An Investigation Into Machine Learning Regression Techniques for the Leaf Rust Disease Detection Using Hyperspectral Measurement

Davoud Ashourloo, Hossein Aghighi, Ali Akbar Matkan, Mohammad Reza Mobasheri, and Amir Moeini Rad

***Abstract*—The complex impacts of disease stages and disease symptoms on spectral characteristics of the plants lead to limitation in disease severity detection using the spectral vegetation indices (SVIs). Although machine learning techniques have been utilized for vegetation parameters estimation and disease detection, the effects of disease symptoms on their performances have been less considered. Hence, this paper investigated on 1) using partial least square regression (PLSR), ν support vector regression (ν-SVR), and Gaussian process regression (GPR) methods for wheat leaf rust disease detection, 2) evaluating the impact of training sample size on the results, 3) the influence of disease symptoms effects on the predictions performances of the above-mentioned methods, and 4) comparisons between the performances of SVIs and machine learning techniques. In this study, the spectra of the infected and non infected leaves in different disease symptoms were measured using a non imaging spectroradiometer in the electromagnetic region of 350 to 2500 nm. In order to produce a ground truth dataset, we employed photos of a digital camera to compute the disease severity and disease symptoms fractions. Then, different sample sizes of collected datasets were utilized to train each method. PLSR showed coefficient of determination (R^2) values of 0.98 (root mean square error (RMSE) = 0.6) and 0.92 (RMSE = 0.11) at leaf and canopy, respectively. SVR showed R^2 and RMSE close to PLSR at leaf (R^2 = 0.98, RMSE = 0.05) and canopy (R^2 = 0.95, RMSE = 0.12) scales. GPR showed R^2 values of 0.98 (RMSE = 0.03) and 0.97 (RMSE = 0.11) at leaf and canopy scale, respectively. Moreover, GPR represents better performances than others using small training sample size. The results represent that the machine learning techniques in contrast to SVIs are not sensitive to different disease symptoms and their results are reliable.**

***Index Terms*—Disease severity, disease symptoms, Gaussian process regression (GPR), hyperspectral measurement, PLSR, plant disease, ν support vector regression (ν-SVR), wheat leaf rust (WLR).**

I. Introduction

SPECTRAL data at the leaf and canopy scales have been utilized to improve the plant diseases detection techniques from remotely sensed observations [1], [2], where the visible and infrared regions are more sensitive to disease development [3]. The measured spectra can be utilized to early detection of fungus disease [1] and to forecast wheat powdery mildew disease [4]. Moreover, the optimized narrow bands vegetation indices were employed to discriminate various disease of wheat [5].

The wheat rust is an important crop disease which has three types, i.e., wheat yellow rust (WYR), wheat leaf rust (WLR), and wheat stem rust [6]. WYR disease is identified by a single symptom which occurs as a narrow yellow stripes parallel to nervures on the leaf, whereas WLR disease is caused by the *Puccinia triticina* fungus and illustrates numerous symptoms simultaneously in various parts of an infected leaf [7]. The WLR symptoms vary from leaf to leaf but its presents a yellow color earlier, then its changes to orange and dark brown. Finally, the disease symptom ends with the dry leaf [6].

During recent decades, a few numbers of researchers have been working on WYR disease detection using different approaches [8]–[10]. For instance, Hansen [8] employed neural network technique to detect disease severity levels of WYL disease at canopy scale. In another work, Huang *et al.*[9] demonstrated the potential of spectral vegetation indices (SVIs) to accurately detect WYL disease using *in situ* hyperspectral reflectance and airborne hyperspectal imaging. Moreover, the method described in [10] shows the potential of *in situ* hyperspectral reflectance in discrimination of WLR disease from nutrient stress. On the other hand, less attention has been paid on WLR disease detection. Franke and Menz [11] compared the performances of using the hyperspectral and the multispectral data for early WLR disease detection at the canopy scale, and they reported higher accuracy for the hyperspectaral data. Then Ashourloo *et al.*[12] employed narrow band SVIs to estimate the disease severity levels of WLR at the leaf scale. They reported that the accuracy of the SVIs declines as disease severity increases [12]; which could be because of the adverse influences of the different disease symptoms [13], [14]. One of the limitations of SVIs for disease detection is that they only used few number of spectral bands, however, other techniques such as machine learning regression methods can employ full spectrum simultaneously [15]. For instant, partial least square regression (PLSR) showed higher performance than vegetation indices in water content estimation [16]. In other studies, Verrelst *et al.*[17] implemented SVR to accurately retrieve biophysical parameters, and Verrelst *et al.*[18] utilized GPR to estimate chlorophyll content of vegetation. In stress detection, support vector machine (SVM) showed accurate results in early stress classification of

Manuscript received November 19, 2015; revised April 20, 2016; accepted May 26, 2016. Date of publication August 3, 2016; date of current version September 30, 2016. *(Corresponding author: Davoud Ashourloo.)*

D. Ashourloo, H. Aghigh, and A. A. Matkan are with the Remote Sensing and GIS Center, Faculty of Earth Sciences, Shahid Beheshti University, Tehran 653641255, Iran (e-mail: d_ashourloo@sbu.ac.ir; h_aghighi@sbu.ac.ir; a-matkan@sbu.ac.ir).

M. R. Mobasheri and A. M. Rad are with the Remote Sensing Group, Faculty of Geodesy and Geomatics Engineering, K.N. Toosi University of Technology, Tehran 19697, Iran (e-mail: mobasheri@kntu.ac.ir; AmirMoeiniRad@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTARS.2016.2575360

drought using hyperspectral measurement [19]. In this research area, Rumpf *et al.* [20] used SVM to early stress detection of wheat using narrow SVIs. PLSR [21], support vector regression (SVR) and Gaussian process regression (GPR) techniques were employed to estimate the vegetation parameters (VP) [17], [18]. However, few numbers of these approaches have been used for disease detection [19], [20].

This paper is investigated on WLR disease detection using PLSR, SVR, and GPR, as well as evaluation of disease symptoms effects on their predictions. Furthermore, as obtaining training samples is often expensive and also the training sample size is an important factor in the machine learning regression methods performance [22], the effect of training sample size on regression accuracies is also considered. Finally, the performances of the machine learning regression methods and SVIs are compared.

This manuscript is organized as follows. In Section I, a review on disease detection methods, and the objectives of this manuscript are presented. In the next section, experimental setup and data collection are explained. In Section III, PLSR, SVR, and GPR implementation are presented. In Section IV, results of the research are illustrated in detail and the effects of different disease symptoms on the accuracy of the machine learning regression method are discussed. Finally, the conclusion is given in Section V.

II. Experimental Setup

To conduct this work, the spectral dataset from healthy and inoculated wheat leaves was collected at the leaf and canopy scales. Moreover, a series of RGB digital photos from all infected leaves were collected. These datasets were utilized to study the potential of GPR, SVR, and PLSR in disease detection. Each of these methods is trained with 25, 50, 75 100, 125, and 150 samples, at the leaf and canopy scales and their corresponding test samples were used to evaluate the performance of each approach. The cultivation condition and pathogen inoculation were discussed in details in our previous works [12], [13], [23].

A. Data Collection

In order to measure the spectral reflectance of the infected leaves, an ASD spectroradiometer (analytical spectral device, Boulder, CO, USA) with a constant 25° field of view (FOV) was utilized. The spectra were collected in the 350–2500 nm range with a band width of 1–4 nm and the sensor FOV of 25°. However, due to water absorption effects on the wavelengths beyond 1000 nm, we limited our analyses from 450 to 1000 nm. Details about data collection at the leaf scale can be found in [12], [13].

At the canopy scale, each spectrum was recorded with a height of 150 cm above the ground, and footprint of 3472 cm^2 (footprint $=\pi r^2$ and $r=$ height $\times$ tan$(\frac{\text{FOV}}{2})$). A total of 175 samples were collected from the early symptoms of the disease until a full infection symptoms. Then, a moving average filter with the length of 17 data points [24] and a second-order polynomial was applied along the signals to smooth the spectra [16].

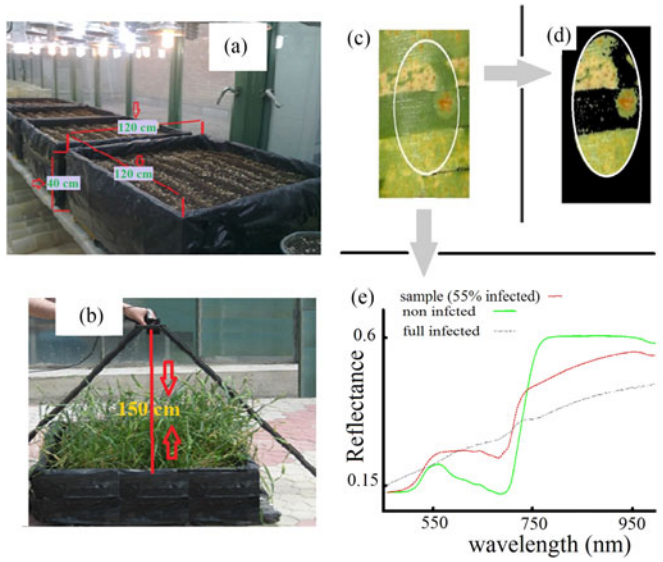

Fig. 1. (a) Illustrates the boxes used to plant wheat. (b) Using ASD spectroradiometer to measure the infected leaf reflectance. (c) Four leaves were placed next to each other and marked by the position of Contact probe (white circle). (d) The infected leaves which separated using green mask.(e) The spectra of the infected and noninfected leaves.

After measuring the spectrum of each sample, an RGB digital photo was taken from 30 and 150 cm height above the samples for the leaf and canopy scales, respectively. These photos were taken by a Canon DIGITAL IXUS 85 IS; F-109 number: $\frac{\text{f}}{3.2}$; Shutter speed: $\frac{1}{60}$. The taken photos were transformed from RGB (red, green, and blue) to HSI (hue, saturation, and intensity) system to remove the effect of outdoor light intensity variations which was discussed in [25]. At the canopy scale, 30 leaves in each photo were randomly selected [26] and the disease severity and disease symptoms were computed for each leaf. In order to extract the disease symptoms fraction accurately in selected leaves, we have used an appropriate digital zoom and texture algorithm [12], [13]. Finally, the mean of the disease symptoms and disease severity of the selected leaves was computed and employed as the ground truth of the disease symptoms and disease severity at canopy scale.

The boxes and spectral measurement at the canopy scale are shown in Fig. 1(a) and (b), respectively. The separated area using a white circle in Fig. 1(c) illustrates the position of Contact probe which was used for spectroradiometry. Then a green mask [25] was applied on Fig. 1(c) to separate the infected portion of the leaves [see Fig. 1(d)] which shows 55% disease severity. The spectrum of these infected leaves (red color) in comparison with spectra of the full infected (gray color) and noninfected leaves (green color) are shown in Fig. 1(e).

III. Multivariate Methods

A. Partial Least Square Regression

PLSR is one of the statistical techniques that has been used to estimate the VP [27]. It is defined as

$$\boldsymbol{X} = \boldsymbol{T} \times \boldsymbol{P}^{\mathrm{T}} + \boldsymbol{E} \tag{1}$$

$$\boldsymbol{y} = \boldsymbol{T} \times \boldsymbol{q} + \boldsymbol{f} \tag{2}$$

where $\boldsymbol{T}$, $\boldsymbol{P}$, and T denote the score matrix, loading matrix for $\boldsymbol{X}$ (e.g., spectral dataset) and transpose, respectively. In (2), $\boldsymbol{y}$ is the response vector (e.g., disease severity) and $\boldsymbol{q}$ is the score vector corresponding to $\boldsymbol{y}$. $\boldsymbol{E}$ and $\boldsymbol{f}$ are residuals for $\boldsymbol{X}$ and $\boldsymbol{y}$, respectively.

PLSR was designed to analyze data with numerous strongly collinear, noisy, and independent variables [28]. In order to explain most of the variations in both predictors and responses, PLSR finds a few components. Factors that explain response variations properly provide good predictive models for new responses and factors that explain predictor variations are well represented by the observed values [21]. In this study, PLSR was used to model the correlation between leaves' recorded spectra and WLR disease. Since the number of components is an important factor in the result of PLSR, visual interpretation and RMSE variations were used to select them. At the leaf and canopy scales, 8th and 6th factors were selected, respectively. Because the first eight components explain 99.8% of the data at the leaf scale and the first six components at the canopy scale explain 99.7% of the data. In comparison between autoscale and mean center methods for data scaling, the autoscale performs better at canopy scale data and the mean center at leaf scales data. PLSR was carried out by a code which has been developed in Matlab by [29].

B. Support Vector Regression

SVR is a regression technique which has been used to estimate the biochemical and the biophysical parameters of plant [18]. SVR estimates the nonlinear relationship between input vectors (e.g., spectral signature) and response variables (e.g., disease severity) by fitting an optimal hyperplane [30]. The most commonly used versions of SVR are "ε-SVR" and "ν-SVR," where ε and ν are two different regularization parameters in SVR to apply a penalty to optimize points which are not accurately anticipated [31]. Here ν is changed between $(0, 1]$ and lets one to control the number of support vectors and training errors [31]. It was proved that ν is an upper bound on the fraction of the badly predicted errors and a lower bound of support vectors [31]; thus, ν-SVR has a more meaningful interpretation than ε-SVR [32]. For this reason, this work only used ν-SVR using a radial basis function kernel to approximate the nonlinear regression function, as (3) [32]. The RBF was chosen, because of less implementation difficulties, faster training rate, and delivering an acceptable accuracy [33], [34]

$$f(x) = w^{\mathrm{T}} \times \Phi(x) + b, \tag{3}$$

Here, w^{T} is the weight vector to the corresponding nonlinear mapping function $\Phi(x)$, and b is a bias. Both parameters w and b should be estimated and $f(x)$ desired to be as close as possible to the output y. In order to solve this problem, the following minimization should be solved [31]:

$$\frac{1}{2} w w^{\mathrm{T}} + C \left\{ \nu\epsilon + \frac{1}{N} \sum_{i=1}^{N} (\xi_i + \xi_i^*) \right\}$$

TABLE I
ACCURACY OF MACHINE LEARNING REGRESSION METHODS (PLSR, GPR, AND ν-SVR) IN DISEASE DETECTION FOR DIFFERENT SAMPLE SIZE AT LEAF AND CANOPY SCALES

Scale	Method		25	50	75	100	125	150
Leaf	PLSR	R^2	87	95	96	97	98	98
		RMSE	0.13	0.12	0.1	0.08	0.08	0.06
	ν-SVR	R^2	95	97	98	98	98	98
		RMSE	0.12	0.11	0.1	0.08	0.07	0.05
	GPR	R^2	97	98	98	98	98	98
		RMSE	0.05	0.04	0.04	0.03	0.03	0.03
Canopy	PLSR	R^2	78	82	82	86	90	91
		RMSE	0.21	0.18	0.16	0.16	0.13	0.12
	ν-SVR	R^2	85	88	91	94	94	95
		RMSE	0.17	0.16	0.14	0.14	0.13	0.12
	GPR	R^2	91	92	94	93	95	97
		RMSE	0.14	0.13	0.12	0.12	0.12	0.11

$$\begin{aligned} &\left(w^{\mathrm{T}} \Phi(x_i) + b\right) - y_i \leq +\xi_i \\ &y_i - \left(w^{\mathrm{T}} \Phi(x_i) + b\right) \leq +\xi_i^* \\ &\xi_i, \xi_i^* \geq 0, i = 1, \cdots, N, \geq 0. \end{aligned} \tag{4}$$

Here, N is the training samples size, C is the regularization parameter, and ξ and ξ_i^* denote slack variables representing upper and lower constraint on the outputs of the system. More detailed information about ν-SVR and also its comparision with ε are provided by [31].

In this work, we employed LibSVM to predict disease severity using ν-SVR [35]. In order to employ this model, we employed ten-fold cross validation to estimate three parameters using training data [22], including C, γ, and ν. The ε value was not adjusted and set to 0.1 [22]. Further details regarding to ν-SVR working are available at [36] and [37]. The accuracy of the trained model using these model parameters are then evaluated using the testing datasets, and their RMSE are presented in Table I.

C. Gaussian Process Regression

GPR is a famous kernel-based machine learning approach for nonlinear regression problems that was formulated and interpreted in a Bayesian context [38], [39]. In order to learn and generate the final function model also known as posterior GPR, it utilizes a training dataset (B-spectral data, $\boldsymbol{x} \in \mathbb{R}^B$) and output (disease severity, $\boldsymbol{y} \in \mathbb{R}$) to elicit a prior GPR to constrain the possible forms of the unknown function. Then, GPR updates the prior using the training dataset to generate a posterior GPR. This procedure requires to tune some parameters related to the covariance or kernel functions known as hyperparameters, i.e., magnitute, characteristic length, and noise variance [18]. In this research, we employed scaled Gaussian function [40]:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \upsilon \exp\left(- \sum_{b=1}^{B} \frac{\left(\boldsymbol{x}_i^{(b)} - \boldsymbol{x}_j^{(b)}\right)^2}{2\sigma_b^2} \right) \tag{5}$$

where υ is the scaling factor and σ_b is the dedicated parameter to control the spread of the relations for each spectral band (b).

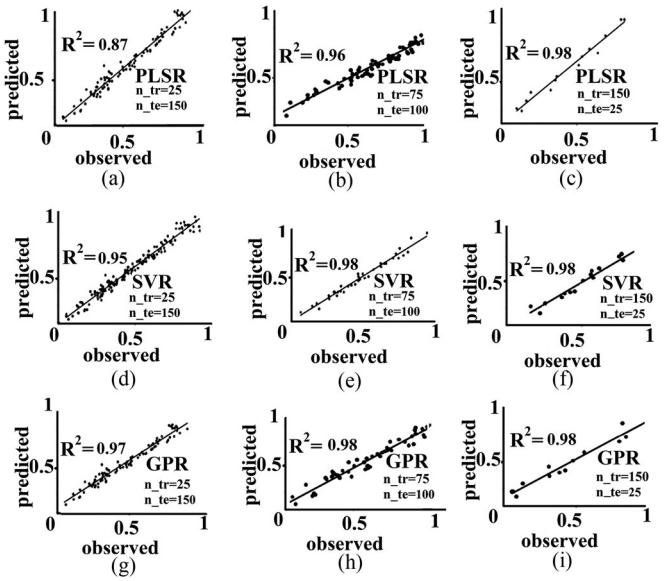

Fig. 2. Relationship between the observed and the estimated disease severity of PLSR (a, b, and c), ν-SVR (d, e, and f) and GPR (g, h, and i) for 25, 75, and 125 training sample size ($\mathbf{n\_tr}$) at leaf scale. $\mathbf{n\_ts}$ is the number of testing sample size.

In order to employ GP as a practical tool in an application, one needs to determine the hyperparameters of the covariance function as training of a Gaussian process [41]. However, Bu and Pan [42] demonstrated that varying the initial hyperparameters does not affect the final optimal results of the GPR. In this study, the squared exponential is determined as a likelihood function, where its hyperparameters can be parameterized in terms of length-scale (ℓ), the signal variance (σ_f), and the noise variance (σ_n). The required hyperparameters are defined as (ℓ, σ_f, σ_n) =(1,1,0.1) [41].

A free implemented software of GPR is freely available at http://www.gaussianprocess.org/gpml/.

D. Comparison of the Machine Learning Regression Methods With SVIs

In our previous work [12], we utilized 22 indices to study WLR detection, where NBNDVI (narrow band normal difference vegetation index) and PRI (photochemical reflectance index) showed the highest accuracy in disease detection. Therefore, this work evaluates the disease symptoms effects on PLSR, ν-SVR, and GPR accuracies after 30 iterations in comparison with NBNDVI (6) and PRI (7)

$$\text{NBNDVI} = \left(\rho_{850} - \rho_{680}\right) / \left(\rho_{850} + \rho_{680}\right) \tag{6}$$

$$\text{PRI} = \left(\rho_{750} - \rho_{531}\right) / \left(\rho_{750} + \rho_{531}\right). \tag{7}$$

Here, ρ_b is the reflectance value at wavelength b (nm).

E. Validation

The root mean square error (RMSE) (8) and coefficient of determination (R^2) (9) are prevalent statistical parameters for evaluating the performance of the regression methods. Moreover, in this work, R^2 is used to evaluate the effect of different disease symptoms on the accuracies of the implemented methods at ten disease severity levels

$$\text{RMSE} = \sqrt{\frac{1}{N}\left(\boldsymbol{y}_i - \boldsymbol{o}_i\right)^2}. \tag{8}$$

Here $\boldsymbol{y}_i$, $\boldsymbol{o}_i$ are the predicted and the observed samples, respectively.

$$R^2 = \frac{\left(\sum_{i=1}^{N}\left(\boldsymbol{o}_i - \overline{\boldsymbol{o}}_i\right)\left(\boldsymbol{y}_i - \overline{\boldsymbol{y}}_i\right)\right)^2}{\sum_{i=1}^{N}\left(\boldsymbol{o}_i - \overline{\boldsymbol{o}}_i\right)^2 \sum_{i=1}^{N}\left(\boldsymbol{y}_i - \overline{\boldsymbol{y}}_i\right)^2} \tag{9}$$

where $\overline{\boldsymbol{o}}_i$ and $\overline{\boldsymbol{y}}_i$ are the average of observed and response data of the model, respectively.

IV. Results and Discussion

A. Regression Performance

In this study, we performed 30 iterations on all sample sizes and 3 employed methods. With respect to the highest value of R^2, we selected the best models for each algorithm and their R^2 and RMSE values are reported in Table I.

The results of the PLSR technique at the leaf and canopy scales are shown in Table I. With increasing the training sample size ($\mathbf{n\_tr}$) from 25 to 150, the R^2 values of the model increase from 0.87 (RMSE = 0.13) to 0.98 (RMSE = 0.06) and from 0.78 (RMSE = 0.21) to 0.91 (RMSE = 0.12) at the leaf and canopy scales, respectively. Although R^2 values at the leaf scale are saturated between 125 and 150 samples, there is no saturation at the canopy scale. This means that the larger sample size at the canopy scale can result in higher performance of the PLSR model.

As illustrated in Table I, the ν-SVR model predictions results in higher R^2 value at the leaf scale, where it increases from 0.95 (RMSE = 0.12) to 0.98 (RMSE = 0.05) as more training data are available. Although R^2 values are saturated between 75 and 150 sample sizes, there is no saturation in RMSE values. At the canopy scale, the R^2 values increase from 0.85 to 0.95 and RMSE values decline from 0.17 to 0.12 as sample size increases from 25 to 150.

The predictions of GPR model in comparison with ν-SVR and PLSR show a smaller variation in accuracy as the training sample size increases. R^2 values change from 0.97 (RMSE = 0.05) to 0.98 (RMSE = 0.03) and from 0.91(RMSE = 0.14) to 0.97 (RMSE = 0.11) at the leaf and canopy scales, respectively. A specific degree of saturation in R^2 values can be observed between 50 and 150 samples at the leaf scale.

At the canopy scale, GPR model shows the highest accuracy in comparison with the other methods using the same training sample size (see Table I). Although no saturation in the accuracies is shown, a gain in R^2 and RMSE values could be observed for all methods. The scatter plots between the observed and the estimated DS of the implemented methods at the canopy scale are shown in Fig. 3. Each method was evaluated using three different sample size (25, 75, and 125 samples).

It should be considered that R^2 and RMSE values of all methods were changed in different iterations. For instance, at canopy

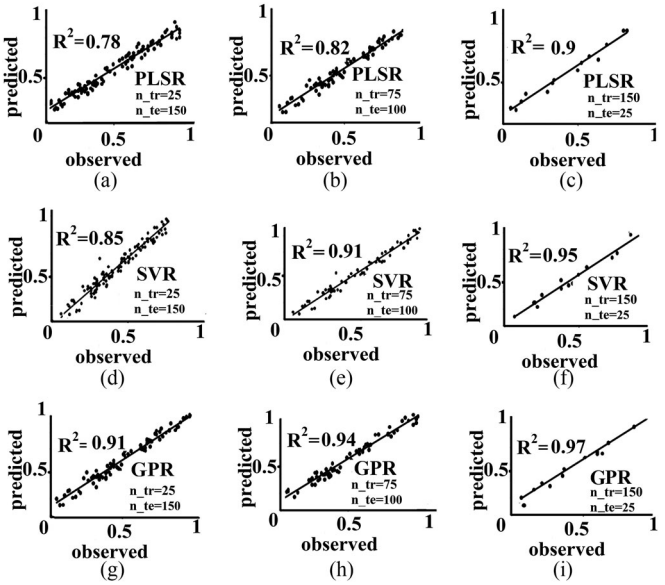

Fig. 3. Relationship between the observed and the estimated disease severity of PLSR (a, b, and c), GPR (d, e, and f) and ν-SVR (g, h, and i) for 25, 75, and 125 training sample size ($\mathbf{n\_tr}$) at leaf scale. $\mathbf{n\_ts}$ is the number of testing sample size at canopy scale.

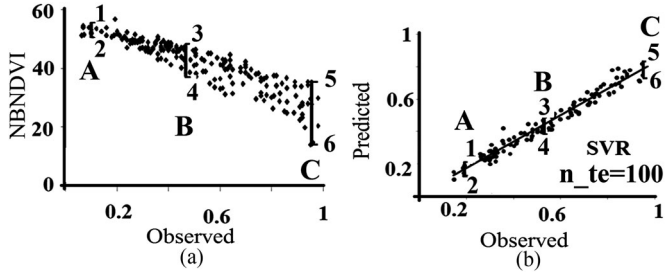

Fig. 4. Relationship between the NBNDVI and predicted disease of ν-SVR with observed disease severity to evaluate scattering of methods, disease severity is constant over eachline ($\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$).

scale, R^2 values variation of the PLSR method were changed from 0.13 (25 training sample size) to 0.02 (150 training sample size). These variations for GPR and ν-SVR were same and they were changed from 0.1 (25 training sample size) to 0.01 (150 training sample size). Also, RMSE values variation for PLSR, SVR, and GPR using 25 training sample sizes are 0.04, 0.02, and 0.02, respectively. The RMSE variations for all methods using 150 training sample sizes were decreased to 0.01. In order to statistically test the difference between the employed techniques results, the F-statistic approach was utilized. The reported p-value < 0.05 rejects the null hypothesis in 95% confidence level which states that there is significant linear correlation between the results using different techniques. At the leaf scale, the p-value just for 25 training sample size was greater than 0.05 in comparisons between the results of PLSR with ν-SVR and also PLSR with GPR. Hence, the null hypothesis for 25 training sample size at 95% level of significance was not rejected. At the canopy scale, the null hypothesis for all methods using different training sample sizes in 95% confidence level were not rejected, except for sample sizes 100, 125, and 150 between ν-SVR and GPR.

In this research, we tested three machine learning techniques to detect WLR disease with various sample sizes. In our data, SVR and GPR showed higher performances than PLSR, because both of them are less sensitive to spectral noise, and also due to use of sophisticated kernel in estimating VP which exhibit nonlinear relation [17]. In this study, GPR performed best with a training sample size of 10% (n=25) which could be due to the better generalization ability of GPR than SVR and PLSR; however it should be considered in future research using new dataset. As stated before, GPR saturated with small size which could be interpreted as the higher potential of GPR model than PLSR and ν-SVR methods at the leaf scale to predict WLR using a smaller sample size. This also can be concluded from Fig. 2 which illustrates the relationship between the observed and estimated WLR for 25, 75, and 125 training sample size.

From Figs. 2 and 3 and Table I, it can be concluded that the accuracies of regression techniques decline from the leaf scale to the canopy scale, which could be due to the strong illumination gradient, shadow and plant geometry at the canopy scale [11]. Although this work considered the plant structure, leaf area index, canopy gaps, litter, soil background and leaf optical properties, the performance of these techniques should be evaluated using airborne and space borne hyperspectral images. Hence, the future study has to consider the effect of atmosphere and imaging systems on the models predictions. There are many imaging systems with spectral resolution from 5 to 10 nm which can be used to implement and test our finding from this research, such as Hyperion, HyMap, etc.

B. Disease Symptoms Effects on WLR Detection Techniques

In order to study the effects of the disease symptoms on the employed techniques, we consider the scattering pattern of NBNDVI [see Fig. 4(a) and ν-SVR Fig. 4(b)] as the disease severity increases. These methods were chosen because Ashourloo [12] introduced NBNDVI as the best predictor in WLR detection using SVIs, and also the broad application of ν-SVR in plants' parameters estimation using different SVIs [18].

As illustrated in Fig. 4, NBNDVI values scatter [see Fig. 4(a)], whereas, ν-SVR values do not scatter [see Fig. 4(b)]. WLR disease has different symptoms during pathogenesis, which can be seen simultaneously in infected leaves. Each symptom has distinct spectrum, therefore, the spectra of the infected leaves are mixed and depend on the proportion of the disease symptoms [12], [20]. On the other hand, the proportions of disease symptoms are different from one leaf to another at specific disease severity. For instance, in a leaf with disease severity of 50%, the proportions of green, yellow, orange, brown, and dry are 50%, 19%, 20%, 7%, and 4%, respectively. However, in another infected leaf with disease severity of 50%, the proportions of green, yellow, orange, brown and dry are 50%, 8%, 11%, 11% and 20%, respectively. Therefore, the spectra of two samples are different, while similar in disease severity levels. As

TABLE II
DISEASE SYMPTOMS PROPORTION AT DIFFERENT DISEASE SEVERITY LEVELS

Point Number	Green (%)	Yellow (%)	Orange (%)	Dark Brown	Dead (%)
1-A [Fig. 4(a)]	90	7	3	0	0
2-A [Fig. 4(a)]	90	5	5	0	0
3-B [Fig. 4(a)]	50	19	20	7	4
4-B [Fig. 4(a)]	50	8	11	11	20
5-C [Fig. 4(a)]	10	38	21	17	14
6-C [Fig. 4(a)]	10	12	16	29	33
1-A [Fig. 4(b)]	90	6	4	0	0
2-A [Fig. 4(b)]	90	4	5	1	0
3-B [Fig. 4(b)]	50	14	23	11	2
4-B [Fig. 4(b)]	50	7	15	19	9
5-C [Fig. 4(b)]	10	28	30	23	9
6-C [Fig. 4(b)]	10	15	19	30	26

TABLE III
R^2 VALUES OF ν-SVR, PLSR, GPR, PRI AND NBNDVI AT TEN DISEASE SEVERITY LEVELS WITH 75 SAMPLE SIZE

Disease severity levels	R^2				
	ν-SVR	PLSR	GPR	PRI	NBMDVI
1–5%	0.91	0.91	0.92	0.56	0.7
5–10%	0.93	0.93	0.94	0.85	0.87
10–20%	0.95	0.96	0.97	0.84	0.88
20–30%	0.95	0.97	0.97	0.84	0.87
30–40%	0.98	0.97	0.98	0.83	0.86
40–50%	0.98	0.98	0.99	0.75	0.77
50–60%	0.97	0.98	0.99	0.69	0.71
60–70%	0.96	0.97	0.98	0.59	0.66
70–80%	0.97	0.96	0.98	0.47	0.53
$>80\%$	0.98	0.97	0.98	0.46	0.51

the recorded spectra of two samples have distinct pattern, their computed NBNDVI values are different as well. Moreover, by increasing the number of samples at specific disease severity, there will be more scatter in NBNDVI values [12]. This scattering is due to different fractions of disease symptoms at a given disease severity. The scattering of data decreases the accuracy of NBNDVI in estimating the disease severity levels. [13].

In order to show the disease symptoms effects on the models performances, we draw three lines ($\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$) at 0.1, 0.5, and 0.95 of disease severity levels [see Fig. 4(a) and (b)]. Then two points are chosen at the both ends of each line to illustrate the disease symptoms fraction at the same disease severity, i.e., points 1 and 2 on line $\mathbf{A}$, points 3 and 4 on line $\mathbf{B}$, and points 5 and 6 on line $\mathbf{C}$ (see Fig. 4). In Fig. 4(a), each point represents the computed NBNDVI value for a sample. It was expected that two samples (two infected leaves) with the same disease severity levels have equal NBNDVI values, however, these samples show different NBNDVI values due to the influence of different disease symptoms. As indicated by lines $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ in Table II, points on each line contains different fraction of green, yellow, orange, dark brown, and dead symptoms in comparison to each other. As can be seen in Fig. 4(a), the difference of NBNDVI values between points on line $\mathbf{B}$ (points 3 and 4) is more than the points on line $\mathbf{A}$ (points 1, 2), and the points on line $\mathbf{C}$ (points 5, 6) have the largest difference.

The differences between the lengths of the lines in Fig. 4(a) represent the scattering, which are due to the change in the disease symptoms fraction and their combinations (see Table II). For instance, line $\mathbf{A}$ in Table II represents the low disease severity levels which contains the yellow and orange symptoms with the similar combinations. But, line $\mathbf{B}$ and line $\mathbf{C}$ contain all symptoms with different combinations (see Table II).

In contrast to NBNDVI in Fig. 4(a), the length of lines $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ in Fig. 4(b) are same which means that the prediction of the ν-SVR model does not scatter as disease severity increases and also disease symptom fractions changes (see Table II). For example, the combinations of disease symptoms at the points of line $\mathbf{C}$ (points 5 and 6) is different from the points on line $\mathbf{A}$ (points 1 and 2); however, the lengths of line $\mathbf{A}$ and $\mathbf{C}$ are the same and the predictions of ν-SVR do not scatter (see Fig. 4).

Additionally, the performance of the regression models and SVIs (NBNDVI and PRI) were evaluated in ten disease severity levels and their R^2 values are reported in Table III. When the disease severity is less than 5%, the R^2 values of NBNDVI and PRI are less than 0.7, whereas R^2 values of ν-SVR, GPR, and PLSR are more than 0.9. These results illustrate the higher performance of regression techniques in early WLR detection. The R^2 values of NBNDVI and PRI increase in disease severity levels between 5% and 40%, and then R^2 values decrease due to the effect of different disease symptoms [12]. In the case of PLSR, GPR and ν-SVR techniques, the R^2 values are always more than 0.93 when the disease severity levels are more than 5%. Our results showed that disease symptoms adversely affected the performance of SVIs in estimating the disease severity levels. Because SVIs developed for remote sensing are sensitive to variability in the carbon, pigment content and biomineral content which are strongly influenced by disease development. In contrast to SVIs which are based on two or few bands, PLSR, SVR, and GPR use full spectrum, where the suitable spectral bands have the greatest weight in the model [16]. Suitable spectral bands are the narrow bands which are more sensitive to disease severity and less affected by disease symptoms [13]. These techniques need to be examined in detection of the other plant diseases which have different symptoms during pathogenesis.

V. CONCLUSION

The objectives of this paper were to employ the machine learning techniques for WLR disease detection as well as evaluating the training sample size and influence of disease symptoms effects on the methods predictions. Moreover, this study compared the performance of PLSR, ν-SVR, and GPR with the PRI and NBNDVI. The experiment was conducted in the greenhouse under controlled conditions to study the different disease symptoms effects on reflectance of the leaves. The combinations of disease symptoms at each disease severity level resulted in very complex spectra which declined the accuracies of PRI and NBNDVI. However, they did not have adverse impacts on PLSR, ν-SVR, and GPR performances. It is evident in this study that GPR's performance using smaller training dataset results in higher accuracy than other implemented methods. Additionally, challenges in early detection of plant disease due to minor

change in reflectance could be reduced using machine learning regression techniques. It must be noted that PLSR, ν-SVR, and GPR need to be tested on various sensors and different varieties of wheat in order to be used in the field.

ACKNOWLEDGMENT

The authors would like to thank anonymous reviewers for their comments and suggestions that have helped to improve this manuscript.

REFERENCES

[1] S. Sankaran, A. Mishra, R. Ehsani, and C. Davis, "A review of advanced techniques for detecting plant diseases," *Comput. Electron. Agriculture*, vol. 72, no. 1, pp. 1–13, 2010.

[2] C. Buschmann and E. Nagel, "In vivo spectroscopy and internal optics of leaves as basis for remote sensing of vegetation," *Int. J. Remote Sens.*, vol. 14, no. 4, pp. 711–722, 1993.

[3] N. K. Poona and R. Ismail, "Using Boruta-selected spectroscopic wavebands for the asymptomatic detection of *Fusarium circinatum* stress," *IEEE J. Select. Topics Appl. Earth Observations Remote Sens.*, vol. 7, no. 9, pp. 3764–3772, Sep. 2014.

[4] J. Zhang, R. Pu, L. Yuan, W. Huang, C. Nie, and G. Yang, "Integrating remotely sensed and meteorological observations to forecast wheat powdery mildew at a regional scale," *IEEE J. Select. Topics Appl. Earth Observations Remote Sens.*, vol. 7, no. 11, pp. 4328–4339, 2014.

[5] W. Huang *et al.*, "New optimized spectral indices for identifying and monitoring winter wheat diseases," *IEEE J. Select. Topics Appl. Earth Observations Remote Sens.*, vol. 7, no. 6, pp. 2516–2524, Jun. 2014.

[6] M. D. Bolton, J. A. Kolmer, and D. F. Garvin, "Wheat leaf rust caused by *Puccinia triticina*," *Molecular Plant Pathology*, vol. 9, no. 5, pp. 563–575, 2008.

[7] C. Robert, M.-O. Bancal, B. Ney, and C. Lannou, "Wheat leaf photosynthesis loss due to leaf rust, with respect to lesion development and leaf nitrogen status," *New Phytologist*, vol. 165, no. 1, pp. 227–241, 2005.

[8] J. G. Hansen, "Use of multispectral radiometry in wheat yellow rust experiments," *EPPO Bull.*, vol. 21, pp. 651–658, 1991.

[9] W. Huang, D. W. Lamb, Z. Niu, Y. Zhang, L. Liu, and J. Wang, "Identification of yellow rust in wheat using *in-situ* spectral reflectance measurements and airborne hyperspectral imaging," *Precision Agriculture*, vol. 8, nos. 4/5, pp. 187–197, 2007.

[10] J. Zhang, R. Pu, W. Huang, L. Yuan, J. Luo, and J. Wang, "Using in-situ hyperspectral data for detecting and discriminating yellow rust disease from nutrient stresses," *Field Crops Res.*, vol. 134, pp. 165–174, 2012.

[11] J. Franke and G. Menz, "Multi-temporal wheat disease detection by multispectral remote sensing," *Precision Agriculture*, vol. 8, no. 3, pp. 161–172, 2007.

[12] D. Ashourloo, M. R. Mobasheri, and A. Huete, "Evaluating the effect of different wheat rust disease symptoms on vegetation indices using hyperspectral measurements," *Remote Sens.*, vol. 6, no. 6, pp. 5107–5123, 2014.

[13] D. Ashourloo, M. R. Mobasheri, and A. Huete, "Developing two spectral disease indices for detection of wheat leaf rust (*Puccinia triticina*)," *Remote Sens.*, vol. 6, no. 6, pp. 4723–4740, 2014.

[14] A.-K. Mahlein, U. Steiner, H.-W. Dehne, and E.-C. Oerke, "Spectral signatures of sugar beet leaves for the detection and differentiation of diseases," *Precision Agriculture*, vol. 11, no. 4, pp. 413–431, 2010.

[15] C. Ryu, M. Suguri, and M. Umeda, "Multivariate analysis of nitrogen content for rice at the heading stage using reflectance of airborne hyperspectral remote sensing," *Field Crops Res.*, vol. 122, no. 3, pp. 214–224, 2011.

[16] M. Mirzaie, R. Darvishzadeh, A. Shakiba, A. Matkan, C. Atzberger, and A. Skidmore, "Comparative analysis of different uni-and multi-variate methods for estimation of vegetation water content using hyper-spectral measurements," *Int. J. Appl. Earth Observation Geoinform.*, vol. 26, pp. 1–11, 2014.

[17] J. Verrelst *et al.*, "Machine learning regression algorithms for biophysical parameter retrieval: Opportunities for sentinel-2 and-3," *Remote Sens. Environ.*, vol. 118, pp. 127–139, 2012.

[18] J. Verrelst, L. Alonso, J. P. R. Caicedo, J. Moreno, and G. Camps-Valls, "Gaussian process retrieval of chlorophyll content from imaging spectroscopy data," *IEEE J. Select. Topics Appl. Earth Observations Remote Sens.*, vol. 6, no. 2, pp. 867–874, Apr. 2013.

[19] J. Behmann, J. Steinrücken, and L. Plümer, "Detection of early plant stress responses in hyperspectral images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 93, pp. 98–111, 2014.

[20] T. Rumpf, A.-K. Mahlein, U. Steiner, E.-C. Oerke, H.-W. Dehne, and L. Plümer, "Early detection and classification of plant diseases with support vector machines based on hyperspectral reflectance," *Comput. Electron. Agriculture*, vol. 74, no. 1, pp. 91–99, 2010.

[21] Z. Ramadan, P. K. Hopke, M. J. Johnson, and K. M. Scow, "Application of PLS and back-propagation neural networks for the estimation of soil properties," *Chemometrics Intell. Laboratory Syst.*, vol. 75, no. 1, pp. 23–30, 2005.

[22] M. Schwieder, P. J. Leitão, S. Suess, C. Senf, and P. Hostert, "Estimating fractional shrub cover using simulated EnMAP data: A comparison of three machine learning regression techniques," *Remote Sens.*, vol. 6, no. 4, pp. 3427–3445, 2014.

[23] D. Ashourloo, A. A. Matkan, A. Huete, H. Aghighi, and M. R. Mobasheri, "Developing an index for detection and identification of disease stages," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 6, pp. 851–855, Jun. 2016.

[24] J. Steinier, Y. Termonia, and J. Deltour, "Smoothing and differentiation of data by simplified least square procedure," *Anal. Chem.*, vol. 44, no. 11, pp. 1906–1909, 1972.

[25] H. Al-Hiary, S. Bani-Ahmad, M. Reyalat, M. Braik, and Z. ALRahamneh, "Fast and accurate detection and classification of plant diseases," *Int. J. Comput. Appl.*, vol. 17, pp. 31–38, 2011.

[26] M. Dmochowska-Boguta *et al.*, "Pathogen-regulated genes in wheat isogenic lines differing in resistance to brown rust *Puccinia triticina*," *BMC Genomics*, vol. 16, no. 1, 2015, Art. no. 742.

[27] T. Udelhoven, C. Emmerling, and T. Jarmer, "Quantitative analysis of soil chemical properties with diffuse reflectance spectrometry and partial least-square regression: A feasibility study," *Plant Soil*, vol. 251, no. 2, pp. 319–329, 2003.

[28] J. Farifteh, F. Van der Meer, C. Atzberger, and E. Carranza, "Quantitative analysis of salt-affected soil reflectance spectra: A comparison of two adaptive methods (PLSR and ANN)," *Remote Sens. Environ.*, vol. 110, no. 1, pp. 59–78, 2007.

[29] M. Daszykowski, S. Serneels, K. Kaczmarek, P. Van Espen, C. Croux, and B. Walczak, "Tomcat: A Matlab toolbox for multivariate calibration techniques," *Chemometrics Intell. Laboratory Syst.*, vol. 85, no. 2, pp. 269–277, 2007.

[30] A. Okujeni, S. Van der Linden, B. Jakimow, A. Rabe, J. Verrelst, and P. Hostert, "A comparison of advanced regression algorithms for quantifying urban land cover," *Remote Sens.*, vol. 6, no. 7, pp. 6324–6346, 2014.

[31] C.-C. Chang and C.-J. Lin, "Training v-support vector regression: Theory and algorithms," *Neural Comput.*, vol. 14, no. 8, pp. 1959–1977, 2002.

[32] V. Anandhi and R. M. Chezian, "Support vector regression to forecast the demand and supply of pulpwood," *Int. J. Future Comput. Commun.*, vol. 2, no. 3, pp. 266–269, 2013.

[33] S. S. Keerthi and C.-J. Lin, "Asymptotic behaviors of support vector machines with Gaussian kernel," *Neural Comput.*, vol. 15, no. 7, pp. 1667–1689, 2003.

[34] D. Bhatt, P. Aggarwal, P. Bhattacharya, and V. Devabhaktuni, "An enhanced MEMS error modeling approach based on nu-support vector regression," *Sensors*, vol. 12, no. 7, pp. 9448–9466, 2012.

[35] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, 2011, Art. no. 27.

[36] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statist. Comput.*, vol. 14, no. 3, pp. 199–222, 2004.

[37] L. Hu, X. Che, and X. Cheng, "Bandwidth prediction based on nu-support vector regression and parallel hybrid particle swarm optimization," *Int. J. Comput. Intell. Syst.*, vol. 3, no. 1, pp. 70–83, 2010.

[38] M. Lázaro-Gredilla, M. K. Titsias, J. Verrelst, and G. Camps-Valls, "Retrieval of biophysical parameters with heteroscedastic Gaussian processes," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 4, pp. 838–842, Apr. 2014.

[39] J. Verrelst, L. Alonso, G. Camps-Valls, J. Delegido, and J. Moreno, "Retrieval of vegetation biophysical parameters using Gaussian process techniques," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 5, pp. 1832–1843, May 2012.

[40] C. E. Rasmussen and H. Nickisch, "Gaussian processes for machine learning (GPML) toolbox," *J. Mach. Learn., Res.*, vol. 11, pp. 3011–3015, 2010.

[41] C. E. Rasmussen and C. K. Williams, *Gaussian Processes for Machine Learning*, Dhaka, Bangladesh: Univ. Press Group Limited, 2006.
[42] Y. Bu and J. Pan, "Stellar atmospheric parameter estimation using Gaussian process regression," *Monthly Notices Roy. Astron. Soc.*, vol. 447, no. 1, pp. 256–265, 2015.

Davoud Ashourloo received the B.Sc. degree in natural resource engineering, the M.Sc. degree in remote sensing and geographic information system from Shahid Beheshti University, Tehran, Iran, in 2001 and 2003, respectively, and the Ph.D. degree in remote sensing from K.N. Toosi University of Technology, Tehran.

His research interests include the areas of remote sensing, conducting research on the wheat leaf rust detection using hyperspectral data and image processing.

Hossein Aghighi received the B.Sc. degree in natural resource engineering, Iran, the M.Sc. degree in remote sensing and geographic information system from Tarbiat Modares University, Tehran, Iran, and the Ph.D. degree in remote sensing from The University of New South Wales, Sydney, Australia in, 2000, 2012, and 2015, respectively.

His current research interests include probabilistic graphical models with application in remote sensing, statistical analysis of polarimetric synthetic aperture radar images, and application of machine learning techniques in diagnosis of plant diseases.

Ali Akbar Matkan received the M.Sc. degree in geography from Tarbiyat Modares University, Tehran, Iran and the Ph.D. degree in remote sensing from University of Bristol, Bristol, England, in 1992 and 2000, respectively.

His current research interests include spectroradiometry of vegetation and information extraction from hyperspectral data.

Mohammad Reza Mobasheri received the B.Sc. degree in physics from Tehran University, Tehran, Iran, the M.Sc. degree in remote sensing from Toledo University, Toledo, OH, USA, and the Ph.D. degree in remote sensing from James Cook University, Townsville, QLD, Australia, in 1975, 1980, and 1995, respectively.

His current research interests include spectroradiometry of vegetation and information extraction from hyperspectral data.

Amir Moeini Rad received the B.Sc. degree in civil engineering, the M.Sc. degree in remote sensing from K.N. Toosi University of Technology, Tehran, Iran, in 2002 and 2006, respectively, and the Ph.D. degree in remote sensing from the same university in 2016.

His research interests include the areas of remote sensing and image processing.