Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

Integrating Satellite and Climate Data for Crop Yield Prediction: Spatiotemporal Analysis and Neural Network based Model

Zena H. Khalil*

College of Computer Science and Information Technology, University of Al-Qadisiyah, Diwaniyah, 58002 Iraq.Email: Zena.khalil@qu.edu.iq

ARTICLE INFO

Article history:
Received: 25 /10/2025
Rrevised form: 09 /11/2025
Accepted : 16 /11/2025
Available online: 30 /12/2025

Keywords:

Artificial Neural Network,
Normalized Difference Vegetation Index,
Histogram learning
Diwaniyah-Iraq.

ABSTRACT

In Diwaniyah-Iraq province, which its economy considerably depends on crop production, the accurate crop harvest prediction is an elementary requirement. The present study has faced some challenges represented as 1) scarce available agricultural data such as the complete absence of cropland maps and also fewer crop health and crop yields statistics; and 2) interannual variation in climatic factors in such semi-arid regions. To overcome these impediments, a novel methodology that concentrates on data fusion and Artificial Neural Networks (ANN) was proposed in this work. The time-series satellite-based Normalized Difference Vegetation Index (**NDVI-MODIS**) was integrated with a multi-source climate database, which showed extra and unique information across the whole season. The strong relation of NDVI with the time of the critical period of the crop growth and its sensitivity to the variability of yield during the growing season provides a powerful feature for prediction. Exploratory data analysis was applied to reveal the temporal patterns of correlations between fused data streams with crop yield. Artificial Neural Network (ANN) was used to construct a winter yield prediction model at the province level, due to the predictive capability of neural networks that come from the hierarchical/multi-layered structure. Several ANN models were tested on 18 years of data for Diwaniyah-Iraq province. Critically, the suggested model enables a reliable yield prediction nine weeks before wheat harvest and seven weeks before barley harvest. Through the testing phase, the best models achieved the results of $R^2 > 0.9$ for both wheat and barley, with $0.01<RMSE<0.1$ for wheat and $0.06<RMSE<0.1$ for barley. The proposed methodology presents a robust and generalizable framework for yield prediction in circumstances of data-scarce, semi-arid environments and small-scale cultivation areas, offering early decision-making and resource management tools across Iraq.

MSC..

https://doi.org/10.29304/jqcsm.2025.17.42559.

1. Introduction

Iraq is one of the earliest farming countries in the world. Despite the fact that at present this sector is no longer the most important contributor to the Iraqi economy, agriculture and farming represent a significant source of livelihood there. The percentage of one-third of the overall population in Iraq lives in rural areas where their incomes depend on agricultural practices [1].

∗Corresponding author: Zena H. Khalil

Email addresses: Zena.khalil@qu.edu.iq

Communicated by 'sub etitor'

The cultivable area of the whole Iraqi region is around 26% and wheat and barley account for 70% to 85% of farmed crops [2], which is usually irrigated using rainwater or irrigated by the Tigris and Euphrates rivers in other areas.; and approximately 67% of wheat and barley comes from irrigation production. Fig. 1 shows the wheat and barley production in Diwaniyah-Iraq over the period (2001-2018).

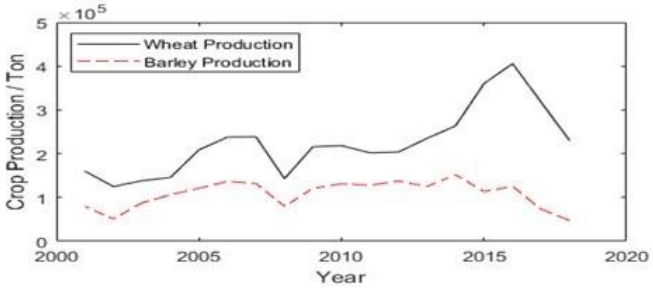


Fig. 1 - crop production in the study area from 2001 to 2018

This work was conducted using different combinations of data sources at different gathering dates during growing seasons. The challenging in crop yield prediction is to get accurately estimation at a higher spatial resolution, for instance territory, province, county or even domestic level. However the researchers in such situation tend to use very high resolution imagery such as Sentinel-2 (10–20 m resolution) [2]. The disadvantages of high-resolution satellite imagery that the requirement of time-consuming and extensive computation due to comprising massive data. Also, the limitation of temporal duration of observations results in a reduction of experiential models' accuracy. In this study, the high temporal Moderate Resolution Imaging Spectroradiometer sensor (MODIS) was used, given its ability to provide land cover maps, vegetation data and crop growth monitoring and estimation. MODIS has narrow spectral bands minimize the effect of water vapor absorption (WVA) in the band of Near-infrared NIR and sensitivity of the red band to chlorophyll, thereby providing a good quality NDVI product. This make NDVI a good sourse to monitor crop growth and progressively capture the crop yield variability within the growing season, and their correlation to yield prediction [3][4]. According to study [5], the grain formation course is sensitive to climate variability not only in the growing season but also during the months before it. for that, climate variables collect for all year months and using 2 data set to check as most as potential climate variables.

Traditionally, the yield prediction was formulated using statistical methods, but these methods have shown limitations in situations where many factors and relationships must be considered [6]. The complicated factors and relationships related to the prediction of yield were handled using artificial neural networks (ANN) in this work.

The work aimed to answer the following questions : (i) How will the combination of satellite and climate data affect the wheat and barley prediction in the irrigated area? (ii) Can it achieve accurate yield prediction at the province level? (iii) Is the histogram learning approach sensitive to the spatial resolution of image data? (iv) How do the bin sizes of the histogram affect prediction results?

The remainder of this paper is structured as follows: Section 2 provides a detailed description of the study area, the datasets and preprocessing, spatiotemporal data analysis, histogram learning approach and the artificial Neural network. Section 3 presents and discusses the experiments and results for both wheat and barley. Finally, Section 4 summarizes the conclusion and highlights limitations and directions for future research.

2. Materials and methods

2.1. Study area

The study area is the province of Diwaniyah-Iraq. It is located in the middle south of Iraq between Latitudes (30.5 ° - 32.5°)N, and Longitudes (44.5° - 46°) E as shown in Fig. 2-a. The predominant climate characterized by hot and dry summers and cold winters with limited rainfall [7]. The geological formations consist of sedimentary plain, Al-Dalmaj marsh, flat sand land and the western plateau. The sedimentary plain constitutes the majority of the province, with roughly 77%. This region is most suitable for farming, particularly the cultivation of grain and crops [8]. This work focused on the main wheat and barley growing areas in Diwaniyah province, where the growing season for these crops typically extends from November to March.

2.2. Datasets and preprocessing

The data used for analysing and modelling in this work was gathered from multiple sources, comprised climate and satellite data, wheat and barley annual yield data and province boundary data, which are represented in detail in Table 1. The study period in this work was restricted from 2001 to 2018, according to the availability of data.

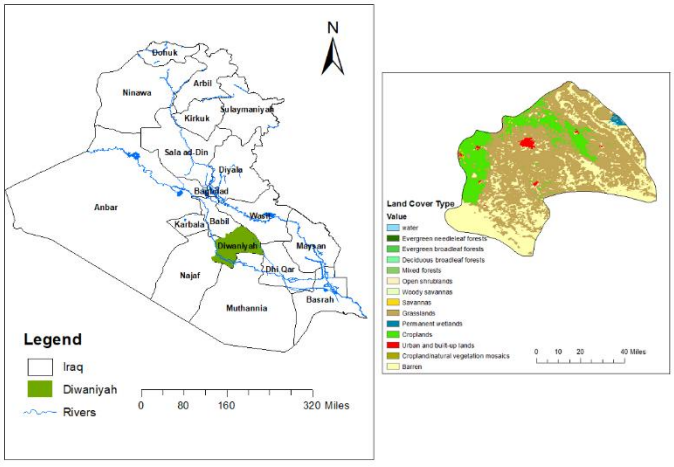


Fig. 2 - Study area: administrative map of Iraq, position of Diwaniyah highlighted by green; to the right, MODIS land cover classification with a spatial resolution of 500 m for 2015 year.

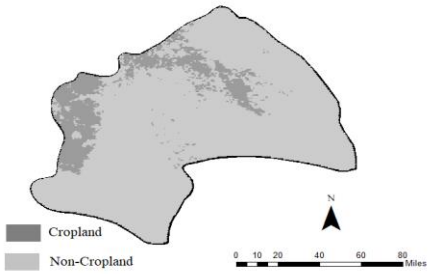
2.2.1. Province boundaries and Yield

First of all the shapefile of Diwaniyah province was provided by Ministry of Planning in Iraq (MOP), while Ministry of Agricultural in Iraq (MOA) provided us with wheat and barley annual crop yield data across Diwaniyah province throuth 2001 – 2018 in the unit of (kg / dunum).

2.2.2. Satellite data

The Terra Moderate Resolution Imaging Spectroradiometer (MODIS) represents a satellite sensor introduced by the National Aeronautics and Space Administration (NASA) specialist for monitoring Earth environment which captures images of the whole Earth's surface every 1–2 days. The MODIS acquire the data in 36 spectral bands at varying spatial resolutions, which the interpretation of global dynamics incident on land, oceans, and the lower atmosphere [9].

MODIS land cover (MCD12Q1) [10] image provides global land cover types derived from six different classification schemes employed with different methodologies at yearly intervals from 2001 to now; in this study, 18 images of (MCD12Q1) throughout the period of 2001-2018 were downloaded and the global vegetation classification scheme of the International Geosphere Biosphere Programme (IGBP) was used Fig. 2, because (IGBP) consistency is the best among the all five schemes of (MCD12Q1) [11].



However, The (IGBP) schem ; eleven classes for natural vegetation, three classes for developed and tated lands. The shapefile of Diwaniyah p n

Fig. 3 - The cropland area in Diwaniyah for 2015 based MODIS land cover/ IGBP classification schemes

extract the pixels which were recorded as cropland (Class = 12). So, the land cover images are only used to create a cropland area mask in Diwaniyah of each year in the study period, which used after then to construct a database of satellite images and climate variables from 2001 to 2018. Fig. 3 shows the distribution of the cropland area in Diwaniyah for 2015 year.

Table 1 - Summary of the dataset used in this study

Data		Spatial Resolution	Temporal Resolution	Source
Remote Sensing	NDVI	250 m	16 days	NASA MODIS / MOD13Q1 product
	Land Cover	500 m	yearly	NASA MODIS/ MCD12Q1 product
Climate	Precipitation, diurnal temperature range, minimum temperature, mean temperature, maximum temperature, vapour pressure, wet day frequency, potential evapotranspiration,	0.5 degree	monthly	CRU
	solar radiation, minimum speed, maximum wind speed	0.5 degree	monthly	NASA POWER
Crop Yield Statistics	Wheat Barley	province	Yearly	Ministry of Agricultural in Iraq(MOA)
Shape file	Al-Diwaniyah boundary	-	-	Ministry of Planning in Iraq (MOP)

In this work, MODIS-NDVI data was utilized in order to predict the yield of wheat and barley in study area; due to the success achieved by MODIS-NDVI not only in crop yield prediction and estimation but also in various studies concern with crop such as crop mapping [12], land use/land cover change [13] and crop classification [14].

The images of the MODIS-NDVI product of MOD13Q1 [15], were acquired from USGS Earth Explorer user interface which introduced by the United States Geological Survey, across sixteen day out of growing season extends from November to May consequent in 12 NDVI images were obtained for each year and 216 NDVI images for the study duration. These images were scaled to a (factor of 0.0001) followed by extracting the cropland pixels from each NDVI image by corresponding cropland shapefile of the same year.

2.2.3. Climate Data

Climate variables were used in this work as previously demonstrated in Tab 1. The climate parameters, inclusived Precipitation (pre) in millimetres per month, diurnal temperature range (dtr), monthly average daily minimum temperature (tmn), daily mean temperature (tmp) and monthly average daily maximum temperature (tmx) in degrees Celsius, vapour pressure (vap) in hectopascals hPa, wet day frequency (wet) in days, potential evapotranspiration (pet) in millimetres per day were obtained from Climatic Research Unit (CRU) [16] ; while the average amount of solar radiation (SR) incident on a horizontal surface in (MJ/m^2/day) , minimum wind speed (minws) and maximum wind speed (maxws) in m/s were obtained by NASA worldwide (POWER) Agroclimatology [17].

All climate variables were downloaded in the period of 2001-2018 as monthly averaged for all months before growing season and during growing season, then these variables masked based on the crop areas.

2.3. Spatiotemporal data analysis

Prior to applying the learning phase, statistical analysis of the data must be performed to select the most appropriate independent input variables as well as dimensionality reduction. In this work an exploratory data analysis was carried out to specify the best potential inputs in addition to temporal analysis which specify the best time of year to accurately predict estimate; by measuring correlations between wheat and barley yields and relative

vegetation health based on monthly anomalies in pixel-level of NDVI. The temporal analysis was done by computing monthly NDVI anomalies for 2001-2018, and correlate the anomalies with crop yield anomalies in Diwanyah provincy. These performed spatially for every cropland pixels in Diwaniah province by computing the NDVI monthly averages and climatologies to finaly find monthly NDVI anomaly images as listed below and represented in fig. 4:

1- First step was to compute maximum-value composites (MVCs) for each month period, which represents the maximum(highest) value on a pixel-by-pixel basis was found for each pixel location over the two extracted cropland images in the same month. This resulting in six NDVI images in each year of study period.

$$NDVI_{MVC} = \max (NDVIs)_{i,m,y} \quad (1)$$

Where, subscript i,m,y represent pixel, month, year indices.

2- Second computed NDVI climatology per pixel which defined as the average NDVIMVC images over 2001–2018 for each month and pixel, resulting in one NDVI image for each year of study period.

$$NDVI_{Climatology} = \frac{1}{N_y} \sum_{y=1}^{Ny} (NDVI_{MVC})_{i,m,y} \quad (2)$$

Where, N_y represents study period and subscript i,m,y represent pixel, month, year indices.

3- Next, the NDVI climatology was spatialy subtracted from each NDVIMVC composit images for every pixel-by-pixel, resulting in the six monthly NDVI anomaly images.

$$\left(NDVI_{anomaly}\right)_{i,m,y} = (NDVI_{MVC})_{i,m,y} - \left(NDVI_{Climatology}\right)_{i,m,y} \quad (3)$$

Where, subscript i,m,y represent pixel, month, year indices.
Then each NDVI anomaly image pixels were aggregated using a zonal mean operation to result in 6 anomaly timeseries, each one corresponding to one month in growing season.

4- Finally, anomaly wheat/barley yields were also computed by subtract average yield over study period from the yield in each year.

$$Yield_{anomly} = Yield_y - \overline{Yield} \quad (4)$$

Yieldy represents the yield in specific year, $\overline{Yield}$ represents average of yield.

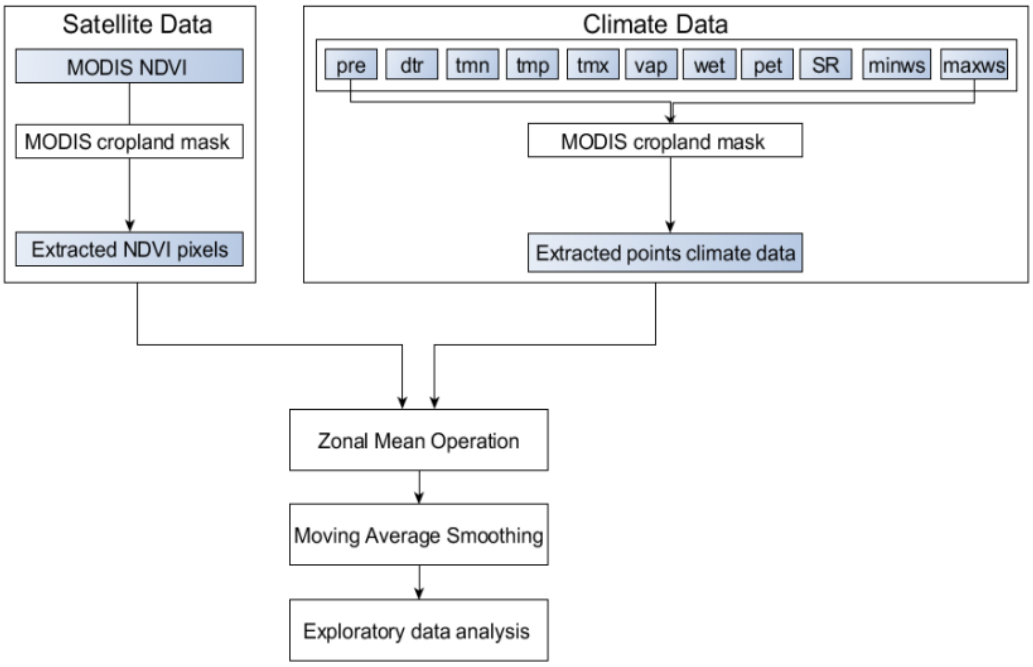


Fig. 4 - The Spatiotemporal data analysis

The NDVI anomaly and yields timeseries were smoothed to reduce noise in the data two point moving average to be then feeded into exploratory data analysis. Exploratory data analysis of satellite data was done by calculated Spearman correlation between the each NDVI anomly timeseries and the yield timeseries of each wheat and barley, then the correlation results in each month during the growing season was used to determine which month of the growing season (with high correlation value) its data will be used as independent variables.
The temporal analysis of correlations results between the anomly NDVI and anomly wheat/barley yields are shown in Fig. 5. These results showed that the peak was around February.

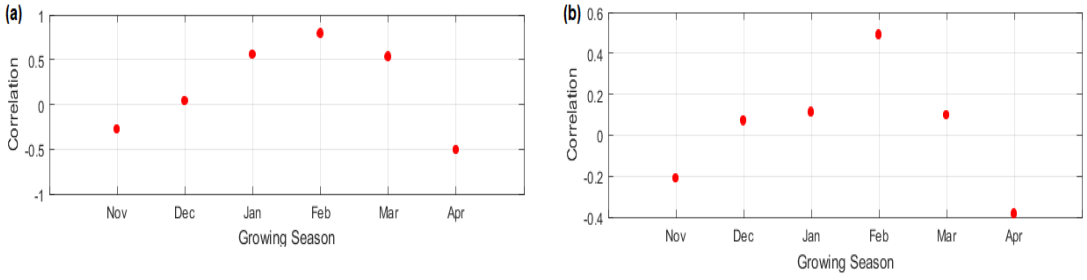
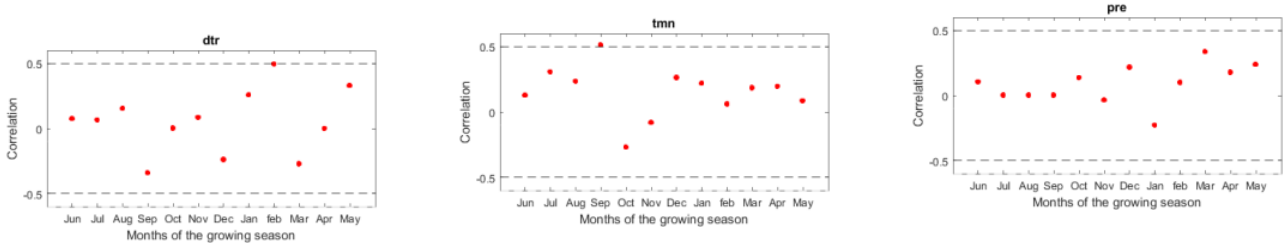


Fig. 5 - Temporal correlation between satellite data and yields, (a) for wheat (b) for barley

Focusing on these results of temporal correlation, noticed that the peak month for both yield was Februray, So the NDVI data in that month have been used as satellite input data in prediction model.



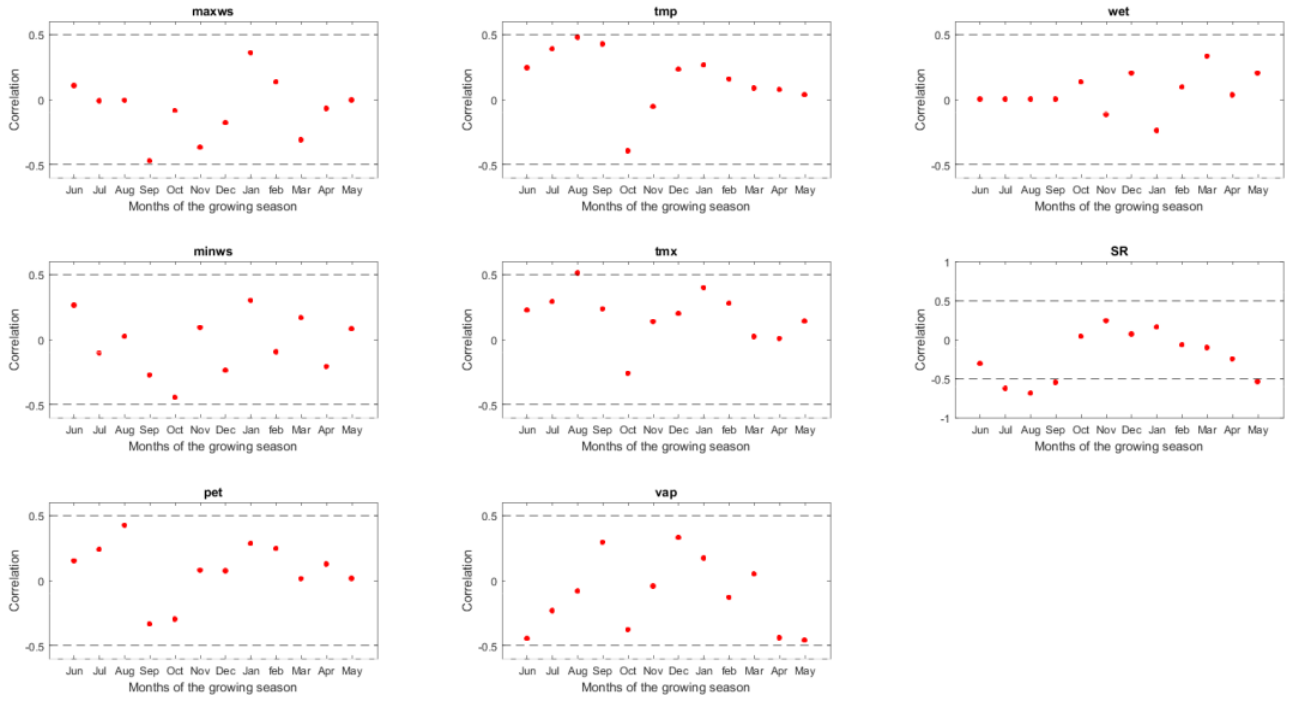
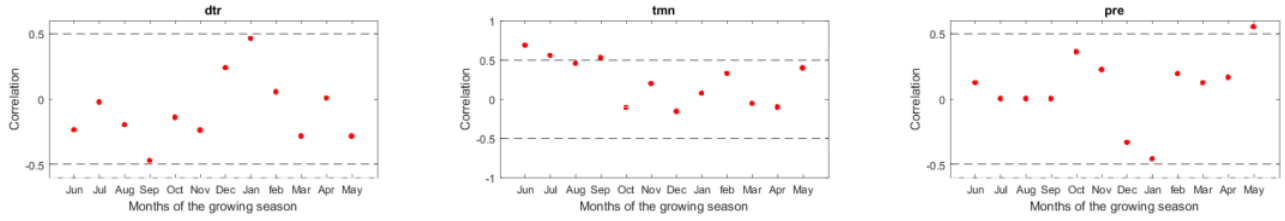


Fig. 6 - The temporal analysis of climate data and wheat yield

For climate data, all the gridded climate variables were aggregated using a zonal mean operation to construct 11 X 12 climate timeseriese, each one corresponding to one variable in one month in the year. The climate timeseries were also smoothed also by 2-point moving average and used in the same exploratory data analysis (see Fig. 6)by conducted Spearman correlation between the climate variable timeseries and yield statistics timeserise for each wheat and barley.

Finally, the climate variables with the absolute correlation values equal to or greater than 5 will be included as independent variables, while the variables with no significant correlation were excluded.

Fig. 6 and Fig.7 showed the results of the temporal analysis for climate data, which determined which climate variable had affection on each yield and in which month. However, the lead time must still be long enough for the prediction to be useful and therefore to ensure a significant lead time, only the temporal data analysis results before March were considered. This resulting in 8 climate variables were considered for wheat as (dtr in Feb; maxws in Sep; SR in Jul, Aug, Sep; tmp and tmx in Aug) and 17 climate variables for barley as (dtr in Feb; pet in Jul,Sep,Oct; SR in Oct, Nov, Dec, Jan; tmn in Jun, Jul, Aug, Sep; tmx in Jul, wet in Jan). The number of climate variables which considered in the barley model are grater than those considered for wheat, this agree with results of multivariate regression analyses we held in work [5], where found that the impact of climate variables effect 63.3% of the barley yield incontrast 50.9% of the wheat yield effected by the climatic variables.



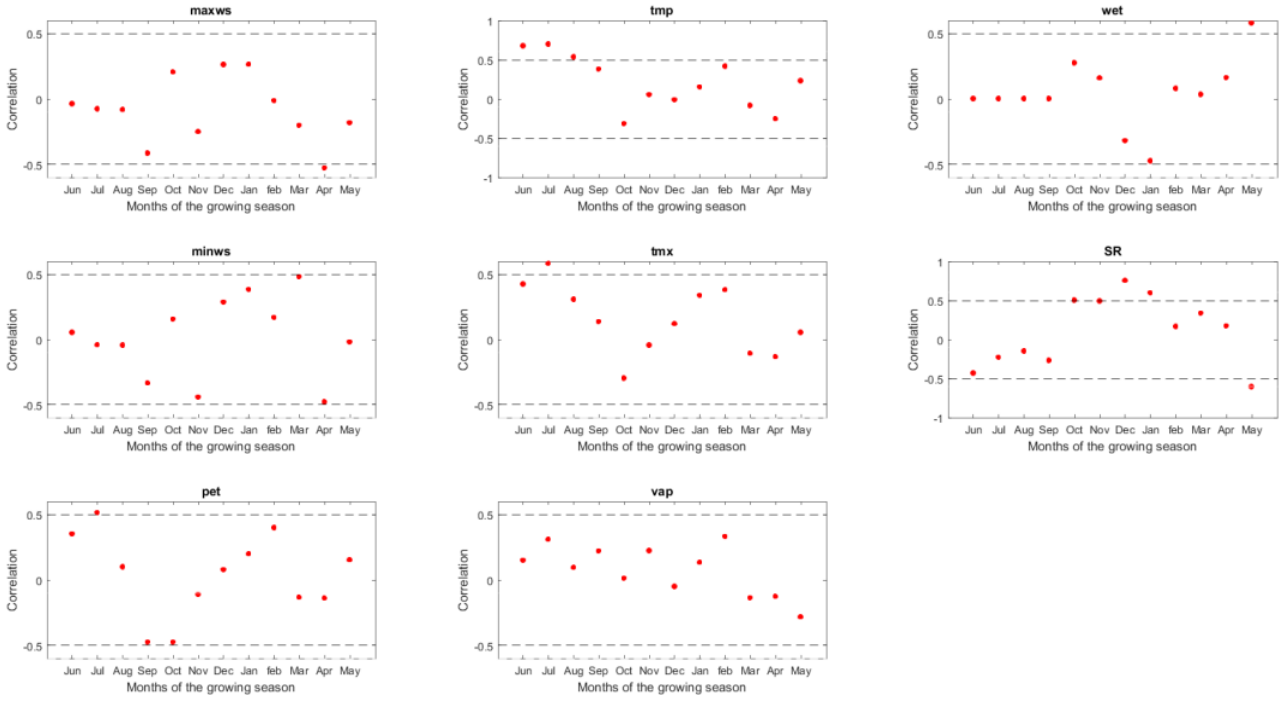


Fig. 7 - Temporal analysis of climate variables and barley

2.4. Histogram learning approach

As mentioned previously, the exploratory analysis resulted in data reduction from 216 NDVI images into only 36 NDVI images; i.e. only captured images in February months along period study, as two images for each year along study period.

Then February NDVI images were segmented into blocks, for each block NDVI and climate data were spatially accumulated to act as features for training of the Neural Network for future prediction. These features which represent the accumulation of all of the NDVI DN pixel values in each region(block) of an image is also needs further processing and representation in a new data structure. That's to combine them into higher-order features which reducing noise effects.

One method of efficient image representation is to use a histogram technique which has been used concerning many remote sensing applications such as land cover classification, land change detection, and crop yield prediction [18][19][20]. The benefits provided by histogram for image description include simplicity in producing, invariant to image translation/rotation, fast querying and most important of the ability to supply more information from the pixels' spatial distribution.

The NDVI data have a potentially long range from -1 to 1 resulting in a histogram with very large quantization levels reach over 5000 bins. To reduce the noise affection on prediction results, the original quantization levels (i.e., the number of histogram bins) were reduced to a specific quantization level. NDVI features based on histogram were performed with two quantization levels 5-bit quantization (32 bins) and 6-bit quantization (64 bins).

Many research represent remote sensing DN raw data as a histogram of pixel counts [20][21][22], but in this work NDVI DN data represented as 2D-histogram include bins of pixel counts and the average of the DN values for each bin. Therefore for each year, 1st week and 3rd week NDVI images in February were utilized, so 3D-histogram for each year exported consist of [time steps X bins no X 2], for each block and 18 years.

In total, climatic and 3D-histogram construct 12,240 wheat features with 32 bins and 23,760 features with 64 bins. Similarly, for barley used 13,050 features with 32 bins and 24,570 with 64 bins.

These features are important for ANN training because with image division the training algorithm will be able to learn from region features rather than just global features, also this makes ANN able to identify and analyze significant features and less significant features in predicting the crop yield which gave better results.

2.5. Artificial Neural network (ANN)

Artificial neural networks defined as nonlinear models with capabilities of capturing nonlinear relationships of data which developed to simulate human brain function and ability to adapt and change as it learns [23].
In this work, a multi-layer neural network was utilized with one hidden layer of neurons (Fig. 8). In order to design the best ANN model and determine the optimum feed forward network architecture, the Grid Search method was used, so the number of histogram bins (32, 64), the numbers of neuron in the hidden layer (2,4,6,8,10), the training algorithm and hidden transfer function were varied systematically and tried one by one. The hyperbolic tangent and logistics activation functions are employed for the hidden layer while the linear activation function is used for the outputs. Two training algorithms are used including Levenberg-Marquardt backpropagation and Scaled conjugate gradient. The learning rate was set to 0.01 and number of epochs was set to 300 epochs.
The input and target data had been normalized into range [-1,1] for efficient training and to get better performance of the ANN model by following equation:

$$\hat{Y} = (Y - min/max - min) \times (new\ max\ -\ new\ min) + new\ min \qquad (5)$$

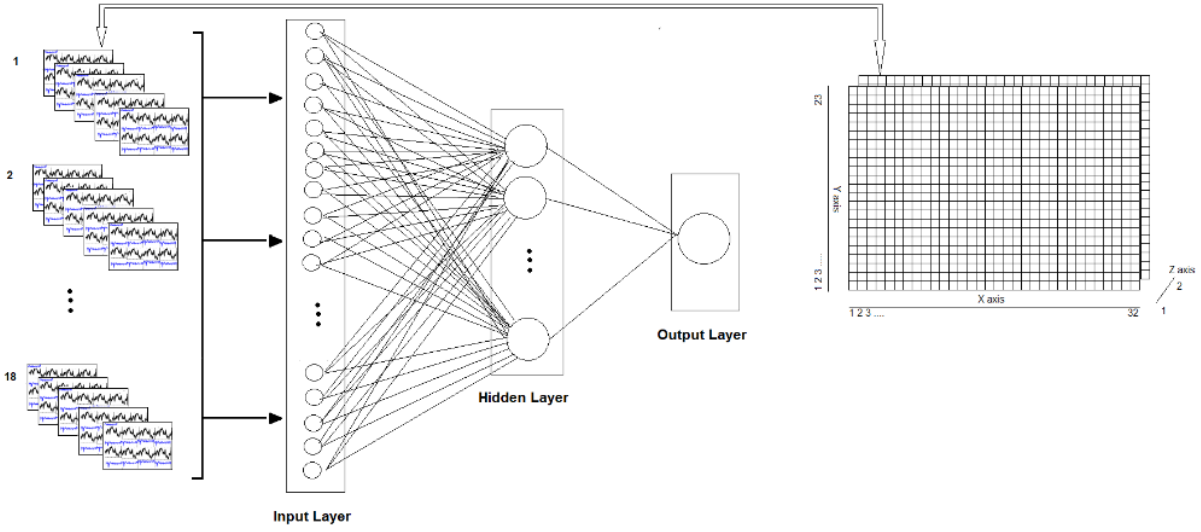


Fig. 8 - Model representation

3. Experiments and results

The ANN models and all analysis were implemented in MATLAB (R2017a) programming language version 9.2.0.538062, in order to evaluate each model's behavior.

The input data were divided into 60 % for training, 20% for validation and 20% for test data i.e. data of 10 years were used for training ANN and data of 4 years for both validation and testing.
The final step of most remote sensing data analysis studies is model validation. That's by comparing the measured and predicted values using common measures such as root mean squared error. So in this paper, validation methods was used considering coefficient of determination (R^2), root mean squared error (rmse) to compare the predicted and observed yield (see Tab. 2 & Tab.3).

Table 2 - prediction model for wheat

Training function	Transfer function in Hidden layer	Neurons no	32-bin						64-bin					
			Training		Validation		Test		Training		Validation		Test	
			RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
trainlm	logsig	2	0.120	0.925	0.065	0.874	0.090	0.812	0.193	0.768	0.094	0.889	0.173	0.946
trainlm	logsig	4	0.009	0.971	0.062	0.998	0.191	0.723	0.091	0.748	0.175	0.805	0.176	0.710
trainlm	logsig	6	0.001	0.997	0.133	0.706	0.120	0.884	0.000	0.977	0.109	0.670	0.228	0.822
trainlm	logsig	8	0.013	0.965	0.111	0.773	0.125	0.826	0.088	0.861	0.116	0.797	0.174	0.668
trainlm	logsig	10	0.025	0.983	0.139	0.823	0.158	0.957	0.020	0.755	0.194	0.840	0.174	0.822
trainlm	tansig	2	0.038	0.921	0.180	0.974	0.084	0.708	0.060	0.792	0.024	0.848	0.097	0.945
trainlm	tansig	4	0.067	0.936	0.127	0.971	0.128	0.960	0.013	0.934	0.172	0.711	0.056	0.938
trainlm	tansig	6	0.064	0.965	0.118	0.946	0.181	0.908	0.058	0.846	0.062	0.824	0.144	0.811
trainlm	tansig	8	0.001	0.838	0.045	0.990	0.220	0.740	0.044	0.739	0.033	1.000	0.169	0.754
trainlm	tansig	10	0.031	0.900	0.246	0.882	0.272	0.875	0.003	0.814	0.108	0.955	0.101	0.935
trainscg	logsig	2	0.095	0.919	0.050	0.861	0.089	0.845	0.102	0.909	0.204	0.902	0.057	0.961
trainscg	logsig	4	0.065	0.969	0.065	0.976	0.077	0.740	0.115	0.815	0.238	0.915	0.002	0.995
trainscg	logsig	6	0.020	0.992	0.026	0.853	0.053	0.954	0.080	0.866	0.173	0.933	0.148	0.835
trainscg	logsig	8	0.097	0.918	0.186	0.944	0.168	0.960	0.093	0.811	0.021	0.967	0.098	0.895
trainscg	logsig	10	0.044	0.874	0.196	0.820	0.131	0.927	0.036	0.905	0.169	0.978	0.151	0.906
trainscg	tansig	2	0.053	0.964	0.056	0.979	0.067	0.972	0.117	0.824	0.050	0.981	0.175	0.795
trainscg	tansig	4	0.088	0.948	0.053	0.973	0.053	0.914	0.076	0.855	0.123	0.842	0.015	0.980
trainscg	tansig	6	0.087	0.946	0.057	0.974	0.077	0.912	0.127	0.735	0.015	0.977	0.018	0.994
trainscg	tansig	8	0.025	0.872	0.183	0.831	0.073	0.948	0.132	0.759	0.081	0.917	0.025	0.928
trainscg	tansig	10	0.077	0.954	0.193	0.765	0.186	0.932	0.097	0.752	0.016	0.973	0.030	0.803

It is clear that, generally, the models trained with 64 bins showed superior performance compared to those trained with 32 bins. This can be interpreted by that, the 64 bins caused in increasing the resolution of the ndvi histogram. This allowed the network to capture more appropriate, subtle and informative features of the wheat canopy distribution and leading to better predictions.

The results in the table 3 showed that the barley models overall are inherently less accurate than the wheat models. This may be occurred because barley is either influenced by more external factors or has data reflect more dispersion than wheat. than wheat.

For barley, the models trained with 64-bit are also shown to have superior results than models trained with 32-bit, as wheat. Where models with 64 bins have max R2 of 0.937, while the 32-bin models have max R2 of 0.932. Also, Barley required fewer hidden neurons than wheat to achieve maximum accuracy, this indicate that the barley's yield function is less complex than wheat's.

Table 3 - ANN prediction model for barley

Training function	transfer function in Hidden layer	Neurons no	32-bin						64-bin					
			training		validation		test		training		validation		test	
			RMSE	R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE	R2
trainlm	logsig	2	0.125	0.869	0.135	0.900	0.164	0.812	0.049	0.975	0.061	0.972	0.162	0.934
trainlm	logsig	4	0.055	0.963	0.156	0.967	0.104	0.893	0.049	0.937	0.290	0.761	0.149	0.662
trainlm	logsig	6	0.015	0.995	0.188	0.920	0.196	0.783	0.047	0.955	0.117	0.780	0.180	0.512
trainlm	logsig	8	0.001	0.976	0.078	0.764	0.168	0.847	0.180	0.726	0.391	0.839	0.230	0.843
trainlm	logsig	10	0.000	0.939	0.149	0.745	0.193	0.818	0.099	0.863	0.304	0.656	0.389	0.711
trainlm	tansig	2	0.004	0.921	0.118	0.851	0.189	0.749	0.144	0.714	0.069	0.986	0.112	0.826
trainlm	tansig	4	0.002	0.922	0.055	0.989	0.145	0.768	0.062	0.819	0.112	0.552	0.081	0.983
trainlm	tansig	6	0.161	0.797	0.224	0.764	0.069	0.951	0.019	0.898	0.288	0.916	0.249	0.734
trainlm	tansig	8	0.018	0.971	0.180	0.973	0.178	0.968	0.021	0.792	0.037	0.893	0.186	0.859
trainlm	tansig	10	0.155	0.854	0.104	0.881	0.256	0.608	0.055	0.806	0.231	0.541	0.219	0.605
trainscg	logsig	2	0.089	0.904	0.091	0.793	0.078	0.848	0.086	0.874	0.114	0.885	0.189	0.801
trainscg	logsig	4	0.097	0.805	0.219	0.659	0.122	0.983	0.055	0.924	0.127	0.893	0.162	0.745
trainscg	logsig	6	0.085	0.850	0.059	0.997	0.124	0.961	0.072	0.793	0.158	0.770	0.166	0.859
trainscg	logsig	8	0.007	0.939	0.147	0.916	0.161	0.911	0.008	0.962	0.078	0.789	0.123	0.912
trainscg	logsig	10	0.103	0.833	0.094	0.860	0.123	0.711	0.081	0.835	0.170	0.786	0.162	0.906
trainscg	tansig	2	0.151	0.703	0.143	0.745	0.103	0.945	0.134	0.851	0.196	0.811	0.200	0.763
trainscg	tansig	4	0.009	0.992	0.078	0.655	0.080	0.887	0.005	0.963	0.217	0.753	0.134	0.828
trainscg	tansig	6	0.007	0.990	0.089	0.802	0.116	0.751	0.012	0.931	0.130	0.836	0.275	0.833
trainscg	tansig	8	0.026	0.932	0.188	0.705	0.095	0.913	0.010	0.978	0.127	0.833	0.085	0.713
trainscg	tansig	10	0.015	0.986	0.274	0.913	0.073	0.981	0.010	0.955	0.114	0.884	0.275	0.945

Experiences were divided into 8 groups, depending on the number of neurons, number of bins, used training function and used transfer function in the hidden layer. Then the best models (see Tab. 4) were choosed for further comparative analysis using relative error RE (%) and cross validation. The leave one year out cross validation (LOOCV) [24], which also known as (Jackknife), was implemented to examine the accuracies of the predicted yield by each model. The best models were frequently refit, leaving out one year observations, and then utilized to obtain a prediction for this left-out year. At result, the predicted crop yield for 18 years was provided, and this enabled to implement the comparative analysis of predicted yields aganist the historical yields for all 18 years. The accuracy of a prediction was measured by the magnitude of relative error RE (%), which is defined as the absolute ratio of the error to the actual observed value [25].

The LOOCV validation technique was conducted for further model evaluation due to its ability to generate more robust performance analysis across a limited sample size. Moreover, it ensures a more trustworthy assessment compared to other techniques because of its ability to reduce information loss and variability. Also, in connection with time series data, where the prediction goal in the future time points, it is essential to utilizied the cross-validation method to ensure the computation of dependable scoring rules [26][27].

Table 4 - The best accuracy models for wheat/barley prediction

Model	Wheat				Barley			
	Training function	Transfer function in Hidden layer	Neurons no.	Bins no.	Training function	Transfer function in Hidden layer	Neurons no.	Bins no.
M1	trainlm	logsig	10	32	trainlm	logsig	2	32
M2	trainlm	tansig	6	32	trainlm	tansig	6	32
M3	trainscg	logsig	6	32	trainscg	logsig	10	32
M4	trainscg	tansig	2	32	trainscg	tansig	2	32
M5	trainlm	logsig	2	64	trainlm	logsig	2	64
M6	trainlm	tansig	4	64	trainlm	tansig	4	64
M7	trainscg	logsig	8	64	trainscg	logsig	4	64
M8	trainscg	tansig	4	64	trainscg	tansig	8	64

The experimental results of relative errors (%) between the predicted and actual yields are presented in Tab. 5. From the results shown in Tab. 5, we can notice that the RE (%) between the predicted and actual yield reached as low as 0.001% for wheat and 0% for barley, which confirms the accuracy of suggested methodology. However, in some instances of wheat experiments, we can notice that the RE is greater than 15%, particularly for model 5 of wheat, which is indicate to the inadequacy of their structures. Also, the models 5&6 of barley experments showed high RE values referred to less efficient of these models.

From the work results we can answer all questions, where the suggested artificial neural network model with integrated satellite and climate data at province level achieved good result.

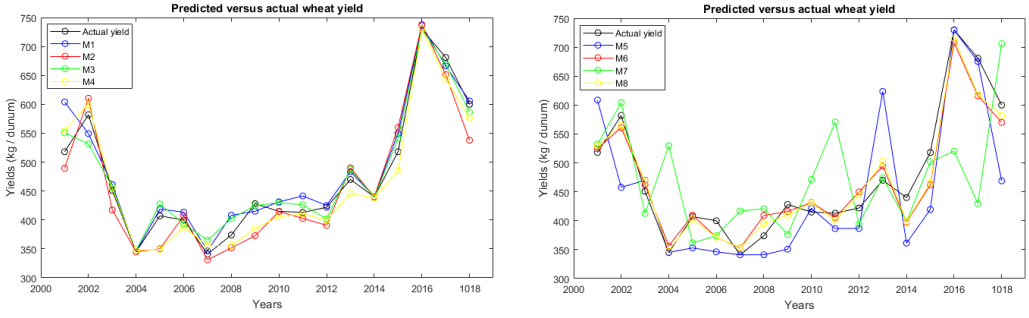


Fig. 9 - Graphical analysis of the predicted versus the actal wheat yield for every year (2001 to 2018) using different models, the left panel using 32bin histogram while right using 64-bin histogramm

For fifth question, the bin sizes of histogram had showed significant affection on prediction results, where increasing the no of bins caused in decreasing the accuracy of prediction.

A graphical analysis of the LOOCV results for best models for wheat and barley for all of the years from 2001 to 2018 are showed in Fig9 & Fig. 10. By looking to graphical analysis it is clear that the suggested technique give best results with wheat yields rather than barley yield, and this is perhaps due to that the bigger part of cropland area is dedicated to growing wheat while barley is grown in smaller areas.

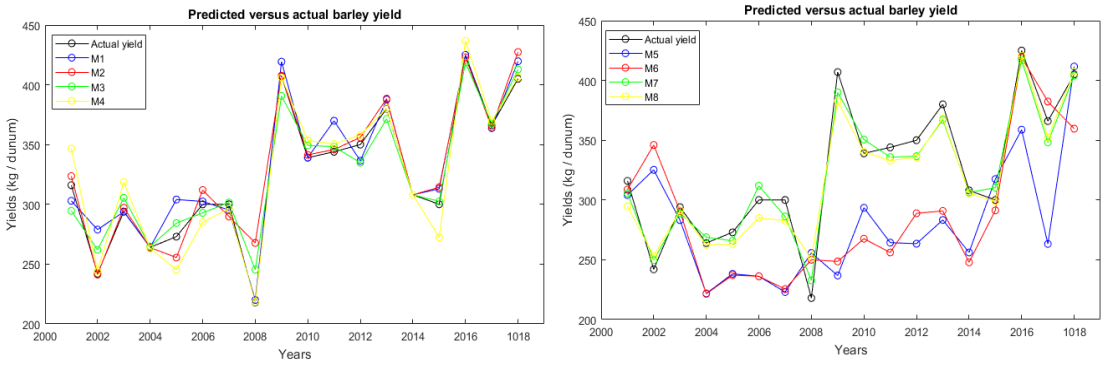


Fig. 10 - Graphical analysis of the predicted versus the actal barley yield for every year (2001 to 2018) using different models, the left panel using 32-bin histogram while right using 64-bin histogramm.

Table 5 - Comparsion of predicted and actual yields of best models using relative error RE(%) and LOOCV technique

Year	M1		M2		M3		M4		M5		M6		M7		M8	
	barley	wheat	barley	wheat	barley	wheat	barley	wheat	barley	wheat	barley	wheat	barley	wheat	barley	wheat
2001	4.13	16.61	2.41	5.51	6.77	6.28	9.71	7.09	3.75	17.48	2.32	1.29	3.28	1.9	6.73	2.22
2002	15.23	5.64	0.31	4.89	8.25	8.66	0.88	3.25	34.38	21.36	42.99	3.62	3.18	8.46	4.62	3
2003	0.07	2.27	1.07	7.44	3.91	1.3	8.36	0.74	3.69	4.2	1.12	2.55	1.75	0.61	0.96	4.03
2004	0.002	0.01	0.04	0.001	0.28	0.68	0.32	0.65	15.97	0.07	16.09	3.32	1.84	4.89	0.58	0.88
2005	11.37	2.9	6.41	13.95	4.13	5.1	10.15	14.22	12.73	13.26	13.14	0.66	2.62	14.18	3.61	1.12
2006	0.81	3.36	3.99	1.73	2.37	1.71	4.98	3.11	21.26	13.44	21.28	7.13	3.97	17.82	5	7.11
2007	1.6	1.65	3.4	2.92	0.52	7.07	1.21	4.33	25.66	0.002	24.85	3.43	4.6	10.21	5.76	3.62
2008	0.92	9.13	22.81	5.97	12.57	7.55	0.18	4.87	17.28	8.77	14.68	9.37	6.78	12.84	15.2	5.63
2009	3.03	3.01	0.16	12.9	4.02	0.73	0.16	9.95	41.81	18.01	38.91	2.82	4.17	14.52	6.02	4.28
2010	0.05	3.93	0.69	0.13	3.04	3.6	4.26	2.2	13.43	1.59	21.05	3.96	3.4	2.88	0.39	3.84
2011	7.52	6.93	0.55	2.48	1.22	3.17	1.88	0.65	23.16	6.43	25.55	1.83	2.37	12.44	3.23	2.97
2012	3.88	0.73	1.71	7.4	4.34	4.8	2.25	5.22	24.72	8.36	17.47	6.58	3.82	9.05	4.06	5.35
2013	2.02	2.87	2.21	4.26	2.25	3.8	0.08	5.14	25.44	32.66	23.46	5.23	3.41	5.96	3.07	7.04
2014	0.001	0.02	0.0004	0.001	0	0.31	0	0.39	16.84	17.89	19.52	9.68	0.54	15.58	0.77	9.4
2015	4.34	6.15	4.78	8.09	0.84	4.5	9.17	6.36	5.83	19.01	2.9	10.79	3.43	13.72	0.4	10.2
2016	0.003	0.83	0.46	1.05	1.52	0.01	2.79	0.06	15.58	0.05	1.35	3.03	1.89	7.47	1.01	2.07
2017	0.61	2.15	0.47	4.36	0.41	1.72	0.99	5.45	28.07	0.83	4.47	9.54	4.87	14.94	3.74	9.25
2018	3.62	0.96	5.52	10.37	1.93	2.24	0.24	3.97	1.63	21.83	11.2	5.01	0.26	11.36	0.56	3.03

4. CONCLUSION

In this work, an ANN model was presented to the timely prediction of winter crop yield in Diwaniyah province based on climate and satellite data. To select of most appropriate inputs, dimensionality reduction and results interpretation, an exploratory data analysis analysis was conducted before building the models.

The model processed satellite data by using the spatial accumulation of MODIS-NDVI data with histogram technique and introducing this method as a promising technique in yield prediction based remote sensing. The integrating satellite and climate data achieved good performance for crop prediction in irregated area, where climate information served as a perfect contribution to prediction model. The prediction models trained by ANN were able to extract significant features from high dimensional data. In addition, the suggested method of designing input data also affected the accuracy of models which developed by learning algorithms.

Moreover, the work demonstrated the conceivability to develop efficient yield estimation model using the suggested technique in such an area with fewer available agricultural data.

According to the experimental results, the accuracy of crop yield prediction using MODIS-NDVI and climate data was satisfactory. We only used the NDVI from the first and third weeks in February to make a good prediction about 9 weeks ahead of harvest time for the wheat yield and 7 weeks ahead of harvest time for the barley yield.

The limitation of this study was the restriction of study period, that is because unavailability of long-term historical crop yield data in the Diwaniyah-Iraq, due to war circumstance the region has been exposed to.

The future work includes employing use a deep leaning approach such as CNN model that has power ability to automatically learn efficient spatial or localized feature dependencies. This can lead to refine RMSE and accuracy, especially for the barley. Also, the proposed model can be applied to other crops and other areas.

References

[1] P. Lucani and M. Saade, “Iraq Agriculture sector note,” (2012).

[2] M. Burke and D. B. Lobell, “Satellite-based assessment of yield variation and its determinants in smallholder African systems,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 114, no. 9, (2017), pp. 2189–2194, doi: 10.1073/pnas.1616919114.

[3] A. R. Huete and H. Q. Liu, “An Error and Sensitivity Analysis of the Atmospheric and Soil-Correcting Variants of the NDVI for the MODISEOS,” *IEEE Trans. Geosci. Remote Sens.*, vol. 32, no. 4, (1994), pp. 897–905, doi: 10.1109/36.298018.

[4] A. Huete, K. Didan, T. Miura, E. P. Rodriguez, X. Gao, and L. G. Ferreira, “Overview of the radiometric and biophysical performance of the MODIS vegetation indices,” *Remote Sens*, vol. 83, no. Nov, (2002), pp. 195–213.

[5] Z. H. Khalil and Abdullaev.S.M., “Sensitivity of Winter Crops to climate variability in the irrigated subtropics of Iraq (Al-Diwaniyah),” *Procedia Comput. Sci.*, vol. 167, (2020), pp. 1066–1079, doi: DOI:10.1016/j.procs.2020.03.405.

[6] S. Conradt, R. Bokusheva, R. Finger, and T. Kussaiynov, “Yield trend estimation in the presence of farm heterogeneity and non-linear technological change,” *Q. J. Int. Agric.*, vol. 53, no. 2, (2014), pp. 121–140, doi: 10.22004/ag.econ.195732.

[7] Ministry of Agriculture, “Iraqi Agrometeorological network.” [Online]. Available: https://agromet.gov.iq/en/

[8] Ministry of Municipalities and Public Works, “Structural plan for Diwaniyah stage 5 report. Final report,” (2014).

[9] “NASA MODIS Home Page.” [Online]. Available: https://modis.gsfc.nasa.gov/about/

[10] Damien Sulla-Menashe and M. Friedl, “MCD12Q1 MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V006 [Data set],” NASA EOSDIS Land Processes DAAC. Accessed: Feb. 27, 2020. [Online]. Available: https://doi.org/10.5067/MODIS/MCD12Q1.006

[11] D. Liang, Y. Zuo, L. Huang, J. Zhao, L. Teng, and F. Yang, "Evaluation of the Consistency of MODIS Land Cover Product (MCD12Q1) Based on Chinese 30 m GlobeLand30 Datasets: A Case Study in Anhui Province, China," *ISPRS Int. J. Geo-Information*, vol. 4, no. 4, (2015), pp. 2519–2541, doi: 10.3390/ijgi4042519.

[12] X. Xiao *et al.*, "Mapping paddy rice agriculture in southern China using multi-temporal MODIS images," *Remote Sens. Environ.*, vol. 95, no. 4, (2005), pp. 480–492, doi: 10.1016/j.rse.2004.12.009.

[13] R. S. Lunetta, J. F. Knight, J. Ediriwickrema, J. G. Lyon, and L. D. Worthy, "Land-cover change detection using multi-temporal MODIS NDVI data," *Remote Sens. Environ.*, vol. 105, no. 2, (2006), pp. 142–154, Nov., doi: 10.1016/j.rse.2006.06.018.

[14] B. D. Wardlow, S. L. Egbert, and and J. H. Kastens, "Analysis of Time-Series MODIS 250 M Vegetation Index Data for Crop Classification in the U.S."

[15] K. Didan, "MOD13Q1 MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V006.," *NASA EOSDIS L. Process. DAAC.*, 2015, doi: 10.5067/MODIS/MOD13Q1.006.

[16] I. Harris, P. D. Jones, T. J. Osborn, and D. H. Lister, "Updated high-resolution grids of monthly climatic observations - the CRU TS3.10 Dataset," *Int. J. Climatol.*, vol. 34, no. 3, pp. 623–642, Mar. 2014, doi: 10.1002/joc.3711.

[17] T. Stackhouse, P.W., Westberg, D., Hoell, J.M., Chandler, W.S. and Zhang, "Prediction of Worldwide Energy Resource (POWER)-Agroclimatology methodology-(1.0 latitude by 1.0 longitude spatial resolution).," 2015, *Hampton, NASA Langely Research Center*. [Online]. Available: https://power.larc.nasa.gov/

[18] P. M. Atkinson, "Super-Resolution Land Cover Classification Using the Two-Point Histogram," *geoENV IV — Geostatistics Environ. Appl.*, vol. 13, (2006), pp. 15–28, doi: 10.1007/1-4020-2115-1_2.

[19] Y. Kita, "A study of change detection from satellite images using joint intensity histogram," in *2008 19th International Conference on Pattern Recognition*, IEEE, Dec. (2008), pp. 1–4. doi: 10.1109/ICPR.2008.4761020.

[20] J. You, X. Li, M. Low, D. Lobell, and S. Ermon, "Deep Gaussian Process for Crop Yield Prediction Based on Remote Sensing Data," in *31th AAAI Conference on Artificial Intelligence*, Association for the Advancement of Artificial Intelligence, (2017), pp. 4559–4565.

[21] H. Russello, "Convolutional Neural Networks for Crop Yield Prediction using Satellite Images," IBM Center for Advanced Studies, (2018).

[22] J. Sun, L. Di, Z. Sun, Y. Shen, and Z. Lai, "County-Level Soybean Yield Prediction Using Deep CNN-LSTM Model," *Sensors*, vol. 19, no. 20, Oct. (2019), p. 4363, doi: 10.3390/s19204363.

[23] W. W. Hsieh and B. Tang, "Applying Neural Network Models to Prediction and Data Analysis in Meteorology and Oceanography," (1998). doi: 10.1175/1520-0477(1998)079<1855:ANNMTP>2.0.CO;2.

[24] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model," in *International Joint Conference on Artificial intelligence (IJCAI)*, (1995), pp. 1137–1143.

[25] C. Tofallis, "A better measure of relative prediction accuracy for model selection and model estimation," *J. Oper. Res. Soc.*, vol. 66, no. 8, (2015), pp. 1352–1362, doi: 10.1057/jors.2014.103

[26] J. Quille-Mamani, L. Ramos-Fernández , J. Huanuqueño-Murillo, D. Quispe-Tito, L. Cruz-Villacorta, E. Pino-Vargas, L. Flores del Pino, E. Heros-Aguilar, L. Ángel Ruiz, "Rice Yield Prediction Using Spectral and Textural Indices Derived from UAV Imagery and Machine Learning Models in Lambayeque, Peru", Remote Sensing; 17(4):632, (2025), https://doi.org/10.3390/rs17040632.

[27] A. Adin, E.T. Krainski, A. Lenzi, Z. Liu, J. Martínez-Minaya, and H. Rue, "Automatic cross-validation in structured models: Is it time to leave out leave-one-out?", Spatial Statistics, 62, (2024), p.100843, https://doi.org/10.1016/j.spasta.2024.100843.