NOVEL DEEP LEARNING-BASED MODELS FOR AIR QUALITY PREDICTION: ADDRESSING NON-STATIONARY AND SPATIO-TEMPORAL ISSUES

ZHANG RUI

UNIVERSITI SAINS MALAYSIA

2025

NOVEL DEEP LEARNING-BASED MODELS FOR AIR QUALITY PREDICTION: ADDRESSING NON-STATIONARY AND SPATIO-TEMPORAL ISSUES

by

ZHANG RUI

Thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

June 2025

ACKNOWLEDGEMENT

Firstly, I would like to express my sincere gratitude to my supervisor, Dr. Norhashidah Awang, for her continuous support of my PhD studies and research, as well as for her patience, motivation, and immense knowledge. Her guidance has been invaluable throughout my research and the writing of this thesis. I also want to thank my friends and colleagues for the stimulating discussions, the sleepless nights we spent working together to meet deadlines, and all the fun we shared over the past four years. Last but not least, I would like to thank my family—my parents, my husband, and my daughters—for their unwavering spiritual support throughout my thesis journey and my life in general.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

$\mathbf{f}_t$	forget gate vector at time t in the LSTM cell
$\mathbf{i}_t$	input gate vector at time t in the LSTM cell
$\mathbf{c}'_t$	candidate cell state vector at time t in the LSTM cell
$\mathbf{o}_t$	output gate vector at time t in the LSTM cell
$\mathbf{h}_t$	hidden state vector at time t in the LSTM cell
$\mathbf{c}_t$	cell state vector at time t in the LSTM cell
$\mathbf{W}_f$	weight matrix for forget gate
$\mathbf{W}_i$	weight matrix for input gate
$\mathbf{W}_c$	weight matrix for cell state
$\mathbf{W}_o$	weight matrix for output gate
$\mathbf{b}_f$	bias vector for forget gate
$\mathbf{b}_i$	bias vector for input gate
$\mathbf{b}_c$	bias vector for cell state
$\mathbf{b}_o$	bias vector for output gate
$\odot$	element-wise multiplication operator
σ	sigmoid activation function
tanh	hyperbolic tangent activation function
$\mathbf{h}_t^{L-1}$	hidden state of the L-th layer at time t
$\mathbf{c}_t^{L-1}$	candidate cell state of the L-th layer at time t
$\mathbf{x}_t^{L}$	input vector to the L-th LSTM layer at time t
L	The number of layer or Transformer encoder block
$PE(t)_i$	positional encoding at position t for dimension i

$\mathbf{Q}$	query matrix
$\mathbf{K}$	key matrix
$\mathbf{V}$	value matrix
$\mathbf{Q}^{(i)}$	query matrix for i-th attention head
$\mathbf{K}^{(i)}$	key matrix for i-th attention head
$\mathbf{V}^{(i)}$	value matrix for i-th attention head
$\mathbf{W}^{O}$	output projection matrix for multi-head attention
$\text{Attention}(\cdot)$	scaled dot-product attention function
$\text{MultiHead}(\cdot)$	multi-head attention function
$\text{Concat}(\cdot)$	concatenation operation for attention heads
$\text{FFN}(\cdot)$	feed-forward network function
$\text{LayerNorm}(\cdot)$	layer normalization operation
ϵ	very small constant to prevent division by zero
$\mathbb{E}_t[]$	The mean
$\text{Var}_t[]$	The variance
γ	learnable affine parameter
β	learnable affine parameter
n	lookback length
I	The first principal components
k	prediction length
m	number of monitoring stations
$\mathbf{F}_{\text{intra\_t}}$	forget gate matrix of the intra-sequence temporal channel
$\mathbf{I}_{\text{intra\_t}}$	input gate matrix of the intra-sequence temporal channel
$\mathbf{O}_{\text{intra\_t}}$	output gate matrix of the intra-sequence temporal channel
$\mathbf{X}_{\text{intra\_t}}$	input value matrix in the intra-sequence temporal channel

$\mathbf{H}_{\mathrm{intra\_t}}$	hidden state matrix of the intra-sequence temporal channel
$\mathbf{C}_{\mathrm{intra\_t}}$	cell state matrix of the intra-sequence temporal channel
$\mathbf{F}_{\mathrm{inter\_t}}$	forget gate matrix of the inter-sequence spatial channel
$\mathbf{I}_{\mathrm{inter\_t}}$	input gate matrix of the inter-sequence spatial channel
$\mathbf{O}_{\mathrm{inter\_t}}$	output gate matrix of the inter-sequence spatial channel
$\mathbf{X}_{\mathrm{inter\_t}}$	input value matrix in the inter-sequence spatial channel
$\mathbf{H}_{\mathrm{inter\_t}}$	hidden state matrix of the inter-sequence spatial channel
$\mathbf{C}_{\mathrm{inter\_t}}$	cell state matrix of the inter-sequence spatial channel
$\mathbf{S}$	spatial relation matrix
$\mathbf{L}_s$	geographical matrix
$\mathbf{L}'_s$	embedded geographical matrix
$\mathbf{T}_{st}$	spatio-temporal trend matrix
$\mathbf{S}_{st}$	spatio-temporal seasonal matrix
$\mathbf{R}_{st}$	spatio-temporal remainder matrix

LIST OF ABBREVIATIONS

1D	one-dimensional
3D	three-dimensional
AE	Absolute Error
ANN	Artificial Neural Networks
AQI	Air Quality Index
ARIMA	Autoregressive Integrated Moving Average model
AttenSTLSTM	Spatio-temporal Long-Short Term Memory with Attention model
Bi-LSTM	Bidirectional Long Short-Term Memory
CMAQ	Community Multiscale Air Quality
CNN	Convolutional Neural Network
CO	Carbon monoxide
ConvLSTM	Convolutional LSTM
CTS-LSTM	Correlated Time Series-oriented LSTM
CWT	Continuous Wavelet Transform
DEWP	Dew Point Temperature
DSWT	Discrete Stationary Wavelet Transform
DWT	Discrete Wavelet Transform
EMD	Empirical Mode Decomposition
FC	Fully Connected
FFN	Feed-forward Network
GNN	Graph Neural Network
GRU	Gated Recurrent Unit

GPS	Global Positioning System
GWR	Geographically Weighted Regression
Hum	Relative Humidity
iTransformer	inverted Transformer
Jing-Jin-Ji	Beijing-Tianjin-Hebei
KNN	K-nearest neighbor
LSTM	Long Short-Term Memory
Loess	Locally weighted scatter plot smoothing
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MLP	Multi-Layer Perceptron
MTMC-NLSTM	Multi-Task Multi-Channel Nested LSTM
NAQPMS	Air Quality Prediction Model System
NLSTM	Nested Long Short-Term Memory
NO2	NO_2, nitrogen dioxide
O3	O_3, ozone
PCA	Principal Component Analysis
PCs	Principal Components
PM	Particulate Matter
PM10	PM_{10}, particles with a diameter of 10 micrometers or less
PM2.5	$PM_{2.5}$, fine particulate matter with a diameter of 2.5 micrometers or less
Pres	Atmospheric Pressure
Rain	Precipitation
ReLU	Rectified Linear Unit

RF	Random Forests
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Networks
R^2	R-squared
SARIMA	Seasonal Autoregressive Integrated Moving Average
SD	Secondary Decomposition
SO2	SO_2, sulfur dioxide
ST^2-LSTM	Seasonal-Trend Spatio-Temporal Long Short-Term Memory
ST-attention	Spatio-temporal attention
ST-fusion	Spatio-temporal fusion
ST-iTransformer	Spatio-Temporal inverted Transformer
STL	Seasonal-Trend decomposition based on the Locally weighted scatterplot smoothing
ST-LSTM	Spatio-temporal LSTM
SVM	Support Vector Machines
Temp	Average Temperature
VMD	Variational Mode Decomposition
WHO	World Health Organization
WRF-Chem	Weather Research and Forecasting model coupled with Chemistry
WS	Average 2-minute wind speed
WSPM	Wind Speed
WT	Wavelet Transform

MODEL BAHARU BERASASKAN PEMBELAJARAN MENDALAM UNTUK RAMALAN KUALITI UDARA: MENANGANI ISU KETIDAKPEGUNAN DAN RUANG-MASA

ABSTRAK

Pemantauan dan ramalan kualiti udara yang tepat adalah penting untuk kesihatan awam, kawalan alam sekitar, dan pengurusan bandar. Walau bagaimanapun, ramalan tetap mencabar disebabkan oleh ciri-ciri tidak pegun yang wujud dalam data bahan pencemar, yang dipengaruhi oleh sumber pelepasan yang dinamik, variabiliti meteorologi, dan aktiviti manusia. Kebergantungan ruang-masa yang kompleks, termasuk kesan tertunda antara stesen pemantauan, turut merumitkan proses pemodelan. Tambahan pula, keseimbangan antara ketepatan model dan kecekapan pengiraan mesti ditangani dengan teliti. Tesis ini menangani cabaran dalam meramal kualiti udara dengan mencadangkan tiga model inovatif. Model pertama, dinamakan PCA-DSWT-NLSTM, direka untuk ramalan masa dengan menggabungkan Analisis Komponen Utama (PCA), Transformasi Gelombang Diskret Pegun (DSWT), dan Memori Jangka Panjang Pendek Bersarang (NLSTM). Model kedua, iaitu Memori Jangka Panjang Pendek Ruang-masa Tren-Bermusim (ST^2-LSTM), dibangunkan untuk ramalan ruang-masa dengan mengambil kira tren dan kebermusiman. Model ketiga, iaitu Transformer Terbalik Ruang-masa (ST-iTransformer), dicadangkan untuk ramalan ruang-masa dengan pemboleh ubah eksogen. Ketiga-tiga model ini dinilai secara menyeluruh menggunakan data sebenar bahan pencemar udara melalui eksperimen perbandingan, kajian penghapusan, dan analisis komprehensif. Hasil menunjukkan bahawa model yang dicadangkan secara konsisten mengatasi model penanda aras dalam pelbagai metrik prestasi, sekali gus meningkatkan keupayaan ramalan kualiti udara secara signifikan. Secara khusus, PCA-DSWT-NLSTM cemerlang dalam ramalan PM2.5 dengan pengekstrakan ciri yang cekap dan keupayaan memori jangka panjang. ST^2-LSTM berkeupayaan mengenal pasti corak bermusim dan ruang dengan berkesan, menghasilkan keputusan yang stabil dan tepat. ST-iTransformer pula menunjukkan ketepatan yang tinggi di pelbagai stesen dengan memodelkan kesan tertunda reruang,

dan terbukti kukuh untuk ramalan jangka pendek dan jangka panjang. Kemajuan ini menyumbang kepada strategi pengurusan kualiti udara yang lebih berkesan dengan membolehkan ramalan yang tepat pada masanya, yang menyokong keputusan termaklum untuk kawalan dan pengurangan pencemaran. Kajian masa hadapan boleh menumpukan kepada peningkatan kecekapan model, penggabungan faktor input tambahan yang relevan, dan peluasan aplikasi model ke domain lain seperti aliran trafik dan ramalan cuaca.

NOVEL DEEP LEARNING-BASED MODELS FOR AIR QUALITY PREDICTION: ADDRESSING NON-STATIONARY AND SPATIO-TEMPORAL ISSUES

ABSTRACT

Accurate air quality monitoring and forecasting are essential for public health, environmental regulation, and urban planning. However, prediction remains challenging due to the inherent non-stationarity of pollutant data, influenced by dynamic emission sources, meteorological variability, and human activities. Complex spatio-temporal dependencies, including lead-lag effects among monitoring stations, further complicate the modeling process. Balancing model accuracy and computational efficiency is crucial. This thesis addresses the challenges of predicting air quality by proposing three innovative models. The first model, named PCA-DSWT-NLSTM, combines Principal Component Analysis (PCA), Discrete Stationary Wavelet Transform (DSWT), and Nested Long Short-Term Memory (NLSTM) for temporal prediction. The second model, named the Seasonal-Trend Spatio-Temporal Long Short-Term Memory (ST^2-LSTM), is developed for spatio-temporal prediction while considering trend and seasonality. The third model, named Spatio-Temporal inverted Transformer (ST-iTransformer), is proposed for spatio-temporal prediction with exogenous variables. All three models are rigorously evaluated on real air pollutant data through ablation studies, comparative experiments, and detailed analyses. Results demonstrate that the proposed models consistently outperform benchmark models across various performance metrics, significantly enhancing air quality forecasting capabilities. Specifically, PCA-DSWT-NLSTM excels in PM2.5 prediction with efficient feature extraction and long-term memory. ST^2-LSTM captures seasonal and spatial patterns effectively, showing stable and accurate results. ST-iTransformer delivers high accuracy across multiple stations by modeling spatial lead-lag effects, proving robust for both short-term and long-term forecasts. These advancements contribute to more effective air quality management. They enable timely and accurate forecasts, supporting informed decisions for pollution control and mitigation. Future work can focus on enhancing model efficiency, incorporating

additional relevant input factors, and extending the application of the models to other domains such as traffic flow and weather forecasting.

CHAPTER 1

INTRODUCTION

This chapter provides an overview of the thesis structure. Section 1.1 introduces the research background to set the context, while Section 1.2 describes the research problem and highlights gaps in current knowledge that this study seeks to address. Section 1.3 outlines the specific research objectives. It aims to examine different aspects of the main research question systematically. Section 1.4 defines the scope of the study, establishing its focus and boundaries. Section 1.5 discusses the significance of the research, emphasizing its contributions to both theory and practice. Section 1.6 acknowledges the limitations of the study and offers a balanced view of its applicability. Finally, Section 1.7 presents a structural outline of the thesis and guides readers through the subsequent chapters.

1.1 Background of Study

With rapid urbanization and growing industrial activities, air quality in many regions has been steadily declining, posing serious risks to the environment and public health (Kocak et al., 2023; Zhan et al., 2023). Air pollution is linked to many adverse health outcomes, including cardiovascular and respiratory diseases, various cancers, and neurological disorders, as highlighted by the World Health Organization (WHO). Given these risks, developing effective air quality prediction systems is vital for protecting human health and managing environmental quality. Accurate forecasts enable early warnings, which can help reduce exposure for vulnerable populations, such as children, the elderly, and those with existing health conditions. Reliable air quality predictions are also crucial for guiding researchers and policymakers in implementing effective pollution control measures. Beyond public health, these predictions support urban planning, industrial emissions regulation, and traffic management, all aimed at improving urban environments. Thus, advancements in air quality prediction are essential for mitigating health risks and creating healthier, more sustainable urban spaces.

To predict air quality effectively, it is important to monitor common pollutants, including particulate matter (PM), sulfur dioxide (SO2), nitrogen dioxide (NO2), carbon monoxide (CO), and ozone (O3). Particulate matter is categorized into PM10 and PM2.5 based on the size of the particles. PM10 refers to particles with a diameter of 10 micrometers or less. PM2.5 refers to fine particulate matter with a diameter of 2.5 micrometers or less. Each of these pollutants has distinct sources and impacts. PM2.5 and PM10, which originate mainly from combustion processes, construction activities, and unpaved roads, can lead to respiratory problems, cardiovascular diseases, and aggravate conditions such as asthma and bronchitis. SO2, largely emitted by burning fossil fuels in power plants and industries, can cause throat and eye irritation, as well as lung damage, especially in vulnerable populations. NO2 and CO, primarily generated by motor vehicles and industrial activities, have severe health effects. NO2 can irritate the respiratory system and reduce lung function, while CO can interfere with oxygen transport in the bloodstream, causing dizziness, confusion, and potentially life-threatening conditions at high concentrations. O3, formed through complex chemical reactions involving volatile organic compounds and nitrogen oxides in the presence of sunlight, can lead to chest pain, and coughing, and exacerbate chronic respiratory diseases. These pollutants can penetrate deep into the lungs and even enter the bloodstream, posing significant health risks and contributing to increased hospital admissions and premature mortality (WHO, 2022).

Building on the understanding of these pollutants and their health impacts, predicting air quality effectively has become a key focus area for mitigating risks and enhancing public health. As a result, air quality prediction models have gained prominence as fields of study in recent decades. Numerous strategies have been employed by researchers, including knowledge-based models and data-driven models. Knowledge-based models employ atmospheric dispersion and chemical reaction principles to understand the mechanisms behind pollutant emission, transfer, and dispersion processes. However, the significant computational resources and specialized expertise required by such models limit their practicality for real-time applications (Lv et al., 2016; Zhou et al., 2019). In contrast, data-driven models provide more computational efficiency and offer notable benefits for predicting future unknown air quality values based on existing data.

Most air pollutant and meteorological datasets are typically imbalanced, a challenge that has been widely recognized in the literature. These data-driven models can be divided into three categories: classical statistical models, traditional machine learning models, and deep learning models. Among them, deep learning models have demonstrated significant superiority in terms of performance and accuracy (Zhang et al., 2022). This is largely due to their flexible architectures, which allow for the construction of complex, multi-layered networks that represent intricate relationships within big data. Deep learning models also possess strong representation learning capabilities, which help identify both spatial and temporal dependencies in air quality data without requiring manual feature engineering. Another notable advantage of deep learning is its high nonlinear processing ability. This capability allows it to model the complex and dynamic behaviors inherent in air pollution data (Zhang et al., 2024). Unlike classical statistical methods or traditional machine learning models, deep learning can automatically learn and adapt to hidden patterns, even when dealing with large-scale datasets. This quality is essential for capturing the intricacies of environmental phenomena (Feng et al., 2023; Zeng et al., 2023).

Moreover, advancements in deep learning neural network architectures, such as Long Short-Term Memory (LSTM) and its variants, as well as Transformer and its variants, have significantly improved the ability of models to process and integrate heterogeneous data sources. These data sources include meteorological variables and historical pollutant values. Such advancements ultimately lead to more accurate and robust air quality predictions, making deep learning an ideal choice for addressing the challenges of air quality forecasting. Therefore, this work focuses on recent deep learning research to enhance the accuracy and reliability of air quality predictions.

1.2 Problem Statement

The non-stationary nature of air quality dynamics presents significant challenges for accurate deep learning predictions. Factors such as changing emission sources, varying weather conditions, seasonal influences, and unpredictable human activities introduce inconsistencies and fluctuations in pollutant levels over time. This makes historical data essential yet challenging for models like LSTM and its variants, such as Nested LSTM (NLSTM), to interpret effectively. Consequently, relying solely on these models is insufficient, and hybrid models incorporating decomposition techniques are necessary to improve prediction accuracy. Additionally, meteorological factors—such as temperature, humidity, precipitation, and wind—significantly influence air quality by affecting pollutant dispersion, accumulation, and removal. However, incorporating decomposition methods and meteorological data into predictive models increases complexity and computational demands, creating a trade-off between prediction accuracy and efficiency. Existing feature selection methods struggle to fully capture these intricate relationships, highlighting a need for comprehensive frameworks that effectively integrate decomposition techniques with robust feature selection to develop more accurate and computationally efficient air quality prediction systems.

In spatio-temporal forecasting, both temporal and spatial factors play pivotal roles. The influence of different time points on the current prediction changes over time. Spatial variations mean that the impact of neighboring stations on a target station also differs across locations. Simply selecting or omitting stations based on proximity does not adequately capture these dynamic relationships. It is essential to consider the evolving effects of both time and space to ensure accurate predictions. Furthermore, current time series decomposition techniques, such as Wavelet Transform (WT), Variational Mode Decomposition (VMD), and Empirical Mode Decomposition (EMD), face significant constraints in their implementation. These methods require complex parameter tuning and often fail to adequately consider inherent time series characteristics, including seasonality patterns, trend components, and random variations. Additionally, existing approaches typically separate spatial and temporal feature extraction into distinct modules, such as employing Convolutional Neural Networks (CNNs) or Graph Neural Networks (GNNs) for spatial features while using Long Short-Term Memory (LSTM) networks for temporal patterns, which inevitably results in information mixing problems. This separation makes it challenging for higher-level layers to effectively differentiate between various types of correlations, highlighting the need for a more unified framework that can handle both spatial and temporal dynamics simultaneously.

Spatio-temporal forecasting lies in the inconsistent movement of air pollutant concentrations between different stations. For instance, imagine that Station A is situated on the west side of the city, while Station B is on the east side. When the wind blows from west to east, common air pollutants like PM2.5 will show up at Station B later than at Station A. Specifically, as the wind carries the pollutants from west to east, the levels of PM2.5 at Station A will increase or decrease first, followed by similar changes at Station B after a certain period. This phenomenon emphasizes that air pollutant concentrations exhibit spatial correlations and temporal lead-lag relationships. Therefore, accurately capturing these lead-lag dynamics and the intricate spatio-temporal correlations is crucial for enhancing the predictive performance of air quality forecasting models. The emerging Transformer-based approaches, while promising, face their own set of distinct challenges in predicting the complex dynamics of spatially diverse and correlated air pollutants. Current Transformer architectures struggle with capturing intricate spatio-temporal relationships and the non-stationary nature of air pollution. These models face difficulties in integrating multiple influential factors, including historical and present pollutant concentrations, meteorological conditions, and spatial dependencies. Additionally, current approaches face technical limitations in their data embedding and token processing capabilities. They demonstrate inadequate utilization of spatial background information, with methods like Time2Vec falling short in fully leveraging spatial context. Furthermore, the practice of combining all variables into time-based tokens can affect prediction accuracy, as it fails to account for the natural lead-lag relationships that exist between sequences.

1.3 Research Objectives

This research aims to address the challenges of non-stationarity and spatio-temporal characteristics in air quality prediction through the development of advanced deep learning frameworks. Specifically, the objectives of this research are:

1) To develop an NLSTM-based model for temporal air quality prediction by employing decomposition techniques to address non-stationarity while incorporating meteorological factors.

2) To develop an LSTM-based model for spatio-temporal single-pollutant prediction by extending the original LSTM into a spatio-temporal framework while giving special attention to the essential time steps and monitoring stations.

3) To develop a Transformer-based model for spatio-temporal multi-pollutant prediction by providing a more flexible approach to handling complex spatio-temporal features while accounting for meteorological factors.

1.4 Scope and Limitation

This thesis aims to propose innovative deep learning models to address the challenges of non-stationarity and spatio-temporal variability in air quality prediction. The primary goal is to develop enhanced variants of LSTM and Transformer architectures, incorporating novel modifications that make them more suitable for capturing the dynamic and complex patterns inherent in air quality data. By employing these advanced models, the study aims to effectively leverage the concentrations of common air pollutants, including PM2.5, PM10, SO2, NO2, and O3, which are known to have significant health and environmental impacts. The dataset used for experiments focuses on the northern regions of China, particularly cities and areas such as Taiyuan, Beijing, and the Jing-Jin-Ji region (Fang et al., 2022; Yan et al., 2019; Zhang et al., 2016). These locations are characterized by rapid urbanization and extensive industrial activities, which contribute substantially to air pollution and create significant challenges for air quality management. Ultimately, this research aims to provide more accurate air quality forecasting models that can assist policymakers and environmental agencies in mitigating the adverse effects of air pollution and improving public health outcomes.

Despite continuous advancements in air quality time series forecasting, significant limitations remain. The dynamic nature of air quality is not an isolated phenomenon. It is influenced by historical data from both the same site and other locations, as well as by various factors such as gas emissions, agricultural activities, and traffic conditions. However, these factors are often random and difficult to quantify statistically, which presents challenges in enhancing prediction accuracy. Additionally, air quality patterns vary widely across different countries, terrains, and even cities. As a result, it is impractical to develop a single model capable of accurately predicting air quality indices for all monitoring sites globally. These complexities often pose significant challenges to air quality time series forecasting.

1.5 Significance of the Study

This research carries significance across theoretical, practical, and methodological dimensions. It provides a comprehensive foundation that advances scientific understanding and addresses real-world challenges. The study bridges gaps in knowledge. It offers practical solutions to pressing environmental issues and proposes innovative methodologies applicable across diverse domains.

1) From a theoretical perspective, this study enhances the understanding of complex air quality patterns in environmental data using innovative deep learning architectures. It contributes to the theoretical framework for managing non-stationary time series in environmental modeling by incorporating data decomposition techniques, and introduces novel approaches for integrating multiple data sources while effectively capturing multi-scale temporal and spatial dependencies.

2) In terms of practical applications, accurately forecasting air quality is crucial for effective urban planning, public health protection, and environmental management (Janarthanan et al., 2021). Specifically, accurate prediction enables evidence-based decision-making for urban planning and development. It facilitates the implementation of targeted emission control strategies and supports smart city initiatives. For public health protection, it provides early warning systems for adverse air quality events, supports preventive healthcare measures for vulnerable populations, and helps healthcare facilities prepare for potential increases in respiratory cases. In environmental governance, these forecasting capabilities support the dynamic adjustment of industrial production schedules, enable proactive traffic management during high-pollution periods, and promote cross-regional collaboration in pollution control.

3) The methodological contributions of this research extend beyond air quality prediction. The developed framework has significant potential for applications across various domains, including traffic flow prediction, weather forecasting, tourism demand analysis, hydrological predictions, industrial systems monitoring, energy consumption forecasting, and economic time series analysis. The technical innovations improve methods for handling multi-source heterogeneous data, capture complex spatial dependencies, and enhance techniques for multi-step prediction.

1.6 Thesis Outlines

The research framework is illustrated in Figure 1.1, consisting of three main objectives: temporal prediction, spatio-temporal single-pollutant prediction, and spatio-temporal multi-pollutant prediction. Each objective includes model development, data analysis, and experimental validation. Collectively, these components contribute to the study's findings, limitations, and directions for future work.

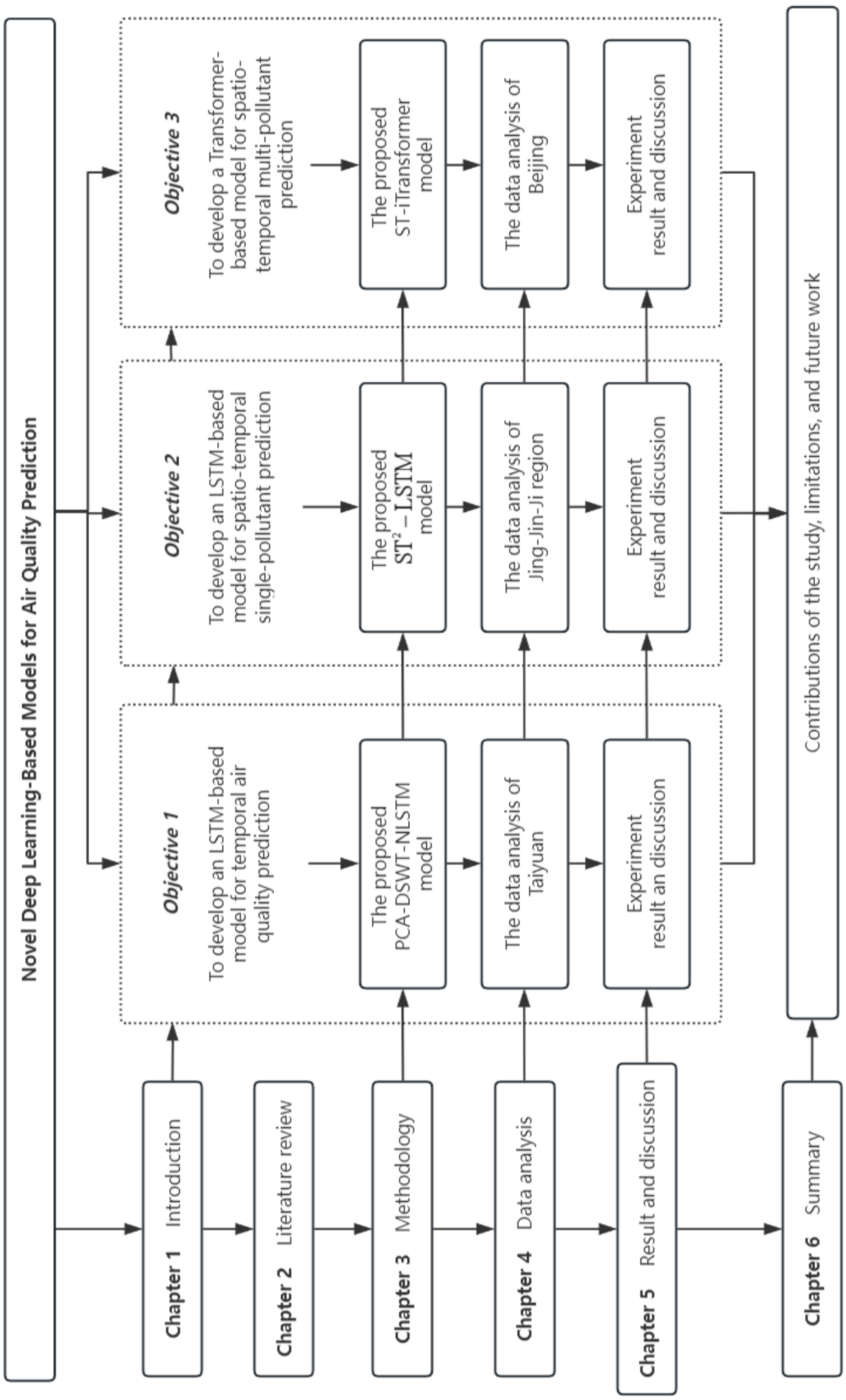

Figure 1.1: The research framework

This thesis is structured into six chapters. Chapter 1 provides a comprehensive introduction, including the research background, problem statement, objectives, study scope, significance, and limitations of the research. Chapter 2 presents an extensive literature review on air quality prediction. It covers classification, deep learning sequence models, temporal air quality models, spatio-temporal air quality models, and decomposition techniques. Chapter 3 outlines the methodology, with a particular focus on three proposed models. It details foundational components and data preprocessing. Next, it introduces the three proposed models: the PCA-DSWT-NLSTM model, Seasonal-Trend Spatio-Temporal Long Short-Term Memory (ST^2-LSTM) model, and Spatio-Temporal inverted Transformer (ST-iTransformer) model. Chapter 4 describes the datasets used: Taiyuan, Jing-Jin-Ji, and Beijing. It provides an analysis of their locations and data characteristics. Chapter 5 presents the experimental results and discusses model performance for each study area, offering insights into their effectiveness. Chapter 6 concludes the thesis by summarizing the key findings, discussing the limitations of the study, and suggesting potential directions for future research.

CHAPTER 2

LITERATURE REVIEW

This chapter provides a comprehensive overview of the current state of research in air quality monitoring and forecasting, with a particular focus on deep learning approaches. Specifically, Section 2.1 presents the classification of air quality forecasting models. Section 2.2 reviews the advancements in deep learning methods for sequence forecasting. Section 2.3 reviews the advancements in deep learning methods for temporal air quality forecasting. Section 2.4 elaborates on the progress in spatio-temporal sequence forecasting. Section 2.5 reflects on the advancements in time series decomposition techniques. The feature selection and extraction are discussed in Section 2.6. Finally, Section 2.7 shows the summary and research gap.

2.1 Classification of Air Quality Forecasting Models

Time series forecasting involves identifying patterns from historical observations and modeling these patterns to predict future values. This research area plays a crucial role in various fields such as economics, finance, transportation, and meteorology. In air quality studies, time series forecasting is essential for understanding the temporal and spatial dynamics of pollutants, contributing to sustainable development, and predicting extreme pollution events. Many researchers have explored various time series forecasting methods to predict different air quality indicators, enabling proactive environmental management and early warnings for extreme pollution events. Figure 2.1 presents a hierarchical classification of air quality prediction models. As illustrated, these models can be broadly categorized into two main branches: knowledge-based models and data-driven models.

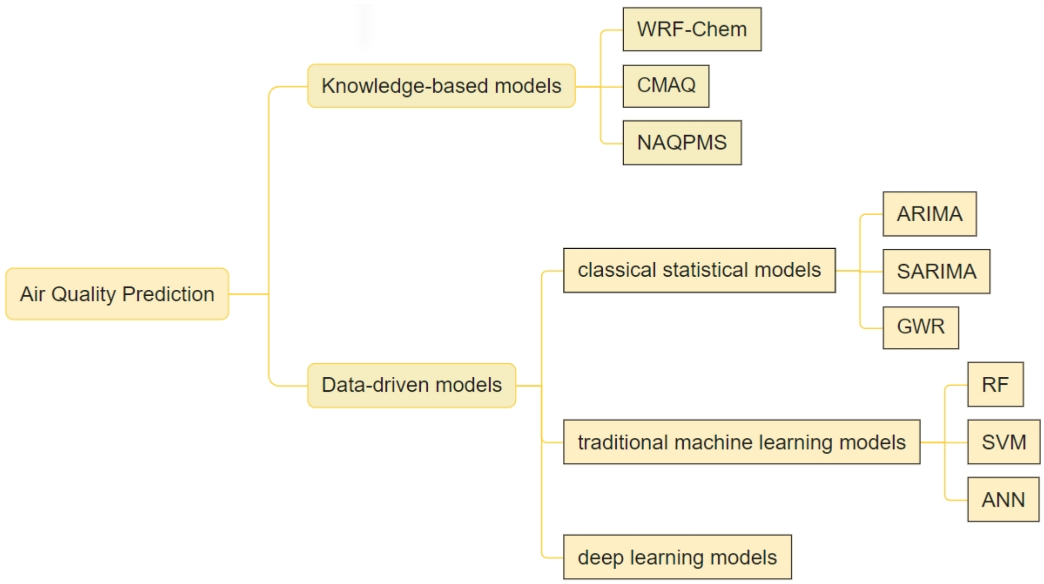

Figure 2.1: Taxonomy of air quality prediction models: knowledge-based and data-driven Approaches

2.1.1 Knowledge-based models

Knowledge-based models use atmospheric dispersion and chemical reaction models to simulate the processes of pollutant emission, transfer, and dispersion. Tie et al. (2007) utilized the Weather Research and Forecasting model coupled with Chemistry (WRF-Chem) to study the formation and evolution of high ozone concentration events in the urban area of Mexico City. The San Joaquin Valley's winter PM2.5 formation is simulated by Chen et al. (2014) using Community Multiscale Air Quality (CMAQ) model. Their study showcases the utility of CMAQ as a tool for developing and evaluating regulatory strategies aimed at reducing air pollutant emissions. Wang et al. (2014) employed the Nested Air Quality Prediction Model System (NAQPMS) to examine the spatio-temporal dynamics of PM2.5 concentrations in the troposphere over central eastern China during January 2013. Pouyaei et al. (2021) developed a Kain–Fritsch convection module within CMAQ model to improve the representation of convective mixing of atmospheric pollutants, effectively capturing both direct and indirect impacts of convection on CO distribution across different altitudes. Wang et al. (2024) developed the Modular Emission Inventory Allocation Tool for the Community Multiscale Air Quality model (MEIAT-CMAQ), which enhances CMAQ performance by efficiently refining emission inventories through detailed spatial, temporal, and species allocation, including vertical distribution.

The successful application of knowledge-based models requires highly professional skills and faces several challenges (Cobourn, 2010). Due to the use of ideal theory in model structure determination, these models often make simplifying assumptions about complex atmospheric processes that may not fully capture real-world conditions. The estimation of model parameters largely relies on empirical experience rather than precise measurements, introducing subjective uncertainties into the modeling process. Moreover, the heavy dependence on expert knowledge and experience-based parameter estimations can lead to inconsistent results across different implementations. As a result of these limitations, the predictive performance of knowledge-based models is often limited (Pak et al., 2020; Stern et al., 2008; Vautard et al., 2007; Zhang et al., 2022). In addition, chemical and transportation regulations may change as the models are applied to different situations, requiring continuous updates to model parameters and assumptions to maintain accuracy and relevance for specific regional contexts.

2.1.2 Data-driven models

A data-driven model is a model that learns patterns or makes predictions based on data, rather than being programmed with fixed rules. It uses historical or observed data to find relationships and make decisions. Data-driven models are recognized by many researchers because the models alleviate sophisticated theories. Research has shown that data-driven models perform better than knowledge-based models (Lv et al., 2016; Zhou et al., 2019). Data-driven models are composed of three categories: classical statistical models, traditional machine learning models, and deep learning models.

Classical statistical models, such as the Autoregressive Integrated Moving Average model (ARIMA), Seasonal Autoregressive Integrated Moving Average (SARIMA), and Geographically Weighted Regression (GWR), are essential tools in time series forecasting. These models analyze and predict future data points based on historical data. Hu et al.

(2013) utilized the GWR model that integrated aerosol optical depth, meteorological parameters, and land use information, achieving higher accuracy in estimating PM2.5 concentrations. Gocheva-Ilieva et al. (2014) applied factor analysis and SARIMA modeling to analyze and forecast concentrations of six air pollutants in Blagoevgrad, Bulgaria. Zhang et al. (2018) employed the ARIMA model specifically tailored to forecast PM2.5 concentrations in Fuzhou. The results demonstrated a high degree of fit between the actual and predicted values. Liu et al. (2021) proposed a combined prediction model, Principal Component Regression–Support Vector Regression–Autoregressive Moving Average (PCR–SVR–ARMA), which integrates principal component regression, support vector regression, and autoregressive moving average to improve air quality forecasting accuracy.

However, classical statistical models often face challenges with time series data due to issues like collinearity and non-stationarity, which can reduce their accuracy. These models may not adequately capture the complex, non-linear interactions present in atmospheric processes (Chaloulakou et al., 2003). Missing data and outliers in air quality time series present significant challenges for classical statistical approaches. These issues can bias parameter estimates and reduce model robustness (Zhang et al., 2024). Despite such limitations, classical methods maintain their popularity due to their simplicity and interpretability, and they often serve as a baseline to compare more complex models. The transparent methodology and computational efficiency of these approaches make them particularly valuable for initial data analysis and rapid assessment of air quality trends.

Traditional machine learning methods for time series prediction typically rely on manually engineered features, such as those used in models like Random Forests (RF), Support Vector Machines (SVM), and Artificial Neural Networks (ANN). Yu et al. (2016) introduced an innovative RF-based approach that significantly improved the precision of urban air quality prediction through effective utilization of urban sensing data. Leong et al. (2020) proposed a model using SVM and a radial basis function kernel to accurately and efficiently predict the air pollution index. He et al. (2022) enhanced monthly PM2.5 prediction accuracy by integrating Bayesian regularization with ANN modeling. Their approach achieved a correlation coefficient of 0.9570 and effectively identified complex variable relationships in Liaocheng.

These traditional machine learning methods demonstrate several notable advantages in air quality prediction. They excel at extracting non-linear characteristics from raw data and provide interpretable results through transparent model structures. The methods also show robust performance with limited datasets due to their efficient learning mechanisms and straightforward parameter tuning processes. However, these approaches face significant limitations in their predictive capabilities. When confronted with large-scale datasets, the models often exhibit reduced efficiency and accuracy. Their capacity to capture complex spatio-temporal correlations remains restricted, particularly in cases where air pollutants show intricate long-term dependencies and complex interactions with meteorological factors (Yan et al., 2021; Zhang et al., 2020).

2.2 Deep Learning Models for Sequence Prediction

Deep learning, a subfield of machine learning, has emerged as the primary focus of research. Its algorithms employ a multi-layer structure to extract essential features and patterns from massive datasets. Recurrent Neural Networks (RNN), Transformers, and their variants have demonstrated exceptional capabilities in sequence prediction tasks (Ahmed et al., 2023; Al-Selwi et al., 2024; Sun et al., 2024; Wen et al., 2022).

2.2.1 The RNN model and its variants

Elman (1990) introduced RNN for sequence prediction tasks and demonstrated the achievement of implicit temporal representation in connectionist models through recurrent connections. This approach enables networks to develop dynamic memory through a feedback mechanism of hidden unit patterns, which creates rich, context-dependent internal representations. These representations effectively combine task processing with memory requirements, establishing RNN as a powerful tool for sequential data analysis.

The fundamental strength of RNN lies in their recurrent network architecture. Internal connections provide various capabilities for temporal data processing and memory retention (Yu et al., 2019). However, despite these advantages and significant success in handling sequential data, RNN exhibit critical limitations in practical applications. The model particularly struggles with long-term data memorization and faces fundamental challenges of gradient explosion and gradient vanishing (Hochreiter & Schmidhuber, 1997). These limitations have necessitated substantial improvements to the original RNN structure, leading to the development of various advanced variants. The advancement of RNN can be examined from two main perspectives: improvements in cell structure and enhancements in network architecture. These developmental directions specifically target the limitations of traditional RNN while preserving their fundamental advantages in sequential data processing.

Hochreiter and Schmidhuber (1997) proposed the LSTM to tackle the issue of "long-term dependencies". This innovation improves the memory capacity of standard recurrent cells by integrating a "gate" mechanism into the cell architecture. The original LSTM cell design incorporates an input gate and an output gate. Gers et al. (2000) enhanced the original LSTM design by adding a forget gate to the cell structure. The LSTM memory cell, typically referring to the version with a forget gate, incorporates three gates: the forget gate, input gate, and output gate. This structure enables the cell to selectively forget, retain, and output temporal information.

The Gated Recurrent Unit (GRU), introduced by Cho (2014), is a variant of LSTM designed to address the problem of slow convergence. The GRU simplifies the LSTM structure by integrating the forget gate and input gate into a single update gate, resulting in only two gates: an update gate and a reset gate. This reduction in parameters aims to improve computational efficiency. Chung et al. (2014) demonstrated that gated units like LSTM and GRU outperform traditional RNN, with GRU showing similar performance to LSTM in music and speech modeling tasks. However, in some scenarios, GRU may be less effective than LSTM in extracting features from time series data (Britz et al., 2017; Weiss et al., 2018). Nina and Rodriguez (2015) introduced a simplified LSTM unit that combines the input and forget gates into a single gate, improving accuracy in the semantic description of images. Unlike GRU, which merges the hidden and cell states and removes the output gate, the simplified LSTM retains the output gate and the separate cell state. Dey and Salem (2017) proposed three variants of the GRU (GRU1, GRU2, and GRU3), which perform as well as the original GRU model while reducing computational costs by systematically reducing parameters in the update and reset gates. Neil et al. (2016) introduced the Phased LSTM model, which enhances the standard LSTM by adding a time gate controlled by an oscillation. This innovation allows the model to process irregularly sampled data more efficiently. Moniz and Krueger (2017) presented Nested LSTM (NLSTM), a unique LSTM architecture that achieves greater depth by nesting memory cells, proving to be highly effective in character-level language modeling tasks compared to traditional LSTMs. Yadav and Thakkar (2024) proposed a novel model, No Output Activation LSTM (NOA-LSTM), which introduces an efficient LSTM cell architecture by modifying the traditional LSTM cell to remove the output activation function, thereby improving forecasting accuracy and training efficiency for time series data.

In addition to improvements in cell structure, advancements have also been made in network architecture. Researchers have developed neural networks primarily composed of cells, focusing on optimizing the connections between these cells to enhance overall network properties. The networks aim to leverage the strengths of cells while improving the ability to process complex sequential data through refined architectural designs (Nie et al., 2016; Zhao et al., 2016).

Schuster and Paliwal (1997) introduced Bidirectional Recurrent Neural Networks, which process information in both forward and backward time directions. It achieves enhanced performance in regression and classification tasks. Graves and Schmidhuber

(2005) introduced Bidirectional Long Short-Term Memory (Bi-LSTM) networks along with a full-gradient version of the LSTM learning algorithm. Their results show that Bi-LSTM networks outperform unidirectional models. They are faster and more accurate than standard RNN and time-windowed Multi-Layer Perceptron (MLP) in phoneme classification. This highlights the critical role of contextual information in speech processing. In a similar vein, Shi et al. (2015) extended the LSTM by introducing new structures in the Convolutional LSTM (ConvLSTM) model, which captures spatio-temporal correlations more effectively and significantly boosts accuracy over traditional methods. Zhu et al. (2015) extended the conventional chain-structured LSTM to a tree-structured model called Structural Long Short-Term Memory (S-LSTM), enabling the model to capture long-distance dependencies across hierarchical structures. Their approach demonstrated superior performance over state-of-the-art recursive models in semantic composition tasks for natural language understanding. Liu et al. (2016) proposed a deep architecture based on two coupled-LSTM to model strong interactions between sentence pairs, demonstrating that it outperforms state-of-the-art methods on large datasets. Mai et al. (2021) proposed the Acoustic and Visual LSTM with Channel Aware Temporal Convolution Network (AV-LSTM-CA-TCN) model, which enhances language representations using modality-specific LSTM variants with gated mechanisms and introduces a channel-aware temporal convolution network to capture both temporal and channel-wise interdependencies for improved multimodal sentiment analysis.

2.2.2 The Transformer model and its variants

The Transformer model, a novel method of sequence modeling, is presented in the publication "Attention is All You Need" by Vaswani et al. (2017). Unlike RNN and its variants, the Transformer eliminates the need for recurrent layers by relying entirely on self-attention mechanisms. It models sequence information through positional encoding, which is added to the input embeddings, allowing the model to capture the order of the sequence without recurrence. The Transformer employs an encoder-decoder structure with multiple identical layers. The process begins with embedding layers that convert input tokens into vectors and add positional information. Each encoder layer contains a multi-head self-attention mechanism and a position-wise feed-forward network. Decoder layers have a similar structure but include an additional cross-attention mechanism between the self-attention and feed-forward components, allowing the decoder to incorporate the output of the encoder. The model concludes with a linear layer and softmax function to generate predictions. As a result, the computational efficiency of the model is much increased, and it can handle entire sequences in parallel.

The Transformer has demonstrated better performance in several domains, such as natural language processing, computer vision, speech, and more recently, time series analysis. It is excellent for sequential modeling problems because of its attention mechanism, which can recognize correlations between elements in a sequence automatically (Nie et al., 2022). To enhance the capability of the Transformer in sequence forecasting tasks, researchers have proposed various modifications to the original architecture. These improvements can be broadly categorized into three main approaches: architectural component modification, intrinsic processing mechanism optimization, and other novel approaches.

The first and most common approach focuses on adapting key architectural components, primarily targeting attention mechanisms and positional embedding. Regarding attention mechanism improvements, several innovative approaches have emerged. Zhou et al. (2021) introduced Informer, an efficient Transformer-based model for long sequence time series forecasting. It features a novel ProbSparse self-attention mechanism that reduces complexity, significantly improving performance and efficiency in handling long-term dependencies. Wu et al. (2021) proposed Autoformer, a novel time series forecasting model that replaces traditional self-attention with an auto-correlation mechanism. Liu et al. (2022) introduced Pyraformer, which utilizes a novel pyramidal attention mechanism with a multi-resolution tree structure. The model achieves linear complexity while significantly improving accuracy on long-term forecasting tasks.

In terms of positional embedding improvements, researchers have developed various methods to enhance temporal representation. Kazemi et al. (2019) proposed the Time2Vec embedding method, which encodes time using detailed calendar components such as year, month, day, season, hour, and minute. Building upon this concept, Shankaranarayana and Runje (2021) introduced a timestamp embedding technique that captures both periodic cycles and aperiodic trends by decomposing timestamps into multiple components. Deihim et al. (2023) further advanced this field by combining position embedding with Time2Vec embedding. Another significant development direction involves fully learnable positional encodings. Zerveas et al. (2021) developed a framework where embedding vectors for each position index are learned jointly with other model parameters. Taking a different approach, Lim et al. (2021) utilized an LSTM network to encode positional embeddings within the Transformer model.

The second approach focuses on fully leveraging the capabilities of the Transformer model, placing greater emphasis on the intrinsic processing of time series data. Kim et al. (2021) designed a flexible and efficient Reversible Instance Normalization (RevIN) layer, which is generally applicable for accurate time series forecasting, particularly in addressing distribution shift issues. Liu et al. (2022) proposed Non-stationary Transformers, incorporating series stationarization and de-stationary module to boost the performance of mainstream Transformers. Han et al. (2024) discussed channel-dependent and channel-independent strategies and introduced a modified channel-dependent method, called Predict Residuals with Regularization, to enhance the prediction accuracy of Transformer-based models. Nie et al. (2022) introduced a Transformer-based model for multivariate time series forecasting, which segments time series into patches and uses a channel-independent approach to improve long-term forecasting accuracy and self-supervised representation learning.

Beyond the previously mentioned ideas, researchers have explored various novel perspectives to enhance Transformer models. Zhang and Yan (2023) advanced the architectural design of Transformer models through the introduction of Crossformer.

This model incorporates two key innovations: the Dimension-Segment-Wise embedding and the Two-Stage Attention mechanism. These design elements effectively capture cross-dimension dependencies in multivariate time series data and lead to superior performance compared to previous state-of-the-art models. Liu et al. (2023) introduced an inverted Transformer (iTransformer), a novel approach that applies attention and feed-forward networks to inverted dimensions, capturing multivariate correlations and learning nonlinear representations.

2.3 Deep Learning Models for Temporal Air Quality Prediction

Deep learning has emerged as a powerful approach to processing sequential data, including time series prediction tasks. This field has been primarily driven by two architectural innovations: RNN and Transformer models, both specifically designed to capture temporal dependencies and sequential patterns. The successful applications of these deep learning approaches span across numerous domains, from traffic flow prediction (Huang et al., 2022) and industrial soft sensor development (Yuan et al., 2020) to flood forecasting (Ding et al., 2020), tourism demand forecasting (Li et al., 2022), and power system applications (Dai et al., 2022). In the specific context of air quality prediction, these deep learning approaches have consistently demonstrated superior performance (Jin et al., 2021; Xu & Yoneda, 2019). The following sections provide a detailed examination of how these two architectural paradigms (RNN and Transformer) have been adapted and applied to air quality temporal prediction tasks. Figure 2.2 presents the taxonomy of deep learning models for temporal air quality prediction.

2.3.1 RNN-based models for temporal air quality prediction

Recent advancements in deep learning techniques have led to significant improvements in air quality prediction models, with researchers exploring various RNN architectures to enhance prediction accuracy. The development of RNN-based air quality prediction models has progressed along four main directions. First, studies demonstrate the direct application of existing RNN variants, such as LSTM and GRU, to

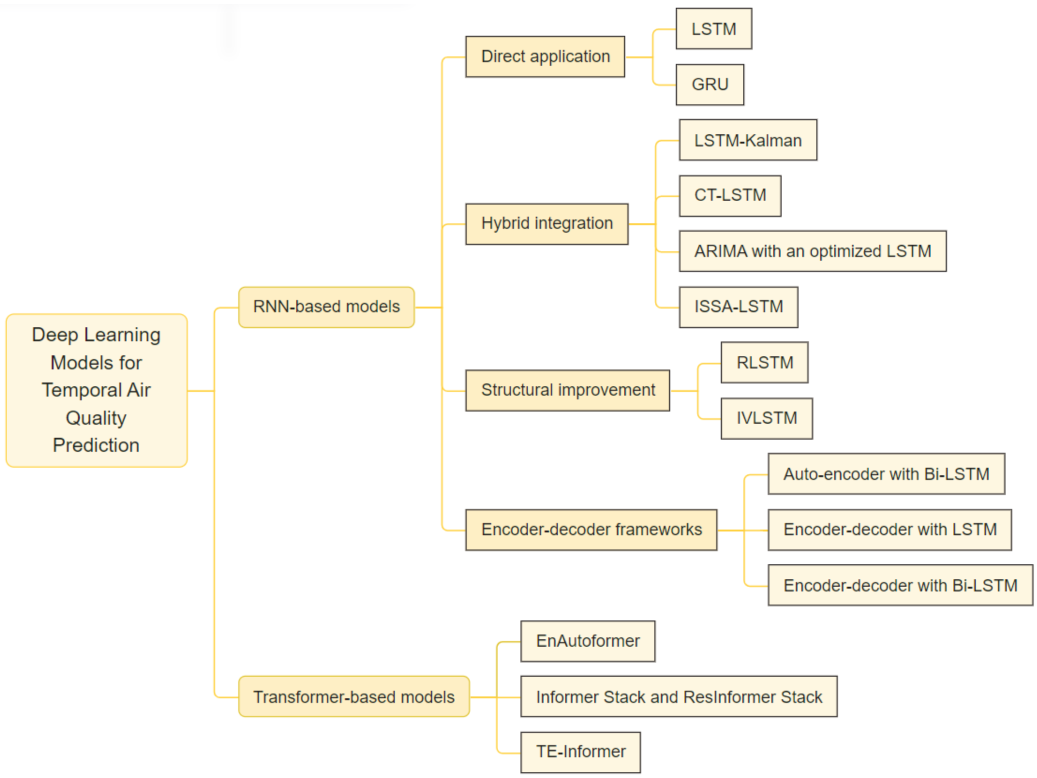

Figure 2.2: Taxonomy of deep learning models for temporal air quality prediction

air quality forecasting tasks. Second, researchers combine RNN with other techniques to create hybrid models that better capture temporal patterns, including statistical methods and optimization algorithms. Third, studies focus on structural improvements to RNN architectures themselves, developing novel variants specifically tailored for air quality prediction. Additionally, the incorporation of encoder-decoder architectures represents a significant advancement in this field, offering enhanced capabilities in processing air quality data and capturing complex temporal dependencies.

Some studies focus on the direct application of existing RNN variants for air quality prediction. Athira et al. (2018) developed DeepAirNet, a comprehensive framework that evaluates different RNN architectures including standard RNN, LSTM, and GRU for PM10 concentration forecasting based on air pollutants and meteorological time series data. Loy-Benitez et al. (2019) presented a multiple sequence prediction model using GRU networks for indoor PM2.5 forecasting, showing superior performance over standard RNN and LSTM models. Rao et al. (2019) developed an air quality prediction model using RNN with LSTM to forecast hourly pollutant concentrations.

Meanwhile, many studies actively improve RNN through hybrid approaches. Song et al. (2019) introduced the LSTM-Kalman model for air quality prediction, combining LSTM with Kalman filtering. This hybrid approach utilizes LSTM to capture long-term patterns in historical data, while Kalman filtering dynamically adjusts the LSTM-processed time series, resulting in more accurate air quality forecasts compared to standard LSTM models. Wang et al. (2021) introduced CT-LSTM, a novel method for air quality prediction that combines the chi-square test and LSTM. In this approach, chi-square test is employed to identify the key factors influencing air quality, while the LSTM network processes these factors to generate forecasts. Gunasekar et al. (2022) developed a hybrid model combining ARIMA with an optimized LSTM network. Wu et al. (2023) developed the ISSA-LSTM model for accurate air quality index (AQI) prediction. This model consists of three main components. First, a feature selection module combines RF with the maximum relevance minimum redundancy algorithm to identify influential variables. Next, an optimization module employs an Improved Sparrow Search Algorithm (ISSA) to tune LSTM hyperparameters. Finally, a prediction module utilizes the optimized LSTM for final AQI concentration forecasting.

Studies are also focused on improving the structure of RNNs themselves. Zhang et al. (2021) developed the novel EDSModel for air pollutant prediction. The model combines Read-first Long Short-Term Memory (RLSTM) and traditional LSTM structures to effectively extract temporal correlation features from historical pollutant and meteorological data. Fang et al. (2023) introduced the Improved Vanilla LSTM (IVLSTM), an optimized variant of the LSTM tailored for air quality prediction. This model enhances the performance of the LSTM by incorporating several key improvements, notably by making the forget and input gates complementary, which effectively reduces the number of parameters.

Building upon RNN approaches, researchers have further refined air quality prediction models by incorporating encoder-decoder architectures. These architectures encode observed data into a fixed-length internal representation before decoding it into final outputs, offering enhanced capabilities in processing sequential data and capturing complex temporal dependencies.

Zhang et al. (2020) developed an advanced deep learning model for PM2.5 concentration forecasting, integrating an auto-encoder with a Bi-LSTM network. The architecture of model is structured into three key components: a data preprocessing stage, an auto-encoder layer for feature extraction and dimensionality reduction, and a Bi-LSTM layer for capturing temporal dependencies in both forward and backward directions. Du et al. (2020) introduced a novel temporal attention encoder-decoder model for multivariate time series forecasting. The model employs Bi-LSTM layers enhanced with a temporal attention mechanism in the encoder network. Kristiani et al. (2021) achieved superior results in predicting PM2.5 indicators by utilizing the LSTM encoder-decoder combined with statistical methods, such as correlation analysis. Jia et al. (2021) introduced an encoder-decoder architecture based on GRU for ozone concentration prediction. Nguyen et al. (2021) developed an innovative approach for accurate PM2.5 prediction by combining a genetic algorithm with an encoder-decoder model utilizing LSTM. The genetic algorithm optimizes feature selection and removes outliers, while the encoder-decoder model with LSTM effectively predicts PM2.5 concentrations.

In summary, the evolution of RNN-based approaches in air quality prediction demonstrates significant progress in addressing the challenges of temporal forecasting. From the direct application of traditional RNN architectures to the development of sophisticated hybrid models and structural improvements, researchers have continuously enhanced the capability of these models to capture complex temporal patterns in air quality data. The integration of complementary techniques, such as statistical methods and optimization algorithms, has further improved the robustness and accuracy of predictions. Notably, the emergence of encoder-decoder architectures represents a significant advancement, offering enhanced abilities to process sequential data and capture long-term dependencies.