МАТЕМАТИЧЕСКИЕ МЕТОДЫ СИСТЕМНОГО АНАЛИЗА, УПРАВЛЕНИЯ И МОДЕЛИРОВАНИЯ

EVALUATING THE PERFORMANCE OF SUPPORT VECTOR MACHINES BASED ON DIFFERENT KERNEL METHODS FOR FORECASTING AIR POLLUTANTS

© 2020 A. Masih✉, A. N. Medvedev

Ural Federal University named after the first President of Russia B. N. Yeltsin
51, Lenina Avenue, 620075 Yekaterinburg, Russian Federation

Annotation. The alarming level of air pollution in urban centres is an urgent threat to human health. Its consequences can be measured in terms of health issues experienced by children, an increasing numbers of heart and lung diseases, and, most importantly, the number of pollution related deaths. That is why a lot of attention has recently been paid to air pollution monitoring and prediction modelling. In order to develop prediction models, the study uses Support Vector Machines (SVM) with linear, polynomial, radial base function, normalised polynomial, and Pearson VII function kernels to predict the hourly concentration of pollutants in the air. The paper analyses the monitoring dataset of air pollutants and meteorological parameters as input variable to predict the concentrations of various air pollutants. The prediction performance of the models was assessed by using evaluation metrics, namely the correlation coefficient, root mean squared error, relative absolute error, and relative root squared error. To validate the model, the accuracy of the predictive algorithm was tested against two widely and commonly applied regression approaches called multilayer perceptron and linear regression. Furthermore, back check prediction test was performed to examine the consistency of the models. According to the results, the Pearson VII function and normalised polynomial kernel yield the most accurate results in terms of the correlation coefficient and error values to predict the concentrations of atmospheric pollutants as compared to other SVM kernels and traditional prediction models.

Keywords: Support Vector Machines, pollution prediction modelling, Pearson VII universal kernel, normalised polynomial kernel, linear kernel, multilayer perceptron, linear regression.

INTRODUCTION

According to World Health organization (WHO) air pollution is a biggest health risk because besides killing millions of people every year, it shortens the life expectancy as well [1]. In view of a report from American Lung Association [2], it is concluded that: a) a slight increase of 10 parts per billion (ppb) in Ozone (O_3) mixing can cause over 3700 premature deaths; b) the increased concentrations of $PM_{2.5}$ in air has a serious concern because, these particles due to their tiny size can deposit into the wind pipe and lung exchange region called as alveoli, and c) SO_2 is also an important precursor due to its strong association with respiratory diseases [3]. Crossing a certain limit, all air pollutants become dangerous to human health, however, due to number of reasons atmospheric concentrations of SO_2 and NO_2 in particular are considered extremely harmful to public health because a short exposure to such pollutants can aggravate human respiratory system [4]. Therefore, for air quality management and effective policy making, besides strict monitoring, the development of accurate prediction models is equally essential.

✉ Masih Adven
e-mail: adven.masikh@urfu.ru

According to several studies, environmental parameters, regional and synoptic meteorology can seriously influence air pollutant concentrations [5]. For example the ground level O_3 concentration over Chicago according to Holloway [6] is sensitive to air temperature, wind speed and direction, relative humidity, incoming solar radiation, and cloud cover. Various meteorological compositions and their interaction with light are important e.g. higher ambient temperature and solar intensity speed up photochemical reactions which lead to the formation of air pollutants, likewise wind speed and humidity can directly affect air pollutants dispersion [7].

Pollution prediction modelling deals with the concentration of atmospheric gases and their connection with the regional meteorological parameters for scientific application [8]. It helps in measuring the level of air pollution and assesses its impact on living beings [9]. Considering the strong association of emission sources with air pollutants as well as with regional and meteorological parameters, the role of such models is indispensable [10]. Because not just they help determining the actual emission sources, but future mitigation solutions is the other major contribution of such models [11].

There are two types of approaches which regulate atmospheric concentration of air pollution: a) Chemical Transport Models (CTMs) and b) data driven approaches. CTMs generally deal with the emission process, mixing and transportation of atmospheric gases with respect to the regional weather parameters [12]. These models are based on multiprocessing techniques which use real time and updated emission and meteorological records. However, the implementation of such models at times is held by the lack of primary emission and meteorological data in areas with initial boundary conditions [13]. No doubt the accuracy achieved by regression based models is reasonable, however, several studies revealed that non-linear behaviour of air pollutants and other influential regional features leads to a complex system of pollution especially in regions with complex terrain. And traditional deterministic models find it difficult to capture this non-linear complex system of air pollution [14]. Therefore, to deal with these problems, data driven approaches based on machine learning such as Artificial Neural Networks (ANN), and SVM seem promising for their ability to efficiently overcome the issue of non-linearity. These approaches are generally based on statistical techniques using historical data to make future predictions. These models are trained by using emission data, meteorological conditions, land use, anthropogenic activity etc. [15].

Literature review conducted in the context of this study suggest that recent researches in the field of environment science based on machine learning techniques such as ANNs, SVM etc. show a superior predictive performances over classic statistical models without knowing the chemical mixing, dispersion and transportation details of atmospheric gases [16–19]. It further revealed that, though traditional machine learning tools are able to handle non-linearity and complexity of emission datasets, however, ANNs based algorithms usually fall in traps of overfitting, local minima and best network architecture [20, 21]. Therefore, SVMs based algorithms can be used as alternative approaches, for their capability to deal with drawbacks of ANNs [22, 23]. Although these techniques are more efficient and reported to have promising performances in other research areas for prediction purposes, however, a lack of research using SVM based on different kernels especially Pearson VII Universal and normalized polynomial Kernel Functions in the field of atmospheric pollution modelling was identified.

Following the observations, the work comprehensively inspects about the performance of SVMs under different kernels against the leading regression approaches based on classic statistical and traditional machine learning algorithms.

1. RESEARCH METHODS
1.1. Study area

Kazakhstan is a growing economy. It generally relies on natural resource extraction such as oil and natural gas. It has a long history of environmental issues. After 28 years of independence, air pollution in most cities remains one of the major urban problem of the country. During recent years, air pollution has become a key focus

due to its serious health effects. Due to the direct involvement of NO_2 and SO_2 concentrations in Air Quality Index value both are deemed to be extremely dangerous for human health. Hence, the accurate prediction will help policy makers to take precautionary measures and formulate effective policies in time.

1.2. Data collection and pre-processing

The study makes use of historical atmospheric dataset gathered in Ust-Kamenogorsk — an administrative centre located in East Kazakhstan where several air quality and meteorological monitoring stations are installed. The dataset was provided by the Institute of Industrial Ecology Ural branch Russian Academy of Science. The datasets were recorded at 8 air monitoring sites located near Gastello Street (station 1), Delegatskaya Street (station 2), Kazakhstan Street (station 3), Auezov Avenue (station 4), Pogranichnaya street (station 5), Kuybysheva Street (station 6), Mendeleev Street (station 7), and Abay Avenue (station 8). These monitoring stations measure the concentration of various pollutants such as nitrogen dioxide (NO_2), carbon monoxide (CO), Sulphur dioxide (SO_2), hydrochloric acid (HCl), Formaldehyde (HCOH) and total hydrocarbon amount (CXHY). Since, wind speed and wind direction are reported to have a vital role in the transportation of pollutants, while regional temperature, precipitation and relative humidity aid the chemical mixing and dilution of air pollutants, therefore under this work meteorological records gathered at a meteorological station near Astana Agro-technika building were also taken into account. It captures wind speed (m/s), wind direction (degree), amount of precipitation (mm/h), ambient temperature (°C), relative humidity (%), and atmospheric pressure (mm-Hg).

A careful preliminary analysis unearthed several important features of the dataset e.g. the concentration values of CXHY were very low to be considered for pollution modelling, and atmospheric pressure values recorded during the study period were more or less static, therefore, for prediction modelling features considered include; (1) concentration of SO_2; (2) concentration of NO_2; (3) concentration of CO; (4) concentration of HCl; (5) concentration of HCOH; (6) ambient temperature; (7) wind speed; (8) wind direction; (9) amount of Precipitation; and (10) relative humidity, to characterize the hourly concentrations of air pollutants. Moreover it was observed that around 9 % meteorological values are missing, and a number of outliers were found in air quality dataset. Which necessitates the thorough cleaning of the data. Prior to in-depth cleaning, incomplete/missing values with discrepancies were carefully replaced, whereas the noisy data containing outliers were also removed. After data pre-processing a dataset containing 18000 instances were prepared for modelling. Lastly the dataset were split into two subsets containing 80 % – 20 % of instances for training and testing purpose respectively.

1.3. Experiment setting

Support vector machines are one of the best and widely known data mining tools for solving problems related to classification and regression. The technique works on a principle of optimizing hyperplane in a kernel to maximize the boundary between two classes. To achieve the maximum margin, hyper-parameters (or kernel parameters) are need to be selected carefully. These hyper-parameter the hyperplane/boundary line largely depend on the selection and value of support vectors.

In order to classify non-linear behaviour of a data point, SVMs generally adopts a kernel trick. Different kernels have different nonlinear mapping prospects which implies that SVM performances are often hindered by the right choice of the kernel. A number of such kernels are commonly adopted for regression purposes, however, this work only analyses the prediction performances of linear, polynomial, normalized polynomial, Gaussian radial basis function (RBF), and Pearson VII universal function kernel (PUFK) by using the equations (1)–(4) respectively. While the prediction performances of classifiers were evaluated on the basis of predicted and observed concentration value of pollutants obtained from test dataset by using evaluation measures such as Coefficient of Determination (R^2), Root Mean Squared Error (RMSE), Relative

Absolute Error (RAE) and Relative Root Squared Error (RRSE) by following equations (5)–(8);

$$K(a,b) = a \times b \qquad (1)$$

$$K(a,b) = (< a,b > +1)^d \qquad (2)$$

$$K(a,b) = \frac{< a,b >}{\sqrt{< a,a >< b,b >}} \qquad (3)$$

$$K(a,b) = e^{(-\gamma(a-)^2)} \qquad (4)$$

$$K(a,b) = \frac{H}{\left[1 + \left(\left(2(a-b)\sqrt{2^{(1/\omega)} - 1/\sigma}\right)^2\right)\right]^\omega} \qquad (5)$$

where $a, b \in R$, d is the degree of polynomial; γ is the peak height at the center, σ is the tailing factor of the peak and ω is the Pearson width in equations (2)–(5).

$$R^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2(y_i - \overline{y})^2}} \qquad (6)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2} \qquad (7)$$

$$RAE = \frac{\left|\sum_{i=1}^{n}(y_i - x_i)\right|}{\left|\sum_{i=1}^{n}(\overline{y} - x_i)\right|} \qquad (8)$$

$$RRSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - x_i)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}} \qquad (9)$$

where y_i and x_i are predicted and observed values respectively; $\overline{x}$ and $\overline{y}$ are the averages of targeted and predicted values respectively and n is the index number in equations (6)–(9).

For a fair prediction modelling, an appropriate adjustment of kernel parameters is a key to efficient performance of SVM kernels therefore, kernel parameters such a degree (d) for polynomial kernel, γ — width selection for RBF kernel and optimal selection of σ and ω for PUFK, all were fairly accustomed by using a technique called "grid search" during data training. The technique is known to determine the optimal values for all SVM parameters over a specified search range, hence, the study reports the best predic-

tion results achieved under different SVM based algorithms. All prediction models were trained and tested following a train/test ratio of 80 %–20 % respectively. While, to evaluate the proficiency of the adopted prediction models the experimental design calculates Correlation Coefficient (R^2) and error different error values like RMSE, RAE and RRSE.

Like SVM, study also employs Multilayer Perceptron (MLP) because, it is a supervised machine learning algorithm ranked among the list of top current age algorithms used for classification and regression purposes. The approach is robust having several eye-catching characteristics such as big datasets management and generalization abilities etc. Therefore, the study is aimed at a comprehensive investigation involving the development of several SVM based algorithm using different kernels.

The main contributions of the work include; (1) the application of SVM using PUFK function to predict atmospheric concentration of air pollutants in Ust-Kamenogorsk region; (2) the prediction performance comparison of SVMs with other well-known classic machine learning tools such as LR, MLP and SVMs.

2. RESULTS
2.1. Model Evaluation

Different support vector machine kernels have successfully applied in a number of fields like textile, electricity, bioinformatics, and atmospheric sciences etc. for various purposes such as regression analysis, time series prediction, condition monitoring, optimal control and fault diagnosis [23–27]. However, the review conducted suggest that the application of Pearson VII Universal function kernel and normalized polynomial kernel in environmental sciences have been limited especially for the prediction of air pollutant concentrations. Therefore, the work is aimed at assessing the prediction accuracies of SVM kernels, PUFK in particular, against other kernels such as RBF and polynomial and normalized polynomial kernels as well as the most commonly adopted approaches i.e. MLP and LR.

Besides choosing the right pollutants and environmental factors, to develop a robust predic-

Table 1. Predictive performances of MLP, LR and SVMs using various kernels

		SVM performance under different kernels					MLP	LR
		Linear	Polynomial	PUFK	RBF	Normalized polynomial		
Station 1 SO$_2$	R^2	0.8022	0.9143	0.9767	0.8703	0.9484	0.9557	0.7749
	RMSE	0.0164	0.011	0.0059	0.0141	0.0086	0.0094	0.0177
	RAE	44.9462	27.7634	17.7485	47.832	26.2675	30.468	48.751
	RRSE	60.9271	40.8141	21.87	52.5722	32.1015	31.6128	65.7607
NO$_2$	R^2	0.8993	0.8994	0.9544	0.9104	0.9496	0.9388	0.906
	RMSE	0.003	0.003	0.002	0.0028	0.0021	0.003	0.0029
	RAE	39.072	39.1148	23.4966	36.8143	26.3928	42.0521	40.7934
	RRSE	44.3209	41.5676	29.7657	41.4243	31.3837	39.9076	43.1971
Station 3 SO$_2$	R^2	0.8993	0.8994	0.9544	0.9104	0.9496	0.9388	0.906
	RMSE	0.003	0.003	0.002	0.0028	0.0021	0.003	0.0029
	RAE	39.072	39.1148	23.4966	36.8143	26.3928	42.0521	40.7934
	RRSE	44.3209	41.5676	29.7657	41.4243	31.3837	39.9076	43.1971
NO$_2$	R^2	0.8717	0.8715	0.925	0.8628	0.9278	0.8646	0.889
	RMSE	0.0049	0.0049	0.0038	0.005	0.0037	0.005	0.0045
	RAE	43.7042	43.7753	35.1922	45.0235	34.0623	45.5328	44.738
	RRSE	49.3905	49.4538	37.8584	50.3213	37.1364	50.4308	45.5383
Station 6 SO$_2$	R^2	0.7407	0.6267	0.782	0.7567	0.782	0.7718	0.7591
	RMSE	0.0098	0.0127	0.0091	0.0096	0.0091	0.0096	0.0095
	RAE	52.1616	51.036	42.2638	50.5722	39.3236	53.3295	52.4193
	RRSE	67.527	87.1683	62.3441	65.9006	62.8419	66.1719	65.1206
NO$_2$	R^2	0.6471	0.7765	0.794	0.6564	0.7832	0.7156	0.7144
	RMSE	0.036	0.0304	0.0285	0.0366	0.0292	0.0378	0.0328
	RAE	56.8884	45.7197	42.9555	56.0728	46.2838	83.8154	67.0048
	RRSE	76.7554	64.8455	60.6465	78.018	62.1944	80.5678	70.0049

tion model, the work also considers the fine tuning of all SVM kernel parameters and a careful selection of network architecture for MLP prior to data modelling. For example, during RBF parameter optimization $\gamma = 0.01$ produced a fine accuracy, similarly, SVM with PUFK best performed with $\sigma = 1.2$ and $\omega = 2.01$ whereas polynomial and normalized polynomial kernels were assessed at $d = 2$.

Similarly, for MLP a series of trails were performed to determine the number of neurons in hidden layer, while the optimization of MLP was carried under 80 % training and 20 % testing data points. The initial MLP architecture uses 10 input variables, having 5 neurons in hidden layer, learning rate equal to 0.3, moment coefficient of 0.2 and the number of epochs =500 to predict the concentration of gaseous pollutants.

A. Masih, A. N. Medvedev

The accuracy performances of all classifiers to predict air pollutant concentrations are presented in table 1. It is a fact that the prediction performance of an algorithm is largely dependent upon predictive variables, nonetheless, results compiled in table 1 are evident that overall SVM with PUFK has clearly achieved better performances in terms of high correlation coefficient and low error values as compared to other widely adopted prediction models. It reflects the superiority of PUFK to grasp the complex relationship between air pollutants and meteorological records to make precise future predictions regarding air pollution. Normalized polynomial kernel with slightly low accuracy has registered the second best performance, while the performances of LR, SVM with linear and polynomial kernels were mediocre

Considering the practical application of PUFK in air pollution modelling, its prediction performance was tested using different test datasets acquired at 3 different monitoring locations to predict two different air pollutants namely SO_2 and NO_2. It is recognizable from the table 1 that SVM based on PUFK is the only modelling approach that is steady with top performances regardless the type of pollutant predicted or location of the test dataset used. While the performance of other algorithms to predict air pollutants remained wobbly with respect to the changing test dataset. The prediction model using normalized polynomial kernel and MLP have shown some notable performances to accurately predict

the concentration of both the air pollutants at a number of occasions for example (1) at station 3, the normalized polynomial kernel has achieved an accuracy of 96.6 % to predict SO_2 that is nearly equal to what PUFK has accomplished (97.05 %) at this station; (2) normalized polynomial kernel's accuracy to predict NO_2 concentrations (92.8 %) was slightly better than that of PUFK which yielded 92.5 % accuracy; (3) whereas at station 1 normalized polynomial kernel shared almost a similar performance like PUFK to measure the concentration value of NO_2 yet another time; (4) and lastly both MLP and normalized polynomial kernel have secured the second and third best spots by predicting atmospheric concentration values of SO_2 with a precision equal to 95.5 % and 94.5 % respectively. However, the prediction performance of SVM having PUFK kernel remains the best with highest correlation coefficient values that no other algorithm selected under this study could achieve at all 3 test data stations.

In order to clearly see the overall performance of each model the author draws an average of correlation coefficient value using results obtained from 3 test stations to predict the amount of SO_2 and NO_2 in air as presented in Fig 1. It shows that on average the prediction accuracy attained by employing PUFK kernel in support vector machines is the best to predict SO_2 concentrations with $R^2 = 0.91$ followed by normalized polynomial kernel with an average R^2 value equal to 0.90. On the other hand, to predict the exact con-
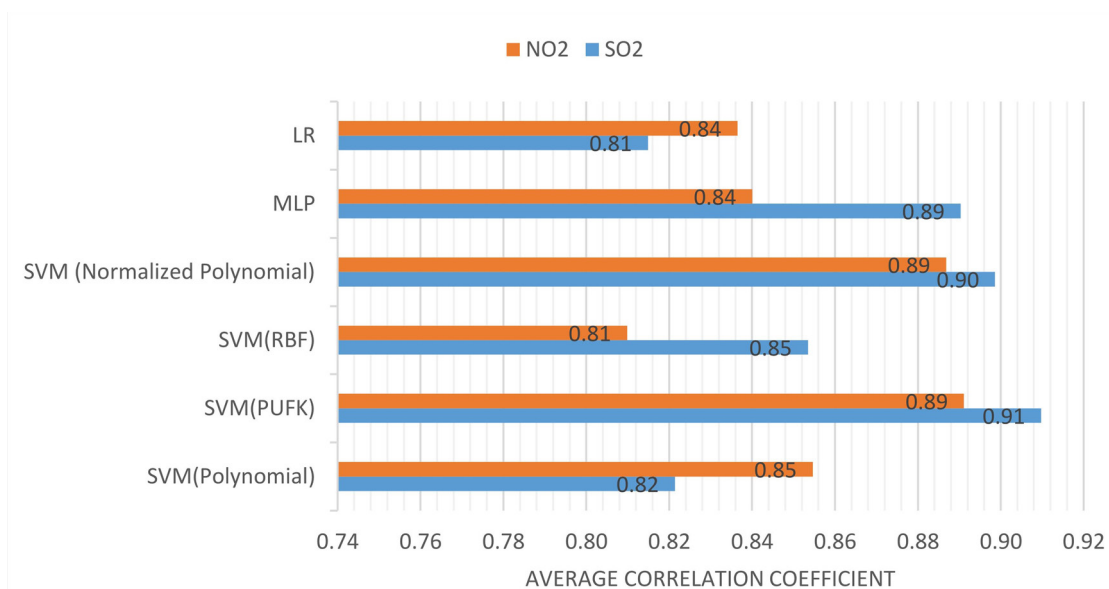

Fig. 1. Averaged R² value of 3 test data stations achieved by different prediction models
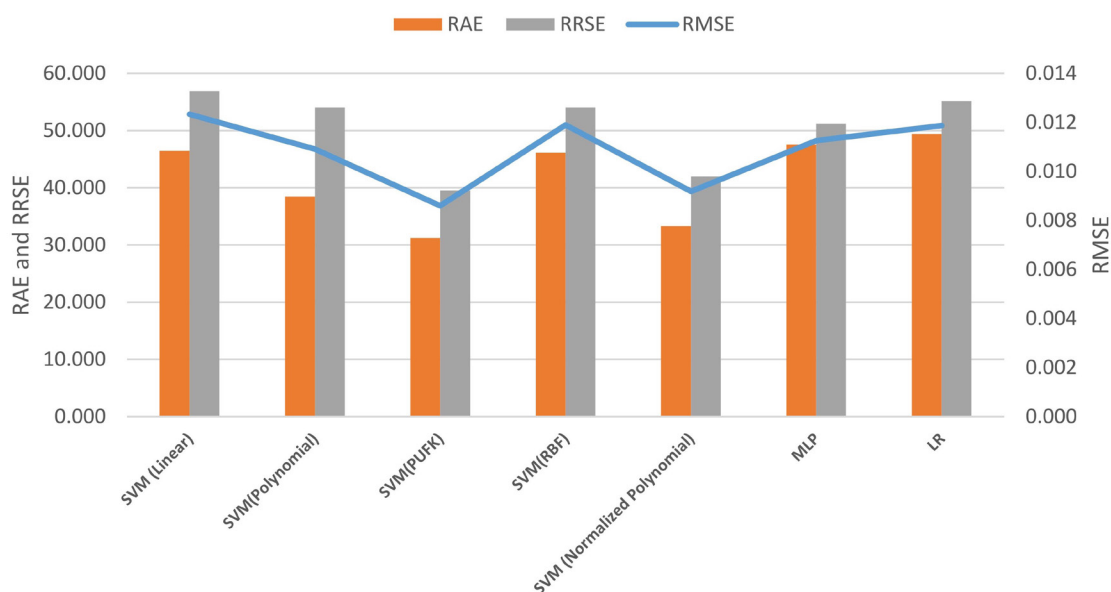
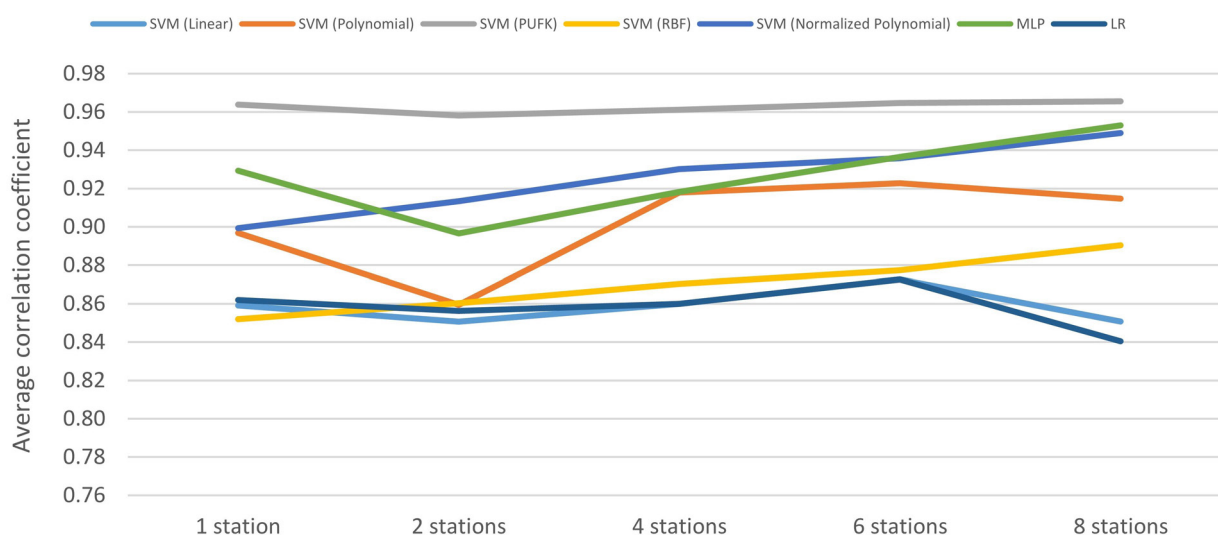*Fig. 2. Averaged RMSE and RAE achieved to predict SO$_2$ and NO$_2$*

Fig. 3. Prediction performance of different algorithms with respect to data station records

centration of NO$_2$ in air, PUFK and normalized polynomial kernels have shown an analogous efficiency with coefficient value 0.89 which is significantly better than that of other algorithms selected under this work. Although MLP have also shown a character to predict SO$_2$ with a reasonable coefficient of determination value 0.89 that is not far behind what PUFK and normalized polynomial kernel have attained, however it fails to continue its run to make an equally efficient prediction of NO$_2$ concentrations. Interestingly, to predict NO$_2$ concentrations, the average accuracies accomplished by polynomial kernel (0.85) is

better than LR (0.84) and is exactly equal to that of MLP accuracy i.e. 0.84.

In addition to accuracy measurements, algorithm performances with respect to three important error functions called RMSE, RAE, and RRSE were also gauged using equations (8), (9) and (10). The section analyses the average error values of RMSE, RAE and RRSE to predict pollutant concentrations of SO$_2$ and NO$_2$ drawn under different algorithms. The data presented in Fig. 2 indicate that, PUFK based prediction model uphold the top spot for having lowest error values (RMSE = 0.009, RAE = 31.1 %,

RRSE = 39.5 %). Normalized polynomial kernel shares an equal RMSE value i.e. 0.009 with PUFK, but due to its relatively high RAE and RRSE values of 33.3 % and 41.9 % respectively as compared to PUFK, the kernel earns the second position. No doubt the performance chart of MLP presented in Fig. 1 is reasonably good, but Fig. 2 reflects that its error values are listed among the bottom performers like LR, SVMs with linear, RBF and polynomial kernels.

2.2. Back propagation test

Since more attributes in a dataset make it more difficult for an algorithm to correctly learn the complicated relationship of input and output variables. Therefore, In view of back check prediction or self-consistency test the study divides the above main dataset of 17,216 instances (acquired at 8 pollution monitoring stations) into 5 sub-datasets for 5 different tests using data collected at monitoring stations 1, 2, 4, 6, and 8 respectively. In other words, besides embracing meteorological data, 1st test only uses the data gathered at station 1 i.e. 2152 to predict the amount pollutants in air, 2nd test assesses the prediction performance of algorithms based on 2 stations data (4304 instances). Similarly, next tests (i.e. 3rd, 4th and 5th) train pollution records obtained from 4, 6 and 8 stations having data instances equal to 8608, 12912 and 17216 instances respectively, to construct prediction models. The test was aimed at examining the consistency of MLP, LR, SVM based PUFK, RBF, normalized polynomial and polynomial kernels to predict the concentration of SO_2 and NO_2 present in air.

It is important to note that, the results presented in Fig. 3 display the average correlation coefficient value obtained to predict NO_2 and SO_2 concentrations acquired by different modelling schemes. It visibly demonstrates the pure dominance, character, and consistency of SVM algorithm attained by Pearson VII Universal Function Kernel to predict the atmospheric concentration of air pollutants namely SO_2 and NO_2 with high accuracy, low variance and low error values irrespective the size of the dataset. Under similar circumstances other algorithms have underperformed. The prediction performance of nor-

malized polynomial kernel is comparatively low during first experiment when 1 station's dataset were taken into account, however, the consistent upward trend, and smoothness of the curve is an answer to why the kernel should be rendered as the second best model in this specific experiment setting. Although overall SVM with RBF kernel hasn't done well nevertheless the sleekness of its curve confirms the stability of the kernel with the increased data instances. Fig. 3 also illustrates the inability of LR and linear kernel to capture the non-linear behavior between input and output predictors when size of the data is big. Admitting to the fact that the performance of MLP algorithm has been average, yet it has shown multiple times i.e. table 1, Fig. 1 and 3, that, the algorithm is capable of attaining a realistically high correlation coefficient value irrespective of the type of the modelling schemes adopted and varying size of the dataset. Contrary to that, it was quite strange to see the high enough error values produced by MLP algorithm (Fig. 2) due to which it could be categorized among the models with worst prediction accuracies such as linear regression, SVMs using linear, polynomial, and RBF kernels, hence confines its adoption for pragmatic purposes.

CONCLUSIONS

In this work 7 prediction models namely LR, MLP, SVMs using PUFK, RBF, linear, polynomial, and normalized polynomial kernels were developed. The study uses air pollutants data of NO_2, SO_2, CO, HCl, and CHOH gathered at 8 different monitoring stations and weather-related regional parameters such as ambient temperature, wind speed, wind direction, amount of precipitation, relative humidity, and atmospheric pressure to predict the atmospheric concentration of NO_2, and SO_2. The results obtained suggest that overall support vector machine using Pearson VII Universal Function Kernel has the characteristics of predicting atmospheric concentration of SO_2 and NO_2 with high accuracy, low error values and low overfitting probability. The work also affirms that, PUFK can outsmart both classic (LR) as well as the state of the art machine learning algorithms such as MLP and SVM having polynomial, normalized polynomial and RBF kernels adopted,

irrespective of data size or type of modelling technique applied. Normalized polynomial and MLP both have shown high accuracies under specific conditions but MLP doesn't qualify to the list of best predictors for overtraining and producing a significantly high error values. Other notable observations of the work include the winning performances of normalized polynomial kernel at occasions to predict the concentration of NO_2 and SO_2 with high accuracy, and finally the stability of RBF and PUFK kernels which remained fascinatingly steady and solid during consistency test.

CONFLICT OF INTEREST

The authors declare the absence of obvious and potential conflicts of interest related to the publication of this article.

REFERENCES

1. WHO. Air Pollution Levels Rising in Many of the World's Poorest Cities. 2016. Available at: https://www.who.int/news-room/detail/12-05-2016-air-pollution-levels-rising-in-many-of-the-world-s-poorest-cities

2. American Lung Association. "The state of the air 2019". 2019. Available at: https://www.lung.org/our-initiatives/healthy-air/sota/key-findings/.

3. *Xie Y., Zhao L., Xue J., Hu Q., Xu X., Wang H.* A cooperative reduction model for regional air pollution control in China that considers adverse health effects and pollutant reduction costs. Science of the Total Environment. 2016. 573. P. 458–469. Available at: doi: org/10.1016/j.scitotenv.2016.08.077

4. *Brunekreef B. and Holgate S. T.* Air pollution and health. The lancet, 2002. 360. P. 1233–1242. Available at: doi: org/10.1016/S0140-6736(02)11274-8.

5. *Gardner M. W. and Dorling S. R.* Neural network modelling and prediction of hourly NO_2 and NO_2 concentrations in urban air in London. Atmospheric Environment. 1999. 33. P. 709–719. Available at: doi: org/10.1016/s1352-2310(98)00230-1.

6. *Holloway J.* Legend-tripping in spooky spaces: ghost tourism and infrastructures of enchantment. Environment and Planning D: Society and Space, 2010. 28(4). P. 618–637. Available at: doi: org/10.1068/d9909.

7. *Bedoui S., Gomri S., Samet H. & Kachouri A. A.* Prediction distribution of atmospheric pollutants using support vector machines, discriminant analysis and mapping tools (Case study: Tunisia). Pollution, 2016. 2. P. 11–23.

8. *Daly A. and Paolo Z.* Air pollution modelling – An overview. Ambient air pollution, 2007. P. 15–28.

9. *Cohen A. J., Brauer M., Burnett R. et al.* Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study. The Lancet. 2017. 389. P. 1907–1918. Available at: doi: org/10.1016/s0140-6736(17)30505-6

10. *Lelieveld J., Evans J. S., Fnais M. et al.* The contribution of outdoor air pollution sources to premature mortality on a global scale. Nature. 2015. 525. P. 367. Available from: doi: org/10.1038/nature15371

11. *Pannullo F., Lee D., Neal L., Dalvi M., Agnew P., O'Connor F. M., Mukhopadhyay S., Sahu S., Sarran C.* Quantifying the impact of current and future concentrations of air pollutants on respiratory disease risk in England. Environ. Health, 2017. 16. P. 29. Available at: org/10.1186/s12940-017-0237-1

12. *Kinney P. L.* Climate change, air quality, and human health. American journal of preventive medicine. 2008. 35. P. 459–467. Available at: doi: org/10.1016/j.amepre.2008.08.025

13. *Jiménez P. A. and Dudhia J.* On the ability of the WRF model to reproduce the surface wind direction over complex terrain. J. Appl. Meteorol. Climatol. 2013. 52. P. 1610–1617. Available at: doi: org/10.1175/jamc-d-12-0266.1

14. *Ritter M., Müller M. D., Tsai M. Y., Parlow E.* Air pollution modelling over very complex terrain: An evaluation of WRF-Chem over Switzerland for two 1-year periods. Atmos. Res. 2013. 132. P. 209–222. Available at: doi: org/10.1016/j.atmosres.2013.05.021

15. *Masih A.* Machine learning algorithms in air quality modelling. Global Journal of Environmental Science and Management. 2019. 5(4). P. 515–534.

16. Rahimi A. Short-term prediction of NO_2 and NOx concentrations using multilayer perceptron neural network: a case study of Tabriz, Iran. Ecological Processes. 2017. 6(1). P. 4. Available at: doi: org/10.1186/s13717-016-0069-x

17. *Yu R., Yang Y., Yang L., Han G., Move O.* Raq--a random forest approach for predicting air quality in urban sensing systems. Sensors. 2016. 16. P. 86. Available at: doi: org/10.3390/s16010086

18. *Wang W., Men C., & Lu, W.* Online prediction model based on support vector machine. Neurocomputing. 2008. 71. P. 550-558. Available at: doi: org/10.1016/j.neucom.2007.07.020

19. *Grivas G., & Chaloulakou A.* Artificial neural network models for prediction of PM10 hourly concentrations, in the Greater Area of Athens, Greece. Atmospheric environment. 2006. 40. P. 1216–1229. Available at: doi: org/10.1016/j.atmosenv.2005.10.036

20. *Masih A.* Application of ensemble learning techniques to model the atmospheric concentration of SO_2. Global Journal of Environmental Science and Management. 2019. 5(3). P. 309–318.

21. *Jiang N., & Riley M. L.* Exploring the utility of the random forest method for forecasting ozone pollution in SYDNEY. Journal of Environment Protection and Sustainable Development. 2015. V. 1. P. 245–254.

22. *Elangasinghe M. A., Singhal N., Dirks K. N., Salmond J. A., Samarasinghe S.* Complex time series analysis of PM_{10} and $PM_{2.5}$ for a coastal site using artificial neural network modeling and k-means clustering. Atmospheric Environment. 2014. 94. P. 106–116. Available at: doi: org/10.1016/j.atmosenv.2014.04.051

23. *Baawain M. S. & Al-Serihi A. S.* Systematic approach for the prediction of ground-level air pollution (around an industrial port) using an artificial neural network. Aerosol and air quality research. 2014. 14. P. 124–134. Available at: doi: org/10.4209/aaqr.2013.06.0191

24. *Guo H., Zhao J., Yin J., Yao L.* Structural testing of polyimide nanocomposite films with SAXS and SVM-PUK. Polymer Testing. 2018. 70. P. 30–38. Available at: doi: org/10.1016/j.polymertesting.2018.06.025

25. *Liu Zhiliang and Hongbing Xu* Kernel parameter selection for support vector machine classification. Journal of Algorithms & Computational Technology. 2014. 8.2. P. 163–177. Available at: doi: org/10.1260/1748-3018.8.2.163

26. *Abakar Khalid A. A. and Chongwen Yu.* Performance of SVM based on PUK kernel in comparison to SVM based on RBF kernel in prediction of yarn tenacity. Indian Journal of Fiber and Textile Research. 2014. 39. P. 55–59.

27. *Zhang G. and Huihua G.* Support vector machine with a Pearson VII function kernel for discriminating halophilic and non-halophilic proteins. Computational biology and chemistry. 2013. 46. P. 16–22. Available at: doi: org/10.1016/j.compbiolchem.2013.05.001

28. *Juhos I., Makra, L. and Tóth B.* Forecasting of traffic origin NO and NO_2 concentrations by Support Vector Machines and neural networks using Principal Component Analysis. Simulation Modeling Practice and Theory. 2008. 16. P. 1488–1502. Available at: doi: org/10.1016/j.simpat.2008.08.006

Masih Adven — postgraduate student, Ural Federal University named after the first President of Russia B. N. Yeltsin.
E-mail: adven.masikh@urfu.ru
ORCID iD: https://orcid.org/0000-0001-6124-9422

Medvedev Alexander N. — PhD in Technical Sciences, Assistant Professor, Department of Big Data Analysis and Methods of Video Analysis, Ural Federal University named after the first President of Russia B. N. Yeltsin.
E-mail: a.n.medvedev@urfu.ru
ORCID iD: https://orcid.org/0000-0002-5160-2816