ASTANA IT UNIVERSITY

Department of Computer Science

DIPLOMA THESIS

Development of a System for Analyzing and Visualizing Air Quality in Astana Using Data Analysis Techniques

Student: _____

Scientific Supervisor:

_____

Astana, 2025

Contents

Introduction

Research Relevance

Air pollution is recognized as one of the most serious environmental threats to public health on a global scale. According to the World Health Organization (WHO), exposure to polluted air causes millions of premature deaths annually, and the burden of disease attributable to air pollution is comparable to risk factors such as unhealthy diet and tobacco smoking [1]. In 2015, the World Health Assembly adopted a resolution recognizing air pollution as a risk factor for non-communicable diseases, including ischemic heart disease, stroke, chronic obstructive pulmonary disease, asthma, and cancer.

Astana, the capital of the Republic of Kazakhstan, is characterized by a unique combination of factors affecting air quality: a sharply continental climate with extreme seasonal temperature variations (from $-40°$C in winter to $+40°$C in summer), intensive urban development, significant vehicular traffic, and proximity to industrial facilities. Studies conducted in Kazakhstan demonstrate substantial spatiotemporal variability in pollutant concentrations [2, 3].

Traditional monitoring methods based solely on data from stationary posts of state networks do not always provide sufficient spatial resolution and timeliness for management decision-making. At the same time, the rapid development of Internet of Things (IoT) technologies, low-cost sensors, and machine learning methods opens new opportunities for creating intelligent air quality monitoring and forecasting systems [4, 5].

The relevance of this work is driven by the need to develop a comprehensive system capable of integrating heterogeneous data sources, providing accurate pollution level forecasting, and delivering scientifically-based information to regulators and the public.

Object of Study

The object of study is the process of monitoring and analyzing ambient air quality in the city of Astana.

Subject of Study

The subject of study is the application of machine learning methods and data analysis techniques for:

1. **Machine Learning (ML) methods** for complex nonlinear spatiotemporal analysis:
    - Time series forecasting models: Support Vector Regression (SVR), Long Short-Term Memory (LSTM), and their hybrid variants;
    - Statistical pattern analysis methods: correlation analysis, seasonal decomposition (STL), and anomaly detection.
2. **Architectural strategies** for integrating highly heterogeneous air quality data:
    - Data from mandatory state networks (RGP Kazhydromet with 127+ automatic posts);
    - International reference stations (U.S. Embassy post);
    - Low-cost sensor systems and commercial APIs.
3. **Data Quality Assurance (DQA) methodologies** and ETL (Extract-Transform-Load) pipelines for unifying heterogeneous information flows.

Research Goal

The primary goal is the design, development, and validation of a robust, data-driven Air Quality Intelligence System (AAQIS) for the city of Astana. The system should provide:

1. Accurate short-term (24–48 hours) forecasts of key pollutant concentrations, primarily $PM_{2.5}$;
2. Identification and quantification of seasonal and diurnal pollution patterns through correlation analysis with meteorological factors;
3. Clear visualization of monitoring data, forecasts, and analytical results through a web interface for the public.

Research Tasks

To achieve the stated goal, the following tasks must be completed:

Task 1: Systematic Literature Review. Conduct a systematic review of the current state of air quality monitoring technologies (including IoT sensor networks and

open data APIs) and the application of ML/DL models for air quality forecasting. The review should cover at least 20 peer-reviewed scientific publications from IEEE, Scopus, Springer databases and authoritative institutional reports (WHO, U.S. EPA).

Task 2: Monitoring Ecosystem Analysis. Analyze the existing data monitoring ecosystem in Astana (RGP Kazhydromet, U.S. Embassy post, local sensor networks) to identify specific problems: heterogeneity of data formats and temporal resolutions; limited programmatic accessibility (absence of public APIs); data quality issues (gaps, outliers, sensor drift).

Task 3: DQA/ETL Methodology Development. Define and formalize a robust Data Quality Assurance (DQA) methodology and ETL pipeline necessary for integrating and harmonizing multi-parametric data streams into a Unified Data Model (UDM).

Task 4: ML Model Development and Implementation. Develop and implement specialized ML models: AAQIS-Forecast (SVR and LSTM models for forecasting $PM_{2.5}$ concentrations at 24–48 hour horizons); AAQIS-Patterns (correlation analysis and seasonal decomposition to identify pollution patterns and their relationship with meteorological factors).

Task 5: Model Validation. Conduct rigorous validation of the predictive capabilities of the developed forecasting models on historical data, with benchmarking against reference data from the U.S. Embassy monitoring station. Compare model performance using standard metrics (RMSE, MAE, R^2).

Research Hypothesis

Hypothesis: The application of modern machine learning methods, specifically Long Short-Term Memory (LSTM) neural networks and Support Vector Regression (SVR), trained on validated historical $PM_{2.5}$ data from the U.S. Embassy reference monitoring station in Astana (2018–2023), will provide accurate short-term (24–48 hour) air quality forecasts with prediction errors (RMSE) at least 20% lower than baseline statistical methods (persistence model, moving average).

Furthermore, correlation analysis between $PM_{2.5}$ concentrations and meteorological factors (temperature, humidity, wind speed) will reveal statistically significant seasonal and diurnal patterns that can inform public health advisories and urban planning decisions.

Scientific Novelty

1. For the first time for the city of Astana, a comprehensive web-based air quality monitoring and forecasting system has been developed, integrating real-time data from international APIs (AQICN, OpenAQ) with machine learning prediction models.

2. A comparative analysis of LSTM and SVR models for $PM_{2.5}$ forecasting in Astana's specific climatic conditions (sharply continental climate with extreme seasonal variations) has been conducted.

3. A methodology for automated data collection, cleaning, and harmonization from heterogeneous air quality APIs has been developed and implemented.

4. Seasonal and diurnal patterns of $PM_{2.5}$ concentrations in Astana have been quantitatively characterized through correlation analysis with meteorological factors.

Practical Significance

The practical significance of the work is determined by the possibility of using the developed AAQIS system for:

1. Operational public information about current and forecasted air quality in Astana through an accessible web dashboard with intuitive visualizations;

2. Providing 24–48 hour $PM_{2.5}$ forecasts to help citizens plan outdoor activities and protect vulnerable groups (children, elderly, people with respiratory conditions);

3. Demonstrating the feasibility of building cost-effective air quality monitoring systems using open data sources and modern ML techniques;

4. Creating a foundation for future expansion to include additional pollutants and cities across Kazakhstan as more data becomes available.

Thesis Structure

The diploma thesis consists of an introduction, three chapters, a conclusion, a list of references, and appendices.

The first chapter presents the theoretical foundations of intelligent air quality analysis. The second chapter presents the analysis of the subject domain and system design methodology. The third chapter describes the practical implementation of the system and experimental results.

Chapter 1

Theoretical Foundations of Intelligent Ambient Air Quality Analysis

The first chapter is devoted to systematizing theoretical knowledge and modern scientific achievements in the field of monitoring, analysis, and forecasting of ambient air quality. The chapter is structured as follows: Section 1.1 examines key atmospheric pollutants and their health impacts; Section 1.2 is devoted to reviewing modern monitoring technologies; Section 1.3 analyzes data quality and heterogeneity problems; Section 1.4 contains an overview of machine learning methods for forecasting and pattern analysis; Section 1.5 examines architectural approaches to building intelligent monitoring systems.

1.1 Ambient Air Pollution and Public Health Risks

1.1.1 Definition and Classification of Atmospheric Pollutants

Air pollution refers to the presence of substances in the atmosphere at concentrations exceeding natural background levels and capable of adversely affecting human health, ecosystems, and material objects. The World Health Organization and national regulatory bodies identify six main ("criteria") pollutants subject to mandatory monitoring [1, 6]:

1. **Particulate Matter (PM):**
   - $PM_{2.5}$ — particles with an aerodynamic diameter of less than 2.5 micrometers;
   - PM_{10} — particles with a diameter of less than 10 micrometers.

   Particulate matter is the most dangerous component of urban pollution due to its ability to penetrate deep into the respiratory system and even into the bloodstream (in the case of ultrafine particles $PM_{0.1}$).
2. **Nitrogen Dioxide (NO_2):** Formed during high-temperature combustion processes. A marker of vehicular emissions.

3. **Sulfur Dioxide (SO_2):** Characteristic of emissions from coal and oil combustion. A marker of industrial and energy sector emissions.

4. **Carbon Monoxide (CO):** Formed during incomplete combustion. High concentrations are hazardous to human health.

5. **Ozone (O_3):** Secondary pollutant formed in the atmosphere as a result of photochemical reactions involving NO_x and VOCs.

6. **Lead (Pb):** Heavy metal, currently significantly reduced due to the phase-out of leaded gasoline.

In addition, for specific regions, the following may be relevant: Hydrogen Sulfide (H_2S) — characteristic of oil and gas industry regions and the Caspian region of Kazakhstan [2]; formaldehyde and other volatile organic compounds (VOCs).

1.1.2 Impact of Air Pollution on Public Health

A meta-analysis of epidemiological studies demonstrates a statistically significant relationship between air pollutant concentrations and a wide range of health effects [1]:

Respiratory System: Exacerbation of asthma; chronic obstructive pulmonary disease (COPD); reduced lung function; increased susceptibility to respiratory infections.

Cardiovascular System: Ischemic heart disease; stroke; hypertension.

Carcinogenic Effects: Lung cancer (outdoor air pollution classified by IARC as a Group 1 carcinogen).

Neurological Effects: Emerging evidence on neurodevelopmental effects and cognitive decline.

Vulnerable Groups: Children; elderly; people with pre-existing respiratory and cardiovascular conditions; outdoor workers.

1.1.3 WHO Air Quality Guidelines

In 2021, the World Health Organization published updated Global Air Quality Guidelines (AQG), significantly tightening recommendations for acceptable pollutant concentrations [1]:

Table 1.1: WHO Air Quality Guidelines 2021

Pollutant	**Averaging Period**	**AQG Level**
$PM_{2.5}$	Annual mean	5 $\mu g/m^3$
$PM_{2.5}$	24-hour mean	15 $\mu g/m^3$
PM_{10}	Annual mean	15 $\mu g/m^3$
PM_{10}	24-hour mean	45 $\mu g/m^3$
NO_2	Annual mean	10 $\mu g/m^3$
NO_2	24-hour mean	25 $\mu g/m^3$
SO_2	24-hour mean	40 $\mu g/m^3$
O_3	8-hour mean	100 $\mu g/m^3$
CO	24-hour mean	4 mg/m^3

It should be noted that WHO guidelines are recommendations and are not legally binding. National standards may differ (typically being less stringent).

1.1.4 Air Quality Indices

To communicate air quality status to the public, aggregate Air Quality Index (AQI) scales are used. The most widely used indices include:

U.S. EPA AQI: A 6-level scale (from "Good" to "Hazardous") based on concentrations of criteria pollutants, widely used internationally [6].

European CAQI: Common Air Quality Index developed for EU countries.

China AQI: National index accounting for local conditions.

The general formula for calculating sub-index for pollutant p:

$$\text{AQI}_p = \frac{I_{hi} - I_{lo}}{BP_{hi} - BP_{lo}} \times (C_p - BP_{lo}) + I_{lo} \tag{1.1}$$

where C_p is the pollutant concentration, BP are breakpoints, and I are corresponding index values. The overall AQI is the maximum of all pollutant sub-indices.

1.2 Modern Air Quality Monitoring Technologies

1.2.1 Classification of Monitoring Systems

Air quality monitoring systems can be classified according to several criteria:

By Measurement Principle:

- Reference (regulatory-grade) equipment;
- Low-cost sensors (LCS);

- Satellite observations (remote sensing);
- Mobile monitoring platforms.

By Deployment Model:

- Fixed (stationary) networks;
- Mobile measurements;
- Personal exposure assessment devices.

By Data Accessibility:

- Government (closed systems);
- Open data platforms;
- Commercial systems.

1.2.2 Regulatory Monitoring Networks

Regulatory networks are operated by government agencies and form the basis for assessing air quality status and compliance with standards.

Characteristics of Regulatory Stations:

- High accuracy (meeting metrological requirements);
- Regular calibration and quality control;
- Complete set of criteria pollutants;
- High cost (from tens of thousands to hundreds of thousands of dollars per station);
- Limited spatial coverage (typically 1 station per 100,000–500,000 population).

Data Availability: In many countries (including Kazakhstan), regulatory data is not always available in real-time through public APIs, limiting its use for operational systems.

1.2.3 Low-Cost Sensors and IoT Systems

The development of low-cost sensor (LCS) technology opens new opportunities for densifying monitoring networks [5, 7]:

Sensor Types: Optical particle counters (for PM); electrochemical sensors (for gases NO_2, SO_2, CO, O_3); metal oxide sensors (MOS).

Advantages: Low cost (from tens to several hundred dollars); compact size; Internet connectivity (IoT); possibility of dense network deployment.

Limitations and Challenges: Lower accuracy compared to reference equipment; susceptibility to humidity, temperature, cross-sensitivity effects; measurement drift over time; need for calibration and collocation studies.

Low-Cost Sensor Calibration: To ensure data quality, calibration using reference equipment is necessary [7, 8]: collocation with reference stations; statistical calibration (linear/polynomial regression); ML calibration (Random Forest, neural networks); accounting for meteorological parameters (temperature, humidity).

1.2.4 Mobile Monitoring

Mobile platforms equipped with sensors are mounted on: public transport vehicles; postal service vehicles; dedicated monitoring vehicles; drones and UAVs.

Advantages: Increased spatial coverage; ability to identify pollution hotspots; lower cost compared to dense fixed networks.

Challenges: Temporal mismatch of data; complex spatial analysis; need for sophisticated data fusion algorithms.

1.2.5 Satellite Monitoring and Remote Sensing

Satellite instruments provide a global view of air quality through retrieval of aerosol and gas parameters from reflected/emitted radiation.

Satellite Systems: MODIS (Terra/Aqua); Sentinel-5P (TROPOMI); GOES-ABI (geostationary).

1.3 Data Quality and Heterogeneity Challenges

1.3.1 Sources of Data Heterogeneity

Multi-source data integration faces several challenges:

Temporal heterogeneity: Different temporal resolutions (1-minute, hourly, daily averages); temporal misalignment; timezone differences.

Spatial heterogeneity: Point measurements vs. area averages (satellite); different site representativeness.

Measurement heterogeneity: Different measurement principles; different accuracy and precision; different units and reference conditions.

1.3.2 Data Quality Issues

Common issues include: missing values (equipment malfunction, power outages, maintenance); outliers and anomalies (calibration errors, interference); drift (gradual sensor

degradation); inconsistencies between sources.

1.3.3 Data Quality Assurance Methods (DQA)

Automated Checks: Range checks; rate-of-change checks; consistency checks between parameters; statistical outlier tests (z-score, IQR).

Missing Value Imputation Methods: Linear interpolation (for short gaps); seasonal decomposition (STL); k-NN imputation considering spatial correlation; ML methods (Random Forest, MICE).

Calibration and Correction: Linear regression against reference data; multivariate correction (accounting for T, RH); ML calibration.

1.3.4 Data Accessibility Challenge

A critical barrier for intelligent system development is limited data accessibility. Government networks often lack public APIs, and data may be available only through manual requests or web scraping. International platforms like OpenAQ and AQICN address this gap by aggregating and standardizing data from multiple sources.

1.4 Machine Learning Methods for Air Quality Analysis

1.4.1 Overview of ML Applications in Air Quality Tasks

Main ML Tasks in the Air Quality Domain:

1. Concentration forecasting (regression/time series forecasting) — *primary focus of this work*;
2. Pattern analysis and correlation with meteorological factors — *secondary focus of this work*;
3. Pollution level classification (AQI category prediction);
4. Source apportionment (requires multi-pollutant data);
5. Low-cost sensor calibration;
6. Spatial interpolation (spatial prediction);
7. Anomaly detection.

1.4.2 Models for Time Series Forecasting

Support Vector Regression (SVR): A kernel-based method for regression problems, widely used in environmental time series forecasting [9].

Principle: Mapping data to high-dimensional feature space via kernel function; finding a hyperplane that minimizes prediction error within an ε-tube.

Kernels: Linear; Polynomial; Radial Basis Function (RBF) — most common for air quality.

Advantages: Effective for medium-sized datasets (10,000–100,000 samples); good generalization with proper regularization; works well with nonlinear relationships.

Limitations: Scales poorly to very large datasets; sensitive to hyperparameter selection (C, γ, ε).

Long Short-Term Memory (LSTM) Networks: A specialized type of recurrent neural network designed to capture long-range dependencies in sequential data [4].

Architecture: Comprises memory cells with input, forget, and output gates; cell state for long-term information storage; hidden state for short-term information.

LSTM Advantages for Air Quality Forecasting: Ability to capture long-term temporal dependencies; accounting for seasonal patterns (daily, weekly, annual cycles); capability to work with multiple input features; scalability for large datasets.

A study by Sharipova et al. (2025), conducted on Astana data, demonstrated the effectiveness of LSTM for forecasting atmospheric pollutant emissions with high accuracy [10].

Hybrid and Ensemble Models: Modern research demonstrates the benefits of combining models [11]: CNN-LSTM for extracting spatiotemporal features; attention mechanisms for weighting the importance of time steps; Graph Neural Networks for accounting spatial relationships between stations; ensemble methods for improving robustness.

1.4.3 Statistical Methods for Pattern Analysis

Correlation Analysis: Quantifying relationships between pollutant concentrations and influencing factors (meteorological parameters, temporal features) is essential for understanding pollution dynamics [1].

Methods: Pearson correlation coefficient for linear relationships; Spearman rank correlation for non-linear monotonic relationships; partial correlation to control for confounding variables; cross-correlation for time-lagged relationships.

Seasonal Decomposition: Time series decomposition methods separate observed data into interpretable components:

- STL (Seasonal and Trend decomposition using Loess): robust decomposition into trend, seasonal, and residual components;

- Fourier analysis: identification of periodic patterns at different frequencies (diurnal, weekly, annual cycles);
- Moving averages: smoothing for trend extraction.

Anomaly Detection: Identifying pollution episodes and unusual events is important for public health alerts:

- Statistical methods: z-score, IQR-based detection;
- Machine learning approaches: Isolation Forest, Local Outlier Factor;
- Domain-specific thresholds: AQI category transitions, health guideline exceedances.

1.4.4 Model Quality Evaluation Metrics

Standard metrics are used for objective model performance evaluation [9]:

Regression Metrics (Forecasting):

- R^2 (coefficient of determination): proportion of explained variance;
- MSE (Mean Squared Error);
- RMSE (Root Mean Squared Error);
- MAE (Mean Absolute Error);
- MAPE (Mean Absolute Percentage Error).

For time series forecasting, RMSE is typically the primary metric as it penalizes larger errors more heavily, which is important for air quality applications where missing a pollution episode can have health consequences.

1.4.5 Feature Selection and Feature Engineering

Prediction quality significantly depends on input feature selection:

Feature Types for Air Quality Models:

1. Pollutant concentrations ($PM_{2.5}$, PM_{10}, NO_2, SO_2, O_3, CO);
2. Meteorological parameters (T, RH, pressure, wind speed and direction);
3. Temporal features (hour of day, day of week, month, holidays);
4. Spatial features (distance to roads, industrial zones, elevation);
5. Lagged values for time series.

1.5 Architectural Approaches to Intelligent Monitoring Systems

1.5.1 Architectural Patterns

Intelligent air quality monitoring systems typically follow layered or microservices architectures:

Layered Architecture: Data ingestion layer; data processing layer; analytics/ML layer; presentation layer.

Microservices Approach: Independent services for data collection, processing, and API delivery; enables scaling and independent deployment.

1.5.2 Data Pipeline Design

Modern systems employ ETL (Extract-Transform-Load) or ELT patterns:

ETL Components: Extract from heterogeneous sources (APIs, files, databases); Transform (cleaning, normalization, feature engineering); Load into unified storage.

1.5.3 AAQIS Conceptual Architecture

Based on the literature review, the following conceptual architecture is proposed for the Air Quality Intelligence System (AAQIS):

Data Sources Layer: OpenAQ API (historical reference data); AQICN API (real-time data and forecasts); OpenWeatherMap API (meteorological context).

ETL Layer: Scheduled data collection (hourly); format normalization and validation; storage in PostgreSQL database.

Analytics Engine Layer: AAQIS-Forecast (LSTM/SVR models for prediction); AAQIS-Patterns (correlation analysis, seasonal decomposition); AQI Calculator (EPA standard implementation).

ML/DL Platforms: TensorFlow/Keras for LSTM implementation; scikit-learn for SVR and statistical analysis; statsmodels for time series decomposition.

Presentation Layer: Django web application; Plotly.js for interactive visualizations; RESTful API for data access.

1.5.4 Forecasting Module (AAQIS-Forecast)

Input Data: Historical pollutant concentrations; meteorological parameters (current and forecast); temporal features.

Models: LSTM for capturing temporal patterns; SVR for reliable short-term forecasts; ensemble for improved robustness.

Output Data: Concentration forecast for 24–48 hours; AQI category forecast; confidence intervals.

1.5.5 Pattern Analysis Module (AAQIS-Patterns)

Input Data: Historical $PM_{2.5}$ concentration data; meteorological parameters (temperature, humidity, wind speed, pressure); temporal features (hour of day, day of week, season).

Methods: Correlation analysis (Pearson, Spearman) to identify relationships between pollutants and meteorological factors; seasonal decomposition (STL) to extract trend, seasonal, and residual components; anomaly detection to flag pollution episodes.

Output Data: Correlation coefficients with confidence intervals; seasonal and diurnal pattern visualizations; identified pollution episodes with potential meteorological triggers.

1.5.6 User Interface and Visualization

For the Public: Simplified interface with AQI and health recommendations; map with color-coded air quality indication; forecast for the coming days; alerts for high pollution levels.

Dashboard Components: Current AQI display with color coding; historical time series charts; forecast visualization for 24–48 hours; meteorological context (temperature, wind, humidity); trend analysis and seasonal patterns.

1.6 Chapter 1 Conclusions

The first chapter conducted a systematic analysis of theoretical foundations and current state of the field of intelligent ambient air quality analysis. Main conclusions:

1. Air pollution, particularly $PM_{2.5}$, poses a serious threat to public health. Updated WHO recommendations (2021) establish stricter threshold values, reflecting accumulated evidence about the harm of even low concentrations.

2. Modern air quality monitoring is implemented through a multi-level system: regulatory networks provide accurate reference measurements (such as the U.S. Embassy station in Astana), while open APIs (AQICN, OpenAQ) enable programmatic access to air quality data.

3. Data integration from heterogeneous sources requires robust ETL pipelines for timestamp normalization, unit conversion, and handling of missing values.

4. Machine learning methods, particularly LSTM neural networks and SVR, demonstrate high effectiveness for time series forecasting of pollutant concentrations, capturing both short-term dynamics and seasonal patterns.

5. For the city of Astana, the development of an air quality forecasting system is highly relevant given the sharply continental climate (extreme seasonal temperature variations from $-40°$C to $+40°$C), which significantly impacts pollution levels, particularly during the heating season.

6. The AAQIS conceptual architecture should include: multi-source data collection via APIs, robust ETL pipeline, ML forecasting models, and an interactive web dashboard for public access.

Further work (Chapter 2) will be devoted to detailed analysis of available data sources, justification for tool selection, and development of the system architecture for practical AAQIS implementation.

Chapter 2

Analysis of the Subject Domain and System Design Methodology

Chapter Introduction

The second chapter is devoted to the analysis of the subject domain specific to Astana's air quality monitoring ecosystem, the examination of available data sources, justification for the selection of methods and tools, and the development of the system architecture. The chapter is structured as follows: Section 2.1 analyzes the current state of air quality monitoring in Astana; Section 2.2 examines available data sources and their characteristics; Section 2.3 provides justification for the selection of methods and tools; Section 2.4 presents the system architecture design; Section 2.5 describes the data processing methodology.

Note on Scope: Given the limited availability of multi-pollutant historical data for Astana (only $PM_{2.5}$ measurements are consistently available from the U.S. Embassy reference station), the system focuses on **forecasting** and **pattern analysis** rather than complex source apportionment, which would require tracer pollutants (SO_2, NO_2, H_2S) not available in open datasets.

2.1 Analysis of Air Quality Monitoring in Astana

2.1.1 Climatic and Geographic Context

Astana, the capital of the Republic of Kazakhstan, is located in the northern part of the country on the banks of the Ishim River. The city is characterized by a sharply continental climate with extreme temperature variations.

Geographic Coordinates:

- Latitude: 51.1694° N
- Longitude: 71.4491° E

- Elevation: approximately 347 meters above sea level

Climate Characteristics:

- Winter temperatures: down to −40°C (January average: −14.2°C);
- Summer temperatures: up to +40°C (July average: +20.8°C);
- Annual precipitation: 300–350 mm;
- Prevailing winds: north and northwest directions;
- Frequent dust storms in spring and autumn periods.

These climatic conditions significantly impact air quality:

1. Extreme cold in winter leads to increased heating loads and associated emissions from residential and power generation sectors;
2. Temperature inversions during winter months trap pollutants near the surface;
3. Dust storms from surrounding steppe regions contribute to elevated PM_{10} concentrations;
4. Strong winds can both disperse and transport pollutants over long distances.

2.1.2 Primary Emission Sources in Astana

Understanding emission sources is critical for interpreting pollution patterns. The main pollution sources in Astana include:

1. **Transportation Sector:** Over 500,000 registered vehicles; intensive traffic on major arterial roads; characteristic pollutants: NO_2, CO, $PM_{2.5}$ (from diesel), BC.
2. **Power Generation and Heating:** Central heating plants primarily using coal and natural gas; individual heating systems in private residential areas; seasonal pattern with maximum emissions in heating season (October–April).
3. **Construction Activities:** Extensive urban development projects; dust emissions from construction sites (characteristic pollutants: PM_{10}, TSP).
4. **Industrial Facilities:** Industrial zone in the northern part of the city; characteristic pollutants: specific VOCs, SO_2, particulates.
5. **Transboundary and Natural Sources:** Dust transport from arid regions; emissions from neighboring industrial areas; Sand and Dust Storms (SDS).

2.1.3 Current Monitoring Infrastructure

The air quality monitoring infrastructure in Astana consists of several components with different capabilities and coverage:

RGP Kazhydromet Network: Kazhydromet operates the official state monitoring network in Astana, consisting of 5–7 automatic monitoring posts. It measures a wide range of parameters including $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, O_3, H_2S, and formaldehyde with hourly resolution. Data is published on the official website but lacks a public API.

U.S. Embassy Air Quality Monitor: An independent reference-grade BAM-1022 monitor located at the U.S. Embassy compound. It provides hourly $PM_{2.5}$ measurements according to U.S. EPA protocols and is available via AirNow/AQICN platforms.

Commercial and Research Sensors: Various low-cost sensor deployments exist, including the Sergek system sensors and individual PurpleAir units.

2.1.4 Identified Challenges and Gaps

Analysis of the current monitoring ecosystem reveals several critical challenges:

Data Accessibility Challenges:

- No standardized public API for Kazhydromet data;
- Limited spatial coverage (5–7 stations for 1.3+ million population);
- Data format heterogeneity across different sources.

Data Quality Challenges:

- Measurement discrepancies between sources;
- Data gaps due to equipment maintenance or power outages;
- Limited tracer measurements needed for robust source apportionment.

2.2 Data Sources and Their Characteristics

2.2.1 Primary Data Source: RGP Kazhydromet

RGP Kazhydromet serves as the primary authoritative data source. Table 2.1 lists the measured parameters.

Table 2.1: Pollutants Measured by Kazhydromet Network

Category	**Pollutant**	**Typical Units**
Criteria	$PM_{2.5}$, PM_{10}	$\mu g/m^3$
Pollutants	SO_2, NO_2, O_3	$\mu g/m^3$
	CO	mg/m^3
Tracer	H_2S, Phenol	$\mu g/m^3$
Components	NH_3, Formaldehyde	$\mu g/m^3$
Heavy Metals	Lead (Pb)	$\mu g/m^3$
	Cadmium (Cd), Nickel (Ni)	ng/m^3

Since no public API is available, data acquisition requires manual download or web scraping. Historical data is available from 2015 onwards with hourly resolution.

2.2.2 Reference Benchmark: U.S. Embassy Monitor

The U.S. Embassy PM2.5 monitor (MetOne BAM-1022) provides high-quality reference data. It serves as ground truth for model validation and enables calibration of low-cost sensors. Data is accessible via the AirNow API and AQICN API.

2.2.3 Supplementary Data Sources

External APIs for Gap-Filling:

- **OpenWeatherMap Air Pollution API:** Global coverage, gap-filling for missing periods.
- **Google Air Quality API:** Visualization enrichment and health recommendations.
- **AQICN (World Air Quality Index):** Real-time aggregated data for alerts.

Meteorological Data Sources:

- **Kazhydromet Meteorological Stations:** T, RH, Pressure, Wind Speed/Direction.
- **OpenWeatherMap Weather API:** Hourly conditions and forecasts.
- **ERA5 Reanalysis Data (ECMWF):** Historical analysis for model training.

2.2.4 Data Integration Strategy

The AAQIS system implements a priority hierarchy:

1. Kazhydromet (primary regulatory data);
2. U.S. Embassy (reference benchmark);
3. Commercial APIs (gap-filling);
4. Meteorological services (feature enrichment).

Integration solutions include normalizing all timestamps to UTC, converting units to SI ($\mu g/m^3$), and using ML imputation for missing values.

2.3 Justification for Selection of Methods and Tools

2.3.1 Programming Language Selection: Python

Python (version 3.10+) is selected as the primary development language due to its dominant position in data analysis and machine learning. Key libraries include **pandas** and **numpy** for data manipulation, **scikit-learn** and **TensorFlow/Keras** for machine learning, and **Django** for web development.

2.3.2 Machine Learning Model Selection

For Forecasting (AAQIS-Forecast Module):

- **Long Short-Term Memory (LSTM):** Selected for its ability to capture long-term temporal dependencies and seasonal patterns in time series data.
- **Support Vector Regression (SVR):** Selected as a robust alternative for medium-sized datasets, effective in modeling nonlinear relationships.

For Pattern Analysis (AAQIS-Patterns Module):

- **Correlation Analysis:** Pearson and Spearman correlation coefficients to identify relationships between $PM_{2.5}$ concentrations and meteorological factors (temperature, humidity, wind speed).
- **Seasonal Decomposition:** STL (Seasonal-Trend decomposition using LOESS) to separate trend, seasonal, and residual components in the time series.
- **Anomaly Detection:** Statistical methods (Z-score, IQR) to identify pollution episodes and unusual patterns.

2.3.3 Database Selection

Design Principle: Simplicity Over Premature Optimization. For the estimated data volume (approx. 50,000–70,000 records per year), a specialized distributed time-series database is not required.

Primary Database: PostgreSQL. Selected as the single unified database for all data types (time series, metadata, user accounts). It is robust, supports spatial queries via PostGIS, and integrates natively with Django. Time-series performance is optimized through proper B-tree indexing and partitioning strategies.

2.3.4 Web Framework Selection: Django

Django is selected for its "batteries-included" approach, providing a built-in admin interface, ORM, and authentication out of the box. **Django REST Framework (DRF)** will be used to expose API endpoints.

2.3.5 Summary of Technology Stack

Table 2.2 summarizes the selected technologies, prioritizing implementability for a diploma project.

Table 2.2: AAQIS Technology Stack

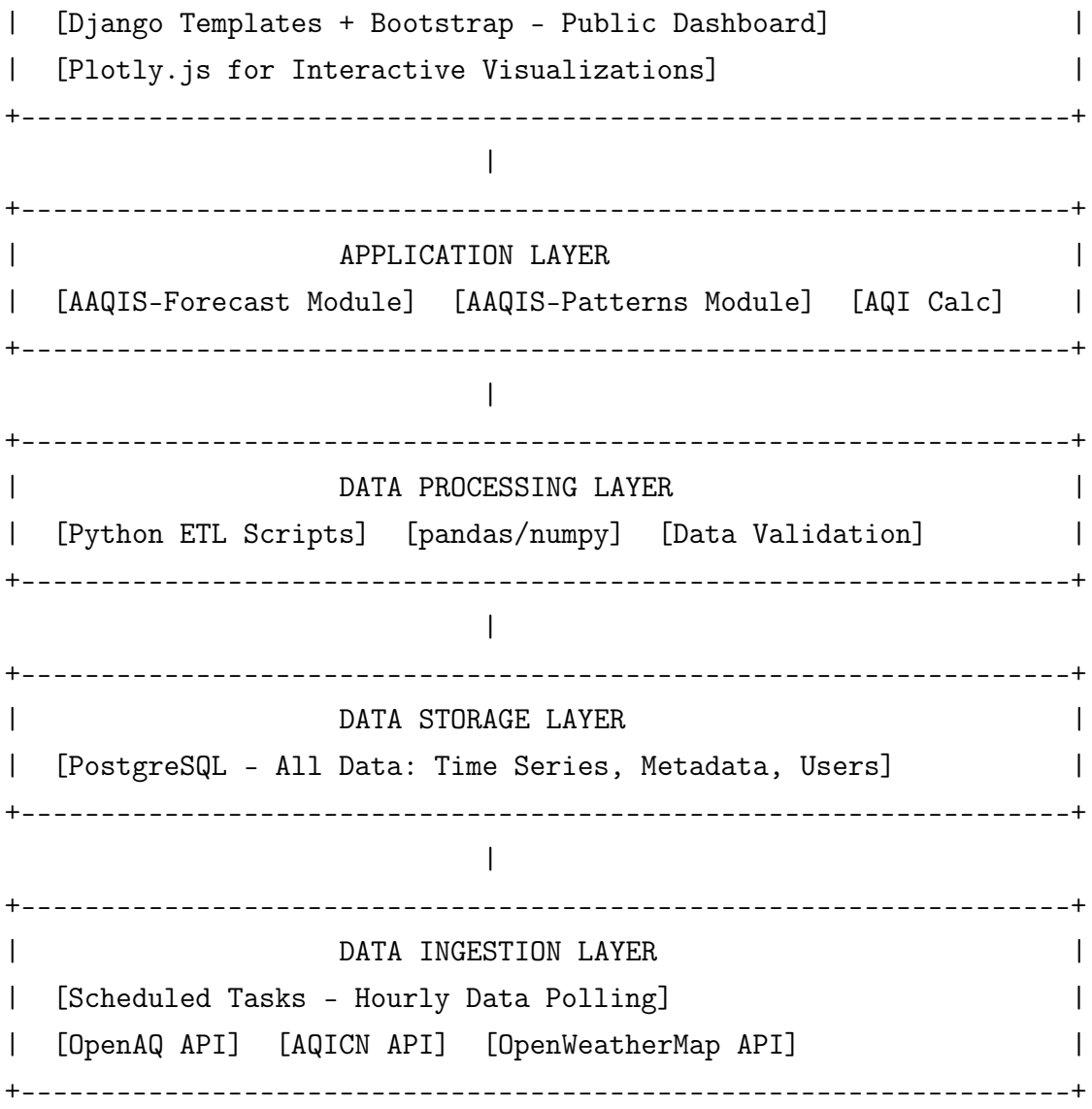
Component	**Technology**	**Purpose**
Language	Python 3.10+	Core development
Data Processing	pandas, numpy	ETL, Data manipulation
ML	scikit-learn, Keras	SVR, LSTM models
Database	PostgreSQL	Unified storage
Web Framework	Django	Backend, Templates
Frontend	Bootstrap, Plotly.js	UI, Visualizations
Deployment	Docker Compose	Local containerization

2.4 System Architecture Design

2.4.1 High-Level Architecture Overview

The AAQIS system follows a simplified layered architecture pattern, prioritizing maintainability.

```
+-------------------------------------------------------------+
|                    PRESENTATION LAYER                       |
```



2.4.2 Layer Descriptions

Data Ingestion Layer: Responsible for collecting data via scheduled tasks (cron or Django management commands). It includes API clients for OpenAQ (historical and reference data), AQICN (real-time AQI and forecasts), and OpenWeatherMap (meteorological context).

Data Processing Layer: Implements the ETL pipeline. It handles format standardization, timestamp normalization (UTC), unit conversion, and data quality checks (range limits, outlier detection).

Application Layer: Contains the core analytical logic.

- **AAQIS-Forecast:** Uses historical $PM_{2.5}$ data, lag features, and meteorological variables to predict pollutant concentrations for 24–48 hours using LSTM and SVR models.

- **AAQIS-Patterns:** Performs correlation analysis between $PM_{2.5}$ and meteorological factors, seasonal decomposition to identify trends, and anomaly detection to flag pollution episodes.
- **AQI Calculator:** Computes Air Quality Index based on U.S. EPA and WHO standards, with health recommendations.

Presentation Layer: Provides the user interface using **Django Templates** with Bootstrap for responsive design and Plotly.js for interactive visualizations.

2.4.3 Deployment Architecture

The system is designed for **Local Containerized Deployment** using Docker Compose. This ensures reproducibility for development and demonstration, and allows for easy migration to a cloud VPS if required in the future.

Docker Services:

- `web`: Django application (development server);
- `postgres`: Database.

Deployment Steps:

1. Clone the repository
2. Configure environment variables in `.env` file
3. Start services: `docker-compose up -build -d`
4. Apply database migrations
5. Access the application at `http://localhost:8000`

2.5 Data Processing Methodology

2.5.1 Data Collection and Cleaning

Data is collected hourly. The cleaning process involves:

1. **Format Standardization:** Parsing JSON/CSV/HTML into a unified schema.
2. **Timestamp Handling:** Alignment to hourly intervals in UTC.
3. **Missing Value Treatment:** Linear interpolation for gaps < 3 hours; ML-based imputation for longer gaps.

2.5.2 Feature Engineering

Temporal Features: Hour of day, Day of week, Month, Season, Holiday flags — capturing daily and seasonal pollution patterns characteristic of Astana's continental climate.

Lag Features: $PM_{2.5}$ values at $t-1, t-2, ..., t-24$ hours — leveraging autocorrelation in air quality time series for improved forecast accuracy.

Meteorological Features: Temperature, Humidity, Wind Speed, Atmospheric Pressure — key factors influencing pollutant dispersion and accumulation.

Derived Features: Rolling averages (24h, 7d), rate of change, deviation from seasonal mean — providing additional context for pattern recognition.

2.5.3 Model Training and Validation

Training uses the historical $PM_{2.5}$ dataset from OpenAQ (U.S. Embassy station, 2018–2023, approximately 16,000 hourly measurements). The data is split temporally: 70% training (2018–2021), 15% validation (2022), 15% test (2023) to prevent data leakage and ensure temporal validity.

Evaluation Metrics:

- **RMSE** (Root Mean Square Error): Primary metric for forecast accuracy;
- **MAE** (Mean Absolute Error): Robust to outliers;
- R^2 (Coefficient of Determination): Proportion of variance explained;
- **MAPE** (Mean Absolute Percentage Error): Interpretable percentage error.

Time-series cross-validation with expanding window is employed to ensure model robustness across different time periods.

2.6 Chapter 2 Conclusions

The second chapter presented the system design and methodology. Main conclusions:

1. A simplified Python-based stack (Django, PostgreSQL, Scikit-learn, Keras) was selected to prioritize implementability within the diploma project scope.
2. The system architecture focuses on two core modules: **AAQIS-Forecast** for $PM_{2.5}$ prediction using LSTM/SVR models, and **AAQIS-Patterns** for correlation and seasonal analysis.
3. Primary data source is the U.S. Embassy reference station (via OpenAQ API), providing 5+ years of validated hourly $PM_{2.5}$ measurements — sufficient for training robust forecasting models.

4. Real-time data integration via AQICN API enables live dashboard updates and current AQI display.

5. A containerized deployment strategy using Docker Compose ensures reproducibility for development and thesis defense demonstration.

The next chapter will present the practical implementation, including model training results, performance metrics, and the web dashboard interface.

Chapter 3

System Implementation and Experimental Results

This chapter will be completed after the practical implementation phase. It will include:

- Detailed description of the implemented system components
- Exploratory Data Analysis (EDA) results
- LSTM and SVR model training and hyperparameter tuning
- Comparative analysis of model performance
- Web dashboard screenshots and functionality description
- Discussion of results and limitations

Conclusion

To be completed after implementation.

Bibliography

[1] World Health Organization, "WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide," tech. rep., World Health Organization, Geneva, 2021.

[2] A. Kerimray *et al.*, "Spatiotemporal variations and contributing factors of air pollutants in Almaty, Kazakhstan," *Aerosol and Air Quality Research*, vol. 20, no. 6, pp. 1340–1352, 2020.

[3] D. Bissengaliyeva *et al.*, "Determination of the reliability of air pollution measurement data based on vehicular emission recognized as concomitant in Astana," *Scientific Journal of Astana IT University*, vol. 13, pp. 42–51, 2023.

[4] S. Shahid *et al.*, "Innovations in air quality monitoring: Sensors, IoT and future research," *Sensors*, vol. 25, 2025.

[5] L. Morawska *et al.*, "Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: How far have they gone?," *Environment International*, vol. 116, pp. 286–299, 2018.

[6] U.S. Environmental Protection Agency, "A guide to air quality and your health," Tech. Rep. EPA-456/F-14-002, U.S. Environmental Protection Agency, 2014.

[7] D. D. Drajic and N. R. Gligoric, "Reliable low-cost air quality monitoring using off-the-shelf sensors and statistical calibration," *Elektronika ir Elektrotechnika*, vol. 26, no. 3, pp. 32–41, 2020.

[8] W. Mui *et al.*, "Development of a performance evaluation protocol for air sensors deployed on a Google Street View car," *Environmental Science & Technology*, vol. 55, no. 19, pp. 12965–12974, 2021.

[9] J. Kerckhoffs *et al.*, "Performance of prediction algorithms for modeling outdoor air pollution spatial surfaces," *Environmental Science & Technology*, vol. 53, no. 3, pp. 1413–1421, 2019.

[10] S. Sharipova *et al.*, "Development of a neural network-based module for forecasting atmospheric pollutant emissions," *Scientific Journal of Astana IT University*, vol. 17, pp. 15–28, 2025.

[11] X.-B. Jin *et al.*, "Deep spatio-temporal graph network with self-optimization for air quality prediction," *Entropy*, vol. 25, no. 2, p. 247, 2023.