PM10 FORECASTING THROUGH APPLYING CONVOLUTION NEURAL NETWORK TECHNIQUES

PIOTR A. KOWALSKI[1,2], KASPER SAPAŁA[3] & WIKTOR WARCHAŁOWSKI[3]
[1]Faculty of Physics and Applied Computer Science, AGH University of Science and Technology, Poland
[2]Systems Research Institute, Polish Academy of Sciences, Poland
[3]Airly sp. z o.o, Poland

ABSTRACT

The World Health Organization (WHO) estimates that air pollution kills around 6.5 million people around the world every year. The European Environment Agency, in turn, points out that about 50,000 people die annually in Poland due to this. PM10 pollution arises in the form of smog (smoke and fog) and is an unnatural phenomenon created by adverse weather conditions and human activity. The aim of this article is to assess the possibilities of tasking modern neural networks to predict PM10 air pollution levels in the following hours of the subsequent day. In evaluating the prediction task, several types of error are considered, and machine learning algorithms and structures are utilized as learning models. Of note, the algorithm selected for stochastic optimization is a form of convolutional neural networking and deep learning neural networking that is used in machine learning when considering Big Data issues. The obtained results were then analysed and compared with other methods of prediction. As a result of this research, the proposed convergent neural network could be used effectively as a tool for calculating detailed air quality forecasts for the subsequent 24-h period.

Keywords: air pollution prediction (forecasting), big data, convolutional neural networks, machine learning, regression task, neural network, particulate matters.

1 INTRODUCTION

PM10 is a mixture of suspended particles, the diameters of which do not exceed 10 μg. This dust is harmful due to the content of such elements as benzopyrenes, furans, dioxins and carcinogenic heavy metals. The norm of average daily concentration of this dust is 50 μg/m^3, and the annual safe set limit is 20 μg/m^3, according to the WHO organization. It is worth noting, however, that information about exceeded norms is announced when the daily PM10 concentration is 200 μg/m^3 – one can see how often we think that the air is fine, while it is very polluted, but the alarm level has not yet been reached.

Human health depends, among other things, on the purity of the air breathed in. Unfortunately, Poland is not a country that could boast of excellent air quality. This is evidenced by the WHO report, which shows that among 50 cities and towns of the EU with the highest level of pollution, the report lists as many as 33 from Poland.

Measurement, modelling and prediction of concentrations of air suspended particulates are very important challenges, especially in areas where exceedance of standards is very common. Therefore, as part of this study, it was decided to develop a machine-learning model based on deep learning of neural structures, which allows prediction of PM10 dust concentration covering the approaching 24 h – hour by hour.

Deep learning is a specific subfield of machine learning. It uses a new approach to learning from data, and places emphasis on teaching subsequent layers of neurons of more and more significant representations. The term ‘deep’ refers to the many layers between the input data and the obtained result. This allows the algorithm to use numerous processing possibilities – both linear and non-linear. The convolutional neural network (CNN) employed in the presented research is a form of multilayer neural networking; its main purpose is to recognize

ISSN: 2398-2640 (paper format), ISSN: 2398-2659 (online), http://www.witpress.com/journals
DOI: 10.2495/EI-V3-N1-31-43

different patterns, especially on images. This network consists of an input layer, an output layer and many hidden layers. In the hidden layers, convolutional and pooling operations are performed.

In order to overcome the PM10 forecasting problem, the present paper will present models of PM10 air pollution prediction that have been based on artificial neural network procedures, in particular, on the most modern technique – ‘convolutional neural networking’. The evaluation of learning process for the investigated models was mostly based upon the mean square error criterion, however, during the model validation, a number of other methods of quantitative evaluation were taken into account. The presented model of pollution prediction has been verified by way of real weather and air pollution data taken from the Airly sensor station network (Fig. 1).

The distributed network of Airly measurement devices enables access to current and archival data on air pollution, temperature, suspended particulate matter PM1.0, PM2.5 and PM10, CAQI levels, atmospheric pressure and air humidity (Fig. 2). Currently (May 2019), the dense network of sensors is composed of 2,458 stations. This degree of coverage puts Airly into the leading reportage position in Poland and makes it one of the dominant providers in the world. Outside Poland, the sensor network includes several test locations in Europe, such as that of Spitzbergen. Indeed, the number of measuring sensors located in Europe is

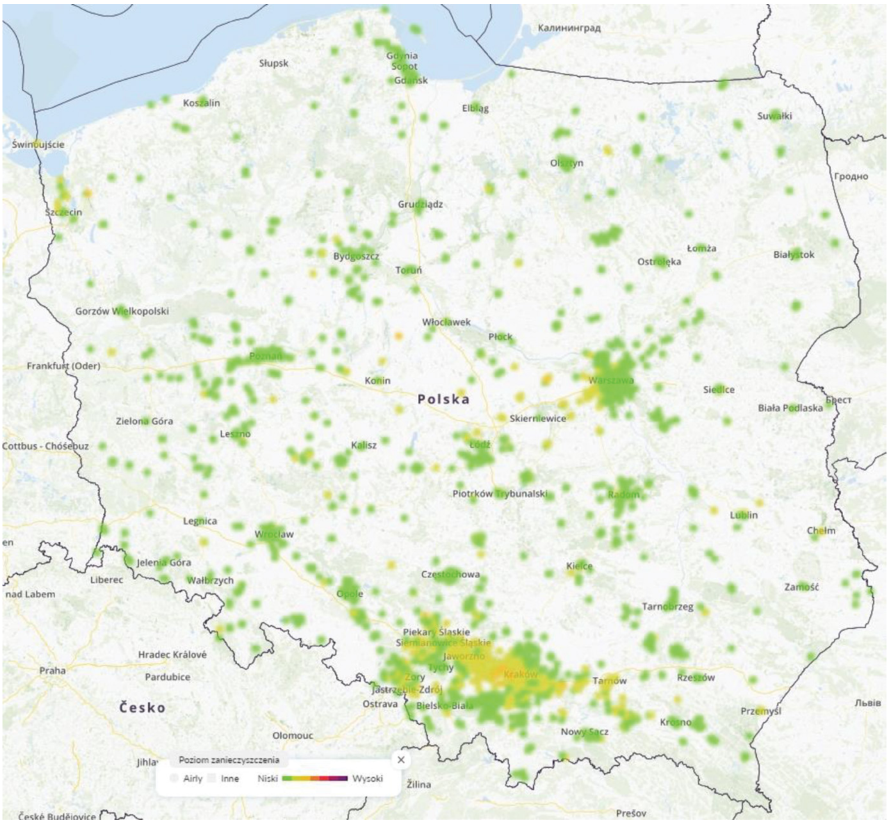
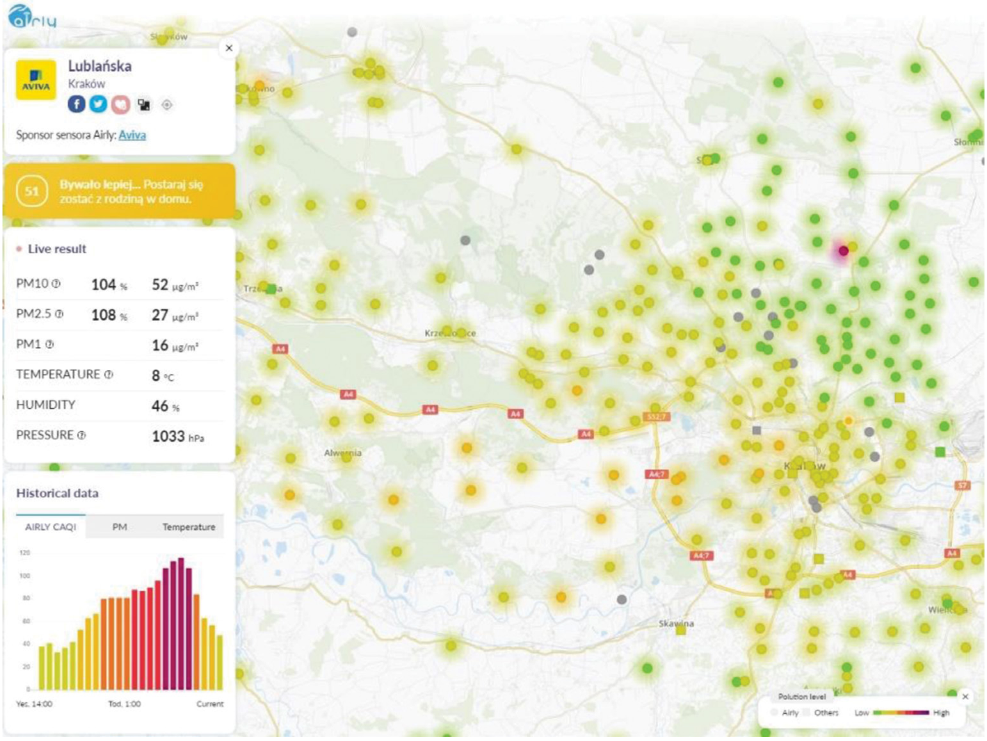


Figure 1: The Airly sensor network in Poland.



Lublańska
Kraków

Sponsor sensora Airly: Aviva

51 Bywało lepiej... Postaraj się zostać z rodziną w domu.

Live result

PM10	104 %	52 µg/m³
PM2.5	108 %	27 µg/m³
PM1		16 µg/m³
TEMPERATURE		8 °C
HUMIDITY		46 %
PRESSURE		1033 hPa

Historical data

Figure 2: The sensor locations near Krakow and the displayed left information panel seen on their website.

growing dynamically (outside of Poland), e.g. in Austria, Germany, Great Britain, Greece, Ireland, Italy, North Macedonia, Romania and Spain.

The problem of modelling and predicting air pollution levels due to PMx particles is a very important but difficult task. The impressive number of scientific articles that deal with this issue evidences this. What is more, this issue is completely different for various geographic and environmental conditions, even though it would seem that it can be easily determined. When reviewing literature related to this problem, it is possible to distinguish several groups of machine learning algorithms that allow for its solution, e.g. Linear Regression [1], Fuzzy Logic [2], Neural Network Procedures [3], Neural Networks with ARMA models [4], Deep Neural Networks [5,6] and Autoencoders [7].

Interestingly, CNN type neural networks [8] are more often applied in information processing. This type of neural network was predefined to issues related to image analysis [9]-[11] and video sequences [12], and classical algorithms using shallow type of neural networks were employed to create solutions to prediction issues [4]. Recently, deep neural network application, particularly CNN, has expanded into the domain of analysis and modelling of time series [13]-[15].

The paper is organized as follows. In Section 2, convolutional neural network models are introduced and investigated. In Section 3, the real data set used for the presented numerical experiments is set out. Moreover, procedures for the learning neural model Adaptive Moment Estimation algorithm [16] are advanced. In Section 4, results of numerical verification are presented and discussed. Finally, Section 5 contains concluding remarks and plans for future investigations.

This work is a continuation of the research presented in [1].

2 INVESTIGATED MODELS

Deep learning is a subcategory of machine learning, the field of computer programming science that enables a device to learn from data. Deep learning describes a set of techniques for teaching the brain of the device. These are models of neural networks (deep networks) built of multiple layers that are inspired by the activities performed by the human brain.

Deep networks are based on a hierarchical model (Fig. 3); the output of the lower layer is the entry of the next one. The lower layers represent the simplest characteristics of the input signal, the higher layers generate more general data features based on the previous layers. The deeper layers extract the basic information, while the higher correspond to the application of this information to the modelled problem. The main feature of deep networks that distinguishes these from typical neural networks is the number of layers. A deep network can have dozens or even hundreds of hidden layers, while a typical network usually has up to three hidden layers. Skilful application of deep learning in data processing has revolutionized this dynamically developing machine learning field.

The advantage is the use of models of CNN convolution. The multiplication of the weight matrix and the input signal of traditional ANN are replaced by the convolution – the console layer. The application of convolution significantly limits the number of parameters by sharing weights and using filters. In regular networks using fully connected layers, it is necessary to assign at least several weights to each input element. Unfortunately, this solution does not scale well for large data sets; the number of necessary neurons explodes when increasing the size of the considered data. Using of CNN increases the size of the input without the need to expand the architecture. Thanks to the significant reduction in the number of parameters compared to previous solutions, it is possible to run such a network on any device; hence, the networks are universal and have a large spectrum of applications.

3 LEARNING PROCEDURES AND DATA SET

The proposed prediction algorithm, based on CNN, is learned using the adaptive moment estimation algorithm (ADAM). It is a supervised learning algorithm, i.e. a data set consisting of pairs in which an input pattern and a correct model response are to be provided. The

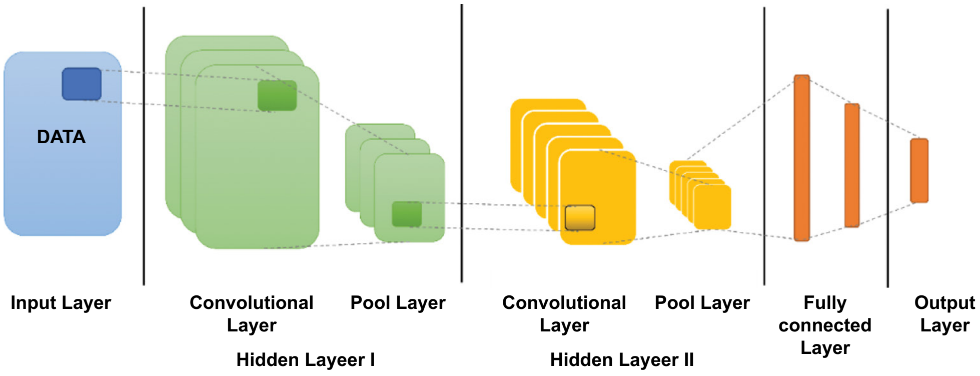


Figure 3: Hierarchical structure of convolutional neural network with two hidden and two fully connected layers.

algorithm learns to match the output to the actual value based on the training data. This is a method that allows modification of weights in a network with a layered architecture, in all its layers. Data is usually divided into three sets: learning, testing and validation. The largest, the learning set, is used in the network teaching process. A test set is then employed during learning to assess the quality of the classification. Finally, a validation set is applied, albeit less frequently than the test set, to determine the quality of the model under real conditions. The undoubted advantage of ADAM is the estimate of the first and second moment of gradients, i.e. the mean and variance, respectively [16].

In our study, all time series simulations were used, including hourly average pollution measurements, as well as weather condition, both of which were obtained from Airly sensors. The measurements on which the numerical experiments are based were for a period of about 1 year. Each measurement (at time *t*) includes the dust concentration PM10 [μg/m^3], temperature [°C] and average wind force [m/s], which were marked as x_{10}^{t} , x_{temp}^{t} and x_{w}^{t}, respectively.

The data set subject to numerical calculations contains 5,969 hourly measurements. Due to the specificity of the process of preparing the model in the Machine Learning domain, these data were divided into two separate sets. The first is a training set and is used to build a model that includes the refinement of its particular parameters. The second is the test set, the task of which is to evaluate the quality of the finished model. In the present case, the data is divided into 5,242 elements that are designated towards the training data (x_{learn}), with the remaining 727 being the test data (x_{test}). In addition, it should be emphasized that the data is discontinuous. This is a consequence of the use of data from real measuring sensors, and these discontinuities result from minor faults, service breaks or breakdowns of external computer networks transmitting data to the servers aggregating them.

In the present research, the historical data described above, as well as the prediction data associated with the meteorological measurements used were employed as inputs to the prediction model. The second group included the temperature and wind strength received from the DarkSky portal. Thus, the input vector to the predictor was formulated in the form:

$$x = \left\{ x_{\mathrm{PM10}}^{-25}, \ldots, x_{\mathrm{PM10}}^{-1}, x_{temp}^{-25}, \ldots x_{temp}^{24}, x_{w}^{-25}, \ldots x_{w}^{24} \right\}. \tag{1}$$

The output from the prediction task is a PM10 concentration forecast for the next 24 h. This is presented in the form:

$$y = \left\{ y_{\mathrm{PM10}}^{1}, \ldots, y_{\mathrm{PM10}}^{24} \right\}. \tag{2}$$

Network overtraining is a situation when the network learns to classify the examples perfectly from the training set, but the network loses the ability to generalize, i.e. the network's ability to correctly classify the data on which the network was not trained upon. Overtraining may come about due to the use of too many parameters in comparison to the complexity of the problem. The overtrained model can also classify on the basis of insignificant details within the data, e.g. the background image scheme. A drop in the accuracy of tests upon the validation set may indicate overtraining as learning progresses, or more directly by generating lower accuracy values for the test set relative to the training set. In order to eliminate this unfavourable phenomenon, many solutions can be applied, e.g. momentum, dropout, regularization L1 and L2, batch normalization, gradient clipping, data augmentation. All the aforementioned are dedicated to neural algorithms.

In this work, standardization of input data and procedure dropout were applied. This algorithm relies on a technique where randomly selected neurons are ignored during training. This task is a very important element that allows dispersing the tasks performed by dropped-out neurons on neighbouring cells. In this manner, the neuron weights that are the carrier of knowledge are tuned to specific features related to the issues studied. What is more, this method ensures that neurons do not specialize in the recognition of learning data, that is, they do not undergo a process of overfitting.

In the learning process, the basic concept is the measure of the quality of the solution. This is the foundation for calculating the gradient that affects the direction of the change of weights in subsequent iterations of the training procedure. In the present case, due to the continuous nature of the exit, the mean square error is applied. This can be expressed as:

$$MSE = \frac{1}{P}\sum_{i=1}^{P}\left(\widehat{y_{\mathrm{i}}^{\tau}} - y_{\mathrm{i}}^{\tau}\right)^{2}, \tag{3}$$

where $\widehat{y_l^{\tau}}$ is n predictions of investigated forecast (for τ time), and y_i^{τ} consists of P observed values of the predicted variable. However, for very fast variable data such as hourly concentrations of PM10, a much better formula for analytical purposes is:

$$R^2 = \frac{Cov\left(y^{\tau}, \widehat{y^{\tau}}\right)}{std\left(y^{\tau}\right)\, std\left(\widehat{y^{\tau}}\right)}, \tag{4}$$

where $std\left(y^{\tau}\right)$ and $std\left(\widehat{y^{\tau}}\right)$ are the standard deviations of observed value and the investigated forecast, respectively, and $Cov\left(y^{\tau}, \widehat{y^{\tau}}\right)$ denotes the covariance coefficient.

Herein, simultaneously, both types of measures have the property of forecast compliance. This can be defined as the enabling aggregation and holding assurance of minimization of error at each level of the hierarchy. The above corresponds directly to the selection of the forecast assessment measures, as obtaining the conditional value expected for the forecasting result is possible only by minimizing square errors, such as, among others, adopted MSE and R^2.

4 EMPIRICAL STUDY

As a part of the research on the prediction of air pollution in the form of PM10 dust, several hundred CNN were used. The obtained results were dependent on many factors related to this type of neural network. The research focused on the topological structure itself, such as choosing the learning algorithm, the length of the training process, the value of the learning rate parameter and the dropout procedure. This section shows the results from the best network that was obtained. All graphs below were generated only for test data.

For research purposes, several tens convolution neural networks were synthesized. For each of these, the possibilities of prediction for each following hour from the first to the 24th of the following day were separately examined. Figure 4 shows the correlation graphs for the PM10 concentration forecast for the first, second, sixth, twelfth and 24-h intervals, respectively. The horizontal axis shows the forecast value, while the real value is on the vertical axis. In the ideal case, i.e. for full compliance of the forecast with real data, all points would be on the green line. Therefore, it can easily be concluded that in the presented data, the envelope width indicates a prediction error.

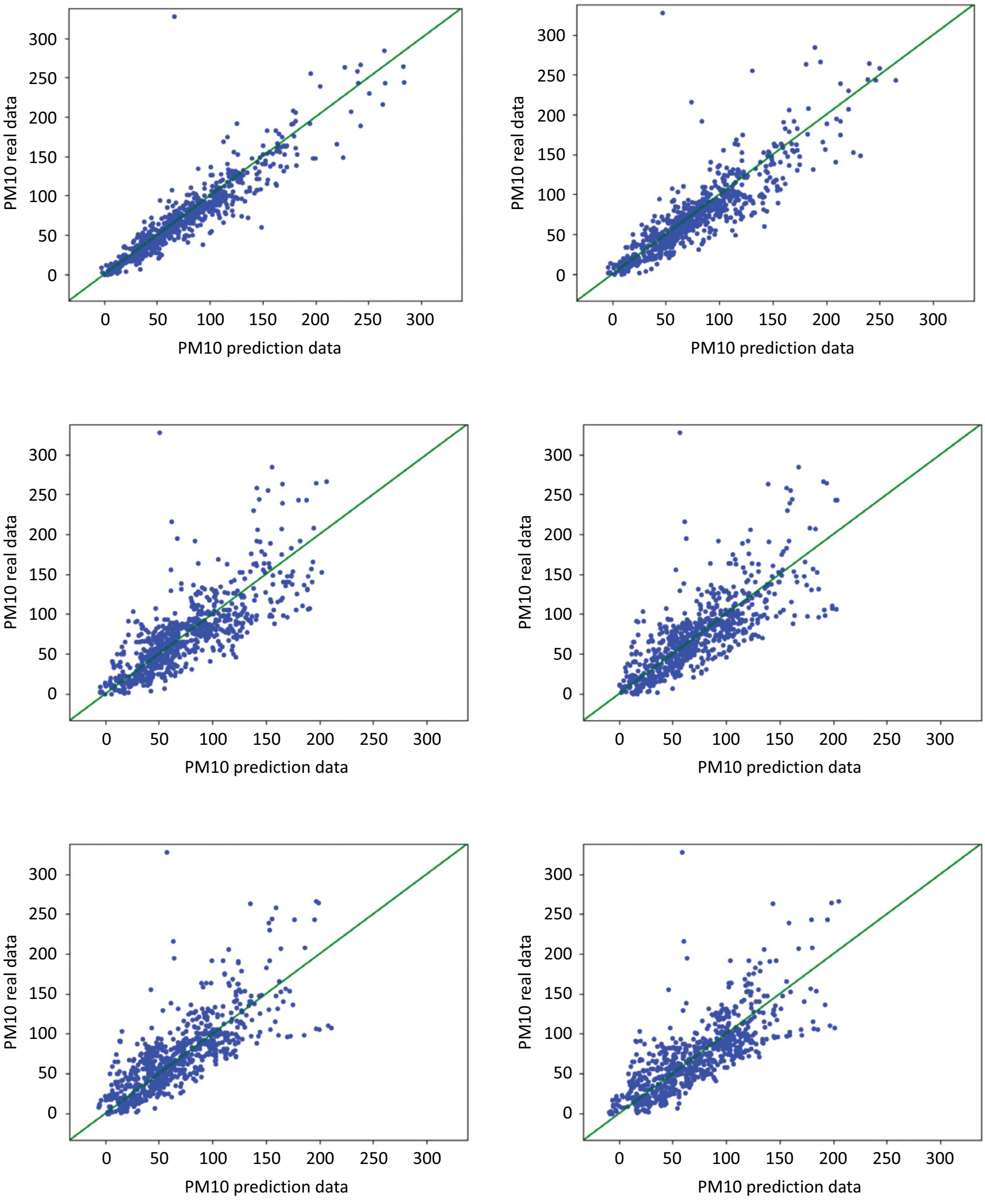


Figure 4: Results of correlation based on the testing set for predictor built upon CNN for the 1st, 2nd, 6th, 12th, 18th and 24th hour of forecast.

The individual charts show that the smallest error was achieved for the first hours of prediction, and this error increased in the subsequent intervals. Another observation is that the smallest error is for minute concentrations of PM10. For the range from 0 to 50, this error is practically negligible, while for larger values, prediction values significantly differ to the real. A further observation is the tendency to underestimate the forecast rather than to overestimate.

This effect is visible in the bottom row of results – through the occurrence of a specific hump in the points over the green line that indicates the ideal prediction. This hump introduces asymmetries in the envelopes of the individual charts. The outlier elements form another evident grouping and go beyond the envelope of consistency. Herein, the prediction value is completely different from the actual data. From the presented figures, it is safe to say that the number of these points is about 1 part-per-thousand of the data considered.

The above facts are also indicated in Fig. 5, which shows the temporal dependence for the 1st, 6th, 12th and 24th forecast hour, and which is where the compression within the linear model is dissipated. In this group of figures, 727 time measurements for particular prediction times are shown, with the red line being real data, the green being a prediction being based on a linear model and the blue prediction being based on the CNN. In the cases discussed, the first hour of prediction results coincide with real data. However, for the following hours, the prediction error begins increasing. Indeed, by the 6th prediction hour, a deviation between the data is evident. At the same time, the worst prediction barely copes with sudden jumps in PM10 concentration, as noticeable around the 300th hour of the test. Another example of poor coping with such abnormalities is a single peak around the 700th hour of the test.

Compared to the prediction algorithm based on the linear regression method, for the more distant predictions, the CNN-based procedure is much better at changing the temporal character of the time courses.

The relationship between the three models is demonstrated in Figs. 6 and 7. The first prediction – the CNN Model (red line) plots the results obtained in the previous part of paper. The next, the CNN Model with time (yellow line), adds time coordinates. The last line in green is a plot of the linear regression that the prediction is based on (1).

Figure 6 shows the change in the Pearson correlation coefficient depending on the time of prediction. Quite a clear conclusion can be derived from the decrease in this parameter as the PM10 prediction time increases. In this graph, it is noticeable that both CNN models are close to each other, especially at the beginning of prediction, i.e. up to 5 h; in subsequent hours, however, the distance in this relationship increases, as the CNN with time model is characterized by results that are inferior. The linear model behaves differently, it is much worse than the previous two (especially for the later hours of prediction). This is the most interesting and useful result of our study.

Figure 7 reveals the dependence of the MSE error on the subsequent prediction hours. In this figure, the colour convention is the same as in the previous case. For the criterion of the quality of prediction, which is the MSE error, it is clearly evident that the individual algorithms are at the same level only for the first 2 h. Afterwards, the CNN model is distinctly dominant while that of the CNN with time model is of less dominant and the Linear Model is even less so. At the beginning of the simulation, in the results for the Linear Model, the nature of the changes is similar to the CNN models, but with time, saturation comes about, albeit with less and less deceleration of the error. It is worth emphasizing that for the last hours of simulation, the models based on the CNN curve their MSE error characteristics, which means that for these cases, the forecast error slightly decreases. The reason for this may be lie in the cyclical nature of changes related to the periodicity of pollutants, i.e. people's habits, car traffic, factory work hours, etc. Interestingly, classical methods based on linear regression do not uncover such nuances in such a noteworthy manner.

5 CONCLUSIONS

This paper presents a study of predicting the level of air pollution concentration through the use of CNN. The purpose of the prediction was to indicate the condensation value of PM10 particulates for the next 24-h time interval. Many neural CNN have been synthesized for this

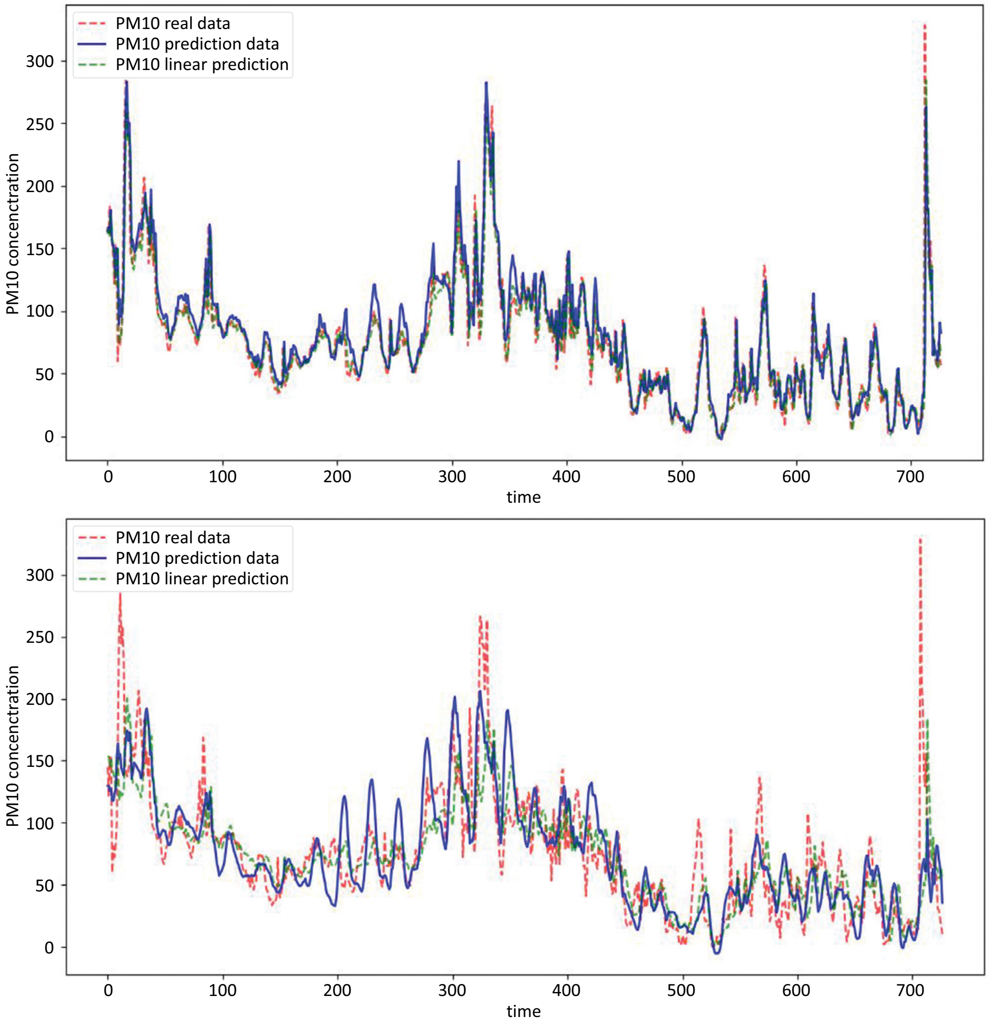


Figure 5: Testing set PM10 results in which the CNN predictor was applied for the 1st, 6th, 12th and 24th forecast hour, as compared with the linear model.

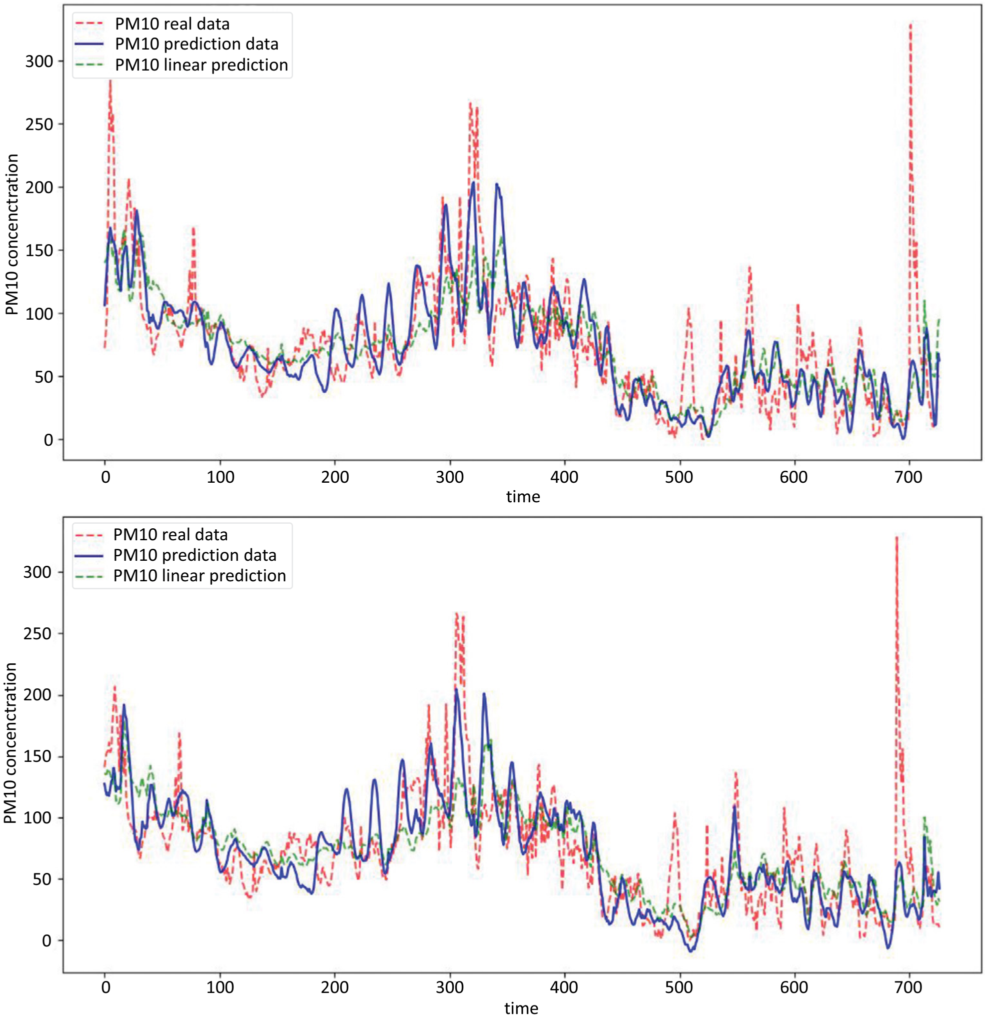


Figure 5: (*Continued*)

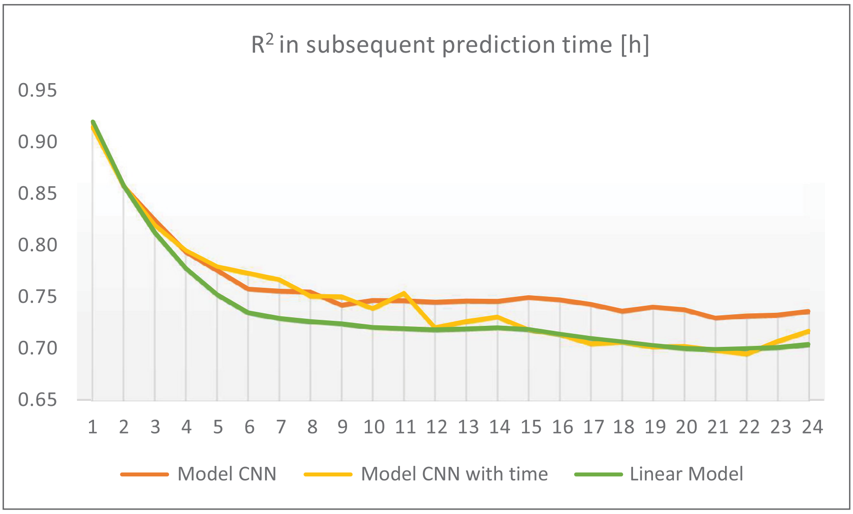


Figure 6: Comparison of PM10 results: the CNN model with and without time, as compared with the linear model, taking into account the R^2 measure.

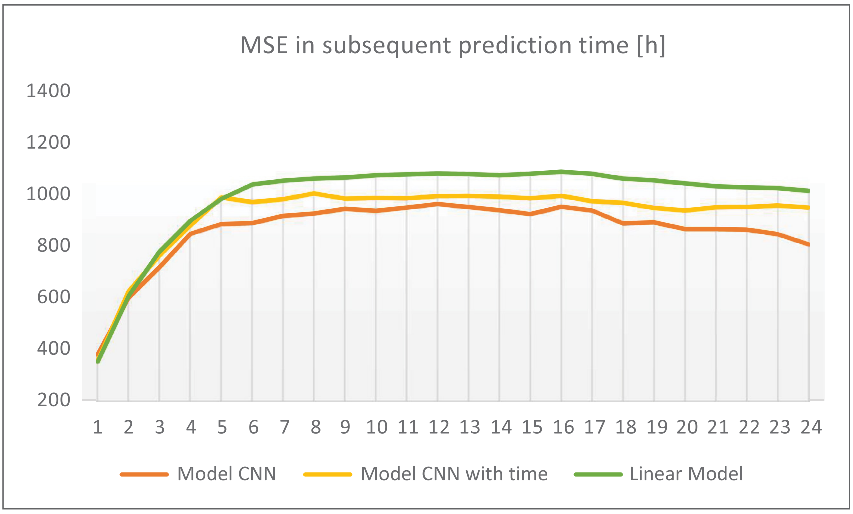


Figure 7: Comparison of PM10 results: the CNN model with and without time, as compared with the linear model, taking into account the MSE measure.

task, but due to the limited length of the articles, only the best solution was presented. For learning and testing the neural network, real data from the Airly sensor station located in one of the villages near Krakow were used. On the basis of the numerical analysis of the obtained results, our results indicate the positive qualities of utilizing neural network applications for the PM10 prediction task. In the comparative analysis carried out between the CNN models, CNN with time and the Linear Model, the CNN-based models showed superiority.

In subsequent studies, it will be prudent to scale the obtained results by applying several best neural networks to a larger scale of data, in particular to that derived from more measurement stations. In addition, the use of a heterogeneous model is proposed. This will include numerous neural networks of various types, from classical linear networks [1], through deep learning [6] to those of the probabilistic type [17,18], as the application of such a hybrid predictive model will ease the task of analysing the sensitivity of individual elements of the input vector [17,19].

ACKNOWLEDGEMENT

This work was supported by The National Centre for Research and Development (project no. POIR.01.01.01-00-0049/17 NCBiR).

REFERENCES

[1] Kowalski, P.A. & Warchałowski, W., The comparison of linear models for PM10 and PM2.5 forecasting, *WIT Transaction on Ecology and the Environment*, **230**, WIT Press: Southampton and Boston, pp. 177–188, 2018.

[2] Domańska, D. & Wojtylak, M., Application of fuzzy time series models for forecasting pollution concentrations. *Expert Systems with Applications*, **39(9)**, pp. 7673–7679, 2012.

[3] Chakraborty, K., Mehrotra, K., Mohan, C.K. & Ranka, S., Forecasting the behavior of multivariate time series using neural networks. *Neural Networks,* **5(6)**, pp. 961–970, 1992.

[4] Faruk, D.O., A hybrid neural network and arima model for water quality time series prediction. *Engineering Applications of Artificial Intelligence,* **23(4)**, pp. 586–594, 2010.

[5] Perez, P., Menares, C. & Ramirez, C., Forecasting in the Most Polluted City in South America, *WIT Transaction on Ecology and the Environment*, **230**, WIT Press: Southampton and Boston, pp. 199–204, 2018.

[6] Grover, A., Kapoor, A. & Horvitz, E., A deep hybrid model for weather forecasting. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* ACM, pp. 379–386, 2015.

[7] Xie, J., Wang, X., Liu, Y. & Bai, Y., Autoencoder-based deep belief regression network for air particulate matter concentration forecasting. *Journal of Intelligent & Fuzzy Systems*, **34(6)**, pp. 3475–3486, 2018.

[8] Schmidhuber, J., Deep learning in neural networks: an overview. *Neural Networks*, **61**, pp. 85–117, 2015.

[9] Ma, X., Dai, Z., He, Z., Ma, J., Wang, Y. & Wang, Y, Learning traffic as images: a deep convolutional neural network for large-scale transportation network speed prediction. *Sensors*, **17(4)**, pp. 818, 2017.

[10] Lawrence, S., Giles, C.L., Tsoi, A.C. & Back, A.D., Face recognition: a convolutional neural-network approach. *IEEE Transactions on Neural Networks*, **8**, pp. 98–113, 1997.

[11] Ji, S., Yang, M. & Yu, K., 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis Machine Intelligence*, **35**, pp. 221–231, 2013.

[12] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R. & Li, F.F., Large-scale video classification with convolutional neural networks. *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, pp. 1725–1732, 23–28 June 2014.

[13] Langkvist, M., *Modeling Time-Series with Deep Networks*, Örebro Studies in Technology, 2014.

[14] Liu, J.N., Hu, Y., He, Y., Chan, P.W. & Lai, L., Deep neural network modeling for big data weather forecasting. *Information Granularity, Big Data, and Computational Intelligence*, Springer, pp. 389–408, 2015.

[15] Liu, J.N., Hu, Y., You, J.J. & Chan, P.W., Deep neural network based feature representation for weather forecasting. *Proceedings on the International Conference on Artificial Intelligence (ICAI)*. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), p. 1, 2014.

[16] Kingma, D.P. & Ba, J.L., Adam: a method for stochastic optimization. *International Conference on Learning Representations*, pp. 1–13, 2015.

[17] Kowalski, P.A. & Kusy, M., Sensitivity analysis for probabilistic neural network structure reduction. *IEEE Transactions on Neural Networks and Learning Systems*, **19**(**5**), pp. 923–937, 2018.

[18] Kowalski, P.A. & Kulczycki, P., Interval probabilistic neural network. *Neural Computing and Applications*, **28**(**4**), pp. 817–834, 2017.

[19] Kowalski, P.A. & Kusy, M., Determining significance of input neurons for probabilistic neural network by sensitivity analysis procedure. *Computational Intelligence*, **34**(**3**), pp. 895–916, 2018.