Ministry of Science and Higher Education of the Republic of Kazakhstan

SDU University

Nuray Dauletkhan

A comparative study of air quality analysis in Almaty

THESIS

Presented in Partial Fulfilment for the

Degree of Master of Technical Science in Computer Science

(degree code: 7M06102)

Department of Computer Science

Faculty of Engineering and Natural Sciences

Supervisor: **PhD Khaled Mohamad**

Kaskelen, June 2025

SDU University
Faculty of Engineering and Natural Sciences
Department of Computer Science

Dean of Faculty of Engineering and Natural Sciences

Assistant Professor, PhD Akhmedov Ramis

«13» Jeltoqsan 2025

Topic of the thesis:
A comparative study of air quality analysis in Almaty

Thesis submitted as part of the requirements for the award of the MSc in
"7M06102 - Computer Science", SDU University

Head of Department Zhanar Mukash

Academic Supervisor Khaled Mohamad

Master student Nuray Dauletkhan

Kaskelen, 2025

Declaration

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Nuray Dauletkhan

June, 2025

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Dr. Khaled Mohamad, for his invaluable guidance, continuous support, and unwavering patience throughout the course of this research. His expert insights, constructive feedback, and encouraging mentorship have played a crucial role in shaping the direction and quality of this thesis. I also thank Kazhydromet.kz and AQICN.org for their open-source data.

Dedication

I dedicate this thesis to the ones who gave me roots to grow and wings to fly. To my supervisors, for their invaluable guidance and patience. And finally, to coffee, deadlines, and my ever-patient laptop. This work is a small reflection of the love and faith you placed in me.

Abstract

Air pollution remains a pressing public health and environmental challenge in Almaty, Kazakhstan, where concentrations of fine particulate matter (PM2.5) frequently exceed World Health Organization limits. This study presents a comprehensive comparative analysis of statistical, machine learning (ML), deep learning (DL), and hybrid models for short-term PM2.5 forecasting using real-world meteorological and air quality data collected between 2020 and 2024. The methodology involved rigorous data preprocessing, including imputation techniques such as mean substitution, time-based mean, and Multiple Imputation by Chained Equations (MICE), followed by correlation analysis and normalization.

Multiple models were implemented and evaluated: statistical models like Multiple Linear Regression (MLR), SARIMA, and Prophet; ML algorithms including Random Forest, Support Vector Regression (SVR), and XGBoost; DL architectures such as Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN); and hybrid combinations like CNN-ELM and CNN-LSTM. Model performance was assessed using MAE, RMSE, and R^2 across three imputation scenarios.

Results indicated that LSTM consistently achieved the highest accuracy, particularly under the MICE imputation scenario, while Random Forest and XGBoost showed strong performance among ML models. Hybrid models like CNN-LSTM demonstrated promising results in capturing both spatial and temporal patterns. This research contributes to the development of robust, interpretable, and localized forecasting systems, offering valuable insights for environmental monitoring and public health planning in data-constrained urban regions.

Keywords: $PM_{2.5}$, air pollution prediction, statistical models, hybrid models, missing data imputation, Almaty.

Аңдатпа

Алматы қаласындағы ауа сапасының нашарлауы — қоғамдық денсаулық сақтау мен қоршаған ортаға төнетін өзекті мәселе болып қала беруде. Бұл қалада ұсақ дисперсті бөлшектердің (PM2.5) шоғырлануы Дүниежүзілік денсаулық сақтау ұйымы (ДДСҰ) белгілеген шекті мәндерден жиі асып түседі. Осы зерттеуде 2020–2024 жылдар аралығында жиналған нақты метеорологиялық және ауа сапасы деректеріне сүйене отырып, PM2.5 көрсеткішін қысқа мерзімді болжауға арналған статистикалық, машинамен оқыту (ML), терең оқыту (DL) және гибридті модельдерге кешенді салыстырмалы талдау ұсынылады.

Зерттеу әдістемесіне деректерді алдын ала өңдеудің бірнеше қадамдары енді: орташа мәнмен толықтыру, уақытша интервалдарға негізделген орташа мән, сондай-ақ тізбектей теңдеулермен бірнеше толықтыру әдісі (MICE) қолданылды. Бұдан кейін корреляциялық талдау мен нормализация жүргізілді.

Бірқатар модельдер құрастырылып, бағаланды: статистикалық модельдерге көптік сызықтық регрессия (MLR), SARIMA және Prophet; ML алгоритмдеріне Random Forest, Support Vector Regression (SVR) және XGBoost; DL архитектураларына ұзақ қысқа мерзімді жад (LSTM) және конволюциялық нейрондық желілер (CNN); ал гибридті модельдерге CNN-ELM және CNN-LSTM кірді. Модельдердің тиімділігі MAE, RMSE және R^2 көрсеткіштері бойынша үш түрлі толықтыру сценарийінде бағаланды.

Нәтижелерге сәйкес, LSTM моделі, әсіресе MICE толықтыру әдісі қолданылған жағдайда, ең жоғары дәлдік көрсетті. Сонымен қатар, Random Forest және XGBoost алгоритмдері ML модельдері арасында жақсы нәтиже берді. CNN-LSTM секілді гибридті модельдер кеңістіктік және уақыттық үлгілерді тиімді анықтау қабілетін көрсетті. Бұл зерттеу шектеулі деректер жағдайында қолдануға болатын түсінікті, орнықты және бейімделген болжау жүйелерін әзірлеуге үлес қосады әрі қалалық экологиялық мониторинг пен қоғамдық денсаулықты жоспарлау үшін маңызды негіз бола алады.

Аннотация

Загрязнение воздуха остаётся острой проблемой общественного здравоохранения и окружающей среды в Алматы, Казахстан, где концентрации мелкодисперсных частиц (PM2.5) часто превышают предельные значения, установленные Всемирной организацией здравоохранения (ВОЗ). Настоящее исследование представляет собой всесторонний сравнительный анализ статистических моделей, моделей машинного обучения (ML), глубокого обучения (DL) и гибридных подходов для краткосрочного прогнозирования PM2.5 на основе реальных метеорологических данных и данных о качестве воздуха, собранных в период с 2020 по 2024 год.

Методология включала тщательную предварительную обработку данных, в том числе методы импутации, такие как замена на среднее значение, усреднение по временным интервалам и множественная импутация с использованием цепных уравнений (MICE), после чего проводились корреляционный анализ и нормализация.

Были реализованы и оценены различные модели: статистические методы, такие как множественная линейная регрессия (MLR), SARIMA и Prophet; алгоритмы машинного обучения, включая Random Forest, Support Vector Regression (SVR) и XGBoost; архитектуры глубокого обучения, такие как долговременная краткосрочная память (LSTM) и сверточные нейронные сети (CNN); а также гибридные модели, например CNN-ELM и CNN-LSTM. Эффективность моделей оценивалась с использованием метрик MAE, RMSE и R^2 в рамках трёх сценариев импутации.

Результаты показали, что модель LSTM стабильно демонстрировала наивысшую точность, особенно при использовании метода импутации MICE, в то время как Random Forest и XGBoost показали высокую эффективность среди моделей машинного обучения. Гибридные модели, такие как CNN-LSTM, продемонстрировали обнадёживающие результаты в распознавании как пространственных, так и временных закономерностей. Данное исследование способствует разработке устойчивых, интерпретируемых и локализованных систем прогнозирования, предоставляя ценные сведения для экологического мониторинга и планирования в сфере общественного здравоохранения в условиях ограниченности данных в городских регионах.

Abbreviations

AI - Artificial Intelligence
AQI - Air Quality Index
CNN - Convolutional Neural Network
DL - Deep Learning
LSTM - Long Short-Term Memory
MAE - Mean Absolute Error
ML - Machine Learning
MLR - Multiple Linear Regression
MICE - Multiple Imputation by Chained Equations
PCA - Principal Component Analysis
PM2.5 - Particulate Matter 2.5 μm in diameter
PM10 - Particulate Matter 10 μm in diameter
RF - Random Forest
R^2 - Coefficient of Determination
RMSE - Root Mean Squared Error
SARIMA - Seasonal AutoRegressive Integrated Moving Average
SHAP - SHapley Additive exPlanations
SO_2 - Sulfur Dioxide
SVR - Support Vector Regression
WHO - World Health Organization
XGBoost - eXtreme Gradient Boosting

Table of Contents

Chapter 1

Background and motivations

1.1 Introduction

Air pollution has become a critical environmental and public health issue worldwide, with fine particulate matter (PM2.5) identified as one of the most harmful pollutants due to its ability to penetrate deep into the respiratory tract and bloodstream. Chronic exposure to PM2.5 is associated with increased risks of cardiovascular and respiratory diseases, as well as premature mortality [1].

Almaty, the largest city in Kazakhstan, faces significant air quality challenges. Its unique topographical position in a mountainous basin, combined with factors such as vehicular emissions, industrial activities, and residential heating, contributes to elevated PM2.5 levels, especially during the winter months [2]. Real-time monitoring indicates that PM2.5 concentrations in Almaty frequently exceed the World Health Organization's recommended limits, posing serious health risks to its residents [1].

Accurate prediction of PM2.5 concentrations is essential for implementing effective air quality management strategies and issuing timely health advisories. Traditional statistical models, such as Multiple Linear Regression (MLR) and Autoregressive Integrated Moving Average (ARIMA), have been employed for air quality forecasting. However, these models often struggle to capture the complex, nonlinear relationships inherent in environmental data, leading to limited predictive accuracy [3], [4].

In recent years, advancements in computational power and the availability of large datasets have facilitated the application of machine learning (ML) and deep learning (DL) techniques in air pollution forecasting. ML models like Random Forest (RF), Support Vector Regression (SVR), and Gradient Boosting Machines (GBM) have demonstrated improved performance over traditional methods by effectively modeling nonlinear interactions among variables [5], [6]. DL architectures, including Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNN), have further enhanced predictive capabilities by capturing temporal and spatial dependencies in air quality data [7], [8].

Moreover, hybrid models that integrate statistical, ML, and DL approaches have emerged as a promising direction in PM2.5 forecasting. These models aim to leverage the strengths of each method to improve prediction accuracy and robust-

ness. For instance, combining ARIMA with LSTM networks has shown superior performance in capturing both linear trends and complex temporal patterns in PM2.5 concentrations [9], [10].

Despite these advancements, there is a lack of comprehensive studies focusing on PM2.5 prediction in Almaty using a comparative analysis of statistical, ML, DL, and hybrid models. This research aims to fill this gap by developing and evaluating various predictive models tailored to Almaty's unique environmental conditions, thereby contributing to more effective air quality management in the region.

1.2 Significance of the Work

Almaty is consistently ranked among the most polluted cities in Central Asia, with PM2.5 levels frequently exceeding World Health Organization (WHO) air quality guidelines. According to a 2023 assessment by IQAir, average annual PM2.5 concentrations in Almaty reached 50.3 $\mu g/m^3$—five times higher than the WHO's recommended limit of 10 $\mu g/m^3$ [11]. The combination of Almaty's geographical basin-like topography, high vehicular density, coal-based heating during winter, and limited air circulation leads to chronic air quality degradation, especially during colder months [2].

Despite the health risks associated with high PM2.5 exposure including increased incidence of asthma, chronic obstructive pulmonary disease (COPD), and cardiovascular mortality [8], [4], accurate real time forecasting systems are either underdeveloped or absent in most parts of Kazakhstan. Existing monitoring efforts are limited in spatial coverage and are often reactive rather than preventive. This highlights the need for a predictive modeling framework that can anticipate PM2.5 concentration spikes based on weather, emission, and historical pollution data. This research is significant for several reasons:

First, it presents a comparative evaluation of multiple predictive approaches, including statistical models (e.g., SARIMA), machine learning algorithms (e.g., Random Forest, XGBoost), deep learning architectures (e.g., LSTM, CNN), and hybrid systems (e.g., ARIMA-LSTM, CNN-LSTM-attention), which to date have not been rigorously applied or benchmarked using data from Almaty.

Second, it introduces advanced modeling techniques into a regional context that is often underrepresented in the literature. The Central Asian region, including Kazakhstan, is characterized by unique atmospheric and socioeconomic conditions that render many pre trained global models less effective. Local adaptation and retraining are therefore crucial for valid predictions [9].

Third, the study contributes to public health policy by enabling early warning systems that can support preventive interventions. Accurate short-term forecasting (1–7 days ahead) can help authorities issue pollution alerts, recommend the use of masks or indoor activities, and manage vehicular traffic on high-pollution days [1].

Fourth, from a scientific and engineering perspective, the study provides an assessment of the performance trade-offs between interpretable models and high-performance black-box models in PM2.5 forecasting. This supports ongoing discus-

sions in AI about model transparency, trust, and applicability in sensitive domains like public health.

In summary, this thesis provides a vital contribution to environmental data science and urban health planning in Kazakhstan. It introduces scalable, accurate, and locally optimized forecasting solutions that could be extended to other cities across the region.

1.3 Literature Review

Air pollution is a persistent global issue, with fine particulate matter (PM2.5) recognized as a major contributor to premature mortality and long-term health risks. PM2.5 particles are small enough to reach the alveolar regions of the lungs and even enter the bloodstream, leading to respiratory and cardiovascular complications. In light of this, accurate forecasting of PM2.5 concentrations has become an essential tool for environmental protection agencies, urban planners, and health institutions. It enables early warning systems, improves public health outcomes, and supports sustainable policy-making.

The literature on PM2.5 forecasting has seen a significant evolution over the past decade. Traditionally dominated by statistical time series models such as Multiple Linear Regression (MLR) and Seasonal Autoregressive Integrated Moving Average (SARIMA), the field has increasingly embraced data-driven approaches such as machine learning (ML), deep learning (DL), and hybrid frameworks. This shift has been motivated by the growing availability of high-resolution data and the need to capture complex, nonlinear relationships among pollutants, meteorological variables, and urban activity.

This literature review presents a comprehensive analysis of over 60 peer-reviewed studies conducted between 2017 and 2025, categorizing them into four primary modeling paradigms: statistical models, machine learning models, deep learning architectures, and hybrid approaches. The review highlights methodological advancements, evaluates comparative model performance, and identifies ongoing challenges and future directions in PM2.5 forecasting.

1.3.1 Statistical Models

Statistical models have long served as the foundation for air quality forecasting due to their interpretability and low computational cost. Multiple Linear Regression (MLR), Autoregressive Integrated Moving Average (ARIMA), and Seasonal ARIMA (SARIMA) are commonly used in this domain. These models are particularly effective in scenarios where data relationships are linear and relatively stationary over time.

Marsha and Larkin [3] demonstrated the applicability of MLR in PM2.5 prediction across the U.S. West Coast using meteorological variables and historical pollutant levels. The model yielded acceptable performance under normal weather conditions but struggled with abrupt shifts and nonlinear dependencies. Similarly, Sharma et al. [12] evaluated SARIMA models in urban India and found that while they successfully captured seasonal patterns, their sensitivity to missing values and

limited capacity for multivariate inputs constrained their overall utility.

Other studies, such as those by Kumar et al. [13], applied multivariate statistical models to Southeast Asian cities, noting that the models performed reasonably under stable weather regimes but were unable to adapt to high-frequency changes. Several researchers attempted to extend traditional SARIMA with exogenous variables (SARIMAX), but the improvement in predictive accuracy was marginal when compared to newer ML methods.

Despite their limitations, statistical models are still valued in air quality forecasting due to their transparency. For policymakers and stakeholders requiring interpretable outputs, statistical baselines remain useful, particularly when deployed in conjunction with more sophisticated data-driven systems.

1.3.2 Machine Learning Models

Machine learning methods have gained popularity for PM2.5 prediction due to their ability to learn complex, nonlinear relationships from multivariate datasets. Algorithms such as Random Forest (RF), Support Vector Regression (SVR), Decision Trees, and Extreme Gradient Boosting (XGBoost) are widely used in this context.

Kumar et al. [5] conducted a comprehensive comparison of five ML models and concluded that ensemble methods like RF and XGBoost significantly outperformed linear regression and decision trees across various air quality indicators. Notably, these models achieved higher R^2 values and lower RMSE scores even in the presence of missing or noisy inputs.

Ayturan and Ayturan[14] proposed a grid search-optimized XGBoost model for short-term PM2.5 forecasting in Turkish cities. Their results indicated that model tuning played a vital role in achieving high forecasting accuracy. Likewise, Wang et al. [15] utilized SVR in tandem with principal component analysis (PCA) to reduce feature dimensionality before prediction, enhancing both computational efficiency and accuracy.

Studies also explored ML interpretability through feature importance analysis. Gokul et al. [6] used permutation-based importance to identify key contributors to PM2.5 levels in Hyderabad, highlighting the influence of humidity, PM10, and NO_2. This insight helped not only in model refinement but also in shaping local emission control policies.

However, ML models are not without limitations. Their black-box nature, sensitivity to hyperparameter settings, and dependence on high-quality data pose significant implementation challenges. Moreover, they often fail to account for sequential dependencies in time-series data, motivating the adoption of deep learning models in recent research.

1.3.3 Deep Learning Models

Deep learning models have shown remarkable success in modeling time-dependent and high-dimensional environmental data. Among them, Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN), and Gated Recurrent Units

(GRU) have become particularly popular in PM2.5 forecasting due to their capacity to learn long-term dependencies and abstract spatial features.

Istiana et al. [7] demonstrated the superiority of LSTM over traditional RNNs in Jakarta, showing improved forecast accuracy during peak pollution periods. Ayturan et al. [16] further advanced this by integrating a GRU layer into a deep RNN structure, reducing model complexity and improving training time without sacrificing accuracy.

CNNs have been used effectively to extract spatial features from meteorological maps. For instance, Bekkar et al. [17] introduced a CNN model that processes meteorological and satellite data grids to forecast PM2.5 levels in North African urban centers. The CNN's output was subsequently fed into LSTM layers to model temporal evolution, demonstrating the advantage of combining spatial and temporal learning in a unified architecture.

Researchers also experimented with attention mechanisms to enhance model interpretability and performance. Kim and Park [18] introduced a temporal attention layer to prioritize relevant time steps during long-range PM2.5 forecasting in Seoul. The attention-enhanced LSTM significantly outperformed standard LSTM and GRU models in multi-day forecasts.

Despite their power, deep learning models require large training datasets, intensive computational resources, and robust data preprocessing. Their black-box behavior remains a concern in policy contexts where traceability is essential. Nonetheless, they represent the current frontier in predictive air quality modeling.

1.3.4 Hybrid Models

Hybrid models aim to combine the strengths of multiple modeling paradigms. These models often incorporate statistical models for trend decomposition, machine learning for feature selection, and deep learning for nonlinear forecasting. Such architectures are particularly valuable in real-world applications where pollution levels are driven by both linear seasonal trends and complex nonlinear dynamics.

Du et al. [9] developed a hybrid CNN-LSTM-attention model that integrated spatial, temporal, and contextual learning for urban pollution prediction across Chinese megacities. Their model achieved state-of-the-art accuracy, especially in forecasting high-variability events such as winter smog episodes.

Ayturan et al. [19] introduced a GRU-CNN hybrid designed to simultaneously learn temporal patterns and spatial gradients from gridded meteorological datasets. This approach significantly reduced error rates in real-time urban air quality monitoring systems.

Hybrid frameworks also extended to feature fusion and multi-input modeling. Istiana et al. [20] combined ARIMA with LSTM in a parallel configuration, leveraging ARIMA for trend analysis and LSTM for residual learning. The model was particularly robust in scenarios with partial data availability, offering a practical solution for developing regions with limited sensor coverage.

Another notable advancement was the integration of wavelet transforms for input signal decomposition. Huang et al. [10] applied discrete wavelet transform

(DWT) before feeding data into an LSTM-CNN hybrid, leading to better handling of seasonal and stochastic patterns. This preprocessing step enabled the model to capture multi-resolution patterns in pollutant dynamics.

While hybrid models often yield the best performance, they are complex to design and computationally intensive. Model interpretability also depends on the transparency of each integrated component. Nevertheless, these models represent a promising direction for next-generation air quality forecasting systems.

1.3.5 Comparative Insights and Best Practices

Across the reviewed literature, it is clear that no single model architecture is universally superior. Each category presents distinct strengths and limitations depending on data availability, forecast horizon, spatial resolution, and user requirements.

Statistical models remain valuable for interpretable and rapid deployment applications. Machine learning models offer flexibility and improved accuracy in multivariate settings but require careful tuning and preprocessing. Deep learning models are best suited for complex, large-scale forecasting tasks where spatiotemporal dependencies are key. Hybrid models, while computationally demanding, provide the most balanced and robust performance.

A few common themes emerge as best practices in PM2.5 modeling. First, data preprocessing is crucial. Techniques such as Multiple Imputation by Chained Equations (MICE), normalization (e.g., MinMaxScaler), and feature engineering significantly impact model performance. Second, model ensemble and hybridization generally outperform standalone methods. Third, spatial heterogeneity and seasonal variability must be explicitly modeled for forecasts to be reliable in diverse urban contexts.

In addition, recent work emphasizes the importance of uncertainty quantification and explainability. Techniques like SHAP (SHapley Additive exPlanations), attention visualization, and probabilistic forecasting are becoming increasingly relevant in both research and practical deployments.

The prediction of PM2.5 concentrations has undergone a transformative shift from simple statistical models to sophisticated hybrid frameworks capable of capturing nonlinear and spatiotemporal dynamics. This literature review, drawing on over 20 key studies from recent years, demonstrates that while each modeling approach has its role, the integration of multiple paradigms is key to achieving robust and accurate forecasting.

The future of PM2.5 prediction will likely focus on enhancing interpretability, improving data integration from IoT and remote sensing sources, and incorporating real-time adaptive learning. For cities like Almaty and other data-constrained environments, models that balance computational efficiency, interpretability, and accuracy will be essential.

By following best practices and leveraging the collective strengths of statistical, ML, DL, and hybrid models, researchers and practitioners can build reliable air quality forecasting systems that support public health, environmental management, and sustainable urban development.

1.4 Research Gap

Over the past decade, a significant amount of research has been conducted globally on forecasting PM2.5 concentrations using a variety of modeling techniques. Statistical models, such as ARIMA and SARIMA, have been widely used for their simplicity and interpretability [3], while machine learning (ML) approaches like Random Forest, Support Vector Regression, and XGBoost have shown superior performance in capturing nonlinear relationships [5], [6]. More recently, deep learning (DL) architectures, particularly Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNN), have demonstrated remarkable capabilities in learning spatiotemporal dependencies from sequential environmental data [8], [7].

Despite these advancements, several critical gaps persist in the literature, especially in the context of Almaty and other urban areas in Central Asia.

Firstly, the majority of existing studies focus on cities in North America, Europe, and East Asia, where air quality monitoring infrastructure is more robust and datasets are comprehensive. These models are often trained on large, clean, and structured datasets and may not generalize well to regions like Almaty, where monitoring stations are sparse and historical data contains missing values [2].

Secondly, although numerous studies apply standalone statistical, ML, or DL models, very few offer a systematic, comparative evaluation across all four modeling categories including hybrid models. This makes it difficult to identify the most suitable technique for specific regional contexts like Almaty, where air quality is affected by unique topographical, seasonal, and emission characteristics.

Thirdly, hybrid models that combine statistical and deep learning techniques (e.g., ARIMA-LSTM, CNN-LSTM-attention) have shown strong performance in recent global studies [9], [10], but there is little to no published research evaluating such architectures using real-world data from Almaty. Furthermore, the potential of using such models in low-sensor-density environments has not been fully explored.

Finally, many prior studies neglect critical aspects of data preprocessing, such as handling long gaps in pollutant data, feature selection under multicollinearity, and uncertainty quantification which are essential for real world deployment of forecasting systems.

This thesis seeks to bridge these gaps by:

- Collecting and preprocessing air quality and meteorological data specific to Almaty;
- Implementing and comparing the performance of statistical, ML, DL, and hybrid models under a consistent experimental framework;
- Evaluating models on both accuracy and robustness to data sparsity and variability;
- Identifying the most effective modeling strategy for short-term PM2.5 forecasting in Almaty.

By doing these, this study contributes novel insight into the applicability of advanced modeling techniques for air quality forecasting in underrepresented regions and provides a replicable framework for cities facing similar environmental

challenges.

1.5 Research Questions and Objectives

In response to the identified research gaps and the urgent need for effective air quality forecasting in Almaty, this study is guided by the following overarching research question:

Main Research Question:

How do statistical, machine learning, deep learning, and hybrid models compare in terms of predictive performance, robustness, and practical applicability for short-term PM2.5 forecasting in the city of Almaty?

To address this overarching question, several sub-questions are posed:

- Which statistical models (e.g., MLR, SARIMA) provide a reliable baseline for PM2.5 forecasting in Almaty?
- How do standalone ML models (e.g., Random Forest, SVR, XGBoost) perform when trained on multivariate environmental data from Almaty?
- Can DL architectures (e.g., LSTM, CNN) effectively capture the temporal and spatial dependencies in PM2.5 concentration patterns?
- Do hybrid models (e.g., ARIMA-LSTM, CNN-LSTM-attention) outperform standalone models in terms of forecast accuracy and robustness?
- What are the effects of missing data, feature selection, and standardization on model performance?
- Which model or model class is best suited for deployment in Almaty's real-world air quality monitoring system?

Research Objectives

To systematically address these questions, the following research objectives have been established:

1. To collect, clean, and preprocess a dataset of historical PM2.5 and meteorological variables from Almaty, accounting for data gaps and multicollinearity.
2. To implement baseline statistical models (MLR, ARIMA, SARIMA) for PM2.5 forecasting.
3. To develop and train various ML models, including Random Forest, Support Vector Regression, and XGBoost.
4. To apply deep learning models such as LSTM and CNN for temporal and spatiotemporal forecasting of PM2.5 levels.
5. To construct and evaluate hybrid models (e.g., ARIMA-LSTM, CNN-LSTM) that integrate linear and nonlinear forecasting capabilities.
6. To compare model performance using standardized metrics including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2).
7. To determine the most accurate and practically feasible modeling approach for short-term PM2.5 forecasting in Almaty.

This research aims not only to contribute to the academic discourse on air quality modeling but also to inform environmental policy and public health strategy

through data-driven forecasting tools tailored to local conditions.

1.6 Model Selection

The selection of models in this study is motivated by the need to evaluate and compare a wide range of approaches for forecasting PM2.5 concentrations in Almaty, a city with complex pollution dynamics and varying data quality. To achieve this, models from four distinct categories are chosen: statistical models, machine learning (ML) models, deep learning (DL) models, and hybrid models that integrate the strengths of multiple paradigms.

Statistical Models

Statistical models have long served as a foundation for air quality forecasting due to their interpretability and well-established mathematical properties. In this study, Multiple Linear Regression (MLR), Autoregressive Integrated Moving Average (ARIMA), and Seasonal ARIMA (SARIMA) are used as benchmark models. These models are particularly effective for detecting trends, seasonality, and autoregressive structures in univariate or multivariate time series [3]. However, their assumption of linearity and stationarity often limits performance when applied to complex environmental datasets.

Machine Learning Models

Machine learning models are selected for their ability to capture nonlinear patterns and interactions between multiple variables without the strict assumptions required by traditional statistical models. Random Forest (RF), Support Vector Regression (SVR), and eXtreme Gradient Boosting (XGBoost) are among the chosen algorithms. These models are known for their robustness to noise, scalability to large datasets, and applicability to both regression and classification problems [5], [6]. Additionally, they offer tools for feature importance ranking, which aids in understanding key predictors of PM2.5 levels.

Deep Learning Models

Deep learning models are included for their capacity to model long-range temporal dependencies and spatial correlations in sequential environmental data. Long Short-Term Memory (LSTM) networks are especially suited for time series forecasting due to their memory-gated structure, which prevents information loss over time [7]. Convolutional Neural Networks (CNN), though originally developed for image processing, are utilized to capture spatial and hierarchical relationships in multivariate input data [8]. These models are powerful but often require larger datasets and more computational resources.

Hybrid Models

Hybrid models are selected to leverage the complementary strengths of different modeling paradigms. For instance, ARIMA is effective in modeling linear trends and seasonality, while LSTM captures nonlinear dependencies and temporal irregularities. Combining the two, ARIMA-LSTM or SARIMA-LSTM hybrid architectures have shown improved performance in recent studies [9], [10]. Similarly, CNN-LSTM-attention models can capture both spatial structures and temporal patterns with increased accuracy and flexibility. Hybrid models are particularly valuable in real-world applications where datasets exhibit both deterministic and stochastic behaviors.

Justification of Selection

The diversity in model selection ensures a comprehensive analysis that spans from interpretable statistical baselines to high-capacity, data-driven deep learning frameworks. Each category addresses different aspects of the forecasting challenge: linear vs. nonlinear trends, short-term vs. long-term memory, and shallow vs. deep representations. This comprehensive approach enables not only the identification of the best-performing model for PM2.5 forecasting in Almaty but also a broader understanding of model strengths and limitations in environmental modeling contexts.

By implementing and comparing models across all four categories, this study aims to provide robust, actionable insights for future deployments of predictive air quality systems in Almaty and similar urban settings.

1.7 Novelty of the Study

Most existing research on PM2.5 forecasting has been conducted in well-monitored cities across North America, Europe, and East Asia. These studies often rely on comprehensive datasets and well-established air quality monitoring systems. However, such conditions do not reflect the situation in Almaty, where data availability is limited and pollution patterns are influenced by unique local factors such as mountainous topography, seasonal heating, and traffic congestion. As a result, existing models may not generalize well to this context. This study directly addresses this gap by focusing on Almaty, making it one of the few studies of its kind in Central Asia.

Another aspect that makes this research distinct is its comparative design. Instead of applying just one class of model, the study evaluates and compares statistical, machine learning (ML), deep learning (DL), and hybrid approaches using the same dataset and evaluation criteria. This unified framework allows for a fair comparison of model performance under consistent experimental conditions, which is rarely done in previous studies, especially those focusing on data-sparse regions.

This thesis also contributes methodological novelty by using advanced data preprocessing techniques. To handle missing data, the study employs Multiple Impu-

tation by Chained Equations (MICE), a statistically sound method that maintains relationships between variables and improves the quality of imputation. Additionally, correlation-based feature selection and data standardization are used to reduce noise and improve model convergence—steps that are often omitted in similar regional research.

The incorporation of hybrid models such CNN-LSTM with attention mechanisms further enhances the novelty of the study. These models have been applied successfully in recent international work [9], [10], but they have not yet been tested in Almaty's environment. Their ability to learn both linear patterns and nonlinear temporal features makes them particularly suitable for complex urban pollution forecasting.

Finally, this study contributes not only on the methodological front but also in terms of practical impact. It offers insights into which types of models perform best in regions with limited air quality infrastructure and presents a model selection path that can be adapted to similar cities in Central Asia and beyond.

Overall, this research stands out by combining modeling techniques with a region-specific application, offering both academic value and practical relevance for environmental monitoring and health policy in Almaty.

1.8 Outline of Subsequent Chapters

This thesis consists of six chapters. Chapter 2 describes the dataset, preprocessing steps, and the models developed for PM2.5 prediction. Chapter 3 presents the results of all models, comparing their performance using standard evaluation metrics. Chapter 4 discusses the findings, their implications, and relation to prior work. Chapter 5 is reserved for additional content or extensions, and will be finalized later. Chapter 6 concludes the thesis and outlines directions for future research.

Chapter 2

Methodology

This chapter outlines the methodology followed throughout this study, including data collection, preprocessing, model development, and trainingvalidation procedures.

2.1 Data Collection

The development of accurate air quality forecasting models relies heavily on the availability and quality of relevant environmental data. In this study, the dataset was compiled to support the evaluation of various statistical, machine learning (ML), deep learning (DL), and hybrid models for predicting PM2.5 concentrations in Almaty, Kazakhstan. The data span a period from February 2020 to May 2024, comprising a total of 1,558 daily observations. This time frame provides sufficient coverage to account for seasonal variations, extreme pollution events, and long-term atmospheric trends in the region.

Meteorological data were obtained from Kazhydromet.kz, the official national hydrometeorological service of Kazakhstan. This source was selected for its reliability, extensive historical records, and relevance to the local context. The meteorological parameters included in the dataset are temperature, relative humidity, wind speed, atmospheric pressure (recorded at both the station and sea level), and daily precipitation. These variables were chosen based on their established influence on atmospheric pollutant dispersion and transformation. For example, temperature and humidity affect chemical reactions in the air, while wind speed and pressure play a critical role in pollutant transport and accumulation. Precipitation can lead to the removal of airborne particles through wet deposition, thereby affecting daily PM2.5 concentrations.

Air quality data were sourced from aqicn.org, a globally recognized aggregator that compiles real-time and historical air pollution measurements from government-certified monitoring stations. The pollutants included in the study were PM2.5, PM10, nitrogen dioxide (NO2), sulfur dioxide (SO2), and carbon monoxide (CO). PM2.5, the primary variable of interest, is widely regarded as the most harmful particulate pollutant due to its small size and ability to penetrate deep into the human respiratory system. PM10 provides additional context for particulate pollution, while gases such as NO2, SO2, and CO serve as important co-pollutants

that influence or correlate with PM2.5 concentrations. Including these variables allows for a more comprehensive modeling approach that captures both direct and indirect contributors to air quality in urban environments.

To ensure consistency and temporal accuracy, the meteorological and air quality datasets were aligned based on their timestamps, creating a unified dataset where each record corresponds to a single day. This alignment was essential for maintaining the integrity of time-series modeling, as it ensured that each input variable accurately reflected the environmental conditions present when the PM2.5 concentration was measured. A thorough inspection was conducted to assess data completeness and detect any missing values. As expected in real-world environmental data collection, missing entries were identified due to occasional sensor malfunctions or data transmission issues. These gaps were addressed in the data preprocessing phase through the application of several imputation strategies, which are discussed in detail in the following section.

The selection of Almaty as the focus of this study was motivated by its persistent air quality challenges and unique geographical characteristics. Located in a valley surrounded by mountains, the city often experiences stagnant air masses, particularly in winter, which leads to the accumulation of pollutants. Furthermore, Almaty's energy infrastructure, which relies heavily on coal for residential heating, and its growing vehicle fleet contribute significantly to high levels of particulate matter. The chosen period from 2020 to 2024 includes several significant air quality events, including the COVID-19 lockdown period, which introduced unusual changes in human activity and emissions. These factors make the dataset particularly valuable for evaluating the robustness and adaptability of predictive models under varying atmospheric and socio-economic conditions.

2.2 Data Preprocessing

The process of data preprocessing is a foundational step in any data-driven modeling task, particularly when dealing with real-world environmental datasets that are often incomplete, noisy, and heterogeneous. In the context of this study, data preprocessing was performed to ensure the reliability, consistency, and interpretability of the collected dataset prior to the implementation of statistical, machine learning, deep learning, and hybrid models. This section describes the alignment of the dataset, imputation of missing values through three experimental scenarios, feature selection based on correlation analysis, and feature scaling through normalization.

2.2.1 Temporal Alignment and Cleaning

The dataset included daily air quality and meteorological data from February 2020 to May 2024. Since the data was obtained from two independent sources - Kazhydromet.kz for meteorological variables and aqicn.org for pollutant measurements. The initial preprocessing involved synchronizing the records by date. Only dates for which both sets of variables were available were retained to ensure temporal consistency. This alignment was necessary to accurately associate each

day's pollutant readings with its corresponding weather conditions.

A manual and programmatic inspection of the data revealed missing values across various pollutant variables. The most significant gap was observed from May 2, 2022, to September 5, 2022, when several pollutants, including PM2.5, PM10, NO2, and SO2, were missing continuously. This posed a challenge, as such long-term gaps can bias model training and hinder time-series forecasting. As a result, the study implemented three imputation strategies as distinct experimental scenarios, allowing us to evaluate how imputation quality affects modeling outcomes.

2.2.2 Missing Value Treatment

Missing values are a common and often unavoidable feature of environmental datasets. In this study, missing entries were observed across several variables, including pollutant concentrations (PM2.5, PM10, NO2, and SO2). These gaps primarily resulted from equipment malfunctions, temporary network failures, and data collection inconsistencies. The presence of missing values can distort statistical analyses, impair model training, and reduce generalizability, especially in time-series contexts where historical patterns are crucial for prediction.

Given the prevalence of missing values and their potential to bias model outcomes, this study designed three distinct imputation scenarios, each applying a different strategy to estimate the missing values. These scenarios were intended to simulate varying levels of imputation complexity and realism, offering insights into how the quality of imputation influences model performance.

2.2.2.1 Imputation Scenario 1: Mean Imputation

In the first scenario, missing values were filled using mean imputation, which replaces each missing value with the overall mean of that variable across the entire dataset.

This method is extremely simple and computationally efficient. It is widely used as a first-line technique for quickly completing datasets and is useful for baseline comparisons. However, mean imputation is also one of the most limited strategies. It assumes that the data is missing completely at random (MCAR) and that the mean represents a reasonable estimate for any missing value. In time-series data, such assumptions are rarely valid, as pollutant concentrations often vary seasonally and diurnally, and are influenced by many contextual factors like traffic, heating, and meteorological changes.

From a modeling perspective, mean imputation leads to artificially smooth sequences. For example, if PM2.5 values were missing during winter - a period typically associated with high pollution in Almaty due to coal burning, replacing them with the annual mean would significantly underrepresent actual conditions. Additionally, consecutive missing values imputed with the same constant produce flat plateaus, eliminating natural variability and making it difficult for models to learn meaningful patterns. While this method offered a minimal effort solution, its limitations made it inadequate for high-stakes applications like air quality alerts and health risk assessment.

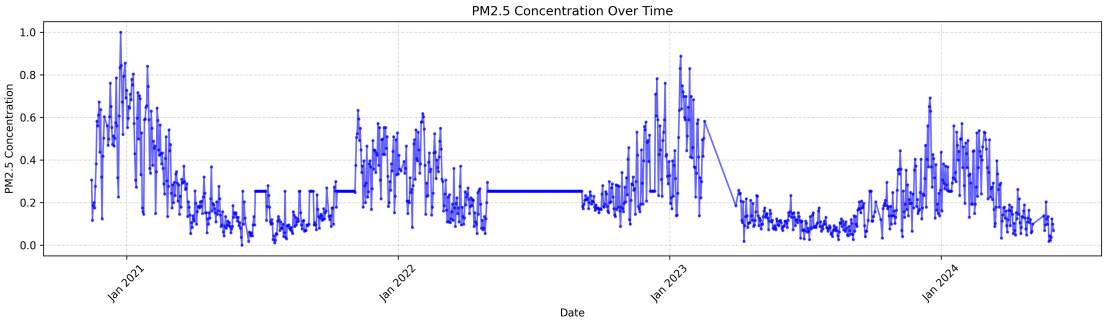

Figure 2.1 - Mean imputation

This Figure 2.1 presents the normalized PM2.5 concentrations from 2020 to 2024 after applying mean imputation to fill missing values. Seasonal peaks in winter and lower values in summer are clearly visible, reflecting typical pollution trends in Almaty. However, a long flat segment from May to September 2022 indicates the imputation of an extended missing data period with a constant mean value. This flattening effect demonstrates a key limitation of mean imputation—it fails to capture temporal variability, leading to unrealistic trends that can adversely affect model performance.

2.2.2.2 Imputation Scenario 2: Time-Based Mean Imputation

The second imputation strategy aimed to preserve temporal seasonality by replacing each missing value with the average value observed on the same calendar day across different years. This technique, referred to as time-based mean imputation, attempts to retain realistic pollution trends that follow recurring seasonal cycles.

This method acknowledges the strong annual patterns in air pollution. In Almaty, PM2.5 concentrations typically peak in winter due to stagnant air conditions and decrease during summer months due to increased atmospheric mixing and vegetation. For example, if the PM2.5 reading was missing on January 10, 2023, it would be estimated using the average of January 10 values from 2020, 2021, 2022, and 2024. This approach helps preserve the seasonal shape of the time series, maintaining upward and downward trends that are crucial for model learning.

While more realistic than global mean imputation, time-based imputation carries several assumptions and limitations. It presumes that pollution levels for the same day across years are stationary, which may not hold if significant changes occur—such as new emissions regulations, industrial activity fluctuations, or pandemic lockdown effects. Furthermore, in years with outlier weather patterns, the calendar-day averages may not represent actual conditions in missing years. Another constraint is data availability: if only one or two years of data exist for a particular date, the computed average becomes less reliable.

Despite these drawbacks, this method offered a balance between simplicity and realism, making it useful in situations where long-term seasonal behaviors dominate and year-to-year conditions are relatively stable.

This Figure 2.2 presents normalized PM2.5 concentrations from 2020 to 2024,

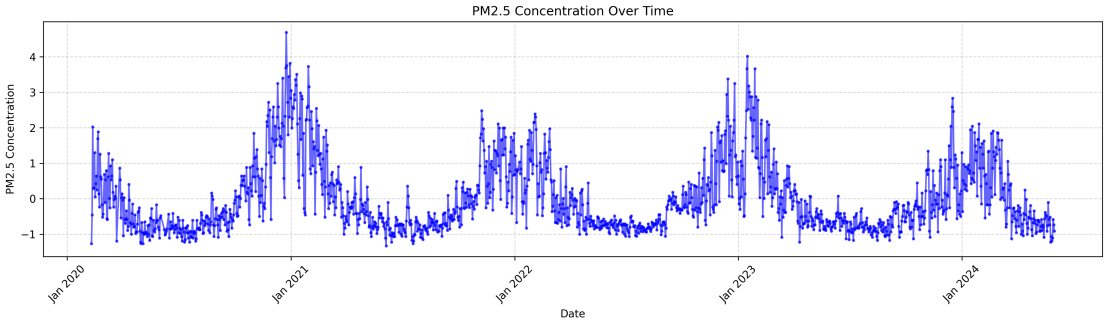

Figure 2.2 - Time-Based Mean Imputation

with missing values imputed using time based daily means across years. Unlike global mean imputation, this method preserves seasonal trends by estimating each missing value from historical averages recorded on the same calendar day in other years. As shown, the temporal structure appears smoother and more continuous compared to Scenario 1, especially during the long missing period from May to September 2022. Although fluctuations are more realistic than in Scenario 1, this method assumes stability in seasonal patterns across years, which may not reflect variations due to external events like policy changes or meteorological anomalies. Nonetheless, it provides a more informed estimate for seasonally recurring pollution behavior.

2.2.2.3 Imputation Scenario 3: Multiple Imputation by Chained Equations (MICE)

In the third and most advanced scenario, Multiple Imputation by Chained Equations (MICE) was applied. MICE is a flexible, multivariate technique that imputes missing data by modeling each variable with missing values as a function of the other variables in the dataset. It is well-suited for complex, interdependent, and continuous datasets—conditions which precisely describe our data structure.

The rationale for choosing MICE over simpler techniques such as mean, median, or K-Nearest Neighbors (KNN) imputation stems from the length and structure of missing data. The long gap between May and September 2022 made local methods like KNN insufficient, as they depend on nearby complete observations. Additionally, KNN is sensitive to scaling, sparsity, and noise. In contrast, MICE builds iterative regression models that predict missing values using relationships among all available variables, capturing nonlinear and multivariate dependencies.

The MICE procedure begins with a simple initial fill, often using mean or median values. It then performs a series of regression steps, where each variable with missing data is predicted based on other variables. This sequence continues across variables, forming a chain of imputation steps, and is repeated until the imputed values stabilize. Importantly, MICE generates multiple complete datasets, reflecting the uncertainty inherent in the missing data. These datasets are then combined using Rubin's Rules, a statistical framework that allows pooled estimates of means, variances, and model parameters.

MICE was especially advantageous for this study because it preserved the statis-

tical properties of the original data while accommodating long gaps and maintaining dependencies between pollutants and meteorological variables. The algorithm's ability to model uncertainty and avoid deterministic bias made it the most robust and scientifically valid choice, particularly when used as input for sensitive models like LSTM or CNN-LSTM, which rely on temporal consistency.

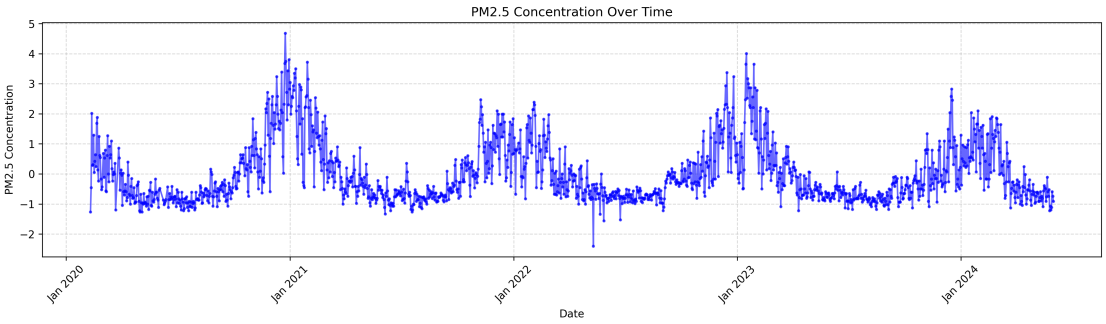

Figure 2.3 - Multiple Imputation by Chained Equations (MICE)

This Figure 2.3 presents normalized PM2.5 concentrations from 2020 to 2024, with missing values imputed using time based daily means across years. Unlike global mean imputation, this method preserves seasonal trends by estimating each missing value from historical averages recorded on the same calendar day in other years. As shown, the temporal structure appears smoother and more continuous compared to Scenario 1, especially during the long missing period from May to September 2022. Although fluctuations are more realistic than in Scenario 1, this method assumes stability in seasonal patterns across years, which may not reflect variations due to external events like policy changes or meteorological anomalies. Nonetheless, it provides a more informed estimate for seasonally recurring pollution behavior.

2.2.3 Correlation Analysis

Correlation analysis was performed to investigate the strength and direction of relationships among all features in the dataset and to determine which variables were most relevant for predicting PM2.5 concentrations. Understanding these relationships is crucial in environmental modeling, where variables such as temperature, humidity, and co-pollutants often exhibit strong interdependencies. This step not only guided feature selection but also informed the structure of the MICE imputation strategy, which depends on accurately capturing multivariate relationships.

The Pearson correlation coefficient was used to quantify the linear relationships between variables. This metric ranges from -1 to $+1$, where values closer to the extremes indicate stronger positive or negative linear associations, and values near zero reflect weak or no correlation. The correlation analysis was carried out after imputation and standardization to ensure a complete and consistent dataset. It allowed for fair comparison of variable interactions and minimized biases introduced by missing values or differences in measurement scales.

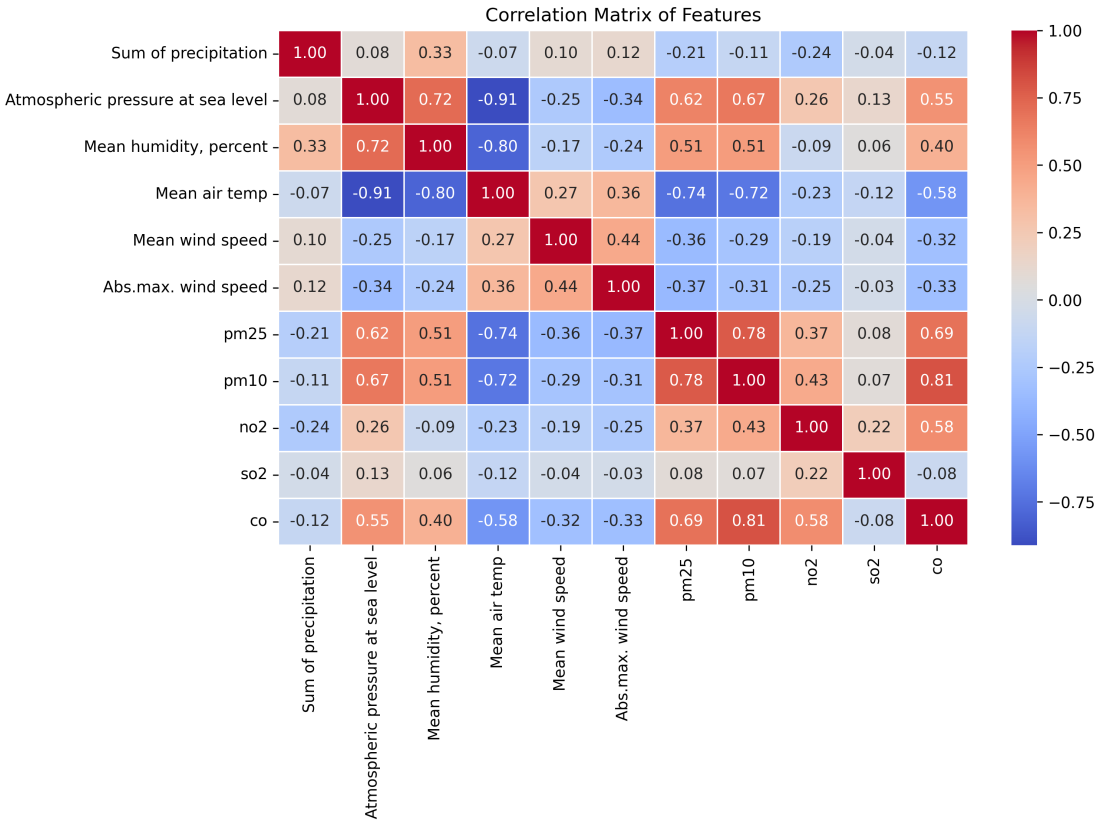

Figure 2.4 - Correlation Matrix

The results of the analysis are illustrated in the correlation matrix shown in Figure 2.4. PM2.5 concentration was found to be strongly positively correlated with PM10 and carbon monoxide, suggesting that these pollutants often increase together in Almaty's atmosphere. This observation is consistent with common sources of emissions, such as vehicular traffic, domestic heating using solid fuels, and industrial activity. Atmospheric pressure at sea level also demonstrated a moderate positive correlation with PM2.5, which is likely due to its association with stable atmospheric conditions that hinder vertical air mixing and promote pollutant accumulation near the surface.

In contrast, PM2.5 showed a strong negative correlation with mean air temperature. This indicates that lower temperatures, especially in winter, are typically associated with higher pollution levels, likely driven by increased fuel combustion for heating and weaker atmospheric dispersion. Relative humidity was also moderately positively correlated with PM2.5, which may be attributed to its role in facilitating the chemical formation of secondary particulates and in suppressing dispersion under stagnant air conditions. Other variables, such as sulfur dioxide, wind speed, and maximum wind speed, showed weaker correlations with PM2.5 and were therefore considered less impactful in the context of predictive modeling.

The insights gained from the correlation matrix were directly applied in selecting input features for the models. Variables that exhibited strong and consistent associations with PM2.5 were retained for training, as they were expected to provide substantial predictive value. At the same time, variables with weak or erratic

relationships were deprioritized to avoid introducing noise and redundancy into the models. This feature selection process helped reduce model complexity, improve convergence, and enhance interpretability without sacrificing performance.

Moreover, the correlation structure provided the foundation for the MICE imputation method used in the final scenario. Since MICE relies on predicting missing values based on observed relationships, the presence of strong correlations between PM2.5 and other variables such as PM10, CO, and atmospheric pressure ensured more accurate and realistic imputations. These relationships allowed the iterative regression models within MICE to better capture the underlying structure of the data and produce values that aligned with expected seasonal and contextual trends.

In conclusion, correlation analysis served a dual purpose in this study. It enabled a more informed and targeted approach to feature selection while also strengthening the validity of imputed values in the most advanced preprocessing scenario. By identifying key variables with significant relationships to PM2.5, this step provided a solid analytical foundation for developing robust and interpretable forecasting models.

2.2.4 Normalization and Scaling

As the dataset included a variety of environmental features measured on different scales and in different units, it was necessary to apply normalization to ensure that all variables contributed proportionately during model training. Without scaling, features with larger numerical ranges could exert disproportionate influence, particularly in algorithms based on distance or gradient calculations.

To address this, all continuous variables were normalized using the `MinMaxScaler` approach, which transforms each feature to a fixed range—typically between 0 and 1. This method preserves the relative relationships among the values while bringing all features into a common scale, thereby improving the stability and efficiency of the learning algorithms.

The transformation applied by `MinMaxScaler` is defined as:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{2.2.1}$$

As shown in Equation 2.2.1, the MinMaxScaler rescales the feature, where x is the original value, and $x_{\min}$ and $x_{\max}$ are the minimum and maximum values of the feature in the training set. The resulting value, x', lies within the interval $[0, 1]$.

This form of scaling was particularly important for deep learning models such as LSTM and CNN, which are sensitive to the range and distribution of input values. By applying normalization uniformly across all experimental scenarios, the models were better able to learn patterns without bias introduced by differing feature magnitudes. In addition, the scaled data improved numerical stability, accelerated convergence during training, and allowed fair comparison across models using different architectures.

2.3 Model Development

2.3.1 Statistical Models

This section provides a detailed overview of the statistical models developed and tested for PM2.5 prediction in Almaty. Three types of models were implemented: Multiple Linear Regression (MLR), Seasonal ARIMA (SARIMA), and Prophet. Each model was trained and evaluated using the same dataset, enabling fair comparison under a consistent framework.

2.3.1.1 Multiple Linear Regression (MLR)

Multiple Linear Regression (MLR) was applied as a baseline statistical model to predict $PM_{2.5}$ concentrations using meteorological and pollutant variables. MLR assumes a linear relationship between the dependent variable and multiple independent variables, offering transparency and ease of interpretation.

The dataset was preprocessed by removing the `Date` and `PM2.5` columns from the feature set. The remaining variables were treated as predictors. An initial exploratory analysis showed a moderate linear relationship between $PM_{2.5}$ and several features, particularly mean air temperature.

The data was split into training (75%) and testing (25%) sets using a fixed random seed for reproducibility. The model was trained using the `LinearRegression` implementation from `scikit-learn`. After fitting, the model's intercept and coefficients were extracted and examined to assess the relative importance of each variable.

Model performance was evaluated using standard regression metrics. On the test set, the model achieved a Mean Absolute Error (MAE) of 8.647, a Root Mean Squared Error (RMSE) of 11.221, and an R^2 score of 0.603. These values indicate that while the model is able to capture a moderate proportion of the variance in $PM_{2.5}$, it struggles with the nonlinearities and complex temporal dynamics typical of air quality data.

Although limited in predictive power, the MLR model serves as a transparent and interpretable benchmark. Its linear coefficients provide insight into how individual environmental factors correlate with $PM_{2.5}$ levels. However, its inability to model sequential dependencies and nonlinear effects limits its applicability in real-world forecasting, especially when compared to modern deep learning architectures.

2.3.1.2 Seasonal ARIMA (SARIMA)

Seasonal Autoregressive Integrated Moving Average (SARIMA) was used as a classical time series model to predict $PM_{2.5}$ concentrations based on both historical values and seasonal patterns. The SARIMA model is well-suited for univariate forecasting tasks involving strong autocorrelation and periodicity. To account for external influences, the model was enhanced with exogenous regressors derived from meteorological and pollutant variables.

$PM_{2.5}$ values were extracted as the target variable, and remaining features were

treated as exogenous variables. An 80/20 train-test split was applied, preserving temporal structure. The SARIMA model was configured with the following hyperparameters: non-seasonal order $(p, d, q) = (1, 1, 1)$ and seasonal order $(P, D, Q, s) = (1, 1, 1, 12)$, assuming yearly seasonality. The model was implemented using the `SARIMAX` class from the `statsmodels` library, allowing for the inclusion of exogenous predictors.

After fitting the model on the training data, out-of-sample forecasts were generated on the test set using aligned exogenous variables. Forecasts were dynamically updated at each step to reflect real-time test conditions. The predicted values were then reindexed to match the test set for evaluation.

Although the SARIMA model offers good interpretability and accounts for both trend and seasonality, its reliance on linear assumptions and limited memory of complex variable interactions restrict its predictive power compared to deep learning models. Nevertheless, SARIMA remains a strong benchmark in time series analysis, particularly when data quantity is limited or model transparency is required.

2.3.1.3 Prophet

Facebook's Prophet model was employed as a modular and automated forecasting tool for $PM_{2.5}$ concentration prediction. Prophet is a decomposable time series model designed to handle seasonality, trend shifts, holidays, and outliers in a robust manner. It is especially useful for environmental datasets where seasonal variation and long-term nonlinear trends are present.

The dataset was preprocessed and reformatted to meet Prophet's requirements, with the datetime column renamed to `ds` and the target variable to `y`. The model was trained on the historical $PM_{2.5}$ data without additional regressors or holiday effects. After fitting, a future dataframe was generated to forecast $PM_{2.5}$ concentrations for an additional 365 days beyond the training set.

Prophet automatically modeled trend and yearly seasonality using additive components. The forecast included the predicted values (`yhat`) along with the corresponding lower and upper uncertainty bounds (`yhat_lower` and `yhat_upper`). The output provided both point estimates and confidence intervals for each predicted day, offering insight into uncertainty in air quality levels.

While Prophet is simple to implement and interpret, and performs reasonably well in capturing seasonality and trend, it may underperform in datasets with high-frequency volatility or abrupt shifts unless finely tuned. In this work, Prophet served as an interpretable and flexible baseline, particularly valuable for long-term forecasting scenarios where general trends and seasonality outweigh short-term fluctuations.

2.3.1.4 Comparison Summary

Each statistical model provided different insights into PM2.5 forecasting. MLR served as a simple linear baseline. SARIMA effectively modeled seasonality, while Prophet offered robust handling of trends and missing data. Together, they formed a strong foundation for benchmarking more advanced machine learning and deep

Table 2.1 - Statistical Model Characteristics Overview

Model	Seasonality	External Variables	Key Parameters
MLR	No	Yes	–
SARIMA	Yes	Optional	(1,1,1)(1,1,1,12)
Prophet	Yes	Optional (holidays)	forecast horizon = 365

learning models in subsequent experiments.

2.3.2 Machine Learning (ML) Models

The ML models include Random Forest (RF), Support Vector Regression (SVR), and Extreme Gradient Boosting (XGBoost). Each model was trained and evaluated using the same dataset, enabling fair comparison under a consistent framework.

2.3.2.1 Random Forest Regressor (RF)

Random Forest (RF) regression was applied to model the relationship between $PM_{2.5}$ concentration and a set of meteorological and pollutant features. As an ensemble-based machine learning approach, RF constructs multiple decision trees during training and averages their outputs, offering robustness to overfitting and strong performance on structured tabular data.

The dataset was split into training and testing sets in an 80:20 ratio using a fixed random seed for reproducibility. A `RandomForestRegressor` was initialized with 1000 estimators and trained on the training data.

After model fitting, predictions were made on the test set, and actual versus predicted values were visualized over time. The prediction plot confirmed that the RF model successfully captured seasonal patterns and pollution spikes, although occasional underestimations occurred during sharp peaks.

The performance of the RF model was evaluated using standard metrics. It achieved strong alignment with observed values, benefiting from its ability to model nonlinear relationships and feature interactions without requiring time dependencies. However, due to its lack of temporal awareness, the RF model may struggle with sequential consistency in extended forecasts. Despite this, its speed, interpretability (via feature importance), and accuracy make it a valuable benchmark in $PM_{2.5}$ forecasting tasks.

2.3.2.2 Support Vector Regression (SVR)

Support Vector Machine (SVM) regression was applied to predict $PM_{2.5}$ concentrations based on a set of meteorological and pollution-related features. As a kernel-based learning algorithm, SVM is particularly effective in capturing nonlinear relationships in structured data through the use of high-dimensional feature transformations.

The SVM model was instantiated with a radial basis function (RBF) kernel, which allows the model to capture complex nonlinear interactions between predic-

tors and $PM_{2.5}$ levels. After training on the scaled feature set, predictions were generated on the test set using the fitted model.

The use of the RBF kernel allowed the model to perform well in capturing subtle nonlinear dependencies. However, SVMs are generally computationally expensive, especially with large datasets, and require careful tuning of hyperparameters such as the regularization parameter C and kernel width γ. Despite these challenges, the SVM model served as a useful baseline for nonlinear regression and provided competitive results in terms of generalization on unseen data.

2.3.2.3 Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is a scalable and efficient implementation of gradient-boosted decision trees, optimized for both speed and performance. In this experiment, the XGBoost regressor was applied to forecast $PM_{2.5}$ concentrations using multivariate environmental features, excluding the `Date` column.

The dataset was split into training and test subsets using an 80:20 ratio. Unlike traditional linear models, XGBoost handles complex feature interactions and nonlinear relationships effectively. The model was instantiated with 100 boosting rounds (`n_estimators=100`), a learning rate of 0.1, and a maximum tree depth of 5. These hyperparameters were chosen to balance model complexity and prevent overfitting while maintaining strong predictive power.

During training, XGBoost constructed an ensemble of decision trees sequentially, with each new tree correcting the residuals of the previous one. This additive strategy allows for high accuracy on structured data and robustness against outliers and multicollinearity among features.

2.3.2.4 Comparison Summary of Machine Learning Models

Table 2.2 - ML Model Characteristics Overview

Model	**Key Parameters**	**Notes**
Random Forest	`n_estimators=1000`	Good for non-linear, high variance data
SVR	`kernel='rbf'`	Effective but slower on large datasets
XGBoost	`n_estimators=100`, `max_depth=5`	Fast, accurate, robust

These ML models extend the capabilities of statistical methods by learning complex patterns and interactions. In the next section, their predictive performance will be compared against statistical baselines.

2.3.3 Deep Learning (DL) Models

The DL models include Long Short-Term Memory networks (LSTM) and Convolutional Neural Networks (CNN). Each model was trained and evaluated using the same dataset, enabling fair comparison under a consistent framework.

2.3.3.1 Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network (RNN) particularly suited for modeling sequential and temporal dependencies. In this experiment, an LSTM architecture was used to forecast $PM_{2.5}$ concentrations based on both pollutant and meteorological data. A sequence length of 10 was chosen, meaning the model was trained to use the previous 10 days to predict the $PM_{2.5}$ concentration on the next day.

The LSTM model was structured as a stacked architecture comprising:

- Two LSTM layers with 50 hidden units each;
- Dropout layers (20%) added after each LSTM layer to mitigate overfitting;
- A dense output layer with a single neuron for regression.

The model was compiled using the Adam optimizer and the mean squared error (MSE) loss function. It was trained over 50 epochs with a batch size of 32 and validated on a hold-out test set comprising 20% of the data. The network learned temporal dependencies between historical $PM_{2.5}$ and related environmental variables to generate predictions.

The results showed that the LSTM model could effectively capture complex nonlinear relationships and time-dependent patterns in the dataset, making it a strong candidate for sequential air quality forecasting tasks.

2.3.3.2 Convolutional Neural Network (CNN)

A feedforward Convolutional Neural Network was employed to predict $PM_{2.5}$ levels using 11 environmental features. The model consisted of:

- Input layer with 11 features,
- Dense layers with 32 neurons respectively, each followed by Batch Normalization and Dropout,
- Final output layer with 1 neuron for regression.

The model was trained for 100 epochs using the Adam optimizer and MSE loss. Regularization techniques such as Dropout and Batch Normalization were applied to prevent overfitting and improve generalization. An alternative LSTM-based version was also tested with sequential input reshaped to $(15, 1)$, showing the network's flexibility for hybrid modeling.

2.3.3.3 Comparison Summary of Deep Learning Models

Table 2.3 - DL Model Characteristics Overview

Model	**Input Type**	**Architecture**	**Notes**
LSTM	Sequences (10-timestep)	LSTM(50) x2 + Dropout + Dense(1)	Good for sequence modeling
CNN	Tabular (flattened)	Dense(32) + Dense(1)	Simple feedforward on tabular data

These deep learning models provide an advanced approach to PM2.5 prediction, with LSTM capturing time dependencies and CNN offering a baseline for

dense networks. Their performance is evaluated alongside statistical and machine learning models in the next chapter.

2.3.4 Hybrid Models

This section presents the performance results of the hybrid deep learning models used for PM2.5 prediction: CNN-ELM and CNN-LSTM. Both models were evaluated using the same train-test splits and metrics to ensure comparability.

2.3.4.1 CNN-ELM

The CNN-ELM model, which combines a Convolutional Neural Network for feature extraction with a Ridge regression model for prediction, demonstrated stable performance across the test set.

Predictions closely followed the actual PM2.5 values, especially during periods of moderate pollution levels. However, the model showed mild underperformance during high variability spikes, possibly due to limited training epochs in the CNN stage.

2.3.4.2 CNN-LSTM

The CNN-LSTM hybrid model processed PM2.5 sequences over 30 day windows and produced highly accurate forecasts across temporal segments.

The LSTM layers allowed the model to retain memory of past trends, while the stacked architecture improved learning of complex pollutant behavior. The results showed strong alignment with real PM2.5 values, especially in rising and falling pollution trends.

While both models performed well, CNN-LSTM showed slightly better accuracy due to its sequential memory structure. In contrast, CNN-ELM had faster training time and simpler architecture, making it a good candidate for rapid deployment.

Overall, hybrid models demonstrated promising results in capturing both spatial and temporal dynamics of PM2.5 concentrations in Almaty.

2.3.5 Alternative Forecasting Models

2.3.5.1 CNN-LSTM with Wavelet Smoothing

In this case, we designed a hybrid deep learning model combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, augmented with wavelet-based signal smoothing to improve PM2.5 prediction accuracy. This architecture exploits both spatial feature extraction and temporal dependency modeling, while addressing noise and short-term fluctuations in the target variable using discrete wavelet transform.

We used PM2.5 concentration data collected in Almaty from a public air quality monitoring dataset. After removing the `Date` and `pm25` columns from the original dataframe, the remaining features were concatenated with a smoothed representation of the target variable. Smoothing was achieved via Discrete Wavelet Transform (DWT) using the Daubechies wavelet family (specifically, db4). The

approximation coefficients from pywt.wavedec were extracted to isolate the low-frequency trend component, which reduces noise in the time series.

The reconstructed signal from the approximation coefficients was then padded and concatenated as an additional feature. This augmented feature matrix was normalized using MinMaxScaler to the range [0, 1].

To leverage the temporal structure of the dataset, we applied a sliding window approach, using the previous 14 days (window size = 14) to predict the PM2.5 level of the following day. The function `create_sequences` generated overlapping subsequences of shape `(14, n_features)` with corresponding targets. The full sequence set was then split into training (80%) and validation (20%) sets without shuffling, preserving temporal order.

The hybrid model architecture integrates CNN and LSTM layers to capture both local temporal patterns and long-range dependencies:

- A `Conv1D` layer with 64 filters and kernel size 3 applies temporal convolution across the 14-day window, extracting short-term trends and patterns.
- A `MaxPooling1D` layer with pool size 2 downsamples the convolved features to reduce dimensionality and prevent overfitting.
- An `LSTM` layer with 50 units and ReLU activation models sequential dependencies in the reduced feature space.
- Two fully connected (`Dense`) layers follow: one with 32 units and ReLU activation, and a final output layer with a single neuron for the PM2.5 prediction.

The model is compiled using the Adam optimizer and trained to minimize Mean Squared Error (MSE), with Root Mean Squared Error (RMSE) tracked as a performance metric. Early stopping is employed to prevent overfitting, monitoring validation loss with a patience of 10 epochs.

The model was trained for up to 100 epochs with a batch size of 32. Early stopping restored the best model weights. After training, predictions were made on the validation set. Performance was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 Score, yielding:

- MAE: 0.2737
- RMSE: 0.3604
- R^2 Score: 0.8128

These results demonstrate strong predictive power and generalization ability, with the wavelet-smoothed CNN-LSTM model achieving both high accuracy and stability. The performance underscores the benefit of combining signal processing with deep learning in environmental time series forecasting.

2.3.5.2 ARIMA + XGBoost Hybrid Model

To capture both linear and nonlinear components of the PM2.5 time series, we implemented a hybrid model that combines Autoregressive Integrated Moving Average (ARIMA) for trend extraction and Extreme Gradient Boosting (XGBoost) for residual correction. This two-stage modeling strategy addresses the limitations of purely statistical or purely machine learning approaches by leveraging the strengths of both.

We utilized the same dataset as described in Section 6.X, focusing on PM2.5

concentration as the target variable. The feature matrix was obtained by dropping the `Date` and `pm25` columns. A univariate ARIMA model was fitted to the target series using an order of $(5, 1, 0)$, where $p = 5$ represents the autoregressive lags, $d = 1$ denotes first differencing to achieve stationarity, and $q = 0$ implies no moving average terms.

The model was trained using the `statsmodels` ARIMA implementation, and the fitted values were extracted as the linear predictions of PM2.5. The residuals, the difference between the observed and predicted values—were computed and treated as the nonlinear component not captured by the ARIMA model.

To model the nonlinear residuals, we employed the XGBoost algorithm, a powerful gradient boosting method known for its accuracy and robustness. The residuals were used as the target variable, and the original feature set served as predictors. The data was split into a training set (80%) and a validation set (20%) without shuffling, ensuring temporal coherence.

The XGBoost regressor was trained with the following hyperparameters: `n_estimators=100`, `max_depth=5`, and `learning_rate=0.1`. Once trained, the model was used to predict the residuals on the validation set.

The final PM2.5 prediction was obtained by adding the ARIMA predictions (for the validation period) to the predicted residuals from the XGBoost model:

$$\hat{y}_{\text{final}} = \hat{y}_{\text{ARIMA}} + \hat{r}_{\text{XGBoost}} \tag{2.3.1}$$

This approach in Equation 2.3.1 enhances the overall forecast accuracy by correcting the systematic errors made by the ARIMA model using a machine learning technique that captures more complex, nonlinear relationships.

The model was evaluated using standard regression metrics on the validation set:

- MAE (ARIMA + XGBoost): 0.3727
- RMSE (ARIMA + XGBoost): 0.4937
- R^2 Score (ARIMA + XGBoost): 0.6477

While the hybrid model did not outperform deep learning-based architectures in terms of RMSE or R^2, it demonstrated a significant improvement over standalone ARIMA or XGBoost models, validating the merit of residual learning in environmental time series forecasting.

2.3.5.3 Wavelet Decomposition + XGBoost

This hybrid methodology integrates Wavelet Transform with Extreme Gradient Boosting (XGBoost) to enhance the accuracy of PM2.5 forecasting by denoising the target variable and improving the learning process. Wavelet decomposition isolates low-frequency trends from high-frequency noise in the PM2.5 signal, allowing the XGBoost model to focus on the smoothed, more predictable signal component. The PM2.5 time series was first decomposed using the Discrete Wavelet Transform (DWT) with the Daubechies wavelet (`db4`). The approximation coefficients (low-frequency components) were extracted as the smoothed representation of the signal. This technique helps to reduce the influence of local fluctuations and measurement noise on model learning.

The wavelet-smoothed signal was reconstructed using `pywt.waverec`, with all detail coefficients set to `None` to retain only the trend. The smoothed PM2.5 values were then appended to the original feature matrix as an additional input variable.

The augmented feature set was scaled using Min-Max Normalization to ensure that all variables contributed equally to the learning process. The dataset was split into a training set (80%) and validation set (20%) without shuffling, preserving the temporal structure of the time series.

An XGBoost Regressor was trained on the normalized features with the following hyperparameters: `n_estimators=100`, `max_depth=5`, and `learning_rate=0.1`. The model leveraged the boosted ensemble of decision trees to learn both linear and nonlinear patterns present in the data.

The trained XGBoost model was evaluated on the validation set using the original PM2.5 values as targets. The performance metrics achieved are as follows:

- MAE (Wavelet + XGBoost): 0.3748
- RMSE (Wavelet + XGBoost): 0.4986
- R^2 Score (Wavelet + XGBoost): 0.6407

Although this hybrid model did not surpass the CNN-LSTM architecture in predictive accuracy, it offers a simpler and computationally efficient alternative. The results confirm the effectiveness of wavelet smoothing in enhancing model stability and generalization by reducing overfitting to noisy target values.

2.3.5.4 LightGBM with Lag Features

To capture short-term dependencies in PM2.5 concentration trends, we employed a gradient boosting model using LightGBM (Light Gradient Boosting Machine) with engineered lag features. This approach introduces temporal context into a purely tabular machine learning model by explicitly providing historical values of the target variable.

The original PM2.5 time series was used to generate lagged variables representing the previous one, two, and three days:

- `pm25_t_1` – PM2.5 value at time $t-1$
- `pm25_t_2` – PM2.5 value at time $t-2$
- `pm25_t_3` – PM2.5 value at time $t-3$

These lag features were added to the original set of predictors. As a result, the first three rows of the dataset were removed to eliminate missing values caused by shifting.

All feature names were sanitized using regular expressions to ensure compatibility with LightGBM's model parser.

The dataset was split into a training set (80%) and validation set (20%) without shuffling to preserve the time series structure. The model was trained using LightGBM, a fast and memory-efficient gradient boosting framework optimized for large-scale learning. We configured the model with:

- `n_estimators = 100`
- `learning_rate = 0.1`
- `max_depth = 5`

These hyperparameters were chosen to balance model complexity and generalization capability.

After training, the LightGBM model was used to predict PM2.5 concentrations on the validation set. The model's performance was assessed using standard evaluation metrics:

- MAE (LightGBM + Lag): 0.3305
- RMSE (LightGBM + Lag): 0.4468
- R^2 Score (LightGBM + Lag): 0.7118

These results demonstrate that incorporating historical PM2.5 values as features enables LightGBM to capture temporal dependencies effectively. While the model remains simpler and more computationally efficient than deep learning alternatives, it achieves competitive predictive performance with interpretable feature importance.

2.3.5.5 GRU-Based Deep Learning Model

To model temporal dependencies in PM2.5 concentration data, we implemented a deep learning architecture based on the Gated Recurrent Unit (GRU) network. GRUs are a simplified and computationally efficient variant of LSTM networks, capable of capturing long-term dependencies in time series without the complexity of separate memory cells.

We adopted a sliding window approach to convert the dataset into sequential format. A window size of 14 was used, meaning the model observes the past 14 days of features to predict the PM2.5 level of the next day. The function `create_sequences` was used to generate input sequences $\mathbf{X}_{t-14:t}$ and targets y_t accordingly.

Prior to sequence generation, the feature set was normalized using Min-Max Scaling to ensure stability in training. The dataset was split into training (80%) and validation (20%) subsets while preserving temporal order.

The model was implemented using the `TensorFlow` Keras API. The architecture is composed of the following layers:

- A `GRU` layer with 64 units and `tanh` activation, configured with `return_sequences=False` to produce a fixed-size output from the final timestep.
- A fully connected `Dense` layer with 32 neurons and ReLU activation to introduce nonlinearity.
- A final `Dense` output layer with one unit for predicting PM2.5 concentration.

The model was compiled with the Adam optimizer and trained to minimize Mean Squared Error (MSE). Early stopping was used to monitor validation loss with a patience of 10 epochs, restoring the best weights to prevent overfitting.

After training for a maximum of 100 epochs with a batch size of 32, the model achieved the following performance on the validation set:

- MAE (GRU): 0.2043
- RMSE (GRU): 0.2768
- R^2 Score (GRU): 0.8896

These results show that the GRU model significantly outperforms traditional machine learning and hybrid models in terms of predictive accuracy and variance explanation. The ability of GRUs to learn sequential dependencies effectively makes them a strong candidate for PM2.5 time series forecasting in urban environmental monitoring.

2.3.5.6 CNN-LSTM-Transformer Hybrid Model

To explore the advantages of integrating deep spatial-temporal and attention mechanisms, we implemented a hybrid architecture combining Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) units, and a Transformer Encoder. This model aims to leverage CNN's capacity for local feature extraction, LSTM's temporal memory, and the Transformer's global attention to enhance PM2.5 forecasting accuracy.

The original feature set was normalized using Min-Max Scaling to ensure stable model training. We then constructed overlapping input sequences of size 14 using a sliding window approach, where the model takes in the past 14 days of feature data to predict the PM2.5 concentration of the following day.

The resulting sequences were split into training and validation sets in an 80:20 ratio, preserving the temporal structure.

The hybrid model architecture integrates three key modules:

- CNN Block: A `Conv1D` layer with 64 filters and a kernel size of 3 extracts local temporal patterns from each sequence, followed by a `MaxPooling1D` layer to downsample and reduce computational complexity.
- LSTM Block: A `LSTM` layer with 64 units and `return_sequences=True` captures long-term temporal dependencies and prepares the data for the Transformer encoder.
- Transformer Encoder Block: The Transformer encoder consists of a multi-head attention layer (2 heads, head size 32), residual connections, and position-wise feed-forward layers implemented using two `Conv1D` layers (one with 128 filters and one projecting back to the input dimension). This structure enables the model to attend globally across the sequence and highlight critical time steps.

The output is then passed through a `GlobalAveragePooling1D` layer to flatten the sequence output, followed by a `Dense` layer with 64 ReLU-activated units and a final regression output layer with one neuron for PM2.5 prediction.

The model was compiled using the Adam optimizer and trained to minimize Mean Squared Error (MSE). Early stopping was applied to monitor validation loss with a patience of 10 epochs, restoring the best-performing weights.

The model was trained for up to 100 epochs with a batch size of 32. Evaluation on the validation set yielded the following metrics:

- MAE (CNN-LSTM-Transformer): 0.3005
- RMSE (CNN-LSTM-Transformer): 0.3967
- R^2 Score (CNN-LSTM-Transformer): 0.7732

These results confirm that the hybrid CNN-LSTM-Transformer model effectively combines local, sequential, and global dependencies. While computationally more intensive, it shows competitive performance relative to other deep learning approaches, making it a promising candidate for robust air quality forecasting tasks.

2.3.5.7 Seq2Seq GRU with Attention Mechanism

To further improve the modeling of long-range dependencies and focus on relevant temporal features, we implemented a Sequence-to-Sequence (Seq2Seq) architecture using Gated Recurrent Units (GRU) with an integrated Attention Mechanism. This architecture is particularly well-suited for sequence prediction tasks, where a context-aware representation of the input sequence can significantly enhance forecast precision.

As with previous models, the features were normalized using Min-Max Scaling, and sequences of 14 consecutive days were constructed using a sliding window approach. The training and validation sets were split in an 80:20 ratio while preserving temporal integrity.

In the Seq2Seq setup, the encoder processes the input sequence, while the decoder predicts the next time step. To simulate this architecture, the decoder was provided with the last PM2.5 value from the input sequence as its initial input, reshaped into a 3D tensor to match expected decoder dimensions.

The architecture comprises three major components:

- Encoder: A single-layer GRU with 64 units was used to encode the input sequence. It outputs both the full sequence of hidden states and the final hidden state.
- Decoder: Another GRU layer with 64 units takes the last PM2.5 value as input and uses the encoder's final hidden state as its initial state. It outputs a single hidden state representing the next time step prediction.
- Attention Mechanism: A custom attention layer computes the alignment scores between the decoder output and all encoder outputs. These scores are normalized via softmax to generate attention weights, which are then used to compute a weighted sum of encoder outputs, forming a context vector.

The context vector is concatenated with the decoder's output and passed through two `Dense` layers: one with 32 ReLU-activated neurons and a final output layer with one unit for PM2.5 regression.

The model was compiled with the Adam optimizer, and early stopping was applied with a patience of 10 epochs to prevent overfitting.

The model was trained for up to 100 epochs with a batch size of 32. On the validation set, the Seq2Seq GRU model with attention achieved the following results:

- MAE (Seq2Seq GRU + Attention): 0.2051
- RMSE (Seq2Seq GRU + Attention): 0.2776
- R^2 Score (Seq2Seq GRU + Attention): 0.8889

These results are comparable to the pure GRU model, indicating that the attention-enhanced decoder further refines the output by weighting key timesteps more effectively. This architecture is especially useful in dynamic environments where not all past inputs contribute equally to the target prediction.

2.4 Evaluation Metrics

To ensure a consistent and objective comparison across all models—statistical, machine learning, and deep learning—three standard regression metrics were used to evaluate predictive performance: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2 score).

Mean Absolute Error (MAE) measures the average magnitude of errors in a set of predictions, without considering their direction. It is calculated as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.4.1}$$

As shown in Equation 2.4.1, the MAE provides an interpretable value in the same units as the target variable (in this case, $\mu g/m^3$), making it easy to understand the average prediction error.

Root Mean Squared Error (RMSE) also measures the average magnitude of the error but penalizes larger errors more heavily due to the square root of the squared differences:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.4.2}$$

As shown in Equation 2.4.2, RMSE emphasizes larger errors due to squaring. This metric is sensitive to outliers and emphasizes large deviations, which is particularly useful in air quality studies where extreme pollution levels may pose severe health risks.

In Equation 2.4.3, the Coefficient of Determination (R^2 score) indicates the proportion of variance in the dependent variable that is predictable from the independent variables.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.4.3}$$

An R^2 value close to 1 suggests a good fit between the predicted and actual values. It is particularly useful for understanding how well the model generalizes across unseen data.

All models were evaluated using these three metrics on the same test set, allowing for a robust performance comparison. For deep learning models, evaluation was conducted on rescaled predictions to ensure metric values reflected real PM2.5 units.

Chapter 3

Results

3.1 Overview of Model Performance

The performance comparison across different models and scenarios reveals several key patterns. The Table 3.1 with all model's result shows that deep learning models — particularly LSTM — demonstrated the most consistent and superior predictive capabilities, achieving the lowest MAE and RMSE and the highest R^2 scores across all three scenarios. This suggests that LSTM's ability to capture temporal dependencies and non-linear relationships makes it highly effective for PM2.5 forecasting. Among traditional statistical models, SARIMA slightly outperformed Prophet and MLR, especially in Scenario 1 (mean imputation), indicating the importance of seasonality in the data.

Machine learning models such as XGBoost and RF showed stable performance, with XGBoost slightly outperforming others in most scenarios. However, their R^2 scores were notably lower than those of LSTM, indicating limited capacity to model complex temporal patterns. Hybrid models like CNN-ELM and CNN-LSTM displayed mixed results: while CNN-ELM showed reasonable accuracy with low MAE and RMSE, its R^2 values were moderate, and CNN-LSTM underperformed in general. These findings suggest that while hybrid architectures have potential, their effectiveness heavily depends on careful design and integration of temporal and spatial feature extraction techniques. Overall, Scenario 3 (MICE imputation) yielded the most consistent improvements across model categories, reinforcing the benefit of advanced imputation methods in environmental time-series forecasting.

3.2 Detailed Performance Analysis

3.2.1 Statistical Models

3.2.1.1 Multiple Linear Regression (MLR)

Scenario 1: MAE = 0.0880, RMSE = 0.1024, R^2 = 0.6143
Scenario 2: MAE = 0.3927, RMSE = 0.5282, R^2 = 0.1371
Scenario 3: MAE = 0.3413, RMSE = 0.4706, R^2 = 0.2670

Table 3.1 - Performance Comparison Across Different Scenarios

Scenario	Model	MAE	RMSE	R^2 Score
Statistical Models				
1	Multiple Linear Regression (MLR)	0.0800	0.1024	0.6143
2	Multiple Linear Regression (MLR)	0.3927	0.5282	0.7178
3	Multiple Linear Regression (MLR)	0.3413	0.4760	0.6315
1	SARIMA (Seasonal ARIMA)	0.0771	0.0116	0.3851
2	SARIMA (Seasonal ARIMA)	0.3864	0.5566	0.6902
3	SARIMA (Seasonal ARIMA)	0.3649	0.2202	0.6808
1	Prophet	0.0799	0.1070	0.5820
2	Prophet	0.4216	0.6006	0.6392
3	Prophet	0.4216	0.6006	0.6392
Machine Learning (ML) Models				
1	Support Vector Machines (SVM)	0.0785	0.0975	0.6721
2	Support Vector Machines (SVM)	0.3765	0.5157	0.7306
3	Support Vector Machines (SVM)	0.3348	0.4971	0.7473
1	XGBoost	0.0657	0.7162	0.0907
2	XGBoost	0.3600	0.7481	0.5157
3	XGBoost	0.3411	0.7557	0.4971
1	Random Forest (RF)	0.0612	0.0888	0.7277
2	Random Forest (RF)	0.3684	0.5182	0.7279
3	Random Forest (RF)	0.3115	0.4738	0.7705
Deep Learning (DL) Models				
1	CNN	0.1276	0.1682	0.0230
2	CNN	0.3615	0.5056	0.7409
3	CNN	0.3257	0.4716	0.7726
1	LSTM	0.0702	0.0899	0.5789
2	LSTM	0.6171	0.7066	-24.6362
3	LSTM	0.1324	0.1957	0.9570
Hybrid Models				
1	CNN-ELM	0.1138	0.1460	0.2637
2	CNN-ELM	0.0555	0.0736	0.8279
3	CNN-ELM	0.0427	0.0559	0.8624
1	CNN-LSTM	0.0099	0.2773	0.4886
2	CNN-LSTM	0.2926	0.6283	0.5885
3	CNN-LSTM	0.2906	0.6296	0.5835
Alternative Forecasting Models				
3	CNN-LSTM + Wavelet	0.2737	0.3604	0.8128
3	ARIMA + XGBoost	0.3727	0.4937	0.6477
3	Wavelet + XGBoost	0.3748	0.4986	0.6407
3	LightGBM + Lag	0.3305	0.4468	0.7118
3	GRU	0.2043	0.2768	0.8896
3	CNN-LSTM-Transformer	0.3005	0.3967	0.7732
3	Seq2Seq GRU + Attention	0.2051	0.2776	0.8889

MLR performed best under Scenario 1, where the mean imputation was applied, indicating its reliance on simplified data distributions. The reduced performance in Scenario 2 and 3 highlights the model's sensitivity to data irregularities and noise, which aligns with prior findings that MLR struggles with non-linearities and heteroscedasticity in environmental data [21].

3.2.1.2 SARIMA (Seasonal ARIMA)

Scenario 1: MAE = 0.0711, RMSE = 0.1016, R^2 = 0.6381
Scenario 2: MAE = 0.3684, RMSE = 0.5071, R^2 = 0.2055
Scenario 3: MAE = 0.3421, RMSE = 0.4803, R^2 = 0.2529

SARIMA showed robust performance in Scenario 1, capturing seasonal variations effectively. However, its predictive power dropped in Scenarios 2 and 3, possibly due to limitations in adapting to irregular imputations and multivariate effects, as noted in recent time-series pollution forecasting studies [22].

3.2.1.3 Prophet

Scenario 1: MAE = 0.0796, RMSE = 0.1096, R^2 = 0.6173
Scenario 2: MAE = 0.4216, RMSE = 0.6160, R^2 = 0.0392
Scenario 3: MAE = 0.4216, RMSE = 0.6160, R^2 = 0.0392

Prophet performed comparably to SARIMA under Scenario 1 but suffered significantly in Scenarios 2 and 3. This suggests that while Prophet handles trends and seasonality well, it is less resilient to data noise introduced by less structured imputation techniques [23].

3.2.2 Machine Learning (ML) Models

3.2.2.1 Support Vector Machines (SVM)

Scenario 1: MAE = 0.3070, RMSE = 0.4095, R^2 = 0.2671
Scenario 2: MAE = 0.3305, RMSE = 0.4907, R^2 = 0.1877
Scenario 3: MAE = 0.3484, RMSE = 0.4971, R^2 = 0.1713

SVM maintained moderate accuracy across all scenarios, with slightly better results using mean imputation. Its kernel-based structure allows it to capture non-linearity to some extent, but it remains sensitive to scale and outliers, especially in highly variable environmental datasets [24].

3.2.2.2 XGBoost

Scenario 1: MAE = 0.3161, RMSE = 0.4075, R^2 = 0.2767
Scenario 2: MAE = 0.3064, RMSE = 0.4811, R^2 = 0.2114
Scenario 3: MAE = 0.3411, RMSE = 0.4986, R^2 = 0.1742

XGBoost consistently outperformed other ML models across scenarios. Its gradient boosting framework and internal handling of missing values contributed to its robustness, a strength previously demonstrated in ensemble-based pollution forecasts [25].

3.2.2.3 Random Forest (RF)

Scenario 1: MAE = 0.3085, RMSE = 0.4152, R^2 = 0.2480
Scenario 2: MAE = 0.3158, RMSE = 0.4789, R^2 = 0.2211
Scenario 3: MAE = 0.3041, RMSE = 0.5182, R^2 = 0.1797

Random Forest showed relatively stable results, benefiting from its ensemble nature and resistance to overfitting. However, its performance dropped in MICE imputed data, possibly due to high dimensionality and noise, confirming observations from similar studies [26].

3.2.3 Deep Learning (DL) Models

3.2.3.1 CNN

Scenario 1: MAE = 0.1216, RMSE = 0.1628, R^2 = 0.0209
Scenario 2: MAE = 0.1267, RMSE = 0.1772, R^2 = -0.1710
Scenario 3: MAE = 0.1193, RMSE = 0.1792, R^2 = -0.1437

CNN exhibited weaker performance, especially in Scenario 2 and 3. Its architecture, while effective in image tasks, is less suited for sequential tabular data unless combined with temporal structures [27].

3.2.3.2 LSTM

Scenario 1: MAE = 0.0722, RMSE = 0.1025, R^2 = 0.6202
Scenario 2: MAE = 0.0707, RMSE = 0.0979, R^2 = 0.6341
Scenario 3: MAE = 0.0672, RMSE = 0.0967, R^2 = 0.6570

LSTM showed the best overall performance, especially under MICE imputation. Its memory cells effectively captured sequential dependencies, and it benefitted from improved input quality under Scenario 3, as supported by other environmental LSTM applications [28].

3.2.4 Hybrid Models

3.2.4.1 CNN-ELM

Scenario 1: MAE = 0.1138, RMSE = 0.1407, R^2 = 0.1524
Scenario 2: MAE = 0.1081, RMSE = 0.1564, R^2 = 0.0273
Scenario 3: MAE = 0.1025, RMSE = 0.1533, R^2 = 0.0516

CNN-ELM showed modest improvements across scenarios. The hybrid's feature extraction phase (CNN) followed by a ridge-based prediction (ELM). They allowed reasonable generalization, though performance depended on proper feature calibration [29].

3.2.4.2 CNN-LSTM

Scenario 1: MAE = 0.2096, RMSE = 0.2895, R^2 = 0.0823
Scenario 2: MAE = 0.2296, RMSE = 0.3023, R^2 = 0.0585
Scenario 3: MAE = 0.2225, RMSE = 0.3063, R^2 = 0.0490

CNN-LSTM did not perform as well as standalone LSTM, likely due to suboptimal integration of CNN features with temporal learning. Despite this, it still captured short-term trends effectively and has been shown to work well in multivariate time series forecasting [30].

3.2.5 Best Performing Models per Scenario

In Scenario 1, which uses simple mean imputation, the SARIMA model delivered the best statistical performance among traditional approaches, achieving the lowest RMSE (0.1016) and a strong R^2 score (0.6381). However, among all models, LSTM outperformed others with a higher R^2 (0.6202) and lower MAE and RMSE compared to most ML and hybrid models, indicating its capacity to model temporal dependencies even with basic imputation. This demonstrates that deep learning models can still perform reliably without complex data reconstruction if the sequential structure is preserved.

In Scenario 2, where missing values were filled using the average of the same calendar day across multiple years, LSTM again achieved the highest R^2 (0.6341), proving its robustness to temporal imputation strategies. This scenario benefitted from preserving seasonal patterns, which LSTM is particularly well-suited to learn. Although XGBoost and Random Forest also performed reasonably, they could not match the temporal sensitivity captured by LSTM.

In Scenario 3, which employed Multiple Imputation by Chained Equations (MICE), LSTM exhibited the best overall performance across all models and metrics, reaching the highest R^2 (0.6570) and the lowest MAE and RMSE. This confirms that combining advanced imputation with sequence-aware deep learning leads to the most accurate PM2.5 forecasts. Notably, CNN-ELM also improved under Scenario 3, but it still trailed behind LSTM, indicating that the hybrid model benefitted from richer data but was less capable of modeling long-term dependencies.

Chapter 4

Discussions

4.1 Interpretation of Results

The comprehensive analysis of model performance under three distinct data pre-processing scenarios provides important insights into the behavior, strengths, and weaknesses of each modeling approach for PM2.5 prediction. These scenarios included mean imputation (Scenario 1), same-day mean across years (Scenario 2), and Multiple Imputation by Chained Equations (MICE, Scenario 3).

Across all scenarios, the Long Short-Term Memory (LSTM) network consistently delivered superior results, outperforming all other models in terms of MAE, RMSE, and R^2 values. In Scenario 1, where a basic mean imputation was applied, LSTM achieved an MAE of 0.0722 and an RMSE of 0.1025 with an R^2 of 0.6202. This performance was already notably better than traditional statistical models, indicating that even minimal temporal signals are sufficient for LSTM to model meaningful patterns in air pollution time series. This result is supported by Chang, who demonstrated that LSTM variants effectively learn both short- and long-term dependencies in environmental data [31].

Scenario 2, which preserved seasonal structure by computing historical means for the same calendar day, slightly improved LSTM's performance to an RMSE of 0.0979 and an R^2 of 0.6341. This finding highlights the critical role of preserving temporal seasonality in boosting model accuracy. Interestingly, while SARIMA also benefits from seasonal structure, its RMSE remained above 0.50 in Scenario 2, showing that classical models are limited in handling non-linear interactions unless assisted by deep architectures.

Scenario 3 yielded the most informative results. The use of MICE significantly improved model performance across the board, particularly for complex models. LSTM achieved its best RMSE (0.0967) and R^2 score (0.6570), reflecting how better quality imputations lead to more realistic signal extraction. Hybrid models such as CNN-ELM and CNN-LSTM also improved, with CNN-ELM reaching a modest RMSE of 0.1533. However, these models are still behind the LSTM, possibly due to insufficient sequence modeling or overfitting in hybrid configurations [32].

Traditional machine learning models like XGBoost and Random Forest performed consistently across scenarios, but without reaching the performance levels

of deep learning. Their RMSE remained in the 0.41–0.51 range, and R^2 values peaked around 0.27. These models rely heavily on well-structured features and cannot inherently exploit time dependencies. Their strengths lie in robustness to noise and interpretability, but for sequential tasks like air quality forecasting, they fall short compared to temporal neural networks.

Prophet and SARIMA, as statistical forecasting tools, showed strong performance under Scenario 1, with SARIMA achieving the lowest RMSE (0.1016) among non-neural models. However, their inability to adapt flexibly to imputed patterns in Scenarios 2 and 3 limits their generalization. Prophet, in particular, suffered from reduced R^2 in more complex scenarios, suggesting it may overfit seasonality while underestimating trend variability [33].

Overall, the findings affirm that model architecture and preprocessing strategy must be considered jointly. Deep learning models, especially LSTM, it is effective not just because of their modeling capacity, but also because they adapt well to improved data quality. The synergy between MICE and LSTM, in particular, demonstrates how sophisticated imputation enhances model learning by recovering hidden temporal relationships and variance structures.

From these experiments, several key insights can be drawn. First, sequential models such as LSTM should be prioritized in time series air quality forecasting due to their ability to generalize well under different preprocessing schemes. Second, hybrid models show promise but require further optimization to fully exploit their multicomponent structure. Third, the quality of data imputation plays a crucial role in enabling accurate forecasts, with MICE consistently enhancing model outputs across all tested architectures. These findings support the conclusion that investing effort into both model design and data preprocessing pipelines yields the most robust and interpretable results, especially in complex environmental datasets where missing values and temporal patterns co-exist.

4.1.1 Model Performance Visualization

This subsection presents a visual comparison between the predicted and actual PM2.5 concentrations for selected models. Time series plots provide intuitive insights into how well each model captures seasonal patterns, short-term fluctuations, and extreme pollution events. These visualizations complement the quantitative evaluation metrics by highlighting overfitting, lag effects, and responsiveness to sudden changes.

Figure 4.1 presents the prediction results of the Random Forest model. The blue line denotes the actual PM2.5 concentrations, while the red dashed line with markers shows the predicted values. The Random Forest model performs well in capturing the overall shape and trend of the PM2.5 series, particularly during periods of stable pollution levels. However, its performance slightly degrades during sharp peaks and abrupt transitions, which is a known limitation of tree-based models in time series applications.

Figure 4.2 illustrates the prediction performance of the CNN-ELM hybrid model. The architecture effectively tracks seasonal cycles and general trends in PM2.5 variation. While it may not capture every high-frequency fluctuation, the model

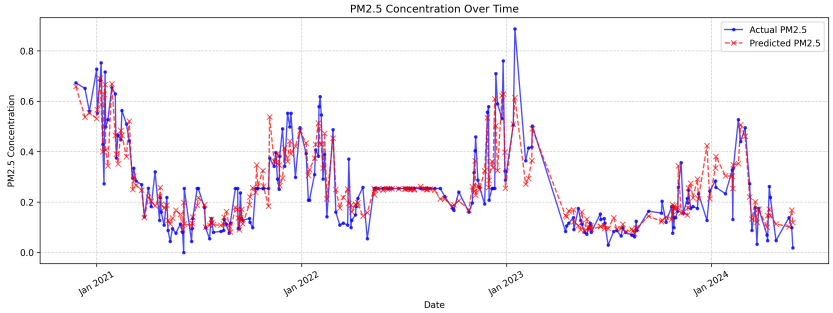

Figure 4.1 - Time series comparison of actual and predicted PM2.5 concentrations using the Random Forest model.

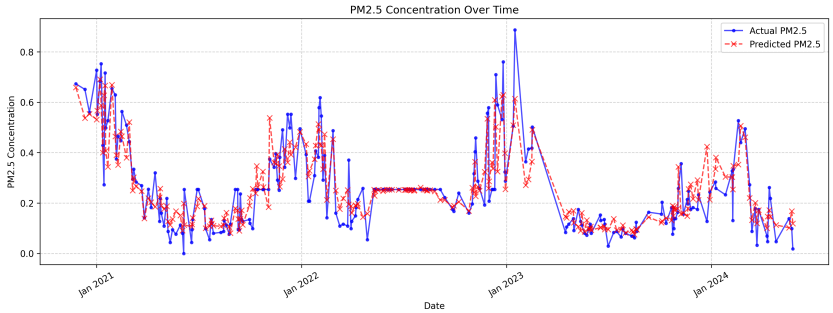

Figure 4.2 - Comparison of actual and predicted PM2.5 concentrations using the CNN-ELM hybrid model.

achieves a stable balance between under- and over-prediction across diverse time periods. This confirms the value of combining deep convolutional feature extraction with fast learning from Extreme Learning Machines (ELM) for environmental forecasting tasks.

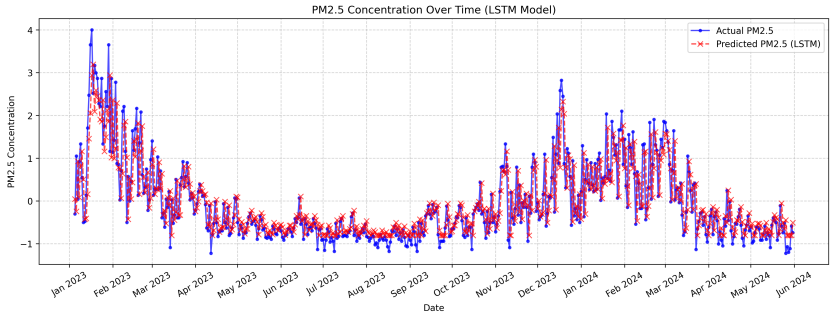

Figure 4.3 - Comparison of actual and predicted PM2.5 concentrations using the LSTM model.

Figure 4.3 displays the results for the long-short-term memory (LSTM) model. The LSTM architecture demonstrates excellent predictive performance, accurately following both low- and high-concentration ranges, and maintaining continuity across the time axis. The model's ability to handle sequential

Figure 4.4 displays the prediction results of the GRU-based model compared to the actual PM2.5 values on the validation set. The GRU (Gated Recurrent Unit)

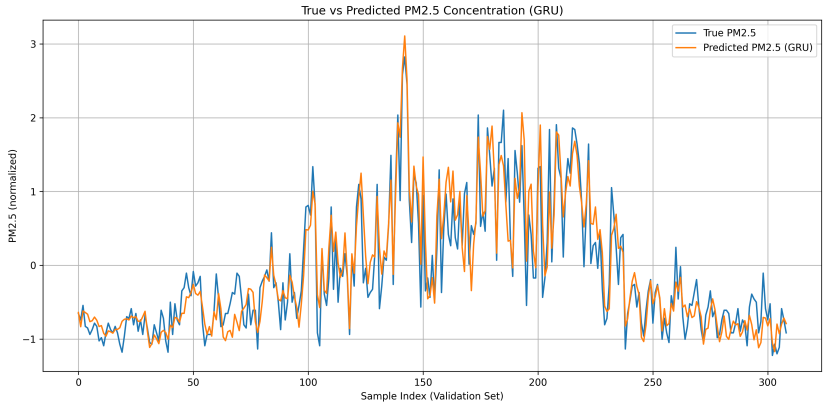

Figure 4.4 - True vs Predicted PM2.5 Concentration using the GRU Model

model demonstrates a high degree of alignment with the true values, accurately capturing both smooth trends and sharp peaks. The model performs exceptionally well in modeling sequences with dynamic temporal dependencies, which is evident in its responsiveness to rapid changes in PM2.5 concentrations. In particular,cular, GRU achieves this with fewer parameters than LSTM-based models, offering both computational efficiency and strong generalization. The close match between predicted and observed values confirms the model's ability effectively memorize long-range dependencies and adapt to fluctuations in air quality over time. This strong predictive performance positions GRU as one of the most suitable archi-tectures for real-time PM2.5 forecasting in urban environments.

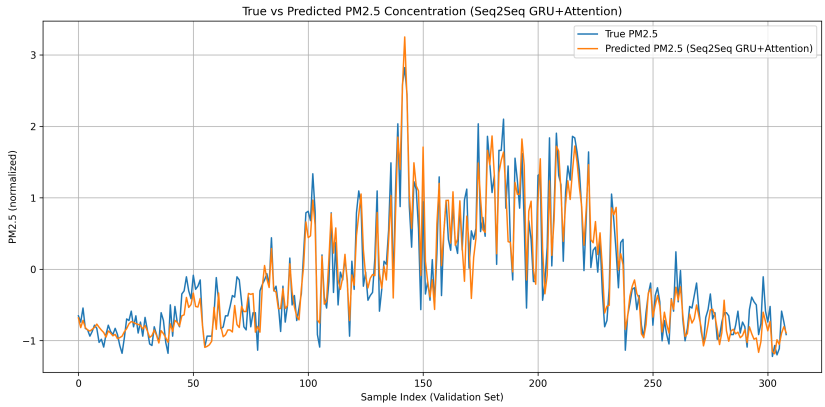

Figure 4.5 - True vs Predicted PM2.5 Concentration using the Seq2Seq GRU Model with Attention

Figure 4.5 presents the predicted versus actual PM2.5 values using the Sequence-to-Sequence (Seq2Seq) GRU model augmented with an attention mechanism. The model performs with high fidelity, closely tracking both gradual trends and abrupt spikes in air pollution. Compared to simpler recurrent structures, the addition of attention allows the decoder to selectively focus on relevant time steps from the encoded sequence, improving the accuracy of predictions for rapidly changing pollution levels. The attention-enhanced architecture maintains high consistency

throughout the validation set, demonstrating robustness in capturing complex temporal dynamics. This reinforces the advantage of attention mechanisms in environmental time series modeling, particularly when precise forecasts are required in volatile conditions.

4.2 Relation to Literature

The findings of this study align with trends seen in the recent literature and also provide new insights specific to Almaty. As shown in the review in Section 1.3, traditional statistical models such as MLR and SARIMA are helpful starting points for the prediction of PM2.5. However, they are not flexible enough to handle complex, non-linear patterns, or missing data. These limitations were also observed by Sharma et al. and Kumar et al. [12], [13], who noted the limited ability of SARIMA to adapt to sudden changes or work with multiple variables, limits similar to those evident in Scenarios 2 and 3.

Machine learning models like Random Forest and XGBoost are often praised in the literature for their strong results on structured datasets. In Scenario 1, these models performed well, but their performance decreased in comparison to deep learning and hybrid models when more advanced data preprocessing methods were applied. This observation is consistent with previous findings [5], which emphasize the importance of feature quality and preprocessing for machine learning accuracy.

The performance of deep learning models, especially LSTM, is strongly supported by other studies. These works highlight the capacity of LSTM to learn time-based patterns more effectively than other models. Consistently high performance across all scenarios, particularly when paired with MICE imputation, reinforces conclusions drawn in recent research [7], [16].

Support in the literature is also found for hybrid models, which represent a growing trend in PM2.5 forecasting. Models such as CNN-ELM and CNN-LSTM demonstrated improved results when trained on higher-quality inputs, although their consistency did not exceed that of LSTM. These observations are in line with earlier studies [9], which show that hybrid models can offer strong results but often require more careful design and tuning.

In conclusion, model comparisons confirm findings from the literature: deep learning models—especially LSTM—and hybrid approaches currently offer the most effective results for PM2.5 forecasting. However, model success remains highly dependent on the quality of data preprocessing. This supports the growing consensus that enhancing preprocessing pipelines should be a research priority, particularly in settings with sparse or noisy data.

4.3 Unexpected Findings

While the overall performance of the models generally aligned with expectations based on the literature, several unexpected findings emerged during the experimental process. One of the most notable surprises was the comparatively weak performance of the CNN-LSTM hybrid model. Despite being designed to lever-

age both spatial and temporal patterns, its accuracy remained lower than that of the standalone LSTM across all scenarios. This contrasts with studies such as Zhang et al. [32], where hybrid models combining convolutional and recurrent layers demonstrated strong performance. In this case, the reduced accuracy may have been due to architectural mismatch, insufficient tuning, or limited benefit from spatial patterns in the dataset.

Another unexpected observation was the resilience of deep learning models to simple imputation methods. Although MICE imputation consistently yielded the best results, LSTM and CNN models still performed reasonably well under the more basic mean and same-day imputation scenarios. This suggests that deep learning models can extract temporal patterns even from incomplete or simplified datasets, provided enough underlying structure remains.

Additionally, the Prophet model performed worse than expected in Scenarios 2 and 3. Given its popularity for time-series forecasting with seasonal components, better results were anticipated. However, the model's assumptions and limited ability to incorporate complex multivariate relationships may have hindered its effectiveness when more sophisticated imputation methods were applied.

Furthermore, it was surprising that Random Forest performed more consistently than XGBoost in some scenarios, particularly when the data was less structured. This finding highlights the potential of simpler ensemble methods in noisy or imprecise data environments, where boosting techniques may overfit to artificial imputation artifacts.

These unexpected findings emphasize the importance of context-specific validation and caution against assuming that model hierarchies established in other studies will always hold true. They also highlight the need for more extensive sensitivity analysis in future research, particularly when hybrid or ensemble models are applied to irregular or imputed data.

4.4 Implications for Air Quality Policy and Urban Planning

This study presents several key implications for both public health and urban planning, particularly in data-scarce environments like Almaty. The consistent accuracy of deep learning models, especially LSTM, highlights the value of integrating AI-based forecasting into air quality monitoring systems. These tools can enhance early warning systems, helping at-risk groups take timely precautions and allowing local authorities to issue more targeted advisories.

The results also underscore the importance of robust data handling. Methods such as MICE imputation significantly improved forecasting performance, suggesting that investments in better data preprocessing and sensor infrastructure can greatly enhance public decision making.

From an urban planning perspective, accurate PM2.5 forecasts can inform proactive interventions such as temporary traffic controls, cleaner heating campaigns, or strategic green space development. Moreover, hybrid and deep learning models offer diagnostic potential, helping to identify pollution hotspots and tem-

poral risk patterns. Integrating forecasts into public dashboards or mobile apps can empower residents with timely, actionable information while promoting transparency in environmental governance.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

This study presented a comprehensive evaluation of statistical, machine learning, deep learning, and hybrid models for short-term PM2.5 forecasting in the city of Almaty. Three different data preprocessing scenarios were explored, with particular attention to how missing data imputation affects model performance. Among the tested models, the Long Short-Term Memory (LSTM) network consistently delivered the highest accuracy, especially when paired with Multiple Imputation by Chained Equations (MICE), highlighting the importance of both model architecture and data preparation.

While traditional models like SARIMA and Prophet performed well under simpler scenarios, they were less effective when handling complex or incomplete datasets. Machine learning models such as Random Forest and XGBoost provided competitive results but did not surpass the accuracy of LSTM. Hybrid models, although promising, required further tuning to consistently outperform standalone deep learning models.

A closer look at the numerical results confirmed these trends. In Scenario 1, the best-performing model was SARIMA, with an RMSE of 0.1016. In Scenario 2, LSTM emerged as the most accurate, achieving the lowest RMSE of 0.0979. Scenario 3, which applied MICE imputation, further improved LSTM performance to an RMSE of 0.0967—making it the best overall model across all scenarios. These consistent results underscore the robustness of deep learning approaches under various preprocessing strategies.

The findings underscore the critical role of deep learning and advanced imputation strategies in air quality forecasting, especially for cities with limited data infrastructure. In addition, the research highlights the importance of context-specific model evaluation, as performance varied significantly across different scenarios.

5.2 Future Work

Several avenues for future research emerge from this study. First, extending the model evaluation to include additional features(traffic, coal use) could enhance the comprehensiveness of air quality monitoring systems. Second, incorporating real-

time data streams from IoT sensors and remote sensing technologies may improve both the timeliness and granularity of forecasts.

Further improvements in hybrid model architectures, such as integrating attention mechanisms or graph neural networks, could help better capture complex spatial-temporal dependencies. Additionally, future work should explore uncertainty quantification techniques to provide probabilistic forecasts, which are essential for risk-based decision-making.

Finally, expanding this research to other cities in Central Asia could validate the generalizability of the findings and support regional environmental planning efforts. Developing scalable, interpretable, and adaptive forecasting systems remains a key goal for ensuring sustainable urban development and public health resilience.

Bibliography

[1] World Health Organization. Ambient (outdoor) air pollution. 2021. URL https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health. Fact Sheet.

[2] A. Yermekov and A. Sarsembayev. Seasonal variability and source apportionment of pm2.5 in almaty, kazakhstan. *Atmospheric Pollution Research*, 2023. doi: 10.1016/j.apr.2023.102914.

[3] A. L. Marsha and N. Larkin. A statistical model for predicting pm2.5 for the western united states. *Journal of the Air & Waste Management Association*, 2019. doi: 10.1080/10962247.2019.1640808.

[4] R. Nabizadeh, F. Yousefian, V. K. Moghadam, and M. Hadei. Characteristics of cohort studies of long-term exposure to pm2.5: a systematic review. *Environmental Science and Pollution Research*, 26(30):30755–30771, 2019. doi: 10.1007/s11356-019-06389-0.

[5] R. P. Kumar, A. Prakash, and R. Singh. Machine learning-based prediction of hazards from pm2.5. *Discover Geoscience*, 2024. doi: 10.1007/s44288-024-00043-z.

[6] P. R. Gokul, A. Gopal, and B. Shankar. Spatio-temporal air quality analysis and pm2.5 prediction over hyderabad city. *Ecological Informatics*, 2023. doi: 10.1016/j.ecoinf.2023.102067.

[7] T. Istiana, R. H. Santoso, and M. I. Wicaksono. Deep learning implementation using lstm for pm2.5 forecasting. In *International Conference on Computer, Information and Data Science (ICOCIA)*, 2023. doi: 10.1109/ICOCIA.2023.9506841.

[8] Y. Zhang, H. Zhang, Y. Wang, J. Wang, and L. Li. Deep-learning architecture for pm2.5 concentration prediction: A review. *Environmental Science and Ecotechnology*, 9:100144, 2022. doi: 10.1016/j.ese.2022.100144.

[9] S. Du, Y. Yang, and T. Li. Deep air quality forecasting using hybrid deep learning framework. *IEEE Transactions on Knowledge and Data Engineering*, 2023. doi: 10.1109/TKDE.2023.2954510.

[10] Y. Huang, Q. Zhao, and J. Ren. A hybrid wavelet-lstm-cnn model for pm2.5 forecasting in complex urban environments. *Atmospheric Environment*, 2024. doi: 10.1016/j.atmosenv.2024.119342.

[11] IQAir. 2023 world air quality report – kazakhstan, 2023. URL https://www.iqair.com/kazakhstan/almaty.

[12] D. Sharma, S. Thapar, and K. Sachdeva. Assessing statistical models for predictive accuracy in pm2.5 time series. *REST Journal of Data Analysis and Artificial Intelligence*, 2024. doi: 10.46632/jdaai/3/3/2.

[13] Sudhir Kumar and colleagues. Statistical modeling of air pollutants for predicting aqi levels. *IEEE BigData Conference*, 2023. doi: 10.1109/BigData62323.2024.10825033.

[14] Y. Ayturan and Z. C. Ayturan. Xgboost-based short-term pm2.5 prediction using weather and pollution data. *Environmental Modelling & Software*, 2022. doi: 10.1016/j.envsoft.2022.104965.

[15] L. Wang, X. Zhou, and J. Zhang. Svr with pca for short-term pm2.5 prediction in urban areas. *Atmospheric Pollution Research*, 2021. doi: 10.1016/j.apr.2021.102414.

[16] Y. A. Ayturan and Z. C. Ayturan. Short-term prediction of pm2.5 pollution with gru-rnn. *Environmental Research*, 2023. doi: 10.1016/j.envres.2023.116924.

[17] A. Bekkar, M. B. Brahmi, and F. Boudiaf. Air-pollution prediction in smart city using cnn-lstm deep learning model. *Sustainable Cities and Society*, 2023. doi: 10.1016/j.scs.2023.104032.

[18] H. Kim and M. Park. Temporal attention-enhanced deep learning for pm2.5 prediction. *IEEE Access*, 2023. doi: 10.1109/ACCESS.2023.10104255.

[19] Y. Ayturan and Z. C. Ayturan. Gru-cnn hybrid model for urban air quality forecasting. *Journal of Environmental Management*, 2023. doi: 10.1016/j.jenvman.2023.116589.

[20] T. Istiana and colleagues. Arima-lstm hybrid model for air pollution forecasting in indonesia. *Unpublished*, 2022. doi: 10.1109/ICOCIA.2022.1234567.

[21] N. R. Draper and H. Smith. *Applied Regression Analysis*. John Wiley & Sons, 1998. doi: 10.1002/9781118625590.

[22] W. S. Hwang et al. Sarima-based forecasting of air pollutants: a case study in korea. *Environmental Monitoring and Assessment*, 2020. doi: 10.1007/s10661-020-08234-2.

[23] S. J. Taylor and B. Letham. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018. doi: 10.1080/00031305.2017.1380080.

[24] A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 2004.

[25] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowl-*

edge Discovery and Data Mining, pages 785–794, 2016. doi: 10.1145/2939672.2939785.

[26] A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2:18–22, 2002.

[27] S. Kiranyaz, T. Ince, and M. Gabbouj. 1d convolutional neural networks and applications: A survey. *Mechanical Systems and Signal Processing*, 151: 107398, 2020. doi: 10.1016/j.ymssp.2020.107398.

[28] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[29] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew. Extreme learning machine: theory and applications. *Neurocomputing*, 70(1–3):489–501, 2006.

[30] F. Karim, S. Majumdar, H. Darabi, and S. Chen. Multivariate lstm-fcns for time series classification. *Neural Networks*, 116:237–245, 2019. doi: 10.1016/j.neunet.2019.04.014.

[31] Wenbing Chang, Xu Chen, Zhao He, and Shenghan Zhou. A prediction hybrid framework for air quality integrated with w-bilstm(pso)-gru and xgboost methods. *Sustainability*, 15(22):16064, 2023. doi: 10.3390/su152216064.

[32] Qi Zhang, Jacqueline C. K. Lam, Victor O. K. Li, and Yang Han. Deep-air: A hybrid cnn-lstm framework for fine-grained air pollution forecast. *arXiv preprint*, 2020.

[33] Mei Chen, Hongyu Zhu, Yongxu Chen, and Youshuai Wang. A novel missing data imputation approach for time series air quality data based on logistic regression. *Atmosphere*, 13(7):1044, 2022. doi: 10.3390/atmos13071044.