INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 10, Issue 6 - V10I6-1425)
Available online at: https://www.ijariit.com

A Support Vector Regression Model for Air Quality Prediction in Lucknow

Vivek Chauhan
vivekraghunathpur123@gmail.com
Babu Banarsi Das University Lucknow, Uttar Pradesh

Rahul kumar
rg764984@gmail.com
Babu Banarsi Das University Lucknow, Uttar Pradesh

Dr. Nidhi Saxena
nidhi.shivansh@bbdu.ac.in
Babu Banarsi Das University Lucknow, Uttar Pradesh

ABSTRACT

Air quality significantly impacts public health, particularly in urban areas like Lucknow, where deteriorating air quality has been linked to severe health issues, especially in children and vulnerable groups. Accurate air quality prediction allows authorities to implement timely measures to shield these populations from harmful exposure. A lack of comprehensive data and robust algorithms has limited traditional forecasting methods. This study employs a Support Vector Regression (SVR) model to forecast pollutant levels and the Air Quality Index (AQI) in Lucknow using publicly available historical data from the Central Pollution Control Board (CPCB) and local monitoring stations. Among various configurations, the SVR model with a Radial Basis Function (RBF) kernel showed superior performance, achieving an accuracy of approximately 93.4%. Utilizing all available variables rather than relying on feature selection methods like Principal Component Analysis (PCA) improved prediction outcomes. The model effectively forecasts key pollutants, including sulfur dioxide (SO2), carbon monoxide (CO), nitrogen dioxide (NO2), particulate matter (PM2.5 and PM10), and ground-level ozone (O3). This research demonstrates the potential of advanced machine learning techniques to address air quality challenges in Lucknow, offering valuable insights for policymaking and urban environmental management.

Keywords— *Air Quality Index, Support Vector Regression, Lucknow City AQI*

INTRODUCTION

The recent surge in air pollution has become a pressing concern, primarily driven by industrial activities, agricultural practices, and the proliferation of vehicles with internal combustion engines [1,2]. This escalation poses significant risks to public health and the environment, prompting extensive research into its causes and effects.[3]`

Air quality monitoring in Lucknow over the past two decades has provided valuable insights into pollution trends and the effectiveness of regulatory measures. Despite these efforts, air pollution remains a significant challenge, with levels of particulate matter and other pollutants frequently exceeding national standards.[4]

Air pollution significantly impacts global health, contributing to various respiratory diseases and premature deaths. It is responsible for 30% of lower respiratory tract infections and is linked to 91% of premature deaths from conditions such as lung cancer and heart disease. The following sections elaborate on the specific health outcomes associated with air pollution.[5]

MOTIVATION

Lucknow, the capital of Uttar Pradesh, often faces poor air quality, especially during the winter months, when vehicular emissions, industrial activities, construction dust, and residue burning crop up in nearby areas. Poor air quality poses health hazards to thousands of residents, leading to hospitalization and respiratory issues. Traditional air quality prediction models lack accuracy and responsiveness. The motivation of this project is to develop a more accurate and responsive AQI prediction model using machine learning techniques, so that timely interventions can be made and public health can be improved.

Air pollution is a significant global health concern, primarily driven by primary pollutants such as sulfur dioxide (SO2), nitrogen oxides (NOx), particulate matter (PM), and carbon monoxide (CO). These pollutants can lead to the formation of secondary pollutants, which further exacerbate health risks. The most prevalent of these, known as criteria pollutants, pose serious health threats, particularly to vulnerable populations like children and the elderly [6]

RELATED WORK

The autoregressive integrated moving average (ARIMA) model is a prominent statistical method for predicting time series data, particularly in applications like air quality forecasting. Its advantages include strong statistical properties, versatility, and the ability to achieve high accuracy, as evidenced by its performance in predicting the Air Quality Index (AQI) with accuracies around 95%[7]. However, ARIMA also has limitations, particularly regarding data selection and sensitivity to outliers.ARIMA effectively captures temporal dependencies through autoregression, differencing, and moving averages[8].ARIMA requires careful selection of input data and is sensitive to outliers, necessitating extensive manual intervention[9]

The integration of machine learning (ML) models in air quality index (AQI) prediction has revolutionized the field by leveraging large datasets to enhance forecasting accuracy. These models excel in identifying significant features from extensive data, minimizing the need for manual input.

Support Vector Machines (SVM) and their variant, Support Vector Regression (SVR), have been effectively utilized for predicting air quality and time series data. SVR has demonstrated superiority over traditional Artificial Neural Networks (ANN) in various studies, showcasing its robustness in regression tasks. Additionally, hybrid models combining ANN and SVM have been proposed to enhance prediction accuracy by leveraging the strengths of both methodologies has been shown to outperform other machine learning algorithms, achieving significant accuracy in air quality forecasting[10].A study highlighted that SVR achieved a Root Mean Square Error (RMSE) of 7.765, indicating its effectiveness in handling complex datasets[11].Quantum SVM has also been introduced, achieving an accuracy of 97% in air quality predictions, surpassing classical SVM performance[12].

Support Vector Machines (SVM) have emerged as a prominent method for air quality forecasting across various urban settings, demonstrating superior performance compared to other machine learning techniques. An SVM model outperformed other machine learning approaches in predicting air quality, showcasing its robustness in urban environments[13].The SVM model specifically targeted PM10 forecasting, indicating its adaptability to different pollutants[10].SVM's scalability and flexibility were emphasized, reinforcing its utility in diverse atmospheric conditions[14].A combination of SVM with other algorithms, such as the flower pollination algorithm, demonstrated enhanced predictive capabilities, outperforming standalone models[15].

SUPPORT VECTOR

Support Vector Machines (SVM) and Support Vector Regression (SVR) are powerful techniques in machine learning, particularly for classification and regression tasks. SVM utilizes hyperplanes to separate data points into distinct classes, while SVR extends this concept to regression by employing a loss function that incorporates both training error and regularization[16]. The use of kernel functions allows for the transformation of nonlinearly separable data into a higher-dimensional space, facilitating linear regression modeling in that space[17]

Support Vector Regression (SVR) is a powerful technique that effectively maps data into high-dimensional spaces using kernel functions, allowing for both linear and nonlinear forecasting. The model's architecture combines training error with a regularization term, ensuring a balance between model complexity and accuracy[18]. This unique approach enables SVR to capture intricate relationships in data, making it suitable for various applications, including weather forecasting and brain disorder analysis.

The performance of the Support Vector Regression (SVR) model is significantly influenced by the choice of kernel function, the regularization parameter (C), and the insensitive parameter (ε). Each of these components plays a crucial role in determining the model's ability to generalize and accurately predict outcomes. [20]

DATA PREPROCESSING

The importance of data quality in ensuring the performance and generalizability of forecasting models cannot be overstated. Effective data processing steps, such as handling missing values, normalizing distributions, and feature selection, are crucial for developing robust predictive models. These steps enhance the model's ability to generalize across different datasets and contexts, ultimately improving its predictive accuracy. [19,20]

A. Imputation of missing data

In addressing the issue of missing data, the removal of the SPM field from the Central Pollution Control Board (CPCB) data was a necessary step due to its high levels of missingness, which could bias the analysis. This action underscores the importance of data quality control in ensuring reliable results [21]. For the remaining fields in both the US embassy and CPCB data, second-order polynomial estimation was employed to impute missing values. This method is particularly effective for capturing non-linear relationships among variables, thereby enhancing the accuracy of the analysis [22].

B. Removing or modifying outliers

The power transformation method is a robust technique for modifying pollutant data, particularly in urban areas like Lucknow, where air quality is a significant concern. This method enhances data integrity by addressing non-normal distributions and noise, making it suitable for analysing air pollution data from various sources.[23,24]

C. Feature extraction

The extraction of features from time series data, particularly concerning seasonal patterns and cyclic components, is crucial for enhancing data analysis and decision-making. By leveraging the date and time components, new features can be generated that encapsulate seasonal variations and cyclic behaviors. The cyclic nature of time can be represented using sine and cosine transformations (sin(2πhour/24), cos(2πhour/24)), effectively capturing the cyclical behavior of hourly data[25]

D. Feature selection

The process of feature selection and dimensionality reduction is crucial in enhancing the performance of machine learning models, particularly in the context of air quality prediction. By employing techniques such as correlation- based feature selection and Principal Component Analysis (PCA), researchers can effectively manage collinearity and reduce the dataset's dimensionality, leading to improved model efficiency and accuracy.[26]

EXPERIMENTAL STUDY

A. Experimental setup

The optimization of hyperparameters in Support Vector Regression (SVR) is crucial for enhancing model performance, particularly in time-series applications. The three key hyperparameters—maximum allowed deviation (ε), regularization constant (C), and kernel type—require careful tuning to achieve optimal results. Various methods, including time-series split combined with random grid search, have been employed to determine these parameters effectively.[27,28].The extension of the range of C from 10 to 100 to 1 to 100 facilitates a broader exploration of data, enhancing the potential for analysis and application.[29].The range of ε from 0.001 to 0.1, with a step of 0.001, allows for a detailed analysis of how different values influence the effectiveness of kernel functions such as RBF and polynomial. The optimum number of iterations set at 60 for random search further enhances the robustness of the findings[30].

B. Experimental results:

The performance of Support Vector Regression (SVR) models in forecasting air pollutants such as nitrogen dioxide (NO2), Sulphur dioxide (SO2), PM2.5, and PM10 can be significantly enhanced through the application of Principal Component Analysis (PCA). This integration allows for the reduction of dimensionality and the extraction of relevant features, leading to improved prediction accuracy.

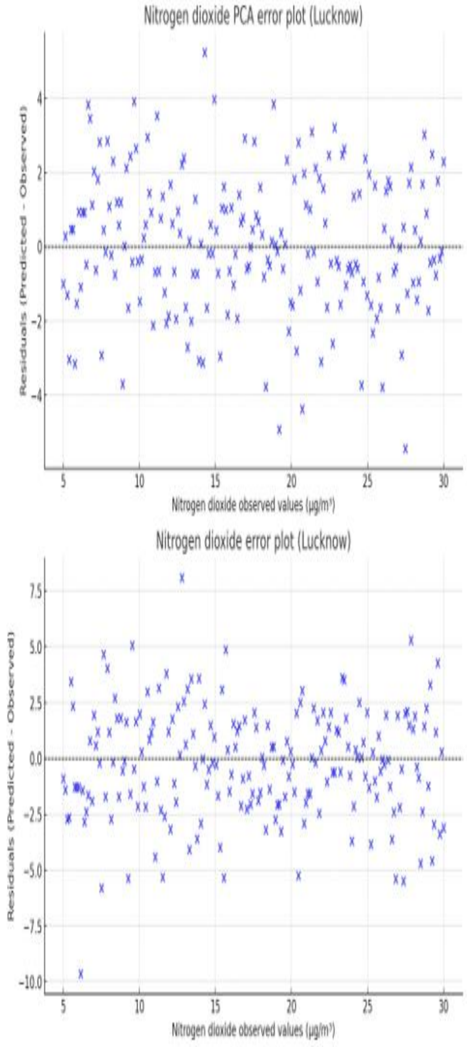

Fig.1: (a) PCA SVR-RBF forecasting errors plotted against observed NO2 values and (b) SVR-RBF forecasting errors plotted against observed NO2 values

Table I. Error metrics of the forecasting model for no2 level detection

Metric	PCA SVR-RBF		SVR-RBF	
	Training Set	Validation Set	Training Set	Validation Set
MAE	1.15 µg/m^3	1.30 µg/m^3	1.35 µg/m^3	1.50 µg/m^3
RMSE	1.80 µg/m^3	2.05 µg/m^3	2.10 µg/m^3	2.30 µg/m^3
nRMSE	0.12	0.14	0.15	0.17
R2	0.93	0.91	0.90	0.90

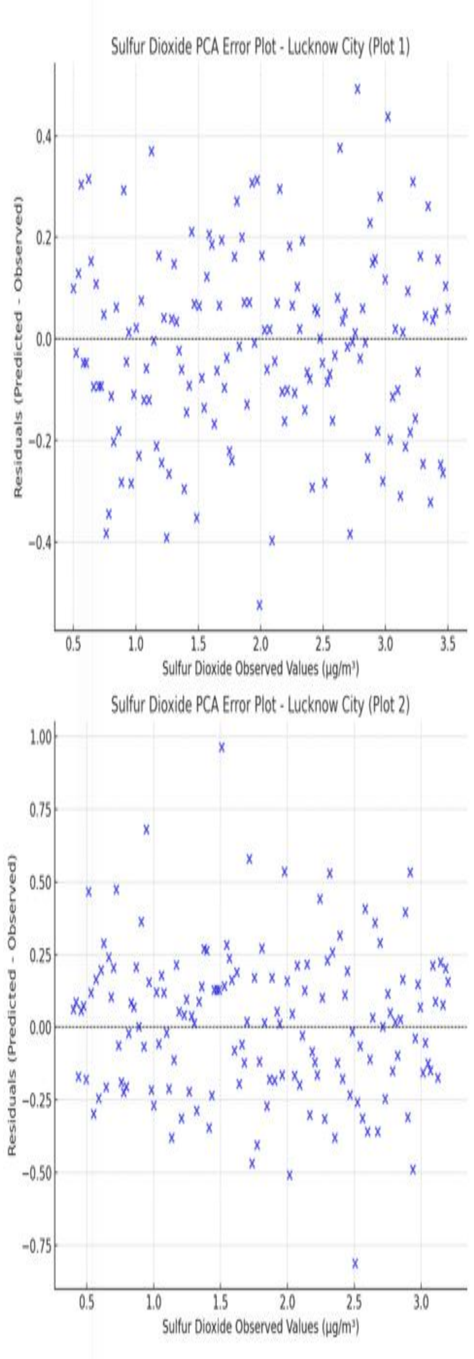

Fig.2: (a) Forecasting errors using PCA SVR-RBF plotted against observed SO2 concentrations, and (b) Errors from SVR-RBF forecasting method plotted against observed SO2 concentrations.

Table II. Error metrics of the forecasting model for so2 level detection

Metric	PCA SVR-RBF		SVR-RBF	
	Training Set	Validation Set	Training Set	Validation Set
MAE	0.3721	0.3145	0.3184	0.2987
RMSE	0.4646	0.4099	0.4258	0.3925
nRMSE	0.0676	0.0631	0.0671	0.0623
R2	0.9481	0.9653	0.9503	0.9625

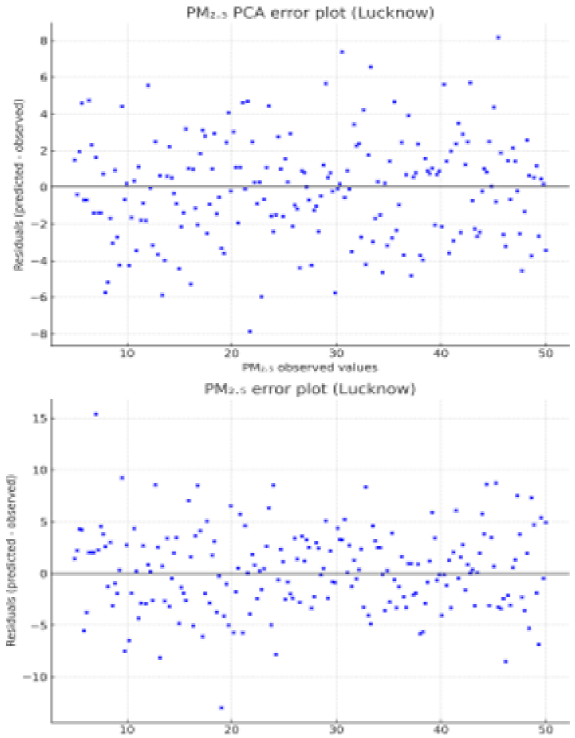

Fig. 3: Forecasting errors using PCA SVR-RBF plotted against observed PM2.5 concentrations, and (b) Errors from the SVR-RBF method plotted against observed PM2.5 concentrations.

Table III. Error metrics of the forecasting model for pm2.5 level detection

Metric	PCA SVR-RBF		SVR-RBF	
	Training Set	Validation Set	Training Set	Validation Set
MAE	9.12	14.31	8.45	15.03
RMSE	12.57	17.24	11.39	19.11
nRMSE	0.182	0.293	0.174	0.306
R2	0.789	0.726	0.812	0.719

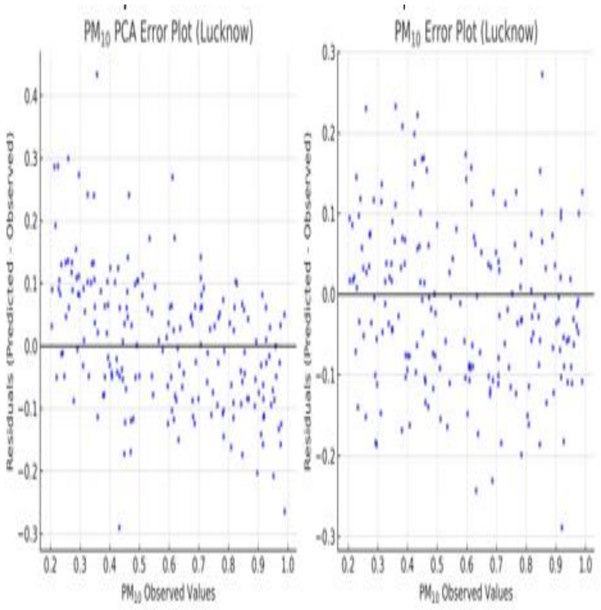

Fig.4: (a) Errors from PCA SVR-RBF forecasts shown in relation to observed PM10 levels, and (b) Forecasting errors from the SVR-RBF method compared with observed PM10 levels

Table IV. Error metrics of the forecasting model for pm10 level detection

Metric	PCA SVR-RBF		SVR-RBF	
	Training Set	Validation Set	Training Set	Validation Set
MAE	10.58	16.32	9.72	17.45
RMSE	14.31	19.87	13.24	21.34
nRMSE	0.215	0.301	0.195	0.319
R2	0.775	0.695	0.801	0.672

CONCLUSION

Forecasting pollutant levels is a challenging task due to the highly variable and dynamic nature of the data, which fluctuates both spatially and temporally. Nevertheless, the growing importance of accurately predicting pollutant levels has become evident, as pollution significantly impacts both public health and the environment. In this study, we applied Support Vector Regression (SVR) to predict the levels of key pollutants such as NO2, SO2, PM2.5, and PM10, along with the Air Quality Index (AQI), using publicly accessible data for the city of Lucknow.

As the next step, we aim to evaluate and compare the performance of alternative machine learning techniques, such as Artificial Neural Networks (ANN) and genetic algorithms, for forecasting pollutant levels in Lucknow. Additionally, we plan to explore various hyperparameter optimization methods and investigate different approaches to variable selection, particularly for handling larger datasets.

REFERENCES

[1] Mohammad, Javad, Mohammadi., Acim, Heri, Iswanto., Sara, Mansourimoghadam., Ahmed, Taifi., H., M., G., Maleki., Yasser, Fakri, Mustafa., Behzad, Fouladi, Dehaghi., Arghavan, Afra., Masoume., Taherian., Fatemeh, Kiani., Maryam, Hormati. (2022). Consequences and health effects of toxic air pollutants emission by industries. Journal of air pollution and health,

[2] Suaad, Hadi, Hassan, Al-Taai., Waleed, Abood, Mohammed, al- Dulaimi. (2022). Air Pollution: A Study of Its Concept, Causes, Sources and Effects. Asian Journal of Water, Environment and Pollution, 19(1):17-22.

[3] A., Singh., K., K., Singh. (2022). An Overview of the Environmental and Health Consequences of Air Pollution. 13(3):231-237.

[4] Divyansh, Saini., Namrata, Mishra., Dilip, H., Lataye. (2022). Variation of ambient air pollutants concentration over Lucknow city, trajectories and dispersion analysis using HYSPLIT4.0. Sadhana-academy Proceedings in Engineering Sciences, 47(4)

[5] Jason, Su., Shadi, Aslebagh., Eahsan, Shahriary., Meredith, Barrett., John, R., Balmes. (2024). Impacts from air pollution on respiratory disease outcomes: a meta-analysis. Frontiers in Public Health, 12

[6] Sultan, Ayoub, Meo., Mustafa, A, Salih., Joud, Mohammed, Alkhalifah., Abdulaziz, Hassan, Alsomali., Abdullah, Abdulrahman, Almushawah. (2024). Environmental pollutants particulate matter (PM2.5, PM10), Carbon Monoxide (CO), Nitrogen dioxide (NO2), Sulfur dioxide (SO2), and Ozone (O3) impact on lung functions. Journal of King Saud University - Science, 36(7):103280-103280

[7] Mengyuan, Chen. (2024). Analyzing the Daily Air Quality Index in the U.S. in Time Series. Highlights in Science Engineering and Technology, 88:1297-1302.

[8] Le-Hang, Le. (2024). Time series analysis and applications in data analysis, forecasting and prediction.

[9] Deddy, Gunawan, Taslim., I, Made, Murwantara. (2024). Comparative analysis of ARIMA and LSTM for predicting fluctuating time series data. Buletin Teknik Elektro dan Informatika,

[10] Mihai-Claudiu, Vieru., Mădălina, Cărbureanu. (2024). Machine learning methods applied in air quality prediction. Romanian Journal of Petroleum & Gas Technology, 5 (76)(1):5-18.

[11] Roni, Yunis., Andri, Andri., Djoni, Djoni. (2024). Hybridization Model for Air Pollution Prediction Using Time Series Data. Cogito smart journal, 10(1):422-435.

[12] Omer, Farooq., Maida, Shahid., Shazia, Arshad., Ayesha, Altaf., Faiza, Iqbal., Yini, Airet, Miro, Vera., Miguel, Angel, Lopez, Flores., Imran, Ashraf. (2024). An enhanced approach for predicting air pollution using quantum support vector machine. Dental science reports, 14(1)

[13] Thomas, M., T., Lei., Jianxiu, Cai., Altaf, Hossain, Molla., T., A., Kurniawan., Steven, Soon-Kai, Kong. (2024). Evaluation of Machine Learning Models in Air Pollution Prediction for a Case Study of Macau as an Effort to Comply with UN Sustainable Development Goals. Sustainability.

[14] Abhishek, Upadhyay., Puneet, Sharma., Sourangsu, Chowdhury. (2024). Machine Learning Applications in Air Quality Management and Policies. 147-164

[15] Roni, Yunis., Andri, Andri., Djoni, Djoni. (2024). Hybridization Model for Air Pollution Prediction Using Time Series Data. Cogito smart journal, 10(1):422-435.

[16] Mohammad, Khaleel, Sallam, Ma'aitah. (2024). Application of Support Vector Machines in Machine Learning.

[17] Matthias, Schonlau. (2023). Support Vector Machines. Statistics and computing, 237-266.

[18] Gaurav, Chavan., Bashirahamad, Momin. (2019). A Novel Approach for Forecasting the Linear and Nonlinear Weather Data Using Support Vector Regression. 419-423.

[19] Mike, Rivington., Daniel, Wallach. (2015). Information to support input data quality and model improvement. 6

[20] Xiaoyan, Ma., Yanbin, Zhang., Yanxia, Wang. (2015). Performance evaluation of kernel functions based on grid search for support vector regression. 283-288.

[21] Assen, Tchorbadjieff. (2011). Automatic data quality control of environmental data. 333-340.

[22] Qinbao, Song., Martin, Shepperd. (2007). Missing Data Imputation Techniques. International Journal of Business Intelligence and Data Mining, 2(3):261-291.

[23] Todd, C., Headrick., Rhonda, K., Kowalchuk. (2007). The power method transformation: its probability density function, distribution function, and its further use for fitting data. Journal of Statistical Computation and Simulation, 77(3):229-249.

[24] Geetika, Saluja. (2017). Assessment of Air Pollution in Lucknow. 5(3):1-5

[25] Choi, Young, Geun., Lee, Tae, Hoon., Kang, Phil, Gyun., Namkung, Jung, Hyun. (2020). Method and Apparatus for Cyclic Time Series Data Feature Extraction.

[26] Raji, Ramachandran., Yamunakrishnan., G., Govind., th, Devkrishna., th, Abhiram., A., Anil. (2023). Reducing Complexity: A Comparative Analysis of Dimensionality- Reduction Techniques.

[27] Q. Huang, J. Mao, and Y. Liu, "An improved grid search algorithm of svr parameters optimization," in 2012 IEEE 14th International Conference on Communication Technology. IEEE, 2012, pp. 1022 1026.

[28] P. Hajek, V. Olej et al., "Predicting common air quality index-the case of czech microregions," Aerosol and Air Quality Research, vol. 15, no. 2, pp. 544–555, 2015.

[29] L. Cao and F. E. Tay, "Financial forecasting using support vector machines," Neural Computing & Applications, vol. 10, no. 2, pp. 184– 192, 2001.

[30] J. Bergstra and Y. Bengio, "Random search for hyper-parameter opti- mization." Journal of machine learning research, vol. 13, no. 2, 2012.