Multi-view Stacked CNN-BiLSTM (MvS CNN-BiLSTM) for Urban $PM_{2.5}$ Concentration Prediction of India's Polluted Cities

A dissertation-II submitted to the Mahatma Gandhi Central University
in partially fulfillment of the requirements
for the award of the degree of

MASTER OF TECHONOLGY

IN

COMPUTER SCIENCE & ENGINEERING

BY

SUBHAM KUMAR

DEPARTMENT OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
MAHATMA GANDHI CENTRAL UNIVERSITY,
MOTIHARI, BIHAR - 845401, INDIA

August 27, 2023

Multi-view Stacked CNN-BiLSTM (MvS CNN-BiLSTM) for Urban $PM_{2.5}$ Concentration Prediction of India's Polluted Cities

A dissertation-II submitted to the Mahatma Gandhi Central University
in partially fulfillment of the requirements
for the award of the degree of

MASTER OF TECHONOLGY

IN

COMPUTER SCIENCE & ENGINEERING

BY

SUBHAM KUMAR
(MGCU2021CSIT4029)

Under the Supervision of

Dr. VIPIN KUMAR

DEPARTMENT OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
MAHATMA GANDHI CENTRAL UNIVERSITY,
MOTIHARI, BIHAR - 845401, INDIA

August 27, 2023

कंप्यूटर विज्ञान और सूचना प्रौद्योगिकी विभाग
Department Of Computer Science And Information Technology
महात्मा गाँधी केन्द्रीय विश्वविद्यालय, बिहार-८४५४०१
MAHATMA GANDHI CENTRAL UNIVERSITY,
MOTIHARI, BIHAR - 845401, INDIA

DECLARATION

This is to certify that the dissertation-II entitled **" Multi-view Stacked CNN-BiLSTM (MvS CNN-BiLSTM) for Urban $PM_{2.5}$ Concentration Prediction of India's Polluted Cities"** is being submitted to the **Department Of Computer Science And Information Technology, Mahatma Gandhi Central University, Motihari, Bihar - 845401, India** in partial fulfillment of the requirements for the award of the degree of **Master of Technology** in **Computer Science & Engineering**, is a record of bonafide work carried out by me under the supervision of **"Dr. Vipin Kumar, Department Of Computer Science And Information Technology, Mahatma Gandhi Central University, Motihari, Bihar - 845401, India."**

The matter embodied in the dissertation has not been submitted in part or full to any University or Institution for the award of any other degree or diploma.

During the preparation of this work, I have not used any AI-based tool to write any part of this dissertation report. I take full responsibility for the submitted content including similarity.

Subham Kumar
(MGCU2021CSIT4029)
Department Of Computer Science And Information Technology
Mahatma Gandhi Central University,
Motihari, Bihar - 845401, India
Email id: - Subh700454@gmail.com

कंप्यूटर विज्ञान और सूचना प्रौद्योगिकी विभाग
Department Of Computer Science And Information Technology
महात्मा गाँधी केन्द्रीय विश्वविद्यालय, बिहार-८४५४०१
MAHATMA GANDHI CENTRAL UNIVERSITY,
MOTIHARI, BIHAR - 845401, INDIA

CERTIFICATE

This is to certify that the dissertation-II entitled **"Multi-view Stacked CNN-BiLSTM (MvS CNN-BiLSTM) for Urban $PM_{2.5}$ Concentration Prediction of India's Polluted Cities"** submitted by **Subham Kumar** to the **Department Of Computer Science And Information Technology, Mahatma Gandhi Central University, Motihari, Bihar - 845401, India** for the award of the degree of **Master of Technology** in **Computer Science & Engineering**, is a research work carried out by him under the supervision of **"Dr. Vipin Kumar, Department Of Computer Science And Information Technology, Mahatma Gandhi Central University, Motihari, Bihar - 845401, India."**

Head of Department

Prof. Vikas Pareek
Department Of Computer Science And Information Technology
Mahatma Gandhi Central University, Motihari, Bihar - 845401, India

Supervisor

Dr. Vipin Kumar
Department Of Computer Science And Information Technology
Mahatma Gandhi Central University, Motihari, Bihar - 845401, India

Abstract

The presence of $PM_{2.5}$ is a significant concern for human well-being and ecosystems. The practical measure of $PM_{2.5}$ is the vital problem worldwide. These tiny particles can quickly enter the respiratory system and deeply infiltrate the lungs, leading to various health issues, including respiratory disorders, cardiovascular diseases, and premature death. The literature shows that hybrid deep learning (DL) models are performing better than stand-alone DL models of time series (i.e., CNN, RNN, GRU, LSTM and BiLSTM) to predict the $PM_{2.5}$ pollutant, but effective performance is not achieved yet. In this research, the author has proposed a hybrid stacked CNN Bidirectional-LSTM model architecture that utilises the multiple views of the data corresponding to seasonal repetitions to induce the multiple models, called Multi-view Stacked CNN Bidirectional-LSTM (MvS CNN-BiLSTM). The proposed model has been deployed over seventeen univariate time series ($PM_{2.5}$) data of highly polluted Indian cities and stand-alone DL models. The performances of the proposed model have been compared using Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) measures. The average enhancement of the proposed model on all datasets has been achieved compared to stand-alone DL models as RMSE: 7.11% (CNN), 5.08% (RNN), 3.80% (GRU), 5.57% (LSTM) and 4.05% (BiLSTM) and MAPE: 27.16% (CNN), 28.52% (RNN), 26.22% (GRU), 27.22% (LSTM), 23.11% (BiLSTM). Moreover, the non-parametric statistical analysis (Friedman and Holmś) have been performed and proves that the proposed model MvS CNN-BiLSTM is performed as distinct and compelling over both performance measures.

Acknowledgements

This M.Tech dissertation-II is the result of hard work, upon which many people have contributed and given their support. I have made this dissertation on the topic **"Multi-view Stacked CNN-BiLSTM (MvS CNN-BiLSTM) for Urban $PM_{2.5}$ Concentration Prediction of India's Polluted Cities."** I have also tried my best in this dissertation to explain all the related detail. I would like to express my sincere gratitude towards my Superviser **Dr. Vipin Kumar**, Department of CS & IT , for providing excellent guidance, encouragement, inspiration, and constant and timely support throughout this M.Tech dissertation work. He taught me how to pursue the right aim towards the work , and showed me differnt ways to approach the research problem. His wide knowledge and logical ways of thinking have been great value for me, and his understanding and guidance have provided the successful completion of the Dissertation work.

First and foremost, I would like to express my gratitude to our beloved Dean of the Computational Sciences, Information and Communication Technology and Head of Department of Computer Science and Information Technology **Prof. Vikas Pareek**, for providing his kind support in various aspects. A special thanks to all the Respected Teachers **Dr. Sunil Kumar Singh**, and **Mr. Shubham Kumar**, of the Department of Computer Science and Information Technology.

I am always grateful to the university, our Honble Vice chancellor **Prof. Sanjay Srivastava** for providing such a good research environment.

Special thanks to Ph.D scholar, especially **Ritika Singh**, **Surbhi Kumari**, **Ibrahim Momin**, **Naushad Ahmad** and my friends **Tej Prakash**, **Gajendra Patel**, **Abhijeet Kumar**, **Amod Kumar**, **Rana Kumar**, **Krishna Murari**, **Rajan Kumar**, **Suraj**, **Md. Aamir Sohail**, **Shahzeb Khan**, and all my lovely juniors for their invaluable feedbacks, care, and moral support during this endeavor.

Mother and **Father**, it is impossible to thanks adequately for everything you have done, from loving me unconditionally to rising me in a stable household,

where your persistent efforts and traditional values taught your children to celebrate and embrace life. I could not have asked for better parents or role-models. You showed me that anything is possible with faith, hard work and determination.

Subham Kumar
(MGCU2021CSIT4029)
M.Tech(CSE)

List of Publications

1. Multi-view Stacked CNN-BiLSTM (MvS CNN-BiLSTM) for Urban $PM_{2.5}$ Concentration Prediction of India's Polluted Cities
juneral of cleaner protecation (Impact Factor: 11.1) Indexed by SCI (Submitted on 14^{th} Aug 2023)
Authors - Subham Kumar and Vipin Kumar

Contents

List of Figures

List of Tables

List of Abbreviations

$PM_{2.5}$	**Particulate Matter less then 2.5 μm**
AQI	**Air Quality Index**
ML	**Machine Learning**
DL	**Deep Learning**
SVM	**Support Vector Machine**
ARIMA	**Autoregressive Integrated Moving Average**
LR	**Linear Regression**
ANN	**Artificial Neural Network**
FL	**Fuzzy Logic**
LSTM	**Long Short Term Memory**
GRU	**Gated Recurrent Units**
CNN	**Convolutional Neural Network**
MVMT	**Multi-view multi-task**
RNN	**Recurrent Neural Network**
BiLSTM	**Bi-directional Long Short Term Memory**
MvS CNN-BiLSTM	**Multi-view Stacked CNN-BiLSTM**
RMSE	**Root Mean Square Error**
MAPE	**Mean Absolute Percentage Error**
MAE	**Mean Absolute Error**
CPCB	**Central Pollution Control Board**

List of Symbols

F	**Filter**
b	**Bias**
σ	**Activation Function**
$\odot$	**Multiplication represented**
W	**Weight**
t	**Time Step**
x_i	**Missing Value**
D	**Dataset**
D_c	**Dataset chunk**
D_s	**No. of Data point available in chunk**
X	**Univariate time series**
X_v	**v^{th}-view of univariate time series**
$\mathscr{L}$	**Lag**
$\mathscr{L}^+_{lowest}$	**Lowest Positive Lag**
X^T	**Tread**
X^S	**Seasonal**
X^R	**Reminder**
$\rho_{\mathscr{L}}$	**Auto correlation function**

Dedicated

to

Maa, and Papajee

STATEMENT OF THESIS PREPARATION

1. Thesis Title:
   "Multi-view Stacked CNN-BiLSTM (MvS CNN-BiLSTM) for Urban $PM_{2.5}$ Concentration Prediction of India's Polluted Cities"
2. Degree for which the thesis is submitted: **Master of Technology .**
3. The thesis Guide was referred for preparing the thesis.
4. Specifications regarding the thesis format have been closely followed.
5. The contents of the thesis have been organized based on the guidelines.
6. The thesis has been prepared without resorting to plagiarism.
7. All sources used have been cited appropriately.
8. The thesis has not been submitted elsewhere for a degree.
9. A copy of Research Article submission/Publication Certificate(s)/Proof of Journal/Conference (National/International) based on dissertation work done during the session is mandatory to submit i.e. forwarded by supervisor(s).

SUBHAM KUMAR
Enrolment No.: MGCU2021CSIT4029
Department Of Computer Science And Information Technology
Mahatma Gandhi Central University, Motihari, Bihar - 845401, India
Email id - Subh700454@gmail.com

Chapter 1

Introduction

1.1 Introduction

Air pollution has become a major global problem due to industrialisation and urbanisation. The rising levels of air pollutants, such as CO, SO, O_3, PM_{10} and $PM_{2.5}$. It has led to environmental issues like soil acidification, fog, haze and severe health problems such as heart attacks and lung diseases. The World Health Organization has revealed that contaminated air affects most of the global population, approximately 90% [1]. $PM_{2.5}$, or 2.5 micrometres or less aerodynamic diameter particulate matter, is an essential factor in calculating the Air Quality Index (AQI). AQI is a numerical scale used to communicate how polluted the air is and its potential health effects to the public. $PM_{2.5}$ is a crucial air pollutant that can infiltrate the respiratory system and have detrimental health consequences. Time series data is a sequential data type collected regularly, with time as the index. It involves examining trends and patterns, with stationarity important, implying statistical properties' constancy over time. Forecasting based on historical patterns finds widespread applications in various fields, providing valuable decision-making and predictive modelling insights.

Three principal methods, namely deterministic, statistical, and Machine learning (ML)/Deep Learning (DL), are widely utilised in predicting air quality. Deterministic methods simulate atmospheric chemistry's dispersion and transport processes, but they can be computationally expensive and less accurate due to limited actual

observations. Statistical methods rely on historical data to forecast pollutant concentrations, but their linear assumptions may limit prediction performance. Researchers are incorporating non-linear machine learning models to surpass these constraints as alternative methods for predicting air quality. Machine learning and Deep Learning models such as Support Vector Machines (SVM) [2], Autoregressive-Integrated Moving Average (ARIMA) [3], Linear Regression(LR) [4], Artificial Neural Networks (ANNs) [5] and Fuzzy Logic (FL) [6] have been applied in air quality prediction studies. ANNs have been particularly popular, showing promising results in various applications. However, the rapid development of deep learning techniques has outperformed traditional ML models. Deep learning models, such as Long Short-Term Memory (LSTM) [7], Gated Recurrent Units (GRU), and Convolutional Neural Networks (CNN) [8], have shown improved prediction performance by capturing long-term dependencies and spatial features in air quality data.

Multi-view learning [9], [10] has emerged as a potent methodology in machine learning and deep learning. It effectively utilises multiple perspectives or representations of data to enhance predictive performance, improve generalisation, and address intricate real-world problems. The technique has garnered significant attention due to its ability to manage varied and complementary information from multiple sources or modalities. This approach holds promise in its application and usability for ML/DL models across different domains, including Text and Image Analysis [11]–[14], Audio and Video Processing [13]–[16], and Environmental Monitoring [17]. Combining multi-view incorporation and hybrid deep learning models is a powerful technique in modern machine learning. This methodology improves predictive accuracy and feature extraction by integrating different data perspectives and utilising diverse neural network architectures.

In time series analysis, the potential of multi-view learning has been shown by [18], which has utilised multivariate heterogeneous features as multiple views such as geographic location and time of day. The author has compared the proposed multi-view multi-task (MVMT) model with a single-view dataset setting and shows better performance as an outcome. In another research [19], a multi-view learning framework has been deployed for adaptive transfer learning to show the effectiveness of the inter-view usability of information transfer. In evaluating the proposed

model, the time series classification task has been performed to get the generalised evaluations. Several multi-view learning for time series approaches corresponding to deep learning with time series [20], [21], machine learning with time series [21], transfer learning [18], [20], etc. It has been observed from the literature that no research has been conducted yet based on a univariate time series dataset. Therefore, the potential of multi-view learning over univariate data may have an opportunity to perform time series prediction effectively.

In this research, the author has proposed a hybrid multi-views stacked CNN-BiLSTM model framework. The hybrid CNN-BiLSTM architecture has been utilised to build the two-stack network. The seasonal characteristics of the univariate data have been utilised to generate the views corresponding to the required number of views (called partitions of the data). Then, a stacked CNN-BiLSTM model was deployed over each data view for the predictions. After this, predictions of view models are ensembled to get the final prediction based on their validation performance at each view. The proposed model has been evaluated over seventeen univariate datasets of $PM_{2.5}$ pollutants and compared their effectiveness based on various performance measures.

Chapter 2

Conclusion

2.1 Conclusion

In this study, a hybrid MvS CNN-BiLSTM model has been proposed. The model utilises stacked 1D CNN & BiLSTM as new architecture to view the univariate data $PM_{2.5}$ pollutant. Additionally, a novel multi-view approach has been proposed for univariate time series, which exploits the seasonal characteristics of the data to construct the views corresponding to the lowest lag. The seventeen datasets ($PM_{2.5}$) of highly polluted cities of India have been used to deploy the proposed MvS CNN-BiLSTM and stand-alone DL models. The evaluation of the models has been performed using RMSE and MAPE. The results have shown that the proposed model has improved the performed than corresponding stand-alone DL models over RMSE: 7.11% (CNN), 5.08% (RNN), 3.80% (GRU), 5.57% (LSTM) and 4.05% (BiLSTM) and MAPE: 27.16% (CNN), 28.52% (RNN), 26.22% (GRU), 27.22% (LSTM), 23.11% (BiLSTM). Moreover, the enhanced performance of the proposed model has also been validated using non-parametric statistical methods, i.e., Friedman ranking and Holm's procedure. The statistical analysis of results concludes that the proposed model performance is better and distinct from stand-alone DL models.

Future research: The proposed model has utilised two-stacked CNN-BiLSTM for this research, where multiple combinations of stand-alone DL models may be investigated along with the stacking of the hybrid for better performance. In a multi-view approach, apart from seasonal characteristics, trends and the remainder may be utilised to generate the views of the dataset, yielding enhanced performance.

Moreover feature set partitioning method of multi-view learning may be utilised for multivariate time series data (single source or multi-source data).

Data availability

Data will be available on CPCB (India) Webportel.The Central Pollution Control Board (CPCB) in India maintains a web portal that offers access to various environmental datasets, including air and water quality, emission inventories, and pollution monitoring data, aimed at promoting environmental awareness and research in the country.

Acknowledgments

We extend our heartfelt appreciation to the Central Pollution Control Board (CPCB), India, for generously providing invaluable environmental data, which significantly enhanced our research and played a pivotal role in completing this study.

References

[1] C. Zhou, G. Wei, H. Zheng, *et al.*, "Effects of potential recirculation on air quality in coastal cities in the yangtze river delta," *Science of the total environment*, vol. 651, pp. 12–23, 2019.

[2] K.-P. Lin, P.-F. Pai, and S.-L. Yang, "Forecasting concentrations of air pollutants by logarithm support vector regression with immune algorithms," *Applied Mathematics and Computation*, vol. 217, no. 12, pp. 5318–5327, 2011.

[3] S. Kumari and S. K. Singh, "Machine learning-based time series models for effective co2 emission prediction in india," *Environmental Science and Pollution Research*, pp. 1–16, 2022.

[4] S. Kumari and S. K. Singh, "Deep learning-based time series models for gdp and ict growth prediction in india," in *2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, IEEE, 2022, pp. 250–256.

[5] O. Taylan, "Modelling and analysis of ozone concentration by artificial intelligent techniques for estimating air quality," *Atmospheric environment*, vol. 150, pp. 356–365, 2017.

[6] B. Wang and Z. Chen, "A model-based fuzzy set-owa approach for integrated air pollution risk assessment," *Stochastic Environmental Research and Risk Assessment*, vol. 29, pp. 1413–1426, 2015.

[7] E. Kristiani, H. Lin, J.-R. Lin, Y.-H. Chuang, C.-Y. Huang, and C.-T. Yang, "Short-term prediction of pm2. 5 using lstm deep learning methods," *Sustainability*, vol. 14, no. 4, p. 2068, 2022.

[8] Y. A. Ayturan, Z. C. Ayturan, and H. O. Altun, "Air pollution modelling with deep learning: A review," *International Journal of Environmental Pollution and Environmental Modelling*, vol. 1, no. 3, pp. 58–62, 2018.

[9] J. Zhao, X. Xie, X. Xu, and S. Sun, "Multi-view learning overview: Recent progress and new challenges," *Information Fusion*, vol. 38, pp. 43–54, 2017.

[10] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *arXiv preprint arXiv:1304.5634*, 2013.

[11] X. Yang, S. Feng, D. Wang, and Y. Zhang, "Image-text multimodal emotion classification via multi-view attentional network," *IEEE Transactions on Multimedia*, vol. 23, pp. 4014–4026, 2020.

[12] F. Nie, G. Cai, J. Li, and X. Li, "Auto-weighted multi-view learning for image clustering and semi-supervised classification," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1501–1511, 2017.

[13] X. Yan, S. Hu, Y. Mao, Y. Ye, and H. Yu, "Deep multi-view learning methods: A review," *Neurocomputing*, vol. 448, pp. 106–129, 2021, ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2021.03.090.

[14] A. Kumar and J. Yadav, "A review of feature set partitioning methods for multi-view ensemble learning," *Information Fusion*, vol. 100, p. 101 959, 2023, ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2023.101959.

[15] E. Garcia-Ceja, C. E. Galván-Tejada, and R. Brena, "Multi-view stacking for activity recognition with sound and accelerometer data," *Information Fusion*, vol. 40, pp. 45–56, 2018.

[16] T. Hussain, K. Muhammad, W. Ding, J. Lloret, S. W. Baik, and V. H. C. de Albuquerque, "A comprehensive survey of multi-view video summarization," *Pattern Recognition*, vol. 109, p. 107 567, 2021.

[17] X. Huang, D. Wen, J. Li, and R. Qin, "Multi-level monitoring of subtle urban changes for the megacities of china using high-resolution multi-view satellite imagery," *Remote sensing of environment*, vol. 196, pp. 56–75, 2017.

[18] J. Deng, X. Chen, R. Jiang, X. Song, and I. W. Tsang, "A multi-view multi-task learning framework for multi-variate time series forecasting," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 8, pp. 7665–7680, 2023. DOI: 10.1109/TKDE.2022.3218803.

[19] D. Zhan, S. Yi, D. Xu, *et al.*, "Adaptive transfer learning of multi-view time series classification.," *arXiv: Learning*, 2019.

[20] S. D. Bhattacharjee, W. J. Tolone, A. Mahabal, *et al.*, "Multi-view, generative, transfer learning for distributed time series classification," 2019.

[21] S. D. Bhattacharjee, W. J. Tolone, A. Mahabal, M. Elshambakey, I. Cho, and G. Djorgovski, "View-adaptive weighted deep transfer learning for distributed time-series classification," 2019.