VayuAnukulani: Adaptive memory networks for air pollution forecasting

Divyam Madaan[1*], Radhika Dua[2*], Prerana Mukherjee[3,4], Brejesh Lall[4]

KAIST[1], Daejeon, South Korea

IIT Hyderabad[2], India

IIIT Sricity[3], India

IIT Delhi[4], India

Equal contribution, work done as an intern at IIT Delhi[*]

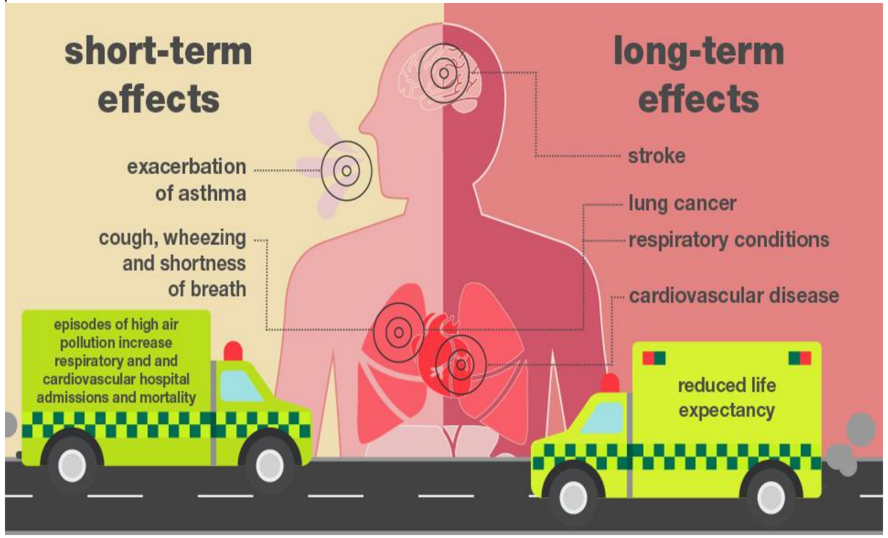
Overview

01	Motivation	Why is this problem important?
02	Challenges	Why hasn’t this already been solved?
03	Problem Statement	Notations and problem statement
04	Approach: Proposed Method	How we make it happen?
05	Dataset and Baselines	Datasets and baselines for measuring success
06	Conclusion	Summary of our work.

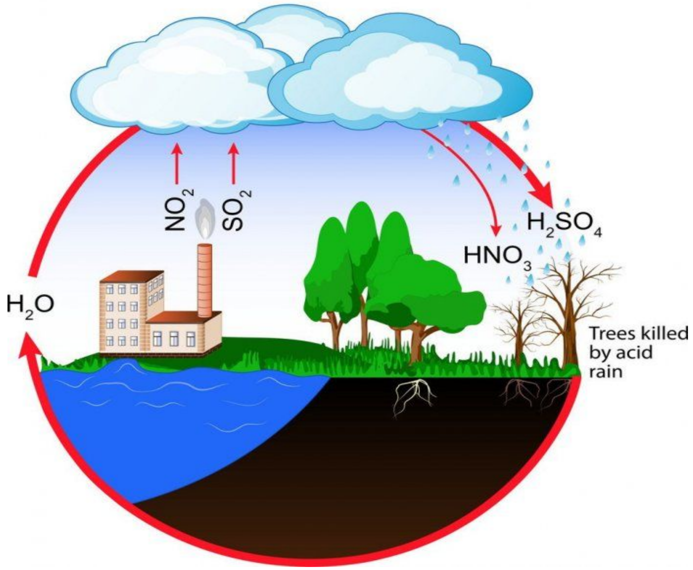
Motivation

Pollution has become an important concern in today's world.

Global warming



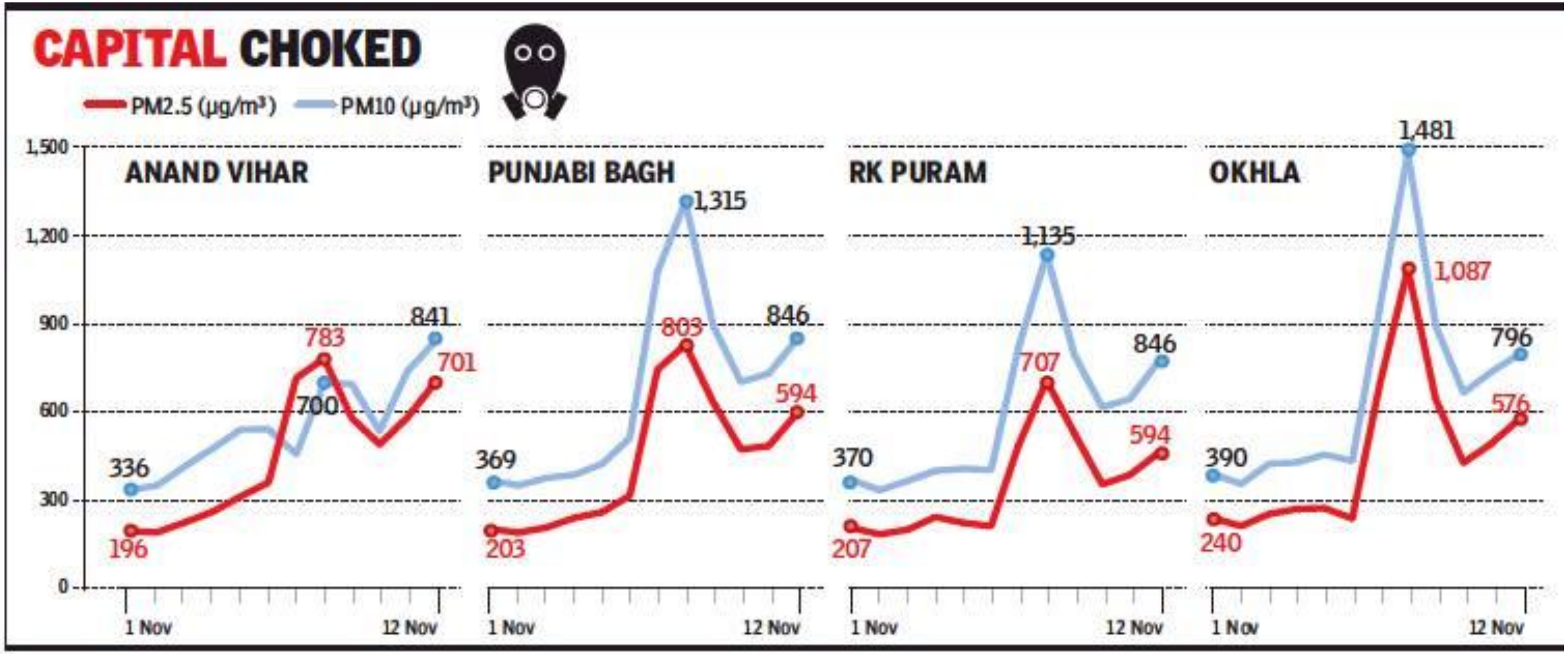
Health problems



Acid rain

Challenges

Air pollution varies with **location** and **time**.



It is essential to have a **separate** solution for each location.

Challenges

There exist various ***outliers*** when pollution increases/decreases.

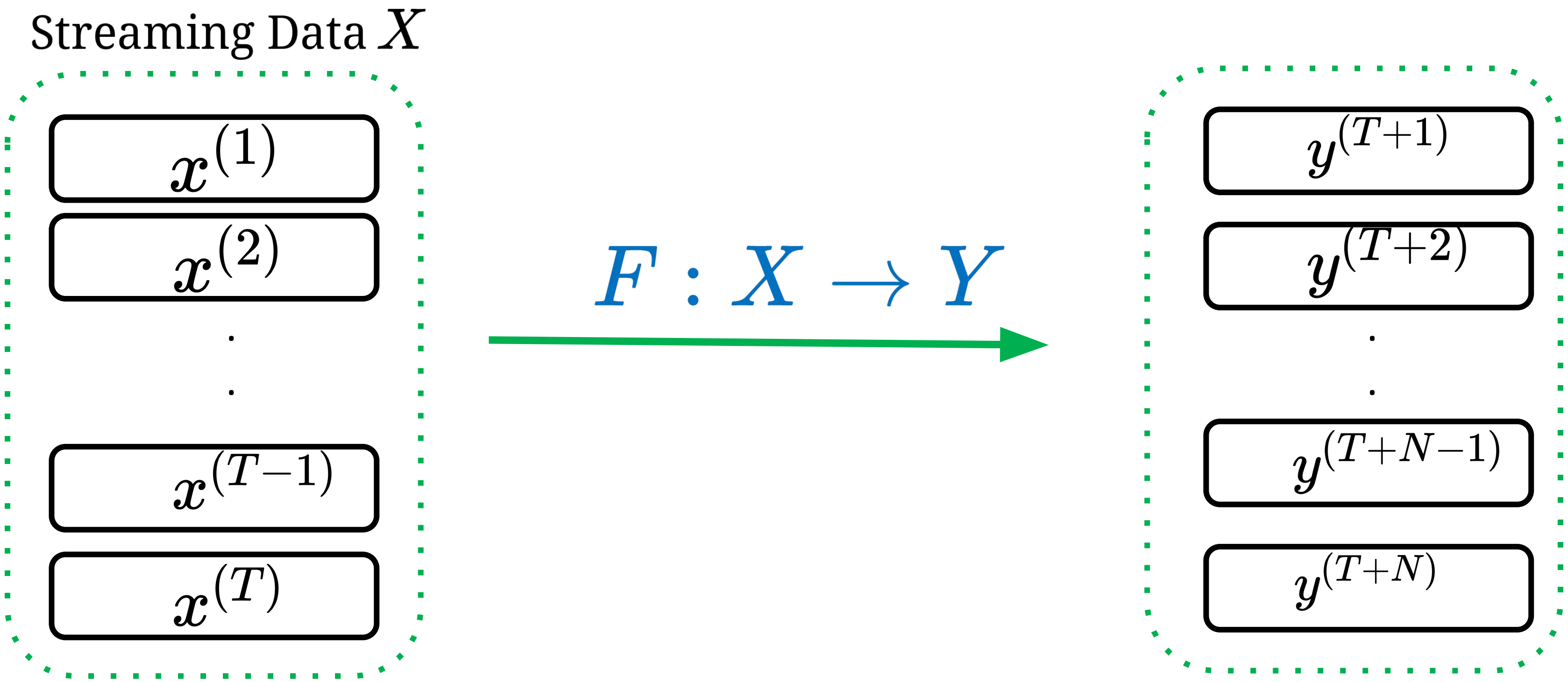
Farm burning

Forest fires

Festivals

Problem Statement

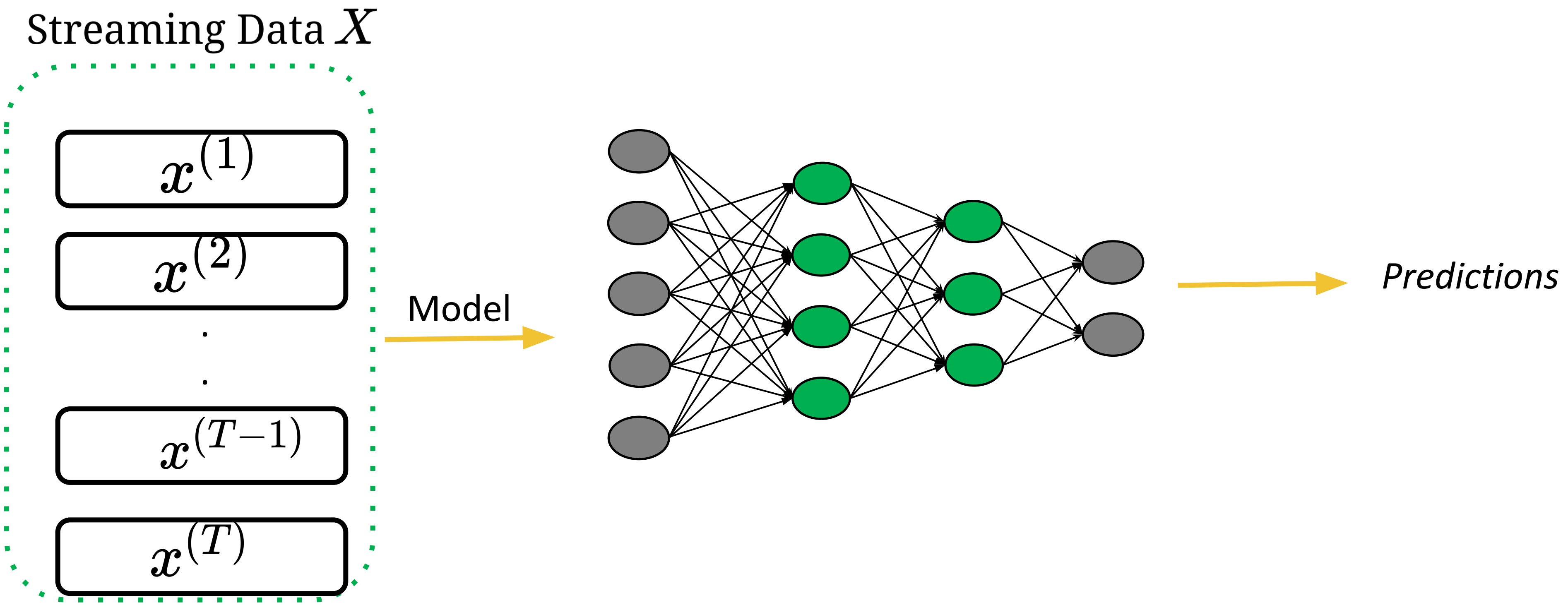
Given the ***input heterogenous urban data*** $X = \{x^{(1)}, x^{(2)} \dots x^{(T-1)}, x^{(T)}\}$, the ***predictive model*** should learn a function $F : X \rightarrow Y$ that maps it to the set of future pollution concentration and levels $Y = \{y^{(T+1)}, y^{(T+2)} \dots y^{(T+N-1)}, y^{(T+N)}\}$.



How can we ***learn*** such a function to predict ***multiple pollutants concentration and levels***?

Difference from Existing models

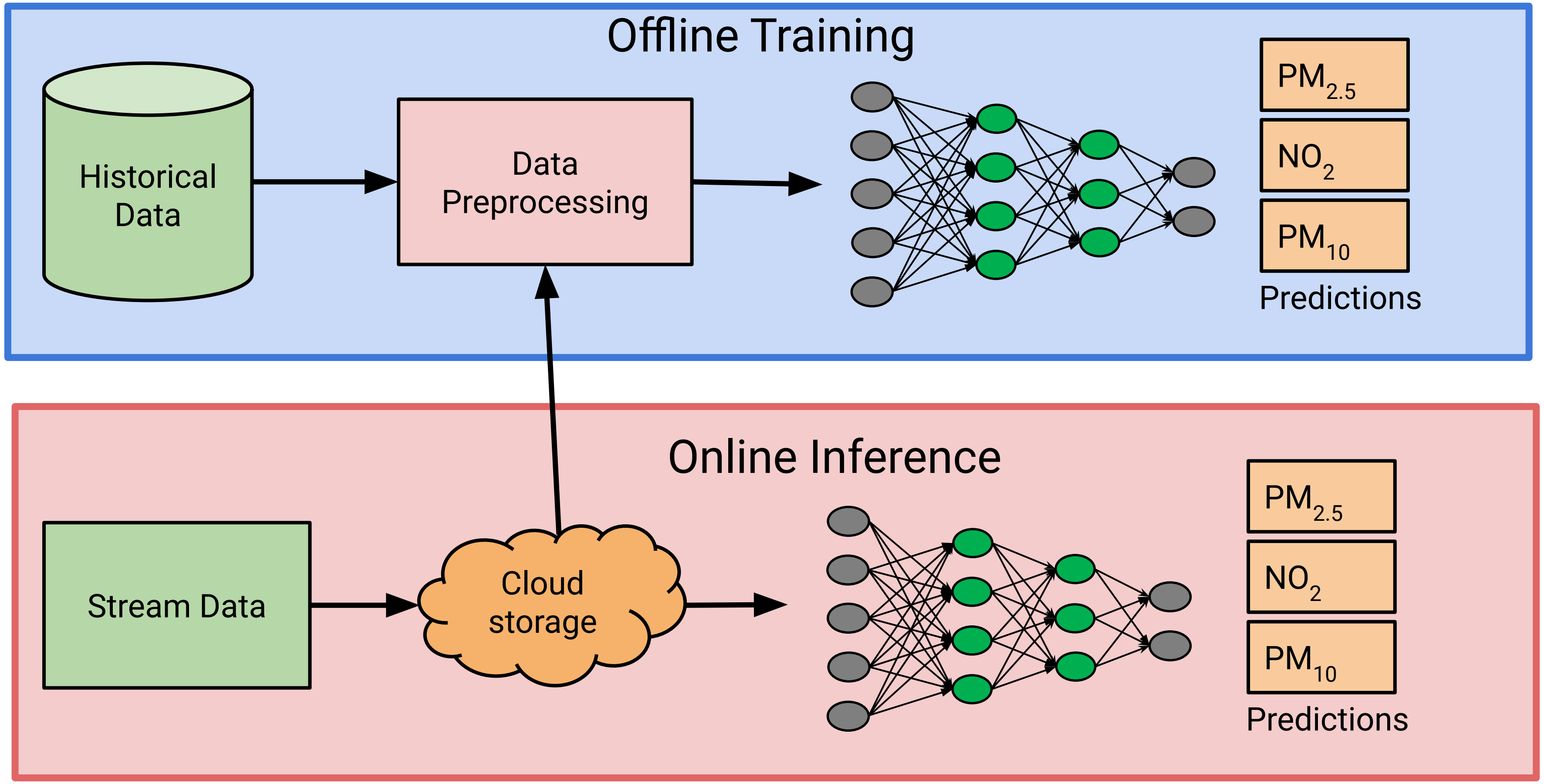
Our pollution prediction task requires a model that can handle sequentially streaming data and perform adaptive updates.



It is ***difficult to solve this problem*** using any existing methods for Delhi due to ***lack of accurate data and scalability***

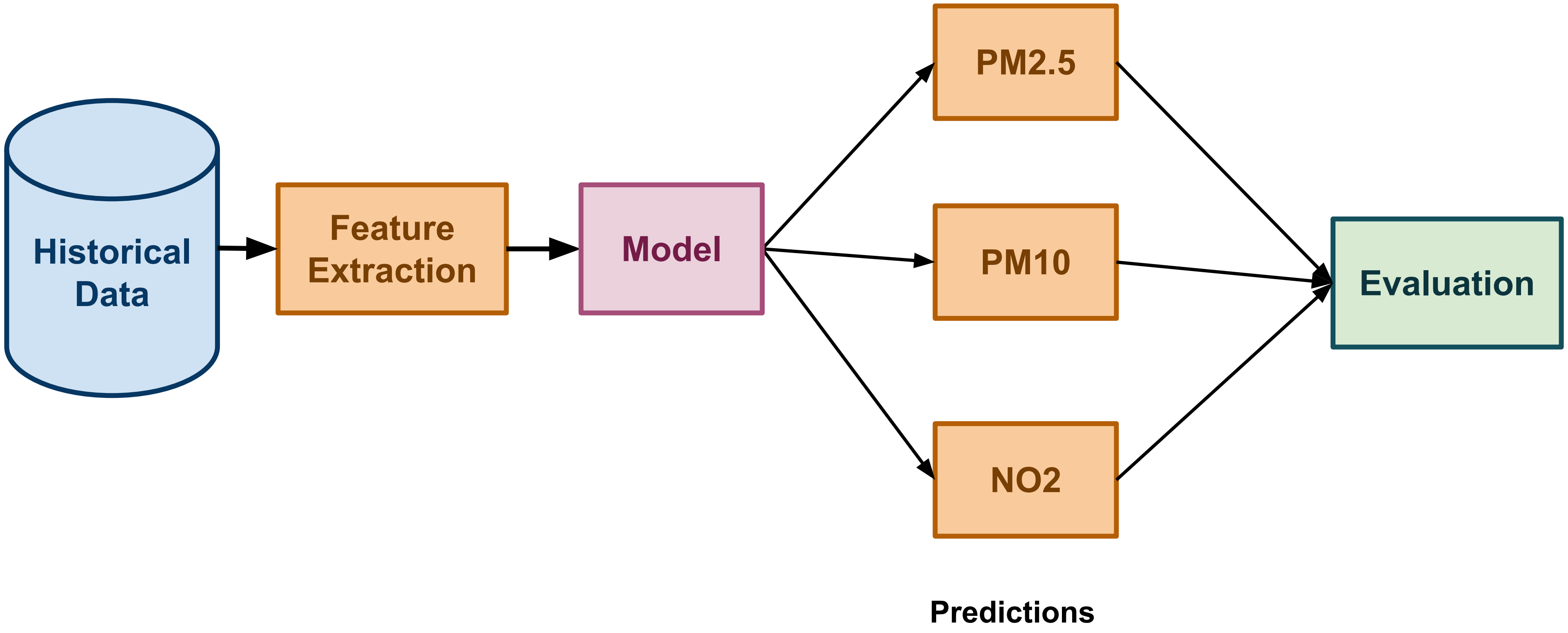
The Components of Proposed Approach

Vayuanukulani consists of ***Offline Training*** module and an ***Online Interface*** module to output the pollutants ***levels*** and ***concentration***.



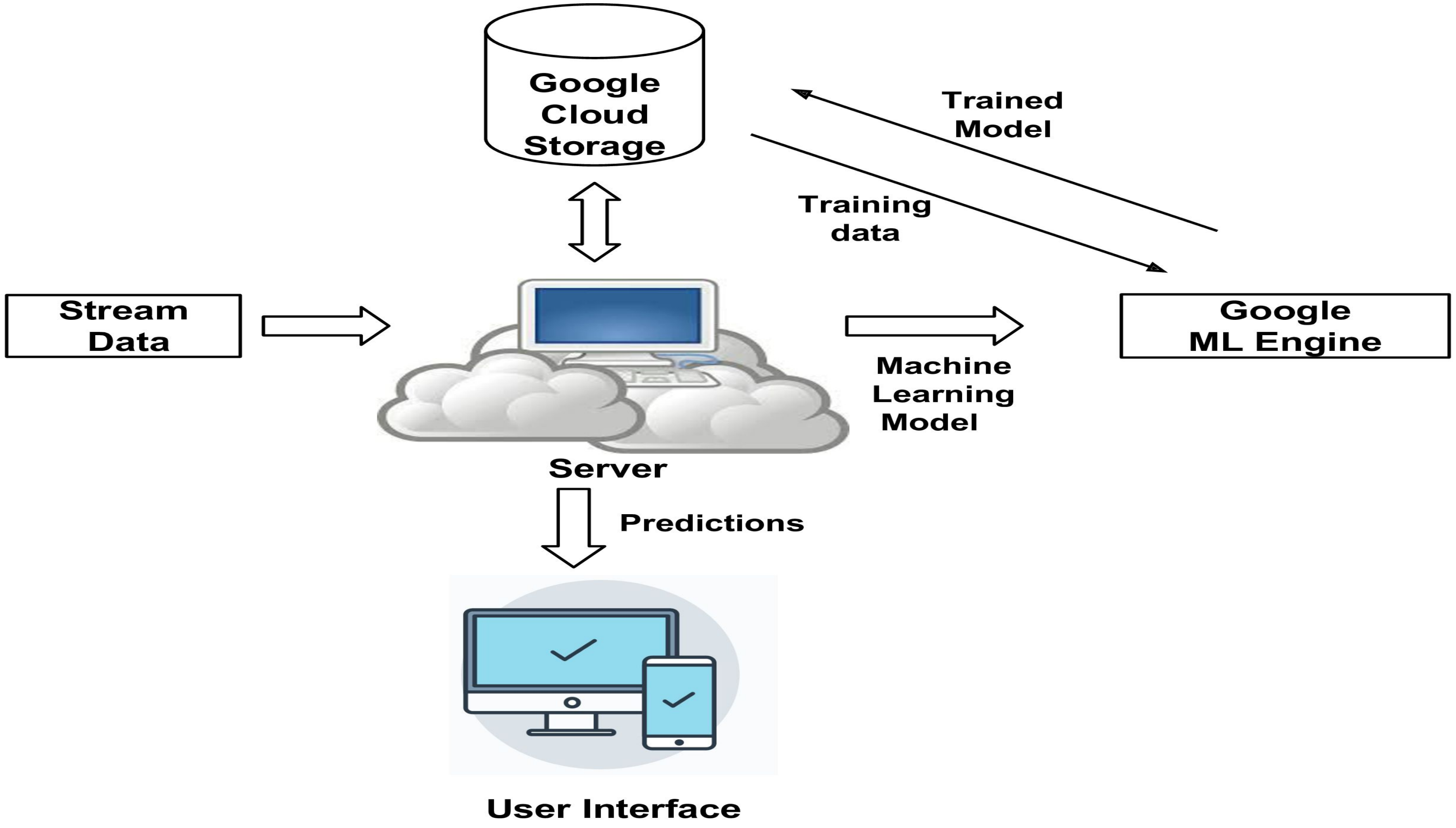
Offline Training

Offline Training module ***extracts features*** from the collected ***historical data*** to ***predict*** the pollutants level and concentration using our ***proposed model***.



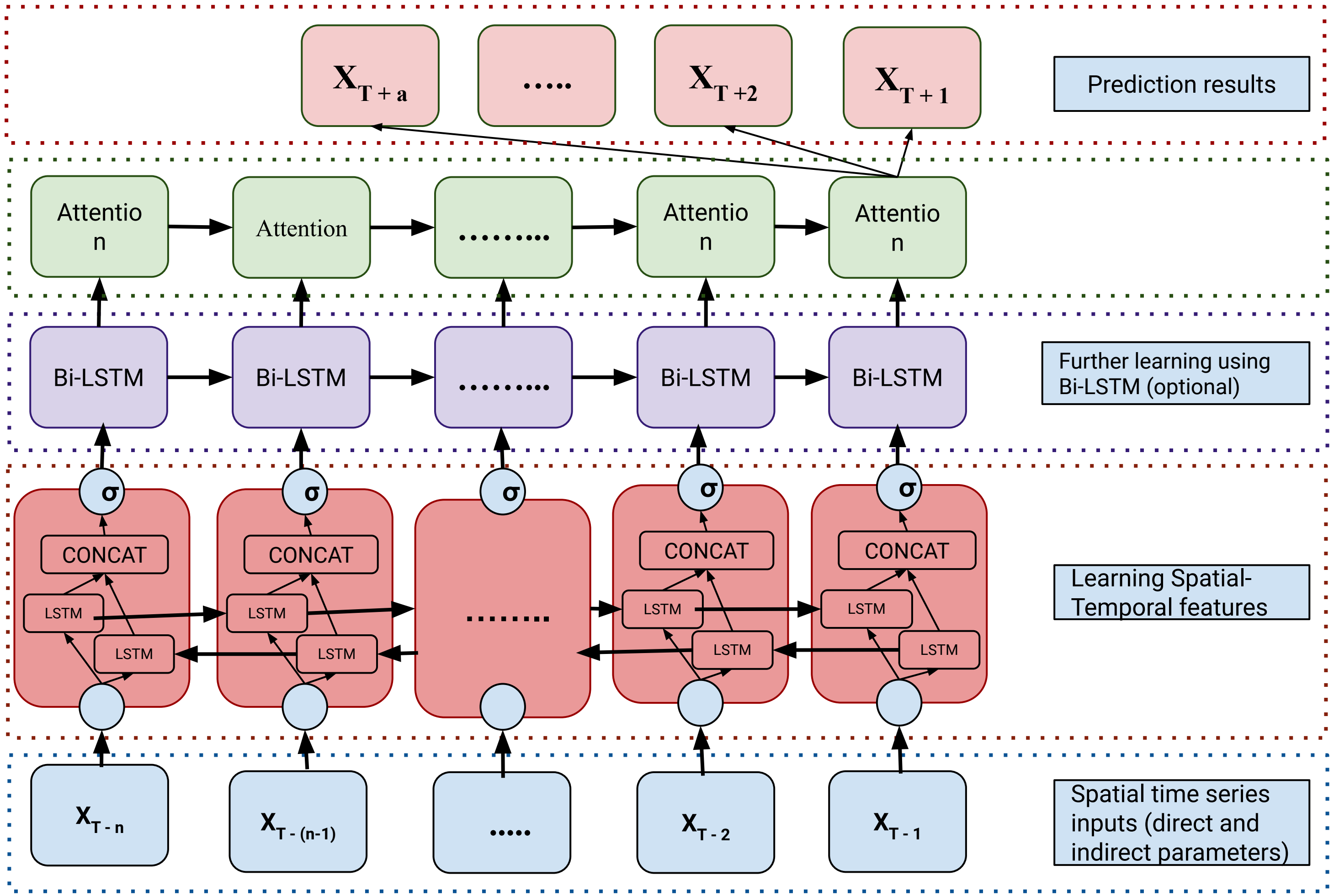
Online Inference

Online Interfaces updates the historical data ***every hour*** and the model ***every week***.



The Proposed Model

Our proposed model consists of a ***Bi-LSTM*** with ***attention*** module.



The Proposed Model

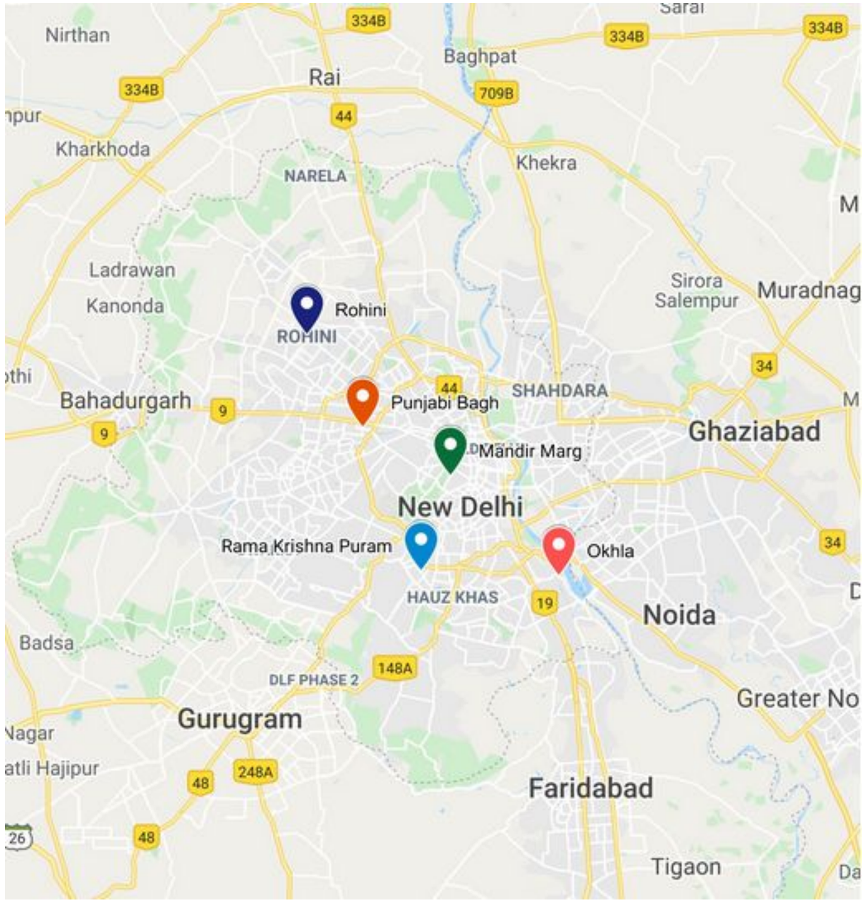
The trained model is ***updated*** every week using the proposed ***adaptive-learning*** approach.

Algorithm 1 Algorithm for proposed adaptive method

Inputs: Data for each location $\{f_1, f_2, ..f_{n-1}, f_n\}$ and learning rate $\alpha = 10^{-3}$.
Initialize $F(x)$ = BiLSTM model with attention mechanism for N pollutants.
for $t \leftarrow 1 ... T$ **do**
    Receive instance: x_t.
    Predict $\hat{y}_t$ for each pollutant for the next 24 hours.
    Receive the true pollutant value y_t.
    Suffer loss: $l_t(w_t)$ which is a convex loss function on both ${w_t}^T x$ and y_t.
    Update the prediction model w_t to w_{t+1}.
end for

Experiments: Dataset

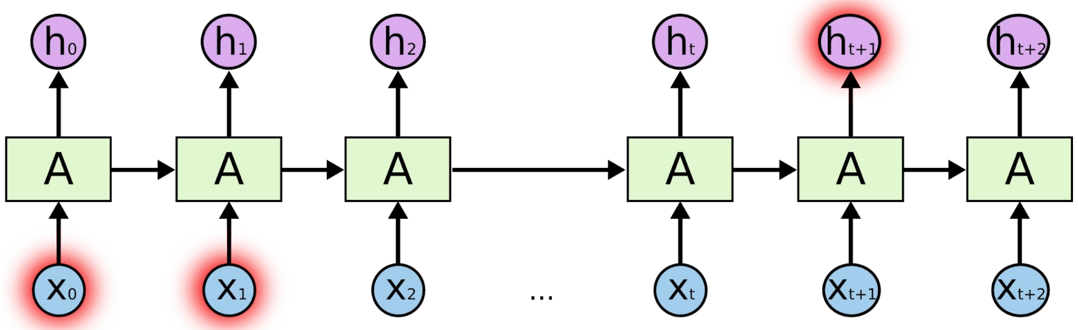
The collected dataset consists of ***direct (air pollutants)*** and ***indirect (meteorological data and time)*** for 3 years.



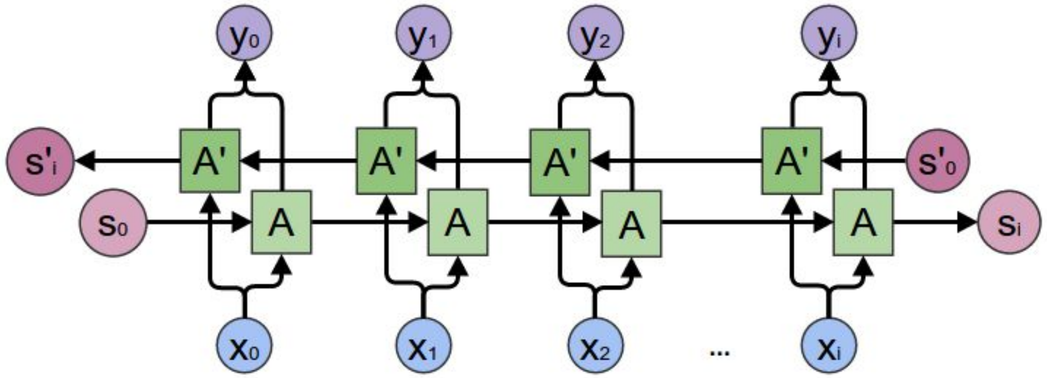
Number of Locations	5
Min number of samples per location	4000
Max number of samples per location	29000
Average number samples per location	7000
Span of data collection	3 years
Number of features per sample	9
Seasons covered	all
Number of hours per day	24

Experiments: Baselines

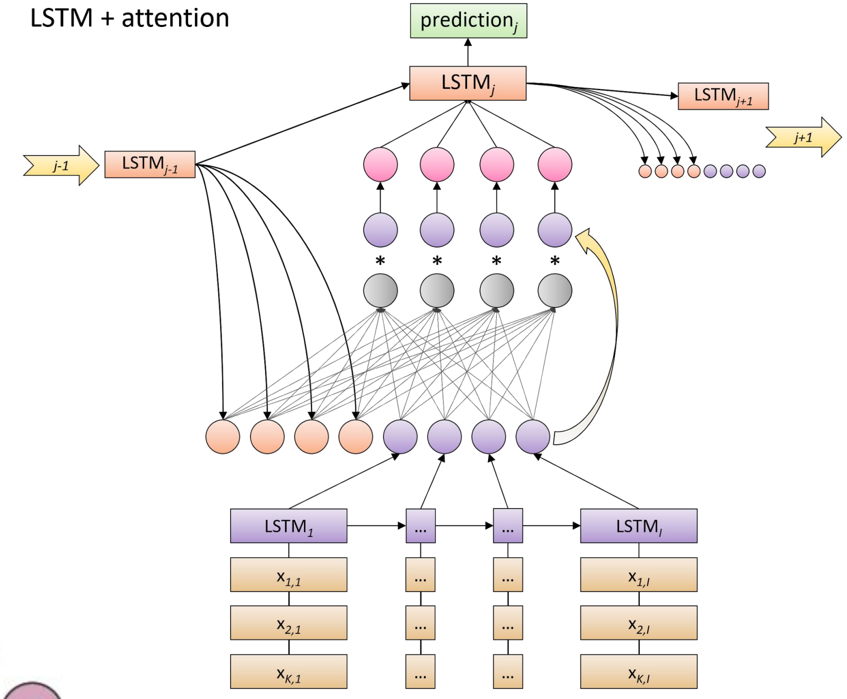
We experiment our ***Vayuanukulani*** against several baselines.



- LSTM



- Bi-LSTM



- LSTM + attention

Results

Our model outperforms the baselines for both the ***pollution levels*** and ***pollutants concentration prediction*** task.

TABLE I: Performance comparison of the proposed model with other baseline models for pollution values forecasting for future 4 hours on the basis of R-squared values and Root mean square error values. The highlighted values indicates the best performance.

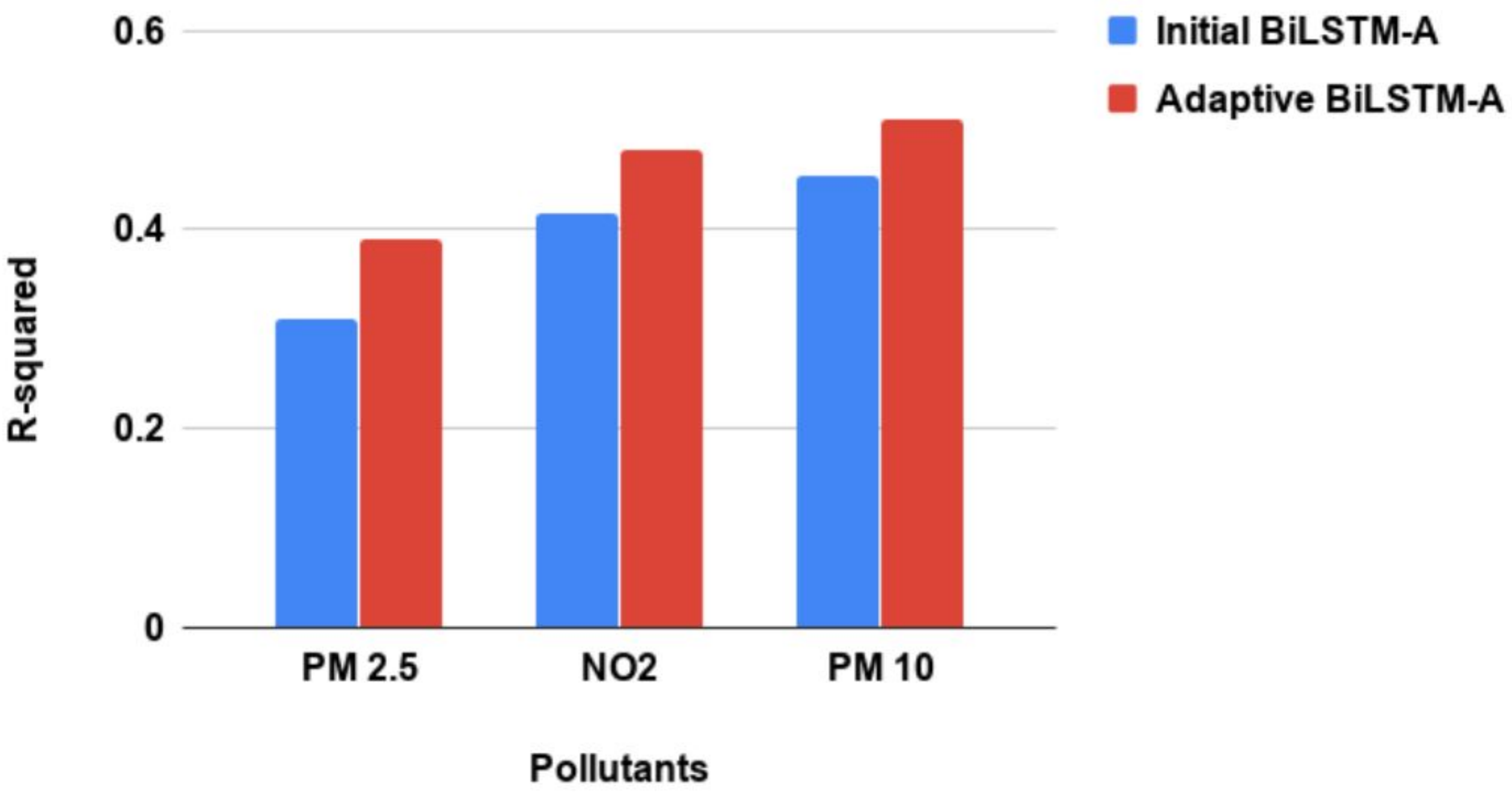
Model	*Pollutants*	*R-square*	*RMSE*
Random Forest	$PM_{2.5}$	**0.35**	**40.69**
	NO_2	0.40	21.12
	PM_{10}	0.42	98.32
LSTM	$PM_{2.5}$	0.31	41.96
	NO_2	0.38	21.52
	PM_{10}	0.44	96.58
LSTM-A	$PM_{2.5}$	0.29	42.52
	NO_2	0.38	21.44
	PM_{10}	0.44	96.49
BILSTM	$PM_{2.5}$	0.30	42.07
	NO_2	0.38	21.47
	PM_{10}	0.44	96.77
BILSTM-A	$PM_{2.5}$	0.31	41.97
	NO_2	**0.41**	**21.08**
	PM_{10}	**0.45**	**96.22**

TABLE II: Performance comparison of the proposed model with other baseline models for pollution levels forecasting for future 4 hours on the basis of Accuracy, average precision and average recall. Higher values of accuracy, precision and recall indicates the better performance of the model. The highlighted values indicates the best performance.

Model	*Pollutants*	*Accuracy*	*Precision*	*Recall*
LSTM	$PM_{2.5}$	67.68	56.15	52.27
	NO_2	76.85	76.29	75.2
	PM_{10}	68.34	71.11	56.31
LSTM-A	$PM_{2.5}$	67.24	56.46	52.56
	NO_2	76.85	76.15	75.65
	PM_{10}	68.71	70.21	57.89
BILSTM	$PM_{2.5}$	67.96	58.35	53.12
	NO_2	77.32	76.75	75.86
	PM_{10}	**68.87**	**70.25**	58.36
BILSTM-A	$PM_{2.5}$	67.96	55.71	52.55
	NO_2	77.66	77.10	**76.26**
	PM_{10}	68.21	69.21	57.73
CBILSTM-A	$PM_{2.5}$	**70.68**	**61.06**	**55.8**
	NO_2	**77.88**	**77.56**	76.14
	PM_{10}	67.45	68.23	**58.52**

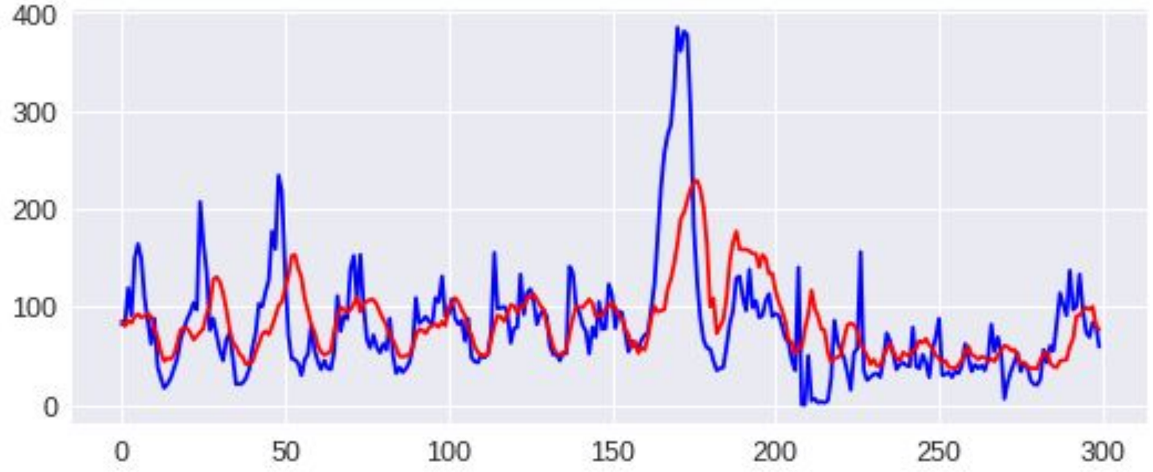
Results

Also, our ***proposed adaptive approach*** outperforms our standard proposed model.

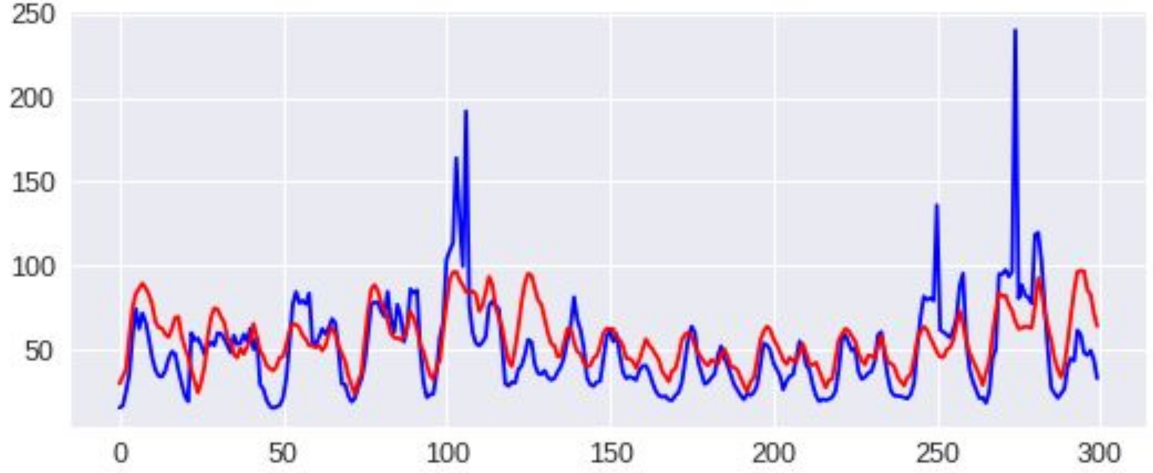


Results

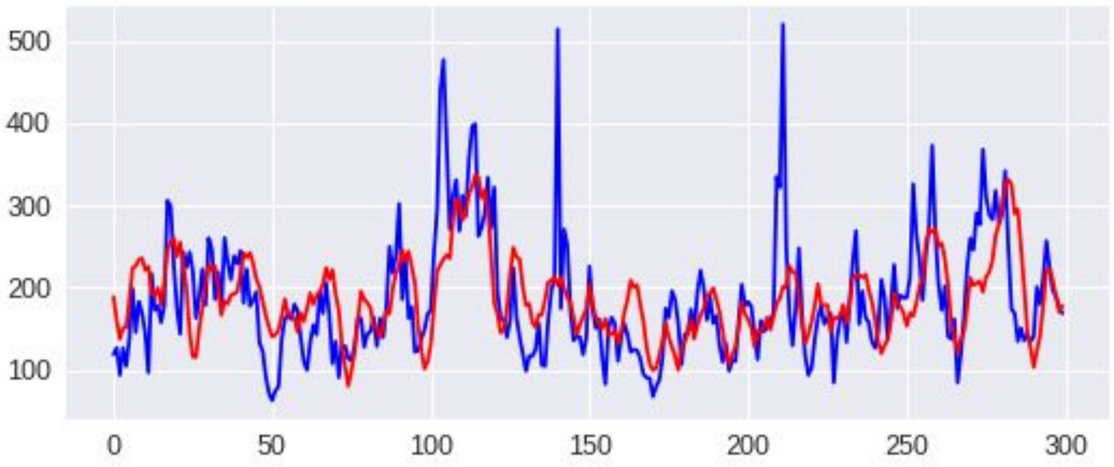
Our model is able to ***predict multiple pollutants*** successfully.



- PM_{10}



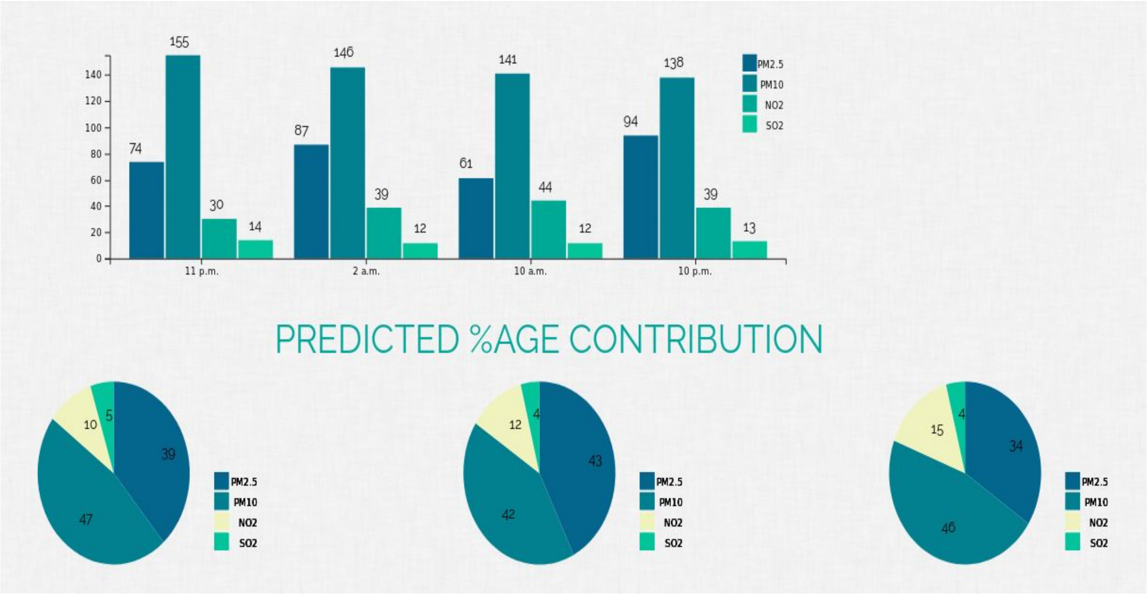
- $PM_{2.5}$



- NO_2

User Interface

Also, we provide an user-interface as a ***Progressive Web Application (PWA)*** to display the predicted results.



Conclusion

- We propose a ***novel end-to-end adaptive system*** that leverages heterogonous urban data to ***predict pollution concentrations and levels.***
- Vayuanukulani ***learns general importance*** by considering the ***relative importance of incoming streaming data*** using the attention mechanism in order to provide accurate predictions.
- Results show that our model ***leverages the incoming information*** and improves predictions for all the pollutants over time.
- We believe that our work can be an ***essential part toward building real-world pollution prediction systems.***

Code available at ***github.com/divyam3897/VayuAnukulani***

Thank you for listening!
Questions?

Acknowledgement: Dr. Aakanksha Chowdhery (Google Brain)
Central pollution control board (CPCB)
The Marconi Society and Celestini Project India