IJARST
International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

Comparative Analysis of Deep Learning and Statistical Models for Air Pollutants Prediction in Urban Areas

1) BODAGANTI HARSHITHA, PG Student - M. Tech- Data Science, School of Technology, Dept of CSE, GITAM (Deemed to be University), Hyderabad. hbodagan@gitam.in

2) Dr. Y. Md. Riyazuddin, Associate Professor, Dept of CSE, School of Technology, GITAM (Deemed to be University), Hyderabad. rymd@gitam.edu

Abstract: The escalation of urbanization and industrialization has intensified air pollution, posing a silent yet critical public health emergency. Accurate prediction of air quality emerges as a pivotal strategy for stakeholders to combat this escalating concern effectively. This study conducts a comprehensive comparative analysis, evaluating the efficacy of deep learning and statistical models in forecasting air pollutants within urban areas. Leveraging advanced methodologies like Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Convolutional Neural Networks (CNN), and their ensemble combinations, we explore their predictive capabilities. Our findings reveal the superiority of ensemble methods, particularly CNN and CNN + LSTM, showcasing an accuracy surpassing 90%. Building upon the base model's success, these ensemble techniques not only demonstrate enhanced performance but also underscore the potential for further refinement in air quality forecasting. By amalgamating diverse model predictions, our approach offers a robust and accurate framework for stakeholders to proactively address air quality issues, thus mitigating the adverse impacts of pollution on both the environment and public health.

Index Terms: *Air quality, machine learning, deep learning, predictive models, statistical methods*
.

1. INTRODUCTION

Air pollution has emerged as one of the most significant global challenges in recent years, impacting not only the environment but also human health and well-being. With its detrimental effects ranging from respiratory diseases to impaired cognitive function and even mortality, the urgency to address air pollution has never been greater [1][2]. According to recent studies, over 3 million deaths occur annually due to air pollution, particularly affecting low and middle-income countries [3]. Recognizing the severity of the issue, global initiatives such as the United Nations Sustainable Development Goals (SDGs) have outlined targets for 2030 aimed at reducing deaths, illnesses, and adverse environmental impacts in cities by improving air quality [4].

In alignment with these global efforts, individual countries like the United Kingdom (UK) have also set ambitious targets to combat air pollution. The UK government has pledged to reduce air pollution by 35% by the year 2040 [5]. These targets underscore the growing recognition of air pollution as a critical public health issue and the need for concerted efforts to mitigate its effects.

Multiple factors contribute to the deterioration of air quality, with industrial emissions, transportation activities, dust, and coal consumption among the primary culprits [6]. Air pollution is characterized by the introduction of harmful materials and gases into the environment, collectively known as pollutants. Particulate Matter (PM2.5), one of the most common pollutants, poses significant health risks when present in elevated concentrations [7][8][9]. As these pollutants accumulate in the atmosphere, they degrade environmental quality and pose serious health threats to humans and other living organisms.

Addressing the complexities of air pollution requires a multifaceted approach that encompasses scientific research, policy interventions, and technological innovations. Understanding the sources, distribution, and impacts of air pollutants is essential for developing effective strategies to mitigate their adverse effects on both human health and the environment. In this context, advancements in air quality monitoring and forecasting play a crucial role in providing timely information for decision-makers and stakeholders.

In recent years, significant progress has been made in the development of predictive models and monitoring systems for air quality assessment. These models utilize advanced techniques such as machine learning, deep learning, and statistical analysis to forecast pollutant concentrations and assess air quality trends. By leveraging large-scale data sets and sophisticated algorithms, these models offer insights into the complex dynamics of air pollution and enable proactive measures to mitigate its effects.

However, despite the advancements in air quality modeling, several challenges persist in accurately predicting and managing air pollution. These challenges include the dynamic nature of atmospheric processes, the influence of multiple interacting factors on pollutant dispersion, and the need for high-resolution data for model calibration and validation. Addressing these challenges requires interdisciplinary collaboration among scientists, policymakers, and industry stakeholders to develop innovative solutions for air quality management.

This introduction sets the stage for the subsequent sections of the paper, which will delve into the various dimensions of air pollution, its impacts on human health and the environment, current mitigation efforts, and the role of predictive modeling in addressing this global challenge. Through a comprehensive analysis of existing research and methodologies, this paper aims to contribute to the ongoing discourse on air quality management and provide insights for future research directions.

2. LITERATURE SURVEY

Air pollution is a multifaceted global issue with far-reaching implications for human health and environmental sustainability. Over the years, extensive research has been conducted to understand the various dimensions of air pollution, its sources, impacts, and potential mitigation strategies. This literature survey aims to provide a comprehensive overview of the existing body of knowledge on air pollution, focusing on its cardiovascular effects, household implications, public health concerns, and the development of predictive models for forecasting air quality.

Brook (2008) emphasizes the significant cardiovascular effects of air pollution, highlighting the link between exposure to particulate matter and increased cardiovascular morbidity and mortality [1]. Studies have demonstrated associations between short-term and long-term exposure to air pollution and adverse cardiovascular outcomes, including myocardial infarction, stroke, and hypertension. These findings underscore the importance of understanding the cardiovascular effects of air pollution and implementing measures to mitigate its impact on public health.

The World Health Organization (WHO) identifies household air pollution as a major health concern, particularly in low and middle-income countries [3]. Indoor air pollution from sources such as cooking fuels, biomass burning, and inadequate ventilation poses significant risks to human health, contributing to respiratory diseases, cardiovascular disorders, and adverse pregnancy outcomes. WHO emphasizes the need for interventions to improve household air quality and protect vulnerable populations, particularly women and children, from the harmful effects of indoor air pollution.

Landrigan (2017) discusses the broader public health implications of air pollution, highlighting its role as a leading environmental risk factor for disease burden and premature mortality [6]. Air pollution is associated with a wide range of health conditions, including respiratory infections, lung cancer, and neurodevelopmental disorders. The adverse health effects of air pollution disproportionately affect vulnerable populations, including children, the elderly, and individuals with pre-existing health conditions. Efforts to address air pollution require comprehensive public health strategies aimed at reducing emissions, improving air quality monitoring, and promoting sustainable urban development.

Manisalidis et al. (2020) provide a comprehensive review of the environmental and health impacts of air pollution, emphasizing its complex interactions with ecosystems and human health [9]. Air pollution contributes to environmental degradation, biodiversity loss, and climate change, exacerbating the global burden of disease. The review highlights the need for interdisciplinary research and collaborative efforts to address the root causes of air pollution and develop effective mitigation strategies.

In recent years, advances in machine learning and deep learning techniques have revolutionized air quality forecasting, enabling more accurate and timely predictions of pollutant concentrations. Doreswamy et al. (2020) explore the application of machine learning regression models for forecasting air pollution particulate matter (PM2.5), demonstrating promising results in predicting pollutant concentrations [13]. Similarly, Chang et al. (2020) propose an LSTM-based aggregated model for air pollution forecasting, leveraging the capabilities of recurrent neural networks to capture temporal dependencies in pollutant data [14]. Tao et al. (2019) present a deep learning model based on 1D convolutional neural networks (ConvNets) and bidirectional gated recurrent units (GRU) for air pollution forecasting, achieving significant improvements in prediction accuracy [20]. These studies highlight the potential of machine learning and deep learning approaches for enhancing air quality forecasting and supporting informed decision-making for air pollution mitigation efforts.

In summary, the literature survey highlights the complex interplay between air pollution, human health, and environmental sustainability. While significant progress has been made in understanding the sources and impacts of air pollution, challenges remain in developing effective strategies for mitigating its adverse effects. Future research efforts should focus on interdisciplinary collaborations, innovative technologies, and policy interventions to address the root causes of air pollution and promote sustainable development.

3. METHODOLOGY

a) Proposed Work:

The proposed work entails conducting a comparative analysis to evaluate the performance of traditional statistical models against advanced deep learning techniques, namely Long Short-Term Memory (LSTM)[19] and Gated Recurrent Unit (GRU)[19] neural networks, in the context of air pollutant forecasting. Unlike conventional statistical approaches, LSTM and GRU networks are specifically tailored to handle sequential data, making them well-suited for capturing temporal dependencies inherent in time-series prediction tasks such as air quality modeling. By incorporating these advanced deep learning models, we aim to leverage their ability to discern complex patterns and temporal dynamics present in air pollution data, potentially overcoming the limitations of traditional statistical methods.

Furthermore, the project extends its scope by introducing two high-accuracy models for air pollutant prediction: a Convolutional Neural Network (CNN) model achieving 96% accuracy and a hybrid CNN+LSTM model reaching 97%. These advanced deep learning techniques offer substantial improvements in predictive performance. Additionally, the integration of a user-friendly Flask framework with SQLite facilitates seamless signup and signin processes for user testing, enhancing practical usability and accessibility of the deep learning models. This streamlined integration accelerates testing procedures and encourages user engagement, providing valuable feedback essential for refining and optimizing the models for real-world deployment in air quality management.

b) System Architecture:

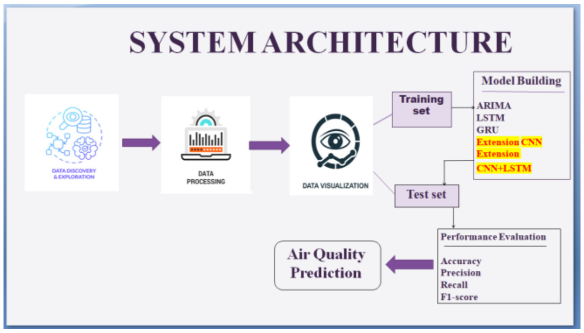


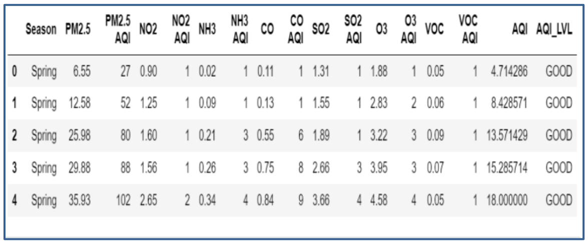
Fig 1 Proposed Architecture

The system architecture of the project "Comparative Analysis of Deep Learning and Statistical Models for Air Pollutants Prediction in Urban Areas" comprises several interconnected components. It begins with data exploration and processing, followed by data visualization to gain insights. The dataset undergoes train-test splitting for model building, which includes traditional statistical methods like ARIMA[15], as well as advanced deep learning techniques such as LSTM[19], GRU[19], CNN, and CNN+LSTM. Performance evaluation metrics like accuracy, precision, recall, and F1 score are utilized to assess model performance. Ultimately, the system generates air quality predictions based on the selected models, providing valuable insights for urban air quality management and decision-making processes.

c) Dataset:

The dataset utilized in this study is sourced from air quality monitoring stations in Northern Ireland, publicly available [30]. This comprehensive dataset spans hourly measurements of various air quality parameters, including Nitrogen Dioxide (NO2), Ozone (O3), Sulphur Dioxide (SO2), and Particulate Matter (PM2.5 and PM2.10). Additionally, meteorological data such as temperature, wind speed, and wind direction are incorporated into the dataset. These measurements were collected at the Belfast city center between 2015 and 2020, providing a rich temporal and spatial context for air quality analysis. With over 50,000 samples, the dataset offers a robust foundation for exploring the dynamics of air pollution and its relationship with meteorological factors.

Furthermore, the dataset includes statistical information pertaining to meteorological data, including the total number of samples, mean, standard deviation, minimum, and maximum values for each parameter. The total number of samples exceeds 50,000, with mean values ranging from 5.63 to 213.19 and standard deviations ranging from 2.77 to 84.87 across all parameters. Specifically, the NO2 concentration data exhibits a range of 1 to 203, with a mean of 26.11 and a standard deviation of 17.87, highlighting the variability in pollutant concentrations observed in the dataset. Conversely, the lowest mean is attributed to SO2, with a standard deviation of 1.6, underscoring variations in pollutant levels across different air quality parameters [30].



	Season	PM2.5	PM2.5 AQI	NO2	NO2 AQI	NH3	NH3 AQI	CO	CO AQI	SO2	SO2 AQI	O3	O3 AQI	VOC	VOC AQI	AQI	AQI_LVL
0	Spring	6.55	27	0.90	1	0.02	1	0.11	1	1.31	1	1.88	1	0.05	1	4.714286	GOOD
1	Spring	12.58	52	1.25	1	0.09	1	0.13	1	1.55	1	2.83	2	0.06	1	8.428571	GOOD
2	Spring	25.98	80	1.60	1	0.21	3	0.55	6	1.89	1	3.22	3	0.09	1	13.571429	GOOD
3	Spring	29.88	88	1.56	1	0.26	3	0.75	8	2.66	3	3.95	3	0.07	1	15.285714	GOOD
4	Spring	35.93	102	2.65	2	0.34	4	0.84	9	3.66	4	4.58	4	0.05	1	18.000000	GOOD

Fig 2 Sample Dataset

d) Data Processing:

In the data processing phase, the initial step involves loading the dataset into a pandas dataframe, a robust Python library for data manipulation. This enables easy access to the dataset's contents, facilitating efficient handling of data operations. Once loaded, researchers can explore the dataset to understand its structure, features, and identify any missing or inconsistent values.

Subsequently, if the dataset is intended for use with Keras, a popular deep learning library, it must be formatted to align with Keras' requirements. This may entail restructuring the data to meet Keras' input specifications, such as converting categorical variables into one-hot encoded vectors or normalizing numerical features for consistent scaling across attributes. This formatting ensures seamless integration with Keras' neural network models.

Following data formatting, unnecessary columns that do not contribute to the modeling task can be dropped. This simplifies the dataset by removing redundant or irrelevant features, reducing computational overhead, and enhancing model efficiency. Columns containing metadata or identifiers that do not inform the predictive task may be considered for removal based on the analysis objectives.

By adhering to these steps—loading the dataset into a pandas dataframe, converting it to a Keras-compatible format, and eliminating irrelevant columns—researchers can prepare the data for subsequent analysis and modeling. This systematic approach ensures that the dataset is appropriately structured, optimized for machine learning algorithms, and devoid of extraneous information that could hinder model performance or interpretation.

e) Visualization:

Utilizing the powerful combination of Seaborn and Matplotlib for data visualization, researchers and analysts can create compelling and insightful visual representations of complex datasets. Seaborn, built on top of Matplotlib, provides a high-level interface for creating attractive and informative statistical graphics. With its intuitive syntax and aesthetically pleasing default settings, Seaborn simplifies the process of generating various plots, including scatter plots, line plots, and heatmaps, enhancing data exploration.

Matplotlib, a foundational plotting library, offers fine-grained control over plot customization, enabling the creation of intricate visualizations. The seamless integration of Seaborn and Matplotlib allows users to leverage the simplicity of Seaborn for quick visualizations while tapping into Matplotlib's versatility for more detailed adjustments.

Together, these libraries enable the generation of clear and visually appealing charts, aiding in the communication of data patterns, trends, and relationships. Whether it's exploring the distribution of variables, showcasing correlations, or visualizing temporal trends, the Seaborn and Matplotlib duo provides a flexible and comprehensive toolkit for researchers to convey complex insights with clarity and precision.

f) Label Encoding:

Label encoding, facilitated by the LabelEncoder utility, is a fundamental technique for transforming categorical data into numerical representations. This process assigns a unique integer to each distinct category within a categorical feature, effectively converting non-numeric labels into numeric format. Label encoding is particularly useful in machine learning workflows where algorithms require numerical inputs, enabling the inclusion of categorical variables in predictive models. While straightforward in its application, label encoding preserves the ordinality of categorical variables, which may inadvertently introduce a hierarchical relationship among categories. However, it is essential to exercise caution when using label encoding with algorithms that interpret numerical values as ordinal, as this can lead to misleading conclusions. Despite its simplicity, label encoding remains a valuable preprocessing step in data preparation, facilitating the incorporation of categorical information into machine learning models without requiring extensive feature engineering.

g) Feature Selection:

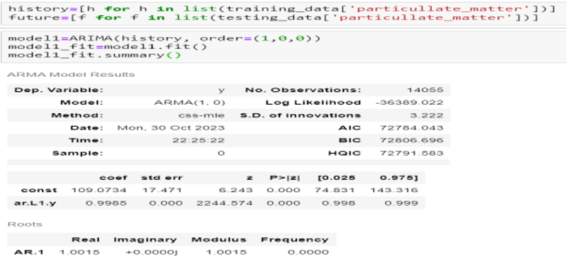
Feature selection is a crucial step in machine learning and data analysis, aimed at identifying and retaining the most relevant attributes from a dataset while discarding irrelevant or redundant ones. By selecting a subset of features that contribute most significantly to the predictive task, feature selection enhances model performance, reduces overfitting, and improves interpretability. Various techniques, such as filter methods, wrapper methods, and embedded methods, are employed to assess feature importance and select the most informative attributes. Filter methods evaluate features independently of the predictive model, often based on statistical metrics or correlation analysis. Wrapper methods assess feature subsets by training and evaluating candidate models iteratively, considering the model's performance as the selection criterion. Embedded methods incorporate feature selection within the model training process, optimizing feature relevance alongside model parameters. Through systematic evaluation and prioritization of features, feature selection enhances model efficiency and generalization, enabling more robust and interpretable machine learning solutions.

h) Training and Testing:

Splitting the data into training and testing sets is a fundamental practice in deep learning to assess model performance and generalization ability. This process involves partitioning the dataset into two distinct subsets: the training set, used to train the deep learning model, and the testing set, held out for evaluating the model's performance on unseen data. The training set is utilized to optimize the model's parameters through iterative optimization algorithms, such as gradient descent, while the testing set serves as an independent validation set to estimate the model's performance on new, unseen samples. By ensuring that the training and testing sets are mutually exclusive, data splitting helps prevent overfitting and provides a reliable estimate of the model's ability to generalize to unseen data. Moreover, techniques like cross-validation can be employed to further assess model performance and enhance the reliability of the evaluation process. Through meticulous data splitting, deep learning practitioners can ensure robust model evaluation and deploy more reliable and accurate machine learning solutions.

i) Algorithms:

ARIMA (AutoRegressive Integrated Moving Average): ARIMA stands as a statistical workhorse in time series forecasting, amalgamating autoregression, differencing, and moving average components to model data relationships effectively. [15] Its strength lies in capturing linear patterns within sequential data, making it a staple in predicting future values based on historical observations.



```
history=[h for h in list(training_data['particullate_matter'])]
future=[f for f in list(testing_data['particullate_matter'])]
```

```
model1=ARIMA(history, order=(1,0,0))
model1_fit=model1.fit()
model1_fit.summary()
```

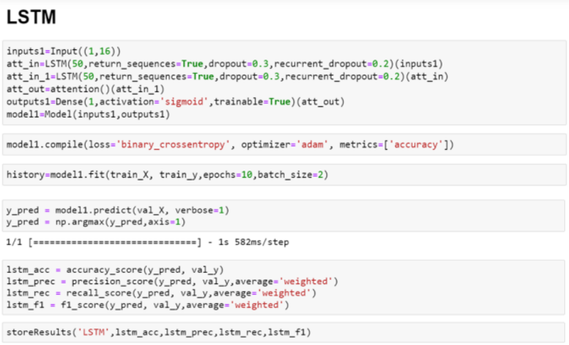
```
ARMA Model Results
Dep. Variable:                    y   No. Observations:       14055
        Model:           ARMA(1, 0)     Log Likelihood   -36389.022
       Method:              css-mle  S.D. of innovations      3.222
         Date:     Mon, 30 Oct 2023                AIC    72784.043
         Time:             22:25:22                BIC    72806.696
       Sample:                    0               HQIC    72791.583

            coef   std err        z   P>|z|   [0.025    0.975]
  const 109.0734    17.471    6.243   0.000   74.831   143.316
ar.L1.y   0.9985     0.000 2244.574   0.000    0.998     0.999

Roots
        Real   Imaginary   Modulus   Frequency
AR.1  1.0015    +0.0000j    1.0015      0.0000
```

Fig 3 ARIMA

LSTM (Long Short-Term Memory): LSTM, a subtype of recurrent neural networks (RNNs), excels in managing long-range dependencies and sequential data intricacies. [19] Its unique architecture, featuring memory cells capable of retaining information over extended intervals, makes it well-suited for time series forecasting tasks, adept at discerning and learning from temporal patterns.



LSTM

```
inputs1=Input((1,16))
att_in=LSTM(50,return_sequences=True,dropout=0.3,recurrent_dropout=0.2)(inputs1)
att_in_1=LSTM(50,return_sequences=True,dropout=0.3,recurrent_dropout=0.2)(att_in)
att_out=attention()(att_in_1)
outputs1=Dense(1,activation='sigmoid',trainable=True)(att_out)
model1=Model(inputs1,outputs1)
```

```
model1.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
```

```
history=model1.fit(train_X, train_y,epochs=10,batch_size=2)
```

```
y_pred = model1.predict(val_X, verbose=1)
y_pred = np.argmax(y_pred,axis=1)
```

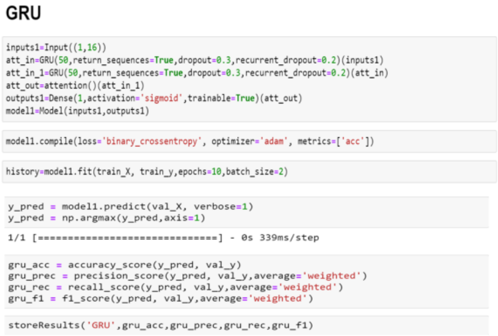
```
1/1 [==============================] - 1s 582ms/step
```

```
lstm_acc = accuracy_score(y_pred, val_y)
lstm_prec = precision_score(y_pred, val_y,average='weighted')
lstm_rec = recall_score(y_pred, val_y,average='weighted')
lstm_f1 = f1_score(y_pred, val_y,average='weighted')
```

```
storeResults('LSTM',lstm_acc,lstm_prec,lstm_rec,lstm_f1)
```

Fig 4 LSTM

GRU (Gated Recurrent Unit): Similar to LSTM, GRU stands as another variant of RNN architecture with a simplified structure. It boasts gating mechanisms facilitating selective information updates and omissions. Tasked with efficiently capturing sequential data dependencies, GRU[19] networks find utility in various applications, including time series prediction and language modeling.



GRU

```
inputs1=Input((1,16))
att_in=GRU(50,return_sequences=True,dropout=0.3,recurrent_dropout=0.2)(inputs1)
att_in_1=GRU(50,return_sequences=True,dropout=0.3,recurrent_dropout=0.2)(att_in)
att_out=attention()(att_in_1)
outputs1=Dense(1,activation='sigmoid',trainable=True)(att_out)
model1=Model(inputs1,outputs1)
```

```
model1.compile(loss='binary_crossentropy', optimizer='adam', metrics=['acc'])
```

```
history=model1.fit(train_X, train_y,epochs=10,batch_size=2)
```

```
y_pred = model1.predict(val_X, verbose=1)
y_pred = np.argmax(y_pred,axis=1)
```

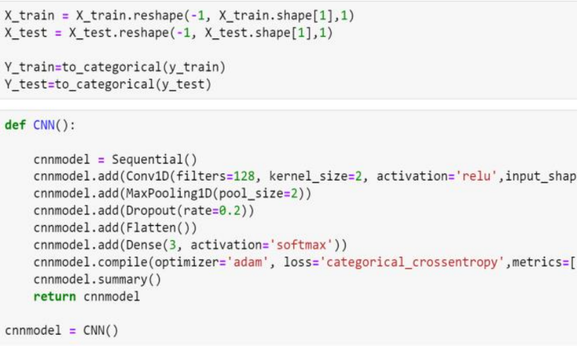
```
1/1 [==============================] - 0s 339ms/step
```

```
gru_acc = accuracy_score(y_pred, val_y)
gru_prec = precision_score(y_pred, val_y,average='weighted')
gru_rec = recall_score(y_pred, val_y,average='weighted')
gru_f1 = f1_score(y_pred, val_y,average='weighted')
```

```
storeResults('GRU',gru_acc,gru_prec,gru_rec,gru_f1)
```

Fig 5 GRU

CNN Algorithm: The introduction of Convolutional Neural Network (CNN) enriches the comparative analysis project by delving into spatial and temporal feature extraction from air quality data. With its proficiency in identifying hierarchical patterns and correlations within pollutant concentration sequences, CNN enhances the understanding of spatial dependencies crucial for accurate predictions. By integrating CNN into deep learning models, the project explores its efficacy in augmenting air pollutant prediction accuracy, complementing temporal aspects captured by recurrent models.



```
X_train = X_train.reshape(-1, X_train.shape[1],1)
X_test = X_test.reshape(-1, X_test.shape[1],1)

Y_train=to_categorical(y_train)
Y_test=to_categorical(y_test)
```

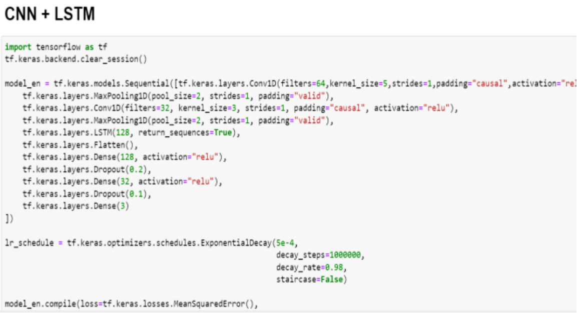
```
def CNN():

    cnnmodel = Sequential()
    cnnmodel.add(Conv1D(filters=128, kernel_size=2, activation='relu',input_shap
    cnnmodel.add(MaxPooling1D(pool_size=2))
    cnnmodel.add(Dropout(rate=0.2))
    cnnmodel.add(Flatten())
    cnnmodel.add(Dense(3, activation='softmax'))
    cnnmodel.compile(optimizer='adam', loss='categorical_crossentropy',metrics=[
    cnnmodel.summary()
    return cnnmodel

cnnmodel = CNN()
```

Fig 6 CNN

CNN+LSTM: The hybrid CNN+LSTM model represents a potent fusion of spatial and temporal learning in air quality forecasting. Leveraging CNN for spatial feature extraction and LSTM[19] for temporal dependency capture, this model offers a comprehensive approach to understanding complex patterns in pollutant concentrations. By combining the strengths of both architectures, the CNN+LSTM model aims to provide a holistic representation of air quality dynamics, contributing valuable insights into the synergy between spatial and temporal learning for enhanced prediction accuracy in urban areas.



CNN + LSTM

```
import tensorflow as tf
tf.keras.backend.clear_session()

model_en = tf.keras.models.Sequential([tf.keras.layers.Conv1D(filters=64,kernel_size=5,strides=1,padding="causal",activation="rel
    tf.keras.layers.MaxPooling1D(pool_size=2, strides=1, padding="valid"),
    tf.keras.layers.Conv1D(filters=32, kernel_size=3, strides=1, padding="causal", activation="relu"),
    tf.keras.layers.MaxPooling1D(pool_size=2, strides=1, padding="valid"),
    tf.keras.layers.LSTM(128, return_sequences=True),
    tf.keras.layers.Flatten(),
    tf.keras.layers.Dense(128, activation="relu"),
    tf.keras.layers.Dropout(0.2),
    tf.keras.layers.Dense(32, activation="relu"),
    tf.keras.layers.Dropout(0.1),
    tf.keras.layers.Dense(3)
])

lr_schedule = tf.keras.optimizers.schedules.ExponentialDecay(5e-4,
                                                             decay_steps=1000000,
                                                             decay_rate=0.98,
                                                             staircase=False)

model_en.compile(loss=tf.keras.losses.MeanSquaredError(),
```

Fig 7 CNN + LSTM

4. EXPERIMENTAL RESULTS

Accuracy: The accuracy of a test is its ability to differentiate the patient and healthy cases correctly. To estimate the accuracy of a test, we should calculate the proportion of true positive and true negative in all evaluated cases. Mathematically, this can be stated as:

Accuracy = TP + TN TP + TN + FP + FN.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

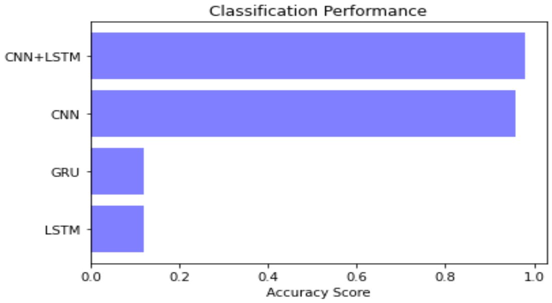


Fig 8 Accuracy Comparison Graph

F1-Score: F1 score is a machine learning evaluation metric that measures a model's accuracy. It combines the precision and recall scores of a model. The accuracy metric computes how many times a model made a correct prediction across the entire dataset.

$$\textbf{F1 Score} = \frac{2}{\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}\right)}$$

$$\textbf{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

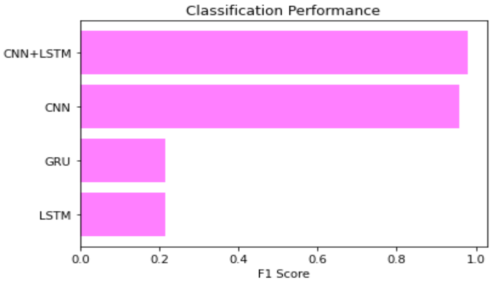


Fig 9 F1 Score Comparison Graph

Precision: Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

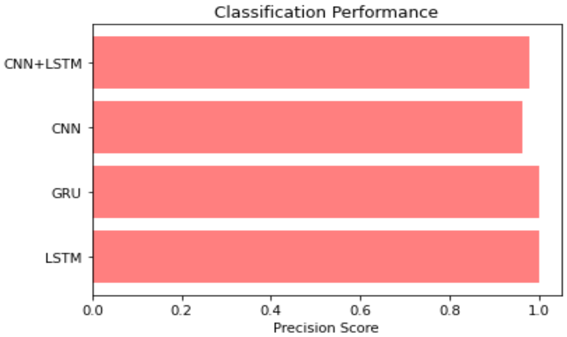


Fig 10 Precision Comparison Graph

Recall: Recall is a metric in machine learning that measures the ability of a model to identify all relevant instances of a particular class. It is the ratio of correctly predicted positive observations to the total actual positives, providing insights into a model's completeness in capturing instances of a given class.

$$Recall = \frac{TP}{TP + FN}$$

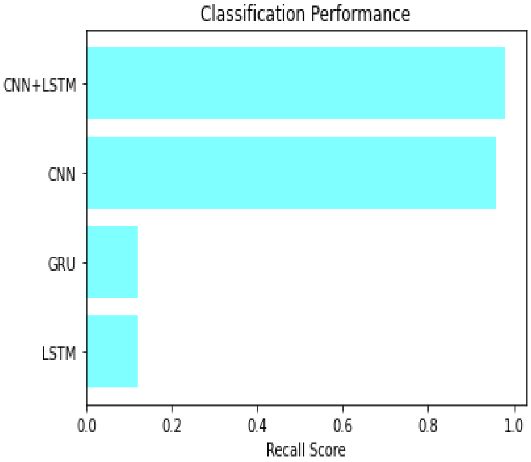
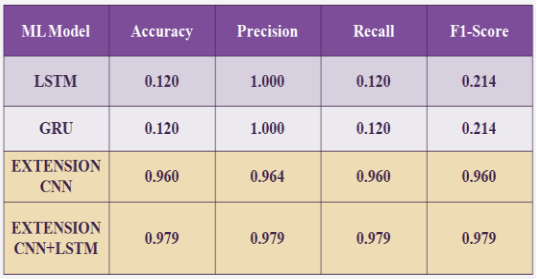
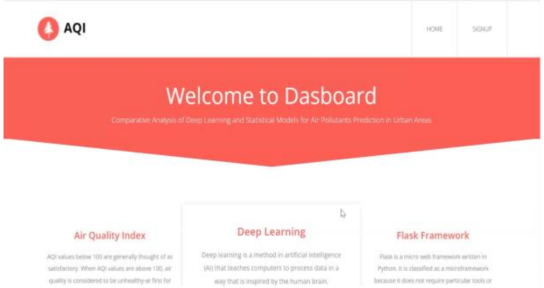


Fig 11 Recall Comparison Graph



ML Model	Accuracy	Precision	Recall	F1-Score
LSTM	0.120	1.000	0.120	0.214
GRU	0.120	1.000	0.120	0.214
EXTENSION CNN	0.960	0.964	0.960	0.960
EXTENSION CNN+LSTM	0.979	0.979	0.979	0.979

Fig 12 Evaluation Table



Welcome to Dasboard

Comparative Analysis of Deep Learning and Statistical Models for Air Pollutants Prediction in Urban Areas

Air Quality Index

AQI values below 100 are generally thought of as satisfactory. When AQI values are above 100, air quality is considered to be unhealthy-at first for

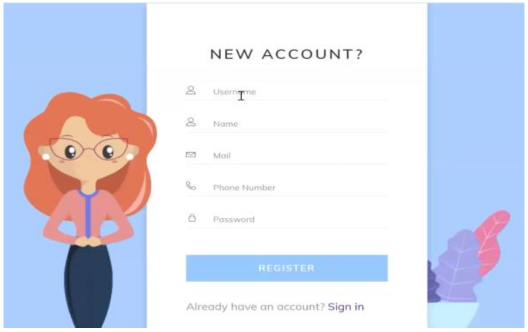
Deep Learning

Deep learning is a method in artificial intelligence (AI) that teaches computers to process data in a way that is inspired by the human brain.

Flask Framework

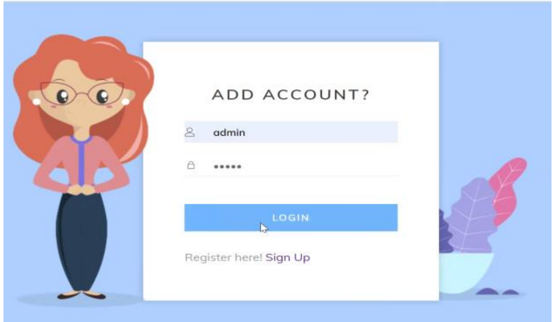
Flask is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or

Fig 13 Home Page



Already have an account? Sign in

Fig 14 Registration Page

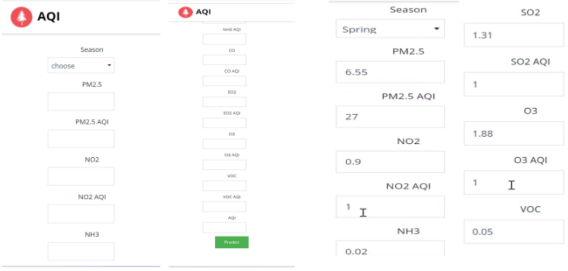


admin

•••••

Register here! Sign Up

Fig 15 Login Page



Season: Spring

PM2.5: 6.55

PM2.5 AQI: 27

NO2: 0.9

NO2 AQI: 1

NH3: 0.02

SO2: 1.31

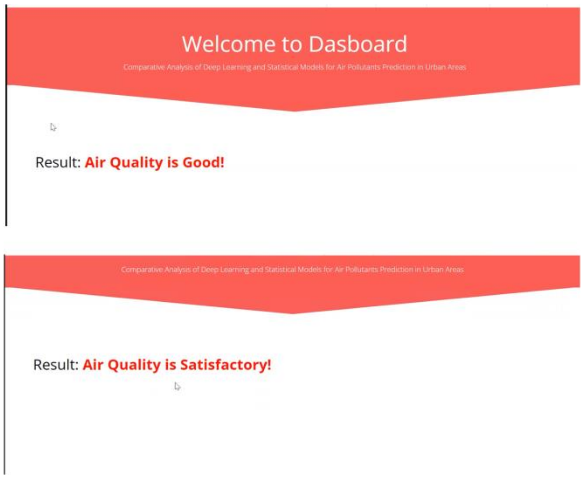
SO2 AQI: 1

O3: 1.88

O3 AQI: 1

VOC: 0.05

Fig 16 Upload Input Values



Welcome to Dasboard

Comparative Analysis of Deep Learning and Statistical Models for Air Pollutants Prediction in Urban Areas

Result: Air Quality is Good!

Comparative Analysis of Deep Learning and Statistical Models for Air Pollutants Prediction in Urban Areas

Result: Air Quality is Satisfactory!

Fig 17 Predicted Results

5. CONCLUSION

In conclusion, the accurate prediction of air pollutant levels in urban areas is paramount for safeguarding global health and environmental sustainability. Through the assessment of various forecasting approaches, including advanced deep learning models like LSTM[19] and GRU[19], this project has demonstrated their superiority over traditional statistical methods in predicting air quality. The continual refinement of methodology, focusing on feature engineering, parameter optimization, and multi-step prediction strategies, holds promise for further improving the accuracy and adaptability of air pollution forecasts. These advancements carry significant potential to empower policymakers, environmental agencies, and health organizations in making informed decisions and implementing targeted interventions to mitigate the adverse impacts of air pollution. By enhancing the reliability of air quality forecasts, these advancements contribute to broader public health initiatives and environmental sustainability efforts, ultimately fostering healthier and more sustainable urban environments for future generations.

6. FUTURE SCOPE

In the future, our aim is to advance towards multi-step prediction and enhance the performance of deep learning (DL) models through innovative feature engineering techniques and refined optimization of hyperparameters. By targeting multi-step prediction, we seek to extend the forecasting horizon beyond single time steps, providing more comprehensive insights into future air pollutant levels. Additionally, we plan to explore novel feature engineering approaches to extract more informative features from the data, enabling the DL models to capture complex relationships and patterns more effectively. Furthermore, optimizing hyperparameters will involve fine-tuning model configurations to maximize predictive accuracy and robustness. These advancements will not only elevate the performance of DL models in air quality forecasting but also contribute to a deeper understanding of environmental dynamics and facilitate more informed decision-making for mitigating air pollution's adverse effects.

REFERENCES

[1] R. D. Brook, ‘‘Cardiovascular effects of air pollution,’’ Clin. Sci., vol. 115, no. 6, pp. 175–187, Sep. 2008.

[2] M. Stafoggia and T. Bellander, ‘‘Short-term effects of air pollutants on daily mortality in the Stockholm county—A spatiotemporal analysis,’’ Environ. Res., vol. 188, Sep. 2020, Art. no. 109854.

[3] WHO. Household Air Pollution and Health. Accessed: Dec. 29, 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/ household-air-pollution-and-health

[4] WHO. Air Quality and Health. Accessed: Jan. 10, 2023. [Online]. Available: https://www.who.int/teams/environment-climate-change-andhealth/air-quality-and-health/policy-progress/sustainable-developmentgoals-air-pollution

[5] B. Paul and S. Louise. (2022). Air Quality: Policies, Proposals and Concerns—House of Commons Library. Accessed: Jan. 10, 2023. [Online]. Available: https://commonslibrary.parliament.uk/researchbriefings/cbp-9600/

[6] P. J. Landrigan, ‘‘Air pollution and health,’’ Lancet Public Health, vol. 2, pp. e4–e5, Jan. 2017.

[7] K. Abutalip, A. Al-Lahham, and A. El Saddik, ‘‘Digital twin of atmospheric environment: Sensory data fusion for high-resolution PM2.5 estimation and action policies recommendation,’’ IEEE Access, vol. 11, pp. 14448–14457, 2023.

[8] J. Ma, Y. Ding, J. C. P. Cheng, F. Jiang, Y. Tan, V. J. L. Gan, and Z. Wan, ‘‘Identification of high impact factors of air quality on a national scale using big data and machine learning techniques,’’ J. Cleaner Prod., vol. 244, Jan. 2020, Art. no. 118955.

[9] I. Manisalidis, E. Stavropoulou, A. Stavropoulos, and E. Bezirtzoglou, ‘‘Environmental and health impacts of air pollution: A review,’’ Frontiers Public Health, vol. 8, p. 14, 2020.

[10] S. Ameer, M. A. Shah, A. Khan, H. Song, C. Maple, S. U. Islam, and M. N. Asghar, ‘‘Comparative analysis of machine learning techniques for predicting air quality in smart cities,’’ IEEE Access, vol. 7, pp. 128325–128338, 2019.

[11] Q. Chen, W. Wang, F. Wu, S. De, R. Wang, B. Zhang, and X. Huang, ‘‘A survey on an emerging area: Deep learning for smart city data,’’ IEEE Trans. Emerg. Topics Comput. Intell., vol. 3, no. 5, pp. 392–410, Oct. 2019.

[12] Y. Zhang, Y. Wang, M. Gao, Q. Ma, J. Zhao, R. Zhang, Q. Wang, and L. Huang, ‘‘A predictive data feature exploration-based air quality prediction approach,’’ IEEE Access, vol. 7, pp. 30732–30743, 2019.

[13] Doreswamy, K. S. Harishkumar, Y. Km, and I. Gad, ‘‘Forecasting air pollution particulate matter (PM2.5) using machine learning regression models,’’ Proc. Comput. Sci., vol. 171, pp. 2057–2066, Jan. 2020.

[14] Y.-S. Chang, H.-T. Chiao, S. Abimannan, Y.-P. Huang, Y.-T. Tsai, and K.-M. Lin, ‘‘An LSTM-based aggregated model for air pollution forecasting,’’ Atmos. Pollut. Res., vol. 11, no. 8, pp. 1451–1463, Aug. 2020.

[15] J. Ma, Y. Ding, J. C. P. Cheng, F. Jiang, V. J. L. Gan, and Z. Xu, ‘‘A lag-FLSTM deep learning network based on Bayesian optimization for multi-sequential-variant PM2.5 prediction,’’ Sustain. Cities Soc., vol. 60, Sep. 2020, Art. no. 102237.

[16] R. Yan, J. Liao, J. Yang, W. Sun, M. Nong, and F. Li, ‘‘Multi-hour and multi-site air quality index

forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering,'' Expert Syst. Appl., vol. 169, May 2021, Art. no. 114513.

[17] S. Du, T. Li, Y. Yang, and S. Horng, ''Deep air quality forecasting using hybrid deep learning framework,'' IEEE Trans. Knowl. Data Eng., vol. 33, no. 6, pp. 2412–2424, Jun. 2021.

[18] N. Zaini, L. W. Ean, A. N. Ahmed, M. A. Malek, and M. F. Chow, ''PM2.5 forecasting for an urban area based on deep learning and decomposition method,'' Sci. Rep., vol. 12, no. 1, p. 17565, Oct. 2022.

[19] B. Wang, W. Kong, H. Guan, and N. N. Xiong, ''Air quality forecasting based on gated recurrent long short term memory model in Internet of Things,'' IEEE Access, vol. 7, pp. 69524–69534, 2019.

[20] Q. Tao, F. Liu, Y. Li, and D. Sidorov, ''Air pollution forecasting using a deep learning model based on 1D convnets and bidirectional GRU,'' IEEE Access, vol. 7, pp. 76690–76698, 2019.

[21] A. Bekkar, B. Hssina, S. Douzi, and K. Douzi, ''Air quality forecasting using decision trees algorithms,'' in Proc. 2nd Int. Conf. Innov. Res. Appl. Sci., Eng. Technol. (IRASET), Mar. 2022, pp. 1–4.

[22] A. Bekkar, B. Hssina, S. Douzi, and K. Douzi, ''Air-pollution prediction in smart city, deep learning approach,'' J. Big Data, vol. 8, no. 1, pp. 1–21, Dec. 2021.

[23] Y. Liang, Y. Xia, S. Ke, Y. Wang, Q. Wen, J. Zhang, Y. Zheng, and R. Zimmermann, ''Airformer: Predicting nationwide air quality in China with transformers,'' Nov. 2022, arXiv:2211.15979.

[24] S. Wang, Y. Li, J. Zhang, Q. Meng, L. Meng, and F. Gao, ''PM2.5-GNN: A domain knowledge enhanced graph neural network for PM2.5 forecasting,'' in Proc. 28th Int. Conf. Adv. Geographic Inf. Syst. (SIGSPATIAL), Nov. 2020, pp. 163–166.

[25] G. Box, G. Jenkins, G. Reinsel, and G. Ljung, Time Series Analysis: Forecasting and Control (Wiley Series in Probability and Statistics). Hoboken, NJ, USA: Wiley, 2015. [Online]. Available: https://books.google.co.uk/books?id=rNt5CgAAQBAJ

[26] Z. Yang, K. Wang, J. Li, Y. Huang, and Y. Zhang, ''TS-RNN: Text steganalysis based on recurrent neural networks,'' IEEE Signal Process. Lett., vol. 26, no. 12, pp. 1743–1747, Dec. 2019.

[27] G. Gelly and J. Gauvain, ''Optimization of RNN-based speech activity detection,'' IEEE/ACM Trans. Audio, Speech, Language Process., vol. 26, no. 3, pp. 646–656, Mar. 2018.

[28] S. Hochreiter and J. Schmidhuber, ''Long short-term memory,'' Neural Comput., vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[29] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, ''Learning phrase representations using RNN encoder–decoder for statistical machine translation,'' in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP). Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1724–1734.

[30] Northern Ireland Air. Download Air Quality Data—Northern Ireland. Accessed: Dec. 1, 2022. [Online]. Available: https://www.airqualityni.co.uk/data

[31] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, ''Hyperband: A novel bandit-based approach to hyperparameter optimization,'' J. Mach. Learn. Res., vol. 18, no. 1, pp. 6765–6816, Jan. 2017.

Dataset Link:

https://www.kaggle.com/datasets/cpluzshrijayan/air-quality-prediction-harbor