Egyptian Journal of Soil Science

http://ejss.journals.ekb.eg/

Integrating Climate and Plant Variables with Machine Learning Models to Forecast Tomato Yield at Different Soil Moisture Levels

Nadia G. Abd El-Fattah[1], Mohamed S. Abd El-baki[1], Mohamed M. Ibrahim[1] and **Salah Elsayed[2,3*]**

[1]Agricultural Engineering Department, Faculty of Agriculture, Mansoura University, Mansoura 35516, Egypt; Nadia_gamal91@mans.edu.eg; mohamedsalah@mans.edu.eg (M.S.A); mohamed_maher@mans.edu.eg

[2]Agricultre Engineering, Evaluation of Natural Resources Department, Environmental Studies and Research Institute, University of Sadat City, Menoufia 32897, Egypt; salah.emam@esri.usc.edu.eg

[3]New Era and Development in Civil Engineering Research Group, Scientific Research Center, Al-Ayen University, Nasiriyah, Thi-Qar 64001, Iraq

ACCURATELY predicting crop yield under different environmental conditions and irrigation regimes plays a vital role in optimizing agricultural practices and ensuring food security. This research aims to develop a tomato yield (TY) estimation model using machine learning (ML) techniques, such as artificial neural networks (ANN), random forests (RF), and decision trees (DT), based on climate and plant variables. The climate variables include Growth Degree Days (GDD), Vapor Pressure Deficit (VPD), solar radiation, Total Sunshine Hours (N), Total Relative Humidity (TRH), and Reference Evapotranspiration (ETo). While plant variables include canopy water content (CWC), dry matter accumulation (DMA), N-sufficiency index (NSI), and Crop Evapotranspiration (ETc). Field experiments were conducted during 2022 and 2023 growing seasons, implementing three irrigation regimes: 100% (T100), 75% (T75), and 50% (T50) of the full irrigation requirements (FIR). The results showed that the highest TY was achieved at T100. In contrast, T75 resulted in a yield reduction of 25.85% and 28.42%, while T50 led to decreases of 54.74% and 55.76% compared to T100 during the first and second seasons, respectively. Also, ANOVA revealed no statistically significant differences in model performance. However, modest improvements in yield prediction accuracy can lead to substantial economic benefits for farmers. The RF model displayed even better accuracy, with an RMSE of 2.39-4.04 and 3.80-4.00 tons/ha, and an R^2 of 0.94-0.98 and 0.95 during the training and testing phases, respectively. These findings highlight the practicality and reliability of utilizing climate and plant variables in combination with ML models to effectively manage tomato crop production, particularly when facing limited water availability for irrigation.

Keywords: Water Stress, Climate Data, Plant Data, Ensemble Models, Yield Prediction, Artificial Neural Network.

1. Introduction

The tomato crop holds a prominent position as one of the most extensively cultivated vegetables worldwide. In recent years, there has been a significant increase in global tomato production, with a growth rate of approximately 10% (**Shalaby and El-Banna 2013**). Notably, Egypt has emerged as a key player in tomato production, securing its position as the sixth-largest tomato producer globally. With an impressive cultivation area spanning approximately 143,618 hectares, Egypt yielded a remarkable total production of about 6.28 million tons, according to the latest data from (**FAOSTAT 2022**). However, the arid climate and limited water resources in the region pose significant challenges to sustaining tomato production. To overcome these hurdles, deficit irrigation strategies have gained considerable attention, aiming to optimize water usage while maintaining satisfactory crop yields. This approach, emphasized by **El-Labad *et al.* (2019)**, has become increasingly prominent in addressing the unique water-related challenges faced in tomato cultivation. Climate variables play a crucial role in predicting crop yield since they directly influence plant growth and development. For instance, temperature, growing degree days (GDD), and vapor-pressure deficit (VPD) have been examined as climate variables that contribute to yield variability due to climate change, as studied (**Meng *et al.* 2017; Aboukota *et al., 2024*; *ElGhamry et al*., 2024)** . These variables provide valuable insights into growing conditions and are commonly used as the primary climate

*Corresponding author e-mail: salah.emam@esri.usc.edu.eg

Received: 29/07/2024; Accepted: 09/09/2024
DOI: 10.21608/EJSS.2024.308236.1829

variables for yield prediction, as mentioned by **Li *et al.* (2019)**. In addition to climate variables, the integration of plant-specific variables enhances the predictive accuracy of the models. Indicators such reference evapotranspiration (ET_O) and crop evapotranspiration (ET_C) provide valuable information about the water requirements of tomato crops (**Kizza *et al.*, 2016)**. The N-sufficiency index, which measures nitrogen availability and utilization efficiency in plants, is another important plant-specific variable for yield prediction, as highlighted by **Alordzinu *et al.* (2021)**. Additionally, dry matter accumulation, reflecting overall biomass production, is closely linked to tomato fruit yield (**El-Labad *et al.*, 2019)**. Furthermore, canopy water content, an indicator of plant water status, provides insights into the physiological response of tomato plants to deficit irrigation and its impact on yield, as explored yield **(Alordzinu *et al.*, 2021; Elsherpiny 2023; Kamara et al., 2023)**.

Predicting crop yield stands as a pivotal task in today's landscape for policymakers and farmers, crucial for ensuring food security and sustainability. Yet, this endeavor poses formidable challenges due to the intricate interplay among soil, plant, and environmental factors that influence crop productivity (**Khaki and Wang 2019)**. Traditional methods like crop growth models and statistical analyses struggle to adapt to the ever-changing biotic and abiotic influences on crop output **(Lobell and Burke 2010)**. Moreover, these conventional models demand copious amounts of data encompassing soil composition, climate patterns, crop specifics, and agricultural practices, along with substantial user expertise to calibrate the model accurately, as noted by **Shahhosseini *et al.* (2021)**. Thankfully, the advent of machine learning (ML) in agriculture has ushered in a revolutionary and refined approach to surmount the constraints inherent in crop forecasting across varying environmental contexts, a development extensively discussed **(Sridhara *et al.*, 2023)**. The application of computational models in ML has brought about a revolutionary transformation in agricultural practices, enabling a deeper understanding of crop production dynamics and facilitating optimal resource management. Among the various ML techniques, artificial neural networks (ANNs), decision trees, and random forests have emerged as formidable tools capable of capturing intricate relationships between input variables and output responses, as demonstrated by **Cedric *et al.* (2022)** and **López-Aguilar *et al.* (2020)**. By incorporating both climate and plant-specific variables, these models offer a comprehensive and holistic approach to accurately forecast crop yield, thereby enhancing the precision of predictions in agricultural contexts.

This innovative approach paves the way for predicting the yield of cultivated crops solely based on existing data. Such a methodology holds the promise of facilitating informed decisions regarding the selection of agricultural technologies, enhancing cropland management practices, assessing future production potential, and crafting climate-smart adaptation strategies to bolster food security. With this background, this study aims to (a) investigate the effects of water stress on some biophysical parameters and tomato yield under different irrigation regimes; (b) coupling plant variables with climate variables using ANNs, decision trees, and random forests to improve yield prediction under deficit irrigation regimes; and (c) comparing the predictive abilities of these models for tomato yield under deficit irrigation regimes to determine the best model.

2. Material and Methods

2.1 Experimental site

Field experiments were conducted for two consecutive spring growing seasons during 2022 and 2023 on a private farm at Talkha, Dakahlia, Egypt. The farm is located at 31.09° N, 31.38° E, and with an elevation of 17 meters. The experimental soil was classified as sandy clay in texture. It had a maximum rain infiltration rate of 30 mm/day.

2.2 Experimental irrigation system

The layout of the drip irrigation network is shown in **Fig. 1**. It includes a control head comprising a centrifugal water pump, a disk filter, a pressure gauge, control valves, and a Venturi-type injector. The main line consists of polyethylene (P.E.) pipes with a diameter of 75 mm. The sub-main line is made of P.E. pipes with a diameter of 63 mm. Lateral lines are also made of P.E. and have a diameter of 16 mm. Built-in emitters are used with an average discharge rate of 6 L/h at an operating pressure of 1 bar. The beginning of each lateral line was provided with a T-shaped 16 mm plastic valve to control the irrigation depth at the desired level for different irrigation treatments. The end of each lateral line was closed by an end cap. A drip irrigation system was constructed and tested before being used in the experimental location using equation (1), according to **Ella *et al.* (2013)**. The distribution uniformity (DU) was estimated to be 92%.

$$\mathrm{DU} = \frac{\text{average of the lowest quartile}}{\text{the average of all readings}} * 100\% \qquad (1)$$

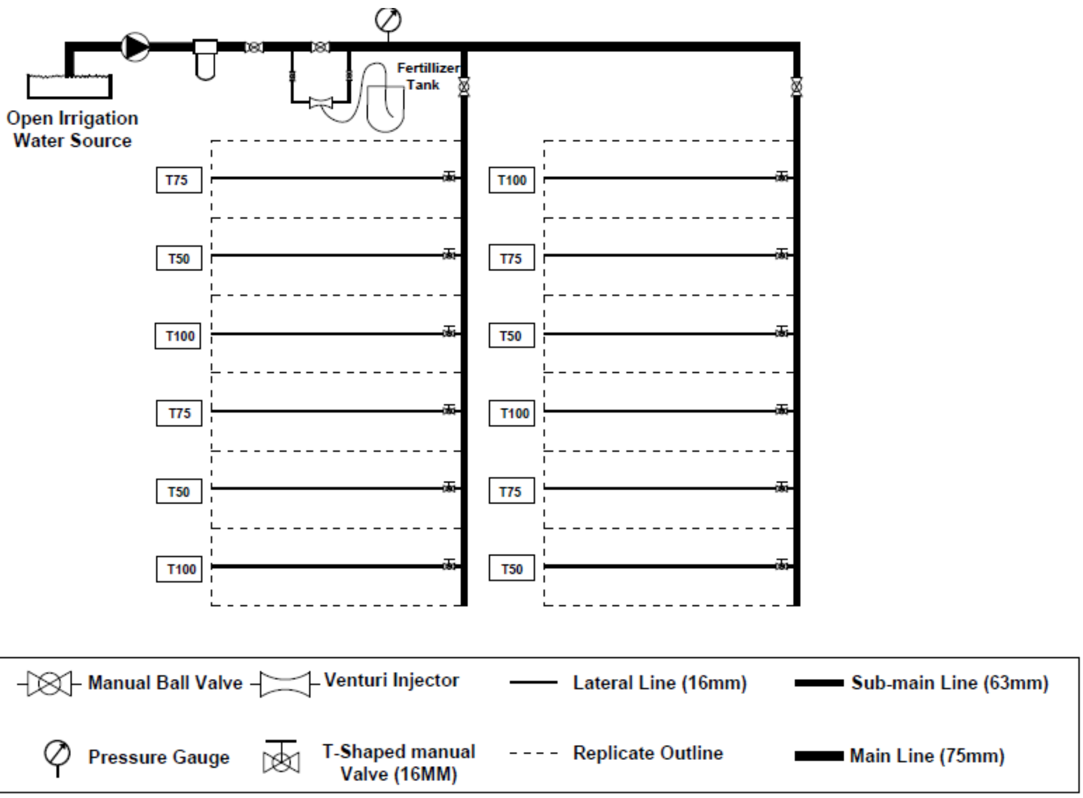

Fig. 1. The layout of experimental design for the different irrigation treatments.

2.3 Tomato plant, agronomic practices, and irrigation regimes

The cultivation of 'Gs12 F1' hybrid tomato seeds embarked on February 23rd after the completion of the initial growth phase, culminating in harvest on June 17th. The subsequent planting season commenced on March 3rd, concluding with harvest on June 25th. The entire growth cycle spanned 150 days, segmented into four distinct stages: initial (35 days), developmental (39 days), middle (46 days), and late (30 days). Throughout these stages, crop coefficients (K_c) were meticulously observed at 0.38, 1.10, 1.10, and 0.65, respectively, as outlined by **Noreldin et al. (2014)**. Drip irrigation was employed to nourish the tomato plants. The experimental setup utilizing a randomized complete block design with four replicates to mitigate spatial variability. Plant spacing within rows was maintained at 0.4 m. The lateral line spacing stood at 1.2 m. Each plot covered an area of 10.8 m^2 (9 m in length by 1.2 m in width). Irrigation commenced 15 days post-transplanting to ensure seedling survival. It continued throughout the growing season, except for the final 10 days when irrigation ceased. Various irrigation regimes, representing 100%, 75%, and 50% of the full irrigation requirements (FIR), were administered to investigate its impacts on biophysical parameters and tomato yield. Fertilizer applications adhered to the guidelines set forth by the Egyptian Ministry of Agriculture. All treatments receiving 357 kg/ha of nitrogen (N) in the form of urea (46.5% N), 60 kg/ha of phosphorus (P) as phosphoric acid (85% P_2O_5), and 238 kg/ha of potassium (K) as potassium sulphate (50% K_2O). The fertilizer was applied through the drip irrigation system utilizing a venturi injector over the course of the two growing seasons.

2.4 Crop evapotranspiration (ET_c)

The reference evapotranspiration (ET_o) values were calculated daily using the FAO Penman-Monteith equation (2), as outlined by **Allen et al. (1998)**. The CROWAT model was employed for this calculation, according to **Gabr (2022)**. Additionally, the crop evapotranspiration (ET_c) values were calculated by multiplying ET_o by the crop coefficient (K_c) for each growth stage, using equation (3), also based on **Allen et al. (1998)**.

$$ET_o = \frac{0.408\Delta(R_n - G) + \gamma(\frac{900}{T + 273}) * U_2(e_a - e_d)}{[\Delta + \gamma(1 + 0.34U_2)]} \qquad (2)$$

Where: ET_o: Reference evapotranspiration (mm/day); R_n: Net radiation at crop surface (MJ/m^2.day); G: Soil heat flux (MJ/m^2.day); T: Average temperature (°C); U_2: Wind speed measured at 2 m above ground (m/s); $e_a - e_d$: Vapor pressure deficit (kpa); Δ: Slope vapor pressure curve (kpa/°C); γ: Sychometric constant (kpa/°C).

$$\mathrm{ET_c = ET_o * K_c} \tag{3}$$

Where:
K_c: Crop coefficient (dimensionless).
ET_c: Crop Evapotranspiration (mm/day).

2.5 Irrigation water applied

The irrigation water applied (IWA) is defined as the amount to replenish crop water used for field capacity. It is calculated by the equation (4), according to **Abdulhadi and Alwan (2021)** for full irrigation requirements. The IWA was employed based on the three regimes: 100%, 75%, and 50% of the full irrigation requirements.

$$\mathrm{IWA} = \frac{\mathrm{d_n * S_e * S_m * K_r}}{\mathrm{E_a}} \tag{4}$$

d_n: The net depth computing using CROPWAT8.0 for full irrigation, mm.

S_e: Lateral spacing along the sub-main, m.

S_m: Dripper spacing along the lateral, m.

E_a: Irrigation application efficiency, (90%).

K_r: wetted area factor, was estimated as 0.33 using the equation (5) according to **YILDIRIM and BAHAR (2017)**:

$$\mathrm{K_r} = \frac{\mathrm{S_e}}{\mathrm{S_m}} \tag{5}$$

2.6 Calculating input variables for models

Our study used two distinct groups of data to calculate input variables for different machine learning (ML) models. The first group encompassed the sum of data from the first day after transplanting (DAT) to 67 DAT for the first season and 68 DAT for the second season. On the other hand, the second group entailed the sum of data from the first day after transplanting to 93 DAT for both seasons.

2.7 Climate variables

The necessary meteorological data can be obtained by downloading it from the NASA POWER | Data Access Viewer website: https://power.larc.nasa.gov/data-access-viewer/. NASA POWER utilizes satellite observations, which can provide a broad view of global climate patterns. These observations contribute to the accuracy of the data, especially for regions with limited ground-based monitoring stations. Reanalysis datasets used in NASA POWER combine observations with numerical models to generate consistent long-term climate records. This approach enhances the accuracy and completeness of the data, as documented by **Power (2022)**. The variables Growth Degree Days (GDD), Vapor Pressure Deficit (VPD), solar radiation, Total Sunshine Hours (N), Total Relative Humidity (TRH), and Reference Evapotranspiration (ET_o) are climate variables. These variables describe the environmental conditions. It encompassed factors such as temperature, radiation, evapotranspiration, and water availability, which are crucial for plant growth and development. The Growth Degree Days (GDD), Vapor Pressure Deficit (VPD), Total Sunshine Hours (N), and Solar Radiation (R_s) were calculated according to the following formula:

2.8 Growing degree days

Growing degree days (GDD) is a valuable metric that quantifies the cumulative heat exposure experienced by plants throughout the growing season, aiding in the assessment of its growth and development. The calculation of GDD involved the utilization of equation (6), according to **(Roberts *et al.* 2013)**.

$$\mathrm{GDD} = \sum_{\mathrm{i=1}}^{\mathrm{n}} (\mathrm{T_{mean} - T_b}) \tag{6}$$

where GDD is the growing degree-day (°C); T_{mean} is the mean air temperature (°C); T_b is the base temperature for tomato grown under open-field conditions = 7 °C, according to **Abdalhi *et al.* (2020)**.

2.9 Vapor pressure deficit

Vapor pressure deficit (VPD) represents the disparity between the moisture content present in the air and its saturation point. It indicates the capacity of the air to hold moisture. This essential metric exhibits an exponential correlation with temperature. As emphasized by **Roberts *et al.* (2013)**, elevated VPD values reflect increased water demands, which are of paramount importance for optimal photosynthesis. The calculation of VPD involved the utilization of equation (7), according to **Roberts *et al.* (2013)**.

$$\mathrm{VPD} = e^{\left(\frac{17.269\mathrm{T_{max}}}{237.3+\mathrm{T_{max}}}\right)} - e^{\left(\frac{17.269\mathrm{T_{min}}}{237.3+\mathrm{T_{min}}}\right)} \qquad (7)$$

Where: VPD: Vapor pressure deficit (ºC); $\mathrm{T_{min}}$ and $\mathrm{T_{max}}$ are the daily minimum and maximum temperatures (°C), respectively.

2.10 Total sunshine hours (N)

The total sunshine hours (N) can be calculated using the equation (8), as provided by **Duffie and Beckman (1980)**.

$$\mathrm{N} = \left(\frac{2}{15}\right) \arccos(-\tan(\delta)\tan(\varphi)) \qquad (8)$$

In this equation, φ represents the latitude of the study site, and δ corresponds to the solar declination angle. The solar declination angle (δ) can be obtained using the equation (9), as detailed by **Duffie and Beckman (1980)**.

$$\delta = 23.45 \sin\big(0.9863(284 + \mathrm{n})\big) \qquad (9)$$

Where, n denotes the number of days from the 1st of January.

2.11 Solar radiation

The solar radiation ($\mathrm{R_s}$) can be calculated by equation (10), according to **Allen *et al.* (1998)**.

$$\mathrm{R_s} = \mathrm{K_{Rs}} * (\mathrm{T_{mean}})^{0.5} * \mathrm{R_a} \qquad (10)$$

Where: $\mathrm{R_s}$: solar radiation (MJ m^{-2} d^{-1}); $\mathrm{R_a}$: extraterrestrial radiation (MJ m^{-2} d^{-1}); $\mathrm{T_{mean}}$: the daily mean temperatures (°C); $\mathrm{K_{Rs}}$: adjustment coefficient (°$C^{-0.5}$) for ‘interior’ locations, where land mass dominates and air masses are not strongly influenced by a large water body, $\mathrm{K_{Rs}} \cong 0.16$.

The extraterrestrial radiation ($\mathrm{R_a}$) can be obtained using the equation (11) according to **Allen *et al.* (1998)**.

$$\mathrm{R_a} = \frac{\mathrm{R_{so}}}{(0.75 + 2 * 10^{-5} * \mathrm{z})} \qquad (11)$$

Where: z: elevation above sea level (m); $\mathrm{R_{so}}$: Clear-sky shortwave radiation (MJ m^{-2} d^{-1}).

2.12 Plant variables

The variables canopy water content (CWC), dry matter accumulation (DMA), N-sufficiency index (NSI), and Crop Evapotranspiration ($\mathrm{ET_c}$) are plant-specific variables. These variables are directly linked to the physiological characteristics and growth of the plant. Measurements for these variables were conducted during both the flowering stage (67 days after transplanting [DAT] in the first season and 68 DAT in the second season) and the fruit-ripening stage (93 DAT for both seasons). For each treatment, measurements were taken from four plants, following its respective adopted formulas for accurate estimation.

2.13 Canopy water content

Canopy water content (CWC) refers to the proportion of water present in the plant canopy. CWC serves as an indicator of the plant's hydration level and transpiration rate. The calculation of CWC can be performed using equation (12), as outlined by **Semananda *et al.* (2016)**.

$$\mathrm{CWC} = \left(\frac{\mathrm{FW} - \mathrm{DW}}{\mathrm{FW}}\right) * 100\% \qquad (12)$$

CWC: represents the canopy water content, %.

FW: corresponds to the fresh biomass weight of the plant canopy, gm.

DW: denotes the dry biomass weight of the plant canopy, gm.

2.14 Dry matter accumulation

After examining the SPAD values, the dry matter accumulation (DMA) weight was estimated for the same plants following the procedures described by **Semananda *et al.* (2016)**. These plants were carefully cut and subjected to dehydration in an oven set at 105 °C for a period of 24 hours. This drying process ensured the removal of any moisture present in the plants. Subsequently, the dried plants were weighed using an electronic balance.

2.15 N-sufficiency index

Leaf relative chlorophyll content was assessed using a handheld Konica Minolta SPAD 502 chlorophyll meter. SPAD readings were conducted on the most recent fully expanded leaf. The readings were taken at a specified position of leaf approximately midway between the leaf edge and the midpoint of the leaf, as suggested by **Bai and Purcell (2018)**. The equation (13) was employed to convert the SPAD data into an N-sufficiency index, according to **Bausch *et al.* (2004)**. This conversion allows for a standardized assessment of nitrogen sufficiency in plants.

$$\mathrm{NSI} = \frac{\mathrm{SPAD_{target}}}{\mathrm{SPAD_{reference}}} \qquad (13)$$

2.16 Tomato fruit yield data

To gather yield data, 8 plants per treatment were randomly chosen and marked for identification. These designated plants were consistently utilized for yield measurements during each pick. Tomato were picked twice, first at 109 and 115 days after transplanting (DAT), and then at 107 and 115 DAT in both the 2022 and 2023 seasons. The total yield of tomato fruits for each treatment was computed by averaging the weights obtained from four replicates, quantified in tons per hectare.

2.17 Machine learning methods for tomato yield prediction

Three machine learning (ML) models were developed namely, an artificial neural network (ANN), a random forest (RF), and a decision tree (DT). These models were created using the Python scikit-learn library and Spyder software to prognosticate tomato yield under deficit irrigation conditions, as shown in **Fig 2**. The input variables for these models encompassed plant and climate data. The dataset was randomly partitioned into training (70%) and testing (30%) subsets a strategy in line with the methodologies of with a fixed random state of 0. This method was employed to regulate randomness and sampling variables during the node-splitting process. Numerous studies have employed analogous methodologies in comparable settings, as evidenced by prior research (**Cedric et al. 2022**). Hyperparameters were predefined before the model training phase. It is crucial for determining model performance. A 5-fold cross-validation approach in conjunction with the grid-search method within the scikit-learn library was employed on the training dataset. It was used to optimize the performance and generalization capabilities of the three ML models. This method facilitated the exploration of diverse hyperparameter combinations. The model exhibiting the most optimal performance was selected based on the lowest RMSE and the highest R^2 value.

2.18 Artificial neural network (ANN)

ANN models were developed consisting of input neural layer, hidden neural layers, and output neural layer. The neurons known as perceptron are similar to multiple linear regression. Stochastic gradient descent (SGD) is chosen as the optimization method, as indicated in equation (14). It is employed to iteratively adjust the connection weights and minimize the discrepancy between the predicted and actual values, as described by **Oymak (2019)**. The hyperparameters that need tuning include the number of neurons in each layer which ranged from 2 to 10. The number of hidden layers, which ranged from 1 to 5. Also, the activation functions that are shown in **Table 1**. The structure of an ANN is typically determined through experience and testing, as mentioned by **Mijwel (2021)**.

$$\theta_{j+1} \coloneqq \theta_j - \alpha.\left(Y_a{}^{(i)} - Y_p{}^{(i)}\right).x_j{}^{(i)} \qquad (14)$$

Θ_{j+1}: Weights of next iteration, Θ_j: Weights of current iteration; α: Learning rate; $x_j^{(i)}$: input feature; $Y_a^{(i)}$: Actual value; $Y_p^{(i)}$: Predict value.

Table 1. Activation Function, as stated by Sharma *et al.* (2017).

Name	Equations
Hyperbolic Tangent (Tanh)	$f(x) = \frac{(e^x - e^{-x})}{(e^x + e^{-x})}$
Logistic (Sigmoid)	$f(x) = \frac{1}{1 + e^{-x}}$
Rectified Linear Unit (ReLU)	$f(x) = \max(0, x)$
Linear (Identify)	$f(x) = x$

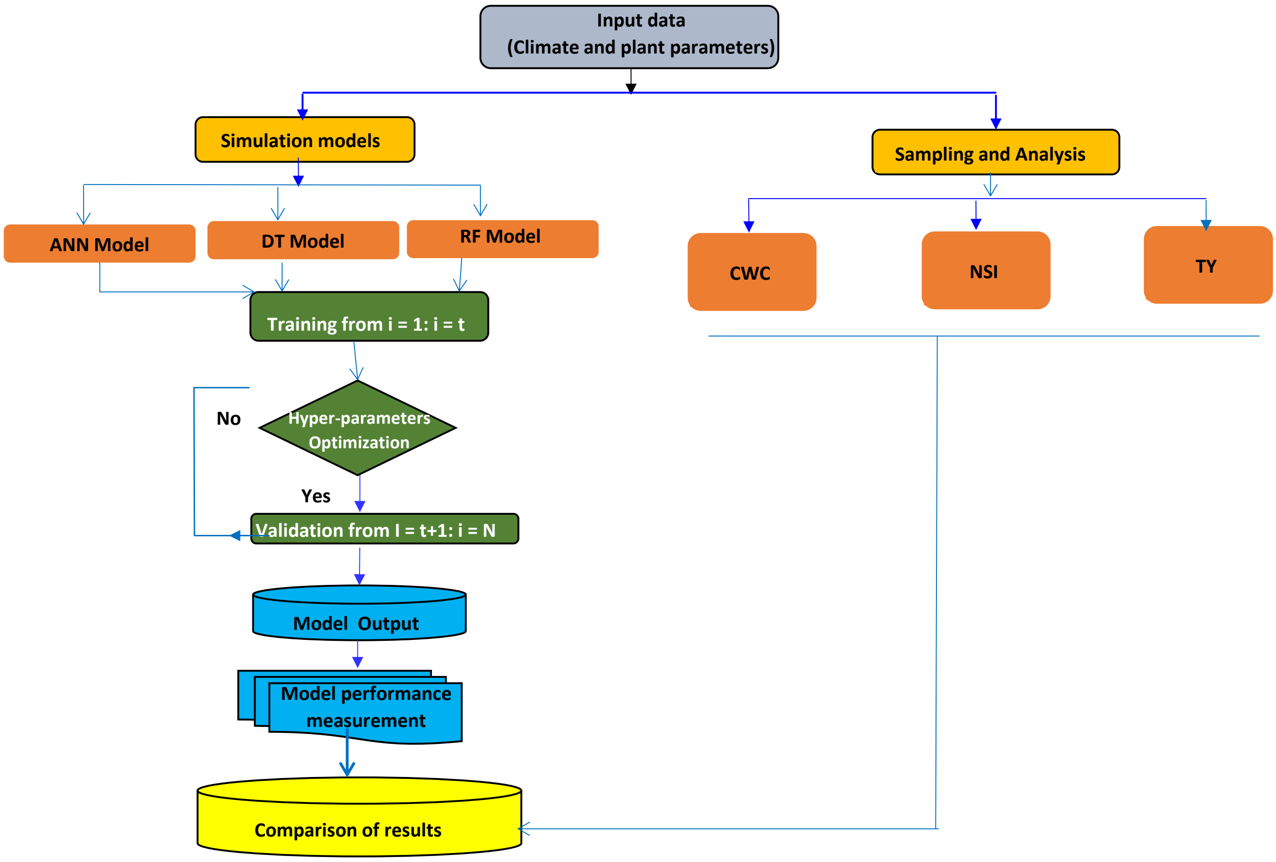
Input data
(Climate and plant parameters)
Simulation models
Sampling and Analysis
ANN Model
DT Model
RF Model
CWC
NSI
TY
Training from i = 1: i = t
No
Hyper-parameters Optimization
Yes
Validation from I = t+1: i = N
Model Output
Model performance measurement
Comparison of results

Fig. 2. Schematic diagram of the methodology presented in this study.

2.19 Decision tree (DT)

The decision tree algorithm is made up of leaf nodes, decision nodes, branches, and a root node. It is organized like a tree. The root node starts the tree. The decision nodes make decisions that decide the path, which moves from one node to another. The decision nodes end with the leaf nodes, **Han et al. (2022)**. Regression rules are easily created using decision trees. Because the DT doesn't require parameter setting or domain expertise, it is appropriate for exploratory knowledge discovery. During training, hyperparameter optimization and the optimal parameters were used to create the top-level model **Xia et al. (2017)**. Two-key hyperparameters including the maximum depth and criterion equation were taken into account during training to optimize the decision tree model. The maximum depth of the tree was varied from 1 to 10. The criterion used to judge the quality of a split. The criterion options were the mean squared error (MSE) and mean absolute error (MAE) methods (refer to equations 13 and 14), as described by **Ahmad *et al.* (2018)**.

$$\mathrm{MSE} = \frac{\sum_{i=1}^{N}(Y_a - Y_p)^2}{N} \qquad (15)$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{N} \left| Y_a - Y_p \right|}{N} \qquad (16)$$

2.20 Random forest (RF)

Random forest is a widely used ML algorithm employed for classification or regression tasks. It utilizes a combination of multiple decision trees to enhance prediction accuracy. Its effectiveness has been demonstrated in various research domains, including crop prediction **Abbas *et al.* (2020)**. A random forest comprises an ensemble of decision trees generated from random subsets of the available data. Three key hyperparameters, including the number of trees, maximum depth, and criterion function, were considered. The number of trees in the forest was varied from 1 to 20. The maximum depth of individual trees was varied from 1 to 10. The criterion functions were included the MSE and the MAE methods (refer to equations 13 and 14). By aggregating predictions from multiple trees, the random forest output is evaluated by averaging the results, as suggested by **Breiman (2001)**. This ensemble technique significantly improves the overall performance and the model's ability to generalize to unseen data.

2.21 Models' evaluation

The evaluation of the three models (ANN, RF, and DT) was conducted utilizing metrics including the R^2 and the RMSE value, as detailed in equations (17) and (18). These metrics were instrumental in quantifying the variance between the actual values and the estimated values produced by the models.

$$R^2 = \frac{\sum (Y_a - Y_p)^2}{\sum (Y_a - \bar{Y})^2} \qquad (17)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (Y_a - Y_p)^2} \qquad (18)$$

Where: Y_a, Y_p, and $\bar{Y}$: represents the actual value, predict value, and mean value, respectively. N: represent the number of data.

2.22 Statistical Analyses

The experiment was laid out in a randomized complete block design (RCBD) with four replicates. All collected data were subjected to analysis of variance (ANOVA) in order to examine the response of plant variables and tomato yield to different irrigation treatments. SPSS statistical software package version 28.0 was used to analyze the data. Significantly different means were separated using Tukey post-hoc test at the $P \leq 0.05$ level of probability.

3. Result

3.1 Plant variables

Irrigation regimes started 15 days after transplanting (DAT) for tomatoes to ensure the survival rate of the seedlings. Following this, irrigation regimes were applied for the rest of growing season, with the exception of the last 10 days for the tomato crop before harvest, when irrigation was stopped. Before initiating the irrigation regimes, equal depths of ET_c were added to each treatment during the first 15 days (53.01 mm in the first season and 75.65 mm in the second season). During the first season, the first group period received total ET_c depths of 367.40 mm, 288.80 mm, and 210.21 mm, corresponding to 100%, 75%, and 50% of FIR respectively. In the second group period of the same season, the tomato plants received 586.10 mm, 452.83 mm, and 319.56 mm for the respective regimes. Moving on to the second season, during the first group period, the tomato plants received total ET_c depths of 448.10 mm, 354.99 mm, and 261.88 mm, corresponding to 100%, 75%, and 50% of FIR respectively. In the second group period of the same season, the tomato plants received 681.30 mm, 529.89 mm, and 378.48 mm for the regimes, respectively.

A one-way ANOVA was performed to compare the effect of deficit irrigation regimes on canopy water content (CWC), Dry Matter Accumulation (DMA), and N-sufficiency Index (NSI). A one-way ANOVA revealed a statistically-significant difference in average CWC according to deficit irrigation regimes ($F(2)= 234.604$, $p <$

0.0001) and ((F(2)= 736.055, p < 0.0001) during flowering and fruit ripening stages, respectively. A Tukey post-hoc test revealed significant pairwise differences between 100% of FIR and 75% of FIR, with an average difference of 0.975 and 1.184% (p < 0.0001), between 100% of FIR and 50% of FIR, with an average difference of 4.20 and 4.768% (p < 0.0001), and between 75% of FIR and 50% of FIR, with an average difference of 3.22 and 3.584% (p < 0.0001) during flowering and fruit ripening stages, respectively. The results depicted in **Table 2** reveal that the highest CWC values were recorded when utilizing 100% of FIR (87.63 and 86.79%), while the lowest values were observed at 50% of FIR (82.95 and 82.57%). Comparatively, the 75% of FIR (86.46 and 85.81%) showed a slight reduction in CWC when compared to the 100% of FIR during flowering and fruit ripening stages, respectively. Also, a one-way ANOVA revealed a statistically-significant difference in average DMA according to deficit irrigation regimes (F(2)= 92.196, p < 0.0001) and ((F(2)= 315.438, p < 0.0001) during flowering and fruit ripening stages, respectively. A Tukey post-hoc test revealed significant pairwise differences between 100% of FIR and 75% of FIR, with an average difference of 25.51 and 19.53 gm/plant (p < 0.0001), between 100% of FIR and 50% of FIR, with an average difference of 51.67 and 47.30 gm/plant (p < 0.0001), and between 75% of FIR and 50% of FIR, with an average difference of 26.16 and 27.77 gm/plant (p < 0.0001) during flowering and fruit ripening stages, respectively. The findings presented in **Table 2** highlight that the highest DMA weights were observed when utilizing 100% of FIR (104.32 and 103.51 gm/plant) during the flowering and fruit ripening stages in both seasons, respectively. Remarkably, when employing 75% of FIR, the DMA weights experienced reductions of 18.20% and 25.54% at the flowering and fruit ripening stages in both seasons, respectively. Furthermore, at 50% of FIR, there was a more substantial decrease of 45.17% and 51.92% compared to the utilization of 100% FIR at the flowering and fruit ripening stages in both seasons, respectively. A one-way ANOVA revealed a statistically-significant difference in average NSI according to deficit irrigation regimes (F(2)= 149.442, p < 0.0001) and ((F(2)= 168.927, p < 0.0001) during flowering and fruit ripening stages, respectively. A Tukey post-hoc test revealed significant pairwise differences between 100% of FIR and 75% of FIR, with an average difference of 0.07 and 0.04 (p = 0.001), between 100% of FIR and 50% of FIR, with an average difference of 0.26 and 0.14 (p < 0.0001), and between 75% of FIR and 50% of FIR, with an average difference of 0.19 and 0.18 (p < 0.0001) during flowering and fruit ripening stages, respectively. The findings observed in **Table 2** highlight that during the flowering stage, the highest NSI values were recorded in 75% of FIR (1.01) for both seasons, followed by 100% of FIR (0.97) and 50% of FIR (0.84). Regarding the fruit ripening stage, the highest NSI values were observed at 50% of FIR (1.18), followed by 75% of FIR (0.99) and 100% of FIR (0.92) in both seasons.

Table 2. Canopy water content (CWC), dry matter accumulation (DMA), and N-sufficiency index (NSI) of tomato crop in both seasons under different irrigation regimes (50%, 75%, and 100%) of FIR.

Plant Variables	Regimes	Flowering Stage	Fruit Ripening Stage
CWC (%)	100% of FIT	87.63±0.44a	86.79±0.10a
	75% of FIT	86.46±0.38b	85.81±0.73b
	50% of FIT	82.95±0.50c	82.57±0.16c
DMA (gm/plant)	100% of FIT	104.32±5.57a	103.51±6.22a
	75% of FIT	85.34±4.73b	77.07±8.45b
	50% of FIT	57.20±3.26c	49.77±4.68c
NSI (dimensionless)	100% of FIT	0.97±0.02b	0.92±0.04c
	75% of FIT	1.01±0.01a	0.99±0.02b
	50% of FIT	0.84±0.02c	1.18±0.03a

3.2 Climate variables

In the first season, the first group exhibited Growth Degree Days (GDD) at 727.90 °C, Vapor Pressure Deficit (VPD) at 235.72 °C, Total Sunshine Hours (N) at 816.50 hours, Total Relative Humidity (TRH) at 4121.00, Solar Radiation (R_S) at 1508.27 MJ m^{-2} d^{-1}, and Reference Evapotranspiration (ET_o) at 327.43 mm. While, the second group data during the first season demonstrated GDD at 1185.80 °C, VPD at 375.13 °C, N at 1168.40 hours, TRH at 5475.00, (R_S) at 2329.67 MJ m^{-2} d^{-1}, and ET_o at 521.80 mm. Moving on to the second season, the first group data showed that the GDD was recorded at 954.55 °C, VPD at 275.49 °C, N at 846.5 hours, TRH at 3529.00, R_S at 1733.34 MJ m-2 d^{-1}, and ET_o at 400.17 mm. While, the second group data during the second season exhibited that the GDD was recorded at 1448.70 °C, VPD at 421.81 °C, N at 1189.30 hours, TRH at 4755.00, R_S at 2569.18 MJ m-2 d^{-1}, and ET_o at 608.17 mm.

3.3 Tomato fruit yield

A one-way ANOVA was performed to compare the effect of deficit irrigation regimes on tomato yield. A one-way ANOVA revealed that there was a statistically significant difference in mean tomato yield between at least two treatments (F(2, 51) = 464.252, $p < 0.0001$). The effect size, eta squared (η^2), was 0.948, indicating a large effect. Tukey's HSD post hoc test showed that the T100 scored significantly higher than both T75 (p = 0.0001, 95% C.I. = [17.47, 24.16]) and T50 (p = 0.0001, 95% C.I. = [38.88, 45.58]). The T75 scored significantly higher than the T50 (p = 0.0001, 95% C.I. = [18.07, 24.76]). These findings suggest that T100 leads to the highest tomato yield, followed by T75, and lastly, T50. The effect size confirms these differences are practically significant. **Table 3** presents the tomato yield values for both seasons. The highest tomato yield was 77.44 ton/ha and 75.50 ton/ha at 100% of FIR, followed by 75% of FIR with yield of 57.42 ton/ha and 54.05 ton/ha. Conversely, the lowest yield recorded was 35.05 ton/ha and 33.40 ton/ha for 50% of FIR during first and second seasons, respectively. During both seasons, 75% of FIR resulted in a reduction in yield by 25.85% and 28.42% respectively, while 50% of FIR led to a decrease of 54.74% and 55.76% compared to 100% of FIR, during first and second seasons, respectively.

Table 3. The effect of deficit irrigation on tomato fruit yield (ton/ha) in both season.

Regimes	First Season	Reduction, (%)	Second Season	Reduction, (%)
100% of FIT	77.44±3.76a	0	75.50±2.77a	0
75% of FIT	57.42±5.67b	25.85	54.05±2.44b	28.42
50% of FIT	35.05±6.19c	54.74	33.40±3.13c	55.76

3.4 ML-models performance to predict tomato yield

3.4.1 Tomato yield prediction using first group data

The study, utilizing data from the first day after transplanting (DAT) to 67 DAT for the first season and 68 DAT for the second season, showcased the remarkable predictive capabilities of three machine learning (ML) models in forecasting tomato yield with impressive accuracy. **Table 4** delineates the performance of various ML models in predicting tomato yield during both the training and testing phases, while **Fig. 3** specifically illustrates the outcomes observed in the testing phase. During the training phase, the artificial neural network (ANN-TFY1) model attained an impressive R^2 value of 0.96 and a RMSE of 3.45 ton/ha. Constructed with a single hidden layer comprising 8 neurons and utilizing the ReLU activation function over 500 iterations, as depicted in **Fig. 4**. The random forest (RF-TFY1) model outperformed the others, achieving an R^2 of 0.97 and the lowest RMSE of 2.78 ton/ha. This model, comprising 10 trees with a maximum depth of 4 and utilizing squared error as the criterion function, exhibited exceptional accuracy. In contrast, the decision tree (DT-TFY1) model displayed slightly lower performance, with an R^2 of 0.95 and an RMSE of 3.81 ton/ha. The DT-TFY1 model, with a maximum depth of 2 and employing squared error as the criterion function, showcased respectable predictive capabilities. Transitioning to the testing phase, all three models sustained commendable accuracy levels. The ANN-TFY1 model yielded an R^2 of 0.95 and an RMSE of 3.98 ton/ha. The RF-TFY1 model achieved an R^2 of 0.95 with an RMSE of 3.80 ton/ha, emerging as the top performer in terms of RMSE. On the other hand, the DT-TFY1 model yielded an R^2 of 0.94 and an RMSE of 4.15 ton/ha. This study conducted a one-way ANOVA to compare the performance of three ML models. The one-way ANOVA revealed that there was not a statistically significant difference in the performance of the three ML models during the training and testing phases. Nevertheless, these results do not detract from the overarching proficiency of the ML models in accurately forecasting tomato yields. Noteworthy is the exceptional performance of the RF-TFY1 model, which consistently registered the lowest RMSE values across both the training and testing phases. This underscores the RF-TFY1 model's exceptional capability in predicting tomato yields, despite the absence of statistically significant distinctions observed among the models.

Table 4. ML models Performance for tomato yield prediction after training and testing using first group data.

Models	Training		Testing	
	R^2	RMSE (ton/ha)	R^2	RMSE (ton/ha)
ANN-TFY1	0.96***	3.45	0.95***	3.98
RF-TFY1	0.97***	2.78	0.95***	3.80
DT-TFY1	0.95***	3.81	0.94***	4.15

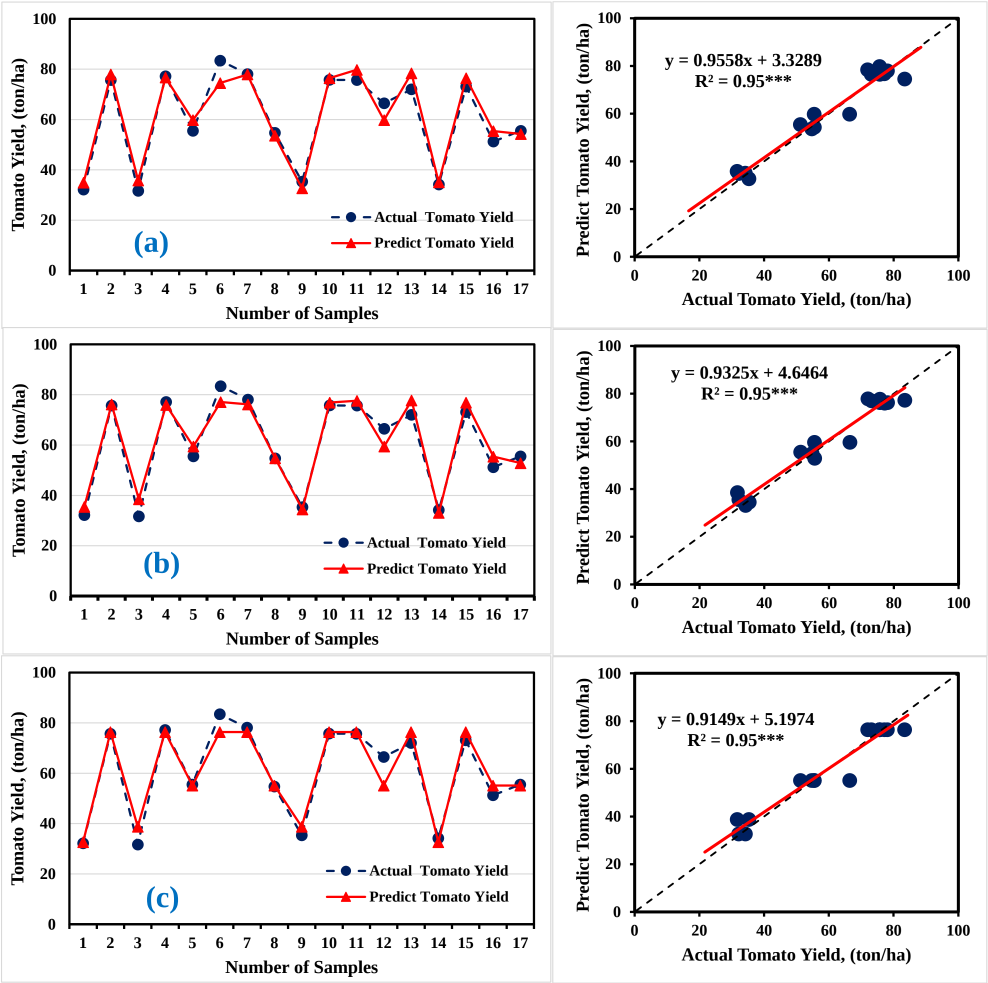

Fig. 3. Comparative Analysis of (a) Artificial Neural Networks (ANN), (b) Random Forest (RF), and (c) Decision Trees (DT) for tomato yield prediction (ton/ha) during testing using first group data.

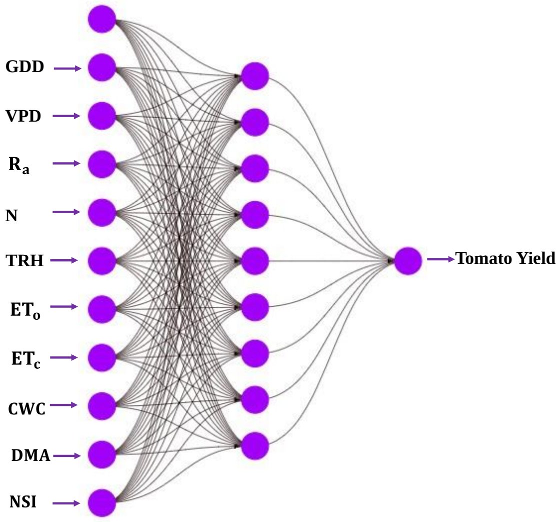

Fig. 4. ANN architecture for tomato yield prediction using first group data.

3.4.2 Tomato yield prediction using second group data

The research employed data spanning from the first day after transplanting (DAT) to 93 DAT across both seasons, showcasing the remarkable predictive capabilities of three machine learning (ML) models in forecasting tomato yield with impressive accuracy. **Table 5** details the performance of different ML models in predicting tomato yield during both the training and testing phases, while **Fig. 5** specifically illustrates the outcomes observed in the testing phase. In the training phase, the artificial neural network (ANN-TFY2) model displayed marginally lower accuracy compared to the other models, yielding an R^2 value of 0.95 and a RMSE of 3.68 ton/ha. Constructed with a single hidden layer comprising 6 neurons, employing the hyperbolic tangent (tanh) activation function, and undergoing 500 iterations, as depicted in **Fig. 6**. The random forest (RF-TFY2) model outshone the others in the training phase, achieving an impressive R^2 of 0.98 and the lowest RMSE of 2.39 ton/ha. Comprising 10 trees with a maximum depth of 4 and utilizing squared error as the criterion function, the RF-TFY2 model demonstrated exceptional accuracy. The decision tree (DT-TFY2) model also showcased high accuracy during training, achieving an R^2 of 0.96 and an RMSE of 3.40 ton/ha. With a maximum depth of 2 and employing squared error as the criterion function, the DT-TFY2 model exhibited robust predictive capabilities. Transitioning to the testing phase, all three ML models maintained reasonably good accuracy levels. The ANN-TFY2 model yielded an R^2 of 0.94 and an RMSE of 4.33 ton/ha. The RF-TFY2 model achieved an R^2 of 0.95 with an RMSE of 3.97 ton/ha. Similarly, the DT-TFY2 model presented an R^2 of 0.95 and an RMSE of 3.75 ton/ha. The study conducted a one-way ANOVA to compare the performance of three ML models. The analysis revealed that there was not a statistically significant difference in the performance of the three ML models during both the training and testing phases. However, these results do not diminish the overall proficiency of the ML models in accurately predicting tomato yields. Of particular note is the outstanding performance of the RF-TFY2 model, which consistently demonstrated the lowest RMSE values during the training phases. The RF-TFY2 and DT-TFY2 models predict with nearly identical R^2 values on the testing phase. The consistency and accuracy of the RF-TFY2 model underscore its ability to forecast tomato yields, despite the lack of statistically significant differences observed among the models.

Table 5. ML models Performance for tomato yield prediction after training and testing using second group data

Models	Training		Testing	
	R^2	RMSE (ton/ha)	R^2	RMSE (ton/ha)
ANN-TFY2	0.95***	3.68	0.94***	4.33
RF-TFY2	0.98***	2.39	0.95***	3.97
DT-TFY2	0.96***	3.40	0.95***	3.75

3.4.3 Tomato yield prediction using merged data from both groups

The study involved three machine learning (ML) models, namely the artificial neural network (ANN-TFY3), random forest (RF-TFY3), and decision tree (DT-TFY3), showcasing its adeptness in predicting tomato yield with remarkable accuracy. **Table 6** presents the performance of these ML models for predicting tomato yield in both the training and testing phases, while **Fig. 7** specifically illustrates the outcomes observed during the testing phase. During the training phase, the ANN-TFY3 model achieved an R^2 value of 0.95 and a RMSE of 4.00 ton/ha, underscoring its robust accuracy. Constructed with a single hidden layer comprising 9 neurons, utilizing the rectified linear unit (relu) activation function, and undergoing 500 iterations, as depicted in **Fig. 8**. The RF-TFY3 model attained an R^2 of 0.94 with an RMSE of 4.04 ton/ha during the training phase. Comprising 10 trees, with a maximum depth of 2 and employing squared error as the criterion function, the RF-TFY3 model showcased respectable accuracy. Similarly, the DT-TFY3 model reached an R^2 of 0.95 with an RMSE of 3.91 ton/ha during training. With a maximum depth of 2 and utilizing squared error as the criterion function, the DT-TFY3 model demonstrated commendable predictive performance. Transitioning to the testing phase, all three models maintained good accuracy levels. The ANN-TFY3 model exhibited an R^2 of 0.95 and an RMSE of 3.98 ton/ha, highlighting its efficacy in predicting tomato yield. The RF-TFY3 model obtained an R^2 of 0.95 and an RMSE of 4.00 ton/ha, showcasing consistent performance. On the other hand, the DT-TFY3 model yielded an impressive R^2 of 0.96 and an RMSE of 3.59 ton/ha during testing, indicating its strong predictive capabilities. The study conducted a one-way ANOVA to compare the performance of three ML models. The analysis revealed that there was not a statistically significant difference in the performance of the three ML models during both the training and testing phases. However, these results do not diminish the overall proficiency of the ML models in accurately predicting tomato yields. Notably, the exceptional performance of the DT-TFY3 model stood out, consistently demonstrating the lowest RMSE values throughout both the training and testing phases. The consistency and accuracy of the DT-TFY3 model highlight its capability in forecasting tomato yields, despite the absence of statistically significant differences observed among the models.

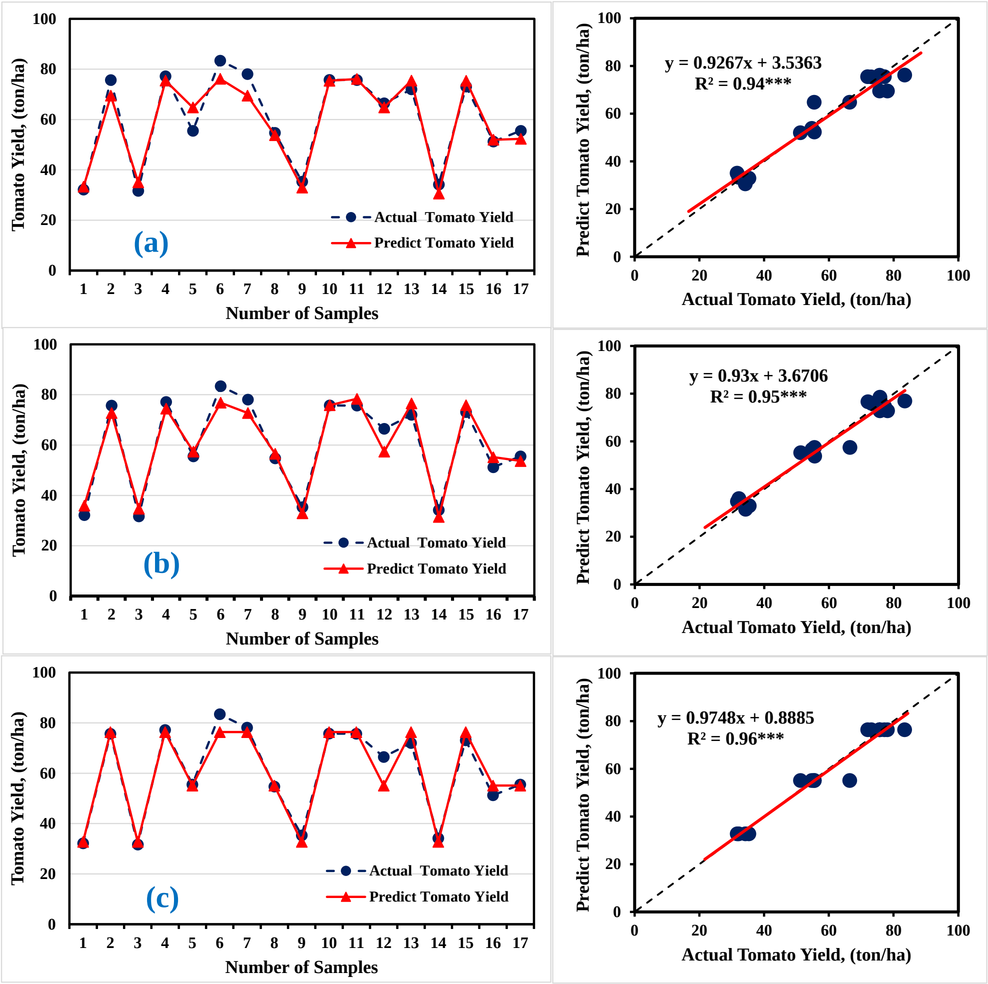

Fig. 5. Comparative Analysis of (a) Artificial Neural Networks (ANN), (b) Random Forest (RF), and (c) Decision Trees (DT) for tomato yield Prediction (ton/ha) during testing using second group data.

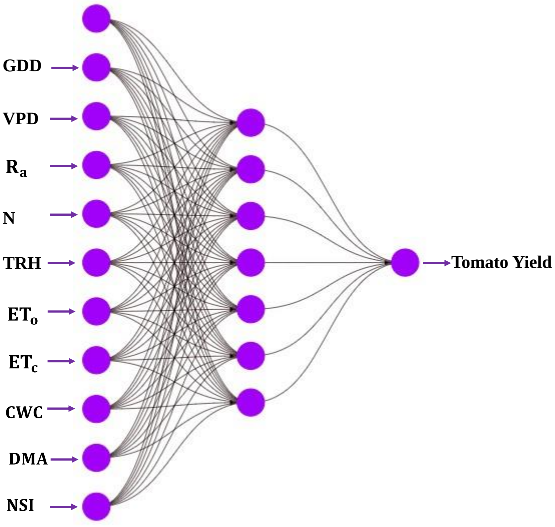

Fig. 6. ANN architecture for tomato yield prediction using second group data.

Table 6. ML models Performance for tomato yield prediction after training and testing using both group data

Models	Training		Testing	
	R^2	RMSE (ton/ha)	R^2	RMSE (ton/ha)
ANN-TFY3	0.95***	4.00	0.95***	3.98
RF-TFY3	0.94***	4.04	0.95***	4.00
DT-TFY3	0.95***	3.91	0.96***	3.59

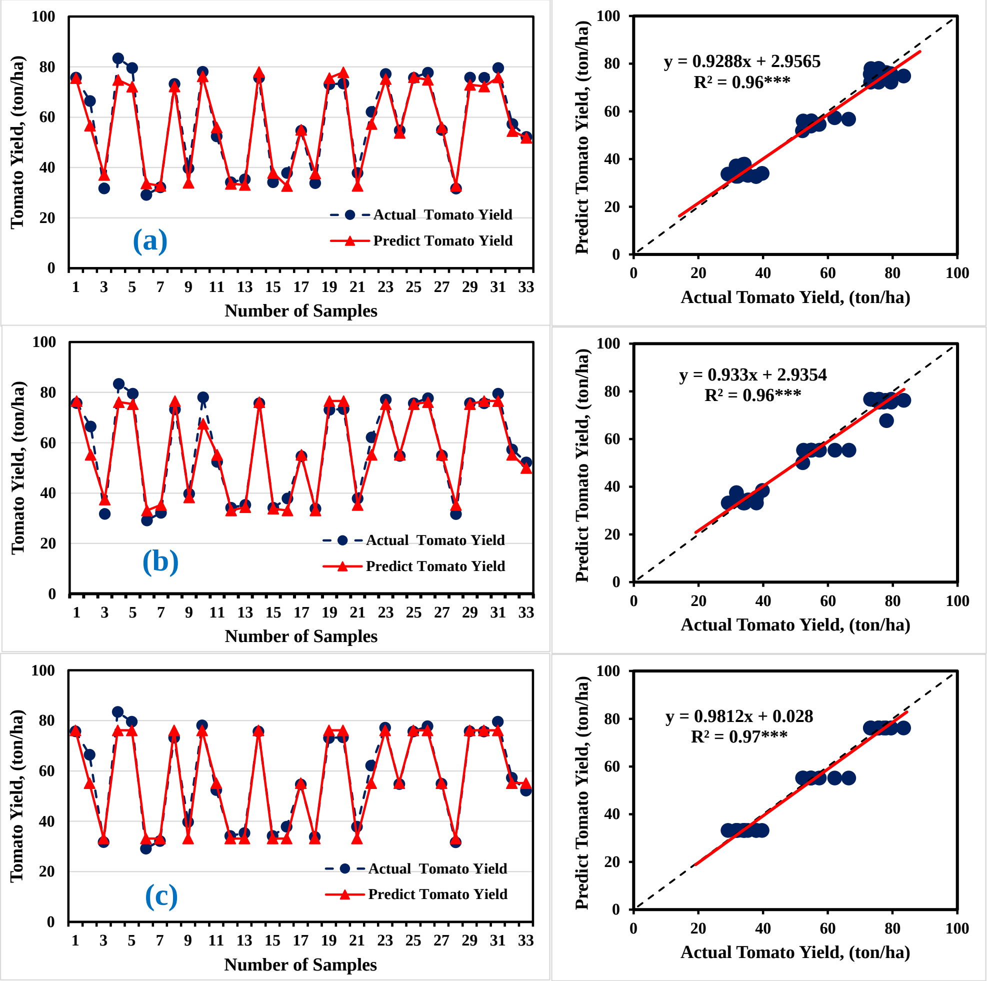

Fig. 7. Comparative Analysis of (a) Artificial Neural Networks (ANN), (b) Random Forest (RF), and (c) Decision Trees (DT) for tomato yield prediction (ton/ha) during testing using merged data from both groups.

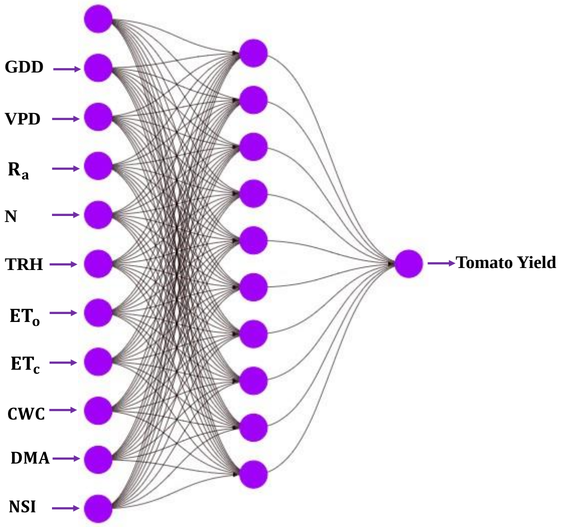

Fig. 8. ANN architecture for tomato yield prediction using merged data from both groups.

4. Discussion

4.1 Plant variables

The results depicted in **Table 2** reveal a clear relationship between different irrigation regimes and CWC. It was observed that reducing the amount of applied water led to a decrease in CWC. These findings are consistent with the studies conducted by **Alordzinu et al. (2021)**. Also, Significant variations in the N-sufficiency Index (NSI) values for tomato plants under different irrigation regimes during the flowering and fruit ripening stages, as presented in **Table 2**. Notably, during the flowering stage, the NSI values peaked at 75% of FIR for both seasons, followed by 100% of FIR and 50% of FIR. The impact of water stress on chlorophyll, the primary pigment crucial for photosynthesis, was apparent. Plants under stress exhibited decreased chlorophyll content due to severely reduced CWC, leading to lower NSI values, as seen in the case of 50% of FIR. Conversely, a slight reduction in CWC could concentrate chlorophyll content in specific leaf areas, resulting in higher NSI values, exemplified by 75% of FIR. These results corroborate **Sarker et al. (2020)** findings that high deficit irrigation reduces NSI values during the flowering stage. Transitioning to the fruit ripening stage, the highest NSI values were observed at 50% of FIR, followed by 75% of FIR and 100% of FIR in both seasons. These results align with **Màtè and SZALÓKINÉ ZIMA (2020)**, emphasizing that NSI values rise with increasing water stress during fruit ripening. Conversely, NSI values decreased in the full irrigation treatment due to its promotion of intensive photosynthesis, ultimately enhancing tomato yield. Moreover, the findings in **Table 2** underscore the substantial influence of different irrigation regimes on DMA during the flowering and fruit ripening stages in both seasons. The observed reductions in DMA can be attributed to the adverse effects of water stress, which hindered photosynthetic processes due to decreased plant water and chlorophyll content, ultimately leading to a decrease in the accumulation of dry matter. These results align with previous studies conducted by **El-Labad et al. 2019)**, which similarly reported that the full irrigation regimes yielded the highest DMA weights for tomato crops when compared to other deficit irrigation approaches.

Lower CWC signifies water stress, leading to decreased photosynthetic activity and poor plant health. Reduced chlorophyll (NSI values) limits photosynthesis efficiency, resulting in stunted growth and fewer flowers and fruits, which are crucial for yield. Additionally, lower DMA indicates diminished overall health and nutrient accumulation, correlating with reduced yield, as plants with insufficient biomass struggle to support fruit development and maturation. Therefore, accurately detecting plant variables is crucial for optimizing crop yield. Previous studies conducted by **López-Aguilar et al. (2020)** have demonstrated that it is possible to predict crop yield by evaluating the total accumulation of dry matter during the early growth phase. These studies highlight the significance of understanding and monitoring plant variables that have a direct correlation with tomato yield. By focusing on these variables, farmers can enhance their ability to predict and improve tomato crop productivity.

4.2 Climate variables

Our results have valuable insights into the relationship between climatic conditions, plant variables, and tomato fruit yield during different seasons. The second season experienced slightly severe conditions, characterized by higher values in GDD, VPD, N, and R_S, which likely contributed to slightly decreased tomato fruit production as well as increase in ET_o values by increasing evaporation rates compared to these values in the first season. Although these values increased during the second season than first season, it still indicates a reasonably favorable environment for tomato fruit production. These proper values improve photosynthesis, enhancing biomass and chlorophyll content (NSI values). Total relative humidity (TRH) helps maintain canopy water content and supports growth, positively influencing DMA and yield. These results are consistent with findings from studies conducted by **Kizza et al. (2016)**. Extensive research conducted by **Siebert et al. (2017)** and **Meng et al. (2017)** has demonstrated the crucial role of these climatic parameters in determining crop yield. Temperature influences plant metabolism, growth rate, and flowering, while sunlight duration and solar radiation affect photosynthesis and energy availability for fruit development. By integrating historical climate data, such as temperature, sunlight duration, and solar radiation, with crop-specific models or algorithms, growers can forecast tomato yield with reasonable accuracy. This predictive capability enables them to anticipate potential challenges, plan resource allocation, and implement targeted management practices to maximize tomato fruit productivity **Li et al. (2019)**.

4.3 Tomato fruit yield

The analysis of the tomato yield data demonstrated a highly significant impact of irrigation regimes ($p < 0.0001$). Water scarcity during the growth period diminishes yield due to reduced fruit weight and number, particularly during flowering stage, where plants are highly sensitive to water stress, leading to flower loss and subsequently fewer fruits, as seen in the case of 50% and 75% of FIR. The outcomes of this study resonate with earlier research conducted by **Djurović et al. (2016)**, and **El-Labad et al. (2019),** which collectively observed that augmenting irrigation practices had a favorable impact on vegetable growth, flowering, and the ultimate yield of tomatoes, as seen in the case of 100% of FIR.

4.4 ML-models performance to predict tomato yield

Tables 4-6 results suggest that the model may be overfit. Optimizing hyper-parameters were searched using the cross-validation approach to make sure our model did not learn excessively from the data. In order to perform cross validation, the dataset is divided into random sets (k-Folds). One set is designated as the test set, and the remaining sets are used to train the model. Every set that is being used as the test set goes through this procedure once more, and the final model is made using the average of the models. The machine learning model, a grid of hyper-parameters, and the selected number of groups (K-Folds or cross-validation value) are entered into the GridsearchCV library, which then outputs the optimal estimator together with its optimal set of hyper-parameters. The data was divided using a 70/30 ratio, meaning that 30 percent was used for testing and the remaining 70 percent was used as training data for cross validation. **Tables 4-6** shows the result of the three ML models. There is a small gap between the training and test phases. Therefore, there is less risk of overfitting for this value. This is why we have chosen K-Folds = 5. The R^2 between the predict tomato yield and actual tomato yield is plotted in three graphs (**Figs. 3**, **5**, and **7**). The data representation indicates a positive slope for the linear regression. Some values were observed to slightly deviate from the large mass. This may be due to the result of bias. Overall, the data representation indicates a linearity between the parameters and thus the possibility of a reduced value of the variance.

The one-way ANOVA conducted revealed no statistically significant differences in the performance of the three models during the training and testing phases. However, it is crucial to consider the practical implications of the observed performance metrics. The random forest models (RF-TFY1 and RF-TFY2) achieved the lowest RMSE using the first and second groups data compared to the decision tree (DT-TFY1 and DT-TFY2) and artificial neural network (ANN-TFY1 and ANN-TFY2) models. Although these differences may not reach statistical significance, it can has substantial real-world consequences. In agricultural contexts, even a modest improvement in yield prediction accuracy can translate into significant economic benefits for farmers, enabling better resource allocation and optimized harvest planning. The superior performance of the random forest (RF-TFY1 and RF-TFY2) models can be attributed to its structure, which consists of 10 trees with a maximum depth of 4. This configuration allows the model to capture complex interactions and nonlinear relationships within the data more effectively than the decision tree (DT-TFY1 and DT-TFY2) models, which, with a maximum depth of just 2, may overlook important patterns. The capacity of the RF model to aggregate predictions from multiple trees enhances its robustness, making it particularly suited for the intricacies of agricultural data. This increased depth in trees can lead to better feature representation, improved predictive accuracy, and reduced risk of overfitting. Conversely, the artificial neural network (ANN-TFY1 and ANN-TFY2) models, constructed with a single hidden layer, demonstrated respectable performance but fell short of the RF model's accuracy. While the ReLU and tanh activation functions can model certain nonlinearities effectively, the relatively small number of neurons might restrict the model's expressiveness. The limited depth and small neuron count in the ANN may hinder its ability to model intricate relationships, particularly in datasets with significant variability, like agricultural yield data. Despite the decision tree (DT-TFY3), with a maximum depth of 2, achieving the lowest RMSE using merged data from both groups compared to the random forest (RF-TFY3) and artificial neural network (ANN-TFY3) models, the superior performance of the DT-TFY3 model can be attributed to its shallow depth, which allows it to focus on the most significant features. This allows it to focus on the most significant features. Such simplicity can lead to better generalization in certain contexts, particularly when the data exhibits clear, dominant patterns and there is a sufficient amount of data, as seen with merged data from both groups. Additionally, this configuration ensures that essential relationships are captured without being influenced by noise. Future research may benefit from exploring more sophisticated architectures or ensemble methods to further enhance predictive accuracy, ensuring that farmers have access to the best tools for informed decision-making.

Machine learning techniques have proven to be highly effective in predicting crop yields, as demonstrated by several studies. **Gholipoor and Nadali (2019)** conducted research on pepper fruit yield prediction using ANN models. By incorporating plant variables such as plant height, fruit number, fruit water content, and canopy width, their study revealed that the ANN model achieved exceptional accuracy with an R^2 value of 0.97 and an RMSE of 0.018 kg/plant. In another study by **Kuradusenge *et al.* (2023)**, data mining techniques were employed to predict future yields of potato and maize crops based on climate data. The researchers employed random forest, polynomial regression, and support vector regressor models for analysis. The outcomes indicated that random forest performed exceptionally well, achieving R^2 values of 0.88 and 0.82 for the potato and maize crop datasets, respectively, making it the superior model for early crop yield prediction. **López-Aguilar *et al.* (2020)** focused on simulating fresh fruit yield in tomato crops using an ANN model. The ANN model incorporated various plant variables such as leaf area, plant height, fruit number, and dry matter of leaves, stems, and fruits, along with growth degree days. The results exhibited a strong correlation between the predicted and actual tomato crop yields, with the ANN model achieving an impressive R^2 value of 0.88. **Cedric *et al.* (2022)** combined climatic data and agricultural yields to create a powerful tool for predicting cassava yields using a DT model. The results of the research showed that the DT model exhibited exceptional performance, with a R^2 reaching 94.1%. The findings from these studies, along with our own research, emphasize the importance of harnessing advanced modeling techniques such as ANN, RF, and DT models to improve crop yield predictions based on climate or plant variables. The successes observed in these studies, as well as our own, highlight the immense potential of ML models in accurately forecasting crop yields. By leveraging these techniques, we gain valuable insights that can be utilized to optimize agricultural practices and foster sustainable crop production.

5. Conclusion

In conclusion, this research developed tomato yield estimation models using artificial neural networks (ANN), random forest (RF), and decision tree (DT) techniques, based on climate and plant variables. The models successfully captured the relationships between input variables and tomato production under deficit irrigation conditions. The one-way ANOVA conducted revealed no statistically significant differences in the performance of the three models during the training and testing phases. However, even modest improvements in yield prediction accuracy can have substantial real-world consequences in agricultural contexts, translating into significant economic benefits for farmers through better resource allocation and optimized harvest planning. The RF model exhibited the highest accuracy, followed closely by the ANN and DT models when using data from the first and second groups. The increased depth of the RF model facilitates better feature representation and improves predictive accuracy while reducing the risk of overfitting. Notably, the DT model demonstrated the highest accuracy when applied to merged data from both groups. This performance can be attributed to its shallow depth, which enables it to focus on the most significant features, leading to better generalization in contexts where the data exhibits clear, dominant patterns and sufficient volume. These findings underscore the practicality and reliability of utilizing climate and plant variables in conjunction with machine learning models to effectively manage tomato crop production, particularly in scenarios of limited water availability for irrigation. The implications of this research extend beyond tomatoes; similar modeling approaches could be applied to other crops, enhancing yield predictions and resource management across various agricultural systems. However, potential challenges and limitations may arise when scaling these models for widespread use. Variability in local climatic conditions, soil types, and crop management practices can influence model performance, necessitating adaptations for different contexts. Additionally, data availability and quality can vary significantly, impacting the accuracy and reliability of the models. Future research should explore more sophisticated architectures or ensemble methods to further enhance predictive accuracy and address these challenges, ensuring that farmers have access to the best tools for informed decision-making across diverse agricultural landscapes.

Acknowledgments: We extend our gratitude to Mansoura University and University of Sadat City for providing the devices utilized in the measurements conducted for this study.
Funding statements: The research was funded by the personal efforts of the authors
Conflicts of interest: Authors have declared that no competing interests exist.

References

Abbas, F., Afzaal, H., Farooque, A. A., and Tang, S. (2020). Crop yield prediction through proximal sensing and machine learning algorithms. *Agronomy*. 10(7), 1046.

Abdalhi, M. A., Jia, Z., Luo, W., Ali, O. O., and Chen, C. (2020). Simulation of canopy cover, soil water content and yield using FAO-AquaCrop model under deficit irrigation strategies. *Russian Agricultural Sciences*. 46, 279-288.

Abdulhadi, J. S., and Alwan, H. H. (2021). Evaluation of the scheduling of an existing drip irrigation network: Fadak Farm, Karbala, Iraq. *In IOP Conference Series: Materials Science and Engineering*. (Vol. 1067, No. 1, p. 012024). IOP Publishing.

Aboukota, M., Hassaballa, H., Elhini, M., & Ganzour, S. (2024). Land Degradation, Desertification & Environmental Sensitivity to Climate Change in Alexandria and Beheira, Egypt.. *Egyptian Journal of Soil Science*, *64*(1), 167-180.

Ahmad, M. W., Reynolds, J., and Rezgui, Y. (2018). Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees. *Journal of cleaner production*. 203, 810-821.

Allen, R. G., Pereira, L. S., Raes, D., and Smith, M. (1998). Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56. *Fao, Rome*. 300(9), D05109.

Alordzinu, K. E., Li, J., Lan, Y., Appiah, S. A., Al Aasmi, A., Wang, H., Qiao, S., et al. (2021). Ground-based hyperspectral remote sensing for estimating water stress in tomato growth in sandy loam and silty loam soils. *Sensors*. 21(17), 5705.

Bai, H., and Purcell, L. C. (2018). Aerial canopy temperature differences between fast-and slow-wilting soya bean genotypes. *Journal of Agronomy and Crop Science*. 204(3), 243-251.

Bausch, W. C., Diker, K., Khosla, R., and Paris, J. F. (2004, November). Estimating corn nitrogen status using ground-based and satellite multispectral data. *In Remote Sensing and Modeling of Ecosystems for Sustainability* (Vol. 5544, pp. 489-498). SPIE.

Breiman, L. (2001). Random forests. *Machine learning.* 45, 5-32.

Cedric, L. S., Adoni, W. Y. H., Aworka, R., Zoueu, J. T., Mutombo, F. K., et al. (2022). Crops yield prediction based on machine learning models: Case of West African countries. *Smart Agricultural Technology*. 2, 100049.

Duffie, J. A., and Beckman, W. A. (1980). Solar engineering of thermal processes (p. 16591). New York: Wiley.

ElGhamry, A., Mosa, A., Elramady, H., GHAZI, D., elsherpiny, M., & helmy, A. (2024). Climate Change and the Possibility of Tea Production in the Egyptian Soils. *Egyptian Journal of Soil Science*, *64*(2), 373-383.

El-Labad, S. A., Mahmoud, M. I., AboEl-Kasem, S. A., and ElKasas, A. I. (2019). Effect of irrigation levels on growth and yield of tomato under El-Arish region conditions. *Sinai Journal of Applied Sciences*. 8(1), 9-18.

Ella, V. B., Keller, J., Reyes, M. R., and Yoder, R. (2013). A low-cost pressure regulator for improving the water distribution uniformity of a microtube-type drip irrigation system. *Applied Engineering in Agriculture*. 29(3), 343-349.

Elsherpiny, M. (2023). Role of compost, biochar and sugar alcohols in raising the maize tolerance to water deficit conditions. *Egyptian Journal of Soil Science*, *63*(1), 67-81.

FAOSTAT, C. 2022. Livestock Products Available online: https://www.fao.org/faostat/en/#data.QCL/visualize (accessed on 29 July 2022).

Gabr, M. E. S. (2022). Management of irrigation requirements using FAO-CROPWAT 8.0 model: A case study of Egypt. *Modeling Earth Systems and Environment*. 8(3), 3127-3142.

Gholipoor, M., and Nadali, F. (2019). Fruit yield prediction of pepper using artificial neural network. *Scientia Horticulturae*. 250, 249-253.

Han, J., Pei, J., and Tong, H. (2022). Data mining: concepts and techniques. *Morgan kaufmann*.

kamara, M., Elgamal, W., Abd El-Aty, M., Mesbah, M., Behiry, S., & abomarzoka, E. (2023). Influence of Foliar Supplied of Some Biostimulants on Physiological, Agronomic Characters and Water Productivity of Rice Under Water Deficit and Normal Conditions. *Egyptian Journal of Soil Science*, *63*(4), 455-464.

Khaki, S., and Wang, L. (2019). Crop yield prediction using deep neural networks. *Frontiers in plant science*. 10, 621.

Kizza, T., Fungo, B., Kabanyoro, R., and Nagayi, R. (2016). Effect of drip irrigation regimes on the growth and yield of tomatoes in Central Uganda. *J. Sci. Res. Adv*. 3(2016), 306-312.

Kuradusenge, M., Hitimana, E., Hanyurwimfura, D., Rukundo, P., Mtonga, K., et al. (2023). Crop yield prediction using machine learning models: Case of Irish potato and maize. *Agriculture*. 13(1), 225.

Li, Y., Guan, K., Yu, A., Peng, B., Zhao, L., et al. (2019). Toward building a transparent statistical model for improving crop yield prediction: Modeling rainfed corn in the US. *Field Crops Research*. 234, 55-65.

Lobell, D. B., and Burke, M. B. (2010). On the use of statistical models to predict crop yield responses to climate change. *Agricultural and forest meteorology*. 150(11), 1443-1452.

Màtè, M. D., and SZALÓKINÉ ZIMA, I. (2020). Development and yield of field tomato under different water supply. *Research Journal of Agricultural Science*. 52(1).

Meng, T., Carew, R., Florkowski, W. J., and Klepacka, A. M. (2017). Analyzing temperature and precipitation influences on yield distributions of canola and spring wheat in Saskatchewan. *Journal of Applied Meteorology and Climatology*. 56(4), 897-913.

Mijwel, M. M. (2021). Artificial neural networks advantages and disadvantages. *Mesopotamian Journal of Big Data*. 2021, 29-31.

Noreldin, T., Ouda, S., Abdou, S. M. M., and KMR, Y. (2014). Using bism model to calculate water requirements for some vegetable crops in egypt. *Fayoum Journal of Agricultural Research and Development*. 28(2), 111-120.

Oymak, S. (2019, June). Stochastic gradient descent learns state equations with nonlinear activations. *In conference on Learning Theory* (pp. 2551-2579). PMLR.

Power, N. A. S. A. (2022). Data Access Viewer Available online: https://power. larc. nasa. gov/data-access-viewer. Last accessed. 11(10).

Roberts, M. J., Schlenker, W., and Eyer, J. (2013). Agronomic weather measures in econometric models of crop yield with implications for climate change. *American Journal of Agricultural Economics*. 95(2), 236-243.

Sarker, M. R., Choudhury, S., Islam, N., Zeb, T., Zeb, B. S., et al. (2020). The effects of climatic change mediated water stress on growth and yield of tomato. *Cent. Asian J. Environ. Sci. Technol. Innov*. 1(2), 85-92.

Semananda, N. P., Ward, J. D., and Myers, B. R. (2016). Evaluating the efficiency of wicking bed irrigation systems for small-scale urban agriculture. *Horticulturae*. 2(4), 13.

Shahhosseini, M., Hu, G., Huber, I., and Archontoulis, S. V. (2021). Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. *Scientific reports*. 11(1), 1606.

Shalaby, T. A., and El-Banna, A. (2013). Molecular and horticultural characteristics of in vitro induced tomato mutants. *Journal of Agricultural Science*. 5(10), 155.

Sharma, S., Sharma, S., and Athaiya, A. (2017). Activation functions in neural networks. *Towards Data Sci*. 6(12), 310-316.

Sridhara, S., Manoj, K. N., Gopakkali, P., Kashyap, G. R., Das, B., et al. (2023). Evaluation of machine learning approaches for prediction of pigeon pea yield based on weather parameters in India. *International Journal of Biometeorology*. 67(1), 165-180.

Xia, Y., Liu, C., Li, Y., and Liu, N. (2017). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert systems with applications*. 78, 225-241.

Yıldırım, M., and Bahar, E. (2017). Water and radiation use-efficiencies of tomato (Lycopersicum esculentum L.) at three different planting densities in open field. *Mediterranean Agricultural Sciences*. 30(1), 39-45.