IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 24, 2022, accepted February 10, 2022, date of publication February 15, 2022, date of current version March 7, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3151632

Machine Learning-Based Estimation of $PM_{2.5}$ Concentration Using Ground Surface DoFP Polarimeters

MAEN TAKRURI[1], (Senior Member, IEEE), ABUBAKAR ABUBAKAR[2], ABDUL-HALIM JALLAD[3,4], (Member, IEEE), BASEL ALTAWIL[5], PRASHANTH R. MARPU[6], (Senior Member, IEEE), AND AMINE BERMAK[2], (Fellow, IEEE)

[1]Department of Electrical, Electronics, and Communications Engineering, American University of Ras Al Khaimah, Ras Al-Khaimah, United Arab Emirates
[2]Division of Information and Computing Technology, College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar
[3]Department of Electrical Engineering, United Arab Emirates University, Al Ain, United Arab Emirates
[4]National Space Science and Technology Center, United Arab Emirates University, Al Ain, United Arab Emirates
[5]Yahsat Space Laboratory, Khalifa University (KU), Abu Dhabi, United Arab Emirates
[6]Group 42, Abu Dhabi, United Arab Emirates

Corresponding author: Abdul-Halim Jallad (a.jallad@uaeu.ac.ae)

ABSTRACT In this paper, we propose a machine learning system for the estimation of atmospheric particulate matter (PM) concentration, specifically, particles with a maximum diameter of 2.5μm. These very fine particles, also known as $PM_{2.5}$ particles, are very dangerous to the human body as they are small enough to penetrate deep areas of the vital organs. The proposed system uses a combination of features from both polarimetric and spectral imaging modalities in training and developing a machine learning model that provides high accuracy $PM_{2.5}$ estimates. Furthermore, acquisition of the polarimetric images is done near the ground surface with a horizontal field of view aiming at standard targets which enables higher accuracy at the surface level. The accuracy of the approach was verified through a study conducted during the summer months of the United Arab Emirates (UAE). The proposed system employs different machine learning techniques such as Support Vector Regression (SVR), Gaussian Process Regression (GPR), and Bagging Ensemble Trees (BET), to provide high accuracy $PM_{2.5}$ estimates. Our proposed system achieves the best performance within the red wavelength with accuracy up to 93.8627% and an R^2 score up to 0.9420.

INDEX TERMS Division of focal plane, environmental monitoring, machine learning, polarization image.

I. INTRODUCTION

The incident light from solar radiation is characterized by intensity, wavelength, and polarization. While intensity and wavelength are respectively perceived as brightness and color, the polarization characteristic is imperceptible to the human eye. As a result, so many applications in the field of applied optics only employ intensity and wavelength. In more recent times, the polarization property of light is shown to provide useful information and as a result, it has been employed in various fields such as food monitoring [1], material classification [2], [3]. Polarimetry is also found to be a promising remote sensing method for the monitoring and characterization of atmospheric aerosols [4].

The associate editor coordinating the review of this manuscript and approving it for publication was Gerardo Di Martino.

Aerosol is a mixture of various small particles of different shapes, morphologies, and composition. The radiative and optical properties of such a mixture are characterized by many complex parameters which need to be recorded for a reliable characterization of aerosols. To record the requisite information about the properties of aerosols, the widely employed instruments are multi-angular multi-spectral polarimeters. Indeed, the sensitivity of observations to detailed aerosol properties could be maximized by the simultaneous spectral, angular and polarimetric measurements of atmospheric radiation [5]–[9].

Aerosol particles' sizes range from a few tenths to several tens of micrometers. Although these particles are invisible to the human eye, their interaction with solar radiation impacts other important parameters such as total atmospheric energy budget, atmospheric visibility, climate dynamics, as well as

air quality [4]. In general, aerosols are mostly characterized by the presence of microscopic particles suspended in the air known as particulate matter (PM). As a result, the term ''aerosol'' is often used to refer to the particulate/air mixture [10]. The particulate matters in aerosols are of two groups – the group with particles having a diameter of 10μm or less, known as PM_{10}; and the group with particles having a diameter of 2.5μm or less, known as $PM_{2.5}$ [11]. These particulates are quite harmful to the human body due to their ability to penetrate deep into the lungs, brain, and blood streams [12].

In a 2013 study involving 312,944 people in nine European countries, the significant danger of particulates was revealed [13]. The study showed that for every increase of $10\mu g/m^3$ in PM_{10} level, the rate of lung cancer rose by 22%; while for the same level of $PM_{2.5}$ increase, the lung cancer rate rose by 36%. From the results of this study, it was also observed how the $PM_{2.5}$ particles are more deadly as they can deposit in deeper parts of the lung causing tissue damage and inflammation. Some other studies to determine the effect of $PM_{2.5}$ concentration levels on the human health have been reported in [14]–[16].

Previous works have shown that polarimetry techniques are promising in the characterization of atmospheric aerosols [5]–[9], [17]. Initially, monitoring of aerosol properties was done by space-borne polarimetry in the late 1980s and early 1990s. Currently, there are several instruments that have already provided polarization observations from space. The first and most extensive record of such space-borne polarimetric imagery was provided by POLDER-I [18], POLDER-II and POLDER/PARASOL multi-angle multi-spectral polarization sensors [19]. More recently, in [20], a multi-angle Stokes vector analyzer was utilized to characterize aerosol particles.

Over the past decades, ground-based polarimetric measurements have been evolving. Some monitoring stations include the CE318 sun/sky-radiometer manufactured by the Cimel Electronique for measuring atmospheric aerosol and water vapor measurements [21]. The most recent of CE318 version, the CE318-DP [22], possess eight wavelengths in addition to its capability to measure polarization. The Degree of Linear Polarization (DoLP) is calculated at each wavelength and the spatial distribution of the sky polarization is essentially related to the optical and microphysical properties of aerosols. Other ground-based observations include the GroundSPEX spectropolarimeter [23] and the GroundMSPI [24]. Although the characterization of aerosol particles - especially fine particles - is improved by polarimetry, major observational networks such as AERONET [25] are reluctant to include the measurements as part of the routine retrievals. This is due to the complexity of acquiring and interpreting polarization data.

To interpret and analyze any recorded data, Machine Learning models using features other than the polarimetric kinds have been utilized. Some of these models include the random forest model [26] to estimate the quantity of $PM_{2.5}$ in China; and [27] utilized a random forest approach for PM predictions in US.

More recently, a geographically and temporally weighted neural network constrained by global training (GC-GTWNN) was proposed in [28], for the estimation of surface $PM_{2.5}$. The proposed model which was tested across China utilized satellite AOD and surface $PM_{2.5}$ measurements in addition to other auxiliary variables to address the nonlinear spatio-temporal relationship between AOD and $PM_{2.5}$. In [29], a deep learning model ''EntityDenseNet'' was proposed to retrieve ground-level $PM_{2.5}$ concentrations. A key feature of this model is its ability to automatically extract $PM_{2.5}$ spatio-temporal characteristics. A common theme to the aforementioned models was the non-consideration and non-utilization of polarimetric features and observations. However, the studies reported in [30] and [31] have indicated the significant potential of polarimetric observations.

In this work, we investigate the use of polarimetry in the estimation of $PM_{2.5}$ with the aid of machine learning techniques. The study is conducted in the United Arab Emirates (UAE) whose desert climate is characterized in summer by dusty winds and sandstorms that significantly contribute to the rising levels of both PM_{10} and $PM_{2.5}$ particles in the air [32]. Furthermore, the region, which is devoid of forests, is also characterized by very minimal average annual rainfall of less than 12cm. Compared to the tropical regions, the minimal annual rainfall in the desert regions results in the PM particles - especially the fine $PM_{2.5}$ particles - to remain suspended in the air for longer periods. Indeed, the study of Engelbrecht *et al.* [33] reported the presence of significant levels of Particulate Matters in the desert environment that are up to three or four times higher than the acceptable United States Army Center for Health Promotion and Preventive Medicine (USACHPPM) 1-Year Air-MEG value of $50\mu g/m^3$. There is therefore a need to carefully conduct new studies and monitor the concentration of the $PM_{2.5}$ particles using novel techniques.

The goal of the paper is to propose the use of a horizontal setup of polarimeters that use machine learning techniques in order to provide a more practical $PM_{2.5}$ estimation instrument than current solutions. Such a setup would allow wide area horizontal accurate measurements, that are not possible neither with satellites nor with in-situ measurement devices. The horizontal setup allows a wide spatial inclusion of the $PM_{2.5}$ measurements taken. With this vision, the paper provides evidence that it is in fact possible to achieve accuracies of up to 93% with such a system, through the use of machine learning based techniques. In addition, the paper explores various options for system implementation. Several machine learning techniques were tested and compared, and the paper shows that GPR model outperformed SVR and BET, and hence forms a good candidate for the proposed measurement system. All three machine learning models tested provided individual accuracies of more than 90%, and R^2 above 0.9, indicating that on average, the predicted values are close to the observed values, and that the predictor variables used

in the paper can precisely lead to a setup that can predict the $PM_{2.5}$ values accurately. The contribution of this work is two-fold:

- Firstly, the paper proposes a system that captures polarimetric images at near-ground level with a horizontal field-of-view aiming at standard targets to estimate $PM_{2.5}$ levels using polarimetric features such as DoLP and AoP. Such a system has the potential of providing accurate estimates of the levels of small $PM_{2.5}$ particles, as opposed to the satellite AOD/PM products that have reported higher accuracy for studying PM_{10} particles [34], [35].
- Secondly, the paper provides evidence that the accuracy of the proposed model can be further enhanced by employing a combination of both polarimetric and spectral features, rather than only polarimetric features. Specifically, it is shown that the use of red wavelength provides relatively better estimation in the study area probably because of the type of aerosols prevalent in the desert environment. In the future, this has to be investigated for different environments.

The rest of this paper is organized as follows: section II describes the proposed system; experimental results are discussed in section III, and conclusions are drawn in section IV.

II. PROPOSED SYSTEM

The proposed system aims to estimate the level of $PM_{2.5}$ concentration in the environment. The proposed system, as illustrated in Figure 1, is broadly divided into two major implementation processes: Data preparation and machine learning. These processes are presented in this section.

A. DATA PREPARATION

This process begins with the capture of polarization images. The polarization images were captured using the "4D PolarCam snapshot micro-polarizer camera", which is a Division-of-Focal-Plane (DoFP) polarization camera with a spatial resolution of 1780 × 1200. The setup for capturing data is illustrated in Figure 2.

1) IMAGE CAPTURE

The acquisition setup which is positioned (1m) above the ground level, involves a DoFP camera horizontally facing a white spectralon board, as seen in Figure 2. This setup is different from other reported setups in the literature where the instruments are either facing upwards from the ground-level [21], [23], [24], or space-borne facing towards the ground [18], [19]. The DoFP camera has the micro-polarizer (MP) array fabricated on top of the imaging sensor. This MP array is a periodic structure arranged in a 2 × 2 pattern to capture polarization information along four distinct directions (0°, 45°, 90°, and 135°). The proposed system takes full advantage of the micro-polarizer array structure to record the full polarization information of the reflected light in a single frame. In the proposed setup, a spectral filter is also positioned in front of the camera to be able to capture the spectral information in addition to the polarization information

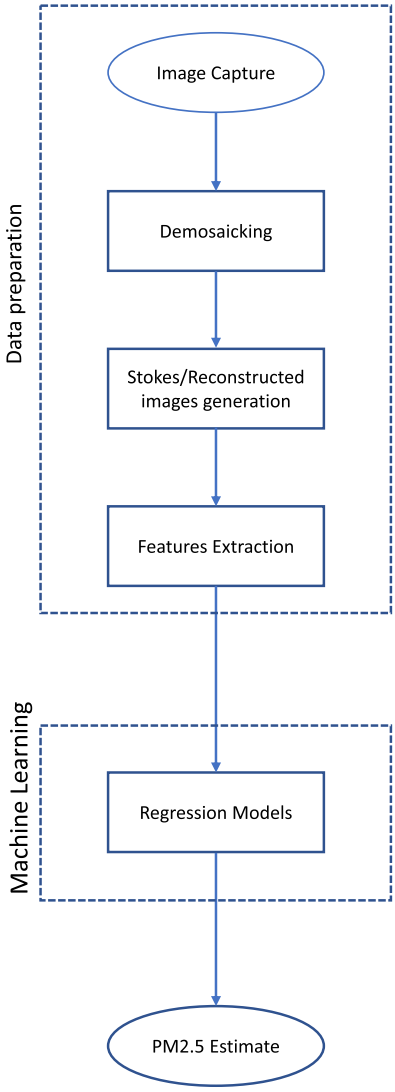

FIGURE 1. Flowchart of the proposed system.

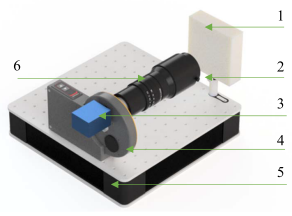

Image Acquisition Setup	
1	White Spectralon Calibration Target
2	Mini-Series Post Holder
3	4D PolarCam G5
4	6-Position Motorized Filter Wheels
5	Optical Breadboard
6	18 - 108 mm EFL, f/2.5 zoom lens

FIGURE 2. Image acquisition setup.

of any incoming light. This spectral filter is mounted on a motorized wheel, as shown in Figure 3, to enable for the capture of spectral properties at different wavelengths: red (620 -750nm), green (520 – 560nm), blue (450 – 490nm) and white (390 – 700nm) where no spectral filter was used (we refer to this case as clear). As light incidents on the white spectralon board, it is reflected and captured by the DoFP camera after passing through the spectral filter. The spectralon board has a very high diffuse reflectance value and, in most cases, assumed to be a Lambertian surface

FIGURE 3. A closer look at the spectral filters mounted on a motorized filter wheel.

with isotropic luminosity [36]. Therefore, the incident light is assumed to retain its property after reflecting from the board.

2) DEMOSAICKING

The image obtained using the DoFP camera is a mosaic image composite of four low-resolution sub-images (I_{0°, I_{45°, I_{90°, and I_{135°). These low-resolution sub-images are extracted, and their respective full-resolution images are generated using Interpolation algorithms. In the proposed system, the nearest neighbor interpolation algorithm [37] is used. This interpolation algorithm involves replacing a missing pixel with its nearest neighbor within a 3×3 block.

3) STOKES/RECONSTRUCTED IMAGES GENERATION

The full-resolution images generated by the demosaicking step will be used in determining the Stokes parameters needed to generate the reconstructed images that have more physical meanings [38]. Mathematically, the Stokes parameters are evaluated using the four full resolution subimages as follows [38]:

$$Intensity/S_0 = I_0 + I_{90} \tag{1}$$
$$S_1 = I_0 - I_{90} \tag{2}$$
$$S_2 = I_{45} - I_{135} \tag{3}$$
$$S_3 = I_{RCP} - I_{LCP} \tag{4}$$

In addition to natural light being typically linearly polarized, the absence of a retarder in the DoFP camera means only linearly polarized light is recorded. As a result, the S_3 term, which is the difference between the Right Circular Polarization (RCP) component and the Left Circular Polarization (LCP) component, is ignored. The other three parameters are dependent on intensity measurements and can therefore be easily computed from the full resolution images. With the determined stokes parameters, two useful images, DoLP and Angle of Polarization (AoP), can be constructed as follows:

$$DoLP = \sqrt{\frac{S_1^2 + S_2^2}{S_0^2}} \tag{5}$$
$$AoP = \frac{1}{2}\, atan\,(\frac{S_2}{S_1}) \tag{6}$$

4) FEATURES EXTRACTION

The input parameters (features) to the machine learning models are the average of each reconstructed image (DoLP and AoP) $DoLP_{avg}$, AoP_{avg}, in addition to S_{0avg}, which is the average of three pre-selected points (pixel (100,100), pixel (200,200) and pixel (300,300)) from the intensity image (S_0). The reason for using the average of the three pre-selected points from the intensity/S_0 image is to represent the brightness or the light intensity as a function of the time in a day, which represents the temporal information about the image within the day. The data spans multiple days over a period of 2 months, and therefore very well caters for the temporal effects.

B. MACHINE LEARNING

In order to model the relationship between the polarization images and the corresponding $PM_{2.5}$ measurements, machine learning based regression is implemented in two phases, namely the training phase and the testing phase.

Taken as an input and target pairs in the training phase, image feature vectors (DoLP, AoP of each filter and S_{0avg}) and the associated $PM_{2.5}$ measurements (taken from the training data), are fed to the machine learning block to model the function $f(\cdot)$, as illustrated in Figure 4. In the testing phase, the trained model $f(\cdot)$, takes the feature vectors of new images (taken from the testing data) as input to estimate the corresponding $PM_{2.5}$ concentrations. MAE between the estimated and measured $PM_{2.5}$ concentrations is calculated to judge the estimation accuracy generated model f(.). In this work, we implement three machine learning algorithms namely, Gaussian Process (GP) method, Support Vector Machines (SVM) and Bagging Ensemble Trees (BET). The three algorithms are used in the regression mode.

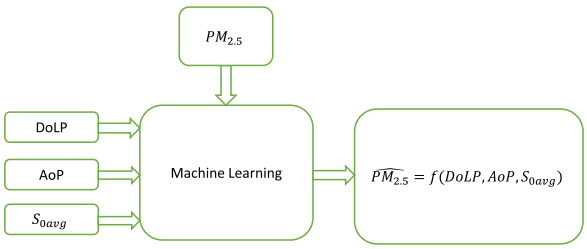

FIGURE 4. Training phase.

Support Vector Regression (SVR) is a supervised machine learning algorithm that uses a kernel function to map the problem in the input space to a higher dimensional space where regression problems that are highly nonlinear in the input space become linear in the higher dimensional space. Based on the structural risk minimization principle [39], it utilizes a risk function consisting of the empirical error and a regularization term and aims to minimize the risk based on Vapnik's e-insensitive loss metric. A detailed formulation of the SVR method can be found in [40].

Gaussian Process Regression (GPR) is a kernel-based supervised machine learning method. It is a non-parametric Bayesian approach [41] that assumes a prior probability

distribution of the input data. Using the training data, a posterior probability distribution is generated as an update for prior probability distribution. Although the posterior probability distribution is completely described by its covariance and mean value, the mean value is the one used for prediction [42], [43]. The key assumption in GP modelling is that our data can be represented as a sample from a multivariate Gaussian distribution [44] which means that a draw from the GP is a function and not a single value [45]. The mathematical formulation of GPR can be found in [41].

Bagging ensemble trees are improved form of decision trees. Decision trees, which are used for both classification and regression purposes are based on the idea of recursive partitioning [46]. They are considered as a computationally simple supervised machine learning methods [47]. Unfortunately, they can suffer from overfitting or under fitting leading to high variance or bias in their predictions [48]. Being applied on decision trees, ensemble methods such as boosting and bagging are used to account for the above mentioned problems [47]. While boosting aims to reduce bias, bagging, which is also known as bootstrap aggregation, results in reducing variance in predictions [48]. An improved and well-known form of bagging ensemble trees where input feature selection is implemented is the random forest algorithm [49]. Because of their ability, to limit prediction variability, ensemble methods including random forests have been widely used in literature for modeling and predicting environmental related phenomena [26], [47], [50]. Since in this work the number of features is low, we use the normal bagging ensemble trees rather than random forest. More information on decision trees and ensemble methods can be found in [46], [48].

III. EXPERIMENTAL RESULTS AND DISCUSSIONS

A. EXPERIMENTAL SETUP

To verify the usefulness of polarization imaging in estimating the concentration of $PM_{2.5}$ particles in the surrounding environment, four experiments - corresponding to the four spectral filters were conducted during the months of July and August 2020 in Ras Al Khaimah, UAE. During this period, DoFP polarization images were captured using the experimental setup illustrated in Figure 2. The actual $PM_{2.5}$ measurements were recorded using the ''Xiaomi Smartmi $PM_{2.5}$ Detector[1]'' and were compared to the measurements retrieved from the ''Air-quality[2]'' website at the same time stamp for further confirmation of the measurements accuracy. The total data accumulated encompasses 544 DoFP images (136 images per filter) in addition to 136 actual $PM_{2.5}$ measurements. A fifth experiment considered a combination of the four spectral features obtained from experiments 1-4. The experiments aimed to relate the captured DoFP polarization images under different wavelengths to the actual $PM_{2.5}$ measurements acquired during the same period. More specifically, experiment one, which employed no spectral filter (white or we refer to it here as clear filter), aimed to evaluate the efficiency of using the polarization properties to estimate the $PM_{2.5}$ concentrations in the surroundings. On the other hand, the objective of experiments 2-5 was to evaluate the efficiency of using the polarization properties under different wavelength to estimate the $PM_{2.5}$ concentrations.

In experiments 1, 2, 3, and 4, the respective filters used were clear, blue, green, and red filters. The features set used in each of these experiments included 3 parameters namely, S_{0avg}, DoLP and AoP for the corresponding filter used in the experiment. Experiment 5 however, used a 9-element feature set that comprises of the DoLP and AoP of each of the four filters together with S_{0avg}. The actual $PM_{2.5}$ measurements on the other hand, formed the training targets.

For each of the five experiments, 75% of the data was utilized to train and validate the system using 5-fold cross validation method. The remaining 25% was used test the performance of the system. Both sets of data included $PM_{2.5}$ measurements ranging from small, medium to high values. It is common practice in machine learning literature to use ''k-fold cross validation'' when the dataset is small to avoid over-fitting. In this work, the training set included 100 measurements on which 5-fold cross validation was applied. 5-fold cross validation partitions the data into 5 groups. It uses 4 groups to train and develop the model, and the fifth group to validate the trained model. This is repeated until each group serves as a validation group. The average of the five iterations is the reported model accuracy. Cross validation ensures no over-fitting occurs and helps in optimizing the machine learning model parameters that we applied to test the performance on the testing set (36 measurements).

The machine learning systems used were Gaussian Process Regression (GPR), Bagging Ensemble Tree (BET) and Support Vector Regression (SVR). Both GPR and SVR use RBF kernel function. Performance of the systems was verified by calculating the overall Mean Absolute Error (MAE), Root Mean Square Error (RMSE), estimation accuracy and the coefficient of determination (R^2). R^2 is a statistical measure that represents the proportion of the variance for a dependent variable (estimated $PM_{2.5}$) that's explained by the independent variable in a regression model. Both RMSE and R^2 are widely used to judge the quality of regression models. As is the case in related $PM_{2.5}$ estimation literature [28], [29], R^2 is used here to evaluate the correlation between the measured $PM_{2.5}$ values and the estimated $PM_{2.5}$ values from the machine learning model, as a function of polarimetric and spectral properties.

B. RESULTS AND DISCUSSIONS

The results of the five experiments when employing each of the three machine learning methods (GPR, BET, and SVR), are reported in Tables 1 - 3. The tables show the performance of each method using four metrics: the Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Estimation accuracy and the Coefficient of determination R^2. The estimation accuracy and the coefficient of determination (R^2) for the

[1] https://xiaomi-mi.com/air-and-water-purifiers/xiaomi-mi-pm25-detector-white/

[2] https://air-quality.com/

TABLE 1. GPR results.

Filter	MAE	RMSE	R^2	Accuracy %
Clear	0.8285	2.6670	0.8550	90.6092
Blue	1.0225	2.5820	0.8580	90.9085
Green	1.0762	2.5500	0.8680	91.0211
Red	0.6565	1.7430	0.9370	93.8627
All	1.0760	2.5500	0.8680	91.0211

TABLE 2. BET results.

Filter	MAE	RMSE	R^2	Accuracy %
Clear	2.2380	3.6520	0.7518	87.1409
Blue	3.3560	5.1430	0.5608	81.8909
Green	3.1510	4.9810	0.6602	82.4613
Red	1.8410	2.1910	0.9420	92.2852
All	3.1510	4.9810	0.6602	82.4613

three machine learning models are also depicted in figure 5 and figure 6, respectively.

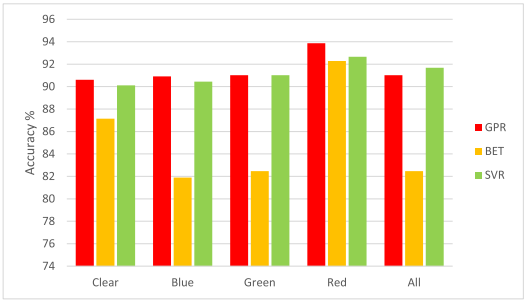

FIGURE 5. $PM_{2.5}$ Estimation accuracy (%) corresponding to each spectral filter for the three models.

Referring to Table 1 and the GPR bars in figure 5 and figure 6, it is evident that the red filter resulted in an estimation accuracy of 93.8627% and a RMSE of 1.743 in addition to having a reported R^2 of 0.9370. This leads to the inference that the system performed best for DoFP images within the red wavelength. Interestingly, when all the spectral features are added by combining all the filters, the resulting accuracy, RMSE and R^2 were respectively 91.0211%, 2.55 and 0.8680. These results clearly show a lower performance compared to the case of the red filter. This can be attributed to the fact that the increased number of features (9) in the combined filter case is relatively high with respect to the number of training vectors (100). It is well known in machine learning literature that such a scenario could lead to over-fitting in the training phase and less generalization ability in the testing phase. This in turn, results in a reduced accuracy when compared to the case of red filter.

Table 2 and the BET bars in figure 5 and figure 6 show the results of the five experiments when BET was used as the machine learning regression method. From the presented results, it is evident that the red filter resulted in an estimation accuracy of 92.2852%, a RMSE of 2.1910 and a reported R^2 of 0.9420 which once again leads to the conclusion that the system performed best for DoFP images in the red wavelength. Similar to the GPR case, it can be seen that, the combination of all the filters did not improve the system performance as the obtained accuracy was 82.4613% while the reported RMSE and R^2 were 4.9810 and 0.6602, respectively.

TABLE 3. SVR results.

Filter	MAE	RMSE	R^2	Accuracy %
Clear	1.3920	2.8060	0.8460	90.1197
Blue	1.4900	2.7140	0.8600	90.4437
Green	1.3770	2.5500	0.8680	91.0211
Red	1.0730	2.0860	0.9083	92.6549
All	1.2260	2.3630	0.8820	91.6796

The SVR results of the five experiments are presented in Table 3 and the green bars in figure 5 and figure 6. The use of SVR as the regression method resulted in an estimation accuracy of 92.6549%, a RMSE of 2.0860 and a reported R^2 of 0.9083 ranking the best among the 5 cases. The SVR results also concur with the GPR and BET results to indicate how the combination of all the filters did not rank highest. In the SVR, it ranked second at 91.6795% in terms of accuracy and last at 0.8820 in terms of R^2.

Referring to Tables 1 – 3, as well as figure 5 and figure 6, it can be seen that the use of the polarization properties alone (clear case) to estimate $PM_{2.5}$ concentrations proved to be successful resulting in an estimation accuracy ranging from 87.1408% in the case of BET to 90.6091% in the case of GPR. It resulted, as well, in R^2 ranging from 0.7518 in the case of BET to 0.8550 in the case of GPR. Also, the addition of spectral features did not necessarily improve the performance except for the case of the red filter. Including the polarization properties at the red wavelength considerably improved the estimation accuracy by a range of 2.5352% in the case of SVR to 5.1444% in the case of BET. It also improved R^2 by a range of 0.0623 in the case of SVR to 0.1902 in the case of BET.

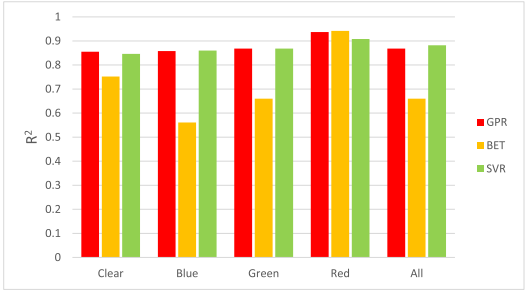

FIGURE 6. R^2 corresponding to each spectral filter for the three models.

Figure 7 and Figure 8 give a closer comparison among the three methods in terms of accuracy and R^2 when the red filter is used. It is clear from Figure 7 that the GPR model, at the red wavelength, outperformed the other two models by yielding the highest accuracy value of 93.8627%. On the other hand, figure 8 shows that that the BET model, at the red wavelength, outperformed the other two models by yielding the highest R^2 of 0.9420. Fortunately, the three models have individual accuracies above 91% and R^2 above 0.9 which indicate that on average, the predicted values are close to the observed values; and that the predictor variables (S_{0avg}, DoLP and AoP) can precisely predict the $PM_{2.5}$ values.

As noticed above, the use of a red filter before the DoFP camera resulted in the best estimation of $PM_{2.5}$ measurement, as compared to the study when any of the other individual filters, or their full combination, is used. It is believed that smaller aerosols contributing to $PM_{2.5}$ are characteristically different in a dusty region compared to other regions originating from a variable combination of natural and anthropogenic sources [32], [51]. The natural dust sources from the surrounding desert form a significant part of the $PM_{2.5}$ composition and contribute to higher reflectance in the Red wavelength band [52].

The proposed approach showed a stable performance. It was tested on 3 different machine learning methods. All of them resulted in the best estimation performance for $PM_{2.5}$ when utilizing the Polarimeteric properties in the red wavelengths range. This also agrees with literature that reported best performance in the red range in a desert environment [52]. Another indicator of the stability of the system is the high reported values of R^2 with low values of RMSE; and that R^2 and RMSE resulting from training and testing were comparable. We tried different values of K during "k-fold cross validation" and got comparable results. This shows that the system can generalize properly.

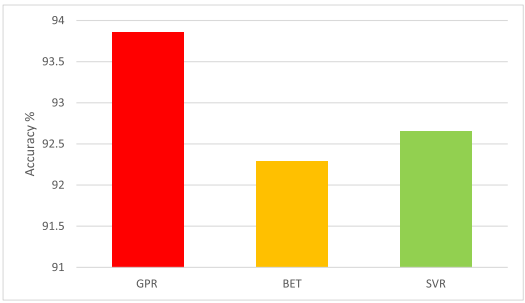

FIGURE 7. Red results for the accuracy (%) metric.

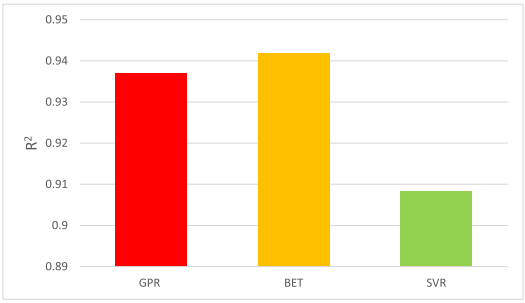

FIGURE 8. Red results for the R^2 metric.

The discussion above proves that the proposed system can estimate $PM_{2.5}$ near the ground level. To our knowledge, this is the first system to employ polarimetry and spectral characteristics to estimate $PM_{2.5}$ near the surface.

IV. CONCLUSION

In this work, we introduced a machine learning based system for the estimation of atmospheric particulate matter (PM) concentration, specifically, particles with a maximum diameter of 2.5μm. Unlike the other reported setups in the literature where the instruments are either facing upwards from the ground-level, or space-borne facing towards the ground, this system enabled the acquisition of the polarimetric images near the ground surface with a horizontal field of view aiming at standard targets which enables higher accuracy at the surface level. The proposed system uses a combination of features from both polarimetric and spectral imaging modalities developing three machine-learning models to estimate $PM_{2.5}$ concentrations in the surrounding environment. The experiment was conducted in Ras Al Khaimah, UAE during the months of July and August, where the weather tends to be hot, dusty, and humid. Evaluation of the proposed system showed high estimation accuracies up to 93.8627% and an R^2 score up to 0.9420 for the $PM_{2.5}$ concentrations. The highest estimation accuracy was reported for the red wavelength over all the used machine-learning approaches. While the current acquisition setup was at close distance to the reference point, there is a plan as part of future work to accommodate more spatial features by placing the reference point farther away, or by rotating the sensor and having multiple reference points.

REFERENCES

[1] M. Takruri, A. Abubakar, N. Alnaqbi, H. A. Shehhi, A.-H.-M. Jallad, and A. Bermak, "DoFP-ML: A machine learning approach to food quality monitoring using a DoFP polarization image sensor," *IEEE Access*, vol. 8, pp. 150282–150290, 2020.

[2] Z. Guan, F. Goudail, M. Yu, X. Li, Q. Han, Z. Cheng, H. Hu, and T. Liu, "Contrast optimization in broadband passive polarimetric imaging based on color camera," *Opt. Exp.*, vol. 27, no. 3, pp. 2444–2454, Feb. 2019. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-27-3-2444

[3] M. Wan, G. Gu, W. Qian, K. Ren, and Q. Chen, "Stokes-vector-based polarimetric imaging system for adaptive target/background contrast enhancement," *Appl. Opt.*, vol. 55, no. 21, pp. 5513–5519, Jul. 2016. [Online]. Available: http://ao.osa.org/abstract.cfm?URI=ao-55-21-5513

[4] O. Dubovik *et al.*, "Polarimetric remote sensing of atmospheric aerosols: Instruments, methodologies, results, and perspectives," *J. Quant. Spectrosc. Radiat. Transf.*, vol. 224, pp. 474–511, Feb. 2019.

[5] M. I. Mishchenko and L. D. Travis, "Satellite retrieval of aerosol properties over the ocean using polarization as well as intensity of reflected sunlight," *J. Geophys. Res., Atmos.*, vol. 102, no. D14, pp. 16989–17013, Jul. 1997. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/96JD02425

[6] M. I. Mishchenko and L. D. Travis, "Satellite retrieval of aerosol properties over the ocean using measurements of reflected sunlight: Effect of instrumental errors and aerosol absorption," *J. Geophys. Res., Atmos.*, vol. 102, no. D12, pp. 13543–13553, Jun. 1997.

[7] M. I. Mishchenko, L. D. Travis, W. B. Rossow, B. Cairns, B. E. Carlson, and Q. Han, "Retrieving CCN column density from single-channel measurements of reflected sunlight over the ocean: A sensitivity study," *Geophys. Res. Lett.*, vol. 24, no. 21, pp. 2655–2658, Nov. 1997.

[8] O. P. Hasekamp, "Linearization of vector radiative transfer with respect to aerosol properties and its use in satellite remote sensing," *J. Geophys. Res.*, vol. 110, no. D4, pp. 1–18, 2005. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2004JD005260

[9] O. P. Hasekamp and J. Landgraf, "Retrieval of aerosol properties over land surfaces: Capabilities of multiple-viewing-angle intensity and polarization measurements," *Appl. Opt.*, vol. 46, no. 16, pp. 3332–3344, 2007.

[10] J. H. Seinfeld *et al.*, "Improving our fundamental understanding of the role of aerosol-cloud interactions in the climate system," *Proc. Nat. Acad. Sci. USA*, vol. 113, no. 21, pp. 5781–5790, May 2016. [Online]. Available: https://www.pnas.org/content/113/21/5781

[11] J. H. Seinfeld and J. F. Pankow, "Organic atmospheric particulate material," *Annu. Rev. Phys. Chem.*, vol. 54, no. 1, pp. 121–140, Oct. 2003, doi: 10.1146/annurev.physchem.54.011002.103756.

[12] K. H. Kim, E. Kabir, and S. Kabir, "A review on the human health impact of airborne particulate matter," *Environ. Int.*, vol. 74, pp. 136–143, Jan. 2015.

[13] O. Raaschou-Nielsen *et al.*, "Air pollution and lung cancer incidence in 17 European cohorts: Prospective analyses from the European study of cohorts for air pollution effects (ESCAPE)," *Lancet Oncol.*, vol. 14, no. 9, pp. 813–822, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1470204513702791
[14] R. Atkinson, S. Kang, H. Anderson, I. Mills, and H. Walton, "Epidemiological time series studies of $PM_{2.5}$ and daily mortality and hospital admissions: A systematic review and meta-analysis," *Thorax*, vol. 69, no. 7, pp. 660–665, 2014.
[15] R. Beelen *et al.*, "Effects of long-term exposure to air pollution on natural-cause mortality: An analysis of 22 European cohorts within the multicentre ESCAPE project," *Lancet*, vol. 383, no. 9919, pp. 785–795, 2014.
[16] M. Stafoggia, E. Samoli, E. Alessandrini, E. Cadum, B. Ostro, G. Berti, A. Faustini, B. Jacquemin, C. Linares, M. Pascal, G. Randi, A. Ranzi, E. Stivanello, and F. Forastiere, "Short-term associations between fine and coarse particulate matter and hospitalizations in southern Europe: Results from the MED-PARTICLES project," *Environ. Health Perspect.*, vol. 121, no. 9, pp. 1026–1033, Sep. 2013.
[17] M. I. Mishchenko, B. Cairns, J. E. Hansen, L. D. Travis, R. Burg, Y. J. Kaufman, J. V. Martins, and E. P. Shettle, "Monitoring of aerosol forcing of climate from space: Analysis of measurement requirements," *J. Quant. Spectrosc. Radiat. Transf.*, vol. 88, nos. 1–3, pp. 149–161, Sep. 2004.
[18] P.-Y. Deschamps, F.-M. Bréon, M. Leroy, A. Podaire, A. Bricaud, J.-C. Buriez, and G. Seze, "The POLDER mission: Instrument characteristics and scientific objectives," *IEEE Trans. Geosci. Remote Sens.*, vol. 32, no. 3, pp. 598–615, May 1994.
[19] D. Tanré, F. Bréon, J. Deuzé, O. Dubovik, F. Ducos, P. François, P. Goloub, M. Herman, A. Lifermann, and F. Waquet, "Remote sensing of aerosols by using polarized, directional and spectral measurements within the A-train: The PARASOL mission," *Atmos. Meas. Tech.*, vol. 4, no. 7, pp. 1383–1395, Apr. 2011.
[20] R. Liao, W. Guo, N. Zeng, J. Guo, Y. He, H. Di, D. Hua, and H. Ma, "Polarization measurements and evaluation based on multidimensional polarization indices applied in analyzing atmospheric particulates," *Appl. Sci.*, vol. 11, no. 13, p. 5992, Jun. 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/13/5992
[21] B. N. Holben, T. F. Eck, I. Slutsker, D. Tanre, J. P. Buis, A. Setzer, E. Vermote, J. A. Reagan, Y. J. Kaufman, T. Nakajima, and F. Lavenu, "AERONET—A federated instrument network and data archive for aerosol characterization," *Remote Sens. Environ.*, vol. 66, no. 1, pp. 1–16, 1998.
[22] C. Cheng, L. Zhengqiang, L. Donghui, L. Kaitao, Z. Ying, H. Weizhen, and X. Yisong, "Ground-based polarization remote sensing of atmospheric aerosols and the correlation between polarization degree and $PM_{2.5}$," *IOP Conf. Ser., Earth Environ. Sci.*, vol. 17, Mar. 2014, Art. no. 012039, doi: 10.1088/1755-1315/17/1/012039.
[23] G. van Harten, J. de Boer, J. H. H. Rietjens, A. Di Noia, F. Snik, H. Volten, J. M. Smit, O. P. Hasekamp, J. S. Henzing, and C. U. Keller, "Atmospheric aerosol characterization with a ground-based SPEX spectropolarimetric instrument," *Atmos. Meas. Techn.*, vol. 7, no. 12, pp. 4341–4351, Dec. 2014.
[24] D. J. Diner, F. Xu, J. V. Martonchik, B. E. Rheingans, S. Geier, V. M. Jovanovic, A. Davis, R. A. Chipman, and S. C. McClain, "Exploration of a polarized surface bidirectional reflectance model using the ground-based multiangle spectropolarimetric imager," *Atmosphere*, vol. 3, no. 4, pp. 591–619, Dec. 2012.
[25] O. Dubovik, A. Smirnov, B. N. Holben, M. D. King, Y. J. Kaufman, T. F. Eck, and I. Slutsker, "Accuracy assessments of aerosol optical properties retrieved from aerosol robotic network (AERONET) sun and sky radiance measurements," *J. Geophys. Res., Atmos.*, vol. 105, no. D8, pp. 9791–9806, Apr. 2000. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2000JD900040
[26] G. Chen, S. Li, L. D. Knibbs, N. A. S. Hamm, W. Cao, T. Li, J. Guo, H. Ren, M. J. Abramson, and Y. Guo, "A machine learning method to estimate $PM_{2.5}$ concentrations across China with remote sensing, meteorological and land use information," *Sci. Total Environ.*, vol. 636, pp. 52–60, Sep. 2018.
[27] X. Hu, J. H. Belle, X. Meng, A. Wildani, L. A. Waller, M. J. Strickland, and Y. Liu, "Estimating $PM_{2.5}$ concentrations in the conterminous United States using the random forest approach," *Environ. Sci. Technol.*, vol. 51, no. 12, pp. 6936–6944, 2017.
[28] T. Li, H. Shen, Q. Yuan, and L. Zhang, "A locally weighted neural network constrained by global training for remote sensing estimation of $PM_{2.5}$," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.
[29] X. Yan, Z. Zang, N. Luo, Y. Jiang, and Z. Li, "New interpretable deep learning model to monitor real-time $PM_{2.5}$ concentrations from satellite data," *Environ. Int.*, vol. 144, Nov. 2020, Art. no. 106060. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0160412020320158
[30] A. A. Kokhanovsky *et al.*, "The inter-comparison of major satellite aerosol retrieval algorithms using simulated intensity and polarization characteristics of reflected light," *Atmos. Meas. Techn.*, vol. 3, no. 4, pp. 909–932, 2010.
[31] K. Knobelspiesse, B. Cairns, M. Mishchenko, J. Chowdhary, K. Tsigaridis, B. van Diedenhoven, W. Martin, M. Ottaviani, and M. Alexandrov, "Analysis of fine-mode aerosol retrieval capabilities by different passive remote sensing instrument designs," *Opt. Exp.*, vol. 20, no. 19, pp. 21457–21484, 2012.
[32] N. M. Hamdan, H. Alawadhi, and N. Jisrawi, "Particulate matter pollution in the United Arab Emirates: Elemental analysis and phase identification of fine particulate pollutants," in *Proc. 2nd World Congr. New Technol.*, 2016, pp. 1–9.
[33] J. P. Engelbrecht, E. V. McDonald, J. A. Gillies, and A. W. Gertler, "Department of defense enhanced particulate matter surveillance program (EPMSP)," Desert Res. Inst., Reno, NV, USA, Tech. Rep. W9124R-05-C-0135/SUBCLIN000101-ACRN-AB, 2008.
[34] Y. Li, S. Yuan, S. Fan, Y. Song, Z. Wang, Z. Yu, Q. Yu, and Y. Liu, "Satellite remote sensing for estimating $PM_{2.5}$ and its components," *Current Pollut. Rep.*, vol. 7, pp. 1–16, Jan. 2021.
[35] J. P. Kerekes, M. M. Patel, and C. C. D'Angelo, "Review of global near real time $PM_{2.5}$ estimates and model forecasts," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Sep. 2020, pp. 5570–5573.
[36] D. Doxaran, N. C. Cherukuru, S. J. Lavender, and G. F. Moore, "Use of a Spectralon panel to measure the downwelling irradiance signal: Case studies and recommendations," *Appl. Opt.*, vol. 43, no. 32, pp. 5981–5986, 2004.
[37] A. Ahmed, X. Zhao, and A. Bermak, "Performance evaluation of interpolation algorithms for division of focal plane polarization image sensors," in *Proc. 4th Int. Conf. Inf. Sci. Control Eng. (ICISCE)*, Jul. 2017, pp. 1587–1590.
[38] D. Goldstein and E. Collett, *Polarized Light*, 2nd ed. New York, NY, USA: Marcel Decker, 2003.
[39] L. J. Cao and F. E. H. Tay, "Support vector machine with adaptive parameters in financial time series forecasting," *IEEE Trans. Neural Netw.*, vol. 14, no. 6, pp. 1506–1518, Nov. 2003.
[40] A. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statist. Comput.*, vol. 14, no. 3, pp. 199–222, 2014.
[41] E. Schulz, M. Speekenbrik, and A. Krause, "A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions," *J. Math. Psychol.*, vol. 85, pp. 1–16, Aug. 2018.
[42] C. Hultquist, G. Chen, and K. Zhao, "A comparison of Gaussian process regression, random forests and support vector regression for burn severity assessment in diseased forests," *Remote Sens. Lett.*, vol. 5, no. 8, pp. 723–732, Aug. 2014.
[43] C. E. Rasmussen, "Gaussian processes in machine learning," in *Summer School on Machine Learning*. Berlin, Germany: Springer, 2003, pp. 63–71.
[44] M. M. Sawant and K. Bhurchandi, "Hierarchical facial age estimation using Gaussian process regression," *IEEE Access*, vol. 7, pp. 9142–9152, 2019.
[45] M. Ebden, "Gaussian processes: A quick introduction," 2015, *arXiv:1505.02965*.
[46] A. J. Izenman, *Recursive Partitioning and Tree-Based Methods*. New York, NY, USA: Springer, 2008, pp. 281–314.
[47] M. A. Hassan, A. Khalil, S. Kaseb, and M. A. Kassem, "Exploring the potential of tree-based ensemble methods in solar radiation modeling," *Appl. Energy*, vol. 203, pp. 897–916, Oct. 2017.
[48] A. J. Izenman, *Committee Machines*. New York, NY, USA: Springer, 2008, pp. 505–550.
[49] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[50] J. Fan, W. Yue, L. Wu, F. Zhang, H. Cai, X. Wang, X. Lu, and Y. Xiang, "Evaluation of SVM, ELM and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of China," *Agricult. Forest Meteorol.*, vol. 263, no. 1, pp. 225–241, Dec. 2018.
[51] K. W. Brown, W. Bouhamra, D. P. Lamoureux, J. S. Evans, and P. Koutrakis, "Characterization of particulate matter for three sites in Kuwait," *J. Air Waste Manage. Assoc.*, vol. 58, no. 8, pp. 994–1003, Aug. 2008.
[52] A. Sadiq and F. Howari, "Remote sensing and spectral characteristics of desert sand from Qatar Peninsula, Arabian/Persian Gulf," *Remote Sens.*, vol. 1, no. 4, pp. 915–933, Nov. 2009.