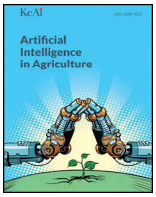
Artificial Intelligence in Agriculture 6 (2022) 156–166

Contents lists available at ScienceDirect

Artificial Intelligence in Agriculture

journal homepage: http://www.keaipublishing.com/en/journals/artificial-intelligence-in-agriculture/



Durum wheat yield forecasting using machine learning

Nabila Chergui *

Faculty of Technology, Ferhat Abbas University, Setif 1. MISC Laboratory Abdelhamid Mehri University, Constantine 2, Algeria.

ARTICLE INFO

Article history:
Received 19 November 2021
Received in revised form 9 September 2022
Accepted 11 September 2022
Available online 19 September 2022

Keywords:
Machine learning
Yield forecast
Deep learning
Data augmentation
Regression
Climate data

ABSTRACT

A reliable and accurate forecasting model for crop yields is crucial for effective decision-making in every agricultural sector. Machine learning approaches allow for building such predictive models, but the quality of predictions decreases if data is scarce. In this work, we proposed data-augmentation for wheat yield forecasting in the presence of small data sets of two distinct Provinces in Algeria. We first increased the dimension of each data set by adding more features, and then we augmented the size of the data by merging the two data sets. To assess the effectiveness of data-augmentation approaches, we conducted three sets of experiments based on three data sets: the primary data sets, data sets with additional features and the augmented data sets obtained by merging, using five regression models (Support Vector Regression, Random Forest, Extreme Learning Machine, Artificial Neural Network, Deep Neural Network). To evaluate the models, we used cross-validation; the results showed an overall increase in performance with the augmented data. DNN outperformed the other models for the first Province with a Root Mean Square Error (RMSE) of 0.04 q/ha and R_Squared (R^2) of 0.96, whereas the Random Forest outperformed the other models for the second Province with RMSE of 0.05 q/ha. The data-augmentation approach proposed in this study showed encouraging results.

1. Introduction

The estimation of crop yield is a crucial task for decision-makers as it enables efficient planning of resources (Chergui et al., 2020; Gyamerah et al., 2020; Kim et al., 2019). Economically, yield prediction can help decision-makers react in the right way on how much to export in the case of a surplus or make early decisions about quantities, contracts, agreements and planning of imports in the case of shortage. It can also help determine the optimal profile of crops to plant and appropriately allocate government resources. On the other hand, it can help farmers decide on what and when to grow and plan their harvest and storage (Chergui et al., 2020). Besides, cereals are considered the most important crops in Algeria and play a crucial role in maintaining the food supply. According to official reports, the estimated need of Algeria for cereals is around 15 million tons per year. In 2019, production for the harvest season was estimated at 6 million tons, while it imported nearly 12 million tons. Consequently, timely and accurate yield information is fateful to decision-makers and sustainable developments. Moreover, crop yield prediction permits studying factors that influence and affect production, such as climate and weather, irrigation and fertilisation practices, agricultural policies, etc.

Advanced information and communication technologies have emerged together with machine learning (ML) to create new opportunities to understand processes in agricultural systems where crop yield prediction is one of their crucial tasks that consists of several complicated steps. ML is a practical approach that can provide better yield prediction based on several features (Klompenburg et al., 2020); ML can determine patterns and correlations and discover knowledge from data sets. (Klompenburg et al., 2020) present a systematic review of the literature on crop yield prediction using ML techniques, where it extracts the major ML algorithms, features and evaluation metrics used for this task. Several models from traditional ML models (Gyamerah et al., 2020; Guo et al., 2021; Shahhosseini et al., 2020) to complex deep learning (DL) models (Alibabaei et al., 2021; Shook et al., 2021; Schwalbert et al., 2020; Wolanin et al., 2020; You et al., 2017) have been successfully developed and proven their efficiency.

However, a good and powerful ML model requires equally good data quality to generate quality predictions, where the data quality can be affected by two factors, impurities and size. In our case, regrettably, directorates and agricultural services in Algeria became aware of the importance of recording the farming data, farmers' knowledge and practices (including seeding, irrigation, fertilisation, soil components, etc.), and their significant role in enhancing the production and food sustainability till the last few years. Therefore, the prediction process in such a situation is challenging.

* Corresponding author.
E-mail address: nabila.chergui@univ-setif.dz (N. Chergui).

https://doi.org/10.1016/j.aiia.2022.09.003

Fortunately, impurities in data, including noise, outliers, missing data, duplicate data or any other anomalies, can be handled using data cleansing methods. While one possible solution to solve the problem of small data sets is data augmentation. This latter includes techniques of increasing training data diversity without explicitly collecting new data (Feng et al., 2021), in an attempt to improve the generalisation ability of training models. For real-world data, generalisation is significant because it can help networks overcome small data sets (Olson et al., 2018), and it is used to regularise and reduce over-fitting when training ML models (Shorten and Khoshgoftaar, 2019). The most widely used data augmentation techniques involve either adding slightly modified copies of existing data or creating synthetic ones (Shorten and Khoshgoftaar, 2019).

Several artificial intelligence (AI) applications apply data augmentation in the case of small training data. In computer vision and pattern recognition for generating more labelled data, where techniques like cropping, flipping, and colour jittering are standard components of model training (Feng et al., 2021). In natural language processing (NLP), in this case, techniques of interpolation and rule-based are applied to augment patterns (Feng et al., 2021). In addition to time series classification based on techniques of random transformation (Iwana and Uchida, 2021).

Agricultural applications based on computer vision like crop classification, diseases and weed detection also used data augmentation based on techniques to artificially increase the number of images in hand based on ideas of (rotation, light shade's variation, colour inversion, translation, improving the resolution and changes in intensity and so on) (Arsenovic et al., 2019; Aravind and Raja, 2020; Chen et al., 2019; Kamal et al., 2019; Liu et al., 2017; Sladojevic et al., 2016; Yamamoto et al., 2017). Yet, this solution is not applicable to crop yield forecasting. Transfer learning is another alternative solution for increasing small data set, where the knowledge obtained from solving a task in a given domain is transferred to the target domain (Aravind and Raja, 2020; Barbedo, 2018; Cruz et al., 2017; Wang et al., 2018). But, this solution can be effective if the source and target domains share certain similarities; and if we initially have a good model for the source domain, which is not always possible.

In this work, we address the problem of crop yield forecasting and the performance of ML methods in the presence of small data sets where we use data augmentation techniques to increase the data sets and enhance the performance of ML models. In which we conduct a comparative study using ML regression models, such as Support Vector Regression (SVR), Random Forest regression (RF), Artificial Neural Network (ANN), Extreme Learning Machine (ELM) and Deep Neural Network (DNN), under three experimental scenarios by evaluating the performances achieved by the different methods.

2. Related works

ML and DL techniques are of increased use to estimate different crop yields. (Elavarasan et al., 2018) discuss the yield estimation by integrating agrarian factors in ML techniques. Besides, several works have been proposed to estimate crop yield using ML (Abbas et al., 2020; Guo et al., 2021; Ji et al., 2021; Piekutowska et al., 2021; Rahman and Robson, 2020; Rezapour et al., 2021; Shahhosseini et al., 2020; Xu et al., 2020) and others have been used DL models (Alibabaei et al., 2021; Cao et al., 2021; Gong et al., 2021; Nevavuori et al., 2020; Shook et al., 2021; Schwalbert et al., 2020).

(Chergui et al., 2020) demonstrate possible applications of ML and DL for crop management, including crop yield forecasting. They distinguish two trends for crop estimation; the first one is related to data sources that directly affect the crops(soil data, weather data, environmental parameter data). These are typically used to offer a pre-season estimate of crop yield (Crane-Droesch, 2018; Ehret et al., 2011; Fukuda et al., 2013; Gonzalez-Sanchez et al., 2014; Gyamerah et al., 2020; Kouadio et al., 2018). The second trend is related to the data sources' collected using advanced technologies and tools such as multi-spectral and hyper-spectral satellite images, remote sensing and sensors. These are usually applied to provide in-season predictions of crop yields (Ji et al., 2017; Maimaitijiang et al., 2020; Pantazi et al., 2016; You et al., 2017). Where other studies use both types of data sources (Abbas et al., 2020; Filippi et al., 2019; Han et al., 2020; Jeong et al., 2016; Kim et al., 2019; Kamir et al., 2020; Oliveira et al., 2018; Schwalbert et al., 2020; Sakamoto, 2020).

Other proposed studies used data augmentation via the aggregation of different small data sets for crop yield forecasting. Among others, we found the work of (Meroni et al., 2021) which predicted the monthly yield of three types of cereals at the national level, where it integrated data sets by crop from all provinces and added an identifier for each Province as an additional predictor variable. (Filippi et al., 2019) combined data on three types of cereals over multiple fields and years into one data set. Then they created three separate models based on factors like (pre-sowing, mid-season and late-season conditions); to study the effect of the availability of within-season information on the predictive ability of their models. They created a single model for the three crops and included the crop type as a predictor variable. In our study, we integrate small data sets from two Provinces to form a more considerable base for training. Rather than creating a single predictor for the two Provinces, we seek to construct a specific predictor for each one that better corresponds to its specificity.

In Table 1, we resume and compare recent works on forecasting crop yield, where we focus on the first trend that used to offer a pre-season estimate of crop yield.

3. Material and methods

3.1. Region and crop selection

The study was carried out in two sites: the Province of Constantine and the Province of Setif 1 (see Fig. 1). Constantine is located in the northeast of Algeria; at latitude: 36.2833 and Longitude: 6.61667, 36° 16′ 60” North and 6° 37′ 0″ East, and it has a Mediterranean climate. Setif's latitude: 36.15 and longitude: 5.43333, 36° 11′ 28.03” North and 5° 24′ 49.43″ East, it has a semi-arid climate.

These regions are known for the cultivation of cereals. The crop chosen for this study is durum wheat because it presents the most important crop and occupies the amplest agricultural area. The cropped wheat in both regions is rainfed or not irrigated.

3.2. Data sources

We have obtained historical data on durum wheat yields from the Department of Agricultural Services of the Province of Constantine[1] and the Province of Setif,[2] data is not available online. The yield data obtained for the Province of Constantine covers the seasons from 1991 to 2019 and starting from 2005 to 2020 for the Province of Setif. The two data sets contain information on the recorded annual yields and the acreage; they present the crop yield as the harvested production of a crop per unit of the harvested area, measured in metric quintals per hectare (q/ha). However, they do not provide information on geographic identification, soil nature and components or fertilisation programs.

Additionally, we used climatic data obtained from different sources, initially from the principal used source,[3] that archives meteorological data at airports of the provinces of Constantine and Setif. However, this site offers downloadable data only from February 2005; therefore, we have got the weather data from 2005 to 1996 from a second source.[4]

[1] www.dsa-constantine.dz

[2] www.dsa-setif.dz

[3] https://rp5.ru/

[4] https://dz.freemeteo.com/

Table 1
ML approaches for crop yield estimation.

Ref	Crop type	Algo	Evaluation metrics	Data type	Extracted features	Results
Shahhosseini et al. (2020)	corn	RF	RRMSE, MBE MDA,RMSE	yield, soil weather management variables	min-temperature max-temperature precipitation, vapor Shortwave radiation Snow water soil organic matter sand/clay content & 12 other features	RMSE = 1.113 kg/ha PRMSE = 9.56% MBE = -116 kg/ha
Wang et al. (2020)	winter wheat	Adaptive Boosting	MAE, RMSE R^2	satellite images, climate data soil maps historical yield	Vegetation indices,mean-temperaturemax-temperature min-temperature, Soil properties, 6 other features	R^2 = 0.86 RMSE = 0.51 t/ha MAE = 0.39 t/ha
Piekutowska et al. (2021)	potato	MLP	RAE, RMSE MAE, MAPE	historical yield agronomic meteorological data	mean-temperature precipitation, sum of NPK planting date, 6 other features on soil fertility	RMSE = 2.121 t/ha RAE = 0.099 MAE = 1.626 t/ha MAPE = 7.203%
Guo et al. (2021)	rice	SVM	RMSE, R^2	climate and geographical data	mean-temperaturemax-temperature min-temperature humidity, min- relative humidity mean-wind speed max-wind speed	RMSE = 760 kg/ha R^2 = 0.33
Stepanov et al. (2020)	soybean	LR	RMSE,MAPE	climate satellite data	Vegetation indices Precipitation, soil-temperature humidity photosynthetic active radiation	RMSE = 0.05 t/ha MAPE = 0.94%
Cao et al. (2021)	winter wheat	LSTM	R^2 , RMSE	climate satellite data soil surveys	max-temperature min-temperature precipitation, soil depth soil texture, pH,...	RMSE = 561 kg/ha R^2 = 0.83
Khaki et al. (2020)	corn soybean	CNN-RNN	RMSE	yield management weather soil	precipitation, solar radiation max-temperature min-temperature soil water content & organic matter, snow, 10 other features	RMSE = 4.15 bu./acre
Alibabaei et al. (2021)	tomato potato	Bidirectional LSTM	MSE, R^2	climate data irrigation scheduling soil water content	mean-temperature max-temperature min-temperature min, max & average relative humidity, average solar radiation and others	MSE = 0.017 R^2 = 0.99
Shook et al. (2021)	soybean	LSTM	MAE, RMSE R^2	weather data	mean-temperature max-temperature min-temperature mean-precipitation, 3 other features	RMSE = 7.226 bu./acre MAE = 5.453 R^2 = 0.795
Gyamerah et al. (2020)	groundnut millet	RF	RMSE, MAPE, MBE R^2	climate data yield	daily sunlight humidity precipitationmin-temperaturemax-temperature mean-temperature	RMSE = 0.0173 t/ha MAPE = 0.909% R^2 = 0.9805 MBE = 0.01 t/ha
Wolanin et al. (2020)	wheat	CNN	Nash-Sutcliffe efficiency	weather variables	vegetation indices max-temperature min-temperature 5 other features	NSE = 0.868

Due to the scarcity of meteorological data for the pre-1996 period, we have reduced the study interval for the province of Constantine from 1996 to 2019.

The climatic data from the first source is rich and downloadable; it offers daily-level records (at least three records per day). The second source provides data from 1945 at the daily level, but it was hard to reuse because it was not downloadable. Additionally, it includes only a few features (air temperature, precipitation, wind and pressure); it contains missing records for a few months. Thence, we selected attributes in common between the two data sources (the monthly mean temperature and the monthly mean precipitation) as predictor variables from November to June for each year. It exists a national climate centre that provides more detailed data and includes more expanded information on solar radiation, humidity, pressure, wind, etc. Because of the absence of agreements between the institutions, this data was helpless.

We used a third data source[5] for the following purposes: to complete the data for the missing months and to extract the relative air humidity for the period before February 2005 to construct the data set used for the second experiment. Data from this source is not directly downloadable.

Durum wheat is sown in autumn (from late October to November, subject to autumn rainfall); harvesting acquires at the beginning of summer (from June to July). Therefore, we collected climatic data from November to May.

3.3. Data pre-processing

The historical yield data was complete with no missing data.

We obtained climatic data from different sources; therefore, we first extracted all the needed features and transformed them into a single form. Next, we cleaned up the data from some redundant records. Then, we imputed the rest of the missing records by using the median approach. After that, we gathered and calculated the monthly means of temperature, precipitation and relative humidity.

Later, we calculated the standard deviation (SD) of the yield time series in each zone and then defined the observed values as outliers and removed them if they were not within the range of ±3SD.

Finally, we normalised the data and defined it into an identical scale in the interval [0,1] using the min_max normalisation function.

3.4. AI models and evaluation metrics

To investigate how the data augmentation approach proposed in this work can improve the accuracy of yield prediction, we conduct a comprehensive set of experiments using the main types of AI regression models for crop yield modelling. Based on the study by (Klompenburg et al., 2020), which claimed that neural networks were the most used and accurate tools for crop yield estimation, we selected the Extreme Learning Machine (ELM), Artificial Neural Network (ANN) and Deep Neural Network (DNN) as predictive models. In addition, we also test the accuracy of the approach using some algorithms commonly used for small data set: Random Forest (RF) and Support Vector Regression (SVR) (See Table 2).

We adjusted the hyper-parameters of each model by experiments; the properties of neural network models like their structure, learning rate and activation functions, in addition to those of RF and SVR.

Finally, we normalised the data and defined it into an identical scale in the interval [0, 1] using the min_max normalisation function.

Root mean squared error (RMSE), mean absolute percentage error (MAPE), coefficient of determination (R-squared or R^2), mean absolute error (MAE), and mean biased error (MBE) were the statistical parameters used for evaluating the accuracy of the forecast.

[5] https://www.wunderground.com/history/monthly/dz/constantine/

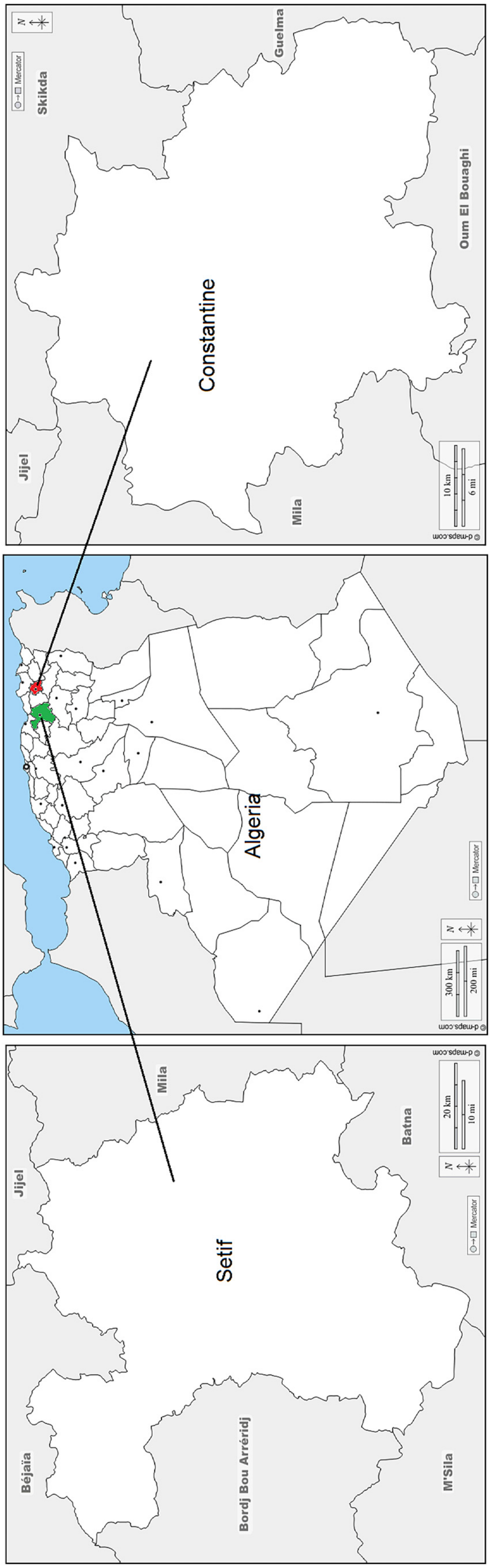


Fig. 1. Study area: Province of Constantine in red (the right sub-figure) and Province of Setif in green (the left sub-figure).

These metrics are defined as follows:

$$RMSE = \sqrt{\frac{\sum (y_i - \widehat{y}_i)^2}{N}} \tag{1}$$

$$MAPE = \frac{1}{n} \sum \frac{|y_i - \widehat{y}_i|}{y_i} \times 100\% \tag{2}$$

$$R^2 = 1 - \frac{\sum (y_i - \widehat{y}_i)^2}{\sum (y_i - \overline{y})^2} \tag{3}$$

$$MAE = \frac{1}{N} \sum |(y_i - \widehat{y}_i|) \tag{4}$$

$$MBE = \frac{1}{N} \sum (y_i - \widehat{y}_i) \tag{5}$$

Where y_i, $\widehat{y}_i$, $\overline{y}$, N are the actual, predicted, mean, and the data set size.

3.5. Methods

ML algorithms, especially neural networks, are known to be data-hungry; small data sets degrade their performance. Therefore, forecasting crop yield using small data sets may not achieve high accuracy.

After pre-processing the data, we obtained a data set containing 23 samples for the Province of Constantine and 14 samples for the Province of Setif.

We are going to conduct three experiments under three different data sets where the component of each one is as follows:

- We first refer by data-sets1, data-sets2 and data-sets3 to the data sets used respectively for experiment1, experiment2 and experiment3;
- We respectively refer by (data-set1A, zone-A) and (data-set1B, zone-B) to the Province of Constantine and its data and the Province of Setif and its data. Both data sets contain the mean monthly precipitation and the mean monthly temperature;
- Data-sets2 used for experiment2 is the data-sets1 enhanced by an additional feature (the relative humidity) to improve the prediction's accuracy by first adding a new feature to data-sets. We built data-set2A for zone-A and data-set2B for zone-B;
- In experiment3, we merged data-set2A with data-set2B to build data-set3 (data-set3A for zone-A and data-set3B for zone-B), the merging process will be discussed later (see Fig. 2).

We divided data sets into training data (70%) and test data (30%). We chose this partition to obtain more cases to test the generation ability of the ML algorithms, therefore, to get more accurate results because too little test data can lead to an optimistic estimation of the model's performance. Then, we used three-fold cross-validation to train the ML algorithms.

3.5.1. Data augmentation through merging of data sets

To increase the performance of models' predictions when only small data sets are available, we augment raw versions of data-sets2 (with humidity) used to train/test ML algorithms for the two zones (A and B) by merging them to build a larger one (data-set3).

Then, we pre-processed the resulting data set (data-set3) by removing the outlier using ±3SD. After that, we normalise it using the min-max function.

To get a more transparent comparison between the three experiments, we keep the same test sets used for experiment1 and experiment2. Therefore, we split data-set3 into three subsets: test-setA for zone-A, test-setB for zone-B and train-setAB, and the training set for the two zones, which means that we exclude test-setA and test-setB from train-setAB.

Table 2
Summary of the main features, strengths and weaknesses of AI models used in this study.

Algo	features	Strengths	Weaknesses
ANN	A network model consisting of input hidden and output layers based on the back-propagation	Performs well with larger data sets	Suffers from local minima problem Hard to determine the proper network structure. Computationally expensive
DNN	A network model consisting of input and several hidden layers and output layer based on the back-propagation	Avoids problems of over-fitting and local minima through the optimisation process in a deep network structure with the combination of activation functions and dropout method. Improves the accuracy by training complicated and huge input data	Requires large amount of data in order to outperform other algorithms Training is expensive due to complex data models Requires a high-end computer
ELM	Is a single hidden layer feed forward NN and based on the empirical risk minimisation theory and its learning process needs single iteration only	Avoids local minimisation and multiple iterations. Reduces the computational burden. Lower computational cost	May cause non-optimal solution
RF	Based on the bagging algorithm and uses Ensemble Learning technique and bootstrap. It creates trees on the subset of the data and combines the output of all the trees. Uses rule based approach instead of distance calculation.	Performs well with high dimensional data. Robust to outliers and nonlinear data. Works well with categorical and continuous values. Automates missing values in data. Doesn't require data normalising. Less impacted by noise	Not interpret-able (black boxes) Memory consuming with large data sets. Susceptible to over-fitting because it does not predict beyond the range in the training data. Cannot deal with outliers when is trained by small data sets.
SVR	Used to predict discrete values with the basic idea of finding the best fit line (hyperplane) that has the maximum number of points within a threshold value, which is the distance between the hyperplane and boundary line.	Robust to the outliers. Performs lower computation. Easy to implement	Sensitive to noise. Not suitable for large data sets. Under-perform if: nbr of features > nbr of samples

Cross-validation is applied to train ML algorithms. To train models for zone-A, we exclude the validation records of zone-B from its training set and the same for training zone-B. In another sense, we adjust the learning of the ML algorithm according to the target zone. Fig. 2 presents the data augmentation process.

4. Results and discussions

4.1. Results

After implementing the ML models in Python using existing libraries like the Scikit-Learn library, we obtained the results presented in Table 3 where the text in bold font indicates the best results for each metric in each experiment. Generally observed, each of the regression methods: ANN, DNN, ELM, RF and SVR, exhibited better performance for experiment3 over the two other experiments for both zones in most cases. Besides, they also showed better-to-comparable performance with experiment2 and experiment1, while experiment2 (data with the additional feature) was favourable. Algorithms known to perform well with larger data sets (ANN, DNN, RF, ELM) performed very well with experiment3 for both zones. However, SVR showed weaknesses in generalising high-quality estimates for all three experiments, especially for experiment3 (zone-A). It is because SVR does not perform well with high-dimensional data sets.

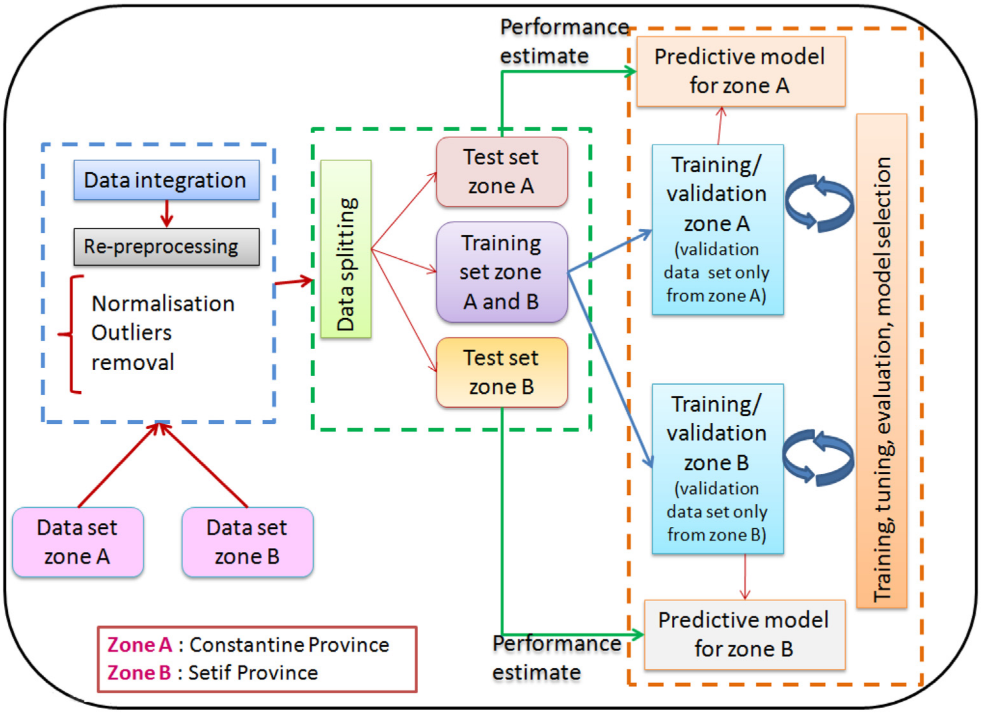

Data integration
Re-preprocessing
Normalisation
Outliers
removal
Data set zone A
Data set zone B
Zone A : Constantine Province
Zone B : Setif Province
Data splitting
Test set zone A
Training set zone A and B
Test set zone B
Performance estimate
Performance estimate
Predictive model for zone A
Training/ validation zone A (validation data set only from zone A)
Training/ validation zone B (validation data set only from zone B)
Predictive model for zone B
Training, tuning, evaluation, model selection

Fig. 2. Learning process using data augmented by merging.

Table 3
Yield forecasting results for zone -A "Constantine province" and zone-B "Setif province" for different ML methods.

	Zone-A						Zone-B				
Test	RMSE(q/ha)	MAPE(%)	MAE(q/ha)	MBE(q/ha)	R^2		RMSE(q/ha)	MAPE(%)	MAE(q/ha)	MBE(q/ha)	R^2
						ANN					
Experiment1	0.1007	**13.9394**	0.0819	0.0789	0.9308		0.1771	45.9472	0.1639	−0.1016	0.1247
Experiment2	0.1388	22.2904	0.1133	0.1133	0.8685		0.1396	33.8436	0.1333	−0.0540	0.3879
Experiment3	**0.0688**	14.5846	**0.0625**	**0.0527**	**0.9510**		**0.1100**	**21.4110**	**0.0908**	**0.0184**	**0.5730**
						DNN					
Experiment1	0.1600	28.8617	0.1557	−0.0614	0.8888		0.1863	21.4180	0.1474	0.0923	0.7721
Experiment2	0.1354	15.1215	0.1015	0.0423	0.9204		0.1571	**15.0768**	0.1288	−0.0081	**0.8381**
Experiment3	**0.0378**	**5.9003**	**0.0285**	**0.0001**	**0.9658**		**0.0854**	18.7099	**0.0739**	**0.0058**	0.7151
						RF					
Experiment1	0.2381	28.0641	0.1862	**0.0311**	0.1521		0.1342	**20.4430**	0.1122	**0.0459**	0.7936
Experiment2	0.2123	**19.2608**	**0.1476**	0.1467	**0.3255**		0.1332	21.3681	0.1095	0.0754	0.7965
Experiment3	**0.1804**	28.8019	0.1646	0.1450	0.0020		**0.0513**	24.8798	**0.0430**	−0.0418	**0.8356**
						ELM					
Experiment1	0.0889	14.7851	0.0805	**0.0675**	0.9461		0.1607	**25.7153**	0.1214	−0.0859	−0.0187
Experiment2	0.0873	14.5597	0.0792	−0.0578	0.9480		0.1539	35.8346	0.1344	0.0963	0.0661
Experiment3	**0.0693**	**9.4766**	**0.0549**	−0.0430	**0.9502**		**0.1109**	44.5886	**0.0858**	**0.0535**	**0.6610**
						SVR					
Experiment1	**0.2251**	**36.1959**	0.1851	**0.1467**	**0.2420**		0.2895	73.8247	0.2400	−0.0016	0.0387
Experiment2	0.2976	43.8391	0.2630	0.2518	−0.3250		0.2678	71.2152	0.2206	−0.0362	0.1774
Experiment3	0.3014	39.8550	**0.1841**	0.2496	−0.97		**0.0985**	**48.0947**	**0.0859**	**0.0291**	**0.3945**

Bold indicates the best results for each metric in each experiment.

Furthermore, the best performance for durum wheat yield prediction was observed when using augmented data (experiment3) for both zones, achieving a (RMSE = 0.037 q/ha, MAPE = 5.9%, R^2= 0.965) with DNN algorithm for zone A, and a (RMSE = 0.051 q/ha, MAPE = 24.879%, R^2= 0.835) with RF for zone-B, as shown in Table 3.

On the other hand, when increasing the number of input features (experiment2), the evaluation metrics gradually improved (RMSE, MAPE, MAE, MBE, R^2) for almost all methods, especially for zone-B. Indicates that all tested regression methods, to some extent, were able to benefit from the additional input feature. In addition, the relative humidity feature has helped improve the prediction accuracy in many cases.

However, ML methods such as ANN, RF and ELM outperformed DNN regression when fewer input variables were available (experiment1 and experiment2). Conversely, DNN achieved higher performance than other methods when a more significant number of input variables were available (augmented data of experiment3). It is because deep learning often goes beyond ordinary machine learning models for larger data sets (Maimaitijiang et al., 2020).

Bar chart that visualises performance of the evaluation metrics (RMSE, MAPE, R^2, MAE and MBE) of the five methods using the test data are given in Fig. 3 for zone-A, and Fig. 4 for zone-B. When predicting the yield of durum wheat using data sets of experiment1, comparative to the five benchmark methods, it is evident that for zone-A, ELM performed better in terms of RMSE, MAPE, MAE and R^2. Yet RF outperformed all the other models in terms of all metrics for zone-B.

Moreover, we can observe that ANN, DNN and ELM performed very well concerning all evaluation metrics when predicting the durum wheat yield using data-set3A from experiment3 compared to RF and SVR. In addition, we perceive that for zone-B, almost all methods reached a high level of precision for RMSE, MAE, MBE and R^2 (except the DNN showed a bit of degradation with R^2). RF outperformed other methods in RMSE, R^2 and MAE, while DNN surpassed them in MAPE and MBE.

On the other side, if we take the MBE metric that explains the average amount by which the observation is greater than the prediction, and allows us to see how much the model overestimates or underestimates. We can perceive that the DNN model outperformed all the other models for both zones. Also, there is an improvement in the value of MBE from an underestimation in experiment1 to a slight overestimation in experiment2. Besides, it reached a neglected level in experiment3 for both zones, indicating that the DNN model in experiment3 for both zones is not biased. If we look at an overview of all evaluation metrics; we can observe that using data-set2A from experiment2 (with the additional feature) does not improve the prediction accuracy for zone-A compared to experiment1 regarding ANN and SVR methods, where DNN recorded better performances but still exceeded by ELM.

However, for zone-B in experiment2, all methods performed better than those in experiment1 in terms of all metrics. RF outperformed other methods considering RMSE and MAE, and DNN performances' surpassed those of the remaining algorithms in terms of R^2, MAPE and MBE. This fact conducts us to understand that the impact of the relative humidity feature is more significant for forecasting wheat grown in zone-B (characterised by semi-arid climate) than zone-A (characterised by Mediterranean climate).

5. Discussions

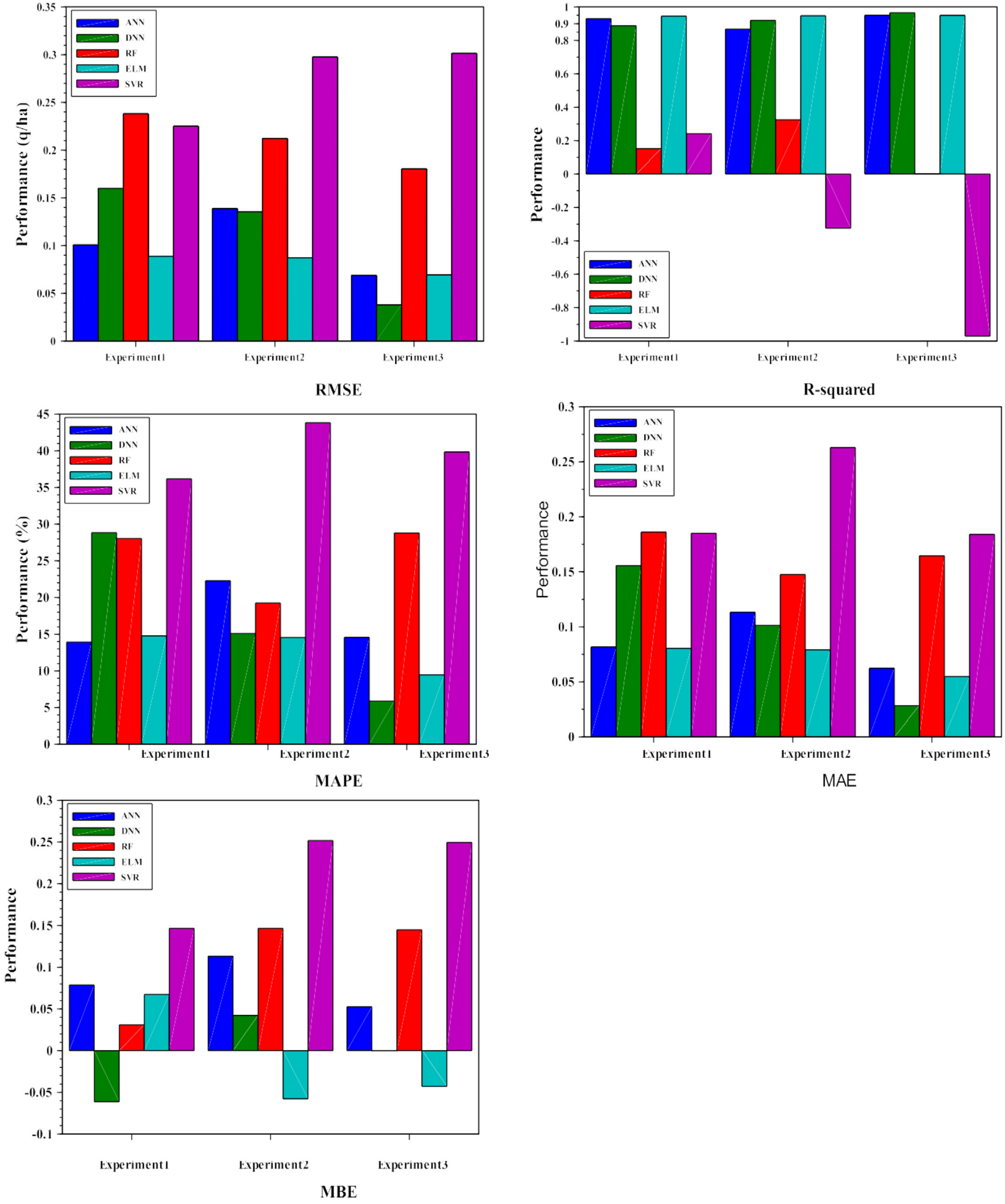
For a clearer vision, we further investigate the relationship between relative humidity and yield using the Partial Dependence Plots (PDP) along with the Individual Conditional Expectation plots (ICE) to visualise the marginal effect of relative humidity on yield prediction and the dependence between them for both zones.

But since there is an average correlation between the relative humidity and rainfall (knowing that $correlation$ = 0.37 for zone-A, $correlation$ = 0.43 for zone-B), PDP may become a misleading. To overcome this limitation, we used the Accumulated Local Effects (ALE) plot, an unbiased model that describes how much the feature influences the prediction of an ML model.

The Figs. 5, 6 show two PDP, ICE and ALE plots for zone-A and zone-B, with a RF regressor.

The PDPs in Figs. 5a and 6a respectively depict the average effect of the relative humidity on the yield for zone-A and zone-B. We can observe that an increase in the relative humidity increases the yield value until reaching some levels where it becomes steady (from 0.55 for zone-A and 0.68 for zone-B). Besides, the effect of relative humidity on zone-B is more significant than its influence on zone-A; this is evident when we look at the size of the change in the yield. Through the curve of zone-A, we notice that the amplitude of variation is equal to 0.035, which is considered negligible compared to 0.20 of zone-B.

The ICE plots are presented in Figs. 5b and 6b, in which the dashed lines in the plots correspond to the PDP line that overlaid on ICE lines (each sample of the feature has a separate line). These plots highlight



RMSE

R-squared

MAPE

MAE

MBE

Fig. 3. Bar chart of RMSE, R-Squared, MAPE, MAE and MBE, predicted durum wheat yield by ANN, DNN, RF, ELM, SVR, using zone-A test data.

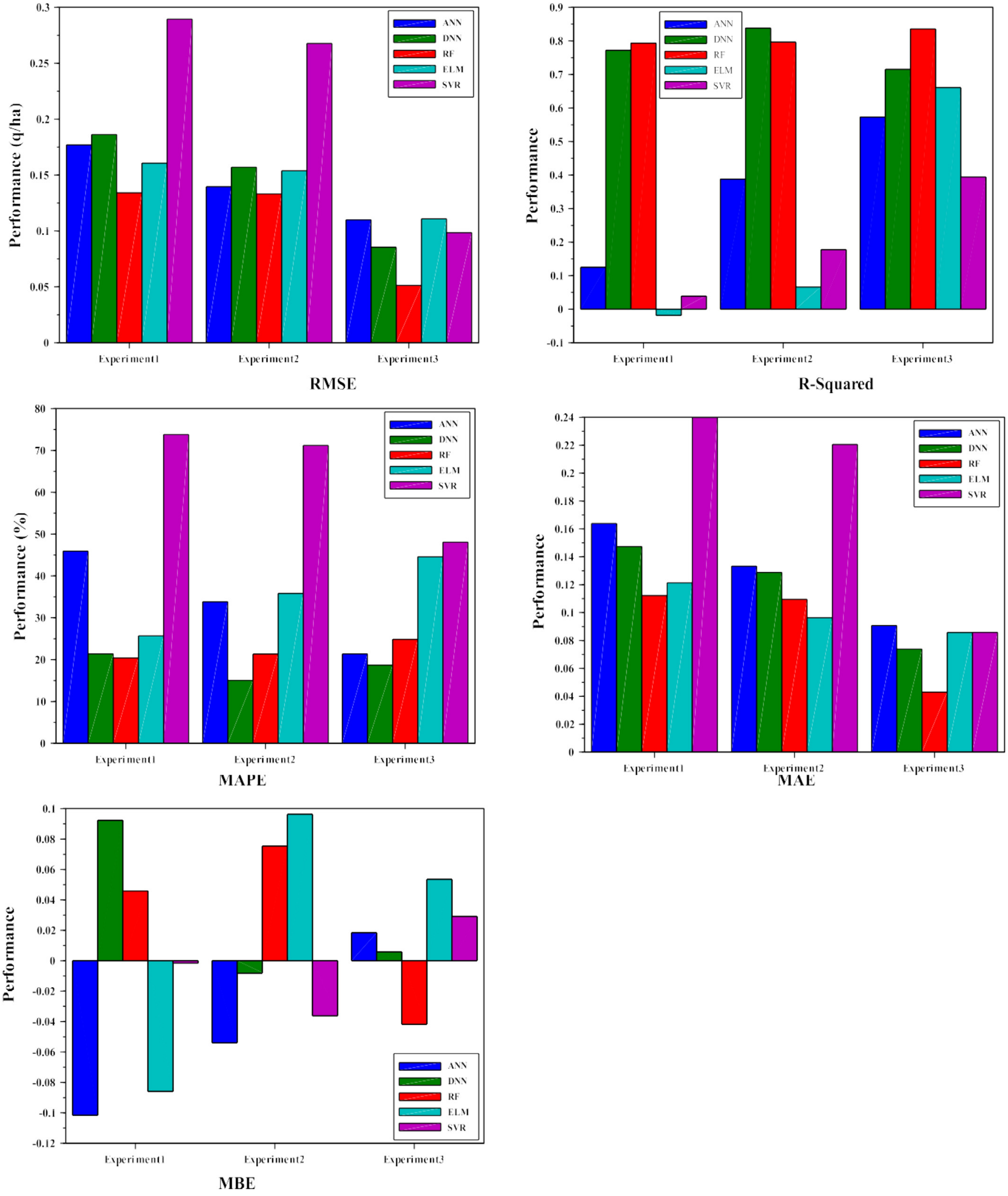
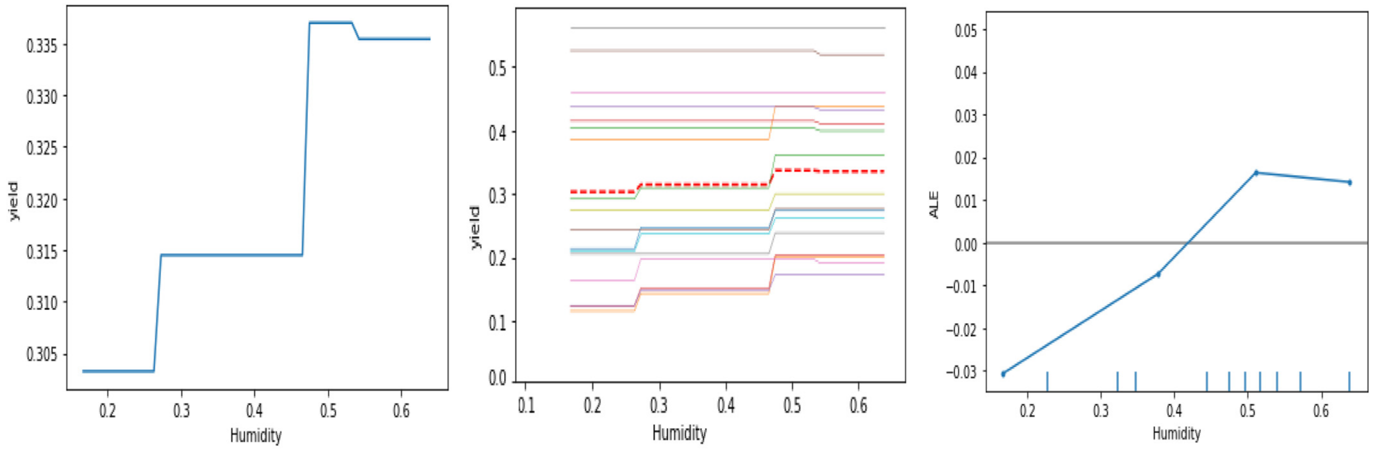


Fig. 4. Bar chart of RMSE, R-Squared, MAPE, MAE and MBE, predicted durum wheat yield by ANN, DNN, RF, ELM, SVR, using zone-B test data.

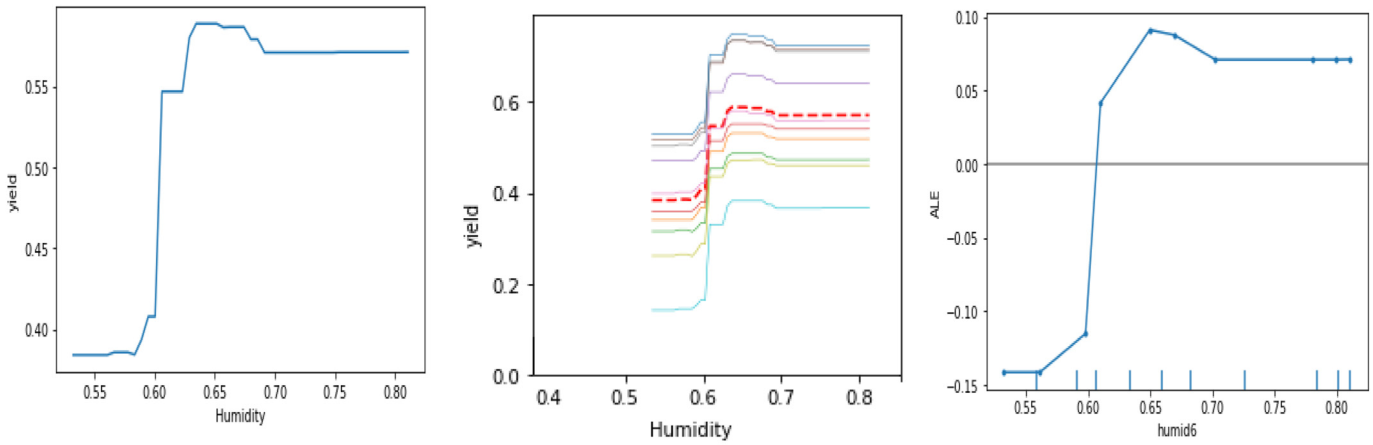
the dependence of the prediction on a feature for each entry separately. These ICE plots confirm the previous observation where we can see that in zone-A, several samples have flat shapes to little curved, whilst in zone-B, all instances have curved shapes. This indicates that relative humidity is more relevant for predictions of zone-B than zone-A.

Figs. 5c and 6c showed the ALE plots for zone-A and zone-B. These plots are zero-centred which signifies the mean prediction, and then the value at each point of the ALE curve is the difference from this mean. The higher the difference is, the higher the effects of the feature on the yield prediction. Accordingly, we conclude that the influence of



(a) PDP (b) ICE (c) ALE

Fig. 5. The partial dependence plot (PDP), the individual conditional expectation plot (ICE) and the accumulated local effects (ALE) plots of humidity for zone-A.



(a) PDP (b) ICE (c) ALE

Fig. 6. The partial dependence plot (PDP), the individual conditional expectation plot (ICE) and the accumulated local effects (ALE) plots of humidity for zone-B.

relative humidity on yield prediction of zone-B is more significant than for zone-A.

6. Limitation and perspectives

The data augmentation approach has remarkably helped improve the quality of predictions for many models. This fact is not due only to larger data sets, but the latter may contain additional information and varieties in weather conditions with different intrinsic yield potential, and this variety would be a necessary predictor variable (Filippi et al., 2019). The drawback of this study is that we used the entire Province (an aggregation of several scattered fields and farms) without considering sub-regional spatial variability (Pham et al., 2022), resulting from climatic factors, the different topography and soil nature. In addition, various farmers manage these fields following different technical paths for seeding, harvesting and so on. Therefore, although we did not use most of this information in this study because of their unavailability, weather data taken at one station per Province are irrelevant for some fields and may affect the models' output. Therefore, if we had accurate climate variable information and ground-truth data at the field level within each Province, we could likely get more data points that would automatically increase the data set and possibly improve the accuracy of results.

Furthermore, the field-level information would allow us to use remote sensing images derived from satellite sensors like sentinel-2. It would make it possible to provide temperature values closer to the actual one on the farm compared to the employed values, which describe a sensed temperature that may be very far from several fields and farms within the same Province.

On the other hand, it would permit us to use additional variables like vegetation indices which have proven their efficiency in enhancing prediction accuracy (Maimaitijiang et al., 2020; Abbas et al., 2020; Filippi et al., 2019; Han et al., 2020; Kim et al., 2019; Kamir et al., 2020; Oliveira et al., 2018; Schwalbert et al., 2020; Sakamoto, 2020).

Another potentially obtained benefit from the availability of field-level information is the exploration of other deep learning models like CNN for yield forecasting using the sensed images. Additionally, the exploitation of data augmentation techniques used for image classification purposes to augment the data size of samples if needed (Arsenovic et al., 2019; Aravind and Raja, 2020). Besides, image-based deep learning approaches can employ spectral, spatial, and temporal information from satellite data and may reduce the need for feature engineering (Sagan et al., 2021).

7. Conclusion

To solve the problem of crop yield estimation in the presence of data scarcity and its impacts on forecasting quality, we conducted an approach to predicting the yield of durum wheat using only two climatic variables for two different provinces in Algeria.

The method was to increase the size and dimension of the data. Firstly, we increased the dimension of the initial data by adding a new climatic feature. Secondly, we augmented its size by merging data sets from the two provinces to create a more significant input. Next, we performed extensive experiments to study the effects of data size and select the best model, particularly to examine and test the effectiveness of the data augmentation-by-merging approach. We carried out a comprehensive comparison of five major AI regression models based on the three different data sets for the two provinces. Then, we proposed a cross-validation training/test process to learn algorithms to develop the best model for crop yield prediction. As a result, we achieved the best performance when utilising data augmentation-by-merging data sets for both Provinces. While the DNN model employed for the first Province outperformed the other four AI models, the RF model outperformed the related models for the second Province.

The data augmentation approach proposed in this study has shown promising results and can be adopted in other regions or for other crops. In addition, the suggested method is generic, allowing it to be applied to other agricultural applications or for other fields where data is scarce.

Credit statement

Nabila CHERGUI: conceptualisation, methodology, data collection and processing, performing the experiments, visualising the results, preparing tables and images, writing draft preparation, writing-reviewing and editing,

Declaration of Competing Interest

We declare that there is no conflict of interest associated with this research paper, and there has been no financial support for this work.

Acknowledgements

The author thanks the Constantine and Setif Agriculture Departments for their cooperation and for providing us with the required data on the durum wheat crop.

References

Abbas, F., Afzaal, H., Farooque, A., Tang, S., 2020. Crop yield prediction through proximal sensing and machine learning algorithms. Agronomy 10, 7. https://doi.org/10.3390/agronomy10071046.

Alibabaei, K., Gaspar, P., Lima, T., 2021. Crop yield estimation using deep learning based on climate big data and irrigation scheduling. Energies 14, 11. https://doi.org/10.3390/en14113004.

Aravind, K., Raja, P., 2020. Automated disease classification in (selected) agricultural crops using transfer learning. J.Automatika. J. Control, Measur. Electr. Comp. Communicat. 62, 260–272. https://doi.org/10.1080/00051144.2020.1728911.

Arsenovic, M., Karanovic, M., Sladojevic, S., Anderla, A., Stefanovic, D., 2019. Solving Current Limitations of Deep Learning Based Approaches for Plant Disease Detection. Symmetry. 11. MDPI, 7. https://doi.org/10.3390/sym11070939.

Barbedo, J.A., 2018. Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification. Comput. Electron. Agric. 153, 46–53. https://doi.org/10.1016/j.compag.2018.08.013.

Cao, J., Zhao, Z., Luo, Y., Zhang, L., Zhang, J., Li, Z., Tao, F., 2021. Wheat yield predictions at a county and field scale with deep learning, machine learning, and google earth engine. Eur. J. Agron. 123. https://doi.org/10.1016/j.eja.2020.126204.

Chen, J., Liu, Q., Gao, L., 2019. Visual tea leaf disease recognition using a convolutional neural network model. Symmetry 11, 3. https://doi.org/10.3390/sym11030343.

Chergui, N., Kechadi, T., McDonnell, M., 2020. The impact of data analytics in digital agriculture: A review. 2020 International Multi-Conference on: "Organization of Knowledge and Advanced Technologies". February 6–8, 2020 Tunis (Tunisia). IEEE, pp. 1–13 https://doi.org/10.1109/OCTA49274.2020.9151851.

Crane-Droesch, A., 2018. Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. Environ. Res. Lett. 13, 11. https://doi.org/10.1088/1748-9326/aae159.

Cruz, A., Luvisi, A., Bellis, L.D., Ampatzidis, Y., 2017. X-Fido: An Effective Application for Detecting Olive Quick Decline Syndrome with Deep Learning and Data Fusion. Technical Advances in Plant Science a Section of the Journal Frontiers in Plant Science. Frontiers https://doi.org/10.3389/fpls.2017.01741.

Ehret, D., Hill, B., Helmer, T., Edwards, D., 2011. Neural network modeling of greenhouse tomato yield, growth and water use from automated crop monitoring data. Comput. Electron. Agric. 79, 82–89. https://doi.org/10.1016/j.compag.2011.07.013.

Elavarasan, D., Vincent, D., Sharma, V., Zomaya, A., Srinivasan, K., 2018. Forecasting yield by integrating agrarian factors and machine learning models: a survey. Comput. Electron. Agric. 155, 257–282. https://doi.org/10.1016/j.compag.2018.10.024.

Feng, S., Gangal, V., Wei, J., Vosoughi, S., Chandar, S., Mitamura, T., Hovy, E., 2021. A survey on data augmentation approaches for nlp. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP*. - , pp. 968–988. https://doi.org/10.48550/arXiv.2105.03075.

Filippi, P., Jones, E., Wimalathunge, N., Pallegedara, D., Pozza, L., Ugbaje, S., Jephcott, T., Paterson, S., Whelan, B., Bishop, F., 2019. An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning. Precis. Agric. https://doi.org/10.1007/s11119-018-09628-4

Fukuda, S., Spreer, W., Yasunaga, E., Yuge, K., Sardsud, V., Muller, J., 2013. Random forests modelling for the estimation of mango (mangifera indica l. cv.Chok Anan) fruit yields under different irrigation regimes. J. Agric. Water Manag. 116, 142–150. https://doi.org/10.1016/j.agwat.2012.07.003.

Gong, L., Yu, M., Jiang, S., Cutsuridis, V., Pearson, S., 2021. Deep learning based prediction on greenhouse crop yield combined tcn and rnn. Sensors 21, 13. https://doi.org/10.3390/s21134537.

Gonzalez-Sanchez, A., Frausto-Solis, J., Ojeda-Bustamante, W., 2014. Predictive ability of machine learning methods for massive crop yield prediction. Span. J. Agric. Res. 12, 313–328. https://doi.org/10.5424/sjar/2014122-4439.

Guo, Y., Fu, Y., Hao, F., Zhang, X., Wu, W., Jin, X., Bryant, C., Senthilnath, J., 2021. Integrated phenology and climate in rice yields prediction using machine learning methods. Ecol. Indic. 120. https://doi.org/10.1016/j.ecolind.2020.106935.

Gyamerah, S., Ngare, P., Ikpe, D., 2020. Probabilistic forecasting of crop yields via quantile random forest and epanechnikov kernel function. Agric. For. Meteorol., 280. https://doi.org/10.1016/j.agrformet.2019.107808.

Han, J., Zhang, Z., Cao, J., Luo, Y., Zhang, L., Li, Z., Zhang, J., 2020. Prediction of winter wheat yield based on multi-source data and machine learning in china. Remote Sens. 12, 2. https://doi.org/10.3390/rs12020236.

Iwana, B., Uchida, S., 2021. An empirical survey of data augmentation for time series classification with neural networks. PLoS One 16. https://doi.org/10.1371/journal.pone.0254841.

Jeong, J., Resop, J., Mueller, N., Fleisher, D., Yun, K., Butler, E., Timlin, D., Shim, K., Gerber, J., Reddy, V., Kim, S., 2016. Random forests for global and regional crop yield predictions. PLoS One, 11. https://doi.org/10.1371/journal.pone.0156571.

Ji, R., Min, J., Wang, Y., Cheng, H., Zhang, H., Shi, W., 2017. In-season yield prediction of cabbage with a hand-held active canopy sensor. Sensors 17, 10. https://doi.org/10.3390/s17102287.

Ji, Z., Pan, Y., Zhu, X., Wang, J., Li, Q., 2021. Prediction of crop yield using phenological information extracted from remote sensing vegetation index. Sensors 21, 4. https://doi.org/10.3390/s21041406.

Kamal, K., Yin, Z., Wu, M., Wu, Z., 2019. Depthwise separable convolution architectures for plant disease classification. Comput. Electron. Agric. 165. https://doi.org/10.1016/j.compag.2019.104948.

Kamir, E., Waldner, F., Hochman, Z., 2020. Estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods. ISPRS J. Photogramm. Remote Sens. 160, 124–135. https://doi.org/10.1016/j.isprsjprs.2019.11.008.

Khaki, S., Wang, L., Archontoulis, S., 2020. A cnn-rnn framework for crop yield prediction. Front. Plant Sci. 10. https://doi.org/10.3389/fpls.2019.01750.

Kim, N., Ha, K., Park, N., Cho, J., Hong, S., Lee, Y., 2019. A comparison between major artificial intelligence models for crop yield prediction: case study of the midwestern united states, 2006–2015. ISPRS Int. J. Geo Inf. 8. https://doi.org/10.3390/ijgi8050240.

Klompenburg, T., Kassahun, A., Catal, C., 2020. Crop yield prediction using machine learning: A systematic literature review. Comput. Electron. Agric. 177. https://doi.org/10.1016/j.compag.2020.105709.

Kouadio, L., Deo, R., Byrareddy, V., Adamowski, J., Mushtaq, S., Nguyen, V.P., 2018. Artificial intelligence approach for the prediction of robusta coffee yield using soil fertility properties. Comput. Electron. Agric. 155, 324–338. https://doi.org/10.1016/j.compag.2018.10.014.

Liu, B., Zhang, Y., He, D., Li, Y., 2017. Identification of apple leaf diseases based on deep convolutional neural networks. Symmetry 10, 1. https://doi.org/10.3390/sym10010011.

Maimaitijiang, M., Sagan, V., Sidike, P., Hartling, S., Esposito, F., Fritschi, F., 2020. Soybean yield prediction from uav using multimodal data fusion and deep learning. Remote Sens. Environ. 237. https://doi.org/10.1016/j.rse.2019.111599.

Meroni, M., Waldner, F., Seguini, L., Kerdiles, H., Rembold, F., 2021. Yield forecasting with machine learning and small data: what gains for grains? Agric. For. Meteorol. 308-309. https://doi.org/10.1016/j.agrformet.2021.108555.

Nevavuori, P., Narra, N., Linna, P., Lipping, T., 2020. Crop yield prediction using multitemporal uav data and spatio-temporal deep learning models. Remote Sens. 12, 34. https://doi.org/10.3390/rs12234000.

Oliveira, I., Cunha, R., Silva, B., Netto, M., 2018. A scalable machine learning system for pre-season agriculture yield forecast. The 14th IEEE eScience Conference. IEEE https://doi.org/10.1109/eScience.2018.00131.

Olson, M., Wyner, A., Berk, R., 2018. Modern neural networks generalise on small data sets. 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montreal, Canada. 31. -, pp. 3619–3628.

Pantazi, X., Moshou, D., Alexandridis, T., Whetton, R., Mouazen, A., 2016. Wheat yield prediction using machine learning and advanced sensing techniques. J. Comput. Electron. Agric. 121, 57–65. https://doi.org/10.1016/j.compag.2015.11.018.

Pham, H., Awange, J., Kuhn, M., Nguyen, B., Bui, L., 2022. Enhancing crop yield prediction utilising machine learning on satellite-based vegetation health indices. Sensors 22, 3. https://doi.org/10.3390/s22030719.

Piekutowska, M., Niedbasł, G., Piskier, T., Lenartowicz, T., Pilarski, K., Wojciechowski, T., Pilarska, A., Czechowska-Kosacka, A., 2021. The application of multiple linear regression and artificial neural network models for yield prediction of very early potato cultivars before harvest. Agronomy 11, 5. https://doi.org/10.3390/agronomy11050885.

Rahman, M.M., Robson, A., 2020. Integrating landsat-8 and sentinel-2 time series data for yield prediction of sugarcane crops at the block level. Remote Sens. 12, 8. https://doi.org/10.3390/rs12081313.

Rezapour, S., Jooyandeh, E., Ramezanzade, M., Mostafaeipour, S., Jahangiri, M., Issakhov, A., Chowdhury, S., Techato, K., 2021. Forecasting rainfed agricultural production in arid and semi-arid lands using learning machine methods: a case study. Sustainability 13, 9. https://doi.org/10.3390/su13094607.

Sagan, V., Maimaitijiang, M., Bhadra, S., Maimaitiyiming, M., Brown, D., Sidike, P., Fritschi, F., 2021. Field-scale crop yield prediction using multi-temporal worldview-3 and planetscope satellite data and deep learning. ISPRS J. Photogramm. Remote Sens. 174, 265–281. https://doi.org/10.1016/j.isprsjprs.2021.02.008.

Sakamoto, T., 2020. Incorporating environmental variables into a modis-based crop yield estimation method for United States corn and soybeans through the use of a random forest regression algorithm. ISPRS J. Photogramm. Remote Sens. 160, 208–228. https://doi.org/10.1016/j.isprsjprs.2019.12.012.

Schwalbert, R., Amado, T., Corassa, G., Pott, L., Prasad, P., Ciampitti, I., 2020. Satellite-based soybean yield forecast: integrating machine learning and weather data for improving crop yield prediction in southern Brazil. Agric. For. Meteorol. 284. https://doi.org/10.1016/j.agrformet.2019.107886.

Shahhosseini, M., Hu, G., Archontoulis, S., 2020. Forecasting corn yield with machine learning ensembles. Front. Plant Sci. 11. https://doi.org/10.3389/fpls.2020.01120.

Shook, J., Gangopadhyay, T., Wu, L., Ganapathysubramanian, B., Sarkar, S., Singh, A., 2021. Crop yield prediction integrating genotype and weather variables using deep learning. PLoS One 16. https://doi.org/10.1371/journal.pone.0252402.

Shorten, C., Khoshgoftaar, T., 2019. A survey on image data augmentation for deep learning. J. Big Data 6, 60. https://doi.org/10.1186/s40537-019-0197-0.

Sladojevic, S., Arsenovic, M., Culibrk, A.A.D., Stefanovic, D., 2016. Deep neural networks based recognition of plant diseases by leaf image classification. Computat. Intellig. Neurosci. 2016, 2016. https://doi.org/10.1155/2016/3289801.

Stepanov, A., Dubrovin, K., Sorokin, A., Aseeva, T., 2020. Predicting soybean yield at the regional scale using remote sensing and climatic data. Remote Sens. 12, 1936 12 https://doi.org/10.3390/rs12121936.

Wang, A., Tran, C., Desai, N., Lobell, D., Ermon, S., 2018. Deep transfer learning for crop yield prediction with remote sensing data, in: Proceedings of the COMPASS'18. Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies. Menlo Park and San Jose, CA, USA. June 20–22. ACM.

Wang, Y., Zhang, Z., Feng, L., Du, Q., Runge, T., 2020. Combining multi-source data and machine learning approaches to predict winter wheat yield in the conterminous United States. Remote Sens. 12, 1232. https://doi.org/10.3390/rs12081232 8.

Wolanin, A., Mateo-García, G., Camps-Valls, G., Gómez-Chova, L., Meroni, M., Duveiller, G., Liangzhi, Y., Guanter, L., 2020. Estimating and understanding crop yields with explainable deep learning in the indian wheat belt. Environ. Res. Lett. 15, 2. https://doi.org/10.1088/1748-9326/ab68ac.

Xu, J., Ma, J., Tang, Y., Wu, W., Shao, J., Wu, W., Wei, S., Liu, Y., Wang, Y., Guo, H., 2020. Estimation of sugarcane yield using a machine learning approach based on uav-lidar data. Remote Sens. 12, 17. https://doi.org/10.3390/rs12172823.

Yamamoto, K., Togami, T., Yamaguch, N., 2017. Super-resolution of plant disease images for the acceleration of image-based phenotyping and vigor diagnosis in agriculture. Sensors 17, 11. https://doi.org/10.3390/s17112557.

You, J., Li, X., Low, M., Lobell, D., Ermon, S., 2017. Deep gaussian process for crop yield prediction based on remote sensing data. The Thirty-first AAAI Conference on Artificial Intelligence. AAAI publications, pp. 4559–4566.