agriculture

MDPI

Article

Application of Machine Learning and Neural Networks to Predict the Yield of Cereals, Legumes, Oilseeds and Forage Crops in Kazakhstan

Marzhan Sadenova [1,*], Nail Beisekenov [1], Petar Sabev Varbanov [1,2] and Ting Pan [2]

[1] Priority Department Centre «Veritas» D. Serikbayev, East Kazakhstan Technical University, 19 Serikbayev str., Ust-Kamenogorsk 070000, Kazakhstan; varbanov@fme.vutbr.cz (P.S.V.)
[2] Sustainable Process Integration Laboratory, NETME Centre, Faculty of Mechanical Engineering, Brno University of Technology, Technická 2896/2, 616 69 Brno, Czech Republic
* Correspondence: msadenova@edu.ektu.kz

Abstract: The article provides an overview of the accuracy of various yield forecasting algorithms and offers a detailed explanation of the models and machine learning algorithms that are required for crop yield forecasting. A unified crop yield forecasting methodology is developed, which can be adjusted by adding new indicators and extensions. The proposed methodology is based on remote sensing data taken from free sources. Experiments were carried out on crops of cereals, legumes, oilseeds and forage crops in eastern Kazakhstan. Data on agricultural lands of the experimental farms were obtained using processed images from Sentinel-2 and Landsat-8 satellites (EO Browser) for the period of 2017–2022. In total, a dataset of 1600 indicators was collected with NDVI and MSAVI indices recorded at a frequency of once a week. Based on the results of this work, it is found that yields can be predicted from NDVI vegetation index data and meteorological data on average temperature, surface soil moisture and wind speed. A machine learning programming language can calculate the relationship between these indicators and build a neural network that predicts yield. The neural network produces predictions based on the constructed data weights, which are corrected using activation function algorithms. As a result of the research, the functions with the highest prediction accuracy during vegetative development for all crops presented in this paper are multi-layer perceptron, with a prediction accuracy of 66% to 99% (85% on average), and polynomial regression, with a prediction accuracy of 63% to 98% (82% on average). Thus, it is shown that the use of machine learning and neural networks for crop yield prediction has advantages over other mathematical modelling techniques. The use of machine learning (neural network) technologies makes it possible to predict crop yields on the basis of relevant data. The individual approach of machine learning to each crop allows for the determination of the optimal learning algorithms to obtain accurate predictions.

Keywords: yield forecasting; remote sensing; machine learning; cereals; oilseeds; grain legumes; forage crops; sustainable farming practices

Citation: Sadenova, M.; Beisekenov, N.; Varbanov, P.S.; Pan, T. Application of Machine Learning and Neural Networks to Predict the Yield of Cereals, Legumes, Oilseeds and Forage Crops in Kazakhstan. *Agriculture* **2023**, *13*, 1195. https://doi.org/10.3390/agriculture13061195

Academic Editor: Yongchao Tian

Received: 29 April 2023
Revised: 28 May 2023
Accepted: 1 June 2023
Published: 3 June 2023

1. Introduction

Advances in modern technology have facilitated the monitoring of numerous factors that can influence crop yields. The systematic collection and analysis of data relating to these factors makes it possible to establish their specific impact on yield and even predict it. Accurate crop yield forecasting is a vital and complex activity, essential for achieving sustainable intensification and the sustainable use of natural resources.

The importance of crop yield forecasts extends to the various stakeholders involved in the agrifood chain, including farmers, agronomists, commodity traders and policy-makers. These forecasts provide valuable insights that are used in decision-making processes and help to optimise agricultural practices. However, yield forecasting is a complex task because

of the complex interrelationships between crop-specific parameters, environmental conditions and choice of management practices. Developing a reliable and transparent forecast model is a complex task that requires careful study and a comprehensive understanding of these multifaceted influences. Modern technology allows us to monitor many factors that can influence yield. By collecting data on these factors, it is possible to determine their impact on yields and even predict them. Crop yield prediction is an important but complex process and is necessary for sustainable intensification and efficient use of natural resources [1]. Crop yield forecasts are valuable to many stakeholders in the agri-food chain, including farmers, agronomists, commodity traders and policymakers [2]. Crop yields are influenced by many crop-specific parameters, environmental conditions and management decisions [3], and it is difficult to build a reliable and explainable forecast model.

Typically, field surveys, crop growth models, remote sensing data, statistical models and their combinations are used for crop yield prediction. These methods themselves address slightly different aspects of crop yield forecasting. Crop growth models simulate the growth and development of crops according to the agronomic principles of plant–environment interactions and management [4]. Remote sensing techniques rely on satellite imagery to capture the current crop condition and then estimate the final yield [5]. Statistical models use weather variables and results from the previous three methods as predictors to obtain linear relationships between predictors and crop yields, for example [6]. Recent studies have combined different methods in innovative ways to construct yield prediction models. For example, ref. [7,8] used high-resolution remote sensing data and crop modelling to build statistical models to predict actual yields. Similarly, ref. [9] developed a probabilistic yield forecasting system for Canada using remote sensing, crop modelling, Bayesian inference and statistical models.

Crop monitoring and yield forecasting methods are described in detail in [10]. This paper provides information on satellite data, the approaches used to process satellite images, and the methods used for agricultural land use mapping and yield forecasting. A brief introduction to the combination of parameters that can be used for forecasting analysis is obtained. Machine learning algorithms are used to train the data. We also provide a detailed introduction to the different machine learning methods and accuracy assessments of machine learning algorithms.

The purpose of this paper is to develop a machine learning and neural network-based crop yield forecasting methodology for crop yield prediction. Machine learning uses an empirical modelling approach to learn useful patterns and relationships from input data [11] and provides a promising avenue for improving crop yield forecasts. Machine learning algorithms approximate a function that relates functions or predictors to labels such as crop yields. Similar to statistical models, machine learning algorithms can use the results of other methods as functions. In addition, machine learning algorithms have some distinct advantages: they can model non-linear relationships between multiple data sources [12], their performance tends to improve when more model training data are available [13], and they can become robust to noisy data by using regularization techniques that help reduce variance and generalization error [14]. Thus, machine learning can combine the advantages of other methods such as crop growth models and remote sensing (RS) with data-driven modelling to produce reliable crop yield predictions.

Many studies have applied machine learning to predict crop yields in certain locations, but it is unclear whether their data and methods can be used for other crops and soil-climatic zones. Some have used empirical data collected for specific purposes that may not be available for other crops or soil and climate zones [15]. Some studies have used publicly available climate and satellite data but have opted for crops and locations that limit their reusability [16]. To determine the best machine learning method, the authors' work [17] was used to give a brief introduction to machine learning algorithms that can be used for yield prediction. Crop yield prediction allows early detection of problems that reduce overall crop production, and this paper provides a detailed explanation of the machine learning models and algorithms that are required for crop yield prediction.

The paper also compares the accuracy of different yield prediction algorithms. In [18], the authors showed the effectiveness of using Python programming language [19] and the Jupyter Notebook platform [20] to implement machine learning methods in yield prediction. Python has many supported libraries for data analysis and machine learning [21], making it suitable for achieving the goals of this paper. The developed methodology is written in Python programming language based on the Jupyter Notebook platform. Pandas [22], Matplotlib [23], Numpy [24], Seaborn [23] and Scikit-learn [25] libraries were used to develop the yield forecasting methodology. The methodology is based on remote sensing imagery taken from free sources EO Browser and ladsweb.modaps.eosdis.nasa.gov based on experimental sites with cereals, legumes, oilseeds and forage crops in East Kazakhstan. The farmland data were obtained using processed images from Sentinel-2 and Landsat-8 satellites (EO Browser) for the period 2017–2022. Meteorological data were acquired using the OpenWeatherMap API [26], which made it possible to obtain archived data from 2017 to 2022 with batch processing of factors for the area of interest and to use the acquired data for commercial purposes according to the license. The collected data were exported and sorted by crop into xlsx files. These included the following factors: yield, MSAVI [27], NDVI [28], maximum temperature, minimum temperature, surface soil moisture, root zone moisture, wind speed and air humidity. In total, a dataset of 1600 indices was collected, in which the NDVI and MSAVI indices were recorded at a frequency of once a week. The data were trained for two crop growing seasons: the first growing season (May–June) and the second growing season (August–September).

Several studies have demonstrated the feasibility of using RS data for perennial crop yield forecasting. These studies have employed different RS sensors, such as optical, microwave, and hyperspectral sensors, to capture crop growth and physiological parameters, such as vegetation indices, canopy height, and water content. For instance, in a study by Li et al. [29], 10 vegetation indices (VI) derived from PlanetScope and Sentinel-2 time series images were used to examine the feasibility of estimating maize grain yield using different regression methods. The study was conducted in Minnesota (USA). There was good agreement between observed and predicted yields, with a coefficient of determination of 0.81 at day 86 after sowing. The feasibility of using monthly Sentinel-2 satellite image composites was investigated in [30] to predict rice yield one month before the harvest period at the field level using ML techniques in Taiwan. Model validation results from SVM models using data from transplanting to ripening showed RMSPE and MAPE values of 5.5% and 4.5% for the second crop in 2019 and 4.7% and 3.5% for the first crop in 2020.

However, the aforementioned studies faced some limitations. For instance, some studies used limited or outdated RS data, which may affect the accuracy and reliability of yield forecasting. Others did not consider the influence of environmental factors, such as weather, soil, and pests, on crop growth and yield, which can lead to biased predictions. Moreover, some studies used conventional regression or machine learning models, which may not capture the complex and dynamic relationships between RS parameters and yield and may lead to overfitting or underfitting.

The integration of modern technology and advanced data analysis has the potential to bring a breakthrough to agricultural management and support sustainable practices. However, previous studies have faced limitations that have hindered their effectiveness. Some of the studies [31] used limited or outdated remote sensing data, which reduced the accuracy and reliability of yield predictions. In addition, the influence of important environmental factors such as weather, soil conditions and pests were often not taken into account [32], resulting in biased predictions. Moreover, relying on conventional regression or machine learning models [33], it was not possible to capture the complex relationships between remote sensing parameters and yields, leading to over- or under-interpretation.

To overcome these limitations, the research should focus on incorporating relevant and complete remote sensing data, accounting for environmental factors and using advanced modelling techniques. In this way, agricultural management will be able to take advantage of more accurate and reliable yield forecasts, enabling informed decisions on resource

allocation, land management and crop selection. Ultimately, this will optimise agricultural practices, reduce resource wastage and ensure the sustainability of the agri-food sector.

Within that knowledge gap, the current work proposes a new method for predicting the yield of perennial crops based on RS data. The novel scientific contribution of this work consists of the joint use of several data sources—UAV, remote sensing, agrochemical studies, and weather data, within a neural network for yield forecasting. Following this approach increases the accuracy of predictions, leading to informed decision-making, optimised resource allocation and reduced environmental impact.

2. Method Summary

The research concept lies in using remote sensing and meteorological data processing to develop crop yield forecasts. This approach has the potential to revolutionise agricultural management and support sustainable practices. Machine learning algorithms can be used to analyse large data sets and identify patterns in crop growth and development that are difficult to detect using traditional methods. The assessment and elimination of emission data entail the consideration of multiple criteria and factors, thereby ensuring the reliability and accuracy of the dataset. These criteria encompass various key elements that facilitate the identification and removal of undesirable or erroneous emissions-related information. Among the pivotal factors are as follows:

(1) Use of remote sensing and meteorological data: The most important criterion involves a careful assessment of the accuracy and reliability of remote sensing and meteorological data used in the study. Emission data showing inconsistencies or unreliability with regard to remotely sensed and meteorological data can be deleted.
(2) Application of machine learning algorithms: The second criterion concerns the effectiveness of machine learning algorithms in detecting patterns related to crop growth and development. Emission data that deviate significantly from the identified patterns or show anomalies can be considered potential candidates for deletion.
(3) Comparison with traditional methods: The third criterion involves comparing emission data obtained using remote sensing and meteorological data processing with emission data obtained using traditional methods. If emission data derived from traditional methods are found to be more accurate or reliable, the corresponding emission data derived from remote sensing and meteorological data can be excluded.

If these criteria are reasonably met, the study seeks to provide a comprehensive basis for excluding emissions from the data set. This approach ensures that the resulting crop yield forecast model is based on reliable and valid information.

2.1. Modelling Approach and Procedure

The modelling procedure involves collecting historical yield, remote sensing and meteorological data on the target crops. The data are pre-processed and cleaned to remove outliers and missing values. In the context of yield forecasting, it is essential to ensure the integrity and accuracy of the data set by effectively handling outliers, including outliers that can affect the reliability of forecasting models. The study used a modified statistical Z-Score method to identify and remove outliers that deviated significantly from expected models or distributions.

The modified Z-Score method is recognised for its reliability in detecting outliers, allowing the distinction between extreme outliers, which represent true outliers, and data points, which correspond to underlying relationships in the data set. By applying this statistical method to the preprocessing and data-cleaning phase, we successfully identified and excluded 176 outliers from the dataset.

Determining which outliers should be excluded was based on a thorough statistical analysis. Using this analysis, a suitable threshold was established to distinguish true outliers indicating anomalous behaviour or data quality problems from valid data points that were consistent with the expected patterns of the agricultural system under study.

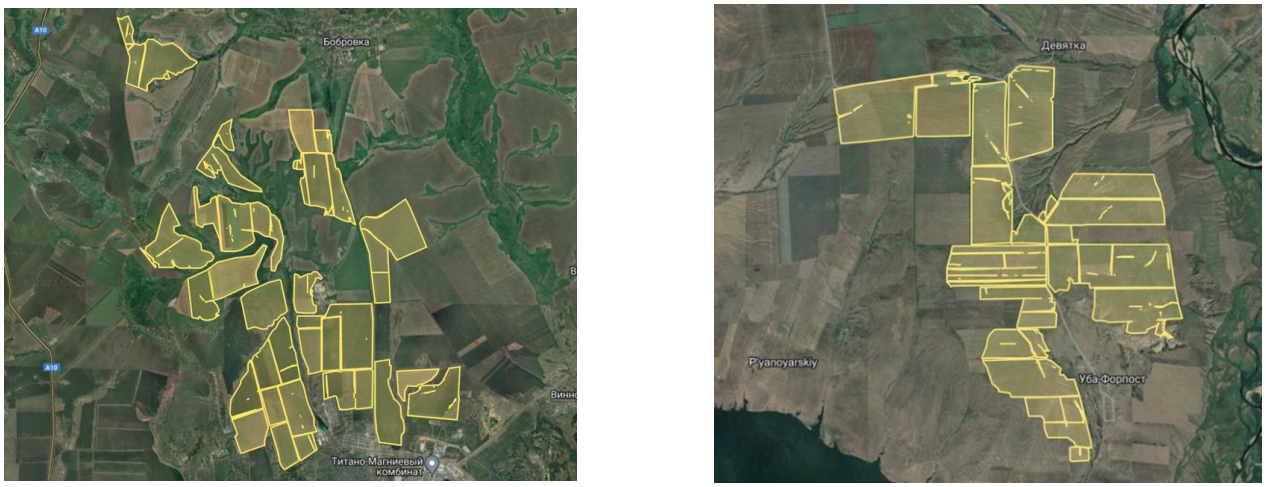
By systematically removing these identified outliers, it was possible to improve the integrity of the data set and minimise the possibility of distortion in the yield prediction models. This approach aimed to ensure that the resulting models were robust and able to reflect the fundamental relationships between the various factors affecting crop yields while mitigating the impact of anomalous outliers.

It is important to note that the exact number of remote emissions, namely 176, highlights the scale of the data points identified as outliers using the modified Z-Score method. Including this information provides transparency and facilitates the reproducibility of the study by allowing other researchers to understand the extent to which emissions were treated as outliers in the data-cleaning process.

Feature engineering techniques are applied to extract relevant information from remote sensing. Meteorological data and machine learning algorithms such as Random Forest and Neural Networks are used to train the yield prediction model. The model is validated using data from recent growing seasons, and the accuracy of the model is assessed using metrics such as mean absolute error and root mean square error. The impact of different data sources, feature construction methods and machine learning algorithms on the accuracy of yield predictions are also evaluated. Finally, the results of the modelling procedure are used to draw conclusions about the potential of machine learning to improve crop yield prediction and support sustainable agricultural practices.

2.2. Remote Sensing Imagery Data and Preparation of the Development Environment

Satellite imagery data were processed using Agisoft Metashape Professional software. The data were obtained from farms in East Kazakhstan, specifically from experimental farmlands in two distinct soil-climatic zones: chernozem (Figure 1A) and chestnut-type soils (Figure 1B). The total number of RS images amounts to 64 tiles.



(**A**) (**B**)

Figure 1. Contours of the experimental polygons for two soil-climatic zones: (**A**) Black soil; (**B**) Chestnut soil.

Earth remote sensing (ERS) data were obtained from the Sentinel Hub EO Browser web platform from the Sentinel-L2A satellite. Spectral indices: normalised vegetation index (NDVI) and modified corrected soil index (MSAVI) were processed using Agisoft Metashape Professional software.

The current methodology was implemented within the Jupyter Notebook development environment [20]. This provides an opportunity to interact interactively with compiled code fragments visualising the work of the programme. The difference from the traditional IDE is that the code can be divided into parts and executed in any order. This development environment allows one to test the function that has been written without compiling the

whole program. A separate memory load function is available so that the resulting part of the content can be checked. This approach saves time and helps to avoid errors.

The effectiveness of the Python programming language in handling geographic data was demonstrated in [34]. Pandas and Numpy libraries were used to handle culture data exported in xlsx format, which handles large volumes of information efficiently and works with the Excel document format. Consistent use of QGIS and Python (Matplotlib, Pandas, Seaborn), incorporating various computational methods, was proposed in [35]. That work emphasises the functionality of a high-level language such as Python and the applicability of machine learning algorithms to earth science. In this study, the Seaborn statistical package built on the Matplpotlib Python library was tested to analyse, model and visualise geospatial data using existing approaches [36]. Efficient and accurate graphing supported by Python demonstrates the undeniable benefits of machine learning in earth sciences.

The field of Data Science is based primarily on the structuring of data; thus, there is a need for data preparation libraries. Currently, the best and most used Python library in this field is Pandas. Pandas has a wide range of capabilities for input/output data formats, such as Excel, CSV, Python/NumPy, HTML, SQL and others. In addition, Pandas has powerful query capabilities, statistical calculations and basic visualisation. Pandas is richly documented, but its syntax is a bit confusing, which is often pointed out as its most significant drawback. Pandas aims to be a fundamental high-level building block for conducting practical, real-world data analysis in Python. In addition, this library has the broader goal of becoming the most powerful and flexible open-source data analysis and manipulation tool available in any language.

3. Sorting the Data and Calculating the Optimum Indicators

3.1. Importing Libraries and Sorting Data

Pandas, Matplotlib and Numpy libraries are imported first, followed by the creation of crop databases from previously downloaded and saved xlsx files. The Excel files are read using the read_excel function, which creates the Pandas database from the existing data in the file.

According to [37], temperature is an important factor for yield modelling. Previous studies have shown the importance of averaged temperature and precipitation of the growing season in explaining crop yield variability [38]. Maximum and minimum temperature parameters need to be combined into an average to reduce the amount of data. Based on the literature review, it was decided to use two data stacks consisting of maximum and minimum temperatures. The average temperature was calculated by determining the arithmetical root of the temperature parameters and creating a new data set from the results (Figure S1).

It was necessary to put the data in order before examining the influence of the factors on each other.

The read_excel function in Figure S2 is the path to the Excel files from which the Pandas database is created. The maximum and minimum temperature columns are extracted from the created database, and their sum is divided by two. The resulting value is stored as the average temperature.

Figure S3 shows a snippet of code to remove block lines in xlsx files with the presence of a hyphen. This method is achieved by checking the columns through an algorithm for data "not equal to" the hyphen. Then, after removing all the empty values, the data are converted to float type via the astype() function for correct data handling in future. To check and clean the data, the info() function is used to display empty null values and the data type. The output of the info() function is shown in Table 1.

Table 1. Output of info().

Column	Non-Null Count	Dtype
Week	93 non-null	int64
NDVI	93 non-null	float64
Average deg. C	93 non-null	float64
Surface soil moisture (%)	93 non-null	float64
Moisture in the root zone (%)	93 non-null	float64
Wind speed (m/s)	93 non-null	float64
Moisture (%)	93 non-null	float64
Date	93 non-null	datetime64[ns]

All data are populated with undefined characters and correspond to an acceptable data type. The output is a list of meteorological data and NDVI at intervals of 1–2 weeks.

3.2. Ratio of Indicators to NDVI

Since we found from previous studies that NDVI is a key indicator for yield, let us consider the behaviour of other factors in relation to it. To do this, plots of the relationship of all indicators to NDVI are output via the plot tool of the Matlpotlib library (Figure 2).

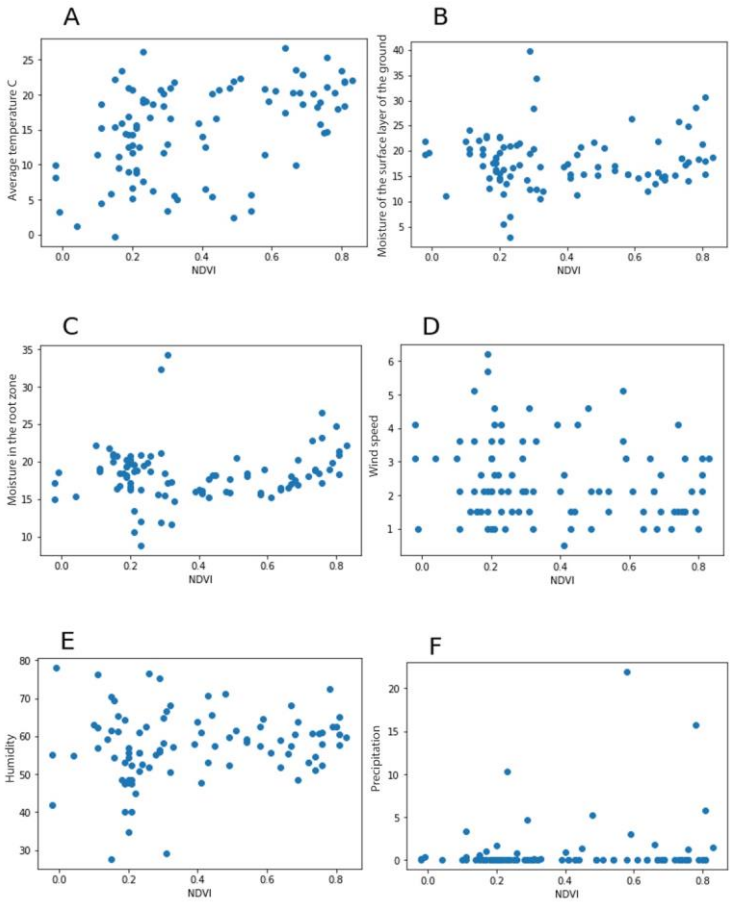


Figure 2. NDVI ratio graphs for (**A**) average temperature, (**B**) surface soil moisture, (**C**) root zone moisture, (**D**) wind speed, (**E**) humidity and (**F**) precipitation.

The correlations and patterns of the NDVI index and weather data are shown in Figure 2. The analysis shows various noteworthy observations. First, the correlation between NDVI index and temperature shows a weak relationship, indicating that temperature alone cannot be a strong predictor of changes in NDVI. Secondly, the behaviour of humidity indices shows similarities with the NDVI index, indicating a potential link between

humidity and vegetation growth. Finally, the impact of precipitation on the NDVI index seems to be insignificant, as the observed effect on the index is limited.

3.3. Data Correlation

The correlation coefficient [39] (denoted as r) determines both the strength and direction of the relationship between the dependent and independent variables. The values of r range from −1.0 (strong negative relationship) to +1.0 (strong positive relationship). A significant correlation between two random variables is always evidence of some statistical relationship in a given sample, but this relationship need not be observed for another sample and be causal. Often, the luring simplicity of correlation studies leads the researcher to draw false intuitive conclusions about the existence of a causal relationship between pairs of variables, whereas correlation coefficients only establish statistical relationships.

The basic intuition underlying information theory is the idea of the characteristic of "unpredictability" of a random variable, also known as information entropy. For a random variable X that takes values in the set $\chi = \{x_1, x_2, \ldots, x_n\}$ with a probability mass function $p(x)$, the entropy $H(X)$ is formulated as (1).

$$H(X) = -\sum_{x \in \chi} p(x) log(p(x)) \quad (1)$$

A negative sign ensures that entropy is always positive or zero. $H(X)$ can be thought of as approximately equal to the information from a single instance of a random variable X. Informativeness will be high when the probability is low, and vice versa. Mutual information (MI) measures how closely a random variable is related to the concept of entropy. The VI for two random variables X and Y, denoted by $I(X;Y)$, is stated as (2).

$$I(X;Y) = H(X) - H(X|Y) \quad (2)$$

where $H(X \mid Y)$ is the conditional entropy for X given by Y. $I(X;Y)$ measures the average reduction in uncertainty in a relation, which leads to a value study [40]. It is a more general form of correlation coefficient, providing an overall measure of the dependence (linear and non-linear) between two variables [41]. The greater the value of the VI, the greater the relationship between the two variables. This is an important statistic when analysing time series from non-linear systems [42]. The VI between two random variables X and Y with joint probability mass function $p(x,y)$ and marginal probability density functions $p(x)$ and $p(y)$ is defined as

$$I(X;Y) = \sum_{x \in \chi} \sum_{y \varepsilon Y} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \quad (3)$$

The correlation method used is the linear correlation coefficient (or Pearson correlation coefficient [43]). The correlation coefficient is calculated using Formula (4).

$$r_{XY} = \frac{cov_{XY}}{\sigma_X \sigma_Y} = \frac{\sum (X - \underline{X})(Y - \underline{Y})}{\sqrt{\sum (X - \underline{X})^2 \sum (Y - \underline{Y})^2}} \quad (4)$$

Here, $\underline{X} = \frac{1}{n} \sum_{t=1}^{n} X_t$, $\underline{Y} = \frac{1}{n} \sum_{t=1}^{n} Y_t$—are the mean values of the samples. The correlation coefficient varies from minus one to plus one. The proof of this is that, dividing both parts of the dual inequality $-\sigma_X \sigma_Y \leq cov_{XY} \leq \sigma_X \sigma_Y$ we get $-1 \leq r_{XY} \leq 1$.

The linear correlation coefficient is related to the regression coefficient, as follows:

$$r_{XY} = a_i \frac{cov_{XY}}{\sigma_X \sigma_Y} \quad (5)$$

where a_i is the regression coefficient, x is the standard deviation of the relevant factor. $x_i^* = a + bx_i$, $y_i^* = c + dy_i$ the linear correlation coefficient will equal Equation (3).

$$r_{X^*Y^*} = \frac{bd}{|bd|} \tag{6}$$

The correlation is automatically calculated by the built-in corr() function of the Pandas library. The correlation graph is called with the Seaborn library via the heatmap function (Figure 3).

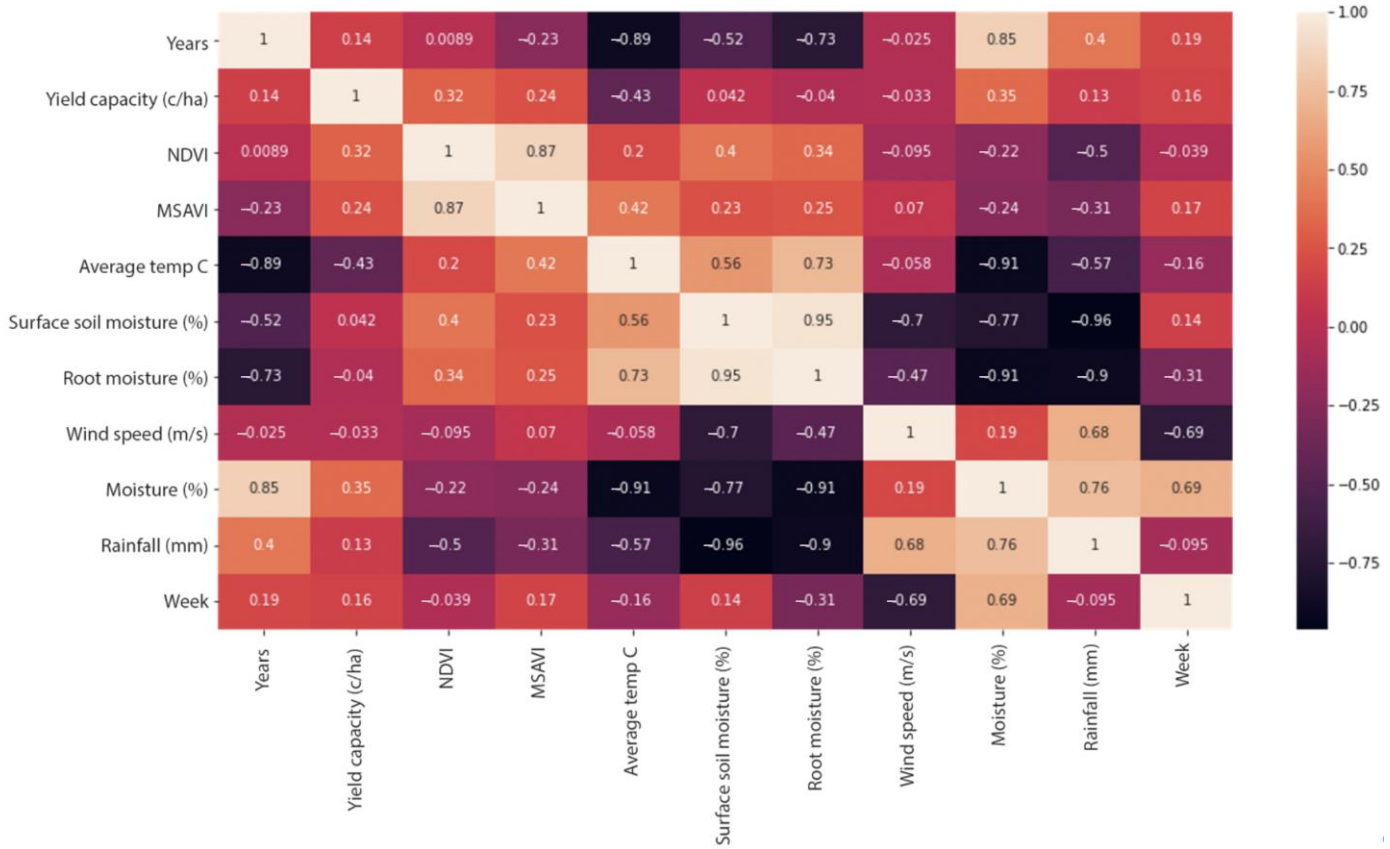


Figure 3. Correlation Matrix.

Figure 3 shows the correlation coefficients of all data studied with one another. The data with high correlation with yield are the priority indicators. In calculating the parameters, attention is paid to their correlation to yield and NDVI. A low correlation to these indicates little influence, while a high correlation to the other parameters means the same thing. This is because if two parameters are highly correlated with each other, they have the same influence on the other factors because they are interrelated. According to research [44], soil health indicators showed no correlation with crop yields. On the contrary, soil moisture is an important factor for predicting yield. The authors of [45] point out that soil moisture content is an effective factor for biological processes and soil profile evolution. In addition, soil moisture affects the distribution of vegetation. Therefore, a lack of moisture can lead to drought and degradation, especially in rainfall-dependent areas [46]. Remote sensing and GIS methods provide reliable alternatives to traditional methods that can cover vast areas and provide information on spatial and temporal variations in soil moisture content. Soil moisture content is a major environmental pressure in areas that suffer from poor soil drainage, along with high groundwater table fluctuations that affect crop survival, growth and productivity [47]. Various authors have developed new methods for estimating soil moisture content based on soil reflection and ground surface temperature [48]. The surface

moisture content of soil can be estimated based on NDVI. The authors of [49] described this NDVI relationship as a linear relationship with soil moisture content. According to these data, the moisture content parameter of the surface layer of soil was taken for machine learning. When analysing the correlation of the data, it was found that the moisture parameters are related to one another; thus, the moisture, precipitation and root zone moisture parameters were removed. The date and week information was also removed, as these parameters are consistent and would only interfere with model training.

Statistical models and machine learning tools were used to simulate crop yield variability using weather indices as inputs. Extreme weather events from the recent past, such as the droughts in Russia in 2010–2011 and in the United States (US) in 2012, and their impact on regional crop production and world commodity markets clearly led to the need to also account for extreme weather conditions when modelling crop yields [50]. For example, winter wheat has been shown to be particularly susceptible to negative temperatures in autumn and to heat stress during grain filling and stem lengthening [51]. This vulnerability to extreme temperatures is thought to be responsible for the decline in wheat yields throughout Europe [52]. According to another study [53], every day above 30 °C causes yield reductions of up to 6% in maize and soybean under rainfed conditions. Similarly, inter-annual variations in rainfall also play a decisive role in crop growth. Although several studies have considered indices of weather extremes in their analyses, their scope has either been limited to measuring the conditional relationship with yield [54] or the types of extremes considered have been limited [55]. A non-linear and threshold relationship has been shown to exist between yield and weather indices [56]. However, most previous studies modelled this non-linearity using regression models with quadratic terms for average weather indices without appropriate justification. Understanding the exact relationship between weather indices and crop yields is important, given that a previous study reported significant stagnation and declines in yields of major cereal crops for more than a quarter of the world's cropland [57].

4. The Use of Machine Learning and Neural Network Technologies for Yield Forecasting

4.1. Scikit-Learn Machine Learning Model

The Scikit-learn library is used for machine learning. Scikit-learn is one of the most widely used Python packages for Data Science and Machine Learning. It allows many operations and provides many algorithms. Scikit-learn also offers excellent documentation about its classes, methods and functions, as well as a description of the algorithms used.

The database is divided into two samples. One is taken as a training sample to train the model, and the second one is used to test the already trained model (Figure S4).

For the best results, several activation functions are used to train the model. To avoid having to rewrite the code to create the model and its outputs, the predict() function is created to take as input only the activator of the desired model function (Figure S5).

Figure 4 shows the flowchart of the yield forecasting methodology. The first flowchart is defined as weather and vegetation index data. The collected data are separated by crop and then sorted, reducing the number of categories and removing empty values. According to the timing of the two growing seasons, two samples are taken from the processed data for the entire growing season and the sowing and harvesting period. The sample is combined with the yield data, and the correlation of all indicators on the collected data is calculated. By examining the correlation of data and studies on the topic, categories are selected that will help in predicting yields.

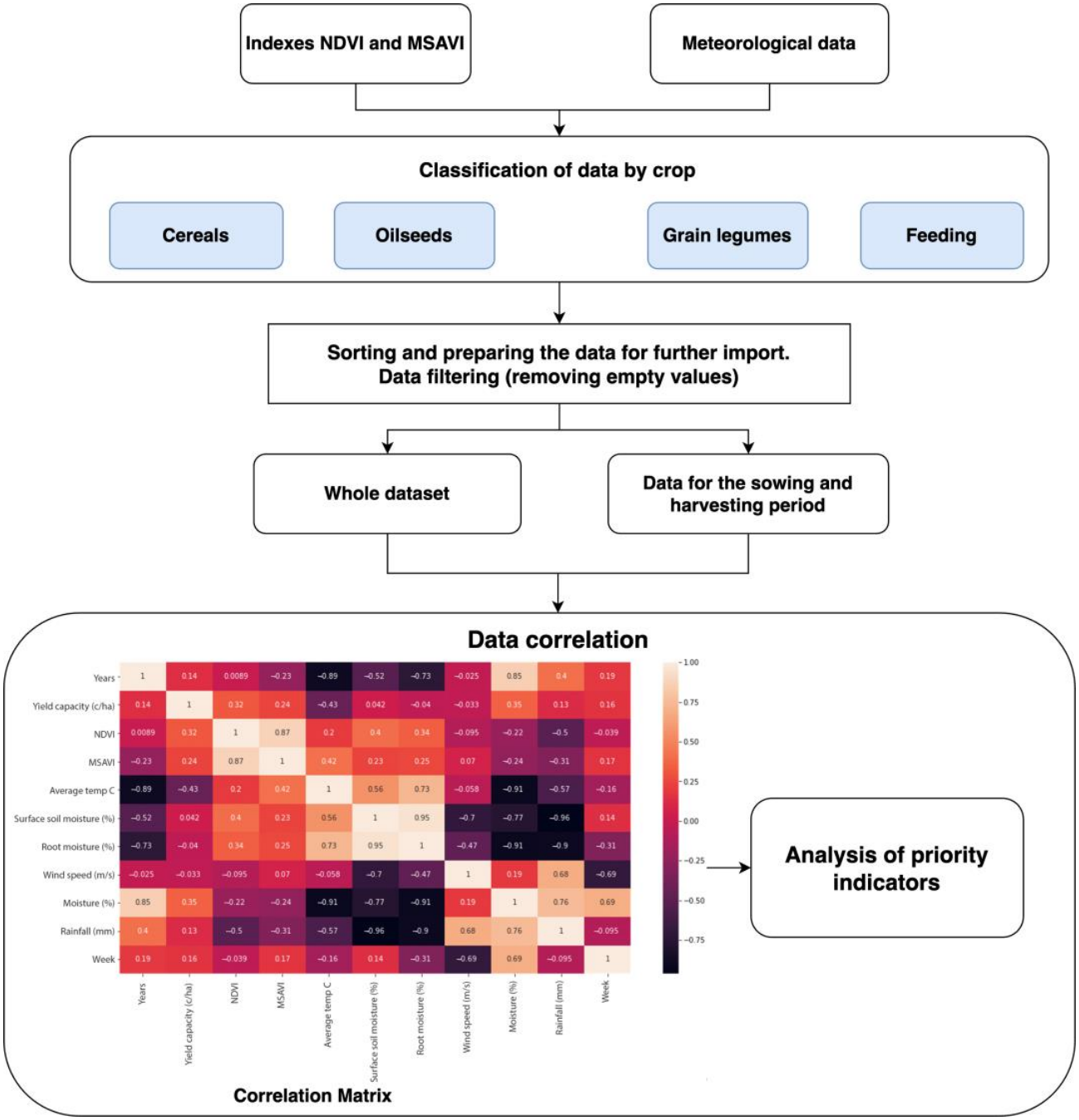


Figure 4. Diagram of the priority indicators analysis stage of the yield forecasting procedure.

Figure 5 shows how the training and test samples are formed based on the selected data. Based on the training sample, the neural network model is trained. During training, the training results of six activation functions are tested on the test sample data. If the prediction accuracy results are low, the neural network automatically corrects its weights and retrains. When the final error value reaches the minimum possible value, the neural network completes the training. If the prediction accuracy of the trained machine learning model is high, the model is stored and used for yield prediction.

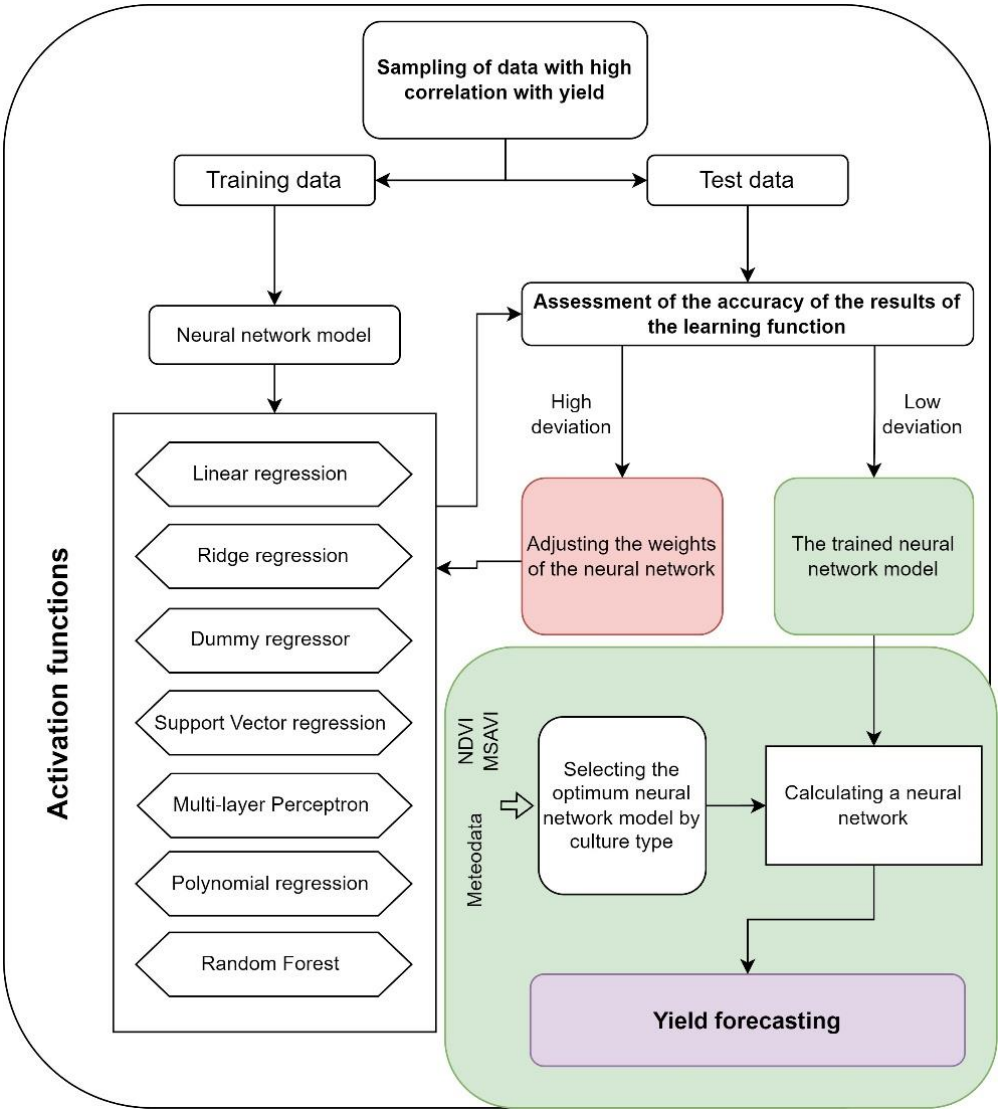


Figure 5. Diagram of the forecasting stage based on identified priority indicators.

The performance of the model was estimated using mean squared error (*MSE*) and root mean square error (*RMSE*). Each output of the function is compared with the actual value, and the prediction accuracy is measured using *MSE* and *RMSE*. These are defined by Equations (7) and (8).

Mean squared error:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \quad (7)$$

where n—number of data points; Y_i—observed values; $\hat{Y}_i$—predicted values.

MSE is the mean squared error [58] used as a loss function for least-squares regression. It is the sum over all data points of the squared difference between predicted and actual target variables divided by the number of data points.

Root mean square error:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2} \quad (8)$$

where i—number of data points; N—number of non-missing data points; x_i—actual observations time series; $\hat{x}_i$—estimated time series.

RMSE is the square root of *MSE* [59]. *MSE* is measured in units that are the square of the target variable, while *RMSE* is measured in the same units as the target variable. Because of its formulation, *MSE*, like the loss squared function from which it derives, effectively penalises more serious errors.

4.2. Activation Functions Used to Train the Neural Network

Artificial neural networks are mathematical tools and flexible structures that can create a non-linear mapping between input and output spaces [60]. Neural networks mimic the function of neurons in the brain to process information and make decisions. With their learning capability, they can identify and model complex non-linear relationships between inputs and outputs of biological processes, providing superior predictive power over traditional statistical methods for understanding complex relationships in communication formulations [61]. The deep learning method promises great advances in natural science data processing because it can account for large input dataset sizes, non-linear relationships and relationships between multiple variables [62]. Convergent neural networks (CNNs), the most popular model among deep learning methods, have been used to estimate and predict crop yields [63]. Deep learning methods with multiple hidden layers tend to perform better than artificial neural network models with a single hidden layer. However, deep learning models are more difficult to train and require more advanced hardware and optimisation techniques [64].

Among various ANN methods and learning algorithms, Multilayer Perception Networks (MLP) and radial basis function (RBF) are the most popular neural network models [65]. MLP can identify the relationship between output and input variables and detect intrinsic knowledge in a dataset without prior physical consideration. An RBF network is a combination of an input neuron layer, a hidden layer of RBF neurons and an output neuron layer that linearly connects the hidden layer to the output node and computes the input variables passed from the input layer to the hidden layer. The non-linear transformation process also takes place in the hidden layer, resulting in a mapping between the input neurons and the hidden layers [66]. Neural networks are widely used in agricultural fields such as grain yield in ajouan [67] and sunflower [68], winter rapeseed yield modelling [69], and yield prediction in winter wheat, maize and rice. In addition, studies have been conducted on predicting average regional wheat yield and production and grain yield in sesame, modelling oil yield in sesame, and studying biological and grain yield in barley.

Linear regression was used as the first activation function. It is known as Linear-Regression() [70]. Linear regression fits a linear model with coefficients $w = (w_1, \ldots, w_p)$ to minimise the residual sum of squares between the observed targets in the data set and the targets predicted by the linear approximation. The equation through which this function is calculated is as follows (9):

$$y = ax + b \tag{9}$$

where a, b are the coefficients of the weights, which are determined by the technique during training.

Ridge regression [71] solves some problems of ordinary least squares by imposing a penalty on the size of the coefficients. The ridge coefficients minimise the penalty residual sum of squares (10):

$$min_{\omega} ||X\omega - y||_2^2 + a||\omega||_2^2 \tag{10}$$

The complexity parameter $a \geq 0$ controls for the amount of shrinkage: the greater the value of a, the greater the amount of shrinkage, so that the coefficients become more robust to collinearity.

By adding a degree of bias to the model coefficients, ridge regression reduces their variance, thereby yielding estimates that are more robust. Equation (11) is a multiple linear regression model between yields and predictors, with β_j representing coefficients. In addition to minimising the deviation from y_i, the target function for the ridge regression shown in Equation (12) also includes a penalty term that reduces the coefficient values

closer to the "true" population parameters. This penalty term, also called L2 regularisation, is equal to the square of the coefficient values. The tuning parameter (λ) controls the strength of the regularization. When λ = 0, the ridge regression reduces to a multiple linear regression, and when λ = ∞, all coefficients drop to 0.

$$y_i = \beta_0 + \sum_{j=1}^{12} \beta_j x_{ij} \tag{11}$$

$$argmin \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{12} \beta_j x_{ij} \right)^2 + \lambda \sum_{12}^{12} \beta_j^2 \tag{12}$$

The Dummy Regressor [72] is a kind of regressor that provides predictions based on simple strategies, without considering the input data.

Support Vector Regression [73] is an algorithm that allows one to choose a range of acceptable errors, both through an acceptable error rate (ε) and through tuning beyond this acceptable error rate. The operation of the function is depicted as $| y_i - w_i x_i | \leq \varepsilon$. Where y_i is the target, w_i is the coefficient, and x_i is the predictor (feature). The methodology derived from a regression of the reference vectors depends only on a subset of the training data, as the cost function ignores samples whose predictions are close to their target. The prediction of this function is calculated by formula (13).

$$\sum_{i \in SV} (a_i - a_i^*) K(x_i, x) + b \tag{13}$$

The parameters are accessed through the attributes dual_coef_, which contains the difference $a_i - a_i^*$, support_vectors_—reference vectors, the intercept and independent term b.

The general form of the regression equation for SVR is shown in Equation (14), where $<\beta, X>$ denotes the dot product and β is the vector of coefficients. The complexity of the model can be controlled by looking for a small β. This can be ensured by minimising the norm $| | \beta | | 2 = < \beta, \beta >$.

$$f(x) = \langle \beta, X \rangle + b \tag{14}$$

Multilayer Perceptron [74] (MLP) is a supervised learning algorithm that learns the function $f(\cdot)$: $R^m \rightarrow R^o$ by training on a data set where $X = x_1, x_2, \ldots, x_m$ is the number of measurements to input, and y is the number of measurements to output. Given a feature set and a target, it can learn a nonlinear function approximation either for classification or for regression. It differs from logistic regression in that there may be one or more non-linear layers, called hidden layers, between the input and output levels. Figure 6 shows an MLP with one hidden layer with a scalar output.

Polynomial regression [75] is a form of linear regression, known as a special case of multiple linear regression, which estimates the relationship as a polynomial of degree n. Unlike ordinary linear regression, polynomial regression is more flexible and forms a natural curve rather than a straight line, depending on the values. It is expressed by Formula (15).

$$y = a_1 x_1 + a_2^2 x_2 + \cdots a_n^n x_n + b \tag{15}$$

Random Forest (RandomForest) [76] is a meta-estimate that matches a number of classifying decision trees on different samples of a dataset and uses averaging to improve prediction accuracy and control overfitting (Figure 7). The subsample size is controlled by the max_samples parameter if bootstrap = True (default); otherwise, the whole dataset is used to construct each tree. In random forests, each tree in an ensemble is built from a sample taken with replacement from the training set. In addition, when partitioning each node during tree construction, the best partition is determined either from all input objects or from a random subset of max_features size. The purpose of these two sources of randomness is to reduce the variance of the forest estimate. Indeed, individual decision trees typically exhibit high variance and a tendency toward overfitting. The introduced randomness in the forest leads to decision trees with somewhat unrelated prediction errors.

By taking the mean of these predictions, a number of errors are negated. Random forests achieve variance reduction by combining different trees, sometimes at the cost of a small increase in bias. In practice, the reduction in variance is often significant, resulting in an overall improvement in the model.

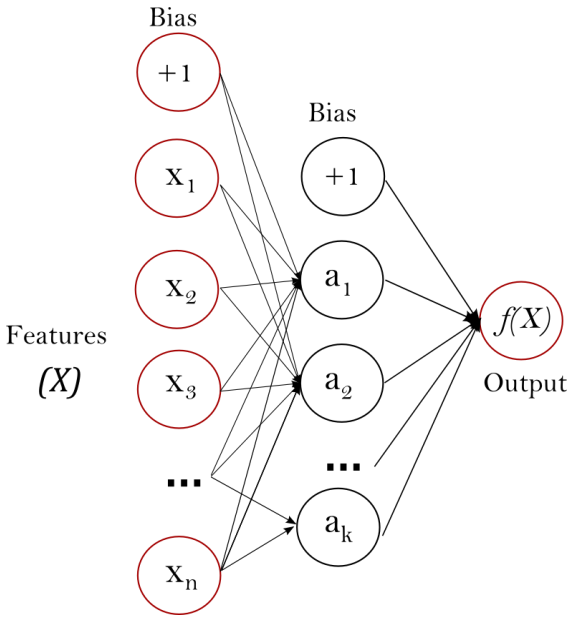


Figure 6. One hidden MLP layer.

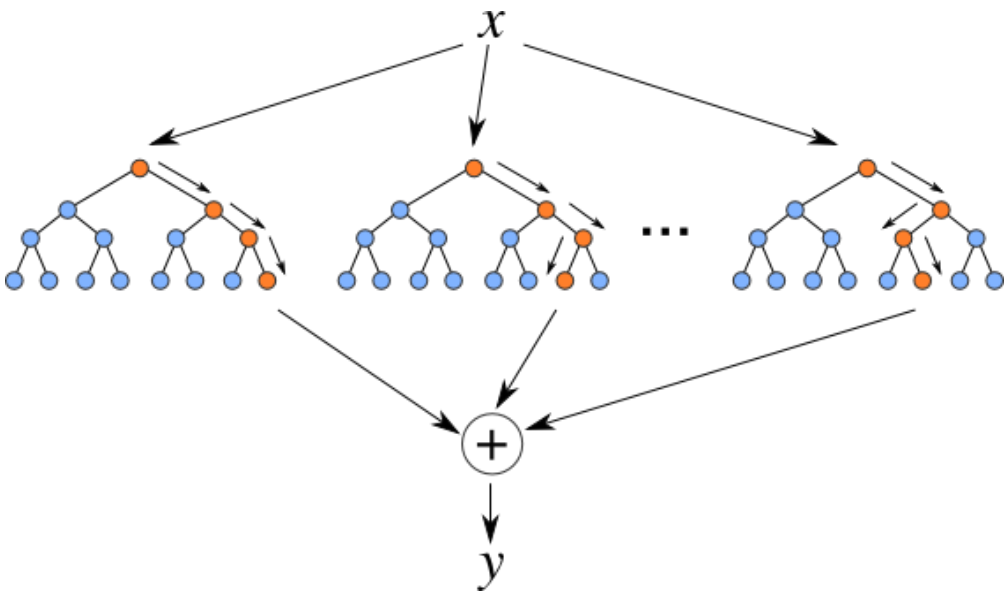


Figure 7. Diagram of a random forest.

4.3. Predicting NDVI and MSAVI Indices

Once the activation functions have been selected, model training and yield prediction starts. Model training will be in two time ranges for different crop growth periods. The first is for the entire growing season, from the beginning of the first sprouts of greens in the fields until they fade (April–November). The second is for the period from sowing the seeds to harvesting the crop (April to July). All activation functions are used on each crop in order to identify the most optimal yield prediction methods. Before training the model to predict yield, the NDVI and MSAVI indices should be predicted.

NDVI is a vegetative index that determines the photosynthetic activity of a plant. It is calculated according to how the plants absorb or reflect infrared rays. The better developed the plants are and the greener the mass they have during the growing season, the greater the NDVI index. The most common use of this index is to track plant development. For the ease of working with this index, a colour scheme from lighter tones to darker tones is used, which is mapped on the field map. The darker the tone, the greater the index. Thanks to the NDVI index, the mistakes of mechanisers and the development of weeds in the period after sowing are clearly visible, because the weeds gain green mass faster and are visible on the spectrum.

Image processing studies were reported on the weed density estimation for soya crops [77] and the modelling of leaf disease using neural networks [78]. The approach of these studies may be useful in improving the accuracy of the current method by accounting for these additional factors.

The authors of [79] point out that crop yields are closely related to NDVI at early reproductive and late maturity stages. For wheat, NDVI values were highly correlated with yield (R^2 = 0.601–0.809) from the medium reproductive to the early maturing stage. Using NDVI values, it was possible to distinguish between fertiliser application levels. Their results showed that small NDVI values are effective in predicting yields and determining fertiliser application levels.

At the peak of the growing season, peak NDVI gives a good prediction. The range of deviation of the forecast from the fact is from 4 to 24%. The average (for many years) forecast error is no more than 20%. Therefore, NDVI prediction is the first important indicator to determine the optimal learning function of the model. An analysis of the accuracy of the NDVI prediction of the learning functions of all the crops studied is shown in Table 2.

Table 2. NDVI prediction error analysis results for the growing season and * for the sowing and harvesting period.

Culture	Activation Function	MSE	RMSE	MSE *	RMSE *
Grain legume crop	Linear regression	0.0448	0.2117	0.0148	0.1216
	Ridge regression	0.0439	0.2097	0.01	0.1
	Dummy regressor	0.0545	0.2335	0.0052	0.0723
	Support vector regression	0.0346	0.1861	0.0154	0.1241
	Multilayer perceptron	1.0408	1.0202	0.0398	0.1995
	Polynomial regression	0.0366	0.1914	0.017	0.1304
	Random forest	0.0524	0.229	0.0152	0.1234
Oilseed crop	Linear regression	0.0517	0.2275	0.0226	0.1506
	Ridge regression	0.0486	0.2205	0.0162	0.1274
	Dummy regressor	0.0449	0.212	0.0175	0.1325
	Support vector regression	0.0363	0.1904	0.0079	0.0889
	Multilayer perceptron	0.3283	0.5729	0.0911	0.3019
	Polynomial regression	0.0397	0.1993	0.0194	0.1396
	Random forest	0.0329	0.1814	0.0123	0.111
Feed crop	Linear regression	0.0244	0.1561	0.0185	0.136
	Ridge regression	0.0253	0.1591	0.0246	0.1568
	Dummy regressor	0.0413	0.2033	0.0448	0.2118
	Support vector regression	0.0164	0.1279	0.016	0.1265
	Multilayer perceptron	0.0477	0.2185	0.6684	0.8175
	Polynomial regression	0.023	0.1517	0.03	0.175
	Random forest	0.0284	0.1685	0.0226	0.1505

Table 2. *Cont.*

Culture	Activation Function	MSE	RMSE	MSE *	RMSE *
Cereal crop	Linear regression	0.0247	0.1573	0.012	0.1098
	Ridge regression	0.0289	0.1699	0.009	0.0952
	Dummy regressor	0.044	0.2097	0.0098	0.0994
	Support vector regression	0.0279	0.1671	0.0134	0.1159
	Multilayer perceptron	0.0398	0.1995	1.6156	1.271
	Polynomial regression	0.0227	0.1506	0.0128	0.1135
	Random forest	0.0243	0.1558	0.0111	0.1056

Histograms of the variance distribution (diff) in the predictions were created by the Seaborn library with the displot tool (Figures 8–11). The activation function histograms in Figure 8 show the distribution of the deviations of the machine learning predictions from the actual values. The minimum value is observed in the Random Forest function.

The activation function histograms in Figure 9 show the distribution of the deviations of the machine learning predictions from the actual values. The minimum value is observed in the Ridge regression function.

The activation function histograms in Figure 10 show the distribution of the deviations of the machine learning predictions from the actual values. The minimum value is observed in the Random Forest function.

The activation function histograms in Figure 11 show the distribution of the deviations of the machine learning predictions from the actual values. The minimum value is observed in the Random Forest function. As a result of this research, a comprehensive yield forecasting methodology was developed, as shown in Figure 12, which demonstrates the significant role of machine learning in providing accurate yield forecasts.

Since NDVI values are given in small numbers (0 to 1), it is difficult to understand from MSE results how accurate the model predictions are; thus, RMSE was used, which is more sensitive to outliers. Analysis of the data error results for the whole growing season revealed the following optimal training functions:

(1) Support Vector Regression with an RMSE error of 0.1861 for the leguminous crops;
(2) Random Forest with an RMSE error of 0.1814 for oilseeds;
(3) Support Vector Regression with an RMSE error of 0.1279 for forage crops;
(4) Polynomial Regression with RMSE error of 0.1506 for grain crops.

For sowing and harvesting period data, the accuracy is higher due to a smaller distribution of NDVI values. The best results for the sowing and harvesting period showed the following activation functions:

(1) Dummy Regressor with an RMSE error of 0.0723 for the leguminous crops;
(2) Support Vector Regression with an RMSE error of 0.08 for oilseed;
(3) Support Vector Regression with an RMSE error of 0.1265 for forage crops;
(4) Ridge Regression with an RMSE error of 0.0952 for grain crops.

Despite the high performance of some of the learning functions, there is good accuracy in some of the studied crops. Arithmetic mean error values are calculated to identify the best activation function for all crops.

In Table 3, the Support Vector Regression function shows the best results (RMSE of 0.1679) for NDVI prediction compared to the others among the prediction results for the whole growing season. The polynomial regression function has high performance (RMSE—0.1732), ranking second in prediction accuracy. Among the data for the sowing and harvesting period, Support Vector Regression (RMSE—0.1138) and Ridge Regression (RMSE—0.1198) showed the best results.

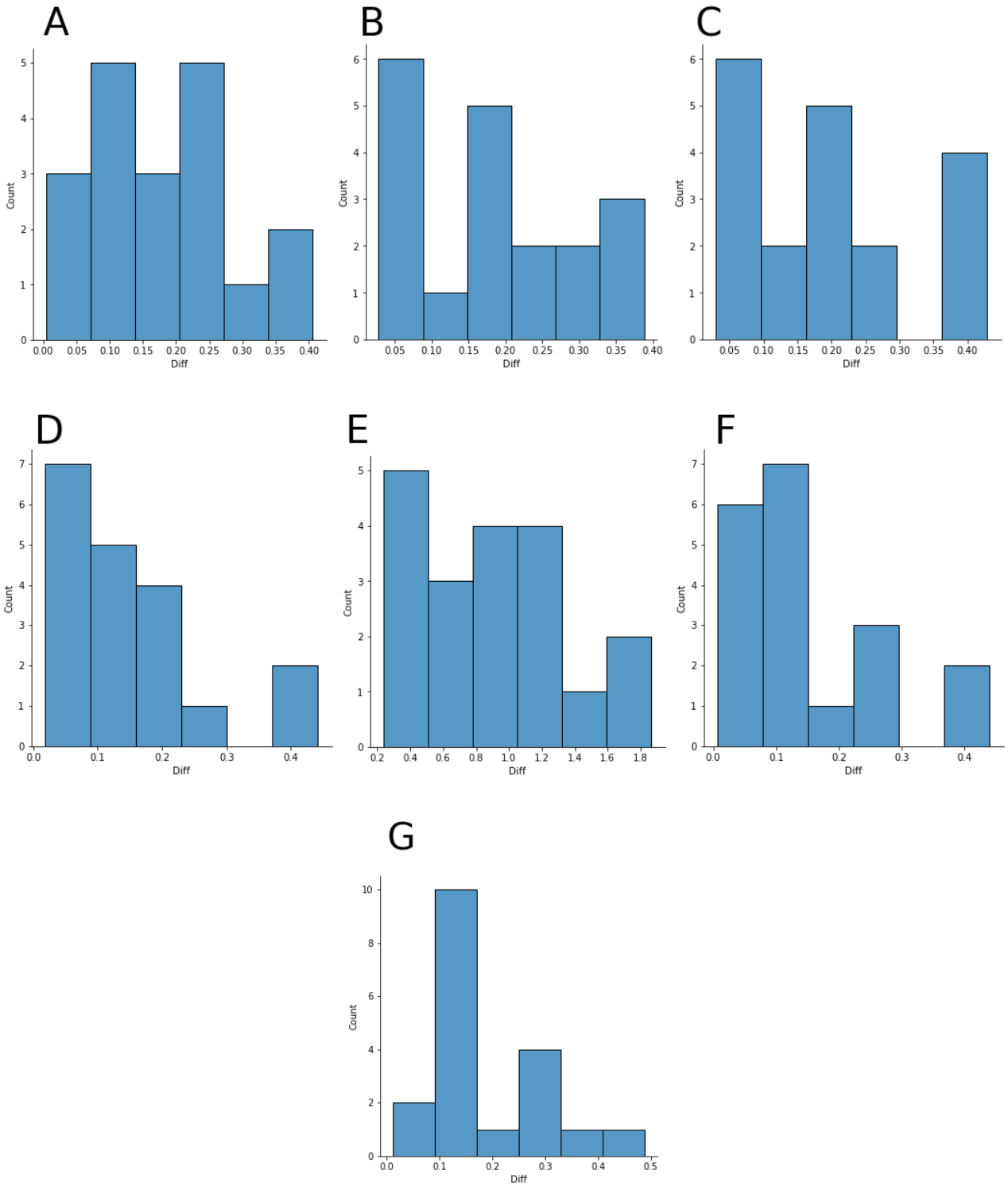


Figure 8. Histograms of the distribution of the variance of the results of the predictions of the bean crop activation functions: (**A**) Linear Regression, (**B**) Ridge Regression, (**C**) Dummy Regression, (**D**) Support Vector Regression, (**E**) Multilayer Perceptron, (**F**) Polynomial Regression, (**G**) Random Forest.

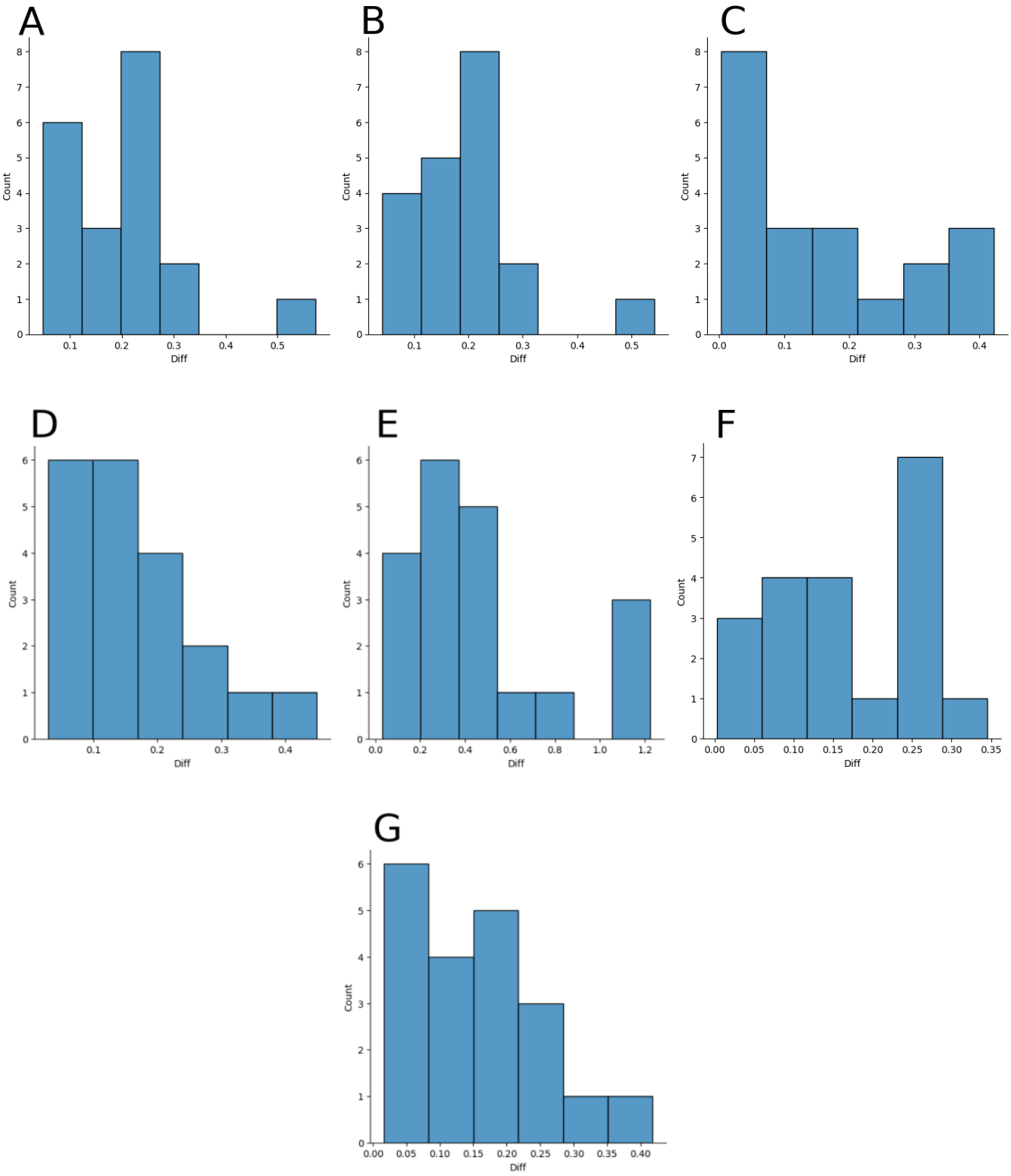


Figure 9. Histograms of the distribution of the variance of the results of the predictions of the oil crop activation functions: (**A**) Linear Regression, (**B**) Ridge Regression, (**C**) Dummy Regression, (**D**) Support Vector Regression, (**E**) Multilayer Perceptron, (**F**) Polynomial Regression, (**G**) Random Forest.

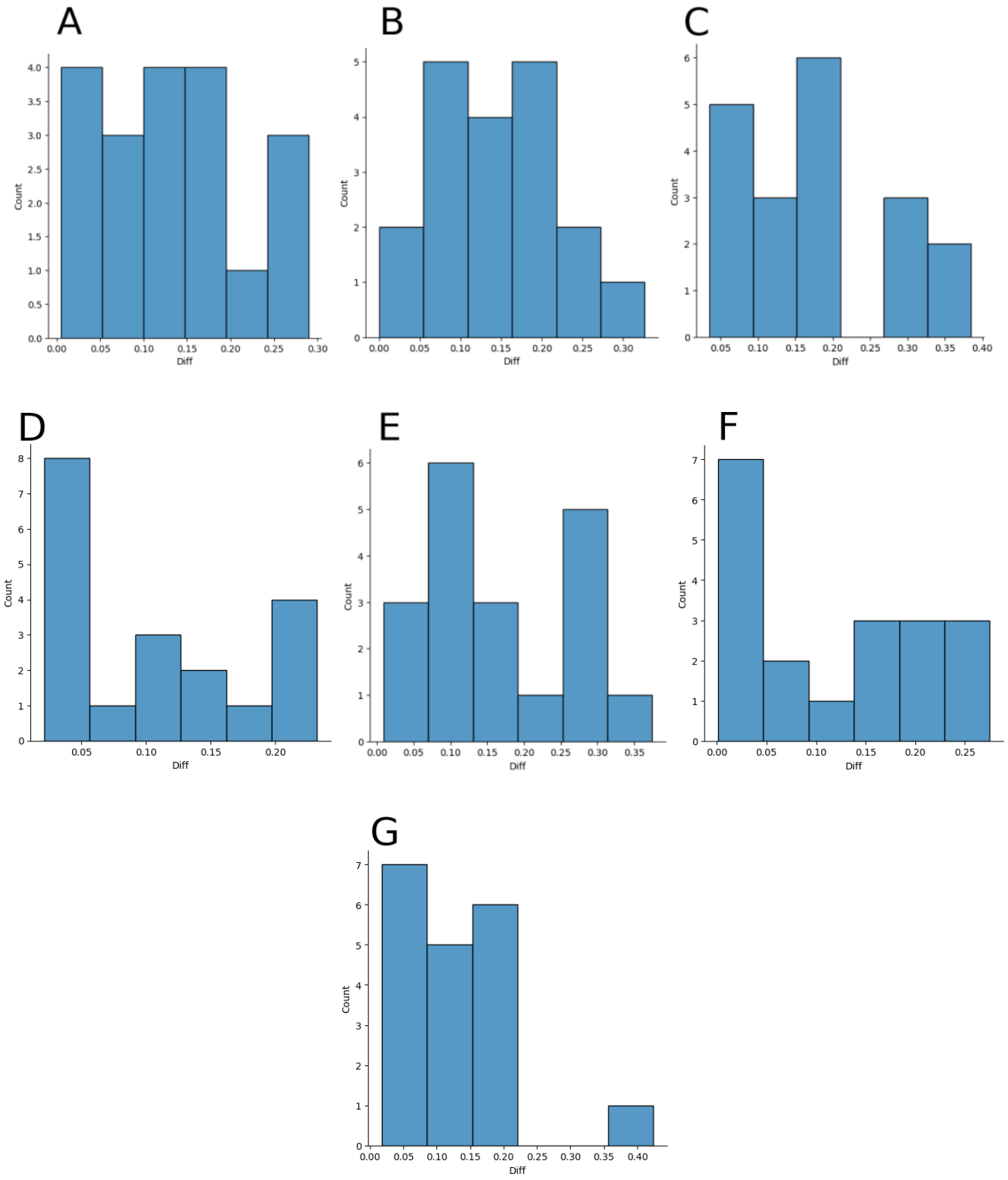


Figure 10. Histograms of the distribution of variance of the results of predictions of the feed crop activation functions: (**A**) Linear Regression, (**B**) Ridge Regression, (**C**) Dummy Regression, (**D**) Support Vector Regression, (**E**) Multilayer Perceptron, (**F**) Polynomial Regression, (**G**) Random Forest.

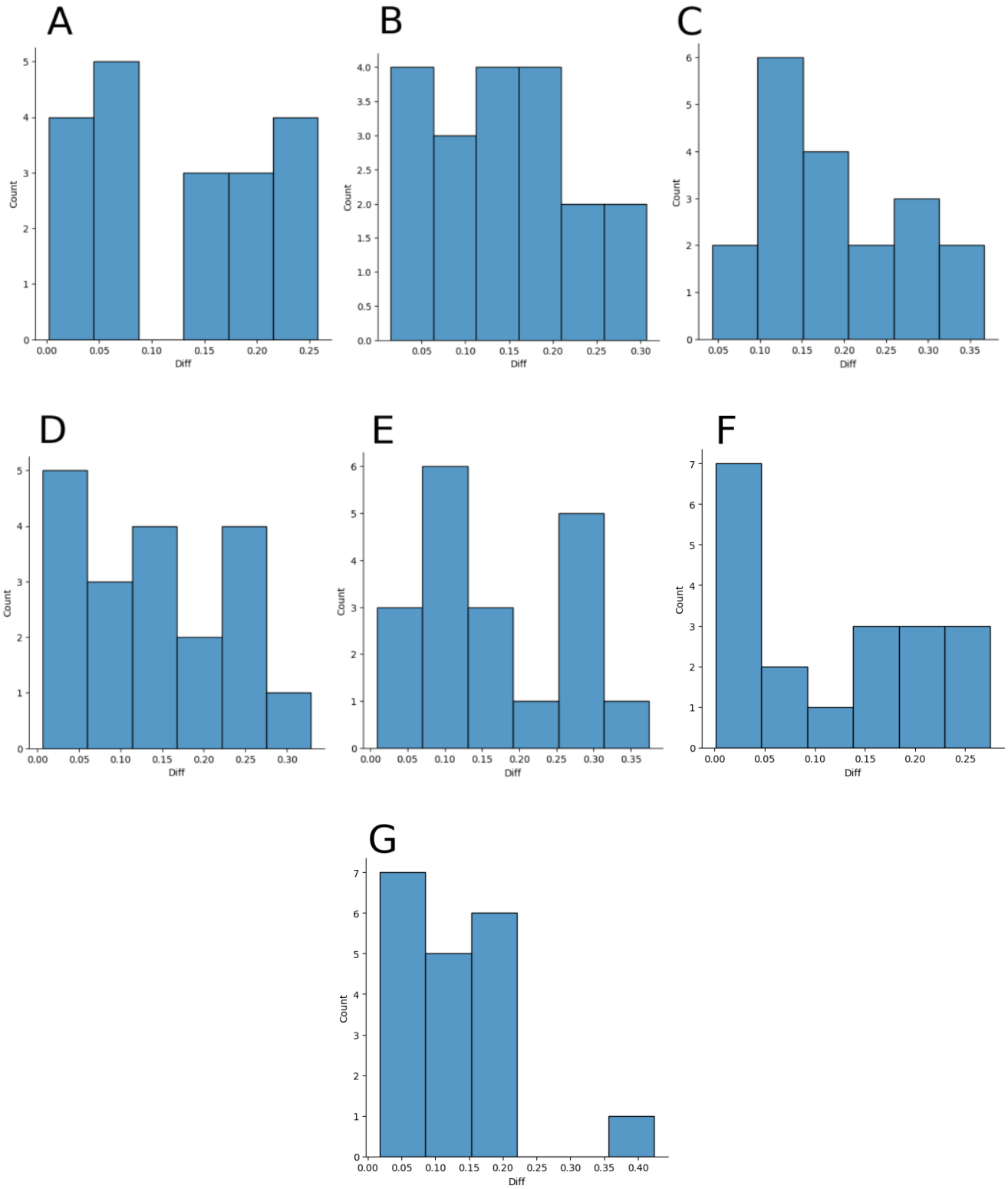


Figure 11. Histograms of the distribution of variance of the results of predictions of the grain crop activation functions: (**A**) Linear Regression, (**B**) Ridge Regression, (**C**) Dummy Regression, (**D**) Support Vector Regression, (**E**) Multilayer Perceptron, (**F**) Polynomial Regression, (**G**) Random Forest.

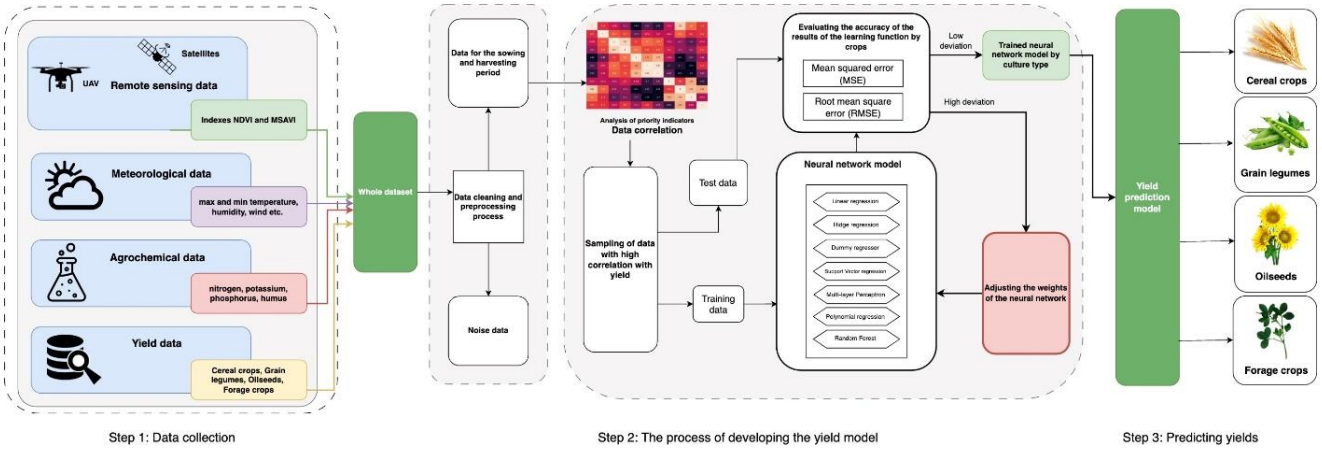


Figure 12. Yield forecasting—the overall procedure and the key outcomes.

Table 3. Arithmetic mean values of NDVI prediction errors for all crops for the growing season and * for the sowing and harvesting period.

Activation Function	MSE	RMSE	MSE *	RMSE *
Linear Regression	0.0364	0.1881	0.0167	0.1295
Ridge Regression	0.0367	0.1898	0.0143	0.1198
Dummy Regressor	0.0462	0.2146	0.0166	0.129
Support Vector Regression	0.0288	0.1679	0.0129	0.1138
Multilayer Perceptron	0.2528	0.5028	0.4192	0.6475
Polynomial Regression	0.03	0.1732	0.0194	0.1396
Random Forest	0.0337	0.1837	0.015	0.1226

The modified soil-adjusted vegetation index (*MSAVI)* is a soil-adjusted vegetation index that aims to remove some of the limitations of *NDVI* when applied to areas with a high degree of soil surface exposure. *MSAVI* is defined in Equation (16):

$$MSAVI = \left(2 \times NIR + 1 - sqrt\left((2 \times NIR + 1)^2 - 8 \times (NIR - RED)\right)\right)/2 \quad (16)$$

Based on the results of the *MSAVI* index studies, the best activation functions with the best predictive accuracy by crop for the agrometeorological growing season are as follows:

(1) Random Forest with an RMSE error of 0.057 for the leguminous crops;
(2) Random Forest with an RMSE error of 0.091 for oilseeds;
(3) Random Forest with an RMSE error of 0.052 for forage crops;
(4) Polynomial Regression with an RMSE error of 0.011 for grain crops.

For the organic growing season:

(1) Ridge Regression with an RMSE error of 0.0117 for leguminous crops;
(2) Ridge Regression with an RMSE error of 0.0124 for oilseed crops;
(3) Random Forest with an RMSE error of 0.0152 for forage crops;
(4) Dummy Regressor with an RMSE error of 0.0293 for grain crops.

4.4. Predicting Yields

By defining the key functions for predicting NDVI and MSAVI indices, yield prediction results will be calculated. Since the yield value is calculated as a single value for the whole year, averaged data values will be used to predict it.

The weather and vegetation index database is sorted by year and concat() is merged with the yield file data. At the time of sorting, the date and week categories are removed, as the data are sorted by year (Figure S5).

Yield predictions of activation functions for each crop for two periods are checked. The results are presented in the Supplementary Materials (Table S1). According to the results shown in Table S1, the best activation functions by crop for the entire growing season are as follows:

(1) Leguminous crops—Polynomial Regression (prediction accuracy 97.1%, RMSE—0.27);
(2) Oilseed crop—Multilayer Perceptron (prediction accuracy 99.24%, RMSE—0.09);
(3) Forage crop—Linear Regression (prediction accuracy 99.58%, RMSE—0.04);
(4) Grain crop—Multilayer Perceptron (prediction accuracy 83.88%, RMSE—1.77).

For the sowing and harvesting period:

(1) Leguminous crops—Multilayer Perceptron (prediction accuracy 90.69%, RMSE—0.33);
(2) Oilseed crop—Multilayer Perceptron (prediction accuracy 97.19%, RMSE—0.09);
(3) Forage crop—Polynomial Regression (prediction accuracy 87.6%, RMSE—1.36);
(4) Grain crop—Multilayer Perceptron (prediction accuracy 82.75%, RMSE—2.00).

5. Conclusions

In conclusion, this paper highlights the crucial power of machine learning, in particular neural network technology, in crop yield prediction. Using the capabilities of the Python programming language libraries, a yield prediction model was developed by extensive testing using weather data and crop vegetation indices. By analysing these datasets, it was found that yield prediction was possible using NDVI vegetation index data and weather parameters such as mean temperature, surface soil moisture and wind speed.

Machine learning programming languages make it possible to calculate complex relationships between these parameters and build a neural network capable of accurately predicting yields. The neural network uses data weights that are adjusted using activation function algorithms, which allows it to make reliable predictions. Consequently, using machine learning technologies (neural networks) allows accurate crop yield prediction if the relevant data are available.

It is important to note that a customised machine learning approach makes it possible to identify the optimal learning algorithms for each specific crop, further increasing the accuracy of the predictions.

In performing the neural network training and selecting the best activation functions, the following results were revealed:

(1) Prediction readings of all activation functions have a higher accuracy if data for training are taken for the whole growing season rather than for the sowing and harvesting period. The average prediction accuracy for the growing season was 95%, whereas the average for the sowing and harvesting period was slightly lower, at 89.5%.
(2) The calculated best functions for the vegetation period for all crops are Multilayer Perceptron, with a prediction accuracy ranging from 66% to 99% (mean 85%), and Polynomial Regression, with a prediction accuracy range from 63% to 98% (mean 82%). It should be understood that the Multilayer Perceptron function generates new random weights at the start of each training, so the results it produces may vary between +5% and −5%. This module takes the average of these ranges.
(3) The Ridge, Dummy Regressor, Support Vector Regression and Random Forest functions gave similar prediction results, with prediction accuracy between 70% and 80% (78% on average).
(4) For the forage crop, the linear learning functions proved to be more efficient compared to the others.

As can be seen, the prediction accuracy ranges sometimes have a difference of up to 30%. This is due to the fact that there are other factors affecting the yield. It is possible to adjust the developed code of the yield prediction methodology by adding new indicators and updating the methodology, which will reduce the range of prediction bias. Similarly,

these technologies are used to predict other values, such as NDVI and MSAVI, which can help to build other relationships.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/agriculture13061195/s1, Figure S1: Calculating the average temperature. Figure S2: Sorting and cleaning the database. Figure S3: Dividing the database into a training database and a test database. Figure S4: code snippet of predict() function. Figure S5: Creating a Pandas database with yields. Table S1: Results of yield predictions.

Author Contributions: M.S.: funding acquisition, concept development, project supervision; writing the original draft; N.B.: mathematical processing of results, artwork, work with satellite images; P.S.V.: methodology check, writing, reviewing and editing; T.P.: editing the original draft. All authors have read and agreed to the published version of the manuscript.

Funding: The research was carried out within the framework of project BR10865102 "Development of scientific and methodological approaches to the introduction of remote sensing (RS) technologies to improve agricultural management", funded by the Ministry of Agriculture of the Republic of Kazakhstan.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data generated in the current study are available in the Supplementary Materials.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Phalan, B.; Green, R.; Balmford, A. Closing yield gaps: Perils and possibilities for biodiversity conservation. *Philos. Trans. R. Soc. B Biol. Sci.* **2014**, *369*, 20120285. [CrossRef]
2. Basso, B.; Liu, L. Chapter Four—Seasonal Crop Yield Forecast: Methods, Applications, and Accuracies. In *Advances in Agronomy*; Sparks, D.L., Ed.; Academic Press: Cambridge, MA, USA, 2019; Volume 154, pp. 201–255.
3. Fischer, R.A. Definitions and Determination of Crop Yield, Yield Gaps, and of Rates of Change. *Field Crops Res.* **2015**, *182*, 9–18. [CrossRef]
4. Chipanshi, A.; Zhang, Y.; Kouadio, L.; Newlands, N.; Davidson, A.; Hill, H.; Warren, R.; Qian, B.; Daneshfar, B.; Bedard, F.; et al. Evaluation of the Integrated Canadian Crop Yield Forecaster (ICCYF) Model for in-Season Prediction of Crop Yield across the Canadian Agricultural Landscape. *Agric. For. Meteorol.* **2015**, *206*, 137–150. [CrossRef]
5. López-Lozano, R.; Duveiller, G.; Seguini, L.; Meroni, M.; García-Condado, S.; Hooker, J.; Leo, O.; Baruth, B. Towards Regional Grain Yield Forecasting with 1 km-Resolution EO Biophysical Products: Strengths and Limitations at Pan-European Level. *Agric. For. Meteorol.* **2015**, *206*, 12–32. [CrossRef]
6. Bussay, A.; van der Velde, M.; Fumagalli, D.; Seguini, L. Improving Operational Maize Yield Forecasting in Hungary. *Agric. Syst.* **2015**, *141*, 94–106. [CrossRef]
7. Lobell, D.B.; Thau, D.; Seifert, C.; Engle, E.; Little, B. A Scalable Satellite-Based Crop Yield Mapper. *Remote Sens. Environ.* **2015**, *164*, 324–333. [CrossRef]
8. Zhao, Y.; Potgieter, A.B.; Zhang, M.; Wu, B.; Hammer, G.L. Predicting Wheat Yield at the Field Scale by Combining High-Resolution Sentinel-2 Satellite Imagery and Crop Modelling. *Remote Sens.* **2020**, *12*, 1024. [CrossRef]
9. Newlands, N.K.; Zamar, D.S.; Kouadio, L.A.; Zhang, Y.; Chipanshi, A.; Potgieter, A.; Toure, S.; Hill, H.S.J. An Integrated, Probabilistic Model for Improved Seasonal Forecasting of Agricultural Crop Yield under Environmental Uncertainty. *Front. Environ. Sci.* **2014**, *2*, 17. [CrossRef]
10. Hollinger, D.L. *Crop Condition and Yield Prediction at the Field Scale with Geospatial and Artificial Neural Network Applications*; Kent State University: Kent, OH, USA, 2011.
11. Willcock, S.; Martínez-López, J.; Hooftman, D.A.P.; Bagstad, K.J.; Balbi, S.; Marzo, A.; Prato, C.; Sciandrello, S.; Signorello, G.; Voigt, B.; et al. Machine Learning for Ecosystem Services. *Ecosyst. Serv.* **2018**, *33*, 165–174. [CrossRef]
12. Chlingaryan, A.; Sukkarieh, S.; Whelan, B. Machine Learning Approaches for Crop Yield Prediction and Nitrogen Status Estimation in Precision Agriculture: A Review. *Comput. Electron. Agric.* **2018**, *151*, 61–69. [CrossRef]
13. Heaton, J. Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep Learning. In *Genetic Programming and Evolvable Machines*; The MIT Press: Cambridge, MA, USA, 2016; 800p, ISBN 0262035618. [CrossRef]
14. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; Springer: Berlin/Heidelberg, Germany, 2013; Volume 112.
15. Pantazi, X.E.; Moshou, D.; Alexandridis, T.; Whetton, R.L.; Mouazen, A.M. Wheat Yield Prediction Using Machine Learning and Advanced Sensing Techniques. *Comput. Electron. Agric.* **2016**, *121*, 57–65. [CrossRef]

16. Cai, Y.; Guan, K.; Lobell, D.; Potgieter, A.B.; Wang, S.; Peng, J.; Xu, T.; Asseng, S.; Zhang, Y.; You, L.; et al. Integrating Satellite and Climate Data to Predict Wheat Yield in Australia Using Machine Learning Approaches. *Agric. For. Meteorol.* **2019**, *274*, 144–159. [CrossRef]
17. González-Sanchez, A.; Frausto-Solis, J.; Ojeda, W. Predictive Ability of Machine Learning Methods for Massive Crop Yield Prediction. *Span. J. Agric. Res.* **2014**, *12*, 313. [CrossRef]
18. Rahimov, N.; Dilmurod, K. The Application of Multiple Linear Regression Algorithm and Python for Crop Yield Prediction in Agriculture. *Harv. Educ. Sci. Rev.* **2022**, *2*, 181–187. [CrossRef]
19. Python, W. Python. Python Releases Wind. 2021. Available online: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=1f2ee3831eebfc97bfafd514ca2abb7e2c5c86bb (accessed on 28 May 2023).
20. Randles, B.M.; Pasquetto, I.V.; Golshan, M.S.; Borgman, C.L. Using the Jupyter Notebook as a Tool for Open Science: An Empirical Study. In Proceedings of the 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), Toronto, ON, Canada, 19–23 June 2017; pp. 1–2.
21. Raschka, S.; Patterson, J.; Nolet, C. Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence. *Information* **2020**, *11*, 193. [CrossRef]
22. McKinney, W. Pandas: A Foundational Python Library for Data Analysis and Statistics. *Python High Perform. Sci. Comput.* **2011**, *14*, 1–9.
23. Bisong, E. *Building Machine Learning and Deep Learning Models on Google Cloud Platform*; Springer: Berlin/Heidelberg, Germany, 2019.
24. Oliphant, T.E. *A Guide to NumPy*; The MIT Press: Cambridge, MA, USA, 2006; Volume 1, 371p.
25. Kramer, O. *Machine Learning for Evolution Strategies*; Springer: Berlin/Heidelberg, Germany, 2016; Volume 20.
26. Musah, A.; Dutra, L.M.M.; Aldosery, A.; Browning, E.; Ambrizzi, T.; Borges, I.V.G.; Tunali, M.; Başibüyük, S.; Yenigün, O.; Moreno, G.M.M.; et al. An Evaluation of the OpenWeatherMap API versus INMET Using Weather Data from Two Brazilian Cities: Recife and Campina Grande. *Data* **2022**, *7*, 106. [CrossRef]
27. Qi, J.; Chehbouni, A.; Huete, A.R.; Kerr, Y.H.; Sorooshian, S. A Modified Soil Adjusted Vegetation Index. *Remote Sens. Environ.* **1994**, *48*, 119–126. [CrossRef]
28. Carlson, T.N.; Ripley, D.A. On the Relation between NDVI, Fractional Vegetation Cover, and Leaf Area Index. *Remote Sens. Environ.* **1997**, *62*, 241–252. [CrossRef]
29. Li, F.; Miao, Y.; Chen, X.; Sun, Z.; Stueve, K.; Yuan, F. In-Season Prediction of Corn Grain Yield through PlanetScope and Sentinel-2 Images. *Agronomy* **2022**, *12*, 3176. [CrossRef]
30. Son, N.-T.; Chen, C.-F.; Cheng, Y.-S.; Toscano, P.; Chen, C.-R.; Chen, S.-L.; Tseng, K.-H.; Syu, C.-H.; Guo, H.-Y.; Zhang, Y.-T. Field-Scale Rice Yield Prediction from Sentinel-2 Monthly Image Composites Using Machine Learning Algorithms. *Ecol. Inform.* **2022**, *69*, 101618. [CrossRef]
31. Joshi, A.; Pradhan, B.; Gite, S.; Chakraborty, S. Remote-Sensing Data and Deep-Learning Techniques in Crop Mapping and Yield Prediction: A Systematic Review. *Remote Sens.* **2023**, *15*, 2014. [CrossRef]
32. Liliane, T.N.; Charles, M.S. Factors affecting yield of crops. In *Agronomy Climate Change and Food Security*; 2020. [CrossRef]
33. Ansarifar, J.; Wang, L.; Archontoulis, S.V. An interaction regression model for crop yield prediction. *Sci. Rep.* **2021**, *11*, 17754. [CrossRef]
34. Reback, J.; McKinney, W.; Van Den Bossche, J.; Augspurger, T.; Cloud, P.; Klein, A.; Hawkins, S.; Roeschke, M.; Tratner, J.; She, C.; et al. pandas-dev/pandas: Pandas 1.0. 5. Zenodo. 2020. Available online: https://zenodo.org/record/3898987 (accessed on 28 May 2023).
35. Lemenkova, P. Python Libraries Matplotlib, Seaborn and Pandas for Visualization Geo-Spatial Datasets Generated by QGIS. *An. Stiintifice Ale Univ. Alexandru Ioan Cuza Din Iasi-Ser. Geogr.* **2020**, *64*, 13–32.
36. Lemenkova, P. Geospatial Analysis by Python and R: Geomorphology of the Philippine Trench, Pacific Ocean. *Electron. Lett. Sci. Eng.* **2019**, *15*, 81–94.
37. Konduri, V.S.; Vandal, T.J.; Ganguly, S.; Ganguly, A.R. Data Science for Weather Impacts on Crop Yield. *Front. Sustain. Food Syst.* **2020**, *4*, 52. [CrossRef]
38. Ray, D.K.; Gerber, J.S.; MacDonald, G.K.; West, P.C. Climate Variation Explains a Third of Global Crop Yield Variability. *Nat. Commun.* **2015**, *6*, 5989. [CrossRef]
39. Godfrey, K.R. Correlation Methods. *Automatica* **1980**, *16*, 527–534. [CrossRef]
40. MacKay, D.J.; Mac Kay, D.J. *Information Theory, Inference and Learning Algorithms*; Cambridge University Press: Cambridge, UK, 2003.
41. Fraser, A.M.; Swinney, H.L. Swinney Independent Coordinates for Strange Attractors from Mutual Information. *Phys. Rev. A Gen. Phys.* **1986**, *33*, 1134–1140. [CrossRef]
42. Moon, Y.I.; Rajagopalan, B.; Lall, U. Lall Estimation of Mutual Information Using Kernel Density Estimators. *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.* **1995**, *52*, 2318–2321.
43. Benesty, J.; Chen, J.; Huang, Y.; Cohen, I. Pearson Correlation Coefficient. In *Noise Reduction in Speech Processing*, 1st ed.; Springer: Berlin/Heidelberg, Germany, 2009; pp. 1–4. [CrossRef]
44. van Es, H.M.; Karlen, D.L. Reanalysis Validates Soil Health Indicator Sensitivity and Correlation with Long-Term Crop Yields. *Soil Sci. Soc. Am. J.* **2019**, *83*, 721–732. [CrossRef]

45. Mohamed, E.S.; Ali, A.; El-Shirbeny, M.; Abutaleb, K.; Shaddad, S.M. Mapping Soil Moisture and Their Correlation with Crop Pattern Using Remotely Sensed Data in Arid Region. *Egypt. J. Remote Sens. Space Sci.* **2020**, *23*, 347–353. [CrossRef]
46. Mohamed, E.; Belal, A.-A.; Ali, R.; Saleh, A.; Hendawy, E.A. Land Degradation. In *The Soils of Egypt*; El-Ramady, H., Alshaal, T., Bakr, N., Elbana, T., Mohamed, E., Belal, A.-A., Eds.; Springer: Cham, Switzerland, 2019; pp. 159–174. [CrossRef]
47. El-Zeiny, A.; El-Kafrawy, S. Assessment of Water Pollution Induced by Human Activities in Burullus Lake Using Landsat 8 Operational Land Imager and GIS. *Egypt. J. Remote Sens. Space Sci.* **2017**, *20*, S49–S56. [CrossRef]
48. El-Zeiny, A.M.; Effat, H.A. Environmental Monitoring of Spatiotemporal Change in Land Use/Land Cover and Its Impact on Land Surface Temperature in El-Fayoum Governorate, Egypt. *Remote Sens. Appl. Soc. Environ.* **2017**, *8*, 266–277. [CrossRef]
49. Petropoulos, G.; Carlson, T.N.; Wooster, M.J.; Islam, S. A Review of Ts/VI Remote Sensing Based Methods for the Retrieval of Land Surface Energy Fluxes and Soil Surface Moisture. *Prog. Phys. Geogr. Earth Environ.* **2009**, *33*, 224–250. [CrossRef]
50. Otto, F.E.L.; Massey, N.; van Oldenborgh, G.J.; Jones, R.G.; Allen, M.R. Reconciling Two Approaches to Attribution of the 2010 Russian Heat Wave. *Geophys. Res. Lett.* **2012**, *39*, 1–5. [CrossRef]
51. Tack, J.; Barkley, A.; Nalley, L.L. Warming Effects on US Wheat Yields. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 6931–6936. [CrossRef] [PubMed]
52. Brisson, N.; Gate, P.; Gouache, D.; Charmet, G.; Oury, F.-X.; Huard, F. Why Are Wheat Yields Stagnating in Europe? A Comprehensive Data Analysis for France. *Field Crops Res.* **2010**, *119*, 201–212. [CrossRef]
53. Schauberger, B.; Archontoulis, S.; Arneth, A.; Balkovic, J.; Ciais, P.; Deryng, D.; Elliott, J.; Folberth, C.; Khabarov, N.; Müller, C.; et al. Consistent Negative Response of US Crops to High Temperatures in Observations and Crop Models. *Nat. Commun.* **2017**, *8*, 13931. [CrossRef] [PubMed]
54. Troy, T.J.; Kipgen, C.; Pal, I. The Impact of Climate Extremes and Irrigation on US Crop Yields. *Environ. Res. Lett.* **2015**, *10*, 054013. [CrossRef]
55. Lesk, C.; Rowhani, P.; Ramankutty, N. Influence of Extreme Weather Disasters on Global Crop Production. *Nature* **2016**, *529*, 84–87. [CrossRef]
56. Lobell, D.B.; Bänziger, M.; Magorokosho, C.; Vivek, B. Nonlinear Heat Effects on African Maize as Evidenced by Historical Yield Trials. *Nat. Clim. Change* **2011**, *1*, 42–45. [CrossRef]
57. Ray, D.K.; Ramankutty, N.; Mueller, N.D.; West, P.C.; Foley, J.A. Recent Patterns of Crop Yield Growth and Stagnation. *Nat. Commun.* **2012**, *3*, 1293. [CrossRef]
58. Schluchter, M.D. Mean Square Error. In *Wiley StatsRef: Statistics Reference Online*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2014; ISBN 978-1-118-44511-2.
59. Chai, T.; Draxler, R.R. Root Mean Square Error (RMSE) or Mean Absolute Error (MAE). *Geosci. Model Dev. Discuss.* **2014**, *7*, 1525–1534.
60. Kashaninejad, M.; Dehghani, A.A.; Kashiri, M. Modeling of Wheat Soaking Using Two Artificial Neural Networks (MLP and RBF). *J. Food Eng.* **2009**, *91*, 602–607. [CrossRef]
61. Gago, J.; Martínez-Núñez, L.; Landín, M.; Gallego, P.P. Artificial Neural Networks as an Alternative to the Traditional Statistical Methodology in Plant Research. *J. Plant Physiol.* **2010**, *167*, 23–27. [CrossRef]
62. Wang, X.; Huang, J.; Feng, Q.; Yin, D. Winter Wheat Yield Prediction at County Level and Uncertainty Analysis in Main Wheat-Producing Regions of China with Deep Learning Approaches. *Remote Sens.* **2020**, *12*, 1744. [CrossRef]
63. Wolanin, A.; Mateo-García, G.; Camps-Valls, G.; Gómez-Chova, L.; Meroni, M.; Duveiller, G.; Liangzhi, Y.; Guanter, L. Estimating and Understanding Crop Yields with Explainable Deep Learning in the Indian Wheat Belt. *Environ. Res. Lett.* **2020**, *15*, 024019. [CrossRef]
64. Khaki, S.; Wang, L.; Archontoulis, S.V. A CNN-RNN Framework for Crop Yield Prediction. *Front. Plant Sci.* **2020**, *10*, 1750. [CrossRef] [PubMed]
65. Mokarram, M.; Bijanzadeh, E. Prediction of Biological and Grain Yield of Barley Using Multiple Regression and Artificial Neural Network Models. *Aust. J. Crop Sci.* **2016**, *10*, 895–903. [CrossRef]
66. Golhani, K.; Balasundram, S.K.; Vadamalai, G.; Pradhan, B. A Review of Neural Networks in Plant Disease Detection Using Hyperspectral Data. *Inf. Process. Agric.* **2018**, *5*, 354–371. [CrossRef]
67. Niazian, M.; Sadat-noori, S.A.; Abdipour, M. Modeling the Seed Yield of Ajowan (*Trachyspermum ammi* L.) Using Artificial Neural Network and Multiple Linear Regression Models. *Ind. Crops Prod.* **2018**, *117*, 224–234. [CrossRef]
68. Abdipour, M.; Younessi-Hmazekhanlu, M.; Ramazani, S.H.R. Artificial Neural Networks and Multiple Linear Regression as Potential Methods for Modeling Seed Yield of Safflower (*Carthamus tinctorius* L.). *Ind. Crops Prod.* **2019**, *127*, 185–194. [CrossRef]
69. Niedbała, G. Simple Model Based on Artificial Neural Network for Early Prediction and Simulation Winter Rapeseed Yield. *J. Integr. Agric.* **2019**, *18*, 54–61. [CrossRef]
70. Hackeling, G. Apply Effective Learning Algorithms to Real-World Problems Using Scikit-Learn. In *Mastering Machine Learning with Scikit-Learn*, 2nd ed.; Packt Publishing: Birmingham, UK, 2017; ISBN 1-78829-987-6.
71. McDonald, G.C. Ridge Regression. *WIREs Comput. Stat.* **2009**, *1*, 93–100. [CrossRef]
72. Johansson, R. An Intuitive Explanation of Gradient Boosting. Available online: https://www.cse.chalmers.se/~richajo/dit866/files/gb_explainer.pdf (accessed on 28 May 2023).

73. Sheremet, O.; Sadovoy, O. Using the Support Vector Regression Method for Telecommunication Networks Monitoring. In Proceedings of the 2016 Third International Scientific-Practical Conference Problems of Infocommunications Science and Technology (PIC S&T), Kharkiv, Ukraine, 4–6 October 2016; pp. 8–10.
74. Parisi, L. M-Arcsinh: An Efficient and Reliable Function for SVM and MLP in Scikit-Learn. *arXiv* **2020**, arXiv:2009.07530.
75. Nunno, L. Stock Market Price Prediction Using Linear and Polynomial Regression Models. 2014. Available online: http://www.lucasnunno.com/assets/docs/ml_paper.pdf (accessed on 28 May 2023).
76. Ahmad, M.W.; Reynolds, J.; Rezgui, Y. Predictive Modelling for Solar Thermal Energy Systems: A Comparison of Support Vector Regression, Random Forest, Extra Trees and Regression Trees. *J. Clean. Prod.* **2018**, *203*, 810–821. [CrossRef]
77. Mishra, A.M.; Harnal, S.; Gautam, V.; Tiwari, R.; Upadhyay, S. Weed Density Estimation in Soya Bean Crop Using Deep Convolutional Neural Networks in Smart Agriculture. *J. Plant Dis. Prot.* **2022**, *129*, 593–604. [CrossRef]
78. Kaur, P.; Harnal, S.; Tiwari, R.; Upadhyay, S.; Bhatia, S.; Mashat, A.; Alabdali, A.M. Recognition of Leaf Disease Using Hybrid Convolutional Neural Network by Applying Feature Reduction. *Sensors* **2022**, *22*, 575. [CrossRef]
79. Guan, S.; Fukami, K.; Matsunaka, H.; Okami, M.; Tanaka, R.; Nakano, H.; Sakai, T.; Nakano, K.; Ohdan, H.; Takahashi, K. Assessing Correlation of High-Resolution NDVI with Fertilizer Application Level and Yield of Rice and Wheat Crops Using Small UAVs. *Remote Sens.* **2019**, *11*, 112. [CrossRef]