tahmin

katmanı ile nihai çıkış elde ediliyor. Bu sıralı füzyon mimarisi, hesaplama verimliliği ve model başarımı açısından önemli avantajlar sunmaktadır.

a) Tartışma ve Avantajları:

Sıralı füzyon stratejisi, LSTM ile kısa vadeli yapıları ilk aşamada yakalayarak Transformer'ın giriş boyutunu ve karmaşıklığını düşürür. Böylece, paralel işleme göre daha düşük hesaplama maliyeti sağlanır.

LSTM'nin dinamik hafızası ve Transformer'ın global dikkat mekanizması, birlikte kısa ve uzun vadeli özelliklerin dengeli öğrenilmesine olanak tanır. Bu, özellikle mevsimsel, trend ve ani dalgalanma içeren karmaşık zaman serilerinde başarılı sonuçlar verir.

Çok başlı dikkat mekanizması, farklı bağımlılıkların farklı alt uzaylarda modellenmesini mümkün kılarak modelin genelleme kapasitesini artırır.

b) Genel Uygulama Alanları

Meteoroloji, hava kalitesi tahmini, enerji tüketimi, finansal zaman serileri gibi alanlarda uygulanabilir.

5.3. Özyinelemeli ve Çok Adımlı Tahmin Yaklaşımları

Zaman serisi modelleme aşamasında kullanılan tahmin statejileri, modelin genel performansını etkilemektedir. Hava kirliliği ve meteoroloji gibi çevresel sistemlerde, zamana dayalı değişkenliğin yüksek olması ve tahmin ufkunun genişletme çalışmaları için, tahminleme stratejilerinin seçimi dikkatli olunması gereken kritik bir aşamadır. Bu bağlamda, çalışmada zaman serisi tahmin stratejisi olan özyinelemeli (recursive) ve çok adımlı (multi-step) tahminleme yöntemleri seçilerek uygulanmıştır. Her iki yöntemin avantajları ve sınırlılıkları literatürde kapsamlı olarak incelenmiş ve uygulamalarda yaygın şekilde kullanılmıştır.

Bu kapsamda, bu çalışmada iki temel zaman serisi tahmin yöntemi olan özyinelemeli çok adımlı tahmin yaklaşımı uygulanmıştır. Her iki yaklaşımın da avantajları ve sınırlılıkları literatürde detaylı şekilde incelenmiştir (Ben Taieb, Bontempi, Atiya ve Sorjamaa, 2012).

A. Özyinelemeli Tahmin Yaklaşımı

Özyinelemeli tahmin yaklaşımı, bir adımlık (one-step-ahead) tahmin yapan bir modelin tekrarlı/ardışık biçimde kullanılması esasına dayanır. İlk adımda model, zaman serisinin t+1 zaman noktasını tahmin eder. Daha sonra bu tahmin değeri, bir sonraki tahmin için giriş (input) olarak kullanılır ve bu süreç her yeni zaman adımı için tekrarlanır (Ben Taieb vd., 2012). Bu yaklaşımda tahmin süreci adım adım/ zincirleme ilerleyen bir yapıdadır:

$$\hat{y}_{t+1} = f(x_t), \quad \hat{y}_{t+2} = f(x_{t+1} = \hat{y}_{t+1}), \quad \dots \tag{5.11}$$

a) Avantajları

Mevcut tek-adımlı modellerin yeniden kullanılmasına olanak tanır. Düşük parametre maliyeti ile daha az karmaşıklığa sahiptir. Eğitim süreci basittir, çünkü sadece tek adım tahmini öğretilebilir.

b) Dezavantajları

Hatalar birikerek sonraki adımlarda tahmin doğruluğunu düşürebilir. Model, çok adımlı bağımlılıkları doğrudan öğrenemez.

Özyinelemeli tahmin stratejisi, özellikle LSTM mimarisi ile kullanıldığında geçmiş zaman adımlarındaki bilgileri bellekte tutarak göreli doğruluk sağlar, ancak tahminlenen zaman noktası uzadıkça birikimli hata kaçınılmaz hale gelebilir.

A. Çok Adımlı Tahmin Yaklaşımı (Multi-Step Prediction)

Çok adımlı tahmin yaklaşımı, modelin aynı anda birden fazla gelecek zaman adımını tahmin etmesini amaçlayan bir yaklaşımdır (Ben Taieb ve J.Hyndman (2014). Bu stratejide, modelin çıktısı bir vektör şeklinde olup, her bir eleman sırasıyla gelecekteki bir zaman adımına karşılık gelir.

$$\hat{Y}_{t+1:t+H} = [\hat{y}_{t+1}, \hat{y}_{t+2}, \ldots, \hat{y}_{t+H}] \tag{5.12}$$

Burada H tahmin ufkunu belirtir (örneğin, 6 saatlik tahminlerde H=6). Bu yöntem ile modelin gelecekteki birkaç zaman dilimini (örneğin, 1 saat, 2 saat, ... n saat sonrası) aynı anda veya sırayla tahmin etmesidir.

a) Avantajları

Hata birikimini azaltır. Uzun vadeli bağımlılıkları daha iyi modelleyebilir ve Transfer öğrenme mimarileri ile doğrudan uyumludur.

b) Dezavantajları

Model karmaşıklığı artar. Eğitim süreci daha zordur ve daha fazla veri gerektirir.

5.4. Optimizasyon Yöntemleri

Derin öğrenme modellerinin etkin biçimde eğitilebilmesi, yalnızca doğru mimari yapıların belirlenmesine değil; aynı zamanda parametrik düzeyde optimize edilmiş bir eğitim stratejisinin uygulanmasına da doğrudan bağlıdır. Bu bağlamda, çalışmada geliştirilen hibrit model (LSTM + Transformer) için genel performansı artırmak, aşırı öğrenmeyi (overfitting) önlemek ve eğitim sürecinin kararlılığını sağlamak amacıyla çok yönlü optimizasyon yöntemleri entegre edilmiştir.

A. Kayıp Fonksiyonu: Mean Squared Error (MSE)

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{5.13}$$

MSE, regresyon tabanlı zaman serisi tahminlerinde yaygın olarak kullanılan ve modelin hatalı tahminleri cezalandırmasını sağlayan bir kayıp fonksiyonudur. Bu çalışmada, hem özyinelemeli hem de çok adımlı çıktılar için ortak değerlendirme metriği olarak MSE kullanılmıştır.

B. AdamW Optimizatörü

Modelin parametrelerini güncellemek için tercih edilen AdamW optimizatörü, klasik Adam algoritmasına göre ağırlık çürümesi (weight decay) terimini daha ayrı ve düzenli biçimde ele alması sayesinde, aşırı öğrenme sorununa karşı daha dengeli çözümler sunmaktadır (Loshchilov ve Hutter, 2019). Loshchilov ve Hutter tarafından önerilen bu optimizasyon yöntemi sayesinde, özellikle yüksek parametre uzaylarında öğrenmenin dengelenmesi ve genelleme performansının iyileştirilmesi için Transformer tabanlı yapılarda genel başarıyı arttırmada etkilidir.

$$\theta_{t+1} = \theta_t - \eta(\hat{g}_t + \lambda\theta_t) \tag{5.14}$$

AdamW, Transformer mimarileriyle uyumlu çalışmakta ve ağırlık çürümesi terimi sayesinde parametrelerin gereksiz büyümesini engellemektedir. Uygulamada, ağırlık normlarını kontrol altında tutarak öğrenme sürecini dengelemekte ve parametrelerin gereksiz büyümesini engelleyerek genelleme performansını artırmaktadır.

C. Öğrenme Oranı Azaltma Stratejileri

Öğrenme oranı zamanla azalacak şekilde bir planlayıcı (learning rate scheduler) kullanmak, modelin eğitim performansını artırabilir. Eğitim sırasında ReduceLROnPlateau ve Cosine Annealing öğrenme oranı planlayıcıları (learning rate schedulers) lokal minimumlara ulaşmak için kullanılmıştır:

- ReduceLROnPlateau: Doğrulama hatası belirli bir eşik seviyede iyileşmezse öğrenme oranı azaltılır.
- Cosine Annealing: Öğrenme oranı periyodik olarak düşürülerek modelin daha geniş çözüm uzayını taraması sağlanır.

D. Dropout ve Batch Normalization

Dropout ve Batch normalization yöntemleri, modelin aşırı öğrenmesi önlemek ve eğitim sürecinin kararlılığını artırmak için kullanılmıştır.

Dropout, eğitim sırasında rastgele nöronları devre dışı bırakan etkili bir yöntemdir. Bu yöntem, modelin belirli nöronlara aşırı bağımlılılığını azaltmaya çalışan güçlü bir yöntemdir. Srivastava ve arkadaşları tarafından geliştirilen Dropout yöntemi, özellikle LSTM katmanlarının ardından uygulanır. Böylece modelin genelleme kapasitesi önemli ölçüde artar (Srivastava, Hinton, Krizhevsky, Sutskever ve Salakhutdinov, 2014).

Batch Normalization (Ioffe ve Szegedy, 2015), katman çıktılarının dağılımını normalize eder. Böylelikle, daha istikrarlı bir öğrenme süreci sağlar ve ağırlık güncellemelerini daha verimli hale getirir. Transformer bileşenlerinde, LayerNorm yapısına alternatif olarak belirli yapılarda kullanılmıştır.

E. Erken Durdurma (Early Stopping)

Derin öğrenme modelleri, özellikle büyük parametre uzaylarına sahip olduğunda eğitim verisine fazla uyum sağlayarak aşırı öğrenme (overfitting) riski taşır. Bu sorunun önüne geçmek için kullanılan en yaygın düzenleme (regularization) tekniklerinden biri olan erken durdurma (early stopping)'dır. Bu teknik, modelin doğrulama (validation) kümesi üzerindeki performansını izleyerek eğitim sürecini kontrol altında tutmaktadır.

Erken durdurma yönteminde, modelin doğrulama kaybı belirli bir eğitim turu (epoch) boyunca düzelmezse, eğitim süreci durdurulur. Bu strateji, eğitim süresini kısaltır ve modelin test verisi üzerinde daha iyi genelleme yapmasına olanak tanır (Prechelt, 1998). Ayrıca, erken durdurma hesaplama maliyetini azaltır ve eğitim kaynaklarının daha verimli kullanılmasına katkıda bulunur (Goodfellow, Bengio ve Courville, 2016).

Erken durdurma, özellikle derin sinir ağlarını optimize etmekte sıkça kullanılan teknikler arasında yer almaktadır. Bu yöntem, overfitting'i azaltmayı, erken genelleme yakalamayı ve eğitim süresini kısaltmayı hedeflediğinden tercih edilmektedir. TensorFlow, PyTorch Lightning gibi bir çok eğitim platformu ve framework bu yöntemi yerleşik olarak desteklemektedir.

5.4.1. Optuna ile hiperparametre optimizasyonu

Günümüzde derin öğrenme modellerinin başarısı, yalnızca model mimarisine değil, aynı zamanda hiperparametrelerin doğru biçimde ayarlanmasına da güçlü biçimde bağlıdır. Bu çalışmada, hiperparametre ayarlaması için açık kaynaklı ve güçlü bir optimizasyon çerçevesi olan Optuna kullanılmıştır. Optuna, Ağaç Yapılı Parzen Tahmincisi (Tree-structured Parzen Estimator ,TPE) algoritmasına dayanmakta olup hiperparametre–kayıp ilişkisini modelleyerek daha verimli bir arama süreci sunmaktadır (Akiba, Sano, Yanase, Ohta ve Koyama, 2019).

A. TPE Algoritması - Teorik Temelleri

Hiperparametre optimizasyonu, makine öğrenmi tabanlı modellerin başarımı üzerinde oldukça kritik bir aşamasıdır. TPE algoritması, geleneksel yöntemlerden klasik ızgara (grid search) ve rastgele arama (random search) yöntemlerinden farklı olarak (Bergstra ve Bengio, 2012), bayesçi optimizasyon yaklaşımına dayalı olarak, bilinçli ve adaptif bir şekilde hiperparametre aramasını gerçekleştirmeyi amaçlamaktadır. Önceki denemelerden öğrenerek arama alanını dinamik olarak günceller. Bu algoritma, hiperparametre ve kayıp ilişkisini modellemek için iki ayrı olasılık yoğunluk fonksiyonu l(x) ve g(x) tanımlar. İyi sonuçlar veren hiperparametre kombinasyonları l(x) ile, kötü sonuçlar ise g(x) ile temsil edilir. Yeni bir deneme örneği seçilirken, beklenen iyileşmeyi (Expected Improvement, EI) maksimize eden oran kullanılır (Bergstra, Bardenet, Bengio ve Kegl, 2011).

$$EI(x) = \frac{l(x)}{g(x)} \tag{5.15}$$

Bu oran, belirli bir hiperparametre kombinasyonunun iyi sonuç üretme olasılığının, kötü sonuç üretme olasılığına oranını temsil eder. TPE algoritması, bu oranı maksimize eden örnekleri seçerek, keşif (exploration) ve sömürü (exploitation) arasında dinamik bir denge sağlamış olur. Keşif aşamasında henüz tam olarak değerlendirilmemiş arama alanları araştırılırken, sömürü aşamasında daha önce başarılı olmuş hiperparametre bölgeleri yeniden test edilir. Bu sayede, daha önce denenmiş başarılı bölgelerin sömürülmesi ile yeni ve keşfedilmemiş bölgelerin araştırılması arasında etkili bir yöntem uygulanmış olur (Shahriari, Swersky, Wang, Adams ve Freitas, 2016). TPE algoritması, grid search algoritmasına kıyasla daha verimli ve daha hızlı sonuçlar üretmektedir.

5.5. Model Performans Değerlendirme Metrikleri

Modelleri değerlendirmek için kullanılan performans metrikleri aşağıdaki gibidir. y ve y^ gerçek ve tahmin edilen değerleri gösterir.

A) Ortalama Kare Hatası (Mean Squared Error-MSE)

Tahminlerin gerçek değerlere ne kadar yakın olduğunu ölçen bir ortalama hata metriğidir. MSE, tahminlerin gerçek değerlere ne kadar yakın veya uzak olduğunu ölçer ve her bir hata farkını kare alarak pozitif hale getirir. Bu nedenle, MSE değeri sıfıra yaklaştıkça, tahminler gerçek değerlere daha yakın ve daha iyi kabul edilir. Ancak, yüksek MSE değerleri, tahminlerin gerçek değerlere göre daha uzak olduğunu ve kötü tahminlerin yapıldığını gösterir.

B) Karekök Ortalama Hata (Root Mean Squared -RMSE)

Ortalama karesel hatanın karekökü (RMSE), kullanılan bir model sonucu tahmin edilen veriler ile gerçek veriler arasındaki farklılıkları değerlendirmek için kullanılır (Cihan, 2020). Hataların kareköklerinin ortalamasının karekökünü ifade eder (Willmott ve Matsuura, 2005). MAE'de olduğu gibi RMSE'nin de sıfıra yakın olması iyi bir modelin göstergesidir. Ancak RMSE'nin sınırlamaları da vardır, özellikle veri dağılımı dengesiz veya aykırı değerler içeriyorsa, bu sınırlamalar daha belirgin hale gelebilir. Bu nedenle RMSE'nin kullanımı, bağlam ve veriye bağlı olarak dikkatli bir şekilde değerlendirilmelidir.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5.16}$$

C) Ortalama mutlak hata (Mean absolute error - MAE)

Ortalama mutlak hata (MAE), tahmin edilen veri ile gerçek veri arasındaki yakınlığı ölçmede kullanılan bir parametredir. Hataların mutlak değerlerinin ortalamasını ifade eder. Ne kadar düşük değere sahip olursa model performansının da o kadar iyi olduğunu göstermektedir (Willmott ve Matsuura, 2005).

$$\mathrm{MAE} - \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{5.17}$$

D) R^2 Puan Hatası (R^2 Score Error)

R-squared (R^2) skoru, bir regresyon modelinin veriyi ne kadar iyi açıkladığını veya tahmin ettiğini ölçen bir istatistiksel metriktir. Bu metrik, tahminleme modelinin veriler üzerindeki varyansın ne kadarını açıklayabildiğini gösterir ve genellikle modelin açıklama gücünü değerlendirmek için kullanılır. R^2 skoru, 0 ile 1 arasında bir değer alır. İşte R^2 skorunun anlamı ve hesaplanma yöntemi

R^2 skoru 1'e yakınsa, bu regresyon modelinin veriyi mükemmel bir şekilde açıkladığı anlamına gelir. Yani, model veriyi %100 açıklıyor demektir. Ancak bu durum, pratikte nadiren görülür. R^2 skoru 0'a yakınsa, bu regresyon modelinin veriyi açıklamada başarısız olduğu anlamına gelir. Yani, model veriyi hiç açıklayamıyor demektir. R^2 skoru 0 ile 1 arasında bir değer alırsa, bu regresyon modelinin veriyi ne kadar iyi açıkladığını gösterir. Yüksek R^2 değerleri, modelin veriyi daha iyi açıkladığını ve tahminlerin daha doğru olduğunu gösterir. Ancak, R^2 skoru tek başına bir modelin başarısını yeterince

açıklayamayabilir. Bu nedenle, diğer hata metrikleri ve model değerlendirme teknikleri ile birlikte kullanılması önerilir.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{5.18}$$

6. HAVA KİRLİLİĞİ TAHMİN UYGULAMASI GELİŞTİRME

Bu bölümde, zaman serisi tabanlı hava kirliliği tahminlerine yönelik geliştirilen Streamlit tabanlı bir web uygulamasının yapısı, işlevselliği ve altyapısı ayrıntılı şekilde anlatılmaktadır. Uygulamanın amacı, hava kalitesi tahminlerini kullanıcı dostu bir arayüz aracılığıyla görselleştirerek, tahmin sonuçlarını anlaşılır hale getirerek yalnızca akademik araştırmalar için değil, aynı zamanda kamu kurumları, belediyeler ve çevresel izleme kuruluşları gibi uygulamalı alanlar için de kullanılabilir, anlamlı çıktılar sunabilmektir. Model çıktıları, grafiksel olarak görselleştirilmekte ve coğrafi konuma bağlı olarak interaktif haritalarla desteklenmektedir. Böylelikle, kullanıcıların çevresel karar verme süreçlerine katkı sağlanması amaçlanmıştır.

Günümüzde hava kalitesinin anlık izlenmesi ve kısa vadeli tahmini, kentsel çevresel yönetim politikalarının etkinliğini artırmakta ve halk sağlığı açısından erken uyarı mekanizmalarının geliştirilmesine katkı sunmaktadır. Bu çalışmada, gerçek zamanlı hava kalitesi tahmini amacıyla geliştirilen sistem, veri akışından tahmin sürecine kadar bütüncül bir yaklaşımla ele alınmış; anlık PM_{10} değerlerinin tahmini üzerine odaklanan çok değişkenli bir makine öğrenimi altyapısı sunulmuştur.

6.1. Uygulama Arayüzü

Streamlit, analiz sonuçlarını paylaşmak, etkileşimli deneyimler oluşturmak ve yeni makine öğrenimi modellerini kullanmak için Python tabanlı web uygulamaları oluşturmaya ve geliştirmeye yardımcı olan bir web uygulaması çerçevesidir (Streamlit Documentation, t.y.). Açık kaynaklı ve ücretsiz olmasının yanı sıra sunduğu hızlı uygulama geliştirme imkanı web uygulaması geliştirme süreçlerinde öne çıkmaktadır.

Hava kalitesi tahminleri, özellikle PM_{10} ve $PM_{2.5}$ gibi kirletici parametrelerin kısa vadeli projeksiyonları, geçmiş gözlemlerin öğrenilmesiyle yapılmakta ve zaman serisi modelleme tekniklerine dayanmaktadır. Bu bağlamda geliştirilen uygulama, son 72 saatlik veriyi girdi olarak alıp, önümüzdeki 3 günlük hava kirliliği seviyelerini hibrit derin öğrenme modeli ile öngörmektedir. Bu uygulama yalnızca bir veri görselleştirme

platformu değil, aynı zamanda arka planda çalışan makine öğrenmesi modellerinin çıktılarının işlenip kullanıcıya sunulmasını sağlayan bütünleşik bir tahmin sistemidir.

- Tahmin Sonuçlarının Görselleştirilmesi: Geliştirilen model tarafından üretilen tahminlerin zaman serisi grafiklerine, istasyon bazlı analiz için coğrafi haritalara entegre edilmesi sağlanmıştır.

- Kullanıcı Dostu Arayüz: Farklı disiplinlerden kullanıcıların (örneğin, çevre mühendisleri, sağlık uzmanları, yöneticiler) rahatlıkla erişebileceği, teknik bilgi gerektirmeyen bir kullanıcı deneyimi sağlanmıştır.

- Coğrafi Analiz: Kirletici seviyelerinin istasyon konumlarına göre harita üzerinde görsel olarak temsil edilmesi sağlanmıştır.

6.1.1. Kullanılan teknoloji altyapısı

Uygulamanın geliştirilmesinde çeşitli Python kütüphaneleri ve açık kaynak teknolojilerden faydalanılmıştır. Her bir bileşen, hem ön yüz (frontend) hem de arka yüz (backend) süreçlerin sorunsuz çalışmasını sağlamak üzere dikkatle seçilmiştir. Tablo 2'de kullanılan kütüphane ve kullanımı açıklanmıştır.

Tablo 2. Hava Kirliliği Tahmin Uygulamasında Kullanılan Teknolojiler

Teknoloji/Kütüphane	Açıklama
Streamlit	Web tabanlı kullanıcı arayüzlerinin hızlı geliştirilmesine olanak tanıyan Python framework'üdür.
Plotly	İnteraktif grafikler oluşturmak için kullanılan gelişmi görselleştirme kütüphanesidir.
Pandas	Zaman serisi verilerinin analizi ve işlenmesi için kullanılmıştır.
GeoPandas	Coğrafi veri analizleri ve harita tabanlı çıktılar üretmek amacıyla kullanılmıştır.
Folium	Leaflet.js tabanlı harita görselleştirme aracı olup, hava kalitesi istasyonlarının konumsal analizinde tercih edilmiştir.
Scikit-learn	Modelleme sürecinde ön işleme, değerlendirme ve istatistiksel analizlerde destekleyici araç olarak kullanılmıştır.
Pydeck	Harita tabanlı 3B görselleştirme bileşenlerinin oluşturulmasında tercih edilmiştir.

Ayrıca, uygulamanın modüler tasarımı sayesinde, farklı veri kümeleri ve yeni modelleme yaklaşımlarına kolayca entegre olmasına olanak sağlamaktadır.

6.1.2. Uygulama mimarisi

Uygulama mimarisi, kullanıcı ile tahmin modeli arasında güçlü bir etkileşim sağlayacak şekilde tasarlanmıştır. Bu yapı, kullanıcı arayüzü (frontend), görselleştirme ve arka plan işlemleri (backend) olarak üç ana bölümden oluşmaktadır.

A. Kullanıcı arayüzü tasarımı

Kullanıcı arayüzü, Streamlit altyapısıyla sade ve işlevsel bir tasarım ilkesiyle tasarlanmıştır. Kullanıcılar, tahmin sürecine aktif şekilde katılabilir ve tahmin sonuçlarını gerçek zamanlı olarak anında görüntüleyebilmektedir. Streamlit, kullanıcı arayüzünü

oluşturmak ve tahmin sonuçlarını etkileşimli bir şekilde sunmak için kullanılan açık kaynaklı bir web 11 uygulaması çerçevesidir (Streamlit Documentation, t.y.).

Aşağıda arayüzde bulunan temel bileşen özellikleri verilmiştir:

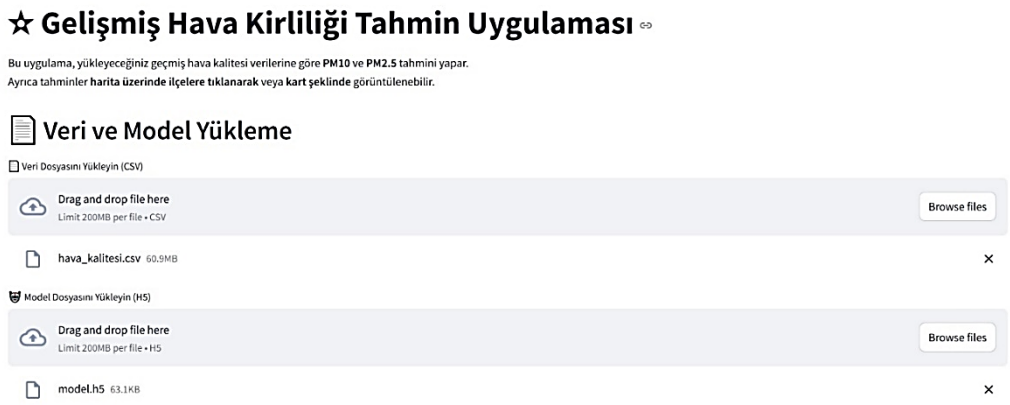
- Veri Dosyasını Yükleme: Kullanıcılar, modelin kullanımasını istedikleri hava kalitesi verilerini .csv formatında yükleyebilir (Şekil 35).

- Model Dosyasını Yükleme: Eğitilmiş derin öğrenme modelinin .h5 formatındaki dosyası sisteme yüklenmiş olur. Yüklemeler başarıyla gerçekleştiğinde tahmin sistemin çalışması sağlanır (Şekil 35).

- Tarih ve saat bilgilerinin biçimleri gözden geçirilmiş ve veriler ölçeklendirilmiştir (Şekil 36).

- İstasyon Seçimi: Kullanıcı, tahmin sonuçlarını görüntülemek istediği istasyonu seçebilir (Şekil 37). Ardından, harita üzerindeki anlık PM tahminlerine ulaşabilir.

Ayrıca, uygulamanın modüler tasarımı, farklı veri setleri ve yeni modelleme yöntemleriyle kolayca entegre olması sağlanmıştır.



Şekil 35. Hava Kirliliği Tahmin Uygulaması: Veri ve Model Yükleme Arayüzü



Şekil 36. Hava Kirliliği Tahmin Uygulaması: Eğitim Aşaması Arayüzü (2)



Şekil 37. Hava Kirliliği Tahmin Uygulaması : Harita Seçimi (3)

B. Uygulama Görselleştirme Aşaması

Tahmin ve analiz sonuçlarını etkili bir şekilde göstermek için çeşitli etkileşimli görselleştirme araçları sisteme dahil edilmiştir.

Zaman Serisi Grafikleri: Plotly kullanılarak tasarlanan bu grafikler sayesinde, belirli bir istasyon için kirletici seviyelerinin zamanla değişimi görselleştirilmektedir.

Grafik Kartları (Metric Cards): PM_{10} ve $PM_{2.5}$ gibi parametrelerin özet değerleri, renkli kartlar aracılığıyla gösterilmektedir (Şekil 38). Yüksek değerli öngörü sonuçları ise açık kırmızı olarak ayrıştırılmıştır.



Şekil 38. Hava Kirliliği Tahmin Uygulaması: Anlık Tahmin Sonucu

Bu görsel araçlar, özellikle teknik olmayan kullanıcılar için karar vermeyi kolaylaştıran sezgisel bir görünüm sunmaktadır.

C. Uygulama arka plan işlemleri

Uygulamanın arka planında (backend) çalışan yapılar, tahmin sürecinin otomatikleştirilmesini ve çıktıların analitik olarak sunulmasını sağlar.

- Veri İşleme: Kullanıcıdan alınan ham veriler ön işleme tabi tutulur, eksik veriler işlenir ve yeniden ölçeklendirme işlemleri gerçekleştirilir.

- Modelin Yürütülmesi: Yüklenen model, işlenmiş veriyi kullanarak tahmin sürecini başlatır ve n-adımlı özyinelemeli tahmin tekniğini kullanarak gelecek günler için tahminler üretir.

- Sonuçların Dönüştürülmesi: Model çıktıları Pandas veri çerçevesi içerisinde analiz edilerek, grafik ve harita bileşenlerine uygun biçimde dönüştürülür ve kullanıcıya sunulur.

6.2. Gerçek Zamanlı Hava Kalitesi Tahmin Sistemi

Bu bölümde, gerçek zamanlı hava kalitesi verilerinin işlenmesi ve tahmin edilmesine yönelik geliştirilen sistemin mimarisi, veri akışı ve uygulama süreci detaylı biçimde ele alınmaktadır. Sistem, özellikle PM_{10} parametresi üzerine odaklanarak, anlık hava kirliliği seviyelerini tahmin edebilen, otomatik veri çekimi ve model entegrasyonu sağlayan bir yapıda tasarlanmıştır. Bu sistem, hava kalitesine yönelik erken uyarı mekanizmalarının geliştirilmesi ve şehir düzeyinde çevresel karar destek altyapısının oluşturulması için önemli bir katkı sağlamaktadır.

Geliştirilen sistem, gerçek zamanlı tahmin altyapısı ile entegre edilmiş olup, bu süreç bir veri işleme pipeline'ı üzerinden yürütülmektedir. Pipeline yapısı, ham verilerin toplanması (API'den veri çekme), ön işlenmesi (normalizasyon), model giriş formatına dönüştürülmesi ve tahmin sonuçlarının eş zamanlı olarak üretilmesi adımlarını kapsamaktadır. Bu bütünleşik yapı sayesinde, sistem kesintisiz veri akışlarını işleyebilmekte ve hava kalitesindeki ani değişimlere karşı hızlı, güvenilir tahminler sunabilmektedir.

6.2.1. Gerçek zamanlı veri sağlayıcıları

Dünya genelinde güvenilir hava kalitesi verileri sağlayan ve yaygın olarak bilinen platformlar incelenerek, sistemde kullanılacak veri kaynağı belirlenmiştir.

AQICN (World Air Quality Index Project) platformu, dünya genelinde 100'den fazla ülkedeki istasyonlardan veri sağlayan açık kaynaklı bir hava kalitesi API'sidir. PM_{10}, $PM_{2.5}$, NOx, SO_2 gibi hava kirleticilerin yanı sıra meteoroloji verileri de sunmaktadır (AQICN API Documentation, t.y.).

OpenWeatherMap platformu, hava kalitesi verileri ve meteoroloji verilerini bir arada sunan kapsamlı bir platformdur (OpenWeatherMap API Documentation, t.y.).

AirVisual ise, gerçek zamanlı hava durumu ve iklim veri kaynaklarını, özellikle düşük maliyetli sensörlerden veri sağlamasıyla dikkat çeken bir platformdur. IQAir'in AirVisual API desteği sayesinde, hava kalitesi bilgileri için makine öğrenmesi destekli güvenilir bir veri kaynak sunmaktadır. Ayrıca 3 günlük tahmin verisi de sunmaktadır (AirVisual API Documentation, t.y.).

Bu çalışmada geliştirilen gerçek zamanlı uygulamada, AQICN API'si platformundan gerçek zamanlı veri akışı sağlanmış ve tahmin sistemine entegre edilmiştir. API'den alınan veriler, Python ile işlenmiş, makine öğrenmesi modellerine uygun biçimde dönüştürülerek tahmin sürecine dahil edilmiştir (AQICN API Documentation, t.y.).

6.2.2. Gerçek zamanlı uygulama mimarisi

Geliştirilen sistem, gerçek zamanlı tahmin altyapısı ile entegre edilmiştir. Bu altyapı, kesintisiz veri akışlarını işleyecek şekilde tasarlanmış olup, hava kalitesindeki ani değişimlere karşı hızlı tahminler üretebilmektedir. Karar destek sistemlerinde anlık karar alınmasını gerektiren durumlarda, bu özellik sayesinde etkin bir kullanım imkanı sağlamaktadır.

Gerçek zamanlı tahmin sistemi, veri toplama, veri ön işleme modülü, model entegrasyonu, tahminleme ve otomatik zamanlama modüllerden oluşmaktadır.

A. Veri toplama ve ön işleme modülü

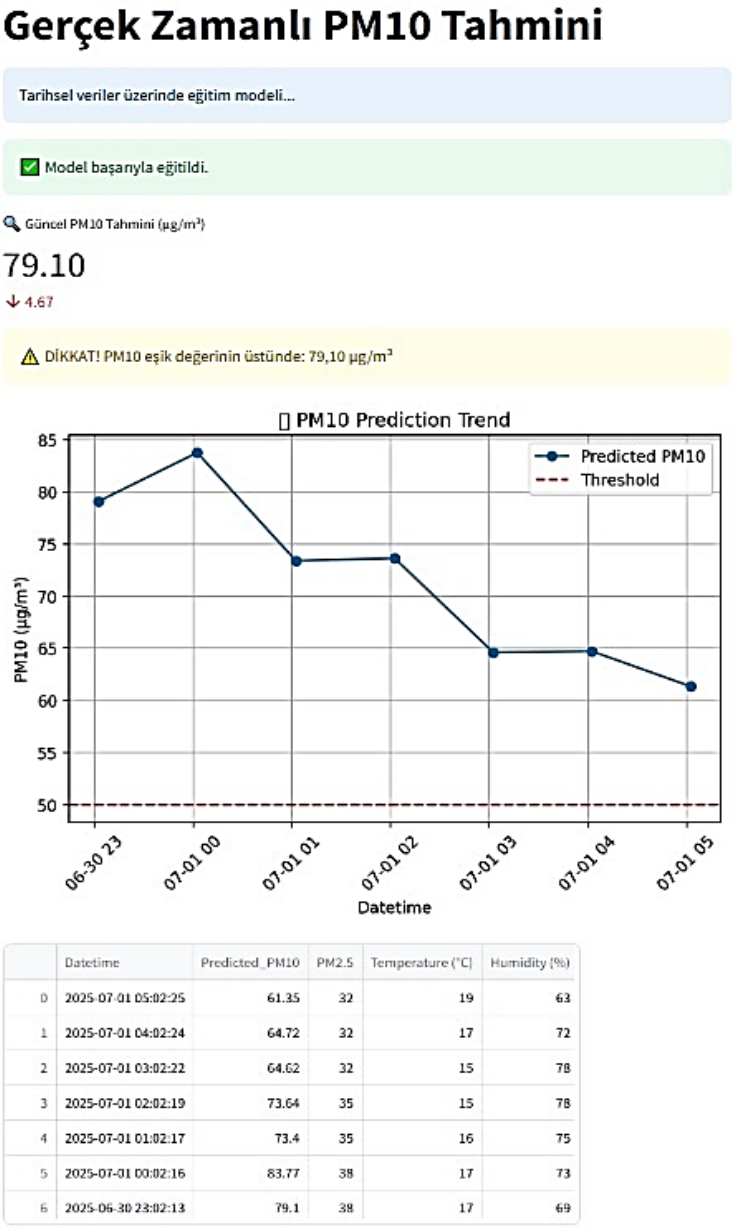
API'den veri çekme işlemi her saat başında HTTP protokolü üzerinden AQICN API'sine istek gönderip, PM_{10}, $PM_{2.5}$, NOx, NO_2, SO_2, CO, O_3 gibi parametreleri JSON formatında alınmaktadır.

Alınan JSON verisi, Python ortamında Pandas kullanılarak DataFrame'e dönüştürülmektedir. Böylelikle, modelin eğitiminde kullanılan veri seti ile eşleşmesi için uygun formata getirilmiştir.

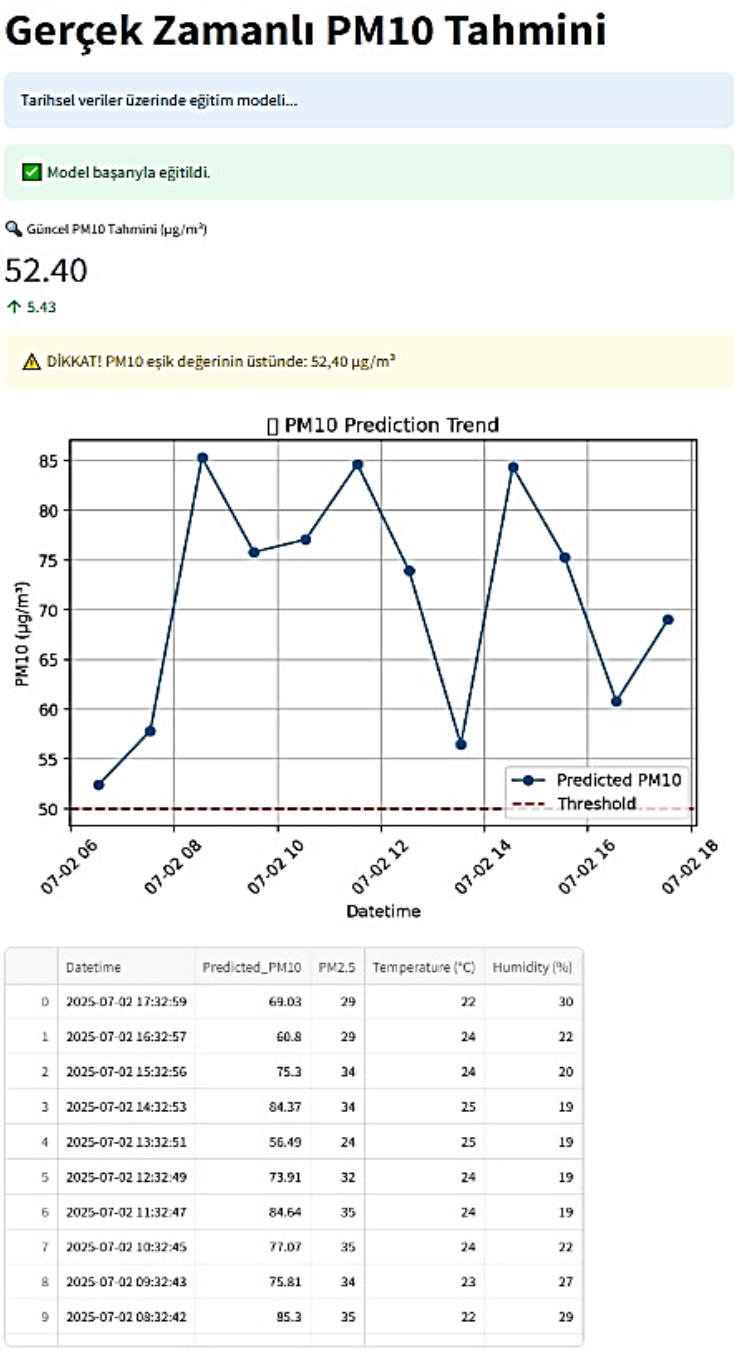
Model tahmin performans başarısını artırmak amacıyla StandardScaler yöntemi ile normalize edilmiştir. Böylece, modelin geçmiş veri dağılımıyla gerçek zamanlı veri seti değerleri uyumlu hale gelmesi sağlanmıştır.

B. Model entegrasyonu ve tahminleme modülü

Modelin tahminleme aşamasında, eğitilen hibrit modeli, ölçeklenmiş anlık veriyi giriş olarak alarak PM_{10} değerini tahmin eder (Şekil 39). Başka bir tarihte tekrar çalıştırılan uygulama ekran görüntüsü Şekil 40'da yer almaktadır.



Şekil 39. Gerçek Zamanlı PM_{10} Tahmin Uygulaması Ekran Görüntüsü



Şekil 40. Gerçek Zamanlı PM10 Tahmin Uygulaması Ekran Görüntüsü (2)

C. Zamanlama modülü

Sistem, Python ile geliştirilen zamanlayıcı (otomatikleştirilmiş zamanlama -scheduler) yapısı sayesinde saatlik periyotlarla otomatik veri çekimi ve tahmin döngüsünü gerçekleştirmektedir. time.sleep(3600) komutu ile sistem her saat başı API'ye istek

göndererek veri güncellemesi yapar. Bu süreçte, real_time_forecasting() adlı fonksiyonu, her yeni veriyi modele sunarak gerçek zamanlı tahmin çıktıları üretmektedir.

Bu yaklaşım, gerçek zamanlı tahmin sistemini otomatik bir yapıya kavuşturmakta ve sürekli güncelleme döngüsünü mümkün kılmaktadır. Sistem ayrıca tahmin sonuçlarını kullanıcıya hem konsol çıktısı hem de veri kayıt dosyaları aracılığıyla sunmaktadır. Bu mekanizma, tahmin performansının zaman içinde izlenmesine ve model güncellemelerinin planlanmasına olanak sağlamaktadır.

Geliştirilen bu sistem, gerçek zamanlı hava kalitesi izleme altyapısı sunarak şehir hava kalitesi yönetim sistemlerine entegre edilebilir bir yapıya sahiptir. Özellikle PM_{10} gibi sağlık açısından kritik öneme sahip parametrelerin erken tahmini ile birlikte, hava kirliliğine karşı önleyici önlemlerin zamanında alınmasına olanak tanımaktadır.

Ayrıca, sistemin sürekli güncellenebilir ve farklı modellerle entegre edilebilir yapısı sayesinde, değişen atmosferik koşullara hızlı adaptasyon sağlayarak tahmin doğruluğunu dinamik olarak optimize etme potansiyeli mevcuttur. Bu yapı, literatürde önerilen adaptif modelleme yaklaşımlarına uygun olarak gerçek zamanlı veriyle eğitilen ve tahmin performansı sürekli geliştirilen çevresel yapay zeka sistemlerinin iyi bir örneğini teşkil etmektedir (Chen vd., 2020).

7. ARAŞTIRMANIN BULGULARI

Bu bölümde, çalışmada önerilen LSTM + Transformer hibrit modelinin eğitim süreçleri, hiperparametre optimizasyonu, performans değerlendirmesi ve diğer yöntemlerle karşılaştırılması ayrıntılı olarak değerlendirilmiştir. Modelin başarısı; MAE, RMSE ve R^2 gibi ölçütlerle değerlendirilmiş ve literatürdeki güncel yaklaşımlarla karşılaştırmalı olarak tartışılmıştır.

7.1. Model Eğitim Süreci ve Optimizasyonu

Hava kirliliği tahmini sorununa yönelik geliştirilen hibrit modelin eğitim süreci, veri hazırlığı, model mimarisi ve hiperparametre optimizasyonu olmak üzere üç ana bölümde ele alınmıştır.

Bu tez çalışması kapsamında, geliştirilen makine öğrenmesi çalışmaları ve derin öğrenme modelleme süreçlerinde, yapılan tüm çalışmalar Google Colab Pro platformunda Python programlama dili (3.11.13 versiyonu) kullanılarak uygunlanmış ve tüm modelleme süreçleri bu platform üzerinde yürütülmüştür. Derin öğrenme modellerinin geliştirilmesi ve eğitim süreçlerinde, TensorFlow (2.18.0 versiyonu) ve PyTorch tabanlı (2.6.0+cu124 versiyonu) ileri düzey model geliştirme çerçevesi olan PyTorch Lightning çerçevesi kullanılmıştır. Model modülleri, eğitim döngüleri ve doğrulama süreçleri bu yapı üzerinden optimize edilmiştir.

Modelin eğitimi ve erken durdurulması için EarlyStopping kullanılmıştır. En iyi model ağırlıkları ModelCheckpoint ile kaydedilmiştir. Ayrıca, öğrenme oranı değişimlerini izlemek için LearningRateMonitor gibi PyTorch Lightning geri çağırım fonksiyonları etkinleştirilmiştir. Öğrenme oranının adaptif olarak güncellenmesi için ise ReduceLROnPlateau planlayıcısı tercih edilmiştir.

Zaman serisi verilerinin özellik mühendisliği ve ön işleme adımlarında STL (Seasonal-Trend Decomposition using Loess) yöntemi için Statsmodels**,** veri ölçekleme işlemlerinde RobustScaler, veri kümesi bölme süreçlerinde train_test_split ve TimeSeriesSplit gibi Scikit-learn fonksiyonları kullanılmıştır (Géron, 2019).

Optimum hiperparametrelerin belirlenmesinde Optuna kütüphanesi kullanılarak sistematik bir hiperparametre optimizasyon süreci yürütülmüştür.

Görselleştirme süreçlerinde Matplotlib, Seaborn, Plotly Express, Plotly Graph Objects ve make_subplots gibi etkili görselleştirme kütüphaneleri kullanılarak model çıktıları ve veri analizleri kapsamlı ve ayrıntılı bir şekilde sunulmuştur.

Tüm bu süreçlerde matematiksel hesaplamalar için NumPy, veri analizi için Pandas ve temel matematiksel işlemler için math kütüphanesi yardımcı araçlar olarak kullanılmıştır.

Derin öğrenme ve transformer yöntemleri için tensorflow ve keras, makine öğrenme yöntemleri için scikit-learn, istatistiksel yöntemler için statsmodels ve benzerlik hesapları için tslearn ve dtaidistance Python kütüphaneleri kullanılmıştır (SlD, t.y.).

7.1.1 Veri işleme

Modelin genelleme yeteneğini güvenilir biçimde değerlendirebilmek için veri setinin belirlenen yöntemlerle ve oranlarda eğitim, doğrulama ve test alt kümelerine ayrılması kritik öneme sahiptir. Zaman serisi problemlerinde, verilerin kronolojik bütünlüğünü korumak, geçmiş verinin geleceği "sızdırmasını" (data leakage) ve sonrasında ortaya çıkabilecek yanıltıcı sonuçları önlemek açısından gereklidir. Bu çalışmada, zaman serisi verilerinin gelecekteki bilgileri geçmişe taşımaması için zamana dayalı bölme (time-based split) kullanılmıştır.

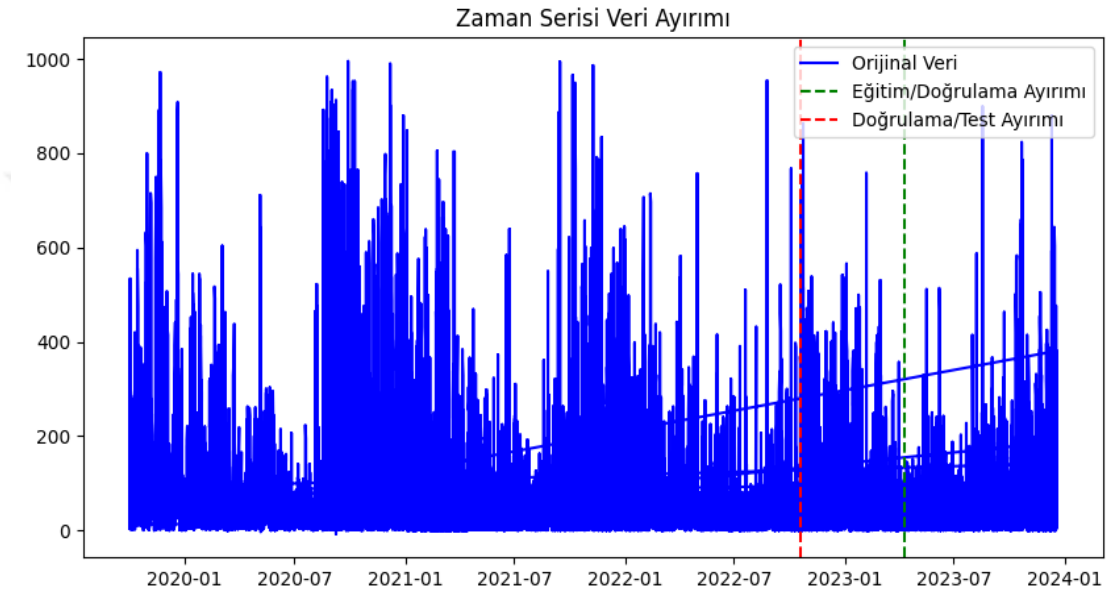
A. Kronolojik Bölme Yaklaşımı

Bu çalışmada, klasik rastgele bölme yerine zamana dayalı bölme kullanılmıştır. Veri kümesinin ilk %70'i eğitim, sonraki %15'i doğrulama ve son %15'i test seti olarak belirlenmiştir. Böylece model; sadece geçmiş verilere dayanarak öğrenmiş ve henüz karşılaşılmamış zamansal desenler üzerinde test edilmiştir.

%70 oranında ayrılan eğitim seti, model ağırlıkları güncellemeleri için kullanıldı. Bu veri seti sayesinde temel örüntülerin öğrenilmesini sağladı.

Doğrulama seti, %15 oranında kullanılmıştır. Bu %15'lik veri seti, hiperparametre ayarlaması, erken durdurma (early stopping) kararı ve öğrenme oranı planlayıcısının (scheduler) optimizasyonu için kullanılmıştır.

Test Seti %15, en son modelin bağımsız performansını ölçmek için kullanılmıştır.Bu ölçüm, eğitim veya ayar sürecinde kullanılmamış veriler üzerinde ölçülmüştür. Çalışma veri setindeki bölme işlemi görselleştirilerek, şekil 41'de sunulmuştur.



Şekil 41. Zaman Serisi Verisinde Veri Bölümleme

Veri bölme işlemi Python'da aşağıda gösterildiği şekilde gerçekleştirilmiştir:

```
# Kronolojik veri bölme işlemi

train_size = int(len(data) * 0.7)

val_size   = int(len(data) * 0.15)

train_data = data.iloc[:train_size]

val_data   = data.iloc[train_size:train_size + val_size]

test_data  = data.iloc[train_size + val_size:]
```

B. Verileri İşleme Yaklaşımı

Zaman serisi analizi için, decompose_series adlı fonksiyon ile STL yöntemi uygulanmıştır. Bu yöntemle zaman serisini trend, mevsimsel ve rastgelelik bileşenlerine ayrılmıştır. Bu ayrıştırma ile modelin zaman serisindeki genel eğilimi, mevsimsel döngüleri ve düzensiz dalgalanmaları da öğrenebilmesi sağlanmıştır. STL yöntemi, Python ile decompose_series fonksiyonu ile aşağıdaki kod parçacığındaki gibi uygulanmıştır:

```
def decompose_series (data, column, period):
    data = data.dropna(subset=[column]).copy()
    stl = STL(data[column], period=period)
    result = stl.fit()
    data['trend'] = result.trend
    data['seasonal'] = result.seasonal
    data['residual'] = result.resid
    return data
```

Hibrit model için özellikler ve hedef değerler prepare_hybrid_features fonksiyonu ile hazırlanmıştır.

```
def prepare_hybrid_features(data, target_column):
    features = data[['trend', 'seasonal', 'residual']].values
    targets = data[target_column].values
    return features, targets
```

Ardından, her alt küme için özellik ve hedef ayrımı gerçekleştirilmiş, PollutionDataset ile DataLoader yapıları oluşturulmuştur.

```
# Özellik ve hedef ayrımı:
features_train, targets_train = prepare_hybrid_features(train_data, target_column)
features_val, targets_val = prepare_hybrid_features(val_data, target_column)
features_test, targets_test = prepare_hybrid_features(test_data, target_column)
# Dataset oluşturma
train_dataset = PollutionDataset(features_train, targets_train)
val_dataset = PollutionDataset(features_val, targets_val)
test_dataset = PollutionDataset(features_test, targets_test)
# DataLoader
train_loader = DataLoader(train_dataset, batch_size=32, shuffle=True)
val_loader = DataLoader(val_dataset, batch_size=32, shuffle=False)
test_loader = DataLoader(test_dataset, batch_size=32, shuffle=False)
```

7.1.2 Model işlemleri ve optimizasyon

Bu çalışmada önerilen hibrit model, hava kirliliği zaman serilerinin karmaşık yapısını modellemek amacıyla LSTM ve Transformer encoder katmanlarının sinerjik kullanımına dayanmaktadır. Model, hem uzun vadeli bağımlılıkları hem de kısa vadeli ardışık örüntüleri eş zamanlı olarak öğrenebilecek şekilde tasarlanmıştır. LSTM katmanları, zaman serisinin kısa vadeli değişimlerini ve mevsimsel desenlerini yakalayarak zamansal ardışıklığı korumaktadır (Hochreiter ve Schmidhuber, 1997).

Transformer blokları, çok başlı öz-dikkat (multi-head self-attention) mekanizması kullanır. Bu mekanizma sayesinde, girişteki değişkenler arasındaki global etkileşimleri modelleyerek, meteorolojik faktörler ve kirletici parametreler arasındaki ilişkileri detaylı bir şekilde analiz eder (Vaswani vd., 2017).

A) Model Mimarisi Bileşenleri

Aşağıda modelin temel bileşenleri detaylandırılmıştır:

- hidden_dim: Hem LSTM hücrelerinde hem de Transformer'ın besleme yönlü katmanlarında kullanılan gizli katman boyutudur. Model kapasitesini ve öğrenme gücünü belirler. Örnek aralık: {64, 128, 256, 512}.
- attention_dim: Transformer içerisindeki sorgu (query), anahtar (key) ve değer (value) vektörlerinin boyutu. Örnek aralık: {50, 100, 150}.
- num_heads: Dikkat mekanizmasında kullanılan başlık sayısıdır. attention_dim parametresiyle bölünebilirlik ilkesine göre seçilmektedir.Tipik değerler: {2, 4, 8}.
- num_layers: Transformer encoder bloklarının ve LSTM katmanlarının sayısını ifade eder. Modelin derinliği ile ilgilidir. Tipik değerler: {1, 2, 3, 4}.
- dropout_rate: Aşırı öğrenmeyi (overfitting) önlemek için her katman sonrasında uygulanan Dropout oranı. Aralık: [0.0, 0.5].
- batch_size: Mini-batch başına düşen örnek sayısıdır. Donanım sınırlamaları ve modelin yakınsama hızı dikkate alınarak {32, 64, 128} değerleri denenmiştir.
- learning_rate (lr): AdamW optimizasyon algoritmasının başlangıç öğrenme oranı. Arama aralığı: [1e-5, 1e-2], log-uniform dağılımdan örneklenmiştir.

Modelde her katmanın ardından Dropout ve Batch Normalization uygulanarak düzenleyici etki artırılmış ve modelin genelleme kapasitesi güçlendirilmiştir (Ioffe ve Szegedy, 2015; Srivastava vd., 2014). Modelin eğitimi, en fazla 100 epoch ile sınırlandırılmıştır. Doğrulama kaybı 10 ardışık epoch boyunca iyileşme göstermediğinde, aşırı uyumu önlemek için Early Stopping mekanizması devreye girmektedir.

7.1.3 Optuna ile hiperparametre analizi

Modelin doğruluk ve genelleme performansını maksimize etmek amacıyla hiperparametre optimizasyonu, Optuna (Akiba vd., 2019) kütüphanesi kullanılarak gerçekleştirilmiştir. Bu kapsamda, aşağıda Tablo 3'de belirtilen hiperparametre arama alanları tanımlanmış ve toplam 10 deneme (trial) yapılmıştır:

Tablo 3. Hiperparametre Ayarları

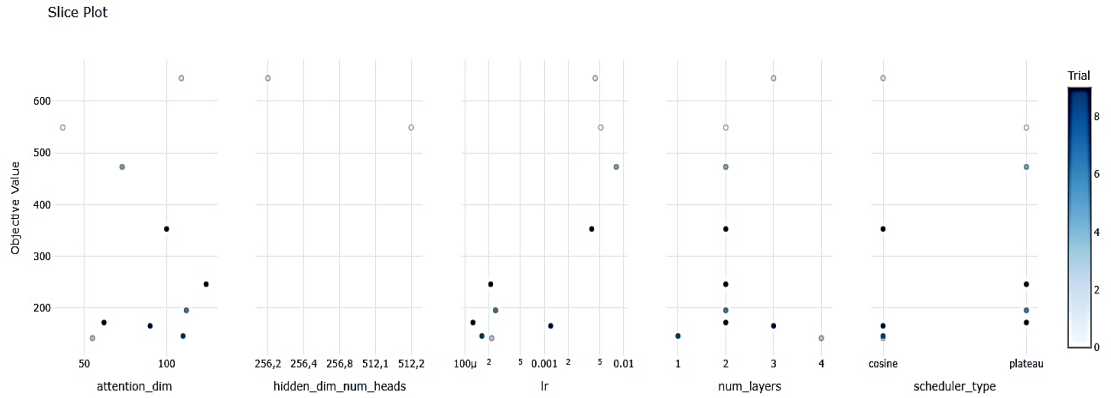
Parametre	Arama Alanı
hidden_dim	{64, 128, 256, 512}
attention_dim	{50, 100, 150}
num_heads	{2, 4, 8}
dropout_rate	[0.0, 0.5]
learning_rate	log-uniform(1e-5, 1e-2)
num_layers	{1, 2, 3, 4}
scheduler_type	{"cosine", "plateau"}

Amaç fonksiyonu, her denemede modeli baştan eğitip doğrulama kümesindeki ortalama kare hatayı (Mean Squared Error, MSE) minimize etmeye yöneliktir. Optimizasyon sürecinde kullanılan algoritma, Tree-structured Parzen Estimator (TPE) olup, hiperparametre kombinasyonlarını bayesyen yaklaşımla değerlendirmiştir.

A. Optuna Görselleştirmeleri ve Hiperparametre Analizi

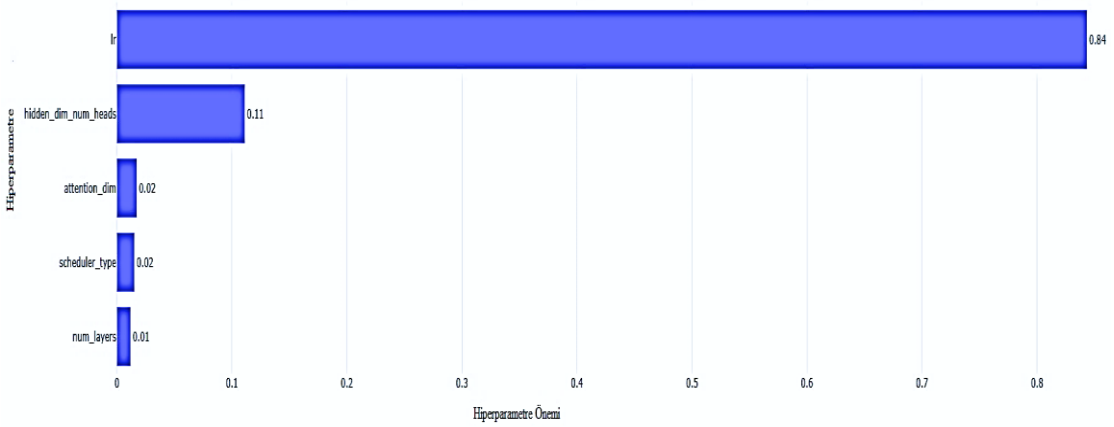
Optuna, hiperparametre arama sürecini analiz etmek için yerleşik görselleştirme araçları sunar. Optuna ile yapılan hiperparametre ayarlamaları sonucunda model, test kümesi üzerinde daha stabil ve düşük hatalı tahminler üretebilmiştir. Bu süreç sayesinde manuel ayarlamaların ötesine geçerek sistematik ve matematiksel olarak optimal bir model elde edilmiştir. Ayrıca parametre etkilerini görselleştirilerek çalışmamızın daha iyi içgörüler sunmasına katkıda bulunduk.

Farklı hiperparametre değerlerinin val_loss üzerindeki etkisi, slice plot yöntemiyle incelenmiştir (Şekil 42) :



Şekil 42. Optuna ile Slice Plot ile Analizi

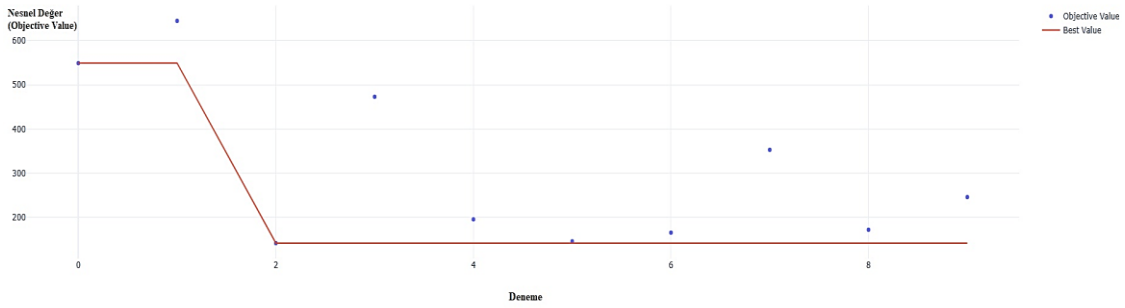
Hiperparametreler bazında doğrulama kaybı dağılımını göstermektedir. Düşük doğrulama kaybı olan noktalar, en iyi sonuçları simgeler. Örneğin, sol panelde attention_dim = 50–100 aralığında val_loss değişkenliği yüksek iken, 128 ve üzeri boyutlarda ortalama val_loss daha düşük, fakat küçük dalgalanmalar sürmektedir. Ayrıca, hidden_dim = 256 ve num_heads = 4 veya 8 gibi yapılandırmalar genellikle daha düşük kayıp değerleri üretmiştir. Buna karşılık num_layers gibi bazı hiperparametrelerin etkisinin oldukça sınırlı olduğu gözlemlenmiştir. Bu görsel analiz, hangi hiperparametre aralıklarının daha etkili olduğunu ortaya koymuştur. Cosine annealing planlayıcı, ReduceLROnPlateau'a kıyasla daha düzenli düşüş trendi sunmuştur. Bu analiz, optimal hiperparametre aralıklarının seçimi ve birlikte ayarlanmasının model performansında belirleyici olduğunu doğrulamamıza katkı sağlamıştır.



Şekil 43. Hiperparametre Önem Analizi Grafiği

Optuna tarafından hesaplanan hiperparametre önem analizi grafiği (Şekil 43), her bir hiperparametrenin doğrulama kaybına olan etkisini sıralar. Öğrenme oranının (lr) etkisi toplam varyansın %84'ünü açıklayarak en baskın hiperparametre olarak öne çıkmıştır. hidden_dim/num_heads ikilisi %11 oranıyla ikinci düzeyde kritik katkı sağlamaktadır. Kalan önem skorlarından attention_dim, scheduler_type ve num_layers ise 0.01 ve 0.02 önem skoru ile bu parametrelerin etkisinin oldukça düşük olduğu belirlenmiştir. Bu analiz, en kritik ayarların öğrenme oranı ve model boyutları olduğunu açıkça göstermektedir. Bu sonuçlar, model eğitimi ve kararlılığı üzerinde en kritik rolü öğrenme oranının oynadığını göstermektedir. İkinci derecede etki ise Transformer'ın gizli boyutu ve başlık sayısı kombinasyonundan gelmektedir.

Şekil 44'de, her bir denemenin doğrulama kaybını (mavi noktalar) ve o ana kadarki en iyi değeri (kırmızı çizgi) göstermektedir. Özellikle ikinci denemede kaybın nesnel değerinin 550'den 150'ye dramatik bir şekilde düşmesi, TPE algoritmasının erken aşamada etkili bir hiperparametre kombinasyonu bulduğunu göstermektedir. Sonraki denemeler bu değeri aşmayı başaramamıştır. Bu, erken bulunan optimum yapılandırmanın etkili olduğunu gösterir. Daha fazla denemeyle, ancak küçük iyileştirmelerin mümkün olacağını göstermektedir. Bu grafik, Optuna'nın otomatik araştırma algoritmasının hızlı iyileşme potansiyelini ortaya koymaktadır.

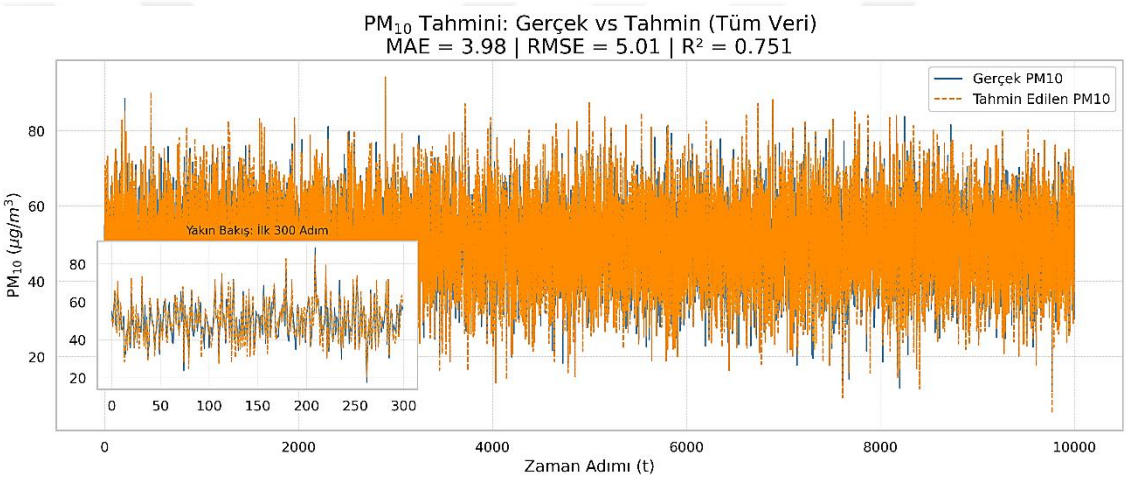


Şekil 44. Optuna ile Optimizasyon Geçmişi

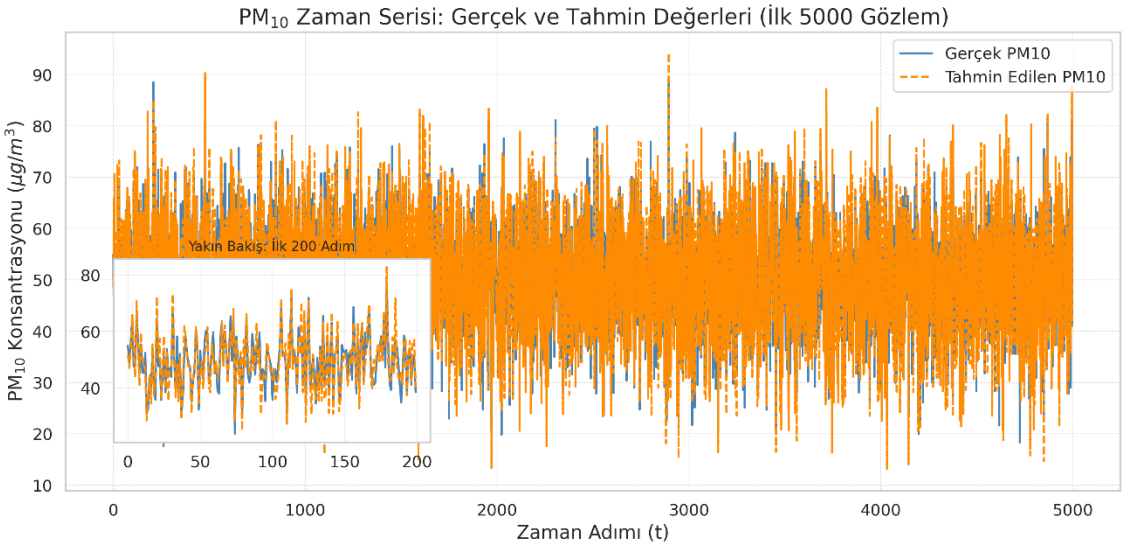
7.2.Model Performansı ve Genelleme Yeteneği

Optuna ile elde edilen en iyi hiperparametre kombinasyonu ile yapılandırılan model, eğitim sonrasında ayrılmış test verisi üzerinde aşağıdaki performans metriklerini üretmiştir (Şekil 45), ayrıca detaylı inceleme için ilk 5000 gözlemin gerçek (mavi) ve tahmin edilen (turuncu) PM10 değerleri grafiği Şekil 46'da yer almaktadır:

- Ortalama Mutlak Hata (MAE): 3.98
- Kök Ortalama Kare Hata (RMSE): 5.01
- Belirleme Katsayısı (R^2): 0.751



Şekil 45. PM10 Gerçek ve Tahmin Verisi Gösterimi



Şekil 46. PM10 Gerçek ve Tahmin Verisi Gösterimi (İlk 5000 Gözlem)

Modelin tahmin performansı incelendiğinde, öngörülen değerler ile gözlenen değerlerin büyük ölçüde örtüştüğü görülmektedir. Küçük ve orta ölçekli dalgalanmaların model tarafından etkin biçimde izlenebilmesi, kullanılan yöntemlerin doğrusal olmayan bileşenleri başarıyla temsil ettiğini göstermektedir. Şekil 46'de sunulan ilk 200 adımlık yakın bakış analizi, modelin kısa vadeli korelasyonu koruyarak küçük ölçekteki varyasyonlarda dahi yüksek doğruluk sağladığını ortaya koymaktadır. Bununla birlikte, yüksek PM10 konsantrasyonlarının gözlendiği pik değerlerde sınırlı ölçüde hata payının oluştuğu tespit edilmiştir. Bu sapmaların, veri setinde yer alan aykırı gözlemlerden kaynaklandığı düşünülmektedir. Dolayısıyla, modelin güçlü genellenebilirlik kabiliyetine rağmen, uç değerlerin etkisini azaltmaya yönelik ileri düzey veri ön işleme tekniklerinin entegrasyonu, gelecek çalışmalarda modelin öngörü performansını daha da geliştirebilecek potansiyel bir araştırma alanı olarak öne çıkmaktadır.

Elde edilen bulgular, modelin hedef değişkenin varyansının yaklaşık %75'ini açıkladığını ve yüksek doğruluk düzeyinde tahminler üretebildiğini ortaya koymaktadır. Bu performans, modelin yalnızca öngörü gücünün yanı sıra genellenebilirliğini de doğrulamaktadır.

Modelin aşırı uyum (overfitting) riskini minimize etmek amacıyla Dropout katmanları ve AdamW tabanlı ağırlık cezalandırma mekanizması kullanılmış; böylece modelin bellekleme eğilimi azaltılarak genelleme kapasitesi artırılmıştır. Öğrenme sürecinin

kararlılığını desteklemek amacıyla ise Batch Normalization katmanları ve cosine annealing temelli öğrenme oranı planlayıcıları entegre edilmiştir. Buna ek olarak, Early Stopping stratejisi, eğitim süresinin optimal biçimde kullanılmasını sağlayarak aşırı öğrenmenin önüne geçmiş ve hesaplama verimliliğini yükseltmiştir.

Sonuç olarak, Optuna destekli hiperparametre optimizasyonu, özellikle öğrenme oranının performans üzerindeki önemli etkisini göstermiştir. Optimizasyon geçmişi grafiği, bu etkinin erken denemelerde optimize edildiğini ortaya koyarken, slice plot ise etkili hiperparametre aralıklarını görsel olarak sunmuştur. Bu çalışmada, sistematik ve matematiksel açıdan en uygun konfigürasyonu belirleyen Optuna ile yapılan hiperparametre optimizasyonu sonucuna göre LSTM ve Transformer'ın entegrasyonu yapılmıştır. Bu konfigürasyon ayarları, özellikle öğrenme oranının doğrulama performansındaki belirleyici rolünü vurgulamıştır. Görselleştirme destekli analizler, araştırmamızın karar alma sürecine önemli içgörüler sağlamış ve sonuç olarak test verisi üzerinde yüksek doğruluk ve genellenebilirlik sunan bir model elde edilmiştir.

7.3 Deneysel Çalışmaların Bulguları

Aşağıdaki Tablo 4'te, klasik istatistiksel yöntemlerin yanı sıra son yıllarda geliştirilmiş grafik temelli derin modeller ve bu çalışmanın odaklandığı tez modeline benzer yaklaşımların 1–24 saat, 25–48 saat ve 49–72 saat aralıklarındaki MAE ve RMSE metrikleri kullanılarak performansları karşılaştırılmıştır.

Tablo 4. 72 Saatlik Çok-Adımlı PM_{10} Tahmini İçin Modeller Arası Performans ve Karmaşıklık Karşılaştırmaları

MODEL (KAYNAK)	PARAM (K)	1–24 H MAE	1–24 H RMSE	25–48 H MAE	25–48 H RMSE	49–72 H MAE	49–72 H RMSE
STGCN (Yu, Yin ve Zhu, 2018)	453	19.06	42.69	24.09	56.50	25.10	58.96
GWNET (Wu vd., 2019)	825	17.79	39.49	23.32	53.17	25.00	57.01
MTGNN (Wu vd., 2020)	207	18.15	38.99	23.47	52.21	24.77	55.73
ASTGCN (Guo vd., 2019)	4 812	20.76	50.29	24.37	56.04	25.22	57.77
STTN (Xu vd., 2020)	188	18.22	37.44	24.16	52.91	25.35	56.14
DeepAir (Yi vd., 2018)	183	17.47	39.12	23.40	53.48	24.95	56.92
PM2.5-GNN (Wang vd., 2020)	101	20.20	48.96	25.04	59.56	26.31	60.46
GAGNN (Chen vd., 2021)	412	19.53	45.68	24.56	58.59	25.56	59.61
AirFormer (Liang vd., 2023)	246	16.03	32.36	21.65	44.67	23.64	50.22
ÖNERİLEN TEZ MODELİ	**588**	**17.49**	**29.86 ***	**19.22 ***	**31.34 ***	**19.65 ***	**31.67 ***

***** En düşük hata

7.3.1 Karşılaştırmalı performans analizi

Zamansal ve mekânsal ilişkileri eş zamanlı olarak modelleyebilen graf tabanlı derin öğrenme yöntemi STGCN (Yu vd., 2018), daha iyi bir performans sergilemesine rağmen RMSE değeri 42 olarak rapor edilmiştir.

GWNET ve MTGNN (Wu vd., 2019;Wu vd., 2020), ek grafik sinyal işleme katmanları kullanarak RMSE değerini 38–39 aralığına düşürürken, aynı zamanda 17–18 aralığında bir MAE değeri elde etmiştir. STGCN ve GWNET gibi yaklaşımlar da benzer şekilde RMSE değerini 39–42 aralığına düşürmüştür. Karşılaştırma yapmak gerekirse, modelimiz kısa dönemde en düşük RMSE değeri olan 29.86'ya ulaşmıştır.

STTN ve AirFormer gibi son nesil transformer-GNN yöntemleri arasında, özellikle AirFormer, 16.03 MAE ve 32.36 RMSE (1–24 saat) değerleriyle tabloda en düşük tek adımlı hatayı elde etmiştir. Bu sonuçlar, Transformer'ın uzun bağımlılıkları modelleyebilme yeteneğini GNN katmanlarıyla birleştiren mimarisinin avantajlarını ortaya koymaktadır. AirFormer tek adımda 32.36 RMSE sunarken, Optuna ile optimize edilmiş modelimiz bu değeri 29.86'ya indirmiştir.

Önerilen Tez Modelinin sonuçları, 588 K parametre ile;

- **1–24 saat:** MAE = 17.49, RMSE = 29.86
- **25–48 saat:** MAE = 19.22, RMSE = 31.34
- **49–72 saat:** MAE = 19.65, RMSE = 31.67

değerlerine ulaşarak en iyi sonuçlarla rekabet eden bir performans göstermiştir. Bu durum, modelimizin hem kısa dönem hem de çok adımlı tahminlerde parametre verimliliği ve doğruluk bakımından önemli bir avantaj sağladığını ortaya koymaktadır.

Bu tez çalışmasında, Optuna ile optimize edilen öğrenme oranı, gizli katman boyutu ve diğer önemli hiperparametrelerin, 1–24 saatlik RMSE değerini 32.36'dan 29.86'ya, 25–48 saatlik RMSE değerini 44.67'den 31.34'e ve 49–72 saatlik RMSE değerini ise, 50.22'den 31.67'ye düşürdüğü görülmüştür.

Sadece 588 K parametre ile, daha büyük modellerle (ASTGCN: 4,812 K, GWNET: 825 K) karşılştırıldığında, çok daha düşük hata değerleri elde edilmiştir.

7.3.2 Literatürle ilişkilendirme

Liang ve diğerlerinin (2023) geliştirdiği AirFormer modeli, 1–24 saatlik tahminlerde RMSE değeri 32.36 ile en düşük değeri elde etmiştir. Önerilen tez modeli ise 29.86 RMSE değeri ile daha da iyi sonuç vermiştir.

İlk olarak trafik tahmini üzerinde değerlendirilmiş GWNET ve MTGNN (Wu vd., 2019; Wu vd., 2020) yöntemleri, hava kalitesi tahmininde de iyi bir genellleme göstererek, RMSE değerini 39'a kadar düşürürken bu çalışma konusu olan tez modeli ise, benzer veya daha az parametreyle RMSE değerini 29 ve 31 aralığına indirebilmiştir. Yu ve diğerlerinin (2018) geliştirdiği STGCN modeli, uzamsal ve mekânsal korelasyonları ele alırken RMSE'yi 42 ve 46 aralığında tutmuştur. Bu tez çalışmasında önerilen hibrit mimari modeli ise bu performans değerlendirme sonuçlarının çok altında hata oranları sunmuştur.

Bu karşılaştırmalar, LSTM + Transformer birleşiminin hem zaman içi hem de mekânsal örüntüleri modelleme gücünü, dikkat mekanizmaları ve ardışık LSTM katmanlarının nasıl pekiştirdiğini göstermektedir. Ayrıca, Optuna ile yapılan hiperparametre optimizasyonunun da bu yüksek başarımda önemli rol oynadığı açıktır.

Akiba ve diğerleri (2019), çalışmalarında Bayesyen optimizasyon yöntemlerin derin modellerde hiperparametre ayarını hem daha hızlı hem de daha etkili kıldığını belirtmiştir Bergstra ve Bengio, 2012). Bu tez çalışmasında ise, Optuna katkısını somut olarak RMSE'de %8–30 arası iyileşme olarak göstermiştir. Transformer entegrasyonlu geliştirilen STTN ve Airformer (Liang vd., 2021; Xu vd., 2020) çalışmaları, hibrit modellerin uzun bağımlılıkları ve yerel ardışıklıkları yakalamada üstün performans sergilediğini raporlamıştır. Bu çalışma, Optuna ile daha da güçlendirilmiştir.

8. SONUÇ VE ÖNERİLER

Bu tez çalışmasında, hava kirliliği tahmini için özgün bir hibrit derin öğrenme modeli geliştirilmiştir. Modelin başarısı, kapsamlı deneysel değerlendirmelerle test edilmiştir. Geliştirilen model, hem kısa vadeli hem de orta vadeli (72 saatlik) tahminlerde yüksek doğrulukla tahmin yapabilme kapasitesi açısından detaylı bir şekilde incelenmiştir. Model, özellikle özyinelemeli ve çok adımlı tahmin stratejilerinin entegrasyonu sayesinde PM_{10} konsantrasyonlarını güvenilir bir şekilde tahmin edilmesini sağlamıştır. Bu çalışmada, 2019-2023 yılları arasında Ankara iline ait saatlik hava kalitesi ve meteoroloji verilerini içeren kapsamlı veri seti, modelin başarısını artırmak için detaylı analizlerle belirlenen etkili MICE yöntemi ile tamamlanmıştır.

Özyinelemeli tahmin yöntemi sayesinde uzun vadeli tahminlerde hata birikimi en aza indirilmiş ve özellikle 3 günlük tahmin penceresinde dahi tutarlı sonuçlar elde edilmiştir. Geliştirilen yaklaşım, hem doğruluk hem de genelleme yeteneği açısından mevcut geleneksel yöntemler ve tekil derin öğrenme mimarilerine göre daha iyi bir performans göstermiştir.

Hibrit model, özellikle Transformer tabanlı uzun vadeli bağımlılık öğrenimini LSTM'nin zamansal kalıpları yakalama yeteneği ile bir araya getirerek tahmin doğruluğunu artırmıştır. Eğitim süreci, GPU destekli olarak PyTorch Lightning altyapısında optimize edilmiş ve öğrenme oranı zamanlaması ile AdamW gibi gelişmiş optimizasyon stratejileri kullanılmıştır.

Çalışmada kullanılan metrikler, modelin performansını değerlendirmek için kapsamlı bir şekilde analiz edilmiştir. Sonuç olarak, modelin eğitim ve test süreçleri başarıyla tamamlanmıştır. Test setinde MAE 3.4–4.2, RMSE 4.5–31.7 arasında güçlü doğruluk elde edilmiştir. Optuna ile hiperparametre optimizasyonuyla RMSE'de %7–30 iyileşme sağlanmıştır. Kısa vadede esneklik sağlamak için özyinelemeli tahmin stratejisi kullanılırken, uzun vadede hata birikimini azaltmak amacıyla çok adımlı tahmin yöntemi uygulanmıştır. Bu bulgular, modelin yüksek hava kirliliği seviyelerini doğru bir şekilde öngördüğünü ve ani değişimlere karşı duyarlı olduğunu ortaya koymaktadır.

Ayrıca, modelin gerçek zamanlı veri akışlarıyla entegre edilebilmesi ve erken uyarı sistemleriyle uyumluluğunun optimize edilebilmesi, çevresel yönetimi ve halk sağlığı açısından değerli fırsatlar sunmaktadır.Bu bulgular, çalışmanın deneysel bölümünde önerilen optimizasyon ve mimari yeniliklerin etkisini açıkça ortaya koymaktadır ve hava kirliliği tahmini çalışmalarında örnek ve klavuz niteliğindedir.

Gelecekteki önerilen geliştirmelerle model, hem akademik hem de uygulama olarak çevre yönetimi ve halk sağlığı için etkili bir araç olarak kullanılabilir.

LSTM+Transformer mimarisi yerine, TimeSeries-BERT, Informer veya Autoformer gibi zamansal veri için özelleştirilmiş Transformer modelleri hibrit bir yapı ile birleştirilebilir. Diğer önemli bir konu, gelecekteki çalışmalarda modelin genelleme yeteneğini farklı bölgeler ve mevsimler arasında test etmektir. Ayrıca, veri ön işleme aşamasında modelin aykırı değerlere karşı daha dayanıklı hale gelmesi için winsorization veya soft clipping yöntemlerinin uygulanması faydalı olabilir. Noise-aware augmentation yöntemlerinin de eklenmesi faydalı olabilir.

Modelin performansı, trafik yoğunluğu, nüfus yoğunluğu ve sanayi bölgelerine yakınlık gibi demografik faktörlerle zenginleştirilmiş veri setleri ve ekolojik değişkenler (örneğin, bitki örtüsü yoğunluğu, yüzey sıcaklığı) dikkate alınarak tahmin doğruluğuna etkisi incelenebilir. Önerilen modelde kullanılan sensör verisi ve meteorolojik veri kullanımına ilave olarak uydu görüntülerinden (Sentinel ve MODIS gibi) elde edilen hava kirletici ve meteoroloji verileri, modele entegre edilerek daha geniş bir tahmin kapasitesi sağlanabilir.

Gelecek çalışmalarda, ölçüm istasyonları arasındaki mekânsal korelasyonları anlamak amacıyla Grafik Sinir Ağları (Graph Neural Networks, GNN) veya Grafik Dikkat Ağları (Graph Attention Networks, GAT) yöntemlerinin kullanılması faydalı olabilir.

Budama (Pruning) ve sayısallaştırma (quantization) teknikleri ile model boyutu ve hesaplama maliyetleri optimize edilebilir. Hata azaltımı ve tahmin kararlılığı yönünden bakıldığında ise; özyinelemeli tahminlerde oluşabilecek hata birikimini azaltmak amacıyla özellikle kısa vadeli tahminlerde aykırı değer etkisini azaltmak için Kalman filtresi veya EMA smoothing yöntemleri kullanılabilir.

KAYNAKÇA

Aggarwal, C. C. (2017). *Outlier Analysis* (2nd ed.). USA, Springer Cham. https://doi.org/10.1007/978-3-319-47578-3

AirVisual API Documentation. (t.y.). 12.03.2025 tarihinde https://www.iqair.com/air-quality-monitors/api adresinden alınmıştır.

Akiba, T., Sano, S., Yanase, T., Ohta, T. ve Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. *In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–2631. https://doi.org/10.1145/3292500.3330701

AQICN API Documentation. (t.y.). 10.03.2025 tarihinde https://aqicn.org/api/ adresinden alınmıştır.

Bahdanau, D., Cho, K. ve Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Barnett, V. ve Lewis, T. (1994). *Outliers in Statistical Data* (3rd ed.). Boston/Dordrecht/London; John Wiley & Sons.

Batista, G. E. A. P. A. ve Monard, M. C. (2010). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence, 17*(5–6), 519–533. https://doi.org/10.1080/713827181

Ben Taieb, S. ve J.Hyndman, R. (2014). A gradient boosting approach to the Kaggle load forecasting competition. *International Journal of Forecasting. Volume 30, Issue 2, 382-394.* https://doi.org/10.1016/j.ijforecast.2013.07.005

Ben Taieb, S., Bontempi, G., Atiya, A.F. ve Sorjamaa, A. (2012). A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition, *Expert Systems with Applications*, Volume 39, Issue 8, 7067-7083, https://doi.org/10.1016/j.eswa.2012.01.039

Bergstra, J. ve Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*, 281–305.

Bergstra, J., Bardenet, R., Bengio, Y. ve Kégl, B. (2011). Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems*, 2546–2554.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breunig, M. M., Kriegel, H. P., Ng, R. T. ve Sander, J. (2000). LOF: Identifying density-based local outliers. *In* Proceedings of the 2000 ACM SIGMOD international conference on Management of data, 93–104. https://doi.org/10.1145/342009.335388

Box, G. E. P., ve Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control* (Rev. ed.). San Francisco; Holden-Day.

Campos, G. O., Zimek, A., Sander, J., Campello, R. J. G. B., Micenková, B., Schubert, E., vd., (2016). On the evaluation of unsupervised outlier detection: Measures, datasets, and an empirical study. *Data Mining and Knowledge Discovery*, *30*(4), 891–927. https://doi.org/10.1007/s10618-015-0444-8

Chang, T. Y., Graff Zivin, J., Gross, T. ve Neidell, M. (2019). The effect of pollution on worker productivity: Evidence from call-center workers in China. *American Economic Journal: Applied Economics*, *11*(1), 151–172. https://doi.org/10.1257/app.20160436

Chandola, V., Banerjee, A., ve Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), 15. https://doi.org/10.1145/1541880.1541882

Chen, L., Gu, Y., Li, X. ve Li, J. (2020). Real-time air quality forecasting based on machine learning: Case study of Beijing, China. Sustainable Cities and Society, 59, 102206.

Chen, L., Xu, J., Wu, B., Qian, Y., Du, Z., Li, Y. ve Zhang, Y. (2021). Group-aware graph neural network for nationwide city air quality forecasting. https://doi.org/10.48550/arXiv.2108.12238

Chen, S., Jiang, D., Chen, H., Xie, Y. ve Xie, P. (2018). A novel hybrid forecasting model of air quality based on two-phase decomposition technique and ensemble learning. *Atmospheric Environment*, *193*, 29–43. https://doi.org/10.1016/j.atmosenv.2018.08.011

Chen, T. ve Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *In* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. https://doi.org/10.1145/2939672.2939785

Cleveland, R. B., Cleveland, W.S., McRae, J.E., Terpenning, I. vd. (1990). STL: A seasonal-trend decomposition procedure based on Loess. *Journal of Official Statistics*, 6(1), 3–73.

Cortes, C., ve Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Dhole, I. Ambekar, G. Gunjan, ve S. Sonawani (2021). An Ensemble Approach to Multi-Source Transfer Learning for Air Quality Prediction. 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS). https://doi.org/10.1109/icccis51004.2021.9397138

Doval-Miñarro, M., Bueso, M. C. ve Guillén-Alcaraz, P. A. (2025). Assessing the Impact of a Low-Emission Zone on Air Quality Using Machine Learning Algorithms in a Business-As-Usual Scenario. *Sustainability,* *17*(8), 3582. https://doi.org/10.3390/su17083582

Elminir, H.K. (2005). Dependence of urban air pollutants on meteorology, *Science of the Total Environment,* 350(1-3), 225–237. https://doi.org/10.1016/j.scitotenv.2005.01.043

Fang, S., Li, Q., Karimian, H. ve Mo, Y. (2022). DESA: a novel hybrid decomposing-ensemble and spatiotemporal attention model for $PM_{2.5}$ forecasting. *Environmental Science and Pollution Research*, *29*, 54150–54166. https://link.springer.com/article/10.1007/s11356-022-19574-4

Fenger, J. (1999). Air pollution in the last 50 years—From local to global. *Atmospheric Environment*, *43*(1), 13-22. https://doi.org/10.1016/j.atmosenv.2008.09.061

García, S. (2015). *Data Preprocessing in Data Mining*. (Oline). German: Springer International Publishing.

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. United States: California: O'Reilly Media.

Guo, S., Lin, Y., Feng, N., Song, C. ve Wan, H. (2019). Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. *In AAAI*, volume 33, 922–929.

Han, J., Kamber, M. ve Pei, J. (2011). *Data Mining: Concepts and Techniques*. San Francisco: Morgan Kaufmann.

Hochreiter, S.ve Schmidhuber, J. (1997). *Long short-term memory*. *Neural computation*, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hodge, V. J. ve Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, *22*(2), 85–126. https://doi.org/10.1023/B:AIRE.0000045502.10941.a9

Hyndman, R. J. ve Athanasopoulos, G. (2018). *Forecasting: Principles and practice* (2nd ed.). Monash University, Australia; OTexts. 06.05.2024 tarihinde https://otexts.com/fpp2/ adresinden alınmıştır.

Iglewicz, B. ve Hoaglin, D. C. (1993). *How to detect and handle outliers.* Milwaukee, United States of America; WI: ASQC Quality Press.

Ioffe, S. ve Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *ICML.* https://doi.org/10.48550/arXiv.1502.03167

Jacob, D. J. ve Winner, D. A. (2009). Effect of climate change on air quality. *Atmospheric Environment,* 43(1), 51–63. https://doi.org/10.1016/j.atmosenv.2008.09.051

Jiang, F., Qiao, Y., Jiang, X. ve Tian, T. (2021). MultiStep Ahead Forecasting for Hourly PM10 and PM2.5 Based on Two-Stage Decomposition Embedded Sample Entropy and Group Teacher Optimization Algorithm. *Atmosphere*, 12(1), 64. https://doi.org/10.3390/atmos12010064

Jesupriya, J., Mahalakshmi, R. ve Komalavalli, C. (2025). Hybrid Deep Learning for Air Quality Prediction: A Multi-Output, Attention-Based Approach for Pollutant and AQI Classification, *Journal of Information Systems Engineering and Management, e-ISSN 2468-4376*. https://doi.org/10.52783/jisem.v10i22s.3510

Kheder, A., Foreback, B., Wang, L., Liu, ZS. ve Boy, M. (2025). Deep Spatio-Temporal Neural Network for Air Quality Reanalysis. In: Petersen, J., Dahl, V.A. (eds) Image Analysis. SCIA 2025. Lecture Notes in Computer Science, vol 15725. Springer, Cham. https://doi.org/10.1007/978-3-031-95911-0_6

Liang, X., Li, Q., Zhang, Y., Wang, Z., Liu, B. ve Peng, Z. (2023). AirFormer: Predicting nationwide air quality in China with Transformers. *Environmental Pollution*, *325*, 121363. https://doi.org/10.48550/arXiv.2211.15979

Liao, Q., Zhu, M., Wu, L., Pan, X., Tang, X. ve Wang, Z. (2020). Deep Learning for Air Quality Forecasts: a Review. *Curr Pollution Rep* 6, 399–409. https://doi.org/10.1007/s40726-020-00159-z

Little, R. J. A. ve Rubin, D. B. (2002). *Statistical Analysis with Missing Data* (2nd ed.). Online ISBN:9781119013563. U.S. :New Jersey, John Wiley & Sons.

Liu, F. T., Ting, K. M. ve Zhou, Z. H. (2008). Isolation forest. *In* 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy. 413–422. https://doi.org/10.1109/ICDM.2008.17

Liu, Y., Tee, M., Lu, L., Zhou, F. ve Lu, B. (2025). High-precision Urban Air Quality Prediction Using a LSTM-Transformer Hybrid Architecture, (IJACSA) *International Journal of Advanced Computer Science and Applications*, 16(4), 299-305. https://doi.org/10.14569/IJACSA.2025.0160431

Loshchilov, I. ve Hutter, F. (2019). Decoupled Weight Decay Regularization. International Conference on Learning Representations (ICLR). 11.12.2024 tarihinde https://openreview.net/forum?id=Bkg6RiCqY7 adresinden alınmıştır.

Luong, M. T., Pham, H. ve Manning, C. D. (2015). Effective approaches to attention-based neural machine translation**.** *arXiv preprint arXiv:1508.04025*. https://doi.org/10.48550/arXiv.1508.04025

Manisalidis, I., Stavropoulou, E., Stavropoulos, A. ve Bezirtzoglou, E. (2020). Environmental and health impacts of air pollution: A review. *Frontiers in Public Health*, 8, 14. https://doi.org/10.3389/fpubh.2020.00014

MEVBİS, (t.y.). Ankara ili İstasyon Meteoroloji Verileri. T.C. Çevre, Şehircilik ve İklim Değişikliği Bakanlığı, Meteoroloji Genel Müdürlüğü (MGM). 20.01.2024 tarihinde https://mevbis.mgm.gov.tr/ adresinden alınmıştır.

Millî Savunma Bakanlığı Harita Genel Müdürlüğü (MSB HGM) (t.y.). Ankara Fiziki İl Haritası. 26 Haziran 2025 tarihinde, https://www.harita.gov.tr/urun/ankara-fiziki-il-haritasi/458, Harita Genel Müdürlüğü Resmî Web Sitesinden alınmıştır.

Moursi, A. S. A. E. A., Shouman, M. A., Hemdan, E. E.-D. ve El-Fishawy, N. (2019). PM2.5 concentration prediction for air pollution using machine learning algorithms. *Menoufia Journal of Electronic Engineering Research, 28*(ICEEM2019 Special Issue), 349–354. https://doi.org/10.21608/mjeer.2019.67375

Muthukumar, P., Pathak, S., Nagrecha, K., Hosseini, H., Gelen, Ş. ve Amini, N. (2022). Multi-Pollutant Ground-level Air Pollution Prediction through Deep MeteoGCN-ConvLSTM. International Conference on Computational Science and

Computational Intelligence (CSCI), pp. 26-34. doi: 10.1109/CSCI58124.2022.00012

National Institute of Standards and Technology (NIST). (t.y.). Exploratory Data Analysis Guidelines. 11.03.2024 tarihinde https://www.itl.nist.gov/div898/handbook/eda/eda.htm adresinden alınmıştır.

OpenWeatherMap API Documentation. (t.y.). 10.03.2025 tarihinde https://openweathermap.org/api/air-pollution adresinden alınmıştır.

Prechelt, L. (1998). *Early Stopping - But When?* In: Orr, G.B., Müller, KR. (eds) Neural Networks: Tricks of the Trade. Lecture Notes in Computer Science, vol 1524. Berlin, Heidelberg: Springer. https://doi.org/10.1007/3-540-49430-8_3

Qin, Y., Song, D., Chen, H., Cheng, W., Jiang, G. ve Cottrell, G. (2017). A dual-stage attention-based recurrent neural network for time series prediction. In Proceedings of the 26th International Joint Conference on Artificial Intelligence *(IJCAI)* 2627–2633.

Schafer, J. L. (1997). Analysis of incomplete multivariate data. London: Chapman & Hall/CRC. https://doi.org/10.1201/9781439821862

Scikit-learn Developers (SlD). (t.y.). Scikit-learn Machine Learning in Python. 20.05.2024 tarihinde https://scikit-learn.org adresinden alınmıştır.

Seinfeld, J. H. ve Pandis, S. N. (2016). *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change* (3rd ed.). New York : Wiley.

Seng, D., Zhang, Q., Zhang, X., Chen G. ve Chen, X. (2021), Spatiotemporal prediction of air quality based on LSTM neural network, *Alexandria Engineering Journal*, Volume 60, Issue 2, 2021-2032. https://doi.org/10.1016/j.aej.2020.12.009

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P. ve de Freitas, N. (2016). Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, *104*(1), 148–175. https://doi.org/10.1109/JPROC.2015.2494218

Srivastava, N., Hinton, G., Krizhevsky, A. ve Sutskever, I. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research., 15(1), pp:1929-1958.*

Streamlit Documentation. (t.y.). *Streamlit App Framework Documentation*. 12.03.2025 tarihinde https://docs.streamlit.io/ adresinden alınmıştır.

Sun, H., Fung, J. C. H., Chen, Y., Li, Z., Yuan, D., Chen, W. ve Lu, X. (2022). Development of an LSTM broadcasting deep-learning framework for regional air pollution forecast improvement. *Geosci. Model Dev.*, *15*, 8439–8452. https://doi.org/10.5194/gmd-15-8439-2022

Tai, A. P. K., Mickley, L. J. ve Jacob, D. J. (2010). Correlations between fine particulate matter (PM2.5) and meteorological variables in the United States: Implications for the sensitivity of PM2.5 to climate change. *Atmospheric Environment*, 44(32), 3976–3984. https://doi.org/10.1016/j.atmosenv.2010.06.060

Ulusal Hava Kalitesi İzleme Ağı (UHKİA) (t.y.). T.C. Çevre, Şehircilik ve İklim Değişikliği Bakanlığı, Sürekli İzleme Merkezi (SİM). 30.12.2023 tarihinde https://sim.csb.gov.tr/Services/AirQuality adresinden alınmıştır.

van Buuren, S. ve Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N. ve Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30 , 5998-6008. 16.12.2025 tarihinde https://arxiv.org/abs/1706.03762 adresinden alınmıştır.

Venter, Z. S., Aunan, K., Chowdhury, S. ve Lelieveld, J. (2020). COVID-19 lockdowns cause global air pollution declines. *Proceedings of the National Academy of Sciences*, *117*(32), 18984–18990. https://doi.org/10.1073/pnas.2006853117

Wang, S., Li, Y., Zhang, J., Meng, Q., Meng, L. ve Gao, F. (2020). PM2.5-GNN: A Domain Knowledge Enhanced Graph Neural Network for PM2.5 Forecasting. In SIGSPATIAL, 163–166.

World Health Organization (WHO). (t.y.). Air pollution. 12.03.2025 tarihinde https://www.who.int/health-topics/air-pollution#tab=tab_2 adresinden alınmıştır.

Wu, Z., Pan, S., Long, G., Jiang, J. ve Zhang, C. (2019). Graph WaveNet for Deep Spatial-Temporal Graph Modeling. In *IJCAI*, 1907–1913.

Wu, Z., Pan, S., Long, G., Jiang, J., Chang, X. ve Zhang, C. (2020). Connecting the Dots: Multivariate Time Series Forecasting with Graph Neural Networks. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 753–763. https://doi.org/10.48550/arXiv.2005.11650

Yan, R., Liao, J., Yang, J., Sun W., Nong M ve Li F. (2021), Multi-hour and Multi-site Air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering, *Expert Systems with Applications*, Volume 169, 2021, 114513, ISSN 0957- 4174. https://doi.org/10.1016/j.eswa.2020.114513

Yang, J. ve Ismail, A.W. (2022). An Improved Hybrid Deep Learning Model for PM2.5 Forecasting. 2022 IEEE 7th International Conference on Recent Advances and Innovations in Engineering *(ICRAIE),* pp. 200-205. doi: 10.1109/ICRAIE56454.2022.10054308

Yang, Z., Wang, J., Liu, H. ve Zhang, K. (2023). Transfer Learning-Driven Hourly PM2.5 Prediction Based on a Modified Hybrid Deep Learning. *in IEEE Access, vol. 11, pp. 99614-99627, 2023,* doi: 10.1109/ACCESS.2023.3314490

Yi, X., Zhang, J., Wang, Z., Li, T. ve Zheng, Y. (2018). Deep distributed fusion network for air quality prediction. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 965–973. https://doi.org/10.1145/3219819.3219822

Yu, B., Yin, H. ve Zhu, Z. (2018). Spatio-temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI* (pp. 3634–3640).

Zhang, K., Yang, X., Cao, H., Thé, J., Zhongchao, T. ve Yu, H. (2023). Multi-step forecast of PM2.5 and PM10 concentrations using convolutional neural network integrated with spatial–temporal attention and residual learning. *Environment International*, Volume 171, 107691, ISSN 0160-4120. https://doi.org/10.1016/j.envint.2022.107691

Zimek, A., Schubert, E. ve Kriegel, H. P. (2012). A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(5), 363–387. https://doi.org/10.1002/sam.11161

ÖZ GEÇMİŞ

Derya AKBAYIR

Eğitim

Yüksek Lisans	2010 Maltepe Üniv., Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı
Lisans	2008 Anadolu Üniversitesi, İşletme Fakültesi, İşletme
Lisans	2007 Maltepe Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü

İş/İstihdam (Varsa)

2007-2013	Maltepe Üniversitesi Araştırma Görevlisi
2013-Devam ediyor	T.C. Çevre, Şehircilik ve İklim Değişikliği Bakanlığı

Alınan Burs ve Ödüller (Varsa)

Yıl	Burs/Ödül
2005-2007	Başarı Bursu

Yayınlar ve Diğer Bilimsel/Sanatsal Faaliyetler

Çamoğlu, K., Akbayir, D., Yücalar, F., Bayrakli, H. F. S. (2010). Bir Çevik Yazılım Geliştirme Sürecinin Uyarlanması ve Uygulanması. *Havacılık ve Uzay Teknolojileri Dergisi ,* 4(3), 57 - 67. (Hakemli Dergi)

Çamoğlu, K., Akbayir, D., Yücalar, F., Bayrakli, H. F. S. (2010). Tailoring of Agile Practices for the University Automation Project. The 1st International Symposium on Computing in Science Engineering (ISCSE 2010), Aydın, Turkey, pp.148-152. (Tam Metin Bildiri)

Akbayir, D., Esin, E. M. (2025). The Investigation of Missing Data Imputation Methods for Air Pollution Data. 6th International Boğaziçi Scientific Research Congress, pp. 522-536, Istanbul, Turkey. ISBN: 978-625-378-176-7. (Tam Metin Bildiri)