agriculture

MDPI

Article

Prediction of Potato (*Solanum tuberosum* L.) Yield Based on Machine Learning Methods

Jarosław Kurek [1,*], Gniewko Niedbała [2], Tomasz Wojciechowski [2], Bartosz Świderski [1], Izabella Antoniuk [1], Magdalena Piekutowska [3], Michał Kruk [1] and Krzysztof Bobran [4]

[1] Department of Artificial Intelligence, Institute of Information Technology, Warsaw University of Life Sciences, Nowoursynowska 159, 02-776 Warsaw, Poland; bartosz_swiderski@sggw.edu.pl (B.Ś.); izabella_antoniuk@sggw.edu.pl (I.A.); michal_kruk@sggw.edu.pl (M.K.)
[2] Department of Biosystems Engineering, Faculty of Environmental and Mechanical Engineering, Poznań University of Life Sciences, Wojska Polskiego 50, 60-627 Poznań, Poland; gniewko.niedbala@up.poznan.pl (G.N.); tomasz.wojciechowski@up.poznan.pl (T.W.)
[3] Department of Botany and Nature Protection, Institute of Biology, Pomeranian University in Słupsk, 22b Arciszewskiego St., 76-200 Słupsk, Poland; magdalena.piekutowska@upsl.edu.pl
[4] Seth Software sp. z o.o., Strefowa 1, 36-060 Głogów Małopolski, Poland; kbobran@seth.software
* Correspondence: jaroslaw_kurek@sggw.edu.pl

Abstract: This research delves into the application of machine learning methods for predicting the yield of potato varieties used for French fries in Poland. By integrating a comprehensive dataset comprising agronomical, climatic, soil, and satellite-based vegetation data from 36 commercial potato fields over five growing seasons (2018–2022), we developed three distinct models: non-satellite, satellite, and hybrid. The non-satellite model, relying on 85 features, excludes vegetation indices, whereas the satellite model includes these indices within its 128 features. The hybrid model, combining all available features, encompasses a total of 165 features, presenting the most-comprehensive approach. Our findings revealed that the hybrid model, particularly when enhanced with SVM outlier detection, exhibited superior performance with the lowest Mean Absolute Percentage Error (MAPE) of 5.85%, underscoring the effectiveness of integrating diverse data sources into agricultural yield prediction. In contrast, the non-satellite and satellite models displayed higher MAPE values, indicating less accuracy compared to the hybrid model. Advanced data-processing techniques such as PCA and outlier detection methods (LOF and One-Class SVM) played a pivotal role in model performance, optimising feature selection and dataset refinement. The study concluded that machine learning methods, particularly when leveraging a multifaceted approach involving a wide array of data sources and advanced processing techniques, can significantly enhance the accuracy of agricultural yield predictions. These insights pave the way for more-efficient and -informed agricultural practices, emphasising the potential of machine learning in revolutionising yield prediction and crop management.

Keywords: machine learning; yield prediction; potato

Citation: Kurek, J.; Niedbała, G.; Wojciechowski, T.; Świderski, B.; Antoniuk, I.; Piekutowska, M.; Kruk, M.; Bobran, K. Prediction of Potato (*Solanum tuberosum* L.) Yield Based on Machine Learning Methods. *Agriculture* **2023**, *13*, 2259. https://doi.org/10.3390/agriculture13122259

Academic Editor: Francesco Marinello

Received: 10 November 2023
Revised: 3 December 2023
Accepted: 7 December 2023
Published: 11 December 2023

1. Introduction

The potato (*Solanum tuberosum* L.) is one of the basic species of cultivated plants in the world. According to FAOSTAT data, the world potato production in 2021 reached 359 million t, and the pioneers in the cultivation of this species were China, India, and Ukraine [1]. In 2021, the highest tuber yield per hectare was recorded in the United States, approximately 51 t, New Zealand, 50.7 t, and Kuwait, 48.7 t. In Europe in 2021, the countries with the highest yield of potato tubers were: France, the Netherlands, Belgium, and Germany. In these countries, the discussed species yields were in the range of 40 to 45 t/ha [2].

In recent years, a change in culinary preferences of consumers towards potatoes has been observed, translating directly to trends in the cultivation of this plant. Currently, the dynamic growth of the market of "convenience" food, as well as fried products can be noted in developing and developed countries around the world [3]. The intensive development of potato processing contributed to the increase in the demand for fast food. Current trends in agricultural development strongly support precision agriculture. The idea behind this is centred on the keywords: observation, measurement, and constant response to inter- and intra-field crop variability [4]. In the case of industrial potato production, precision agriculture—constant monitoring of crop condition—is needed because of the desired good quality of the final raw material while maintaining a high level of commercial yield. Despite the many tools developed to monitor and analyse potato growth and yield over the last 20–30 years, with a peak in sensor development in the last 10 years, many fry potato farms are looking for superior solutions. Widely tested and refined predictive tools for estimating yield and quality prior to final harvest use advanced artificial intelligence methods, among others. It is believed that crop growth monitoring and yield mapping will become mainstream farm research and development in the coming years. Many input parameters specific to specialised crops have become more readily available. For example, remote sensing data on crop emergence date and aboveground biomass are being used to better set model parameters [5,6]. Hybrid forms of sensor systems and crop growth models will provide better information on crop growth during the season. This, combined with weighing systems and cameras on harvesters, will provide site-specific information on the yield and quality of harvested potatoes [7,8].

Research on the use of plant models in predicting potato yield at the field scale has been conducted for over half a century, but their intensity has increased since the beginning of the second decade of the 21st Century [9]. In these studies, it can be observed that classical plant growth models primarily utilise ground-based data, including commonly used factors such as nitrogen fertilisation levels, air temperature values, sunlight exposure, and precipitation levels. This applies to a wide range of models such as SUBSTOR Potato, CROPSYST-SIMPOTATO, and Potato Calculator [10–14]. It is noted that the limitations for the practical application of such models at the field production scale are data availability, the cost of data acquisition, and data quality issues. From another point of view, another source of input data for predictive models is satellite imagery [14]. With the development of satellite Earth observation systems, improved RS data availability for practical applications, and the increased quality of these data in terms of spatial and temporal resolution, satellite data have been increasingly incorporated into potato predictive models [15,16]. When it comes to potato yield forecasting at field scale, there are few publications describing the combined use of ground-based data (soil, agronomy, weather) and satellite data (vegetation indices) as the input parameters for models [12,17–21]. In the practical application of predictive models in agricultural decision support systems, the flexibility of data source selection for modelling becomes an important functional requirement, considering the aforementioned data availability and quality issues. Often, farmers do not have complete sets of ground-based or satellite data. Therefore, there is a need to evaluate and compare ground-based, satellite, and hybrid models that combine data from both types of sources.

Knowledge of yield determinants is important in the development of crop-management-improvement models [1] for both prediction and classification. In the correct construction of production models, it is useful to demonstrate good knowledge of the research object and to have knowledge of yield determinants and potential disturbances, changing the final modelling effect in an independent way [22,23]. The traits explaining varietal yield or supporting potato yield potential in forecasting models in the literature are classified according to the following categories [15,23,24]:

- Weather traits;
- Agricultural traits;
- Traits conditioned by genotype and phytophenological traits;
- Soil environment;

- Spectral data, including vegetation indices;
- Indicators related to plant productivity.

Data on weather conditions during the growing season can be treated as a disturbance over which both producers and predictive model developers have no direct influence. The most-advantageous solution from the point of view of preparing climatic model data is the selection of years in which the weather conditions represent the optimal case for the place of cultivation and observation. The dominant meteorological features in the prediction models are: total precipitation, average, minimum, and maximum daily air temperatures, insolation, relative air humidity, evapotranspiration, etc. [23,25–27].

Agrotechnical features are nothing more than variants of potato cultivation, i.e., the sum of mineral and organic fertilisation, the soil cultivation system, irrigation, plant protection, forecrops, etc. [23,28]. In the case of potato cultivation intended for processing, including French fry varieties, agrotechnical requirements and recommendations are usually prepared by companies purchasing the raw material. Such action guarantees an acceptable level of tuber yield and an even quality of the yield taken from different suppliers.

Features associated mainly with genotype and phytophenological features (meaning successive stages of development achieved by plants while growing in the field) are important for obtaining a satisfactory raw material in terms of quality [29]. Favourable soil conditions are very crucial in potato cultivation. It is known that cultivated plants are "more sensitive" to the abundance of available nutrients in the soil than to the ongoing fertilisation [30,31]. It is also important to maintain the recommended pH and looseness of the soil. Potato cultivation in the desired conditions reduces the occurrence of soil diseases [32], as well as prevents bruises [33]. Features related to potato productivity are usually various indicators, the interpretation of which allows for ongoing analysis of growth, yield, and photosynthetic activity.

In yield-forecasting models, the following are most-often used: Photosynthetically Active Radiation (PAR) and Leaf Area Index (LAI) [34,35]. Measuring these indicators is relatively easy, and the final data are not difficult to interpret. Information obtained using remote sensing and GIS methods is becoming increasingly important in the management of potato cultivation and, thus, in the creation of reliable predictive models. Vegetation indices calculated using these methods are a kind of quantitative measure that is correlated with the amount of biomass or the condition of the vegetation. They are usually formulated as a combination of two or three spectral channels (with red and near-infrared being the most-common). Their values are added, divided, or multiplied in order to obtain one value (index), which tells about the amount and condition of the vegetation [24,36]. A wide application in tuber yield-forecasting models has been confirmed for the Normalised Difference Vegetation Index (NDVI) [37], Normalised Difference Red Edge index (NDRE) [38], Potato Productivity Index (PPI) [16], SAVI, RDVI, and EVI.

In recent years, a departure from the use of classic models for predicting the yield of potato tubers, such as SUBSTOR, POTATO, Lintul-POTATO, etc., can be observed. Classical regression models also do not fully fulfil their role, because the forecast errors generated by such models are very high and, therefore, unacceptable in agriculture [23]. The trend of scientific development in yield modelling runs in two directions. One of them is the improvement by researchers of classic potato models—adapting them to specific climatic or cultivation conditions [10]. The second approach involves modern and reliable modelling techniques—Artificial Neural Networks (ANNs) [23,39], decision trees [37], and deep learning [15,40]. It is worth emphasising that the most-important feature of neural models is their ability to generalise the knowledge they acquire during a specific network learning process. Designing the proper structure of a neural network and determining its parameters requires the use of advanced optimisation algorithms. Solving a specific problem always involves the choice of the type of network. Forecasting issues are usually implemented using MLP models [23,24,41–45]. Neural modelling plays a significant role when solving practical problems requiring a quick response is expected [46]. In addition, analysis carried

out using nonlinear models, which include neural models, are characterised by a smaller forecast error compared to classical methods [47].

Predicting crop yields is an important task for agriculture, and predictive models can be useful in this process. However, there are some limitations and potential sources of error that are worth considering. Here are some of them [48–52]:

- Dependence on historical data.
- Disregarding nonlinear factors.
- Variability of environmental conditions.
- Errors in measurements and other inputs.
- No consideration of changes in agricultural practices.
- Complexity of the interaction between factors.

It is important to understand these limitations and potential sources of error in crop yield prediction. Models can be useful tools, but they should be used carefully, take into account a variety of factors, and evolve with advances in knowledge and data availability.

The aim of this article was to create three models predicting the yield of French fry potatoes grown in Poland using machine learning methods. The research focused on several important scientific aspects, including a thorough analysis of empirical data, which allowed the creation of predictive data aggregates. By subjecting the partial classification results to detailed interpretations, it was possible to reject data introducing distortions and noise in the prediction. Finally, based on the MAPE values, the most-accurate model for predicting tuber yield was indicated.

2. Materials and Methods

2.1. Dataset Description

The data used in this work came from 114 commercial potato fields located in northern Poland. Fields with potato cultivation varied in area from 6.5 to 156 ha. The cultivated potato varieties were Innovator, Ludmilla, Ivory Russet and Zorba. The data covered five growing seasons in the years 2018–2022, containing several types of information, i.e., agronomical data, climatic data, satellite-based vegetation, satellite data, and soil data. Source data were obtained from databases of different natures: public databases as open data, private databases of farmers, and ERP databases of agricultural producers. The field locations are presented in Figure 1. The structure of the potato dataset is shown in Table 1.

The data were divided into two sets, referred to as ground-based data (agronomic and weather data) and satellite-based data. These two sets constituted the sets for the non-integrated terrestrial and non-integrated satellite predictive models, respectively. Both sets were the basis for the creation of the hybrid models.

2.1.1. Data Augmentation

The data augmentation process plays a crucial role in enhancing the performance of machine learning models, especially when dealing with limited datasets. In the context of predicting potato yield, data augmentation involves creating synthetic but realistic data points based on the existing dataset. The augmentation procedure can be broken down into several key steps and has been described below.

Augmentation loop: For each record in the dataset, the algorithm performs the following steps multiple times (5 times), as determined by the number of copies specified:

1. A random change percentage (between 0.01 and 0.05) is chosen within a predefined range, which determines the degree of modification for the augmentation.
2. Noise is generated based on the random change percentage and is applied to both the features and the target variable. This noise addition simulates realistic variability within the data.
3. The new synthetic record, created by applying noise, is then denormalized to bring it back to the original data scale.
4. The synthetic record is appended to the augmented dataset along with its corresponding textual data.

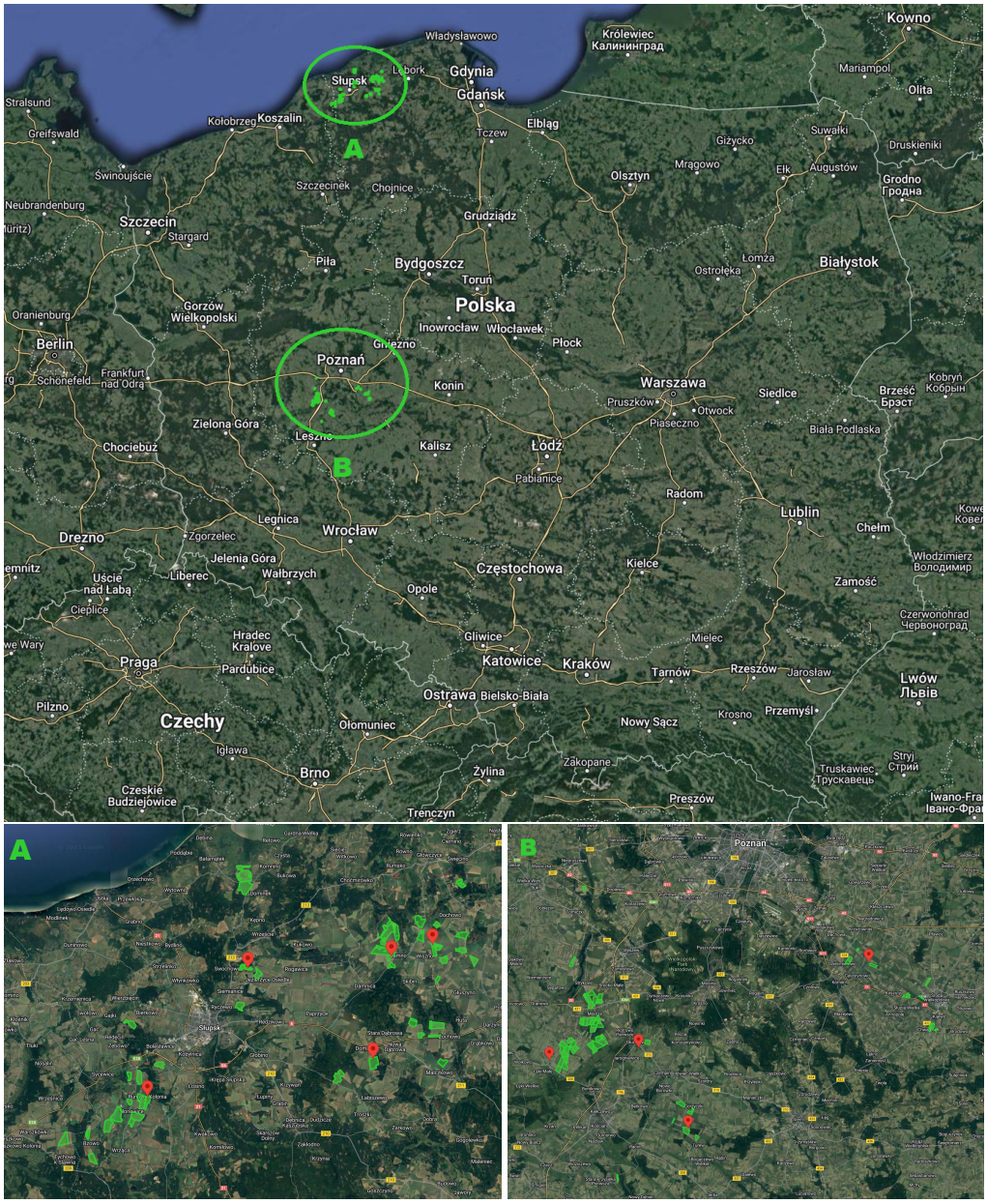

Figure 1. Field locations for the data-collection process: map of Poland, with marked general field placement areas (**Top**), and a close-up map for the area of Słupsk (**A**) and Poznań (**B**).

Table 1. Number of fields for the presented experiments, containing different potato varieties during consecutive years.

Variety	2018	2019	2020	2021	2022	Total
Innovator	60	65	40	55	20	240
Ludmilla	20	30	30	20	15	115
Ivory Russet	5	10	0	0	0	15
Zorba	5	20	15	10	0	50
Total	90	115	95	85	35	420

2.1.2. BBCH-Scale

Part of the data was allocated to the growth stages of the crops cultivated. A universally available BBCH-scale was used in that aspect. The abbreviation BBCH derives from the names of the original participating stakeholders: "Biologische Bundesanstalt,

Bundessortenamt und Chemische Industrie". The BBCH-scale is used to identify the phenological development stages of plants. It was developed for a range of crop species, where similar growth stages of each plant are given the same code. The phenological development stages of plants are used in a number of scientific disciplines (crop physiology, phytopathology, entomology, and plant breeding) and in the agriculture industry (risk assessment of pesticides, timing of pesticide application, fertilisation, and agricultural insurance).

The BBCH-scale uses a decimal code system, which is divided into principal and secondary growth stages and is based on the cereal code system (the Zadoks scale) developed by Jan Zadoks. The phenological development stages obtained from the producer were in following ranges [53,54]:

(1) BBCH 0–10 (from planting to the beginning of emergence);
(2) BBCH 11–50 (from the beginning of emergence to the beginning of tuber setting);
(3) BBCH > 50 (from the beginning of tuber setting to harvest).

Based on the imported data, the BBCH phase limits in the ranges: (1), (2), and (3) were assigned and later used to calculate the aggregated data.

2.1.3. Agronomic Data

The agronomic data were obtained from the Plantator System [55] by Seth Software, as well as from the Plantator System operating during production in regard to: crop register, harvest registration, and registration of hourly work results. The acquired data have different formats, so in some cases, it was necessary to obtain a preprocessed dataset, e.g., from soil test results (pdf) or the locations of crops (jpg). The agronomic data also came from private grower databases. Information on the use of irrigation treatment in the irrigated/non-irrigated structure was also an explanatory feature.

2.1.4. Climate Data

Information on the weather data was gathered from agrometeorological stations situated near the cultivated areas. In cases where there were no local stations, relevant data were acquired from public databases, specifically from weather stations managed by the Institute of Meteorology and Water Management-National Research Institute (IMGW).

2.1.5. Soil Data

Depending on the season, the data regarding soil nutrient content came from different soil testing laboratories. Similarly, the data regarding liquid and solid mineral fertilisation were obtained, resulting in the variability of the analysed components and different units of measurement. All laboratories were nationally accredited for the soil parameters analysed. The soil parameters considered were: the pH and the phosphorus, potassium, and magnesium content. These are the range of parameters most often contracted for soil analyses by agricultural producers in Poland.

2.1.6. Satellite Data

The crop vegetation data were obtained through satellite remote sensing. The primary image database utilised in this study was the European Copernicus Sentinel 2 mission's image database. The Google Earth Engine (GE) platform served as the direct data collection ("COPERNICUS S2 SR"). A Python script was developed by the authors to acquire, filter, and process images and calculate the Vegetation Indices (VIs). The script was executed on a local server, which communicated with the GEE service.

The secondary image database utilised in this study was the PlanetScope images (Planet Labs), geometrically and atmospherically corrected. The images, clipped to the analysed ROI, were downloaded via Planet's dedicated Data API.

In the initial step, the images for each of the potato fields were filtered based on cloud cover (threshold ranging from 7% to 13% depending on the year) using the QA60 band, for Sentinel and from the "cloud_percent" metadata for PlanetScope. The threshold depended on the availability of images for a given ROI. If there were not at least 3 images for

the ROI in the time period analysed, the threshold was automatically increased. The images and corresponding reflectance values were analysed for the period between April 1 and the end of September for each year under study. The VIs were then calculated for the acquisition date using the obtained reflectance values.

The following vegetation indices were applied in this study: Enhanced Vegetation Index (EVI), Normalised Difference Vegetation Index (NDVI), Renormalised Difference Vegetation Index (RDVI), and Soil-Adjusted Vegetation Index (SAVI) [56]. These VIs are widely used in the literature for predicting potato yield and were calculated according to the Index DataBase. Finally, a total of 16 vegetation features were calculated for the 4 VIs, including the minimum, mean, maximum, and standard deviation groups.

2.1.7. Selyaninov Hydrothermal Coefficient

The investigation of climate variation is a subject of keen interest among scientists in various fields, such as hydrology, meteorology, agriculture, and forestry. All of them aim to determine the most-accurate climatic conditions that will prevail in a specific region in the future. Despite having greater computing power, the analysis of increasingly complex models reveals that numerous environmental factors still need to be considered, rendering the issue unresolved.

Central Europe's different climate change scenarios suggest that an increase in temperature will be accompanied by a slight rise in annual precipitation, which will be redistributed throughout the year. Winter precipitation is projected to increase while summer rainfall to decrease. Given the limited retention capacity and a concomitant increase in evaporation, the amount of water available to plants will be reduced during the growing season, and there may be a depletion of reserves from the winter season. Moreover, the growing variance of precipitation and temperatures should not be overlooked as it indicates that unfavourable extreme situations for plant production are likely to occur more frequently.

One aspect that requires close monitoring is the evaluation of water availability in a particular area, particularly in extreme cases such as floods and droughts. Different indicators are used to measure the severity of water scarcity, one of which is the Selyaninov Hydrothermal Coefficient (HTC). This coefficient assesses drought based on the formula [57]:

$$\mathrm{HTC} = 10 \sum n_i \mathbf{P}_i \sum n_i \mathbf{t}_i \quad (1)$$

where:
n—the length of the period considered in days;
$\mathbf{P}_i$—the rainfall on the i-th day (mm);
$\mathbf{t}_i$—the average daily temperature on the i-th day (°C).

Based on the above properties, three aggregated parameters for three vegetation stages (BBCH-based) were generated to be used as additional prediction features. The rainfall and temperature values used to calculate the HTC were taken from the IMGW net and our own agrometeorological stations, as described in Section 2.1.4.

2.1.8. GDD Features

When plants are not subjected to extreme conditions such as abnormal drought or disease, they usually grow incrementally, and the prevailing temperature heavily influences their growth rate. The Growing Degree Days (GDDs) [58] parameter considers various aspects of local weather, enabling farmers to anticipate and even regulate the pace at which their plants mature, particularly in greenhouse settings.

Provided the plants are not affected by other environmental factors such as soil moisture, their developmental rate from emergence to maturity hinges on the daily air temperature. Specific developmental phases of plants and insects depend on the accumulation of specific quantities of heat, allowing the prediction of when these events should occur during a growing season, regardless of temperature differences across years. The GDDs are defined as the number of degrees above the base temperature, which varies depending on the crop species. The base temperature is the temperature at which plant growth is zero.

To calculate the GDDs, each day's maximum and minimum temperatures are added and divided by two, and the base temperature is then subtracted. The GDDs are accumulated by adding each day's GDD contribution as the season progresses. GDDs can be used for various purposes, including:

- Assessing a region's suitability for cultivating specific crops;
- Estimating the growth stages of crops, weeds, or insects;
- Predicting the maturity and cutting dates of forage crops;
- Determining the optimal timing of fertiliser or pesticide application;
- Estimating heat stress on crops;
- Planning the spacing of planting dates to produce separate harvest dates.

These parameters can be calculated as shown in Equation (2):

$$\mathrm{GDD} = \sum_{i=1} n_i \mathbf{T}_{avg} \tag{2}$$

where:
GDD—the Growing Degree Day (°C);
n—the length of the period considered in days;
$\mathbf{T}_{avg}$—the average daily air temperature ≥ 0 (°C).

Similar to the HTC, aggregated parameters for the three vegetation stages (BBCH-based) were generated based on this parameter.

2.1.9. Total Numerical features

After the initial analysis, a set of features was derived for the presented experiments. Apart from the basic crop data (season, variety, acreage, location of cultivation, age of cultivation, yield), we used the BBCH-scale (see Section 2.1.2).

A total of 250 potential explanatory features were derived for the target. The target variable is defined as the total yield of the harvested crop (harvest) (potato). All numerical data (both explanatory and dependent features) were aggregated to full years (2018, 2019, 2020, 2021, and 2022). The target variable was measured in tons (t). A summary of all the numerical feature groups used in the prediction of potato yield before data pruning is presented in Table 2.

Table 2. Summary of all the numerical feature groups used in the prediction of potato yield before data pruning.

Group of Features	No. of Features
Aggregated weather features	7
Weather features	92
Soil features	17
Agrotechnical treatment features	6
Vegetation indexes GE	64
Vegetation indexes PL	64
Total	250

2.1.10. Data Pruning: Addressing Missing Values

In this study, a common challenge of dealing with missing values in the dataset was encountered. Any data analysis, irrespective of the statistical methods applied, is only as robust as the quality and completeness of the data being analysed. In this case, the dataset initially comprised 250 features collected for predicting potato yield.

The initial step was to identify the extent of missing data in the dataset. This process of quantification was carried out by calculating the percentage of missing values in each variable. It is crucial to note that the quantity of missing values can considerably influence the performance and accuracy of AI models.

In order to maintain the integrity of our study, we established a cutoff threshold of 50%. Any variable with more than 50% missing values was deemed unreliable for our analysis due to the massive information gap. The rationale behind this decision was that imputing more than 50% of the data of a variable can introduce a substantial amount of bias and distortion in the prediction model. It also raises concerns about the reliability and validity of the subsequent findings, as more than half of the information would be synthetic or based on estimates. This decision was rooted in a balance between retaining valuable data and ensuring the reliability and robustness of our models. The rationale for choosing this specific threshold was multi-faceted:

- Data integrity: When more than half of the data for a variable are missing, the integrity and representativeness of that variable become questionable. With over 50% missing data, any form of imputation would largely be based on speculation, rather than trends or patterns inherent in the data.
- Statistical significance: Variables with significant missing data can potentially skew the results and lead to unreliable conclusions. By setting the threshold at 50%, we aimed to maintain variables that had a statistically significant amount of data, thereby ensuring that our models were built on solid and representative foundations.
- Balance between data retention and quality: The 50% threshold strikes a balance between retaining as much data as possible and ensuring the quality of the dataset. This threshold allowed us to keep a substantial portion of the dataset while avoiding the pitfalls of basing our analysis on largely imputed or speculative data.
- Benchmarking against standard practices: This threshold is in line with common practices in data science and statistical analysis, where a 50% cutoff is often used as a standard for determining the viability of a variable in a dataset.

By implementing this threshold, we aimed to enhance the robustness and reliability of our predictive models. This approach allowed us to use a dataset that was both comprehensive and credible, leading to more-accurate and -trustworthy outcomes in our study.

After a rigorous examination, it was confirmed that 85 out of 250 features had missing data exceeding the 50% threshold. Therefore, to ensure the reliability of the succeeding analysis, as well as to maintain the robustness of the model, it was decided to exclude those features from the dataset.

Thus, the pruned dataset contained only 86 features, ready for further analysis and AI model training. This data-reduction method helped to maintain the data quality while ensuring that the future predictive model would not suffer from the adverse impacts of missing values and imputation bias.

Moving forward, these 165 features will be used to develop our artificial-intelligence-based prediction models. The retained features were carefully selected from the dataset after excluding those with excessive missing data. The list of the remaining 165 features includes the following:

Taking this strategic approach to data management was meant to ensure the most-accurate and -meaningful results from the AI models in the prediction of potato yield. The final list of numerical feature used in potato yield prediction is presented in Table 3, with a summary of the number of final numerical feature groups provided in Table 4.

Table 3. List of the number of numerical feature groups used in the prediction of potato yield after data pruning.

Variable Type	List of Variables
Agrotechnical treatment features (4 items)	Liquid fertilisation, spraying, planting, broadcast fertilisation
Weather features (23 items)	Average temperature (°C), rainfall (mm), air temperature1 (°C), air temperature2 (°C), air temperature3 (°C), solar panel (mV), precipitation (mm), wind speed AVG (m/s), wind speed Min (m/s), wind speed Max (m/s), battery (mV), leaf wetness time (min), HC serial number, HC air temperature AVG (°C), HC air temperature Max (°C), HC air temperature Max (°C), HC relative humidity AVG (%), HC relative humidity AVG (%), HC relative humidity AVG (%), Dev point temperature AVG (°C), Dev point temperature Max (°C), vapour pressure deficit AVG (mBar), vapour pressure deficit Min (mBar)
Aggregated weather features (6 items)	HTC 0–10, HTC 11–50, HTC > 50, GDD 0–10, GDD > 50, GDD 11–50
Soil features (4 items)	Soil pH H_2O, phosphorus (mg/100 g), potassium (mg/100 g), magnesium (mg/100 g)
Vegetation indices GE (calculated based on Sentinel via Google Earth) (64 items)	EVI_GE_0_10_Max, EVI_GE_11_50_Max, EVI_GE_50_Max, EVI_GE_daily_Max, NDVI_GE_0_10_Max, NDVI_GE_11_50_Max, NDVI_GE_50_Max, NDVI_GE_daily_Max, RDVI_GE_0_10_Max, RDVI_GE_11_50_Max, RDVI_GE_50_Max, RDVI_GE_daily_Max, SAVI_GE_0_10_Max, SAVI_GE_11_50_Max, SAVI_GE_50_Max, SAVI_GE_daily_Max, and so on, for mean, Min, StdDev variants
Vegetation indices PL (calculated based on PlanetScope via Planet Labs) (64 items)	EVI_PL_0_10_Max, EVI_PL_11_50_Max, EVI_PL_50_Max, EVI_PL_daily_Max, NDVI_PL_0_10_Max, NDVI_PL_11_50_Max, NDVI_PL_50_Max, NDVI_PL_daily_Max, RDVI_PL_0_10_Max, RDVI_PL_11_50_Max, RDVI_PL_50_Max, RDVI_PL_daily_Max, SAVI_PL_0_10_Max, SAVI_PL_11_50_Max, SAVI_PL_50_Max, SAVI_PL_daily_Max, and so on, for mean, Min, StdDev variants

Table 4. Summary of the number of final numerical feature groups used in prediction of potato yield after data pruning.

Group of Features	No. of Features
Aggregated weather features	4
Weather features	23
Soil features	4
Agrotechnical treatment features	6
Vegetation indexes GE	64
Vegetation indexes PL	64
Total	165

2.2. Data Imputation

Data imputation, or the process of filling in missing data points in datasets, is a critical aspect of predictive modelling [59], particularly in the field of agricultural yield predictions. The robustness and accuracy of Artificial Intelligence (AI) models depend highly on the quality and completeness of the underlying datasets. In the case of predicting potato yield, incomplete datasets can lead to inaccurate models and predictions, thus impeding the optimisation of crop production.

In the context of AI, missing data could induce significant bias, reduce the statistical power, and ultimately distort the representation of the real-world scenario that the AI model is attempting to capture. This issue is particularly pertinent in agricultural datasets, where factors such as weather conditions, soil properties, and crop health measures can be highly variable and sometimes difficult to measure consistently. Without adequate data in these areas, AI models may not accurately reflect the complex interactions and dependencies among these factors, leading to erroneous predictions of potato yield.

Methods of Data Imputation

Data imputation is a critical step in the preprocessing phase of predictive modelling, especially when dealing with incomplete datasets. In the context of this research, we implemented a hybrid approach that combines regression and mean/median imputation strategies. This method intends to balance the bias introduced by mean/median imputation with the variance captured through regression techniques.

The proposed hybrid procedure, outlined in Algorithm 1, iteratively applies polynomial interpolation to create multiple imputations of the missing data, followed by median aggregation to ensure robustness. This method is particularly suitable for datasets with nonlinear relationships among the variables, such as the one used for predicting potato yield in this study. By applying a polynomial approach, we aimed to capture the intricate patterns inherent in the data, thereby enhancing the accuracy of our imputations. The decision to use this technique was based on preliminary analysis indicating significant nonlinear interactions among the predictive features.

Algorithm 1 Hybrid imputation procedure.

```
procedure HybridImputation(DataFrame, ColumnName)
    ProcessedColumn ← DeepCopy(DataFrame[ColumnName])
    ProcessedColumn ← AddIndexColumn(ProcessedColumn)
    ImputationTargets ← [ColumnName]
    ThresholdValidValues ← 86
    IterationCount ← 0
    while CountNonMissing(ProcessedColumn[ImputationTargets]) < ThresholdValidValues do
        TempColumn ← InterpolateColumn(ProcessedColumn, IterationCount)
        ProcessedColumn ← MergeColumns(ProcessedColumn, TempColumn)
        IterationCount ← Increment(IterationCount)
    end while
    DataFrame[ColumnName] ← ComputeMedian(ProcessedColumn[ImputationTargets])
    return DataFrame
end procedure
```

2.3. Data Normalisation

Data normalisation is an essential preprocessing step while dealing with machine learning or artificial intelligence algorithms. It is performed to bring all features into the range of 0 to 1, maintaining the distribution and relationships of the original raw data. This normalisation process helps to scale down the values of different scale attributes to a standard scale, which, in turn, enhances the performance of the model by allowing it

to converge faster during training. Additionally, it mitigates the risk of the model being influenced disproportionately by different features.

The particular method of normalisation used in this study was the Min–Max normalisation. This method re-scales features to a fixed range, typically 0 to 1, or alternatively -1 to 1 if there are negative values. This transformation preserves the original distribution of the data while ensuring that the impact of outliers is minimised.

The Min–Max normalisation is defined by the following formula [60,61]:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (3)$$

where:

- X_{norm} is the normalised value;
- X is the original value;
- X_{min} is the minimum value in the feature column;
- X_{max} is the maximum value in the feature column.

Each data value in the dataset is replaced by its corresponding normalised value, leading to a new dataset where all feature columns are within the same range. By implementing the Min–Max normalisation method, it is ensured that the model is less biased and, hence, more accurate in predicting potato yield based on the given features.

2.4. Prototyping the 3 AI Models: Non-Satellite, Satellite, and Hybrid

The accurate prediction of agricultural yields such as potato can greatly benefit both the farmers and the supply chain stakeholders. Utilising Artificial Intelligence (AI) methods for these predictions can potentially provide robust and reliable estimations. In this research, we investigated the effectiveness of different AI regression algorithms in constructing the predictive models. The objective was to gain insights into the strengths and limitations of various methods and subsequently guide future applications of AI in agriculture.

Three different models were considered in this study, namely non-satellite, satellite, and hybrid models. These models differed in their variable selections, which influenced their representation of real-world conditions.

The non-satellite model uses 37 features, which do not include vegetation indices. In contrast, the satellite model uses 128 features consisting exclusively the vegetation indices. These two models served as a basis for comparison to evaluate the contribution of vegetation indices in improving prediction accuracy. The hybrid model takes a comprehensive approach by including all 86 features available, thus merging the characteristics of both non-satellite and satellite models.

In constructing these models, a range of regression algorithms was applied, including Linear Regression [62], Ridge [63], Lasso [64], Elastic Net [65], the XGBoost Regressor [66], the Random Forest Regressor [67], the MLPRegressor [68] with different hidden layer sizes, the SGDRegressor [69], and Support Vector Regression (SVR) [70] with different parameters. These algorithms were chosen due to their diverse underlying principles, which provides a broad perspective on the prediction problem.

In the following sections, we detail the construction of these models and discuss the potential implications of our findings. The models were constructed using the Python programming language with the aid of powerful libraries such as scikit-learn and XGBoost.

2.5. Data Partitioning: Training, Validation, and Testing

An essential aspect of building robust and generalizable AI models is data partitioning. This process involves dividing the available dataset into distinct subsets: the training, validation, and test sets.

The training set is utilised to train the model, which essentially involves the adjustment of the model's parameters based on the input–output pairs in the data. The validation set is used during model training to provide an unbiased evaluation of the model's performance.

It allows for the tuning of hyperparameters and helps in model selection. Importantly, the validation set serves as a checkpoint to prevent overfitting, which occurs when the model learns the training data too well and performs poorly on unseen data. Finally, the test set is a separate data subset that is only used once the model has been trained and validated. It offers an objective evaluation of the final model's performance, representing how well the model is likely to perform on unseen, real-world data.

In this study, due to the forecasting nature of the task for upcoming years, we split the data based on the years:

- df_train—data from the years 2018 and 2019;
- df_val—data from the year 2020;
- df_test—data from the year 2021.

The model was trained using "df_train", while its performance was monitored on "df_val". Lastly, "df_test" was set aside for the final evaluation of the model, providing a benchmark of its performance on unseen data that did not participate in training. As such, our main focus was on the results obtained on "df_test".

2.6. Feature Selection

In this study, feature selection was executed using multiple methods including stepwise regression (stepwisefit), the Pearson correlation, the Chi-squared (χ^2) test, and Principal Component Analysis (PCA) [71–74]. These techniques were designed to select the most-relevant features for the task of predicting future years, thus potentially improving model accuracy, computational efficiency, and model interpretability.

2.6.1. Stepwise Regression

The stepwise regression [75] method was applied first for feature selection. Stepwise regression is an iterative process of adding and removing predictor features based on their statistical significance in a regression model. The technique starts from an initial model and takes steps to modify it by adding or removing predictors. The statistical significance of a predictor is typically measured by the *p*-value of the F-statistic when testing the models with and without the predictor.

In general, the stepwise regression process can be described as follows:

1. Fit the initial model.
2. If any predictors not in the model have *p*-values less than the entry tolerance (e.g., 0.05), add the one with the smallest p-value and repeat this step. If not, proceed to the next step.
3. If any predictors in the model have *p*-values greater than the exit tolerance (e.g., 0.10), remove the one with the largest *p*-value, and go back to the previous step. If not, stop.

It should be noted that the stepwise regression method is heuristic and does not guarantee that the final model is globally optimal, meaning that it has the best possible fit to the data. A different initial model or a different sequence of steps could lead to a better fit. In this sense, stepwise models are locally optimal, but not necessarily globally.

In this study, the stepwise regression function, "stepwisefit", was tested using a range of penter and premove values. Specifically, 90 different pairs of penter and premove were used, from (0.01, 0.06) to (0.9, 0.95). The goal of this testing was to explore how different thresholds for adding and removing features would impact the feature sets that the stepwise regression selected. However, only unique feature sets were extracted, which means that there may not necessarily be 90 distinct feature sets as a result of this procedure. The exact thresholds tested in this study are listed in Table 5.

It should be noted that different penter and premove values can have significant impacts on the stepwise regression outcomes. Lower penter values mean that the bar for adding a feature to the model is set higher, as it needs to have a higher level of statistical significance to be included. Similarly, higher premove values mean that the bar for removing a feature from the model is set lower, as it can be excluded even if its

significance is still relatively high. Therefore, different combinations of the penter and premove values can lead to diverse sets of selected features, providing a broad exploration of possible models.

Table 5. penter and premove values for stepwise regression.

No.	Penter	Premove
1	0.01	0.06
2	0.02	0.07
3	0.03	0.08
4	0.04	0.09
...	...	...
87	0.87	0.92
88	0.88	0.93
89	0.89	0.94
90	0.9	0.95

2.6.2. Pearson Correlation

In addition to stepwise regression, the Pearson correlation method was also used for feature selection [71,76]. It measures the linear relationship between two features, ranging from -1 to 1, where 1 means a perfect positive linear relationship, -1 means a perfect negative linear relationship, and 0 means no linear relationship.

In this study, if the absolute value of the Pearson correlation between two features exceeded 0.95, one of the two correlated features was removed from the set of predictive features. This was performed to mitigate the issue of multicollinearity, which can affect the performance and interpretability of the model.

2.6.3. Chi-Squared Test

The Chi-squared test was also applied as a feature-selection method [71]. This statistical test measures the independence between categorical features. In the context of feature selection, the Chi-squared test can be used to select those features that are most likely to be independent of each other and dependent on the target variable.

In this study, if the p-value of the Chi-squared test was greater than 0.05, the corresponding feature was added to the set of predictive features. Otherwise, the feature was blocked and not included in the set of predictive features.

2.6.4. Principal Component Analysis

Lastly, Principal Component Analysis (PCA) was utilised as a feature-selection and dimensionality-reduction method. PCA transforms the original features into a new set of features, which are linear combinations of the original ones [72,77]. These new features (or principal components) are uncorrelated with each other.

In this study, PCA was performed for different numbers of principal components (3, 4, 5, 6, 7, 8, 9, 10). The goal was to assess whether generating artificial features through PCA would enhance the performance of the model. This was performed both for the full set of features and the features selected by stepwise regression.

The advantage of PCA lies in its ability to transform a high-dimensional dataset into a lower-dimensional one while retaining most of the important information. However, the interpretability of the model can be compromised because the new features (principal components) are artificial and are not directly interpretable in terms of the original features.

3. Outlier Detection

Outlier detection is an important step in data preprocessing. Outliers are unusual data points that deviate significantly from the rest of the data. While some outliers may be errors and, hence, require correction, others may carry important information about the

data. In the presented study, two methods for outlier detection were used: Local Outlier Factor (LOF) and One-Class SVM [78].

Both methods have their own strengths and are appropriate for different types of datasets. In general, the LOF is good at detecting outliers that are in low-density regions, while One-Class SVM is effective at identifying outliers that are far away from the majority of the data.

3.1. Local Outlier Factor

The LOF method measures the local density deviation of a given data point with respect to its neighbours [79]. It considers as outliers the samples that have a substantially lower density than their neighbours. The number of neighbours considered (parameter "n_neighbours") is typically set to be 20% of the total number of samples. The outline of the methods is presented in Algorithm 2.

Algorithm 2 Pseudocode for local outlier factor.

procedure LOF(X, *n_neighbours*)
  for $x \in X$ **do**
    Calculate the distance to the *n_neighbours* nearest neighbours of x
    Compute the reachability distance of x
    Compute the local reachability density of x
  end for
  for $x \in X$ **do**
    Compute the LOF of x as the average ratio of the local reachability densities of the neighbours of x to the local reachability density of x
  end for
  Return the LOF of each sample
end procedure

3.2. One-Class SVM

One-Class SVM [77] is a method associated with the SVM family, but it is suited for the problem of outlier detection. The class of interest is modelled with a tight sphere in the feature space characterising the normal behaviour, and those instances that fall outside this sphere are considered outliers. The parameters used in the experiments are kernel = "rbf", gamma = "0.1", and nu = 0.5. Algorithm 3 presents the general overview of this procedure.

Algorithm 3 Pseudocode for One-Class SVM.

procedure ONECLASSSVM(X, *nu*, *kernel*, *gamma*)
  Initialise One-Class SVM with parameters *nu*, *kernel*, and *gamma*
  Fit SVM to the data X
  Predict the labels (1 for inliers, -1 for outliers) for X
  Return the predicted labels
end procedure

4. Results and Discussion

In the presented experiments, a total of three models were prepared: the non-satellite, satellite, and hybrid one, where the first two take into account only subsets of feature, either excluding or including vegetation data, while the final model incorporates all potential features. Table 6 outlines the dataset organisation for different models, while Table 7 shows the parameter configurations used in each case.

Table 6. Dataset organisation for all prepared models.

Model (Number)	Training Set (Samples/Features)	Validation Set (Samples/Features)	Test Set (Samples/Features)
NSM Without Outlier Detection (1)	205/37	95/37	120/37
NSM With Outlier Detection Using Local Outlier Factor (2)	200/37	95/37	120/37
NSM With Outlier Detection Using One-Class SVM (3)	103/37	95/37	120/37
SM Without Outlier Detection (4)	205/128	95/128	120/128
SM With Outlier Detection Using Local Outlier Factor (5)	201/128	95/128	120/128
SM With Outlier Detection Using One-Class SVM (6)	104/128	95/128	120/128
HM Without Outlier Detection (7)	205/165	95/165	120/165
HM With Outlier Detection Using Local Outlier Factor (8)	200/165	95/165	120/165
HM With Outlier Detection Using One-Class SVM (9)	120/165	95/165	101/165

Table 7. Setup of parameters used for the prepared models. Model numbers refer directly to the method organisation presented in Table 6.

Model Number	Is_Stepwise Fit_Used	Penter | Premove	Is_Pearson _Used	Is_Chi2 _Used	Is_PCA	n_PCA_ Components
(1)	True	0.8 | 0.85	False	False	True	5
(2)	False	N/A | N/A	False	False	True	5
(3)	True	0.3 | 0.35	False	False	True	5
(4)	True	0.5 | 0.44	False	False	False	0
(5)	True	0.4 | 0.45	False	False	False	0
(6)	True	0.4 | 0.45	False	False	False	0
(7)	True	0.6 | 0.65	True	False	False	0
(8)	True	0.2 | 0.25	False	False	False	0
(9)	True	0.1 | 0.15	False	False	True	5

4.1. Non-Satellite Model

The Non-Satellite Model (NSM) leverages 37 features excluding the vegetation indices data. The list of features is outlined in Table 8.

Table 8. Summary of the number of numerical feature groups used in non-satellite model.

Group of Features	No.
Aggregated weather features	4
Weather features	23
Soil features	4
Agrotechnical treatment features	6
Total	37

The modelling for the non-satellite data considered three different scenarios: (a) without outlier detection, (b) with outlier detection using the Local Outlier Factor method, and (c) with outlier detection using the One-Class SVM method.

In the case of modelling without outlier detection, the Mean Absolute Percentage Error (MAPE) was found to be 17.31% using SVR. It is important to note that 32 (5 PCs) features were identified as significant in this scenario.

In the case of modelling with outlier detection using the Local Outlier Factor, the Mean Absolute Percentage Error (MAPE) was found to be 16.99% using SVR. It is important to note that only five (PCs) features were identified as significant in this scenario.

In the third scenario, when the One-Class SVM method was incorporated, the Mean Absolute Percentage Error (MAPE) was found to be 18.47% using XGB. In addition, 18 features were identified as significant for the model built using the modified datasets.

4.2. Satellite Model

The Satellite Model (SM) takes into account the vegetation indices, containing a total of 128 features. The model creation for the satellite data considered the same three scenarios as for the non-satellite model.

In the case of modelling without outlier detection, the Mean Absolute Percentage Error (MAPE) equalled 14.87% using Ridge, and 92 features were identified as significant for the model built using the modified datasets.

In the case of modelling with outlier detection using the Local Outlier Factor, the Mean Absolute Percentage Error (MAPE) equalled 15.43% using Ridge, and 83 features were identified as significant for the model built using the modified datasets.

In the final scenario with the One-Class SVM method, the Mean Absolute Percentage Error (MAPE) equalled 16.38% using Ridge, and 102 features were identified as significant for the model built using the modified datasets.

4.3. Hybrid Model

The Hybrid Model (HM) takes into account all 165 features. This includes both vegetation indices and the features used in the non-satellite model. The modelling process for the whole set of data (hybrid model) considered the same three scenarios, including the approach without outlier detection, as well as two additional ones, using the LOF and One-Class SVM for this problem.

In the case of modelling without outlier detection, the Mean Absolute Percentage Error (MAPE) was found to be 6.10% using XGB. It is important to note that 79 features were identified as significant in this scenario.

Before applying the Local Outlier Factor method for outlier detection, the dimensions of the training, validation, and test datasets were as presented in Table 6. After applying the Local Outlier Factor method, the Mean Absolute Percentage Error (MAPE) was found to be 6.94% using Random Forest. It is important to note that 80 features were identified as significant in this scenario.

In the final scenario, with the initial dataset dimensions as with the LOF method, One-Class SVM was applied. In that case, the training dataset was reduced, indicating that the method identified and removed 11 records as outliers. In this case, the Mean Absolute

Percentage Error (MAPE) was found to be 5.85% using XGB. In addition, 57 features were identified as significant for the model built using the modified datasets.

4.4. Models Comparison

Although it is believed in agricultural practice that the potato is one of the plants with low production requirements, potato varieties for frying purposes need cultivation management at a very high level [80]. Choosing a good variety is a key element in determining the plant's behaviour under field conditions. The information related to the traits responsible for the quality of the product—ready to eat—is "written" in the genotype: high nutritional value and good sensory properties. Most quality traits of tubers are strongly influenced by a number of factors acting on the potato during the growing season [81]. It is known that yield plays a key role in the cultivation of potatoes for frying purposes, as it generates farm profitability [82].

Yield, or the product extracted from the crop, can be considered in various aspects. Potential (theoretical) yield is achieved when the main abiotic factors: CO_2 concentration, solar radiation, and air temperature, are used by the plants with the greatest efficiency [81]. To estimate the potential yield, additional aspects must be taken into account. It should be assumed that a particular variety is grown in an environment that is optimal for it, with sufficient water and nutrients, as well as effective control of all biotic stresses. Potential yield is important for crops and environments where irrigation, the amount and distribution of rainfall, or a combination of irrigation and rainfall ensure that water deficits do not reduce yields [83]. Determining the level of potential yield is difficult, but feasible. Simulation modelling, the results of detailed agronomic experiments, yield tests, and knowledge of the maximum yields achieved by farmers are used to achieve this goal [84,85].

Actual yield is the real harvest achieved by most producers under actual production conditions. Real yield is determined relatively easily, but accurate analytical results can only be obtained by ongoing monitoring of yield potential during the growing season. The integration of several methods then comes to the rescue: remote sensing, geospatial analysis, and modelling combined with method validation through field experiments [81,84].

Maintaining high yield potential in the era of climate change is a very difficult task. The relationship between potential, achievable, and real potato yield is well explained by, i.e., yield gap analysis [86]. The yield gap shows the relationship between quantitative differences in potential, attainable, and actual yield at a specific spatial and temporal scale [85]. This analysis makes it possible to reliably identify unused food production capacity [81,87].

The above considerations show that yield prediction, regardless of the purpose of the forecast, is necessary and important [88]. Most valuable, from the point of view of agricultural practice, are models that allow the prediction of pre-harvest yields, in the current agronomic season [23,41,71]. In the case of potato production for French fries, the prediction of the actual yield of tubers before harvest provides the producer with a range of valuable information. They can be the basis for considering the amount of potential profit, the degree of fulfilment of the contract agreement, and the security of storage space [23,37]. The prediction of potential tuber yield, made before harvest, is also crucial for breeders of new varieties and seed companies [23,89]. The results of the analyses will indicate the "fit" of the tested genotypes to local growing conditions while maintaining a high level of controllable factors. The yield gap forecast provides valuable knowledge to institutions that track national and global food resources. It allows estimating food shortages, especially in poor countries with malnourished populations. Currently, it is believed that actual potato yields only reach 2/3 of their potential. Breeders of new varieties are far less likely to fill the gaps with improved, high-yielding genotypes than they could [33,81]. Effective planning and management of potato production now require the use of effective forecasting tools [90]. Tuber-yield-forecasting products must be carefully prepared and well thought out. The greatest difficulties in working with forecasting models are the selection of an appropriate prediction method and the selection of independent

variables that realistically affect tuber yield. It is important that all of the variables tested are readily available to the average user of such models and describe the relationships between phenomena in potato cultivation in a way that is understandable to the producer [16,23,24].

An important measure of prediction quality is the MAPE. The MAPE is defined as the average variance between the significant values in the dataset and the projected values in the same dataset [91]. The interpretation of the magnitude of this error is as follows: a MAPE of less than 10% indicates a very good model fit; when the MAPE is in the range of 10–20%, the degree of model fitness is good. A forecasting model that achieves a MAPE error of more than 30% should be rejected due to the poor mapping of predicted values to the actual ones [45,92]. In agricultural research, an acceptable upper limit for the MAPE's magnitude is around 15% [23,41,42].

Current trends in potato yield forecasting are mainly directed toward the use of various spectral indices and GIS data as independent variables for model construction [16,37,38]. Al Gaadi et al. [93] assessed crop condition and predicted potato tuber yield in Saudi Arabia. Two vegetation indices, NDVI and SAVI, were generated from Landsat-8 and Sentinel 2 satellite images acquired from different stages of potato growth. Yield samples were collected 2–3 days before harvest and correlated with the final yield. Based on this, yield-prediction models and yield maps were developed. The results showed that the difference between predicted yield values and actual yield values (prediction error) ranged from 7.9 to 13.5% for Landsat-8 images and from 3.8 to 10.2% for Sentinel-2 images. Since the prediction errors in the above cases did not exceed 15%, the models created by the authors can be used in practical applications. Li et al. [94] attempted to improve potato yield predictions using Unmanned Aerial Vehicle (UAV) remote sensing by incorporating variety information into machine learning methods. The research was conducted in the state of Minnesota—the northern part of the United States. Although the authors failed to generate accurate predictive models, very interesting research conclusions were drawn. Firstly, it was discovered that UAV-based spectral data from early in the growing season at the tuber initiation stage (late June) were more correlated with the commercial yield of potatoes than spectral data from later in the growing season at the tuber maturation stage. Secondly, it was established that combining high-spatial-resolution UAV images and variety information using machine learning algorithms can significantly improve potato yield prediction, when compared with methods excluding the variety information. The work on yield prediction in potato cultivation is difficult, but research shows that the most-accurate models can be achieved with the compilation of multiple variables: agrotechnical, soil, spectral, and meteorological.

In this study, three distinct models were used—non-satellite, satellite, and hybrid. Each of these models was evaluated in three different scenarios: (a) without outlier detection, (b) with outlier detection using the Local Outlier Factor method, and (c) with outlier detection using the One-Class SVM method. The comparative summary of the non-satellite, satellite, and hybrid models is presented in Table 9.

The comparative analysis of the non-satellite, satellite, and hybrid models in potato yield prediction revealed distinct trends in model performance across various scenarios. The hybrid models consistently showed superior predictive accuracy, evidenced by their significantly lower Mean Absolute Percentage Error (MAPE) values in all scenarios. This enhanced performance is likely attributed to the comprehensive integration of both satellite and non-satellite data features, suggesting the critical role of a diverse feature set in predictive modelling.

In scenarios where Principal Component Analysis (PCA) was applied, particularly in the non-satellite and hybrid models with Support Vector Machine (SVM) for outlier detection, there was a notable reduction in the number of features used. This indicates that PCA is effective at refining feature sets, thereby potentially improving model performance. Specifically, the hybrid model with SVM outlier detection not only achieved the lowest MAPE, but also demonstrated the impactful role of PCA in optimising the feature set for enhanced predictive accuracy.

Table 9. Comparative summary of non-satellite, satellite, and hybrid models.

Type	Model	Outlier Detection	No. of Features	MAPE	PCA Used	PCA No. of Features
Non-satellite	SVR	N/A	32	17.31%	True	5
Non-satellite	SVR	LOF	37	16.99%	True	5
Non-satellite	XGB	SVM	18	8.47%	True	5
Satellite	Ridge	N/A	92	14.87%	False	0
Satellite	Ridge	LOF	83	15.43%	False	0
Satellite	Ridge	SVM	102	16.38%	False	0
Hybrid	XGB	N/A	79	6.10%	False	0
Hybrid	Random Forest	LOF	80	6.94%	False	0
Hybrid	XGB	SVM	57	5.85%	True	5

Conversely, the non-satellite models, which lacked satellite-derived vegetation indices, exhibited higher MAPE values. This observation underscores the importance of vegetation indices in yield prediction, highlighting their contribution to model accuracy.

The satellite models presented an interesting trend, where an increase in the number of features, as seen in the SVM scenario, did not correspond to a decrease in the MAPE. This contrasts with the hybrid models, where a more judicious feature selection yielded better results. This suggests that increasing the number of features does not inherently enhance model performance; rather, the relevance and effective integration of these features are crucial.

The influence of outlier detection methods, namely the Local Outlier Factor (LOF) and SVM, varied across the models. While the hybrid models benefited significantly from SVM outlier detection, the impact on the non-satellite and satellite models was less pronounced. This difference in impact reiterates the necessity of context-specific approaches in outlier management for predictive modelling.

In summary, the hybrid models, especially with SVM for outlier detection, emerged as the most-effective strategy, achieving the lowest MAPE (5.85%) and, thereby, indicating the highest prediction accuracy among the evaluated models. This analysis reinforces the need for the careful selection and integration of features, coupled with appropriate data preprocessing techniques, to enhance the performance of machine learning models in agricultural yield prediction.

5. Conclusions

The comprehensive study on predicting potato yield using machine learning methods, specifically in the context of Polish potato varieties used for French fry production, yielded significant insights. The research highlighted the effectiveness of integrating diverse datasets, including both satellite and non-satellite data, in enhancing the accuracy of yield predictions. The hybrid model, which combined these datasets, demonstrated superior performance over models that utilised either non-satellite or satellite data alone. This superiority was evident in its lower Mean Absolute Percentage Error (MAPE) (5.85%), suggesting a higher prediction accuracy. The results clearly indicated that a multifaceted approach, utilising a broad spectrum of data sources, significantly improved the model's ability to predict yield accurately.

Advanced data processing techniques, such as feature selection and outlier detection, were found to play a pivotal role in the performance of the predictive models. The application of Principal Component Analysis (PCA) and outlier detection methods, including the Local Outlier Factor (LOF) and One-Class SVM, contributed to improvements in model accuracy. This underscores the importance of sophisticated data processing in machine learning applications for agricultural yield prediction.

The comparative analysis of the non-satellite, satellite, and hybrid models, as presented in the table "Comparative summary of non-satellite, satellite, and hybrid models", provided critical insights. The analysis revealed that the hybrid model, especially when coupled with SVM for outlier detection, emerged as the most-effective in predicting potato yield. This model achieved the lowest MAPE, indicating its high accuracy and reliability. In contrast, the non-satellite and satellite models, while beneficial in certain scenarios, did not match the comprehensive accuracy of the hybrid model. The findings from this comparative analysis reinforce the conclusion that a combined approach, utilising an extensive array of features and data sources, is essential for developing robust and accurate agricultural-yield-prediction models.

In conclusion, this study illustrated the potential of machine learning methods in revolutionising agricultural yield predictions. The integration of varied data sources, coupled with advanced data-processing techniques, offers a pathway towards more-efficient, -informed, and -sustainable agricultural practices. As the field of agricultural technology continues to evolve, these findings provide a foundation for further research and development in yield prediction and crop management.

Author Contributions: K.B., G.N. and T.W. conceived of the study design, managed the data collection, built the database, and performed the first data analysis. J.K. and B.Ś. carried out all deep data analyses and built the models until final results were attained. I.A., G.N., T.W., J.K., M.P., M.K. and B.Ś. wrote the manuscript with substantial input from K.B. All authors have read and agreed to the published version of the manuscript.

Funding: The project is co-financed by the European Union from the European Regional Development Fund under the Smart Growth Operational Programme. The project is being conducted under the competition of the National Centre for Research and Development, within the 1.1.1 programme for R&D projects of enterprises "Fast track–Agrotech" Number POIR.01.01.01-00-2298/20.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy issues.

Acknowledgments: We gratefully acknowledge Łukasz Cypcar and Szymon Margański for facilitating the sample collection and organising the data in storage and all the members from Seth Software teams for their assistance and helpful discussions. We thank Joanna Kogut for administrative services, without which we could not provide such fruitful results.

Conflicts of Interest: The authors declare no conflict of interest. Author Krzysztof Bobran was employed by the company Seth Software. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. FAO. FAOSTAT Statistical Database. 2023. Available online: https://ourworldindata.org/grapher/potato-yields (accessed on 7 June 2023).
2. Potatonewstoday. FAO Updates Global Potato Statistics. Available online: https://www.potatonewstoday.com/2022/03/28/fao-updates-global-potato-statistics/ (accessed on 7 June 2023).
3. Popkin, B.M.; Reardon, T. Obesity and the food system transformation in Latin America. *Obes. Rev.* **2018**, *19*, 1028–1064. [CrossRef] [PubMed]
4. Shafi, U.; Mumtaz, R.; García-Nieto, J.; Hassan, S.A.; Zaidi, S.A.R.; Iqbal, N. Precision agriculture techniques and practices: From considerations to applications. *Sensors* **2019**, *19*, 3796. [CrossRef] [PubMed]
5. Vannoppen, A.; Gobin, A. Estimating yield from NDVI, weather data, and soil water depletion for sugar beet and potato in Northern Belgium. *Water* **2022**, *14*, 1188. [CrossRef]
6. Newton, I.H.; Tariqul Islam, A.; Saiful Islam, A.; Tarekul Islam, G.; Tahsin, A.; Razzaque, S. Yield prediction model for potato using landsat time series images driven vegetation indices. *Remote Sens. Earth Syst. Sci.* **2018**, *1*, 29–38. [CrossRef]
7. Cambouris, A.N.; Zebarth, B.J.; Ziadi, N.; Perron, I. Precision agriculture in potato production. *Potato Res.* **2014**, *57*, 249–262. [CrossRef]

8. Hwang, E.; Park, Y.S.; Kim, J.Y.; Park, S.H.; Kim, J.; Kim, S.H. Intraoperative Hypotension Prediction Based on Features Automatically Generated Within an Interpretable Deep Learning Model. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, 1–15. [CrossRef]
9. Renju, R.S.; Deepthi, P.S.; Chitra, M.T. A Review of Crop Yield Prediction Strategies based on Machine Learning and Deep Learning. In Proceedings of the 2022 International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS), Kochi, India, 23–25 June 2022; pp. 1–6. [CrossRef]
10. Šťastná, M.; Toman, F.; Dufková, J. Usage of SUBSTOR model in potato yield prediction. *Agric. Water Manag.* **2010**, *97*, 286–290. [CrossRef]
11. Ahmed, M.; Fatima, Z.; Iqbal, P.; Kalsoom, T.; Abbasi, K.S.; Shaheen, F.A.; Ahmad, S. Potato Modelling. In *Systems Modelling*; Ahmed, M., Ed.; Springer: Singapore, 2020; pp. 383–401. [CrossRef]
12. Divya, K.L.; Mhatre, P.H.; Venkatasalam, E.P.; Sudha, R. Crop Simulation Models as Decision-Supporting Tools for Sustainable Potato Production: A Review. *Potato Res.* **2021**, *64*, 387–419. [CrossRef]
13. Travasso, M.I.; Caldiz, D.O.; Saluzzo, J.A. Yield prediction using the SUBSTOR-potato model under Argentinian conditions. *Potato Res.* **1996**, *39*, 305–312. [CrossRef]
14. Bala, S.K.; Islam, A.S. Correlation between potato yield and MODIS-derived vegetation indices. *Int. J. Remote Sens.* **2009**, *30*, 2491–2507. [CrossRef]
15. Gómez, D.; Salvador, P.; Sanz, J.; Casanova, J.L. Potato Yield Prediction Using Machine Learning Techniques and Sentinel 2 Data. *Remote Sens.* **2019**, *11*, 1745. [CrossRef]
16. Gómez, D.; Salvador, P.; Sanz, J.; Casanova, J.L. New spectral indicator Potato Productivity Index based on Sentinel-2 data to improve potato yield prediction: A machine learning approach. *Int. J. Remote Sens.* **2021**, *42*, 3426–3444. [CrossRef]
17. Sun, J.; Di, L.; Sun, Z.; Shen, Y.; Lai, Z. County-Level Soybean Yield Prediction Using Deep CNN-LSTM Model. *Sensors* **2019**, *19*, 4363 [CrossRef] [PubMed]
18. Gobin, A.; Sallah, A.H.M.; Curnel, Y.; Delvoye, C.; Weiss, M.; Wellens, J.; Piccard, I.; Planchon, V.; Tychon, B.; Goffart, J.P.; et al. Crop Phenology Modelling Using Proximal and Satellite Sensor Data. *Remote Sens.* **2023**, *15*, 2090. [CrossRef]
19. Lin, Y.; Li, S.; Ye, Y.; Li, B.; Li, G.; Lyv, D.; Jin, L.; Bian, C.; Liu, J. Methodological evolution of potato yield prediction: A comprehensive review. *Front. Plant Sci.* **2023**, *14*, 1214006. [CrossRef]
20. Akhand, K.; Nizamuddin, M.; Roytman, L.; Kogan, F. Using remote sensing satellite data and artificial neural network for prediction of potato yield in Bangladesh. In *Remote Sensing and Modelling of Ecosystems for Sustainability XIII*; SPIE: Bellingham, WA, USA, 2016; Volume 9975, pp. 52–66.
21. Al-Gaadi, K.A.; Hassaballa, A.A.; Madugundu, R.; Tola, E.; Fulleros, R.B. Prediction of potato high-yield zones of a field: Bivariate frequency ratio technique. *Curr. Sci.* **2020**, *119*, 992. [CrossRef]
22. Noman, A.M.; Haidar, Z.A.; Aljumah, A.S.; Almutairi, S.Z.; Alqahtani, M.H. Forecasting the Distortion in Solar Radiation during Midday Hours by Analyzing Solar Radiation during Early Morning Hours. *Appl. Sci.* **2023**, *13*, 6049. [CrossRef]
23. Piekutowska, M.; Niedbała, G.; Piskier, T.; Lenartowicz, T.; Pilarski, K.; Wojciechowski, T.; Pilarska, A.A.; Czechowska-Kosacka, A. The application of multiple linear regression and artificial neural network models for yield prediction of very early potato cultivars before harvest. *Agronomy* **2021**, *11*, 885. [CrossRef]
24. Hara, P.; Piekutowska, M.; Niedbała, G. Selection of independent variables for crop yield prediction using artificial neural network models with remote sensing data. *Land* **2021**, *10*, 609. [CrossRef]
25. Li, Q.; Zhang, S. Impacts of recent climate change on potato yields at a provincial scale in Northwest China. *Agronomy* **2020**, *10*, 426. [CrossRef]
26. Rymuza, K.; Radzka, E.; Lenartowicz, T. Effect of weather conditions on early potato yields in east-central Poland. *Commun. Biometry Crop Sci.* **2015**, *10*, 65–72.
27. Nyawade, S.O.; Gitari, H.I.; Karanja, N.N.; Gachene, C.K.; Schulte-Geldermann, E.; Parker, M.L. Yield and evapotranspiration characteristics of potato-legume intercropping simulated using a dual coefficient approach in a tropical highland. *Field Crop. Res.* **2021**, *274*, 108327. [CrossRef]
28. Blecharczyk, A.; Kowalczewski, P.Ł.; Sawinska, Z.; Rybacki, P.; Radzikowska-Kujawska, D. Impact of Crop Sequence and Fertilization on Potato Yield in a Long-Term Study. *Plants* **2023**, *12*, 495. [CrossRef] [PubMed]
29. Pandey, J.; Scheuring, D.C.; Koym, J.W.; Vales, M.I. Genomic regions associated with tuber traits in tetraploid potatoes and identification of superior clones for breeding purposes. *Front. Plant Sci.* **2022**, *13*, 952263. [CrossRef] [PubMed]
30. Singh, B. Are nitrogen fertilizers deleterious to soil health? *Agronomy* **2018**, *8*, 48. [CrossRef]
31. Hasnain, M.; Chen, J.; Ahmed, N.; Memon, S.; Wang, L.; Wang, Y.; Wang, P. The effects of fertilizer type and application time on soil properties, plant traits, yield and quality of tomato. *Sustainability* **2020**, *12*, 9065. [CrossRef]
32. Fiers, M.; Edel-Hermann, V.; Chatot, C.; Le Hingrat, Y.; Alabouvette, C.; Steinberg, C. Potato soil-borne diseases. A review. *Agron. Sustain. Dev.* **2012**, *32*, 93–132. [CrossRef]
33. Vreugdenhil, D.; Bradshaw, J.; Gebhardt, C.; Govers, F.; Taylor, M.A.; MacKerron, D.K.; Ross, H.A. *Potato Biology and Biotechnology: Advances and Perspectives*; Elsevier: Amsterdam, The Netherlands, 2011.
34. Boyd, N.; Gordon, R.; Martin, R. Relationship between leaf area index and ground cover in potato under different management conditions. *Potato Res.* **2002**, *45*, 117–129. [CrossRef]

35. Quiroz, R.; Loayza, H.; Barreda, C.; Gavilán, C.; Posadas, A.; Ramírez, D. Linking process-based potato models with light reflectance data: Does model complexity enhance yield prediction accuracy? *Eur. J. Agron.* **2017**, *82*, 104–112. [CrossRef]
36. Rokhafrouz, M.; Latifi, H.; Abkar, A.; Wojciechowski, T.; Czechlowski, M.; Naieni, A.; Maghsoudi, Y.; Niedbała, G. Simplified and Hybrid Remote Sensing-Based Delineation of Management Zones for Nitrogen Variable Rate Application in Wheat. *Agriculture* **2021**, *11*, 1104. [CrossRef]
37. Salvador, P.; Gómez, D.; Sanz, J.; Casanova, J.L. Estimation of potato yield using satellite data at a municipal level: A machine learning approach. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 343. [CrossRef]
38. Samborski, S.; Leszczyńska, R.; Gozdowski, D. Detecting spatial variability of potato canopy using various remote sensing data. In Proceedings of the Precision Agriculture'21, Budapest, Hungary, 19–22 July 2021; Wageningen Academic Publishers: Wageningen, The Netherlands, 2021; pp. 1786–1798.
39. Prasad Patnaik, P.; Padhy, N. An Approach for Potato Yield Prediction Using Machine Learning Regression Algorithms. In Proceedings of the Next Generation of Internet of Things, Gunupur, India, 5–6 February 2021; Kumar, R., Mishra, B.K., Pattnaik, P.K., Eds.; Springer Nature: Singapore, 2023; pp. 327–336.
40. Sharma, A.K.; Rajawat, A.S. Crop Yield Prediction using Hybrid Deep Learning Algorithm for Smart Agriculture. In Proceedings of the 2022 Second International Approach for Potato Yield Prediction Using Machine Learning Regression Algorithmsional Conference on Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India, 23–25 February 2022; pp. 330–335. [CrossRef]
41. 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]
42. Niedbała, G.; Kurasiak-Popowska, D.; Piekutowska, M.; Wojciechowski, T.; Kwiatek, M.; Nawracała, J. Application of Artificial Neural Network Sensitivity Analysis to Identify Key Determinants of Harvesting Date and Yield of Soybean (*Glycine max* [L.] Merrill) Cultivar Augusta. *Agriculture* **2022**, *12*, 754. [CrossRef]
43. Niedbała, G.; Wróbel, B.; Piekutowska, M.; Zielewicz, W.; Paszkiewicz-Jasińska, A.; Wojciechowski, T.; Niazian, M. Application of Artificial Neural Networks Sensitivity Analysis for the Pre-Identification of Highly Significant Factors Influencing the Yield and Digestibility of Grassland Sward in the Climatic Conditions of Central Poland. *Agronomy* **2022**, *12*, 1133. [CrossRef]
44. Niedbała, G.; Kurasiak-Popowska, D.; Stuper-Szablewska, K.; Nawracała, J. Application of Artificial Neural Networks to Analyze the Concentration of Ferulic Acid, Deoxynivalenol, and Nivalenol in Winter Wheat Grain. *Agriculture* **2020**, *10*, 127. [CrossRef]
45. Hara, P.; Piekutowska, M.; Niedbała, G. Prediction of Protein Content in Pea (*Pisum sativum* L.) Seeds Using Artificial Neural Networks. *Agriculture* **2022**, *13*, 29. [CrossRef]
46. Boniecki, P.; Sujak, A.; Niedbała, G.; Piekarska-Boniecka, H.; Wawrzyniak, A.; Przybylak, A. Neural Modelling from the Perspective of Selected Statistical Methods on Examples of Agricultural Applications. *Agriculture* **2023**, *13*, 762. [CrossRef]
47. Gonzalez-Sanchez, A.; Frausto-Solis, J.; Ojeda-Bustamante, W. Attribute selection impact on linear and nonlinear regression models for crop yield prediction. *Sci. World J.* **2014**, *2014*, 509429. [CrossRef]
48. Maestrini, B.; Mimić, G.; van Oort, P.A.; Jindo, K.; Brdar, S.; Athanasiadis, I.N.; van Evert, F.K. Mixing process-based and data-driven approaches in yield prediction. *Eur. J. Agron.* **2022**, *139*, 126569. [CrossRef]
49. Morales, A.; Villalobos, F.J. Using machine learning for crop yield prediction in the past or the future. *Front. Plant Sci.* **2023**, *14*, 1128388. [CrossRef]
50. Ansarifar, J.; Wang, L.; Archontoulis, S.V. An interaction regression model for crop yield prediction. *Sci. Rep.* **2021**, *11*, 17754. [CrossRef]
51. Kuradusenge, M.; Hitimana, E.; Hanyurwimfura, D.; Rukundo, P.; Mtonga, K.; Mukasine, A.; Uwitonze, C.; Ngabonziza, J.; Uwamahoro, A. Crop Yield Prediction Using Machine Learning Models: Case of Irish Potato and Maize. *Agriculture* **2023**, *13*, 225. [CrossRef]
52. Yun, S.D.; Gramig, B.M. Spatial Panel Models of Crop Yield Response to Weather: Econometric Specification Strategies and Prediction Performance. *J. Agric. Appl. Econ.* **2022**, *54*, 53–71. [CrossRef]
53. Fadón, E.; Herrero, M.; Rodrigo, J. Flower development in sweet cherry framed in the BBCH scale. *Sci. Hortic.* **2015**, *192*, 141–147. [CrossRef]
54. Alcaraz, M.; Thorp, T.; Hormaza, J. Phenological growth stages of avocado (*Persea americana*) according to the BBCH scale. *Sci. Hortic.* **2013**, *164*, 434–439. [CrossRef]
55. Seth Software Sp. z o.o. Plantator System. 2023. Available online: https://plantator.com (accessed on 1 November 2023).
56. Matsushita, B.; Yang, W.; Chen, J.; Onda, Y.; Qiu, G. Sensitivity of the Enhanced Vegetation Index (EVI) and Normalized Difference Vegetation Index (NDVI) to Topographic Effects: A Case Study in High-density Cypress Forest. *Sensors* **2007**, *7*, 2636–2651. [CrossRef]
57. Chmist-Sikorska, J.; Kepinska-Kasprzak, M.; Struzik, P. Agricultural drought assessment on the base of Hydro-thermal Coefficient of Selyaninov in Poland. *Ital. J. Agrometeorol.* **2022**, *1*, 3–12. [CrossRef]
58. McMaster, G.S.; Wilhelm, W. Growing degree-days: One equation, two interpretations. *Agric. For. Meteorol.* **1997**, *87*, 291–300. [CrossRef]
59. Zhang, Z. Missing data imputation: Focusing on single imputation. *Ann. Transl. Med.* **2016**, *4*, 9.
60. Henderi, H.; Wahyuningsih, T.; Rahwanto, E. Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer. *Int. J. Inform. Inf. Syst.* **2021**, *4*, 13–20. [CrossRef]

61. Jegorowa, A.; Górski, J.; Kurek, J.; Kruk, M. Use of nearest neighbors (k-NN) algorithm in tool condition identification in the case of drilling in melamine faced particleboard. *Maderas. Ciencia Y TecnologÃa*. **2020**, *22*, 189–196. Available online: http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-221X2020000200189&nrm=iso (accessed on 1 November 2023). [CrossRef]
62. sklearn.linear_model.LinearRegression—Scikit-Learn 1.0.2 Documentation. 2023. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html (accessed on 1 November 2023).
63. Ridge Regression in Scikit-Learn. 2023. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html (accessed on 1 November 2023).
64. Lasso in Scikit-Learn. 2023. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html (accessed on 1 November 2023).
65. ElasticNet in Scikit-Learn. 2023. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html (accessed on 1 November 2023).
66. XGBoost Python Package. 2023. Available online: https://xgboost.readthedocs.io/en/stable/Python/python_api.html (accessed on 1 November 2023).
67. Random Forest Regressor in Scikit-Learn. 2023. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html (accessed on 1 November 2023).
68. MLP Regressor in Scikit-Learn. 2023. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html (accessed on 1 November 2023).
69. SGD Regressor in Scikit-Learn. 2023. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDRegressor.html (accessed on 1 November 2023).
70. SVR in Scikit-Learn. 2023. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html (accessed on 1 November 2023).
71. Niedbała, G.; Kurek, J.; Świderski, B.; Wojciechowski, T.; Antoniuk, I.; Bobran, K. Prediction of Blueberry (*Vaccinium corymbosum* L.) Yield Based on Artificial Intelligence Methods. *Agriculture* **2022**, *12*, 2089. [CrossRef]
72. Kurek, J.; Świderski, B.; Osowski, S.; Kruk, M.; Barhoumi, W. Deep learning versus classical neural approach to mammogram recognition. *Bull. Pol. Acad. Sci. Tech. Sci.* **2018**, *66*, 831–840. [CrossRef]
73. Swiderski, B.; Kurek, J.; Osowski, S. Multistage classification by using logistic regression and neural networks for assessment of financial condition of company. *Decis. Support Syst.* **2012**, *52*, 539–547. [CrossRef]
74. Osowski, S.; Les, T. Deep learning ensemble for melanoma recognition. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–7.
75. Gil, F.; Osowski, S.; Slowinska, M. Melanoma recognition using deep learning and ensemble of classifiers. In Proceedings of the 2022 23rd International Conference on Computational Problems of Electrical Engineering (CPEE), Zuberec, Slovak Republic, 11–14 September 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–4.
76. Kruk, M.; Kurek, J.; Osowski, S.; Koktysz, R.; Swiderski, B.; Markiewicz, T. Ensemble of classifiers and wavelet transformation for improved recognition of Fuhrman grading in clear-cell renal carcinoma. *Biocybern. Biomed. Eng.* **2017**, *37*, 357–364. [CrossRef]
77. Siwek, K.; Osowski, S.; Kurek, J. Ensemble Neural Network Approach to the Load Forecasting in the Power System. In Proceedings of the International IEEE Conference on ISTET'05, Lviv, Ukraine, 4–7 July 2005; pp. 380–383.
78. Kurek, J.; Osowski, S. Support Vector Machine for diagnosis of the bars of cage inductance motor. In Proceedings of the 2008 15th IEEE International Conference on Electronics, Circuits and Systems, Saint Julian's, Malta, 31 August–3 September 2008; IEEE: Piscataway, NJ, USA, 2008; pp. 1022–1025.
79. Cheng, Z.; Zou, C.; Dong, J. Outlier Detection Using Isolation Forest and Local Outlier Factor. In Proceedings of the Conference on Research in Adaptive and Convergent Systems, RACS'19, Chongqing, China, 24–27 September 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 161–168. [CrossRef]
80. Sawicka, B.; Pszczółkowski, P.; Kiełtyka-Dadasiewicz, A.; Barbaś, P.; Ćwintal, M.; Krochmal-Marczak, B. The Effect of Effective Microorganisms on the Quality of Potato Chips and French Fries. *Appl. Sci.* **2021**, *11*, 1415 . [CrossRef]
81. Haverkort, A.; Struik, P. Yield Levels of Potato Crops: Recent Achievements and Future Prospects. *Field Crops Res.* **2021**, *182*, 76–85. [CrossRef]
82. Cirocki, R.; Gołębiewska, B. Changes in the profitability of production of industrial potatoes in poland—A case study. *Ann. Polish Assoc. Agric. Agribus. Econ.* **2019**, *21*, 19–28. [CrossRef]
83. Licker, R.; Johnston, M.; Foley, J.; Barford, C.; Kucharik, C.; Monfreda, C.; Ramankutty, N. Mind the Gap: How Do Climate and Agricultural Management Explain the 'Yield Gap' of Croplands around the World? *Glob. Ecol. Biogeogr.* **2010**, *19*, 769–782. [CrossRef]
84. Hochman, Z.; Gobbett, D.; Holzworth, D.; McClelland, T.; van Rees, H.; Marinoni, O.; Garcia, J.; Horan, H. Reprint of "Quantifying Yield Gaps in Rainfed Cropping Systems: A Case Study of Wheat in Australia". *Field Crops Res.* **2013**, *143*, 65–75. [CrossRef]
85. Harahagazwe, D.; Condori, B.; Barreda, C.; Bararyenya, A.; Byarugaba, A.; Kude, D.; Lung'aho, C.; Martinho, C.; Mbiri, D.; Nasona, B.; et al. How Big Is the Potato (*Solanum tuberosum* L.) Yield Gap in Sub-Saharan Africa and Why? A Participatory Approach. *Open Agric.* **2018**, *3*, 180–189. [CrossRef]
86. Campos, H.; Ortiz, O. The Potato Crop. In *Its Agricultural, Nutritional and Social Contribution to Humankind*; Springer: Berlin/Heidelberg, Germany, 2020; ISBN 978-3-030-28682-8.

87. Grassini, P.; van Bussel, L.; Van Wart, J.; Wolf, J.; Claessens, L.; Yang, H.; Boogaard, H.; de Groot, H.; van Ittersum, M.; Cassman, K. How Good Is Good Enough? Data Requirements for Reliable Crop Yield Simulations and Yield-Gap Analysis. *Field Crops Res.* **2015**, *177*, 49–63. [CrossRef]
88. Meroni, M.; Waldner, F.; Seguini, L.; Kerdiles, H.; Rembold, F. Yield Forecasting with Machine Learning and Small Data: What Gains for Grains? *Agric. For. Meteorol.* **2021**, *108555*, 308–309.
89. Dwivedi, S.; Goldman, I.; Ortiz, R. Pursuing the Potential of Heirloom Cultivars to Improve Adaptation, Nutritional, and Culinary Features of Food Crops. *Agronomy* **2019**, *9*, 441. [CrossRef]
90. Ahmad, U.; Sharma, L.A. Review of Best Management Practices for Potato Crop Using Precision Agricultural Technologies. *Smart Agric. Technol.* **2023**, *4*, 100220. [CrossRef]
91. Vetrovsky, T.; Siranec, M.; Marencakova, J.; Tufano, J.; Capek, V.; Bunc, V.; Belohlavek, J. Validity of Six Consumer-Level Activity Monitors for Measuring Steps in Patients with Chronic Heart Failure. *PLoS ONE* **2019**, *14*, e0222569. [CrossRef]
92. Hara, P.; Piekutowska, M.; Niedbała, G. Prediction of Pea (*Pisum sativum* L.) Seeds Yield Using Artificial Neural Networks. *Agriculture* **2023**, *13*, 661. [CrossRef]
93. Al-Gaadi, K.; Hassaballa, A.; Tola, E.; Kayad, A.; Madugundu, R.; Alblewi, B.; Assiri, F. Prediction of Potato Crop Yield Using Precision Agriculture Techniques. *PLoS ONE* **2016**, *11*, e0162219. [CrossRef]
94. Li, D.; Miao, Y.; Gupta, S.; Rosen, C.; Yuan, F.; Wang, C.; Wang, L.; Huang, Y. Improving Potato Yield Prediction by Combining Cultivar Information and UAV Remote Sensing Data Using Machine Learning. *Remote Sens.* **2021**, *13*, 3322. [CrossRef]