CHEMICAL ENGINEERING TRANSACTIONS

VOL. 96, 2022

A publication of AIDIC

The Italian Association
of Chemical Engineering
Online at www.cetjournal.it

Guest Editors: David Bogle, Flavio Manenti, Piero Salatino
Copyright © 2022, AIDIC Servizi S.r.l.
ISBN 978-88-95608-95-2; **ISSN** 2283-9216

DOI: 10.3303/CET2296075

Innovative Process for the Geo-Linked Analysis of Air Quality Assisted by AI Techniques

Nicoletta Lotrecchiano[a,b], Diego Barletta[a], Massimo Poletto[a], Daniele Sofia[a,b*]

[a] DIIN-Dipartimento di Ingegneria Industriale, Universitá degli Studi di Salerno, Via Giovanni Paolo II 132, 84084, Fisciano (SA), Italy
[b] Sense Square srl, , Corso Garibaldi 33, 84123, Salerno (SA), 84123, Italy
dsofia@unisa.it

Image recognition is a technology that, through specific algorithms and methodologies, aims to reproduce the typical biological vision systems by identifying particular objects, patterns, colors or geometric shapes. Artificial intelligence was used in this study to define and test an algorithm for image recognition. In recent years it has been applied to the pollutants pattern recognition to evaluate some important phenomena and to catch details in the pollutants dispersion that with the standard measurement systems it is not possible. This study therefore aims to train an artificial intelligence in the automatic recognition of the air pollution level measured in the city of Milan obtained by a new generation sensor network. Air quality data came from the real-time on-road monitoring network operating in Milan. The developed algorithm has been applied to the PM10 concentration maps that divides the area in cells with an area of 1 km^2. The pollutant concentrations are reported on the map according a color scale starting from green to yellow and red according to the law limit value according D.Lgs. 155/2010. So, the color scale defines three levels – good, medium and bad- in which the dataset has been divided. Subsequently, the map where divided into four macro-areas (North West, North East, South West, South East) and the algorithm has been trained to identify the pollution patterns related to each of the four macro-areas. Sensitivity analysis were carried out on the model hyper-parameters to test the model and to increase its performance. In this work, the computer system designed to analyze, process and recognize the images provided by Google, Teachable Machine, was used, which through the use of artificial intelligence, allows the automatic images recognition.

1. Introduction

The human eye captures the signals provided by the surrounding environment and through the processing and analysis of the retina and the visual cortex, the sensory data are returned that allow man the sensation of visual perception and recognition through associations with memories. The methodologies for identifying the image contents are mainly two: the features recognition and the corners recognition (Troung et al., 2016). These are used by computer vision (CV) algorithms to recognize the contents by identifying the main image constituent. The first step is certainly to simplify the image, so as to focus on the subject of interest, leaving out the rest. Through the method used, characteristics are analyzed such as, for example, particular shades of color or lines that meet at an angle. Subsequently, a structural analysis of the image and a segmentation is carried out, in order to understand where the regions of interest are located in order to obtain information on the spatial arrangement of colors or intensities. Following the recognition of the features, the latter must be traced to others already catalogued, so as to create an association between features that allow the classification of the content (Mitchell et al., 1986). For this to happen, the machine set up for image recognition must be trained. Training takes place by showing the machine thousands of images also representing the subject of interest, and labeling them by category. In this way, when the machine is shown a new image with the same subject of interest, then it will return the label of the images that have instructed the machine with more characteristics coinciding with the one to be recognized. This type of training is called supervised machine learning as you already know the category to which the images instructing the machine belong. The means used by learning systems are neural networks. Through external data, they can change their structure by modifying the links between the nodes,

Paper Received: 5 December 2021; Revised: 14 July 2022; Accepted: 24 September 2022
Please cite this article as: Lotrecchiano N., Barletta D., Poletto M., Sofia D., 2022, Innovative Process for the Geo-linked Analysis of Air Quality Assisted by Ai Techniques, Chemical Engineering Transactions, 96, 445-450 DOI:10.3303/CET2296075

thus making the system adaptive (Watanabe, 1985). The learning and reasoning phase is therefore given by the connection of internal information and external data, through the neural network. The most used neural networks in image recognition are convolutional neural networks (CNN or ConvNets) and are precisely those based on the behavior described above. In the field of machine learning, a feature is an individual property and is measurable for an observed phenomenon (Bishop, 2006). Feature learning is defined as the set of techniques that allow a system to independently discover which features to extract from the input patterns and is divided into supervised learning, and unsupervised learning, semi-supervised learning and reinforcement learning. Supervised learning is a machine learning technique that aims to train a computer so that it can perform tasks automatically. In this type of learning, examples are given to the system as input and the correct results that the system should aim for. In order to implement a machine learning algorithm that is performing, a very large and varied training set is needed. In the case of the images relating to air quality, reference is made to the measurements obtained in space. Image analysis using artificial intelligence is widely used in various sectors. In the chemical industry it is used to detect the critical equipment failure (Zhang, 2018), to predict the chemical composition of products (Yao, 2018) or for the process optimization (Jia et al., 2015, Nuhu et al., 2021).

Air quality monitoring systems provide data with a certain spatial resolution connected with the devices technology so they can be viewed on a map and connected to geographic coordinates. Recent real-time on-road measurement systems (Lotrecchiano et al., 2019) allow the measurement of pollutants with a resolution of 1 km^2 with data guaranteed by the implemented blockchain (Sofia et al., 2020). This new technology avoid the problem linked to the installation points of the measuring devices in the monitoring network design (Sofia et al., 2019). These measured values can be extended into space through specially developed spatial interpolation algorithms which take into account not only the pollution levels but also the meteorological parameters that influence their dispersion (Lotrecchiano at al., 2021).

2. Materials and methods

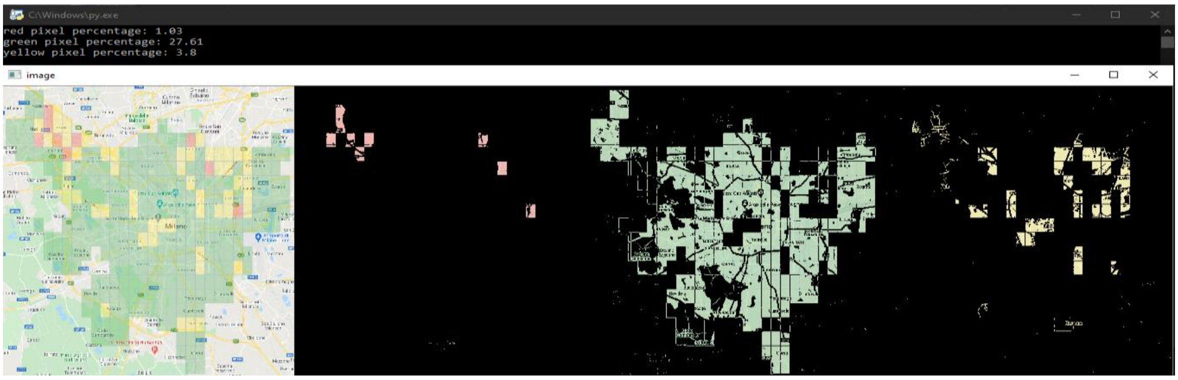
2.1 Air quality data

Air quality data used in this work, come from the real-time on-road monitoring network (ROM) located di Milan (Italy). The network is composed of 53 measuring devices located on couriers that can pass through the city many times per day. The devices can measure the main pollutants concentrations as PM10, PM2.5, PM1, NO_2 and the meteorological parameters as temperature, relative humidity and pressure. Data are stored in a database and available on a web application as images.

2.2 Data preparation and training set definition

Since the images are the object on which the algorithm is based, these must become the input data for the AI. A collection of images is periodically provided which show, each day, the level of pollution of the area considered. The images represent the Milan map divided into colored cells, where the colors correspond to the pollution level: green is good, yellow is medium, red is bad. Furthermore, the images are distinguished by pollutant and each pollutant has its own detection and representation; the study carried out relates to PM10 particulate matter. A color detection script in python environment was implemented to prepare the training set automatically and then classify the images of the urban context in the three categories: good average and bad. The script recognizes colors from images green, yellow, and red. The script analyzes the images and automatically moves them to the different folders according to the set criteria. The input dataset was divided into training, test and validation set in percentages of 60, 20 and 20. To arrive at the final result, various models were tested as the number of images increased. First of all, the input dataset of 426 images was divided into class using the color detection script, and subsequent creation of a model with three classes (good, medium, bad). Given the significantly larger dataset, the models implemented have had better performances, and seven models have been created, the most performing of which has an accuracy of 80%.

The classification was made also dividing the area in 4 macro-areas. The term macro-area refers to the area of the map identified in one of the four quadrants into which the image is divided. Two symmetry imaginary axes, orthogonal to each other and incident in the geometric center divide the image to obtain four equal areas that identify the North-East, North-West, South-East, South-West (NE, NW, SE, SW). Subsequently, it is possible to carry out a qualitative analysis of the images taking into account the prevailing pollution level (good, medium, bad) per quadrant defining 12 different classes.



```
red pixel percentage: 1.03
green pixel percentage: 27.61
yellow pixel percentage: 3.8
```

Figure 1: Image output after the color detection script.

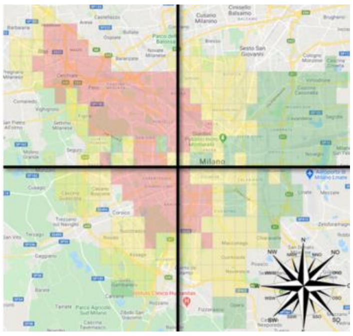


Figure 2: Macro-area identification.

2.3 Tool used

The tool used to define the classification algorithm was the Google Teachable Machine (TM), a web-based tool that allows the creation of machine learning models. The model created by TM is based on Tensorflow.js, an open source library for machine learning from Google. The algorithm inputs are the three classes (good, medium and bad) or the twelve classes (good, medium and bad for NE, NW, SE, SW) and the respective images already classified by the previously created script. A test phase will follow to find the best possible model by modifying the hyperparameters (Epochs, Batch Size, Learning Rate).

2.4 Hyperparameters tuning

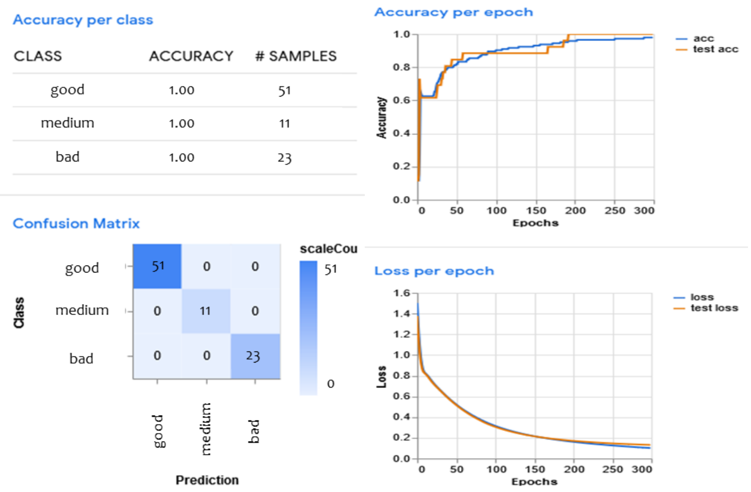
The choice of the hyperparameters and their optimization was a key process to arrive at a well performing definitive model. To obtain the final result, an empirical method by trial was followed, to learn the criticalities at each iteration and improve it in subsequent models. The theory associated with machine learning certainly provides a valuable tool for the effect hyperparameters have on model performance. For example, a smaller learning rate typically requires more epochs for training than a larger learning rate. In addition, small batch sizes are generally suitable for lower learning rates. In this project specifically, the learning rate was set at a predefined value, then, after training, by analyzing the graphs it was possible to understand what were the changes to be made to the model. For example, the final model at first was not performing at all, because the model overfitted. By experience it was concluded that, by lowering the learning rate, the problem of overfitting disappeared. For the optimization of the hyperparameters the metrics and graphs described before were used.

3. Results

3.1 Good, medium and bad classification

The development of an optimal forecasting model was possible by implementing a considerable number of models and analyzing their criticalities. The model is trained through TM followed by the parameter tuning phase. In the search for the most performing model, two other less performing models are also worthy of note. Among the models developed, the third one at the performance level has as hyper-parameters: epochs = 300, batch size = 128 and learning rate = 0.0001. The test accuracy is 90%, while the validation accuracy is 93%. Instead the test loss is 0.11 and the loss is 0.13. The second best performing model has as hyper-parameters: epochs

= 150, batch size = 16, learning rate = 0.0001. The test accuracy is 90% instead the validation accuracy is 97% while the test loss is 0.19 and the loss is 0.04. The most performing model, developed for PM10, has as hyperparameters: epochs = 300, batch size = 64 and learning rate = 0.0001. Looking at Figure 4 it is possible to see that the loss and the accuracy for epochs both converge. The test accuracy is 100%, instead the validation accuracy is 97%, the test loss is 0.13 and the loss is 0.10. In Figure 3 it is also shown that the accuracy per class and the confusion matrix for the test set is perfect achieving one of the best possible results.



Accuracy per class

CLASS	ACCURACY	# SAMPLES
good	1.00	51
medium	1.00	11
bad	1.00	23

Confusion Matrix

Class \ Prediction	good	medium	bad
good	51	0	0
medium	0	11	0
bad	0	0	23

Figure 3: Best model developed with epochs = 300, batch size = 64 and learning rate = 0.0001 for the test set.

3.2 Good, medium and bad for NE, NW, SE, SW classification

After the classification algorithm into three classes, the algorithm for the polluted area classification has been implemented. To have actual feedback following the training of artificial intelligence, it was necessary to derive from the images used as input in the tests, a graph that describes the percentage of the pollution level per quadrant. The graph representing the expected outputs was therefore obtained through a quantitative image. For each quadrant it was first counted the number of total coloured cells, then the number of cells per color, in such a way as to obtain, through simple proportions, the right percentage that quantifies the good, medium or bad level of the quality of the air (Figure 4).

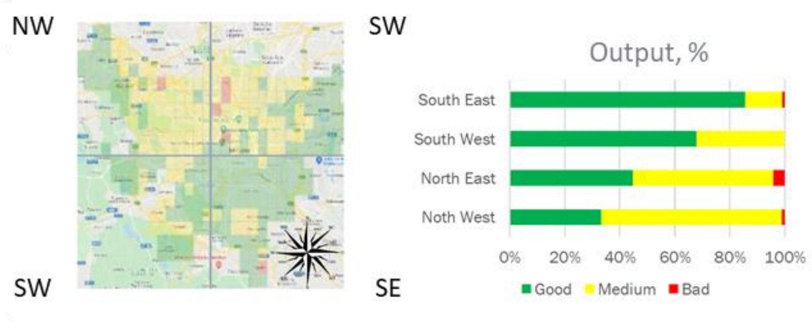


Figure 4: percentage that quantifies the good, medium or bad level of the quality of the air

Compared to the qualitative analysis of the input image, the results obtained from artificial intelligence training are very different and sometimes discordant. The output returned through the training presents a situation in the SE area almost similar to the expected results by means of the epoch parameter set to 75. The mismatch between the results is only for the percentage of the medium indicator, almost zero in the results obtained, higher for the desired results. The situation is similar for the value of epochs equal to 150 where, however, the percentage of the medium indicator is zero and that of good increases. For the other epoch values set, the results obtained show incompatible and misleading values. The behaviour in the SW area is described in a fairly precise manner by the tests carried out with the epochs parameters set to the values 50, 75, 150 and 25. All the results obtained match the zero value of the bad indicator but differ on the values for the classes good and medium, which are respectively greater than the latter in the expected results. More precisely for 25 and 50 epochs, the value of the medium class is greater than the value of the good class while for the other two epochs (75 and 100), the indicator for the medium class is zero for both. With regard to the NW area, the only acceptable results are those relating using 25 epochs as they reflect the values of the expected output. For the other epochs values, ambiguous and discordant results are obtained. Finally, for the NE area, the results that best match the expected results are obtained through with 50 and 75 epochs. However, they both have a zero value for the class bad. It can therefore be deduced from the results obtained that the value of the epochs parameter most suitable for instructing the TM is 75.

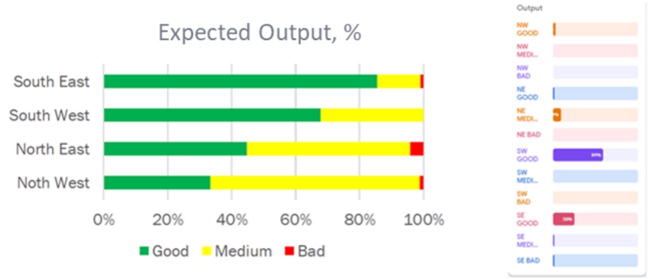


Figure 5: Output expected and resulting

The batch size parameter can be varied in six sizes: 16, 32, 64, 128, 256, 512. For the SE area the results that most correspond to the expected results are those returned by training with the batch size parameter set to 32. The results obtained with the parameter set to 64 also show a similar behavior, but do not consider the value of the bad indicator. The other parameter values return results that deviate from those expected. For the SW area, the batch size value that returns a reliable output is 64, even if it greatly lowers the value of the medium indicator compared to the expected values. The behavior of the NE area is instead better described through the results of the training with the batch size parameter set at 256 and 512, where, however, the difference between the values of the medium indicator, which is very evident, and the barely accentuated good, seems too clear. The only value that returns a behavior similar to that expected in the NW area is 128, in fact these results are the only ones to show the value of the medium indicator higher than that of good even if in a not very evident way. The variation by order of magnitude of the learning rate parameter value quickly excludes the values 1 and 0.1 which present degenerate cases. The significant cases for which the result can be analyzed are those in which the learning rate parameter varies between 0.01, 0.001 and 0.0001. In the SE area the values of the results obtained are close to those hoped for only for the learning rate value equal to 0.01 as it is the only one to show a higher percentage of the good index than the others, but to its detriment it leads the values of the medium and bad indices are zero and therefore do not coincide with the expected ones. For the SW area, the behavior of the results obtained is given for learning rate equal to 0.0001 and 0.01 that show a higher good value.

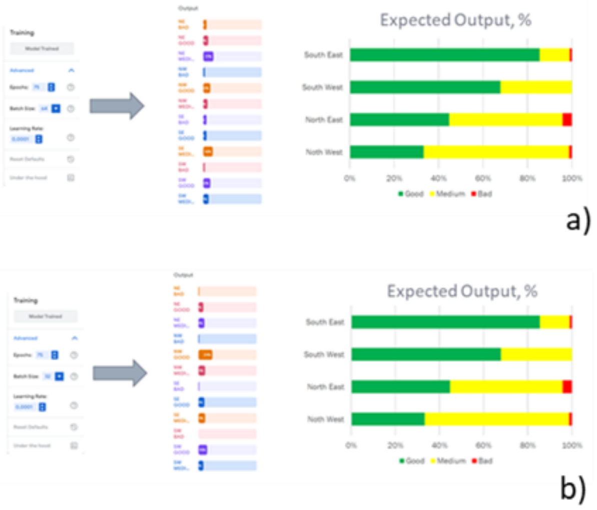


Figure 6: Output expected and resulting in a) test 1 and b) test 2.

The NE area is faithfully described by the results obtained through training with the value of the learning rate parameter equal to 0.0001 even if it greatly lowers the value of the good indicator. Finally, the NW area can also be faithfully represented by the results obtained with 0.0001 of learning rate. Following the analysis carried out, the actual sensitivity of the TM was verified by setting the parameters considered optimal. The results obtained with test 1 (Figure 6a) reflects the behavior of those expected only in the NE and SW zones. In the NW area the good value is greater than the medium value in the results obtained, the opposite is represented by the expected results. In the SE zone the good value is lower than the medium value in the results obtained, while in the hoped-for results we have a prevalence of the medium value over the good one. The results obtained in test 2 (Figure 6 b) corresponds to those expected only in the NE and SW zones. In the NW area the good value is greater than the medium value in the results obtained, the opposite is represented by the expected results. In the SE area the same values mentioned above are almost the same in the results obtained, while in the expected results we have a clear prevalence of the good value compared to that of medium.

4. Conclusions

In conclusion, using the TM for image analysis it was possible to train an artificial intelligence that could recognize the pollution levels defined as good, medium and bad and define the macro areas with the greatest pollution. A forecasting model for air quality was developed with image recognition techniques. During the model training phase, the hyper-parameters were varied until the most performing conditions were reached. The choice of the hyper-parameters was made empirically, by carrying out sensitivity analyzes to varying them. In conclusion, a very performing forecasting model was obtained for the PM10 pollutant. The final model reports 100% accuracy in the test phase, while in the validation phase it reports 97% accuracy. The study will be improved adding information on other pollutants such as PM2.5 which are most dangerous for health.

Acknowledgments

Authors are thankful to Mr. Lucio Bucciero and Mr.Emiliano Nigro for their helpful work.

References

Bishop C.M., 2006, Pattern Recognition and Machine Learning, Springer Science+Business Media, LLC, NY

Jia L., Sun N.G., 2015, A method for determining decision tree complexities based on multiobjective optimization for online learning community, Chemical Engineering Transactions, 46, 193-198 DOI:10.3303/CET1546033

Lotrecchiano N., Sofia D., Giuliano A., Barletta D., Poletto M., 2019, Real-time on-road monitoring network of air quality, Chemical Engineering Transactions, 74, 241–246, doi:10.3303/CET1974041.

Lotrecchiano N., Sofia D., Giuliano A., Barletta D., Poletto M., 2021, Spatial Interpolation Techniques For innovative Air Quality Monitoring Systems, Chemical Engineering Transactions, 86, 391–396, doi:10.3303/CET2186066.

Mitchell T.M., Carbonell J.G, Michalski R.S.,1986, Machine Learning, The Springer International Series in Engineering and Computer Science (SECS), Vol. 12.

Nuhu S.K., Abdul Manan Z., Alwi S.R., Reba M.N.M., 2021, A New Hybrid Modelling Approach for an Eco-Industrial Park Site Selection, Chemical Engineering Transactions, 89, 343-348 DOI:10.3303/CET2189058

Sofia D., Lotrecchiano N., Giuliano A., Barletta D., Poletto M., 2019, Optimization of number and location of sampling points of an air quality monitoring network in an urban contest, Chemical Engineering Transactions, 74, 277–282, doi:10.3303/CET1974047.

Sofia D., Lotrecchiano N., Trucillo P., Giuliano A., Terrone L., 2020, Novel air pollution measurement system based on ethereum blockchain, Journal of Sensor and Actuator Network, 9, 49, doi:10.3390/jsan9040049.

Truong, M. T. N., Kim, S., 2016, A Review on Image Feature Detection and Description, Proceedings of the Korea Information Processing Society Conference, 677–680, https://doi.org/10.3745/PKIPS.Y2016M10A.677

Satoshi Watanabe, 1985, Pattern recognition: Human and Mechanical, John Wiley & Sons Inc

Yao Z., 2018, Prediction of chemical composition in tea based on image processing technology, Chemical Engineering Transactions, 71, 511-516 DOI:10.3303/CET1871086

Zhang Y., 2018, Diagnosis and detection method of critical equipment failure based on electronic nose technology, Chemical Engineering Transactions, 68, 241-246 DOI: 10.3303/CET1868041.