Literature review

The endeavor to forecast the Air Quality Index (AQI) within urban environments, especially in cities such as Delhi, is crucial for comprehending and mitigating the impact of air pollution on public health and the environment. Numerous studies have investigated a variety of modeling approaches and techniques to predict AQI levels, encompassing linear and nonlinear methods to deep learning architectures.

Linear and nonlinear modeling approaches have been thoroughly examined in the context of urban air quality prediction. For instance, research indicated in reference [4] assessed the efficacy of linear modeling techniques like Partial Least Squares Regression (PLSR) against nonlinear methods including Multivariate Polynomial Regression (MPR) and Artificial Neural Networks (ANNs). This study discovered that ANNs surpassed linear models in accuracy, with the Generalized Regression Neural Network (GRNN) demonstrating high correlations for pollutants such as Respirable Suspended Particulate Matter (RSPM), Nitrogen Dioxide (NO2), and Sulfur Dioxide (SO2). Moreover, study [5] concentrated on forecasting AQI using regression models, identifying Support Vector Regression (SVR) as exhibiting superior performance compared to linear models. These models were evaluated using statistical criteria like Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE).

Furthermore, deep learning approaches have emerged as potent tools in AQI forecasting due to their capacity to identify complex temporal patterns. Study [6] employed a deep learning model based on a Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) to predict Tropospheric Ozone (O3) concentrations, acknowledging the complexity of O3 formation processes and sources, and highlighted the appropriateness of RNNs for such intricate predictions. Additionally, research [10] introduced a spatio-temporal deep learning model for air quality prediction, which showed enhanced performance over traditional time series models like Autoregressive Integrated Moving Average (ARIMA) and Support Vector Regression (SVR). This study stressed the significance of considering spatial and temporal correlations in air quality prediction, capabilities effectively embodied by the proposed model.

The integration of time-series forecasting techniques for AQI prediction has attracted attention in recent research endeavors. Study [7] analyzed four time-series analysis methods, including SARIMAX, ARIMA, AR, and LSTM, for forecasting logistics companies’ staffing needs and order volume, finding SARIMAX to excel in order volume predictions, thus underlining its utility in resource planning and management within logistics sectors. Additionally, reference [8] provided an exhaustive literature review on time series forecasting methods, discussing various techniques such as Autoregression (AR), Moving Average (MA), ARIMA, and LSTM, and emphasized the critical role of stationarity in time series modeling and the efficacy of different methods in AQI forecasting.

In summation, the time series forecasting of AQI in urban locales necessitates an integrative methodology that merges linear and nonlinear modeling techniques, deep learning architectures, and advanced time series analysis methods. The insights derived from these studies contribute significantly to the accurate prediction of AQI levels, thereby facilitating improved air quality management and safeguarding public health.

III. METHODOLOGY USED IN THIS PAPER

III. Methodological Framework of This Study

The comprehension of time series data's behavior through temporal progression is paramount. This involves determining the "stationarity" of the dataset, utilizing two distinct tests: the Augmented Dickey-Fuller (ADF) and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests. ADF assesses the presence of unit roots that could disrupt data consistency, whereas KPSS evaluates the constancy of the data's trend. Given their divergent focal points, these tests offer complementary rather than interchangeable insights.

A. Conventional Analytical Techniques

The investigation commences with traditional stalwarts—ARIMA and exponential smoothing. These methodologies stand as foundational pillars in the domain of time series analysis, offering robust analytical tools.

B. Fundamental Machine Learning Approaches

The inquiry extends into the realm of basic machine learning (ML) strategies, spotlighting the adaptability of linear regression and decision trees. These models excel in versatility, capable of navigating a variety of scenarios with efficacy in prediction accuracy, computational speed, and interpretability in practical contexts.

However, these models encounter limitations amidst complex or voluminous datasets, challenging their performance and practical applicability. This segment of the study not only evaluates their predictive capabilities but also examines their resilience against intricate data structures and their feasibility in real-world forecasting scenarios. The objective is to thoroughly assess the efficacy of these conventional ML techniques within the intricate landscape of time series forecasting.

C. Deep Learning Techniques

C. Advanced Deep Learning Methods

Progressing to a more sophisticated segment of our exploration, we venture into the domain of deep learning (DL) models. This segment is characterized by the deployment of elementary deep neural networks (DNN), alongside the more formidable Recurrent Neural Networks (RNNs) and Long Short-Term Memory Networks (LSTMs). Our objective is to tackle the complex patterns within the AQI dataset, utilizing these avant-garde technologies.

In navigating this advanced technological landscape, our primary aim is to evaluate the capabilities of these DL techniques. We are particularly interested in their ability to surpass traditional models in handling the complexities and non-linearities inherent in real-world time series data. This exploration is akin to unraveling the layers of complexity, to gauge how deep learning models can enhance our analytical prowess. While these DL models are endowed with significant power, they also present challenges, including substantial computational demands and a prerequisite for extensive data. Hence, our exploration is balanced, weighing their potential against the challenges they pose. This phase introduces a new dimension of sophistication to our analytical toolkit, keenly observing the contributions of these advanced models.

D. Novel Advancements

Advancing beyond the conventional paths of classical and machine learning models, our investigation delves into the latest advancements in time series forecasting. This journey is not limited to established methods; instead, it extends to exploring the frontier of forecasting innovations. This stage examines groundbreaking techniques such as Time GPT, Patch TST, N-hits, N-beats, TimesNet, TSmixer, and Liquid Neural Networks, aiming to discern how these novel methods redefine forecasting capabilities.

Liquid Neural Networks, in particular, represent a fresh paradigm in adaptive systems, designed for flexibility in accommodating future data through an ODE solver. This feature dynamically adjusts the network’s weights during inference, offering a glimpse into adaptive learning capabilities.

While these new methodologies are intriguing, we scrutinize their limitations, assessing their practicality and potential superiority over traditional forecasting tools. This analysis is not merely focused on novelty but on determining whether these innovations can truly surpass the efficacy of established time series forecasting methods.

IV. Data Set Exploration

A. Dataset Overview and Primary Attributes - AQI Dataset

This sub-section describes the dataset Air Quality Index(AQI). This data set consists of AQI data collected over days spanning 1 to 5 years from 2015-2020 for various metropolitan cities, namely Amaravati, Chandigarh, Hyderabad, Kolkata, Patna, Delhi, Amritsar, Gurugram, Vishakhapatnam, allowing a complete representation of the entire 5 years AQI data.This data set comprises of the attributes such as: City name, Date of measurement, Respirable Suspended Particulate Matter (RSPM) or Particulate Matter (PM10), Particulate matter(PM2.5), Nitric Oxide(NO),Nitrous oxide(NO2), nitrogen oxides(NOx), Ammonia(NH3), Carbon Oxide(CO), Sulphur Dioxide(SO2),Ozone(O3), Benzene, Toluene, Xylene, AQI, AQI Bucket. In handling missing data, we have chosen to discard records where significant attributes, such as AQI Bucket and AQI, are empty for the specified time periods, and consequently, those columns have been dropped. However, certain periods and cities have empty data records due to various reasons. For instance, in Ahmedabad, data from November 2015 to May 2016 and November 2016 to October 2017 is missing. In Amravati, there are gaps in the data for August 2019 to October 2019. Patna’s data records were empty from April 2017 to September 2017, while Vishakhapatnam’s data was unavailable from April 2017 to October 2017. Additionally, for some cities like Chandigarh, the data is available from 2019 onwards, not from 2015. Similarly, data for the city of Kolkata is available from 2018.

B. Analytical Observations

Visual analysis of the RSPM/PM10 data reveals varying densities across years, with no clear patterns emerging within or across the data from 2012 to 2015, suggesting the need for mathematical testing to identify underlying trends. Similarly, the analysis of SO2 and NO2 data indicates a lack of discernible patterns, necessitating appropriate preprocessing to render the dataset suitable for time series analysis.

Data preparation

The process of data preparation involves dividing the dataset into two subsets: one comprising readings from 2012 and 2013, and the other containing readings from 2014 and 2015. This division aims to ensure balance between the sets in terms of the number of observations. Subsequently, it is essential to ascertain whether the data exhibits "stationarity," a condition indicating independence from time and the absence of discernible patterns over time. Stationarity implies freedom from temporal constraints, facilitating more straightforward predictive modeling. However, achieving stationarity is often challenging, particularly in environmental datasets like the one under consideration, namely the Air Quality Index (AQI) dataset. To assess stationarity, two widely employed statistical tests are utilized: the Augmented Dickey-Fuller Test (ADF) and the Kwiatkowski-Phillips-Schmidt-Shin test (KPSS).

The ADF test evaluates stationarity by examining whether a series possesses a unit root, a characteristic prevalent in non-stationary stochastic processes such as random walks. The null hypothesis (HO) posits that the series is non-stationary, while the alternate hypothesis (HA) suggests stationarity. Rejecting the null hypothesis occurs when the test statistic is less than the critical value and the p-value is below 0.05, indicating stationarity. Conversely, failing to reject the null hypothesis implies non-stationarity. Consequently, each unique column reading in the bifurcated dataset must exhibit a p-value below 0.05 for stationarity.

In contrast, the KPSS test operates under the assumption that the series is stationary. The null hypothesis (HO) asserts trend stationarity, while the alternate hypothesis (HA) suggests non-stationarity. Contrary to the ADF test, rejecting the null hypothesis in the KPSS test indicates non-stationarity, while failing to reject it signifies trend stationarity. Similar to the ADF test, a p-value below 0.05, accompanied by a test statistic lower than the critical value, indicates stationarity. Therefore, these tests provide crucial insights into the temporal characteristics of the dataset, facilitating subsequent analysis and modelling endeavours.