Features Extraction Techniques and Support Vector Classifier for Air Particulate Matters Levels Classification

Rayan Awni Matloob, Mohammed Ahmed Shakir *

University of Duhok, College of Engineering, Electrical and Computer Department, Zakho Street 38, Kurdistan Region, Iraq.

Abstract
Smog is a grave environmental issue, especially in urban areas with high tiers of air pollution from traffic, industry, and other sources. It comprises a mix of gases and particles, including PM2.5, which is particulate matter with a diameter of fewer than 2.5 micrometres. In this work, two schemes are separately proposed. The first scheme applies principal component analysis (PCA) to select some suitable image features. The second scheme employs three wavelet transform filters separately for this purpose. Both approaches use a support vector machine classifier (SVC) for the classification stage. These two schemes classify the images into different classes based on their PM2.5 concentration tiers. The used data are 1990 images in total, randomly divided into (80% for the training process, and the remaining images are used to test the schemes and assess their estimation accuracy). The experimental results demonstrate that the first proposed algorithm achieved (87.15%) accuracy, while the second scheme achieved three accuracy results, 90.43%, 87.91%, and 85.90 % for Ricker, Haar, and Daubechies filters respectively.

Keywords:
Air quality, particulate matter (PM2.5), Image Classification, principal component analysis (PCA), wavelet transform, support vector classifier (SVC).

1. Introduction

The adverse effects of air pollution on human health are mainly attributed to particulate matter, which comprises a significant portion of air pollutants. To evaluate the extent and root causes of such damage globally, particulate matter has been utilized as a critical factor [1]. Several countries have previously employed PMs as a benchmark air pollutant to formulate policies aimed at mitigating health damage caused by exposure to air pollution [2], [3]. At certain concentrations, air pollutants can cause significant harm to humans, plants, and animals [4], [5], [6]. The prevalence of air pollutants poses a significant health risk in many regions worldwide, including but not limited to chronic obstructive pulmonary disease, respiratory diseases and cardiovascular diseases [7], [8], [9]. According to the World Health Organization (WHO), exposure to ambient air pollutants is responsible for 7 million premature deaths worldwide every year (https://www.who.int/health-topics/air-pollution#tab=tab_2, last accessed on 25 March 2023). Urban areas, which are home to over 3.5 billion people (more than half of the world's population), account for most of these premature deaths. With urbanization expected to reach 70% by 2050 [10], the situation is only expected to worsen. Since the industrial revolution in the 19th century, urban expansion has led to significant changes in land use [11]. Multiple studies have linked air pollutants to the SARS-CoV-2 (COVID-19) pandemic since its onset [12], [13], [14]. However, due to reduced factory and traffic activities, the emission of pollutants has decreased, thereby improving air quality [15]. According to research conducted in 31 provinces in China [16], the concentration of NO2, PM2.5, PM10, and CO decreased to varying degrees, outweighing the increase in O3 concentration, which resulted in an overall improvement in air quality.

Several major cities, particularly in China and Taiwan, have established monitoring stations for PM2.5 [17], but due to the high cost and resource requirements, this approach may not always be optimal, especially in large cities where multiple stations would be necessary to achieve adequate coverage. Various technical solutions have been proposed to predict air quality using image classification techniques as an alternative to the challenges associated with station installation, maintenance, and monitoring. Due to the ubiquity of smartphones and their camera capabilities, using images has become a convenient and effective means of communication that can be easily utilized by anyone, anytime, and anywhere. This study introduced two image-based PM2.5 analysis schemes that use image feature extraction to categorize the PM2.5 concentration tiers of outdoor images. The first method suggested in this study

Manuscript received May 5, 2026
Manuscript revised May 20, 2026
https://doi.org/10.22937/IJCSNS.2026.26.5.13

leverages the cutting-edge PCA algorithm, which helps in identifying patterns and underlying structures in a dataset that contains a high number of dimensions or features per observation which led to a significantly reduce number of image features used for classification. The second proposed method uses wavelet transform as a feature extraction method and uses a specific filter that represents an edge detector (Ricker, Haar, and Daubechies), that takes an input image with a specific window size as a parameter. The output of the function is an array of features extracted from the image. Both approaches use SVC for training and testing. SVC is the implementation of SVM for classification, The primary objective of the classifier is to locate the hyperplane that maximizes the gap between the two classes, also known as the margin. The Shanghai dataset (1954 photos, one scene) from [18] was used to test our methodology.

The remainder of this paper is structured into discrete sections. In Section 2, we review previous research pertaining to the categorization of aerial images. Section 3 of this paper outlines the contributions of our study and elaborates on the experimental procedures employed. Our recommended methods, utilizing Principal Component Analysis and Wavelets, are detailed in Section 4. In Section 5, we present our experiments' results and highlight our methodology's effectiveness. Finally, we conclude this article in Section 6.

2. Methodology

In **[19]**, The authors of this manuscript proposed two methods for predicting air pollution level from images captured by a smartphone camera. The first method involved feature extraction by pre-processing the images and using Gabor transform, followed by modelling with Random Forest classification and KNN methods. In the second method, a CNN was designed to classify the raw images. The authors evaluated their proposed methods using a dataset of images collected from the city of Tehran and reported an accuracy of 59.38% for the CNN-based method, which was higher by 6% than the traditional combination of feature extraction and classification methods. The author's contribution lies in the development of a novel approach for predicting air pollution levels from images, which could be useful for individuals and organizations interested in monitoring air quality. However, they note that additional datasets with a wider range of locations, days, and pollution levels are needed to improve accuracy.

The manuscript in [20] outlines a novel method for estimating the outdoor images PM2.5 index. The authors' approach involves utilizing deep learning and support vector regression techniques in conjunction with both image and weather information to develop an effective and easily accessible PM2.5 monitoring system. Initially, a convolutional neural network (CNN) is employed to estimate the pollution concentration based on image data. The predicted PM2.5 value is then combined with two weather parameters (namely, humidity and wind speed) using an SVR model to produce the final estimated PM2.5 index. To evaluate their proposed approach, the authors utilized two datasets from Beijing and Shanghai, and the experimental findings demonstrated a high level of accuracy, achieving 86.71% on the Shanghai dataset and 82.49% on the Beijing dataset. Overall, this study provides a promising approach to estimating PM2.5 levels from outdoor images, which could have a significant impact on raising public awareness and improving air quality.

The researchers of **[21]** paper utilized meteorological data and images to predict PM2.5 indices for outdoor photos. They employed support vector regression (SVR) and deep learning techniques and combined two datasets from Beijing and Shanghai cities in China. The proposed approach involved using an SVR model to integrate the predicted PM2.5 from the convolutional neural network (CNN) with two meteorological parameters, namely wind speed, and humidity, to provide the expected PM2.5 index outcomes. In the Shanghai dataset, the proposed model achieved a 26.08% reduction in root mean square error (RMSE) and a 24.57% increase in R-squared. Similarly, for the Beijing dataset, the proposed model reduced the RMSE by 5.27% to 56.03 and increased the R-squared by 8.4% to 0.6046.

The researchers in [**22**] proposed a CNN-RC learning scheme that estimate air quality by combines a regression classifier with the convolutional neural network in specified locations. The model uses the shot extraction feature and feature categorization into air quality categories to accurately calculate air quality levels. To enhance model dependability and estimation accuracy, the authors trained the model on datasets comprising various combinations of the current image, HSV characteristics, and baseline image. The researchers used the Kaohsiung City's monitoring station in Taiwan (Linyuan station), to collect a dataset of 3549 hourly images, besides their PM2.5, and their AQI, to evaluate the model's performance. Their model achieves an estimation accuracy of 76% for R2 for PM2.5 using day (night) photos. By utilizing a single deep learning model, this approach can provide fast and precise image-based estimations of multiple pollutants simultaneously.

In a study conducted by Malaysia's Smart Cities in 2017-2018, a machine learning model was developed to predict PM2 concentrations **[23]**. The dataset underwent preprocessing steps such as data cleaning and normalization. During the feature extraction phase, the dataset was reduced to include location and temporal components. Three supervised machine learning classifiers were employed - long short-term memory (LSTM), artificial neural network (ANN), and random forest (RF) - along with the chi2

function for classification tasks. Several characteristics, including location, station ID, day of the year, and PM2.5 tier, were selected based on their significant correlation with the target feature ("Label"), which was used for training the model. The dataset was divided into a 60-40 train-test split, with labeled data used for classifier training and testing. The study compared the output of the three models using the confusion matrix, with the RF model achieving the highest accuracy at 97.7%, followed by LSTM (61.77%) and ANN (61.14%). Overall, the study demonstrated the effectiveness of machine learning models in predicting PM2 concentrations.

The authors of **[24]** proposed YOLO-AQI, a real-time deep learning model that uses photographs and a dataset with six air quality classifications (AQI tiers, 1-6) to estimate air quality. According to the scheme achieved outcome, the model AQI estimation accuracy reached 75.15% on the NWNUAQI database. The researchers compared the model's performance to that of other models, including ResNet, AlexNet, VGG, GoogleNet, and MobileNet. The results showed that their proposed YOLO-AQI model outperformed the competition, with ResNet having equal accuracy and the others below (65.6) and (72.4). These findings demonstrate the efficacy of the proposed approach.

3. Contribution and Material

3.1 Contributions

The following is a summary of our work's significant contributions:

- This paper proposed two schemes for estimating air quality (PM2.5 tiers) using one scene image.
- The first proposed method uses the concept of PCA to select suitable image features. As each image's dimensions are (389, 584, 3), which is (681528 features per image). Only (1593 features/images) are selected, instead of (681528), which represents a (99.768%) reduction in the actual image features.
- The second method uses three different wavelet transform filters (Ricker, Haar, and Daubechies) for feature extraction.
- After features extraction using the mentioned two methods separately, both use a machine learning model, called Support Vector Machine (SVM) algorithm with a linear kernel (**SVC** class in the scikit-learn library), for class learning and prediction testing. SVC stands for Support Vector Classifier, which is a variation of the SVM algorithm and used for solving the classification problem by constructing an optimal hyperplane.
- This manuscript's primary objective is to provide a classification that considers the following five categories of pollutants: tier 1, tier 2, tier 3, tier 4, and tier 5.

As far as we are aware, no study has been done using the algorithms of principal component analysis or wavelet transform for air pollution PM2.5 levels classification.

3.2 Dataset and Augmentation

The AQI index is an essential tool for monitoring the quality of the air we breathe [25]. This index provides a daily report on the air quality and helps to determine whether the air is clean or polluted. Moreover, it acts as a warning system by alerting individuals to potential health risks associated with inhaling polluted air. The AQI specifically highlights potential daily and hourly health effects, which may occur after exposure to air pollution. Interestingly, the AQI value is inversely proportional to pollution level and the corresponding health issues, indicating that the higher the AQI value, the lower the risk of health problems associated with air pollution. In this study, we have examined the air quality in Shanghai by analysing photographs captured at fixed locations in the city [18], [21], and [26] along with their corresponding PM2.5 levels. To categorize the images, we have classified them into five different tiers based on their respective PM2.5 levels, measured in micrograms per cubic meter (µg/m3). Our analysis revealed that the PM2.5 images fall within tier 1, ranging from (0 to 18.5 µg/m3), and continue up to tier 5, where the PM2.5 concentration is greater than (59.9 µg/m3), as summarized in Table 1.

Table 1, PM2.5 classes and concentration

Class Level	Particulate matter concentration
Tier 1	<= 18.4
Tier 2	18.5 - 30.4
Tier 3	30.5 - 40.4
Tier 4	40.5 - 59.9
Tier 5	>= 60

The figshare platform hosts the Shanghai PM2.5 dataset, which comprises nearly 1954 high-resolution images depicting varying levels of air quality throughout the daylight hours. These images have a resolution of 389

by 584 pixels, offering detailed insights into the air quality tiers in Shanghai.

To effectively employ the Shanghai images, it is necessary to balance the dataset. This entails ensuring that the number of images in each class is relatively uniform to avoid bias during the classification procedure. As the dataset consists of images depicting varying levels of particulate matter, and some levels contain a greater number of images than others, an approach was taken to balance the dataset. This approach involved not using all the images, and also using data augmentation, which facilitated an equal distribution of images across all PM levels. Horizontal flipping of the images was used as a means of augmentation, which resulted in a total of 1990 images. Table (2) displays the count of images in each class, with some images excluded and the used augmented to maintain balance among the data sets.

Table 2, The number of images in each class

Class	No. of images
Tier 1	320
Tier 2	407
Tier 3	410
Tier 4	444
Tier 5	409
All classes' images	1990

4. Methodology

For the first proposed scheme, as of using PCA and SVC, the methodology outlined in the used Python code can be described as follows:

1. The first step in the code is to load the images from the specified directory. The function takes a path as input, reads the subfolders in the directory, and then reads the images in each subfolder. The features of each image are then flattened and stored in a list. The labels of each image are also stored in a separate list.
2. The data is then pre-processed by normalizing the features using the Variance Threshold function. This helps in reducing the noise and improving the accuracy of the model.
3. The data is then split randomly into training and testing sets using an 80-20 split ratio.
4. Principal Component Analysis (PCA) is then used to reduce the dimensionality of the data (reduce the used features from 681528 to 1593 features only, which is a 99.768% reduction ratio). This is done to reduce the computation time required to train the model.
5. A Support Vector Classifier (SVC) with a linear kernel is then trained using the training set (80% of the used images). The SVC classifier is a popular algorithm used for image classification tasks and tries to find the hyperplane that maximizes the margin between the two classes.
6. The trained model is then evaluated using the testing set (20% of the used images). The accuracy score is calculated using the accuracy score function from the scikit-learn library. The accuracy score gives us an idea of how well the model is performing on the unseen data.

For the second proposed scheme, the methodology used in the Python code for image classification using wavelet transform filters and SVC involves the following steps:

1. The first step is to load the data from the given paths. The images are loaded, and the corresponding labels are assigned.
2. The next step is to apply one of the wavelets transforms filters on the loaded images. In this step, each image is divided into smaller windows of fixed size (64,64), and a wavelet kernel is applied to each window. The output of this step is a set of features extracted from each window of the image.
3. The data is then split randomly into training and testing sets using an 80-20 split ratio.
4. The extracted features and corresponding labels are used to train an SVC model with a linear kernel.
5. Once the model is trained, the accuracy of the model is tested using the test set, to predict the labels of the test images based on the extracted features.

The accuracy is calculated by comparing the predicted labels with the actual labels of the test images. The flowchart of both methods is shown in figures (1 and 2).

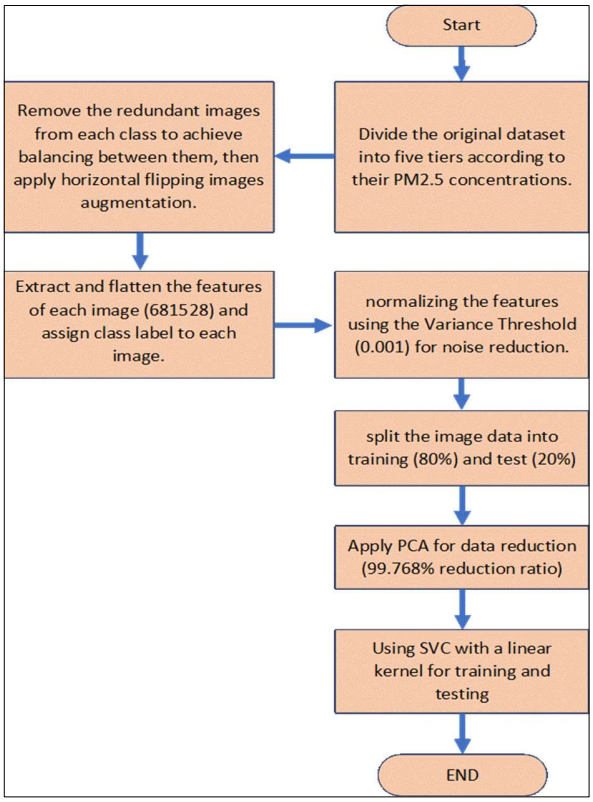

Figure 1, The first scheme process flowchart

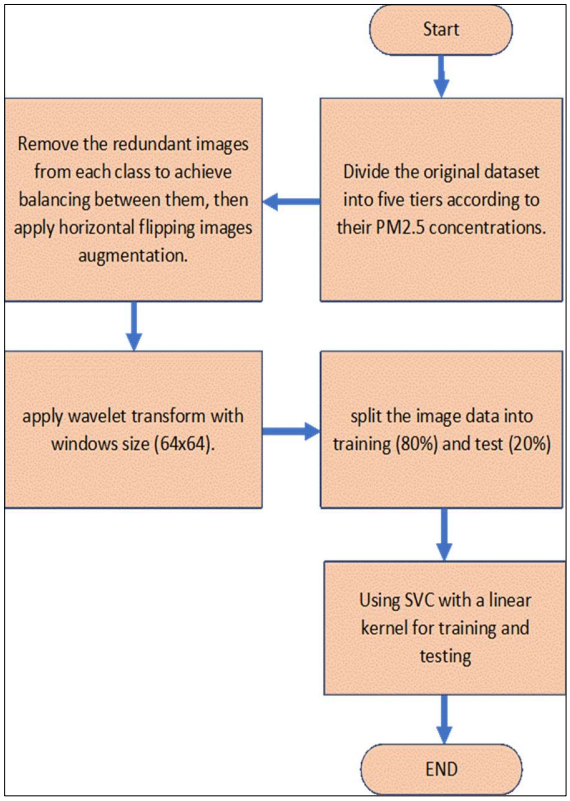

Figure 2, The second scheme process flowchart.

4.1 Variance threshold

The variance threshold is a feature selection method that removes low-variance features from a dataset [27]. It is based on the idea that features with low variance contribute little to the model's predictive power and can be safely removed without affecting the model's performance [28]. In the case of image data, the variance threshold can be used to remove features (i.e., pixels) that have little variance across the entire dataset. For example, consider an image dataset with a large number of images of a blue sky. In this case, the blue pixels in the sky region will have high variance, whereas the pixels in other regions (e.g., clouds or trees) may have low variance. By applying a variance threshold, the pixels in the regions with low variance can be removed, which can reduce the dimensionality of the dataset and improve the model's performance. The variance threshold method is implemented in scikit-learn library using the **VarianceThreshold** class. The **threshold** parameter is used to specify the threshold below which features are removed. Any feature with a variance below the specified threshold is removed from the dataset. The remaining features can then be used to train a model.

4.2 Principal Component Analysis

Dimensionality reduction can be achieved using Principal Component Analysis (PCA), which is a widely used technique [29]. By identifying the principal components (PCs) of a high-dimensional dataset, which are linear combinations of the original features that capture the most variance in the data, PCA transforms the dataset into a lower-dimensional space while retaining as much of the original information as possible [29], [30]. The first PC captures the direction of maximum variance, the second PC captures the direction of maximum variance orthogonal to the first, and so on [30]. PCA is employed for various purposes, including data compression, noise reduction, feature extraction, and visualization. It is often used as a pre-processing step in machine learning to reduce the dimensionality of a dataset before training a model, which can enhance the accuracy and efficiency of the model by reducing the number of features that need to be processed [29], [30]. The basic principle behind PCA is to identify and discard the less important information in the data by finding the directions in which the data varies the most and projecting the data onto those directions. The resulting

lower-dimensional dataset retains most of the information in the original dataset but with fewer dimensions. In summary, PCA is a valuable technique for understanding complex datasets, identifying critical variables or features, and decreasing the dimensionality of the data for further analysis.

4.3 Wavelet transform.

Wavelet Transform is a mathematical technique used for signal processing and image compression. It involves decomposing a signal or image into a set of basic functions called wavelets, which are localized in both time and frequency domains [31]. Next a brief explanation of the set of used filters.

The Ricker wavelet filter, also known as the Mexican hat wavelet, is widely used in wavelet analysis for detecting seismic events in geophysics, as well as in image processing and other fields [32]. It is a type of wavelet with a bell-shaped curve that can effectively capture the features of waveforms with localized oscillations. The Ricker wavelet is defined mathematically as a second-order derivative of a Gaussian function, which gives it its characteristic bell shape [32]. It has a peak frequency that can be adjusted by changing its parameter, known as the central frequency. It is also a useful tool for wavelet analysis due to its ability to accurately capture the high-frequency components of a signal or image.

The Haar filter is a type of wavelet transform that is used for analyzing signals or images. It is named after Alfred Haar, who introduced it in 1909. It is a simple wavelet that consists of a step function (with a value of -1 on one half of the interval and a value of +1 on the other half.) followed by a ramp function [33]. In image processing, the Haar wavelet transform is used to detect edges and features in images and applied on non-overlapping rectangular regions of an image called windows [33]. The wavelet transform of each window is obtained by convolving it with a Haar wavelet kernel, which is a different filter that detects edges. The resulting wavelet coefficients represent the edge information in the window. The written Python code in this research takes an input image and performs the Haar wavelet transform to extract features from the image. The Haar filter is a useful technique for feature extraction because it can highlight edges and other high-frequency information in an image while suppressing low-frequency information such as smooth regions or backgrounds [34]. This makes it well-suited for tasks such as object detection and recognition, where identifying the edges and contours of objects is important.

The Daubechies filter is a type of wavelet transform that is used for image compression and denoising. It is named after Ingrid Daubechies, who introduced it in 1988 [35]. This filter is a family of wavelets that are designed to have compact support and orthogonality properties. In image processing, the Daubechies wavelet transform is used to decompose an image into four sub-bands: approximation (low frequency), horizontal, vertical, and diagonal (high frequency). The transform is applied on non-overlapping rectangular regions of an image called windows. The wavelet coefficients in each sub-band represent the frequency information in the window [35].

4.4 Support vector classifier

Support Vector Classifier (SVC) is a popular machine learning algorithm used for classification tasks. It belongs to the family of supervised learning algorithms, where the model is trained on a labeled dataset to predict the class of unseen data points [36]. The main idea behind SVC is to find a hyperplane that separates the data points into different classes while maximizing the margin between the hyperplane and the closest data points from both classes. The hyperplane that maximizes the margin is known as the optimal separating hyperplane (OSH). To find the OSH, SVC solves an optimization problem that involves minimizing the misclassification error while maximizing the margin. This optimization problem is typically formulated as a quadratic programming problem and solved using numerical methods. However, in many cases, the data may not be linearly separable, i.e., a hyperplane cannot separate the data points into different classes without making any errors. To handle such cases, SVC uses a technique called kernel trick, where the input data is transformed into a higher-dimensional space where a hyperplane can be found to separate the data points. The kernel function is used to define the similarity between two data points in the higher-dimensional space [36].

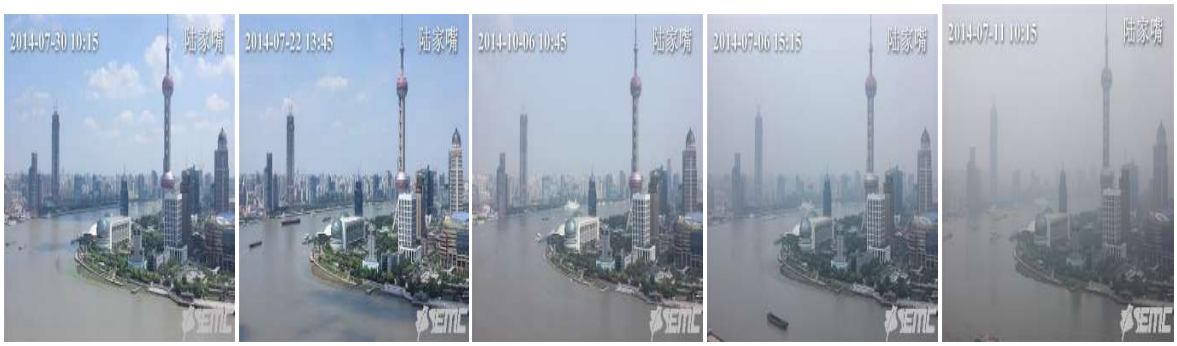

Figure 3, a variability of the five classes starting by tier 1, tier 2, tier 3, tier 4, and tier 5.

5. Experimentation Settings and results

The images are in (.jpg) format and have a size of 389x584x3. Figure (3) highlights the variability of the five classes. Using the PCA technique has a big impact on reducing the used features by over (99%). This reduction plays a big role in reducing the needed processing time. Besides, in the case of reporting the data to external server, this reduction is important from the communication side. The dataset was balanced by excluding some images from the overwhelmed classes and expanding other classes through image data augmentation. By using Python programming language and achieved by flipping the images horizontally to one side only (right). The variance threshold is used before applying the images to PCA to remove low-variance features from a dataset. Finally, the SVC is used as a training machine learning. On the other hand, the wavelet uses the image features without any previous processing by using three types (Ricker, Haar, and Daubechies), then training the SVC with the resulting wavelet features. The proposed model is trained on 80% of the dataset, while the remaining 20% of the data are used for testing progress. The statistics about the training, test, and validation for the sub-images datasets are detailed in Table (3).

Table 3, Datasets statistics for training, validation, and testing.

Set	Percentage	No. of images
Train	80 %	1593
Test	20 %	397

The environment for the experiment used along with Anaconda Navigator (anaconda3) and Python 3.9.15 is as follows:

- System Manufacturer: LENOVO LEGION5
- OS Name: Microsoft Windows 11 Home
- System Type: x64-based PC
- Processor: Intel(R) Core (TM) i7-10750H CPU @ 2.60GHz, 2592 Mhz, 6 Core(s), 12 Logical Processor(s)
- Installed Physical Memory (RAM): 16.0 GB

The first method that employed PCA achieved an accuracy equal to 87.15%. For the second method as three different filters are used, three different results were achieved as follows: The Ricker wavelet transform achieved 90.43% followed by 87.91% and 85.90% for Haar and Daubechies respectively as shown in table (4)

Table 4, The proposed methods and their achieved accuracies

Methods	Achieved accuracy
First method	87.15 %
Second method / Ricker filter	90.43 %
Second method / Haar filter	87.91 %
Second method / Daubechies filter	85.90 %

5.1 Comparison with Previous Works

In the related work in section 2 (related work), discuss different approaches to predicting air pollution levels using different models. they propose different methods of predicting PM2.5 levels by combining image information with weather data or other variables using various machine learning models. This work achieved a remarkable accuracy regarding the small amount of the image dataset compared with some other works as shown in table (5).

Table 5, is a brief comparison of some previous works with this paperwork.

	Method / Model	Dataset	Key findings
[19]	Feature extraction (Gabor transform) + KNN/Random Forest classification; CNN classification	Images from Tehran	CNN-based method: 59.38% accuracy, 6% higher than the traditional method. Need for additional datasets to improve accuracy.
[20]	CNN for PM2.5 prediction + SVR for combining PM2.5 prediction with humidity and wind speed	Datasets from Shanghai and Beijing	Achieved accuracy of 86.71% and 82.49% on respective datasets. A promising approach for estimating PM2.5 from outdoor images to improve public health and awareness of air pollution.
[21]	Merge meteorological data and images for PM2.5 prediction using SVR and deep learning techniques	Datasets from Beijing and Shanghai	Reduced RMSE by 26.08% and increased R-squared by 24.57% for the Shanghai dataset. Reduced RMSE by 5.27% to 56.03 and increased R-squared by 8.4% to 0.6046 for the Beijing dataset.
[22]	CNN-RC model (CNN with regression classifier) for air quality calculation at specified places	Linyuan air station Dataset in Taiwan	Estimation accuracy of 76% for PM2.5 using day (night) photos. A Model trained on datasets comprising different combinations of images, HSV characteristics, and baseline images for improved accuracy.
[23]	Three supervised machine learning classifiers (LSTM,	Smart Cities dataset from Malaysia (2017-2018)	RF model had the highest accuracy (97.7%) compared to ANN (61.14%) and LSTM (61.77%). Chi-squared

	ANN, RF) for PM2.5 prediction		statistics were used for feature selection.
[24]	YOLO-AQI real-time deep learning model for AQI estimation based on six air quality classifications	Dataset of over 5600 photos	Achieved 75.15% accuracy on the NWNUAQI dataset. YOLO-AQI model compared with ResNet model.
This research	PCA and Wavelet transform filters	1954 images from Shanghai	The first method's accuracy is 87.15%. The second method accuracies are (90.43%, 87.91%, and 85.90%) using Ricker, Haar, and Daubechies filters respectively

6. Conclusion And Future Work

In this work, two methods (using PCA and wavelet techniques) are proposed separately for air quality classification. Based on PM2.5 concentration tiers images as utilized with SVC. The two proposed methods achieved remarkable estimation accuracy using the Shanghai dataset. Compared to existing research, the performance of our approaches outperforms state-of-the-art methods. The second method/the Ricker filter achieved the highest prediction accuracy by (90.43%). While less accuracy was obtained by the Second method / Daubechies filter.

Based on the experimental results we conclude the following:

- The Ricker wavelet filter has a good balance between temporal and frequency resolution, which means that it can accurately capture changes in both time and frequency domains. This is important for feature extraction tasks where the goal is to capture fine details and edges in an image. Furthermore, The Ricker wavelet transform has a high degree of symmetry, which makes it more stable and robust in the presence of noise and other distortions. This helps to ensure that the extracted features are more accurate and reliable. Besides, it can be easily adapted to different image processing tasks by adjusting its parameters such as the number of scales and the size of the filter.
- The Haar wavelet transform is well suited to capture the underlying patterns and structures in the data. Besides, it can capture abrupt changes in signals or images, making them a popular choice for image compression and denoising. It also can be more robust to noise since it decomposes the data into different scales and frequencies, which can help filter out the noise.
- The nature of the used image some of which contain the PM atoms, that makes PCA not as effective as the Haar wavelet transform. PCA tries to find the linear combinations of the variables that explain the most variance in the data, but if the data is noisy, these linear combinations might not capture the underlying patterns in the data.
- Daubechies filter provided less accurate than others. That is because it may not have sufficient time-frequency localization, which means that it might not be able to capture fine details and edges in an image as accurately. Besides, it is more susceptible to distortions and noise in the data. Furthermore, the fixed number of scales and coefficients can make it less adaptable to different types of images and feature extraction tasks.

In the future, our approaches can be extended by testing other wavelet filters such as the Coiflets, Symlets, or Biorthogonal wavelets. We could explore the use of deep learning techniques. Also, instead of relying on a single method for image classification, it may be beneficial to combine multiple techniques using ensemble methods.

Reference

[1] Lelieveld, J., Evans, J.S., Fnais, M., Giannadaki, D. and Pozzer, A., 2015. The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature*, *525*(7569), pp.367-371.

[2] Park, G.S., 2006. Basic Plan on the Metropolitan Area Air Quality Management. In *Proceedings of the Korea Air Pollution Research Association Conference* (pp. 37-42). Korean Society for Atmospheric Environment.

[3] Ha J. (2014). Applying policy and health effects of air pollution in South Korea: focus on ambient air quality standards. *Environmental health and toxicology*, *29*, e2014011. https://doi.org/10.5620/eht.e2014011

[4] Baklanov, A., Molina, L. T., & Gauss, M. (2016). Megacities, air quality, and climate. In Atmospheric Environment (Vol. 126, pp. 235–249). Elsevier BV. https://doi.org/10.1016/j.atmosenv.2015.11.059

[5] Kinney, P. L. (2018). Interactions of Climate Change, Air Pollution, and Human Health. In Current Environmental Health Reports (Vol. 5, Issue 1, pp. 179–186). Springer Science and Business Media LLC. https://doi.org/10.1007/s40572-018-0188-x

[6] Pautasso, M., Dehnen-Schmutz, K., Holdenrieder, O., Pietravalle, S., Salama, N., Jeger, M. J., Lange, E., & Hehl-Lange, S. (2010). Plant health and global change - some implications for landscape management. In Biological Reviews (p. no-no). Wiley. https://doi.org/10.1111/j.1469-185x.2010.00123.x

[7] Brauer, M., Freedman, G., Frostad, J., van Donkelaar, A., Martin, R. V., Dentener, F., Dingenen, R. van, Estep, K., Amini, H., Apte, J. S., Balakrishnan, K., Barregard, L., Broday, D., Feigin, V., Ghosh, S., Hopke, P. K., Knibbs, L. D., Kokubo, Y., Liu, Y., … Cohen, A. (2015). Ambient Air Pollution Exposure Estimation for the Global Burden of Disease 2013. In Environmental Science & Technology (Vol. 50, Issue 1, pp. 79–88). American Chemical Society (ACS). https://doi.org/10.1021/acs.est.5b03709

[8] Lelieveld, J., Barlas, C., Giannadaki, D., & Pozzer, A. (2013). Model calculated global, regional and megacity premature mortality due to air pollution. In Atmospheric Chemistry and Physics (Vol. 13, Issue 14, pp. 7023–7037). Copernicus GmbH. https://doi.org/10.5194/acp-13-7023-2013

[9] Manisalidis, I., Stavropoulou, E., Stavropoulos, A., & Bezirtzoglou, E. (2020). Environmental and Health Impacts of Air Pollution: A Review. In Frontiers in Public Health (Vol. 8). Frontiers Media SA. https://doi.org/10.3389/fpubh.2020.00014

[10] United Nations: World Urbanization Prospects: The 2018 Revision, United Nations Department of Economic and Social Affairs, Population Division, New York, 2018.

[11] Seto, K. C., Güneralp, B., & Hutyra, L. R. (2012). Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools. In Proceedings of the National Academy of Sciences (Vol. 109, Issue 40, pp. 16083–16088). Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.1211658109

[12] Filonchyk, M., Hurynovich, V., Yan, H., Gusev, A., & Shpilevskaya, N. (2020). Impact Assessment of COVID-19 on Variations of SO2, NO2, CO and AOD over East China. In Aerosol and Air Quality Research (Vol. 20, Issue 7, pp. 1530–1540). Taiwan Association for Aerosol Research. https://doi.org/10.4209/aaqr.2020.05.0226

[13] Ma, Y., Xu, J., Gao, C., & Tong, X. (2022). Impacts of COVID-19 Travel Restriction Policies on the Traffic Quality of the National and Provincial Trunk Highway Network: A Case Study of Shaanxi Province. In International Journal of Environmental Research and Public Health (Vol. 19, Issue 15, p. 9387). MDPI AG. https://doi.org/10.3390/ijerph19159387

[14] Wang, J., Lei, Y., Chen, Y., Wu, Y., Ge, X., Shen, F., Zhang, J., Ye, J., Nie, D., Zhao, X., & Chen, M. (2021). Comparison of air pollutants and their health effects in two developed regions in China during the COVID-19 pandemic. In Journal of Environmental Management (Vol. 287, p. 112296). Elsevier BV. https://doi.org/10.1016/j.jenvman.2021.112296

[15] Miyazaki, K., Bowman, K., Sekiya, T., Jiang, Z., Chen, X., Eskes, H., Ru, M., Zhang, Y., & Shindell, D. (2020). Air Quality Response in China Linked to the 2019 Novel Coronavirus (COVID-19) Lockdown. In Geophysical Research Letters (Vol. 47, Issue 19). American Geophysical Union (AGU). https://doi.org/10.1029/2020gl089252

[16] Nie, D., Shen, F., Wang, J., Ma, X., Li, Z., Ge, P., Ou, Y., Jiang, Y., Chen, M., Chen, M., Wang, T., & Ge, X. (2021). Changes of air quality and its associated health and economic burden in 31 provincial capital cities in China during COVID-19 pandemic. In Atmospheric Research (Vol. 249, p. 105328). Elsevier BV. https://doi.org/10.1016/j.atmosres.2020.105328

[17] Ying, L.I. and Li, J., 2022. Impact of the Interannual Variability in Large-Scale Circulation on the Ground-Level Ozone Variability Over Eastern China. Authorea Preprints.

[18] Liu, C., Tsow, F., Zou, Y., & Tao, N. (2016). Particle Pollution Estimation Based on Image Analysis. In H. Liu (Ed.), PLOS ONE (Vol. 11, Issue 2, p. e0145955). Public Library of Science (PLoS). https://doi.org/10.1371/journal.pone.0145955

[19] Vahdatpour, M. S., Sajedi, H., & Ramezani, F. (2018). Air pollution forecasting from sky images with shallow and deep classifiers. In Earth Science Informatics (Vol. 11, Issue 3, pp. 413–422). Springer Science and Business Media LLC. https://doi.org/10.1007/s12145-018-0334-x

[20] Won, T., Eo, Y. D., Sung, H., Chong, K. S., Youn, J., & Lee, G. W. (2021). Particulate Matter Estimation from Public Weather Data and Closed-Circuit Television Images. In KSCE Journal of Civil Engineering (Vol. 26, Issue 2, pp. 865–873). Springer Science and Business Media LLC. https://doi.org/10.1007/s12205-021-0865-4

[21] Won, T., Eo, Y. D., Sung, H., Chong, K. S., Youn, J., & Lee, G. W. (2021). Particulate Matter Estimation from Public Weather Data and Closed-Circuit Television Images. In KSCE Journal of Civil Engineering (Vol. 26, Issue 2, pp. 865–873). Springer Science and Business Media LLC. https://doi.org/10.1007/s12205-021-0865-4

[22] Kow P-Y, Hsia I-W, Chang L-C, Chang F-J (2022) Real-time image-based air quality estimation by Deep Learning Neural Networks. Journal of Environmental Management 307:114560. doi: 10.1016/j.jenvman.2022.114560

[23] Palanichamy, N., Haw, S.-C., S, S., Murugan, R., & Govindasamy, K. (2022). Machine learning methods to predict particulate matter PM2.5. In F1000Research (Vol. 11, p. 406). F1000 Research Ltd. https://doi.org/10.12688/f1000research.73166.1

[24] Zhang, Q., Tian, L., Fu, F., Wu, H., Wei, W., & Liu, X. (2022). Real-Time and Image-Based AQI Estimation Based on Deep Learning. In Advanced Theory and Simulations (Vol. 5, Issue 6, p. 2100628). Wiley. https://doi.org/10.1002/adts.202100628

[25] Pandithurai, O., Bharathiraja, N., Pradeepa, K., Meenakshi, D. and Kathiravan, M., 2023, February. Air Pollution Prediction using Supervised Machine Learning Technique. In *2023 Third International Conference on Artificial Intelligence and Smart Energy [ICAIS]* [pp. 542-546]. IEEE.

[26] Bo, Qirong, et al. "Particle pollution estimation from images using convolutional neural network and weather features." 2018 25th IEEE International Conference on Image Processing (ICIP). IEEE, 2018

[27] Ambarwati, Y.S. and Uyun, S., 2020, December. Feature selection on magelang duck egg candling image using variance threshold method. In *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems [ISRITI]* [pp. 694-699]. IEEE.

[28] Guyon, I. and Elisseeff, A., 2003. An introduction to variable and feature selection. *Journal of machine learning research*, *3*[Mar], pp.1157-1182.

[29] Ghojogh, B., Crowley, M., Karray, F. and Ghodsi, A., 2023. Principal Component Analysis. In *Elements of Dimensionality Reduction and Manifold Learning* [pp. 123-154]. Cham: Springer International Publishing.

[30] Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philos Trans A Math Phys Eng Sci. 2016 Apr 13;374[2065]:20150202. doi: 10.1098/rsta.2015.0202. PMID: 26953178; PMCID: PMC4792409.

[31] Zhang D [2019] Wavelet transform. Texts in Computer Science 35–44. doi: 10.1007/978-3-030-17989-2_3

[32] Singh, A., Rawat, A. and Raghuthaman, N., 2022. Mexican hat wavelet transform and its applications. *Methods of Mathematical Modelling and Computation for Complex Systems*, pp.299-317.

[33] Stanković, R.S. and Falkowski, B.J., 2003. The Haar wavelet transform: its status and achievements. *Computers & Electrical Engineering*, *29*[1], pp.25-44.

[34] Gonzalez, R.C. and WOODS 3rd, R.E., 2008. Edition. *Digital Image Processing. Upper Saddle River, USA: Prentice Hall*.

[35] Daubechies, I., 1988. Orthonormal bases of compactly supported wavelets. *Communications on pure and applied mathematics*, *41*[7], pp.909-996.

[36] Feng, Y., 2023. Support Vector Machine for Stroke Risk Prediction. *Highlights in Science, Engineering and Technology*, *38*, pp.9