Object-oriented Model to Predict Crop Yield Using Satellite-based Vegetation Index

https://doi.org/10.3991/ijim.v16i15.33269

Zena H. Khalil[1,2(✉)], Amel H. Abbas[2]
[1]College of Computer Science and Information Technology, University of Al-Qadisiyah, Al-Diwaniyah, Iraq
[2]College of Science, Mustansiriyah University, Baghdad, Iraq
zena.khalil@qu.edu.iq

Abstract—Different models had been developed to predict crop yields based on remotely sensed data. Most approaches were based on developing empirical relationships between the satellite-based normalized difference vegetation index (NDVI) data and the crop yield. This article is proposed to introduce a methodological framework for constructing an object-oriented yield prediction model using satellite data based on the two-level regression models. Here, the trends caused by the influence of technological improvements were considered. Regression models for the wheat and barley crop yield predictions have been developed. The two-level regression model, including the foreword stepwise regression (FSR) technique, firstly selects the set of features that reflect the spatial variations in crops, soil, and agriculture management within districts. After the steps of exploratory data analysis (EDA), object creation, and the zonal average of each object were carried out. The second level consists of yield prediction with multiple linear regression (MLR), least absolute shrinkage, and selection operator (Lasso), support vector machines (SVM) techniques. In the proposed model, the SVM technique outperforms the rest techniques by an average root mean square error (RMSE) of 5.59(4.51) for wheat(barley). The experiments showed that the proposed model provides stability and low prediction error in the vast majority of cases and the used techniques.

Keywords—remote sensing, NDVI, machine learning, feature selection

1 Introduction

The development of earth observation over the last decades aimed to increase the scope of satellite information and decrease the cost at the same time. It also tends to use bands with higher frequency [1]. However, this introduced a chance to improve technologies of automated information processing that solve important problems in related sectors including agriculture. One of the active tasks and field applications in agriculture based on satellite information is yield forecasting systems. Here, normalized difference vegetation index (NDVI) time series have been used for grain yield predictions since the 1980s. The studies found that NDVI variables are very significant

as grain yield predictors for wheat and barley [2]–[4]. This was explained as NDVI reflects a strong correlation with grain yield, especially in the time when the grain productivity becomes sensitive to weather and moisture conditions during the grain development period [5],[6]. This period is called the critical period, which dominates the final productivity. Generally, the NDVI is influenced primarily by some slowly changing environmental factors, such as climate, soil, and topography, which are also called the ecosystem components (EC). The ecosystem components affect the amount and distribution of vegetation on Earth [7]. It is also important to note that every geographic region on Earth contains a specific level of ecosystem resources that indicate the ecosystem potential, which is also known as the carrying capacity of an area. Thus NDVI values can measure the amount of vegetation signifying the carrying capacity of a particular geographic region [8]. In addition, a previous study showed that the NDVI can describe the crops' health and growing conditions in the local administrative region [9]. Also, the crop productivity can be monitored by the NDVI during the entire growing season in the local administrative region to be successfully used to predict crop yields over that area [10]. That is attributed to the spectral information that is embedded within the NDVI and makes it valuable data in examining the vegetation conditions [11]. Based on the mentioned features, we were inspired to use NDVI as a predictor in our recent suggesting yield predicting model over a large area.

To expand the application of the prediction model to more than a single crop within Voronezh and obtain a solid and accurate model, the object-oriented approach was adopted in this work for the first time. However, the object-oriented approach of satellite images is meant to segment into regions, each of which is called an object. A review of using an object-oriented approach for vegetation analysis was carried out in [12]. All the previous works applied the approach of object-oriented satellite images in the object recognition or earth's surface change detection applications as in [13]–[16]. The use of an object-oriented approach showed an increase in the accuracy of target object recognition, which prompted us to examine the use of this approach in the field of yield prediction. So, in our approach, we firstly segmented the cropland area into objects and selected the appropriate objects for each crop to take features from. In order to detect the human-induced factors, we considered the time trends of crop yields. The human-induced factors reveal the influence of technology factors of fertilizer and management improvement that cannot be detected by using NDVI only. Hence, the interannual variability of the NDVI can only reveal crop yield fluctuations caused by weather or climate conditions [2],[17]. Therefore, we tested the combination of the remotely sensed data with the trends of long-term historical crop yields data in the predictor group. The objective of the present work was to build a methodological framework adopted for crop yields prediction by four assumptions: (a) establishing temporal analysis for determining the critical period of the wheat/barley growing season to extract the NDVI values of the images in these periods and use them as predictors; (b) analyze the historical trends in the wheat/barley yield; (c) segment the cropland image of each province into objects; and (d) construct two-level prediction models for wheat/barley yields, included:

- **In the first level**, the optimal predictors(features) would be selected locally for each province.
- **In the second level**, previous features would be train globally and combined with additional constructed features related to yield trends.

2 Research material

2.1 Study area

The Voronezh region is considered one of the main producers of good-quality grain. More than half of the sowing of crops is occupied by cereals including winter wheat, corn, barley and others. Due to the existing rich resources and an appropriate climate for grain crop production. According to [18], the area of agricultural land covers 4005.1 thousand hectares, and the arable land (cropland) represents 79.5% of agricultural land; i.e. 3038.2 thousand hectares. An analysis of the crop yields in the Voronezh region shows that there is an increase in grain yield. For example, the average yield of winter wheat for 2006–2010 reached 24.2 c/ha. In the period of 2011–2012, the winter yield continued in the same trend during the 2013, where it reaches 28.4 c/ha. This increase in productivity was perhaps due to the increasing in the number of applied mineral fertilizationers. The Voronezh Region is part of the Central Federal District and the administrative-territorial division of Russia, the Voronezh region is a territorial organization of the state (administrative-territorial structure of the subject of the Russian Federation), consisting of 31 municipal areas (Figure 1) of rural and urban settlements [19]. We chose 27 from them, which are described as famous region for growing wheat and barley.

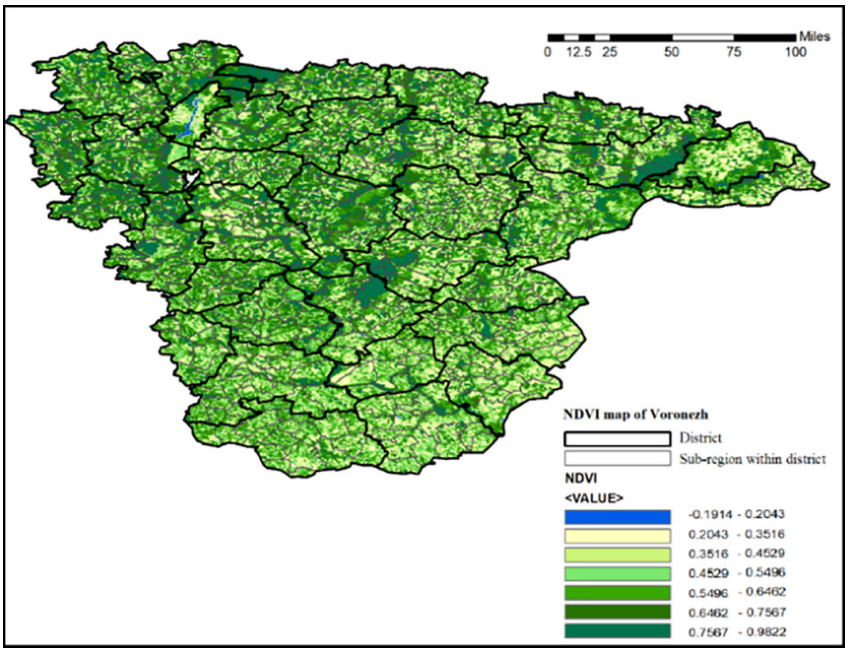

Fig. 1. The NDVI map of Voronezh with plotted boundaries of districts and subregions

2.2 Satellite and yield data

The annual time series of wheat and barley crop data was analysed and modeled in centner per hectare (c/ha) from 2001 to 2014 for 27 districts of the Voronezh region. These data were obtained from tables of official statistics, which is taken from the work [3]. The Terra Moderate Resolution Imaging Spectroradiometer (MODIS) MOD13Q1 product [20], was used to collect NDVI images for all provinces during the growing season (April-August). MODIS land cover (MCD12Q1) [21] images that provide global land cover types were used to create a crop area mask of Voronezh to collect satellite data for that area, by extracting the only pixels which were recorded as cropland from NDVI images of the study area throughout 2001–2014.

3 Methods: data analysis and regression model

3.1 The exploratory data analysis

First of all, it is necessary before applying any machine-learning algorithm to do exploratory data analysis (EDA), to reduce the input dimensionality and select the appropriate inputs. In our case, we conducted EDA in two steps; the first one was to determine the critical period of crop growing season. Where the captured NDVI images consisted of a 16-day composite that span the length of the growing season from April to August. To determine the dates of the critical period, the NDVI images were spatially accumulated (mean of NDVI pixels) within each district, yielding $27_{(\text{districts})} \times 5_{(\text{growing season})} \times 2_{(\text{images per month})} \times 14_{(\text{years})} = 3780$ NDVI values. Then a scatter plot of median_NDVI among the same dates in all years of the study period was done. The scatter plot (Figure 2) showed that the maximum of the annual NDVI cycle in June–July reaches saturation of about 0.7 when the projective coverage is 100%, and generally, when reaching the saturation, NDVI stops working. This means that the critical period in the June-July period. The second step of EDA was to determine the best dates of the growing season to make an accurate crop yield prediction.

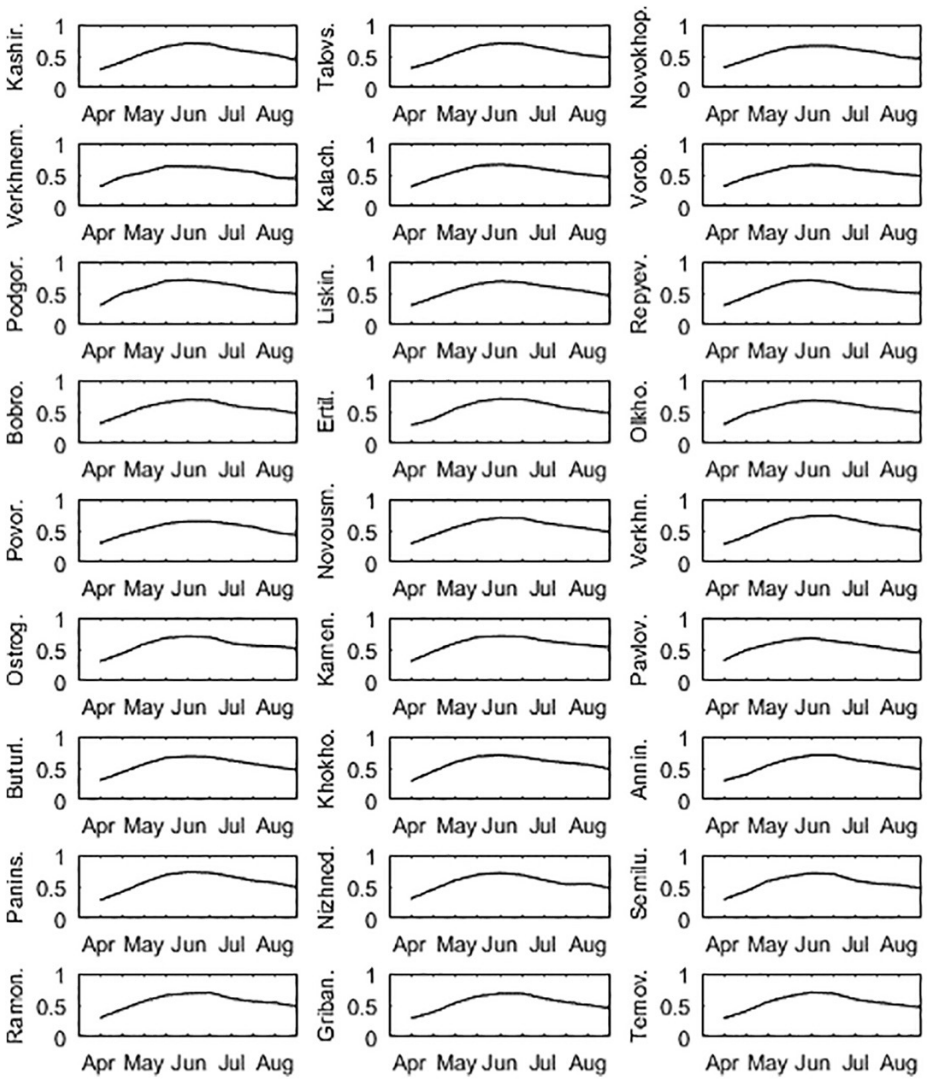

Fig. 2. The temporal analysis of NDVI data in various districts of the Voronezh

However, the lead time must be long enough for the prediction to be useful [22]. So to ensure a significant lead time, not as in other works, we considered the only data from the dates before the critical period. This was done by conducting Spatio-temporal correlation analysis separately for each crop. The correlations were calculated at the district level twice per month across the growing season between the mean_NDVI series and wheat/barley yields. The result of Spatio-temporal analysis showed that the "true crop density" occurred in April and May (Figure 3). Hence, if the winter crops successfully passed the winter in a well way and good tillering. Then in the conditions of usually rainy and warm May and June, the plant will quickly reach the flowering phase

(in only one week). Therefore, as a result, the dates of the first and second part of April and the first part of May are the most suitable date for using their data (NDVI-images) in the prediction model.

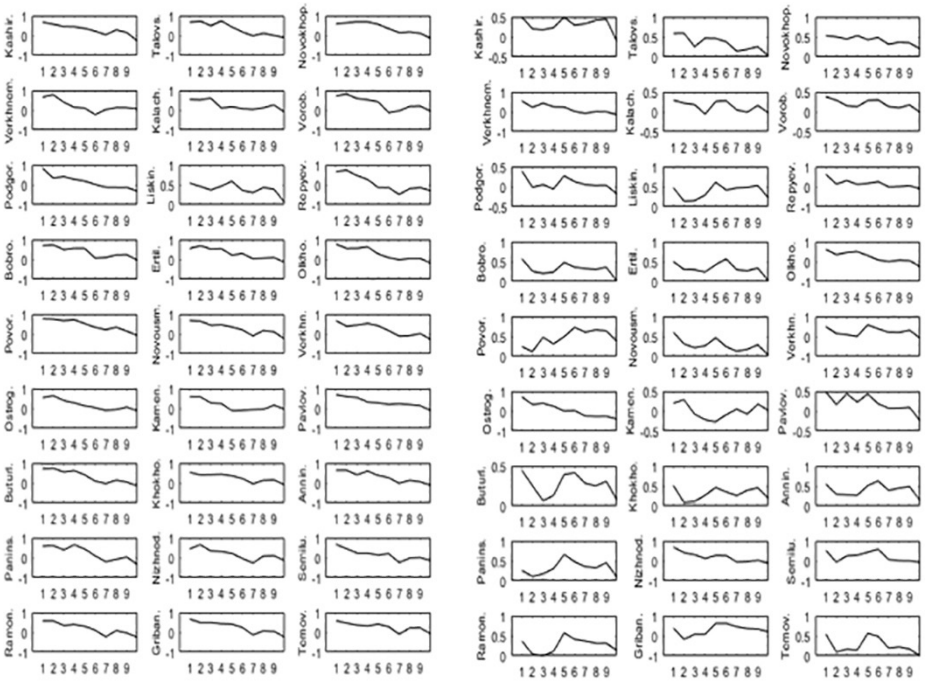

Fig. 3. The correlation coefficient for spatially accumulated NDVI versus the yield of crops in the regions of the Voronezh every 16 days (noted as numbers 1,2..9) from April through August. Wheat on the left, barley on the right

3.2 Crop yield trends analysis

Over the 14 years, the crop yields frequently is experienced a significant statistic long-term trend (upward most often) because of the technological improvements in crop cultivation (Figure 4). This trend can be approximated by a polynomial linear/nonlinear equation, which describes this change over time. While the short-term fluctuations in crop yields around this trend are often caused as a result of variations in the weather/climate over the growing season from one year to the other [23].

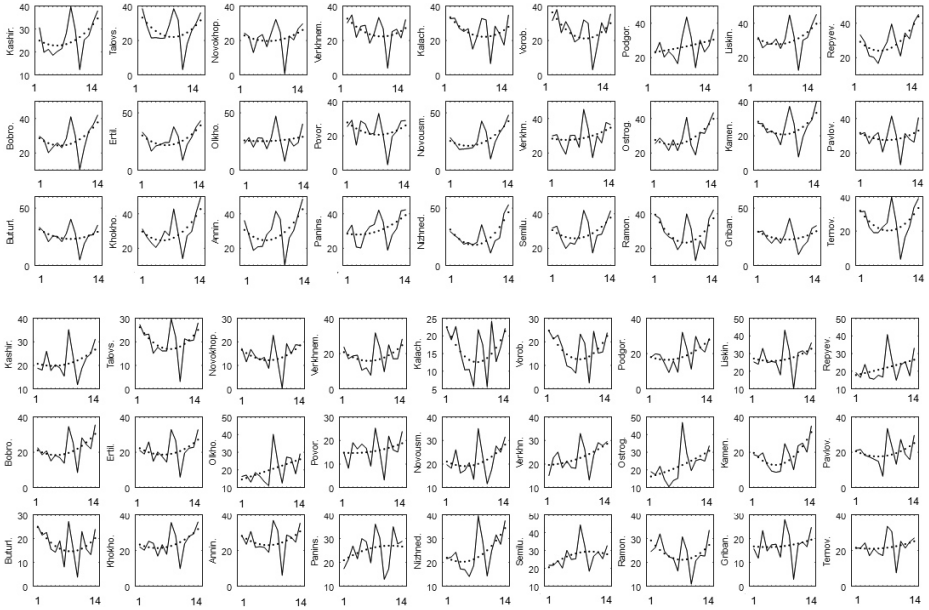

Fig. 4. Yields of winter wheat (upper) and barley (down) in 27 districts of the Voronezh from 2001 to 2014. The dotted curve represents a quadratic trend (polynomial fit of degree 2)

Considering the yield trends analysis, the historical time series of wheat and barley yield Y was decomposed into two components: the first deterministic component Y_t describes the long-term trend related to technological change. The second random component Y_r describes the yield deviation from the technological trend due to weather fluctuations as:

$$Y = Y_t + Y_r \quad (1)$$

Usually, in other works, the first component is approximated by de-trend methods such as moving averages, exponential algorithms, least-squares, or polynomial regressions [24]. While the second component is approximated by the difference or ratio between the actual yield and the trend yield (first component) [25]. In this study, we preferred to un-detrend the yield, to avoid prejudicing the chances of successfully using processed data in estimating relationships and building effective prediction models designed to take into account seasonal fluctuations [26]. Instead, we tended to adopt both components as predictors in training the prediction model. Here, we approximated the second component Y_r using remotely sensed data (NDVI), because NDVI reflects environmental conditions and vegetation as mentioned in section 1. The first component Y_t was approximated by one of these variables: linear time, quadratic time, last year's yield, yield trend (quadratic polynomial) DY, and their combinations.

3.3 Segmentation of districts into subregions (The object's creation)

After the EDA analysis was carried out and determined the best dates to consider their data as predictors, only the three images (two in April and the third in May) for each district were considered. A (3 × 27) NDVI images were collected for each year. Thus over 14 years, we had 1134 MODIS-NDVI images. The most common approach is to spatially accumulate NDVI pixels (by using samples, integrated, summed, or average image pixels) per county or district [27]. These approaches reflect only spatial variations in management and soil properties between districts without any attention to variations between crops characteristics such as greenness, biomass, and planting date; also, without attention to spatial variations in agriculture management and soil within the same district. The contribution is to integrate the principle of object-oriented modeling into the analysis and pre-processing stage. The object-orient in remote sensing involves the process of sensed image separation into separate areas or regions with similar statistical characterization. The areas (regions) obtained at the segmentation stage will be called objects [28]. In our method, the NDVI-images of districts were separated into several sub-region (objects) according to official rural and settlements partitions in each district to ensure control of variations in agriculture management. For example the type of crop and its cultivation conditions, including the phases of the growing season for each crop cultivated in each rural and settlements within each district. After the objects' creation, NDVI values were spatially accumulated (zonal mean operation) for each object resulting in 17304 numeric features (Table 1) representing the initial predictors set.

Table 1. The initial predictors set of Y_r, composed of NDVI objects, the initial set was computed as (number of sub-regions × 3 × 14)

District Name	Sub Regions	Number of Initial Predictors	District Name	Sub Regions	Number of Initial Predictors
Kashirsky	14	588	Verkhnekhavsky	17	714
Talovskiy	24	1008	Ostrogozhsky	20	840
Novokhopersky	11	462	Kamensky	11	462
Verkhnemamonsky	10	420	Pavlovsky	15	630
Kalacheevsky	17	714	Buturlinovsky	16	672
Vorobievsky	11	462	Khokholsky	15	630
Podgorensky	16	672	Anninsky	23	966
Liskinsky	23	966	Paninsky	16	672
Repyevsky	11	462	Nizhnedevitsky	15	630
Bobrovsky	19	798	Semiluksky	15	630
Ertilsky	14	588	Ramonsky	16	672
Olkhovatsky	8	336	Gribanovsky	17	714
Povorinsky	9	378	Ternovsky	14	588
Novousmansky	15	630			

3.4 Two-level prediction model

After extraction of the initial set of NDVI feature predictors, the features are exported into the prediction model. The proposed model consists of two levels:

First level: features selection. Before using the initial set of features, a feature selection step was done to reduce the total number of features and to remove irrelevant and redundant data. In many machines learning models, a wide range of input features (predictors) are available. Which can be used as input predictors to the machine learning model. However, it is difficult to determine which ones are most relevant or useful at all [29]. In meteorological forecasts, one of two approaches to the choice of predictors is usually used [30]. Using the entire possible set of predictors and gradually excluding less significant predictors, or starting from significant features and gradually adding new predictors. The successful application of feature selection must not only reflect the important information for prediction but also must reduce the computational and analytical work for the analysis of high-dimensional data [31]. Variety techniques for finding an optimal subset from features were introduced. For example, the decision tree approach, genetic algorithm (GA), forward feature selection based on a chi-square score or p-value, and multiple linear regression model based on significant feature selection, and backward-elimination procedure for SVM-based feature ranking. Generally, the multiple linear regression MLR models don't consider the nonlinear relations. While the SVM has a limitation in kernel selection caused by inaccurate selection. The GA has an expensive runtime cost and suffers from slow convergence before finding an accurate solution because of the use of minimal prior knowledge and failure in exploiting local information [32]. In the proposed model, a Foreword Stepwise Regression (FSR) technique was employed to identify the best set of features from the given initial dataset of features. The stepwise regression provides a good performance and a way to avoid multicollinearity without high complex and time-consuming because it was originally developed as a feature-selection technique for linear regression. The forward stepwise regression approach is a sequential feature selection technique different from generalized sequential feature selection. It can remove added features or add features that have been removed. Features are sequentially added to the empty set of candidates until the addition of new features leads to a further decrease in the error criterion [33]. For each district, FSR is implemented 'locally' to find the best eight features that predict the yield in that district. The selected features reflect local relation to their district. For example, the features from May were selected for some districts while the another had the features from April. Also the featured from north sub-regions were selected for wheat while other features from other sub-regions were selected for barley. This method will enable the model to deal with the variability characteristics of each district and overcome the drawback of other works.

Second level: prediction. The next step is to use the features produced in the previous level to be training 'generally' in the prediction model. The feature selection in the previous level helped in standardizing the number of features for each district to be then used as predictors of a single general prediction model in this level. The prediction model would deal with each district as the sample, each sample has 9 features (NDVI features with technological improvements trend). Three techniques were employed to train the prediction model, which are Multiple linear regression (MLR), Lasso (L1),

and SVM. Then finally the results were compared for evaluation and determination of the best technique.

4 Results and model measurement

To be that the feature set is independent of the evaluation set and also to avoid prediction bias, the predictors and yield for the first ten years were used for training the model. While the last 4 years (2011–2014) were used for test the model.

The crop yields were first estimated using the NDVI features selected by FSR only. Then by adding one of the features of (last year's yield Y(t – 1), linear time t, quadratic time t^2, yield trend DY as computed in section 3.2) to produce 15 models for both wheat and barley, different in their training data and techniques. The comparison between these models was done based on the coefficient of determination (R^2) and Root Mean Squared Error (RMSE) values. However, the values of R^2 and RMSE showed that the model trained with the combination of NDVI and yield trend DY features gave the best estimations. So, we adopted this training set to predict the wheat/barley yields through 2011–2014 using MLR, Lasso, and SVM techniques. Figure 5 shows the scatter plots of the predicted against measured yields trained with NDVI and yield trend DY data.

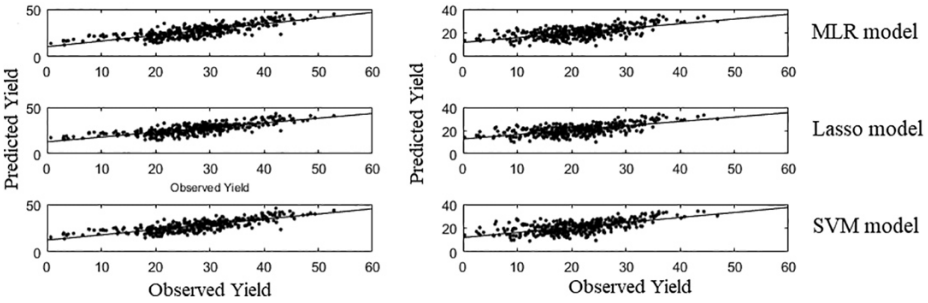

Fig. 5. Scatter plots of predicted vs. observed yield from 2001 to 2014, Wheat on the left, barley on the right

To demonstrate which technique has better results, the comparison between predicted and observed yield through testing intervals can be performed. Hence, each year (2011–2014) has different training data which can output a different model resulting in the prediction of yield from 2011 to 2014. The performance can be evaluated for each district, year-by-year as well as by 4-overall evaluation. To do that, some metrics were selected such as Absolute Error and RMSE. Table 2 and Table 3 showed the actual and prediction errors for wheat and barley through each year and district. The absolute error between the predicted and observed yields of 2011–2014 years was not exceed 19(16) centner/ha for wheat(barley) in the worst case.

Table 2. Absolute errors of wheat c/h for 2011–2014 years

District Name	Absolute Error											
	MLR				Lasso				SVM			
	2011	2012	2013	2014	2011	2012	2013	2014	2011	2012	2013	2014
Kash.	7.9	3.1	3.6	2.4	7.0	2.4	4.0	6.0	7.4	3.0	3.7	2.7
Talov.	0.2	4.0	1.9	3.7	1.1	5.3	4.9	7.8	0.5	4.8	1.9	2.7
Novokh.	3.2	11.2	4.1	0.5	0.8	10.1	4.4	0.1	2.1	11.0	2.8	0.9
Verkhm.	2.8	3.3	3.3	1.8	3.6	2.9	3.7	2.7	1.9	2.1	4.3	2.1
Kalach.	7.0	6.3	6.3	0.4	6.2	7.5	7.1	2.5	6.0	6.7	7.8	0.6
Vorob.	0.9	2.3	0.3	0.2	3.0	1.6	0.6	3.4	1.9	1.7	0.3	0.1
Podgo.	8.8	1.5	0.1	3.1	6.6	0.9	0.0	1.6	7.4	1.4	0.5	2.3
Liski.	0.9	1.3	9.6	5.1	1.4	1.3	10.4	7.8	1.3	1.6	9.7	6.7
Repy.	1.2	1.3	5.9	6.5	3.0	2.8	8.5	9.0	1.6	1.8	6.0	7.7
Bobro.	1.1	2.8	2.0	4.2	4.1	3.3	4.2	7.0	1.9	3.3	2.8	3.9
Ertil.	4.9	6.1	0.2	19.7	3.7	5.3	1.3	18.5	5.5	5.4	1.1	19.6
Olkho.	6.4	3.4	3.2	3.8	5.2	3.2	3.7	6.1	6.2	4.0	3.5	5.3
Povo.	4.8	3.4	1.5	0.2	2.9	3.6	0.6	1.3	3.8	4.0	1.1	0.7
Novousm.	5.6	5.6	6.0	8.9	6.0	3.0	6.9	12.0	5.5	3.9	5.8	8.9
Verkhn.	1.2	5.8	15.4	12.9	0.7	3.6	14.7	11.2	0.5	3.1	12.8	9.1
Ostrog.	5.2	2.7	8.8	0.3	5.5	2.6	8.7	5.4	4.4	3.2	8.4	1.7
Kame.	0.6	0.3	0.2	2.7	2.6	0.4	1.1	3.5	1.3	0.2	0.3	3.0
Pavlo.	8.7	2.9	5.0	7.9	6.3	3.6	2.5	7.4	8.1	1.9	5.4	7.4
Buturl.	0.5	1.0	0.1	1.6	2.0	1.4	0.9	3.6	1.5	1.4	0.8	2.2
Khokh.	5.6	2.6	8.9	8.5	4.2	0.5	9.8	10.9	6.5	0.6	8.6	9.6
Annin.	3.1	0.5	6.9	5.3	3.7	0.7	8.9	7.8	3.2	0.7	5.7	7.8
Panin.	3.9	1.2	9.0	5.0	3.8	1.3	10.4	7.3	4.3	1.3	9.1	6.7
Nizhne.	13.4	2.5	10.7	8.1	11.3	2.8	11.9	12.0	12.5	3.3	10.8	9.2
Semi.	3.9	0.1	6.5	1.5	1.9	1.0	7.8	4.8	3.3	0.5	5.6	3.1
Ramo.	3.1	7.2	7.6	9.1	2.2	6.2	9.0	9.8	3.8	7.6	7.1	11.4
Grib.	7.0	4.6	7.0	5.8	8.0	4.5	7.5	5.7	7.3	5.8	5.4	5.6
Terno.	7.8	5.4	6.2	3.6	7.4	4.4	6.6	5.6	8.8	6.0	5.2	4.0

Table 3. Absolute errors of barley c/h for 2011–2014 years

District Name	Absolute Error											
	MLR				Lasso				SVM			
	2011	2012	2013	2014	2011	2012	2013	2014	2011	2012	2013	2014
Kash.	6.1	3.4	5.7	2.8	5.9	2.9	5.4	2.5	6.6	2.1	3.8	1.6
Talov.	3.3	1.4	1.5	5.6	2.9	1.3	1.0	5.0	3.2	1.0	0.5	4.1
Novokh.	4.3	3.5	2.4	0.9	4.6	3.6	2.1	1.3	5.9	2.2	2.9	1.2
Verkhm.	8.1	0.3	2.0	3.4	8.0	1.1	2.7	3.7	8.2	0.9	2.7	2.3
Kalach.	10.0	5.9	0.1	1.2	9.5	6.4	1.0	1.1	9.7	4.6	0.2	1.7
Vorob.	15.3	6.6	7.3	3.9	14.5	6.3	7.2	3.2	14.9	4.6	5.3	2.0
Podgo.	10.0	4.8	0.5	3.0	10.0	4.0	0.0	2.7	9.7	3.2	1.2	3.5
Liski.	8.6	4.5	3.7	6.0	8.1	4.4	3.2	6.3	5.7	3.6	1.8	4.0
Repy.	1.4	5.5	4.4	8.6	0.7	5.0	4.8	8.1	0.4	4.0	6.2	6.3
Bobro.	7.9	0.7	2.4	8.1	7.4	0.6	2.9	7.7	7.1	0.0	4.7	5.6
Ertil.	1.7	0.6	1.1	15.2	1.7	0.4	1.1	14.2	3.1	0.3	1.7	11.2
Olkho.	9.5	7.7	3.8	0.7	8.6	7.1	4.4	1.2	7.7	5.7	5.5	0.6
Povo.	8.0	2.3	3.0	4.2	7.4	2.2	3.4	3.6	8.0	1.1	2.7	3.9
Novousm.	3.7	5.3	0.8	4.2	3.7	4.6	0.6	4.2	5.3	4.5	0.0	2.4
Verkhn.	0.1	11	6.3	6.4	0.2	10.1	5.5	5.6	0.6	7.7	3.8	2.8
Ostrog.	2.5	5.5	1.0	2.4	2.2	4.3	0.3	2.8	1.1	1.7	1.6	1.4
Kame.	7.1	4.6	5.2	4.3	6.7	4.3	5.2	4.2	7.2	3.6	6.1	2.1
Pavlo.	12.8	1.5	7.7	5.0	12.0	1.9	8.1	4.8	10.4	1.0	7.2	3.1
Buturl.	10.0	2.4	6.8	3.5	9.6	2.5	6.8	3.3	10.2	1.3	4.9	2.6
Khokh.	0.2	2.0	7.5	7.2	0.8	1.9	6.6	6.7	2.1	0.8	4.5	4.2
Annin.	4.6	6.6	1.2	7.7	4.2	5.7	1.0	7.0	3.3	3.7	0.3	4.9
Panin.	8.4	12.5	5.2	4.7	8.6	12.0	4.5	4.3	8.8	10.2	2.7	3.0
Nizhne.	1.5	8.7	1.1	6.2	1.8	7.9	0.9	6.2	2.9	5.9	1.5	4.1
Semi.	1.4	3.7	3.8	4.9	0.5	3.4	3.2	4.6	1.4	1.1	0.9	3.2
Ramo.	2.6	3.5	3.2	5.4	2.3	2.7	2.5	6.0	3.0	1.3	0.2	6.4
Grib.	0.4	1.1	0.1	3.5	0.2	0.7	0.4	3.3	0.1	1.2	0.5	3.7
Terno.	2.5	0.8	5.8	2.8	2.1	0.5	5.4	2.4	1.3	0.1	4.5	0.8

Table 4 shows the performance of the yield prediction based on the different techniques including the MLR, Lasso, SVM. The first four rows correspond to the evaluation between predicted and observed yield made for that year and measured in RMSE (c/ha). The last row is the average RMSE for above four years. The result shows that the SVM model has the advantage of yield prediction in each year for barley, while for wheat Lasso model advantage for 2011, 2012 years. The other years shared between MLR and SVM Models. In general, the SVM outperforms competing techniques significantly for both wheat and barley. However, the average RMSE of the SVM (in barley) has a ~14.09% and ~18.59% reduction of RMSE from the Lasso and

MLR, respectively. While (in wheat) it has a ~1.23% reduction from the Lasso and ~5.29% from MLR.

Table 4. Model performance of wheat/barley yield prediction measured in RMSE

Year	Wheat			Barley		
	MLR	**Lasso**	**SVM**	**MLR**	**Lasso**	**SVM**
2011	5.46	4.91	5.27	6.95	6.63	6.63
2012	4.21	3.88	4.16	5.30	4.93	3.77
2013	6.42	7.08	6.08	4.22	4.07	3.57
2014	6.56	7.78	6.84	5.67	5.37	4.05
Avg	5.66	5.91	**5.59**	5.54	5.25	**4.51**

Also, to show and compare the spatially correlated errors for each technique, in the plot of error map from 2011to 2014 years for both wheat and barley in Figure 6 and Figure 7, the color represents the prediction error in the c/h. The error maps showed that most of the prediction absolute error is less than 10%.

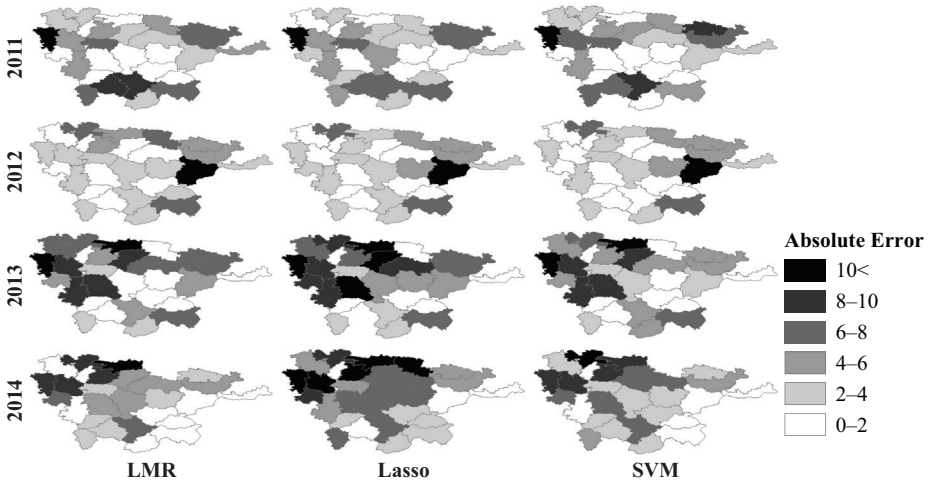

Fig. 6. Error maps at the district level from 2011 to 2014 for wheat yield

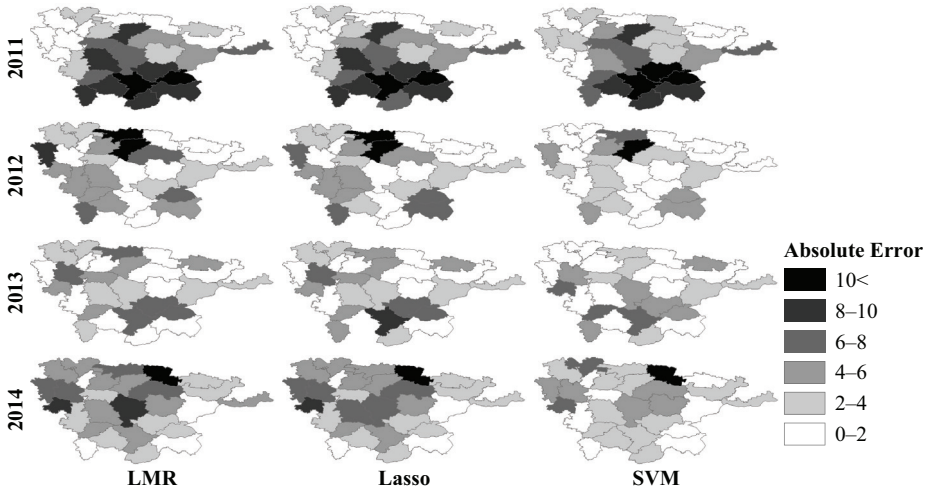

Fig. 7. Error maps at the district level from 2011 to 2014 for barley yield

The dark color in Figure 6 and Figure 7 means high error and vice versa. The prediction error attributes to many factors, such as weather, fertilization conditions, disease, and pests. Also, according to [3] the state subsidies to support agricultural producers varied from year to year. At the same time, starting from 2008, financing for the purchase of mineral fertilizers was unstable, which may affect the production and thus the prediction. In general, Figure 6 and Figure 7 show that the proposed model produces more low absolute errors (<8) across all years and techniques for both yields. Also, the high absolute errors always happen in the same regions. This refers to the stability of the proposed model.

5 Conclusions

This paper presents a framework for crop yield prediction by using remote sensing data. A model was proposed for effective learning representation for object-oriented crop yield prediction, and successfully learn much more effective features from NDVI image series derived from multispectral satellite images. The model is also including the strategy of integrating remote-sensing data with the yield time series analysis to make yield prediction more robust and accurate. The results of R^2 and RMSE were shown that yield trends in combination with NDVI are the best predictors for yield prediction. The proposed model consisted of two-level: in the first level the dimensionality reduction approach based on the FSR technique was employed, whereas the second level consist of yield prediction with MLR, Lasso and SVM techniques. In comparing the MLR, Lasso, and SVM techniques, the SVM in combination with FSR showed a robust method for grain yield prediction in terms of overall accuracy; while MLR and LASSO regression yield similar results. The results also show that the proposed model is stable and produces a high rate of the low absolute errors on the district-level across all years and the used techniques.

6 Acknowledgments

Our sincere thanks to Professor S.M. Abdullaev (South Ural State University), for providing advice to support the work.

7 References

[1] M. M. Alam, M. R. Hasan, M. R. Islam, M. H. Habaebi, A. Basahel, and M. Singh, "Prediction of Time Diversity Gain – Comparison Between ITU-R P.618–13 Using a Concept of Rain Rate with Delay and Synthetic Storm Technique," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 16, no. 11, pp. 178–192, Jun. 2022. https://doi.org/10.3991/ijim.v16i11.30103

[2] M. S. Mkhabela, P. Bullock, S. Raj, S. Wang, and Y. Yang, "Crop Yield Forecasting on the Canadian Prairies using MODIS NDVI data," *Agricultural and Forest Meteorology*, vol. 151, no. 3, pp. 385–393, 2011. https://doi.org/10.1016/j.agrformet.2010.11.012

[3] S. E. Alexandrovich, "Improvement Of Methods For Prediction Of Yield Of Grain Crops," Ph.D. dissertation, Voronezh State Agrarian University named after Emperor Peter I, Voronezh, 2015. [in Russian].

[4] C. J. Weissteiner and W. Kühbauch, "Regional Yield Forecasts of Malting Barley (Hordeum Vulgare L.) by NOAA-AVHRR Remote Sensing Data and Ancillary Data," *Journal of Agronomy and Crop Science*, vol. 191, no. 4, pp. 308–320, 2005. https://doi.org/10.1111/j.1439-037X.2005.00154.x

[5] E. Panek and D. Gozdowski, "Analysis of Relationship Between Cereal Yield and NDVI for Selected Regions of Central Europe Based on MODIS Satellite Data," *Remote Sensing Applications: Society and Environment*, vol. 17, p. 100286, 2020. https://doi.org/10.1016/j.rsase.2019.100286

[6] F. Kogan, L. Salazar, and L. Roytman, "Forecasting Crop Production Using Satellite-Based Vegetation Health Indices in Kansas, USA," *International Journal of Remote Sensing*, vol. 33, no. 9, pp. 2798–2814, 2012. https://doi.org/10.1080/01431161.2011.621464

[7] F. Kogan, W. Guo, A. Strashnaia, A. Kleshenko, O. Chub, and O. Virchenko, "Modelling and Prediction of Crop Losses from NOAA Polar-Orbiting Operational Satellites," *Geomatics, Natural Hazards and Risk*, vol. 7, no. 3, pp. 886–900, 2016. https://doi.org/10.1080/19475705.2015.1009178

[8] F. N. Kogan, "Operational Space Technology for Global Vegetation Assessment," *Bulletin of the American Meteorological Society*, vol. 82, no. 9, pp. 1949–1964, 2001. https://doi.org/10.1175/1520-0477(2001)082<1949:OSTFGV>2.3.CO;2

[9] Z. H. Khalil and S. M. Abdullaev, "Diagnosis of Landscapes of the Province of Al-Diwaniyah (Iraq) by Using of Landsat-8 Multispectral Images," *Bulletin of the South Ural State University.Series: Computational Mathematics and Software Engineering*, vol. 7, no. 3, pp. 5–18, 2018. [in Russian]. https://doi.org/10.14529/cmse180301

[10] Z. H. Khalil and S. M. Abdullaev, "Neural Network Approach To Predict Winter Crop In Diwaniyah-Iraq Based Satellite Data," In *The Twelfth Scientific Conference of Postgraduate and Doctoral Students "Scientific Search" - Section: Natural Sciences* , South Ural State University (SUSU), Mach 2020, pp. 42–49.

[11] B. C. Reed, T. R. Loveland, and L. L. Tieszen, "An Approach for Using Avhrr Data to Monitor U.S. Great Plains Grasslands," *Geocarto International*, vol. 11, no. 3, pp. 13–22, 1996. https://doi.org/10.1080/10106049609354544

[12] T. Blaschke, K. Johansen, and D. Tiede, Object-Based Image Analysis for Vegetation Mapping and Monitoring," In Book *Advances in Environmental Remote Sensing*, 1st Edition, Q. Weng, CRC Press; 2011, p. 241–72. https://doi.org/10.1201/b10599-17

[13] N. Kamagata, Y. Akamatsu, M. Mori, Y. Q. Li, Y. Hoshino, and K. Hara, "Comparison of Pixel-Based and Object-Based Classifications of High Resolution Satellite Data in Urban Fringe Areas," In *Proceedings of the 26th Asian Conference on Remote Sensing,* Red Hook, NY, USA: Curran Associates, Inc., vol. 3, January 2005. pp. 1590–1595.

[14] C. Burnett and T. Blaschke, "A Multi-Scale Segmentation/Object Relationship Modelling Methodology for Landscape Analysis," *Ecological Modelling*, vol. 168, no. 3, pp. 233–249, 2003. https://doi.org/10.1016/S0304-3800(03)00139-X

[15] Y. Han, A. Javed, S. Jung, and S. Liu, "Object-Based Change Detection of Very High Resolution Images by Fusing Pixel-Based Change Detection Results Using Weighted Dempster-Shafer Theory," *Remote Sensing*, vol. 12, no. 6, 2020. https://doi.org/10.3390/rs12060983

[16] G. Chen, G. J. Hay, L. M. T. Carvalho, and M. A. Wulder, "Object-Based Change Detection Techniques," *International Journal of Remote Sensing*, vol. 33, no. 14, pp. 4434–4457, 2012. https://doi.org/10.1080/01431161.2011.648285

[17] D. O. Fuller, "Trends in Ndvi Time Series and their Relation to Rangeland and Crop Production in Senegal, 1987–1993," *International Journal of Remote Sensing*, vol. 19, no. 10, pp. 2013–2018, 1998. https://doi.org/10.1080/014311698215135

[18] Voronezh region. Statistical collection 2012, Voronezh, 2013. [in Russian].

[19] The Federal State Statistics Service (Rosstat). *Sites of territorial bodies of Rosstat*. [Online]. [in Russian]. Available: https://rosstat.gov.ru/. [Accessed: 15-May-2021].

[20] K. Didan, "MOD13Q1 MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V006." NASA EOSDIS Land Processes DAAC, 2015. https://doi.org/10.5067/MODIS/MOD13Q1.006

[21] M. Friedl and D. Sulla-Menashe, "MCD12Q1 MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V006 [Data set]," NASA EOSDIS Land Processes DAAC, 2019. https://doi.org/10.5067/MODIS/MCD12Q1.006

[22] M. D. Johnson, W. W. Hsieh, A. J. Cannon, A. Davidson, and F. Bédard, "Crop Yield Forecasting on the Canadian Prairies by Remotely Sensed Vegetation Indices and Machine Learning Methods," *Agricultural and Forest Meteorology*, vol. 218–219, pp. 74–84, 2016. https://doi.org/10.1016/j.agrformet.2015.11.003

[23] F. Kogan, W. Guo, W. Yang, and S. Harlan, "Space-Based Vegetation Health for Wheat Yield Modeling and Prediction in Australia," *Journal of Applied Remote Sensing*, vol. 12, no. 2, p. 26002, 2018. https://doi.org/10.1117/1.JRS.12.026002

[24] J. Huang, X. Wang, X. Li, H. Tian, and Z. Pan, "Remotely Sensed Rice Yield Prediction Using Multi-Temporal NDVI Data Derived from NOAA's-AVHRR," *PLoS ONE*, vol. 8, no. 8, pp. 1–13, 2013. https://doi.org/10.1371/journal.pone.0070816

[25] F. Kogan, L. Salazar, and L. Roytman, "Forecasting Crop Production Using Satellite-Based Vegetation Health Indices in Kansas, USA," *International Journal of Remote Sensing*, vol. 33, no. 9, pp. 2798–2814, 2012. https://doi.org/10.1080/01431161.2011.621464

[26] S. Pollock, "Trend Estimation and De-Trending," *Optimisation, Econometric and Financial Analysis.*, pp. 143–166, 2007. https://doi.org/10.1007/3-540-36626-1_8

[27] X. LV, "Remote Sensing, Normalized Difference Vegetation Index (Ndvi), And Crop Yield Forecasting," M. S. thesis, University of Illinois at Urbana-Champaign, Urbana, Illinois, 2013.

[28] A. A. Gurchenkov, A. B. Murynin, A. N. Trekin, and V. Y. U. Ignatiev, "Object-Oriented Classification of Substrate Surface Objects in Arctic Impact Regions Aerospace Monitoring," *Herald of the Bauman Moscow State Technical University, Series Natural Sciences*, no. 3, pp. 135–146, 2017. https://doi.org/10.18698/1812-3368-2017-3-135-146

[29] M. Almseidin, A. M. Abu Zuraiq, M. Al-kasassbeh, and N. Alnidami, "Phishing Detection Based on Machine Learning and Feature Selection Methods," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 13, no. 12, pp. 71–183, 2019. https://doi.org/10.3991/ijim.v13i12.11411

[30] A. S. Degtyarev, V. A. Drabenko, and V. A. Drabenko, *Statistical Methods for Processing Meteorological Information*. St. Petersburg: Andreevsky Publishing House LLC, 2015. [in Russian].

[31] A. R. Muhsen, G. G. Jumaa, N. F. AL Bakri, and A. T. Sadiq, "Feature Selection Strategy for Network Intrusion Detection System (NIDS) Using Meerkat Clan Algorithm," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 15, no. 16, p. 158, Aug. 2021. https://doi.org/10.3991/ijim.v15i16.24173

[32] A. Soroush, A. Bahreininejad, and J. Van Den Berg, "A Hybrid Customer Prediction System Based on Multiple Forward Stepwise Logistic Regression Mode," *Intelligent Data Analysis*, vol. 16, no. 2, pp. 265–278, 2012. https://doi.org/10.3233/IDA-2012-0523

[33] N. R. Draper and H. Smith, *Applied Regression Analysis*, Third Edition. Hoboken, NJ: John Wiley & Sons, Inc, 1998. https://doi.org/10.1002/9781118625590

8 Authors

Zena H. Khalil, Assistant lecturer at the department of Multimedia, College of Computer Science and Information Technology, University of Al-Qadisiyah. Research interests include image processing, biometrics and machine learning. https://orcid.org/0000-0002-8290-2707

Amel H. Abbas, PhD and Assistant professor at the department of Computer Sciences, College of Science, Mustansiriyah University. Research interests include image processing and image classification. https://orcid.org/0000-0002-6866-0422

Article submitted 2022-05-18. Resubmitted 2022-06-29. Final acceptance 2022-06-29. Final version published as submitted by the authors.