York St John University

Department of Computer Science
School of Science, Technology and Health

A Machine Learning Approach to Rapeseed Yield Prediction and Shapley Value Analysis

Charlie Rhys Hamer

A thesis submitted in fulfilment of the requirements
for the degree of Master of Science by Research

June 2025

Declaration

I, Charlie Rhys Hamer, confirm that the submission is my own work, that I have not presented anyone else's work as my own and that full and appropriate acknowledgement has been given where reference has been made to the work of others.

I have read and understood the University's 'Research Misconduct Policy'.

I understand that if I commit research misconduct I can be terminated from the University and that it is my responsibility to be aware of the University's 'Research Misconduct Policy' and its importance.

I consent to the University making available to third parties (who may be based outside the European Economic Area) any of my work in any form for standards and monitoring purposes including verifying the absence of plagiarised material.

I agree that third parties may retain copies of my work for these purposes on the understanding that the third party will not disclose my identity.

Copyright

Abstract

Crop harvest prediction depends on many factors, including soil, irrigation, weather, crop genotype, and how all of it is maintained. This study aims to analyse the performance of a pre-existing formula for yield prediction, and build a new machine learning based model to compare it to. This study will use a dataset of rapeseed harvests over three years around the areas of South Wales, Derbyshire, and Lincolnshire to train the model. This dataset will have a dimensionality reduction performed on it, and the effects of that will be evaluated. This study will test the dataset on an independent Nottinghamshire-based dataset for generalisation purposes.

This study proposes a 1-dimensional Convolutional Neural Network, and it will evaluate its performance using the RMSLE prediction matrix.

Finally, this study will use Shapley values to analyse the predictions of the neural network further, and use them to suggest improvements to farmers and data collection researchers. These Shapley values will show which features in the dataset provide the most influence over the final output prediction by the AI model. This research will also discover the importance of the vegetation matrix in such calculations.

The proposed CNN outperformed the original formula, broadcasting results of 96% prediction accuracy under ideal dataset conditions. However, this research observed failures to generalise to new areas with the testing of the Nottinghamshire dataset, and how removing the vegetation index variable can reduce the accuracy measures by almost 10%.

Using Shapley values, the vegetation index, water stress and nitrate density variables were determined to generally be the most important, but some nuance exists based on number of features in the tested dataset and the omission of the vegetation index.

Contents

List of Abbreviations

Abbreviation	Meaning
AI	Artificial Intelligence
CNN	Convolutional Neural Network
ANN	Artificial Neural Network
ML	Machine Learning
VI	Vegetation Index
LR	Linear Regression
RF	Random Forest
GBT	Gradient Boosting Tree
LSTM	Long Short-Term Memory
SVM	Support Vector Machines
NDVI	Normalised Difference Vegetation Index
SAVI	Soil Adjusted Vegetation Index
GVI	Green Vegetation Index
IPVI	Infrared Percentage Vegetation Index
RVI	Ratio Vegetation Index
DVI	Difference Vegetation Index
NDWI	Normalised Difference Water Index
LAI	Leaf Area Index
ReLU	Rectified Linear Unit
KNN	K-Nearest Neighbours
SVR	Support Vector Regression
DNN	Deep Neural Networks
RNN	Recurrent Neural Networks
GRU	Gated Recurrent Unit
MDI	Mean Decrease in Impurity
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
RMSLE	Root Mean Square Logarithmic Error
SHAP	Shapley Additive Explanations
ROE	Return On Equity
ADB	Asian Development Bank
KP	Plant Growth Coefficient
Kw	Water Stress
ET0	Rate Of Transpiration From Plant
pH	Potential Of Hydrogen
NPK	Nitrogen, Phosphorus, and Potassium
CSV	Comma Separated Values
NIR	Near-Infrared
RMSProp	Root Mean Square Propagation
SGD	Stochastic Gradient Descent

1 Introduction

The human race population is expected to grow from 8 billion in 2025 to 10.5 billion by 2080, conservative estimates of the UN claim[1]. With this rise in population, a need for sustainable and efficient food and fuel production arises. Continuous climate change, overpopulation and inefficient and unsustainable agricultural practices have led to a declining productivity in our farming sector and many countries still having high poverty rates[2][3]. As one of the most important aspects of digital agriculture, crop growth monitoring requires real-time access to accurate information on plant growth so as to provide guidance for refined management of crop and significantly improve the crop yields. Usage of a neural network for analysis of this can prove crucial, and much more time efficient for the researchers and farmers.

The cultivation of rapeseed is an important factor for many of the main economies in the world[4], as it provides essential proteins for cooking and can be refined into biodiesel, which may prove essential as the world transitions away from fossil fuels to an alternative that is more sustainable and renewable. Due to the nature of the versatility of this crop, the production and yield of it grows in importance every year, despite growth figures declining in some countries[5].

The Welsh Government has conducted research into the growth of rapeseed in South Wales, and over 30 years they have crafted a formula to estimate the yield of the crop over a given harvest. This formula, while having been adapted as time has gone on and newer methods of data collection have presented themselves, still finds itself outdated and unable to generalise. This research proposes such a formula should be considered outdated, and with modern data collection and analysation techniques, a new alternative should be presented.

As such, this research aims to produce a machine learning model to analyse data collected on rapeseed, within South Wales and other areas in the UK, to either assist or replace the model used by the Welsh Office. A neural network will be used for this task.

This study will also employ the use of Shapley values, a game theory concept, to better analyse how each variable of the soil, irrigation and plants themselves may be affecting the end result produced by the machine learning algorithm. This research proposes this will offer deeper insight into the rapeseed growth process and how farmers can better their crops, which should prove more insightful than the current formula, which only serves to predict the yield, and not why it produces that value. This will give helpful advice and bring awareness to the farmer about what they can do to better their farm's ecosystem to improve future harvests.

As such, the first research question is this:

1. Can AI be used to improve or replace the formula used by the Welsh Office for the prediction of rapeseed yields?

And the second question is:

2. Can Shapley Values be used with such an algorithm to further provide insight into the rapeseed harvest process?

1.1 Structure

First there will be a review of past literature, to ascertain where this research fits in to the current understanding of this subject, and to justify why this research needs to be carried out. This may raise other secondary research questions that can be answered as well.

Then, a methodology segment, presenting the traditional formula used by the Welsh government and the dataset to be used for this research, and the modelling of the neural network and Shapley value implementation.

After that, the results of the test will be presented, and the Shapley value graphs ready for analysis.

Then this study will review how these results have turned out, and discuss how they have affected current viewpoints and what can be done with that information.

Finally, a conclusion will be presented, discussing everything that has taken place and presenting possible future studies that can be conducted off the back of this experiment.

At the end, a bibliography will be compiled.

2 Literature Review

In this section reasons will be detailed for performing this study, and the historical research that has taken place to justify it. Such research will also serve as a foundation for each section of the study, and will give an idea of how best to conduct this investigation.

2.1 Data Collection

For this study, the methods of data collection and their efficiency will dictate how effectively the machine will be able to piece together patterns.

Dahane et al.[6] propose a cost-effective, sustainable and comprehensive irrigation system with a built-in data collection network. This system uses a fog-based IoT and AI framework, utilising an array of sensors to gather information regarding the soil, local climate, and water stress within the fields. This data serves to predict soil moisture levels for specific plots in the system, allowing precise scheduling of irrigation operations.

This system's data collection software is run from an Arduino and a handful of cheap sensors. Despite the simplicity of this design, the fog computing aspect enables a comprehensive detector system, all for roughly $140, or about £100. Its hardware components draw from open-source designs, and the software is similarly open-source, allowing modification on the fly and enabling it to be even cheaper. Extensive testing within farms has demonstrated the system's ability to conserve water and bolster crop yields.

The IoT is connected via a cluster of Raspberry Pi cards forming a local WiFi network. This allows the connectivity between the fog layer and the cloud layer. The researches use a free cloud service to store the data, but for a professional application the cloud service would need to be paid for. This system would work well within the confines of this experiment, as the AI will be built and stored on Microsoft Azure, which offers cloud storage services for results, before and after processing.

There are limitations to consider within this paper. The scope of it was small, and testing was not comprehensive, though the testing that did take place showed promising results for such a cheap system. The paper also fails to consider maintenance costs for the components, and doesn't measure other costs the farm may take on, such as electricity, water, or fertiliser costs, if applicable.

The paper tries to piece together an algorithm that could be trained with incomplete and unclean data. Doing this would streamline the process, as a lot of the raw data could be fed straight into the model without dimensionality reductions and with outliers included. However, the researchers found that even with high quality models, the error value was still quite high. They attributed this to their calibration techniques for the sensors. They aimed to improve this while still keeping the model low cost.

This proposed technology is a great launching point for further research. Integration with other agricultural technologies, such as fertiliser-laying systems and other, more comprehensive crop monitoring solutions could provide a wider

scope of data collection, allowing a well made machine to analyse the results more consistently. This would also help fix the issue of low quality data within the experiment.

This structure also provides cheap data collection system for use in other areas, including monitoring pest and disease pressures. This can provide the basis for early warning systems, allowing farmers and researches to act according to possible threats.

Having access to this sort of data also allows farmers to better seek advice from data analysts if they want a deeper dive into the numbers behind their crops. However, on its face, it should still provide a readable and simple analysis for the farmer to act on.

How data is obtained for this study's research is important as well, as the data needs to be comprehensive enough for a machine learning model to use without overfitting. Davrazos et al.[7] uses various Internet of Things (IoT) sources to feed a machine learning algorithm to determine the most suitable crop for a specific field. This is primarily calculated by factoring the soil characteristics and the predicted sale price of the crop. For the dataset, the soil had seven variables: Nitrogen, Phosphorus, Potassium, temperature, moisture, pH and rainfall. This set is limited compared to others like it, missing important factors such as pesticide usage and soil density.

With this limited data set, Random Forest classifiers emerged as the most effective tool for predicting the crop value. Shapley values ran on this test also state that average rainfall is the most important factor. Satellite imagery notably isn't used at all in this paper. This is often an important part of IoT measurements for crops, as it opens the use of vegetation indices, which Sun et al.[8] used comprehensively in their analysis of the rice rotation effect.

Six different indices were utilised, each using different calculations: normalised difference vegetation index (NDVI), soil adjusted vegetation index (SAVI), green vegetation index (GVI), infrared percentage vegetation index (IPVI), ratio vegetation index (RVI) and difference vegetation Index (DVI).

2.2 Satellite Imagery

Sensors on the ground may not be comprehensive enough for a complete study into a plant. This study has access to aerial technology, giving a bird's eye view of the farms, and allowing the use of vegetation indices to measure near-infrared reflectance associated with the health of the leaves of a plant. This imagery may give a unique insight into how effectively a plant can absorb sunlight, and how effectively it can make use of nutrients in the soil to grow and flourish.

This paper authored by Ibrahim et al.[9] presents an approach to evaluating crop growth by combining Sentinel-2 multispectral imagery, Landsat-8 imagery and local meteorological data. This method involves the use of multiple different vegetation indices, including NDVI, SAVI, and NDWI to monitor the development of wheat, and barley crops over time. The research also incorporates meteorological factors like rainfall and air temperature. This experiment was conducted in Iraq, where sandy soil is found in great quantities. The area is

dry and the yields of crops there needs to be optimised to achieve a worthwhile level of tonnage.

The study identifies the NDVI and NDWI as two of the stronger performers for prediction of crop growth in this area. Overall, the leaf area index (LAI) was the best predictor, as crop mortality was a significant factor here, so plants that could achieve a high leaf area were much more likely to produce well compared to others. This would be less of a factor for British crops, due to the less hostile growing conditions in Britain compared to Iraq. This means a selection of either NDVI or NDWI for the calculation of the vegetation index values would provide the highest possible estimation accuracy.

The paper also identifies problems with the use of satellite imagery. Namely, Landsat images face issues with their limited availability and infrequent visits to certain areas, and their susceptibility to cloud coverage. Sentinel-2 data visits more frequently and uses a superior spatial resolution, so as such can be utilised to greater effect than Landsat. However, both can be used in tandem, so one outclassing the other is not a problem.

Using both Landsat and Sentinel-2 data, combined with local meteorological studies, has been proven to be a very effective way of measuring growth coefficients. However, the researches identify that it can have a more limited effect during all seasons in particularly arid areas. This is not a problem for this study, though may play an effect in the discussion of the potential generalisability of the model.

2.3 Machine Learning Model for Satellite Imagery

Sentinel-2 data alone cannot provide an easy answer to estimations either. Saha et al.[10] state "training deep convolutional architecture with [a Sentinel-2] dataset is still a challenging task in land cover classification due to the absence of ground truth". They go on to claim that the selection of deep architecture for this model is a task in and of itself.

The study by Saha et al. in West Bengal used 10m, 20m and 60m bands stacked as inputs into a BasicConv2d layer each, followed by an adaptive average pooling layer. Following that is a series of 2-D convolutional layers with activation ReLU. They set their mini batches to be as large as possible so as to make sure there is enough diversification within the representation of classes in each batch.

They discover the use of 60m bands with either 10m or 20m produces workable results, but using any by themselves, or just 10m and 20m, does not. They found errors around classification of human settlements and sand river beds can be mixed up. However, this is still an improvement, as general 60m bands can mistake rivers for human settlement, which this CNN with 60m and 20m bands does not. Sub-categories of crops can also be misidentified.

Ultimately, this study finds a combination of 60m and 20m to be the optimal bands to cluster together, using a traditional CNN. However, the quality of land based clustering can depend on the initial parameters of the layers in the

network, so the decisions around that need to be made either manually or a way to automate them must be found.

2.4 Machine Learning Model for Prediction of Crop Growth

The model chosen to evaluate these datasets must be both robust and highly precise. It must be prepared to fairly evaluate the relevancy of many features, whether the dataset has been reduced or not. In an extensive review of different AI models and their competency at predicting crop yields, Shawon et al.[11] compared traditional Linear Regression (LR), Random Forest (RF) and Gradient Boosting Tree (GBT) machine learning algorithms, with deep learning algorithms such as Convolutional Neural Networks and Long Short-Term Memory (LSTM). Due to the growing methods of data collection and the precision at which this data can be processed, a robust and adaptable machine learning model for crop yield production estimation is necessary.

The study by Shawon et al. finds, from a review of previous literature, that ANNs are the most prevalent models when not analysing satellite imagery and vegetation indices, and CNNs were the most prevalent when including those. Within livestock farming, CNNs were also by far the most common, due to the large number of features associated with the maintenance of livestock. In-depth data collection of the livestock's health, grazing land and feed contributed to a series of large datasets that were analysed best by a CNN. The study also notes a growing tendency to use deep learning models, such as CNNs and LSTMs, since 2020. This is due to the advancements made in the general field of AI, and the prospect of moving into using more "complicated" and advanced models has pushed other researchers into utilising deep learning more.

In the period between 2020 and 2022, models utilising Support Vector Machines (SVM) and Random Forest algorithms proved popular still, despite not being deep learning models. This can be attributed to them both excelling at handling large datasets, both in features and values. Deep Neural Networks (DNN) and Recurrent Neural Networks (RNN) proved to be the most prominent deep learning models in this time frame. However, past 2022, there has been a surge of studies primarily using CNNs, and integrating them with other models such as LSTMs and Gated Recurrent Units (GRU), which fix memory loss issues inherent to CNNs and RNNs.

Recurrent Neural Networks have an output that is determined by the inputs at any given time state, as well as the inputs from the time state prior. This allows it to sequence data from the past in a time series better than most other models. Due to crop yield predictions often functioning as a time series, this gives it a low average error rate. However, by themselves, they can only classify image data where each pixel is considered to come lineally "after" the previous one, which is not how satellite data is best processed for vegetation index purposes.

Long Short-Term Memory is an advanced type of RNN that can "remember" or "forget" previous entries using memory gates, deciding what information is relevant and what can be discarded. This solves the problem of RNNs that

can lose data during their calculations. LSTMs can also be integrated into other types of deep learning networks to improve their temporal data sequencing capabilities.

Deep Neural Networks and Convolutional Neural Networks are generally similar. CNNs focus more on weight sharing and local connectivity, so they tend to work in "chunks", while DNNs focus on full connectivity, where each neuron is connected to every neuron in the succeeding layer. Where they differ is in their preferred entry data types. DNNs prefer tabular data that can be sequenced all at once, while CNNs prefer image-based data. Due to DNNs attempting to work fully connected, they can often struggle to "see the bigger picture". By assigning weights to every single data point, they can quickly get bogged down in overfitting their model and as such struggle to work with large datasets, not to mention their inability to consistently work with high-resolution imagery. CNNs perform much better in both of these areas.

Other research[12][13] also concludes that the use of deep learning, specifically in the form of a CNN, works best when integrating satellite or drone imagery into the prediction. The CNN is unparalleled when being used to evaluate vegetation indices. Smaller datasets may instead use less advanced models, such as SVM or RF, but when evaluating large datasets with many different features deep learning models are consistently the most accurate.

In a study conducted in Finland by Nevavuori et al.[14], research is done into the prediction of wheat and barley yield. Their choice of machine learning model is based on the success of CNNs in literature by Chunjing et al.[15], who detail a CNN used for image classification of high-resolution panchromatic images of crops in China, achieving a 99.6% accuracy rating, and Sa et al.[16], who use a CNN in an analysis of multispectral drone imagery to identify weeds and their spread. The model used in the study by Sa et al. was also trained to identify between sugar beet and weed, and it could also identify differences in herbicide levels present on either. They claim that a Fully Convolutional Neural Network would not perform as admirably, due to the sequential max pooling limiting the resolution, yet remain a popular choice in image segmentation.

The study by Nevavuori et al.[14] concluded CNNs using either RGB or NDVI vegetation indices on drone imagery performed to a higher accuracy on average than Artificial Neural Networks (ANN) and traditional regressional models, such as those present in the study by Georg Ruß[17].

2.5 Convolutional Neural Networks in the Prediction of Crop Growth

Convolutional Neural Networks have a history of being the industry standard for the prediction of crop output. They're robust, versatile and are easily configurable. They can take in all kinds of features from data and can identify the emerging patterns hastily. A well-rounded model will be needed for this study, due to the potential large number of features that our planned to be tested. Srivastava et al.[18] introduce a CNN model utilising 1-D convolution operations to predict winter wheat yield. The prediction is based on a comprehensive data set

that includes weather, soil, and crop phenology data. In their study, Srivastava et al. compare the performance of their proposed CNN with that of eight other machine learning models, including K-Nearest Neighbors (KNN), Random Forest, XGBoost, Regression Tree, Lasso and Ridge Regressions, Support Vector Regression (SVR), and Deep Neural Networks.

The study revealed that non-linear models, such as DNN, CNN, and XGBoost, emerge as superior to linear models, given how quickly then can detect emerging patterns and adjust their calculations quicker than them.

Notably, the research highlights performance of their proposed CNN model, surpassing all baseline models across the test years of 2017, 2018, and 2019. The CNN model demonstrates a 7% to 14% reduction in Root Mean Square Error (RMSE), a 3% to 15% decrease in Mean Absolute Error (MAE), and a 4% to 50% higher correlation coefficient when compared to the other top models like the Deep Neural Network and XGBoost. This can be attributed to the CNN's ability to capture the temporal dependencies relating to the meteorological data. The DNN and XGBoost models were found to be inconsistent at this. The DNN generally found the general patterns faster, as is the nature of deep neural networks, but had a tendency to under-predict the end result. XGBoost actually tended to over-predict slightly. All three performed relatively similarly once they had identified the patterns in the data, but overall the CNN came out with the highest accuracy.

This study also highlighted the importance of accurate and comprehensive weather components, including rainfall, temperature readings, the local relative humidity, and wind speed, for the final estimation of growth. This weather data being taken from across each year allows the models to use time frames within their calculations. This can offer insight into exactly the time of year when crop yield is most affected.

The test was carried out across Germany, using a public dataset. Despite a wide range of data, their models failed to generalise well to the eastern and northern counties of Germany, where crop yield was generally lower. All three of the CNN, DNN and XGBoost failed to generalise well to these areas, though the DNN performed slightly better while the CNN was the least generalisable of the three. Despite the lower perceived generalisability, the CNN still outperformed the other two, so such a trade off is still worth it.

Despite the usefulness of the proposed CNN, there are limitations. While it is generally transparent and adaptable, it can still be difficult to understand why it formulates certain patterns, and may limit the direct information that can be fed back to farmers other than pure predictions of their crop yield. Ultimately, it should be our goal to give as much information back to the farmers as possible, so they have time to adapt the current situation of their crops to produce a better, more efficient yield in the future. While a well made CNN may produce good accuracy, an extra metric to deduce what is the cause behind the factors influencing these results will be necessary.

CNNs are widespread[19] neural network systems and are excellent for regressional cases such as this. Compared to a fully connected network, CNNs have a reduced number of parameters to train. Within a CNN architecture,

there are three distinct types of layer: convolutional layers, pooling layers, and fully connected layers. The convolutional layers are composed of filters and feature maps. Filters act as neurones within the layer, possessing weighted inputs and generating an output value. Each feature map can be regarded as the result of a single filter. Pooling layers are employed to down-sample the feature maps from previous layers, offering generalised feature representations and mitigating overfitting. Fully-connected layers are primarily utilised at the network's conclusion to make predictions.

In CNN models, the typical structure involves the sequential arrangement of one or more convolutional layers followed by a pooling layer, and this pattern is repeated multiple times before applying fully connected layers.

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i+1)-\log(a_i+1))^2}$$

Figure 2.1: Root Mean Squared Logarithmic Error

The chosen metric will be the Root Mean Squared Logarithmic Error (RMSLE). It serves as an extension of the Mean Squared Error, particularly useful when predictions exhibit significant deviations[20]. The mathematical function employed is shown above in Figure 2.1, where 'n' represents the total number of observations in the dataset, 'p' denotes the prediction, and 'a' stands for the actual arable efficiency rating in this context. RMSLE incorporates a +1 adjustment to both actual and predicted values before applying the natural logarithm. This adjustment prevents the natural logarithm from being applied to potential zero values in the dataset. Consequently, this function remains applicable even if there are zero values in either the actual or predicted data points. It's worth noting that this function is not suitable for cases where either of the values is negative, although that should not be the case in this dataset.

2.6 Usage of Shapley Values to Evaluate a Machine Learning Model

Shapley values are a set of weights assigned to each feature in a regression model, to observe how the model utilises these features when calculating the final predictive value. By forming coalitions of features, it can calculate the Shapley value of a feature by working out the average contribution of a feature across all possible coalitions. Shapley values are a game theory concept, originally introduced by Lloyd Shapley in 1951, but have since found use in regression analysis with machine learning.

Shapley values as a concept are used for distributing costs or gains among a group of players who have collaborated, based on the input of each player. It calculates each player's contribution by considering how much the overall outcome changes when they join each possible combination of players, and then taking the average of those changes.

Kenji Yamaguchi[21] proposes that SHAP (the machine learning implemen-

tation of Shapley values) is a better measurement metric for the performance of a company than a traditional variable of measurement such as return on equity (ROE). He found there was a linear relationship between the target values and the SHAP values of the predictor variables, despite there being no such relationship between the targets and the raw predictor variables. Mariadass et al.[3] used XGBoost to anticipate crop yields in Malaysia, while also using SHAP to evaluate their implementation of the model. From this, they could determine that average temperature was the leading factor for crop growth for rice and maize, with the use of pesticides affecting the growth the least, unless it is a particularly high tonnage of pesticides, in which case they are the second biggest factor. Details like this would not be immediately obvious to human calculations, and as such is important that this study discovers these with the use of Shapley values.

Relating back to Srivastava et al.[18], they used Shapley values to provide insight into their proposed CNN model for the prediction of winter wheat yields in Germany. They use it as a post-hoc method so the results aren't used to further influence the model manually. That study found that available water capacity within the field was by far the biggest influence over the end results, while wind speed, radiation and water capacity at the point of saturation runners up.

Using Shapley values, that study was able to pinpoint why crops behaved differently depending on the time of year, and were able to understand differences within the geographical areas within Germany. They were able to determine why Shapley values around temperature were high past week 31 compared to other components in the model, as that time coincided with the flowering phase of wheat. This knowledge allowed the researches to identify other key features at this time, to know what was affecting the flowering plants the most. Higher temperatures at this point of the year could lead to a reduced photosynthesis rate, increased respiration, and a higher leaf senescence rate, eventually reducing grain numbers. This knowledge can all be passed back to the farmers to let them know the intricacies of their crops.

They also identified precipitation and water stress as large factors in the determination of crop yields. Comparing yearly rainfall statistics to the temporal field, the researches could identify when rainfall was most helpful for the growing plants, and when drought would be the most punishing. They could also correlate the precipitation data with the type of soil present, whether it had a higher concentration of sand or clay, for example. This factor would affect the water stress variable, which would go on to affect the rate the plant can draw water from the earth, and thus the rate the plant can utilise the water for growth during the most prominent growth period. Water will ultimately be one of the most important factors for the study of plant growth, regardless of the geographical area the growth takes place in and regardless of which species of plant it is. This means a greater factor of importance needs to be placed on water within any model made for this purpose, as Shapley values should be able to identify after the fact.

For a linear regression model such as this, Shapley scores excel at providing

quantitative analysis on the benefits and input of each value present in the model.

The Shapley value of the ith feature for the query point x is defined by the value function v:

$$\varphi_i(v_x) = \frac{1}{M} \sum_{S \subseteq \mathscr{M} \setminus \{i\}} \frac{v_x(S \cup \{i\}) - v_x(S)}{\frac{(M-1)!}{|S|!(M-|S|-1)!}}$$

Figure 2.2: Shapley Formula

Let M be the number of all the features, and let $\mathscr{M}$ be the set of all the features. Let $|S|$ denote the cardinality of the set S. Let $vx(S)$ be the value function of the features in the set S for the query point x.

The use of Shapley values here will provide feedback for the original formula, as knowing which values are of higher worth than others will allow for slight numerical tweaking. Even if replacing part of the formula with a neural network-generated value isn't viable, using the output to modify the other existing parts of the formula could prove useful.

2.7 Water Stress and Supply

Previous literature[18][6] consistently finds water stress and the precipitation variables to be among the most influential when estimating crop yield. This makes sense, due to all crops needing water to live and produce fruit, and needing it in greater quantities compared to other nutrients. Thus, the analysis of a crop's growth potential should entertain a large allowance of attention to the collection of water-based data. Irrigation systems are commonplace in densely packed farms, for crops that need a high amount of water, and for farms that are located in areas with little rainfall or natural groundwater.

For example, winter wheat grown in Shaanxi in China faces numerous irrigation problems. According to a report published for a project by the Asian Development Bank from 2009 to 2015[22], groundwater sources in Shaanxi declined by 5% annually between 2000 and 2007. Furthermore, rainfall had decreased 12% in the Shaanxi area between 1950 and 2000, with a higher rate estimated for the future. This, combined with an average temperature increase of 1.2 °C in the same time frame. According to the ADB, "the proportion of water available for irrigation reportedly decreased from 95% in the 1950s to 59% in 2007". This is a dramatic decrease in available water, and in a country whose population rose from about 552 million in 1950 to 1.28 billion in 2000[23], this represents a staggering potential loss in food production while needing more than double the total tonnage of produce. This water loss is not sustainable, so actions must be taken to ensure that water is used as efficiently as possible.

In their 2011 paper, Wang et al.[24] discuss the automated irrigation solutions that could reduce waste water and increase the efficiency of winter wheat crops in the Baojixia Up-Tableland Irrigation District located in Shaanxi. This

region takes the vast majority of water from the run off of the Weihe river. However, said run off has steadily decreased since the 1980s, leading to a situation where a water allocation system needed to be implemented. Wang et al. factored in groundwater as much as possible, to reduce the need for outside irrigation from the run off and reservoirs. Groundwater can be difficult to measure accurately due to the inconsistencies of aquifers and their potential depth. However, river run off and reservoirs can be measured much easier by data collection services, so despite these generally being more difficult to allocate resources to, such features provide this study with a more quantifiable way of measuring water density in the soil.

Rainfall is considered for optimisation models, but due to the scale of such crops in Shaanxi the precipitation is generally only split into two factors: wet years and dry years. For this study, precipitation will play a larger factor, as British farms have a decidedly smaller scale than Chinese farms, and generally require a smaller amount of water proportionally. Water stress for rapeseed is lower than it is for winter wheat, and Britain has easy access to water from rivers and higher rainfall on average compared to Shaanxi[25].

The model proposed by Wang et al. in Baojixia was tested under two distinct scenarios: a standard year and a drought year. The results deduced that on a standard year, excess water from the Weihe river and surrounding reservoirs could meet the water stress of the winter wheat. However, during drought years, they could only guarantee up to 80% irrigation efficiency with current estimates. This would leave the rice crops in the region, traditionally very thirsty, with less water than they required.

This shows even with efficient usage of outside water sources, crops can still go wanting for sufficient hydration. This goes on to highlight how detrimental a high water stress value for a crop can affect the output of its harvest. Plants are acutely sensitive to differences in water supply, and all water income sources need to be considered when measuring these levels. Precipitation, groundwater, irrigation systems and the sources they draw water from and general climate conditions can all be measured and quantified, and the water content of the plants post-flowering would all offer greater insight into how the crop can best be handled.

For this study, this study predicts that the Shapley values employed will show water-related characteristics to be some of the most important variables in the study. Although this research will not place manual importance on them, there are a significant number of water-based features in the dataset, so finding patterns between them and the rest of the features should be relatively easy.

2.8 Rapeseed

This study will not be analysing winter wheat or rice, it will be analysing rapeseed. Rapeseed takes up different levels of moisture at different periods of the growing stage compared to other crops. For this reason, creating a generalised machine learning algorithm for any crop would be a poor first step, as identifying a consistent pattern of water stress across different plants would be difficult.

As such, access to some factors of this experiment will be limited so as to allow the algorithm the space to formulate consistent patterns. This means limiting the geographical area and the species of crop.

Rapeseed is the second largest source of protein meal and the third largest source of vegetable oil in the world[26]. Rapeseed flowers are about two centimetres in width, and consist of four petals with four sepals. They typically have six stamens in total, and the flower is bright yellow[27].

Rapeseed pods are brown, and turn to black when ripe. The seeds can grow up to ten centimetres long[27].

The Brassica napus belongs to the family Brassicaceae. This encompasses both winter and spring rape. Rape is typically sown in the autumn and harvested in the summer. It typically flowers in the late spring or early summer.

In Europe, it is commonly grown as a break crop in four year rotations consisting of wheat and occasionally barley. Winter rape is generally used for this purpose.

The top producers of rapeseed in the world are China, Canada, and India, according to the United Nations[4].

The UK produces just about a million tonnes of rapeseed a year, which constitutes 1.4% of global production[4].

According to the British Government, winter rape is grown in 98% of all oilseed rape fields in the UK in 2023, comprising 244 thousand hectares of arable land[5]. At the time, this was a low point due to adverse weather conditions. Rapeseed yields were generally considered to be "poor" in their report, due to variable pest and disease damage in growing areas across the UK. The area that saw the largest amount of rapeseed grown was the East Midlands, where this study will mostly take place. Other high yield areas were Yorkshire and East England.

Rapeseed oil can be used for cooking or as a biofuel, but is generally more expensive than traditional diesel due to the growing and refining costs of rapeseed. Despite this, rapeseed is generally the most efficient crop for biodiesel, beating out soybeans by a lot. This is due to the density of yield, and due to the rapeseed oil's low gel point.

Rapeseed has been chosen for this experiment not only because extensive data collection services in the UK are centred around it, but also because it is a globally available crop with a wide range of applications, from cooking to fuel production. This versatility applies itself well to aiming for calculating the method to produce highest, most well-refined and efficient output. This is a plant that is produced almost anywhere in the world, and that will help to underline how important the generalisability of this study is.

3 Methodology

Here will be discussed how this study intends to carry out the experiment and the processes through which it will be performed. The nature of the original formula used for this task and the dataset will also be discussed.

3.1 Formula

Formula: KP=VI/Kw*ET0*KC

KP: The growth coefficient to be calculated and is what estimates how well a plant will harvest.

VI: The vegetation index score of the plant, taken from the satellite imagery over each of the crops and can be averaged over the different years available.

Kw: This is the water stress variable. It represents the water quota which is unfulfilled for the plant. If the plant is receiving full water quota the default value is 1.

ET0: ET0 is the rate of transpiration from the plant given a theoretical infinite water source. It is a reflection of the energy available to evaporate water.

KC: KC is a series of values specific to the crop being researched, representing partly the leaf-area index of the plant and therefore the amount of light that can be intercepted by the plant. The algorithm for determining KC is not disclosed in my data, but since it is closely related to the type of plant, KC is almost constant for all of the data entries.

These values are pre-calculated in the dataset, with only some of them coming with the exact formula for each of them. The vegetation index is obviously an exception, that has been calculated in the traditional way. The NDVI coefficient is used for this variable.

3.2 Dataset

The dataset to be used has been provided by a company that wishes to remain anonymous. It has been collected partly at the behest of the Welsh Government. The data has been collected from just under 20,000 hectares of fields across the UK: all located in either Lincolnshire, Derbyshire, or South Wales. Each subplot of soil will be one hundred square metres in size, and crops inside the subplot are grouped together. A separate variable works to show the density of plants into one square meter so this can still be taken into account. This gives a good area in which to work, and it can be assumed that the rapeseed in each plot is all identical. That is to say, if a small sample is collected from an area of the rapeseed and another sample is collected from a different area, the samples would be indistinguishable.

There will be a whole host of sensors used to take information from these plots, mainly a moisture sensor, a pH sensor and an NPK sensor. Overall, 131 features will be measured of the soil. The sensors will be hooked up to an Arduino that feeds the data into a CSV file. The entries are going to be kept

sorting into their locations (i.e Derbyshire, South Wales) for now, until they are split 80/20 for training and testing, respectively. Then the location may be discarded. This makes sure that each location receives some training, instead of randomly having very few entries being trained on and testing on a lot of entries.

Additionally, a smaller dataset from Nottinghamshire is included. This can be used in testing to see if the model is generalised enough to work for other counties, so it is not overfitted to Lincolnshire, Derbyshire and South Wales. The Nottinghamshire dataset also includes satellite imagery and will not be used during the training phase at all.

This study also has access to satellite imagery of each of these plots, taken twice a year, each year. These years are 2016, 2018 and 2019. The imagery was taken in March and July of each year, using 10m multispectral resolution. This imagery is also what was used to calculate the "VI" in the current formula being used, being the variable that relates to the vegetation index.

A dimensionality reduction will need to be performed, as much of the data being fed into the neural network will not pose much of a bearing on the outcome. The data will need to be cleaned and standardised for use in the model.

The vegetation index will be calculated using the satellite imagery.

Type	Calculation Formula	Implication	References
ExG	$R = \frac{r}{r+g+b}$ $G = \frac{g}{r+g+b}$ $B = \frac{b}{r+g+b}$ $ExG = 2 * G - R - B$	Reflects the chlorophyll content and growth status of vegetation	[20,21]
NDVI	$NDVI = \frac{nir-r}{nir+r}$	Describes the vegetation's cover and heath condition	[22,23]
RENDVI	$RENDVI = \frac{nc-r}{nc+r}$	Indicates the biomass and leaf structure of vegetation	[24,25]

Figure 3.1: Vegetation Index Types

The value wanted is NDVI.

These processes will result in six different datasets (and their testing partners) being produced.

CNN (large dataset) contains the data that has had the dimensionality reduction performed on it.

CNN (Nottinghamshire dataset) contains the data that has had the dimensionality reduction performed on it, but only contains data from Nottinghamshire. This will be used to compare smaller how the machine learning model performs with smaller, less comprehensive datasets.

CNN (without vegetation index) contains the data that has had the dimensionality reduction performed on it, but the satellite imagery aspect has been omitted. This is due to the prediction that satellite imagery may dominate the weightings, and so this can be used to test that.

CNN (Nottinghamshire without vegetation index) is the same, but limited only to the smaller Nottinghamshire dataset. This will likely be an even better indicator whether satellite imagery is over-represented at that makes up a higher percentage of the data.

CNN (no dimensions reduced) uses all data initially obtained. All 131 features will be used, and all entries for those features will be factored in. This

will include the vegetation index.

CNN (Nottinghamshire no dimensions reduced) is again the same. These two datasets will reveal how the less weighted features may still end up influencing the final prediction.

The full list of 131 features is shown in Table 1 in the appendix.

The list post-dimensionality reduction is shown in Table 2.

3.3 Vegetation Index Introduction

The vegetation index is a measure of a given plant's health and density of vegetation obtained from satellite or aerial imaging. It measures this by obtaining the red and near infrared (NIR) wavelengths that reflect off the plant in the imagery. Healthy vegetation generally reflects most NIR wavelengths while absorbing red light for photosynthesis, so the higher the NIR value, the healthier the plant. Soil, water and unhealthy plants will still reflect red light, so it can be observed quite easily the areas where healthy plants are growing. Other calculation can account for very high plant density, such as that found in jungles, and adjusting for the soil index, for crops that are quite sparse, but neither are needed for this study. The rapeseed is generally not too dense, but not enough of the soil shows through on the imagery for a calculation to adjust it to be necessary.

To calculate the vegetation index, this study has received satellite imagery that has been corrected for atmospheric distortion and cloud masking and has been cross-referenced against a global map coordinate system. This imagery is taken and used in a Python script. The Python library "rasterio" converts this kind of imagery into NumPy n-dimensional arrays and GeoJSON. Applying the formula to these rasters gives a 2D array of each pixel and a vegetation index score relating to them.

From there, the split of the fields is applied into one hundred metre squared chunks, and average out the vegetation indices across those areas, to give a series of one-dimensional values that can feed back into the database. Adding each 2D array of vegetation indices for every field into to the neural network would increase the computational time exponentially, so this is the compromise needed to be taken. This will generally invalidate minor outliers, which is not necessarily an upside. The data has been carefully cropped by the supplying company to ensure that there is no overlap with other types of vegetation, such as tree cover, hedgerows, or other produce that could be growing on farms.

Water, soil and fences will be quite easily filtered out by the red light they reflect being vastly different to that of the rapeseed. A few crops will be affected by disease such as light leaf spot, so those will generally be included with general unhealthy plants, rather than having their own categorisation. This is due to the low number of disease crops in the dataset, and the fact their output would be roughly equal to the unhealthy crops, even if that output is discarded afterwards.

3.4 Data Analysis

An initial exploratory data analysis will be needed to evaluate which of the 131 initial features should be discarded. First, the dataset will be checked for missing values. A small number of of features contained incomplete or absent values, which will be filled in using median imputation for continuous variables. Median will be used over mean due to the tendency for outliers within the data. Any values that are categorical will instead be converted with one-hot encoding.

Then, due to the similarity between many features, a correlation analysis using a matrix will be performed. Several groups of variables have strong correlations, particularly relating to the soil nutrients and the irrigation. This indicates that some features may potentially be redundant.

The correlation analysis matrix will use the Pearson coefficient, which measures the strength of a linear relationship between two variables, resulting in a value between -1 and 1. A score of -1 indicates a strong negative relationship, while 1 indicates a strong positive relationship. Values close to 0 indicate no relationship.

The resulting matrix contains correlation values for all features in pairs. Any values greater than 0.8 or less than -0.8[28] indicate a generally high correspondence between the two variables. If there is sufficient proof that a significant number of pairings produce values greater than 0.8 or less than -0.8, then that serves as justification to use a dimensionality reduction.

Weather variables tend to have high correlation coefficients, due to either water or sun being integral to a plant's development, and most weather features contribute to either how much water a plant has access to, or how much sun. Similarly, many of the soil features relating to the texture makeup are closely linked, like sand density and clay density. As one of those values increase, the other will likely decrease.

Multiple variables end up describing the same agronomic processes, meaning the addition of new variables on top of that will be redundant and not contribute to the overall accuracy much at all.

No features will be directly removed because of the results of the correlation matrix. It would be difficult to decide which feature in a pair should be excluded, and would likely come down to the personal preference or knowledge of the researcher. As such, this correlation matrix will be used purely for data analysis purposes, and to justify the use of a dimensionality reduction via random forests that will filter out the less useful features automatically.

Most features within the category of "Remote Sensing & Vegetation" in Table 1 all relate very closely to the NDVI value, leaving at least half of them redundant. Some, such as canopy cover percentage, differ enough that they still prove relevant to the calculation.

Similarly, over half of the features in "Fertiliser & Management" have very high redundancy, down to a few factors. First, fertiliser is relatively standardised in the UK, and only some subsections of Britain require specific usage of other fertiliser, such as Nottinghamshire[29] and North Wales, due to poor soil quality. However, if a wider range of arable land was to be evaluated with this model,

care should be taken to evaluate the soil and fertiliser features carefully, as other features within these groups may prove more relevant for generalisation than in this study, which only focuses on three British land areas. For example, this study will not require the use of the variable "Phosphorus Use Efficiency", but if this model were to be applied to Eastern Europe or Sub-Saharan Africa, areas of the world that feature a drastic lack of phosphorus in their soil[30], then measuring how fertiliser provides phosphorus may prove very relevant.

3.5 Random Forest

Random Forest algorithms work by constructing an ensemble of decision trees. These trees further split the data based on the thresholds of the features that reduce prediction error. Feature importance will be extracted via the mean decrease in impurity (MDI). This value measures how much each feature contributes to reducing prediction error. These features are then ranked from most impactful to least impactful. Then a series of tests using a different number of top features will be conducted to best gather how many features will be used for the final dataset. Models will be trained using 10, 20, 30, 40, 50 and 60 features and the value which strikes the balance with predictive performance without redundancy will be used.

To improve the robustness of the model and to reduce variance from random sampling, each Random Forest model will be trained multiple times using bootstrapped samples of the overall dataset. The feature importance scores will then be averaged over the runs to produce the leader board of features.

3.6 Convolutional Neural Network

The model will be built in Jupyterlab using Python and will be ready for training using the data. The AI will then run a validation dataset to test how well they have been trained. They will also then run test datasets, to give a final estimation at how accurately it can work out the plant growth coefficient.

The model will be run 10 times with the training data to gather the exact accuracy at predicting results. The algorithm will output an average prediction accuracy, which is measured between 0 and 1, with 1 being completely accurate. This research aims to achieve 90%, as the average accuracy for the formula is 85%.

3.7 Tensorflow Keras

To build this model, Tensorflow Keras will be used. Tensorflow is a framework for machine learning that is open source and thus allows the use of other APIs to integrate with it. It offers compatibility with traditional Python data libraries, such as NumPy. NumPy's native datatype, NDarrays, are automatically converted to Tensorflow tensors, for example. Tensorflow also provides integration with SHAP, although not Tensorflow 2.0.

Keras is an open-source Python library that provides a Python based interface for neural networks. Keras is not limited to Tensorflow; it can be used with PyTorch and has integrations with JAX, to name a few. Keras focuses on being user-friendly while maintaining efficient computing speed for machine learning. It contains the commonly used building blocks for neural networks, such as layers, activation functions and optimisers. It also offers tools to convert images and table-based data to usable values for the neural network.

It also supports both recurrent and convolutional neural networks. Keras is ultimately a simple but powerful library that excels at producing, training and testing convolutional neural networks, which this experiment intends to use. It also allows complicated models to run quickly and efficiently, meaning this model that is being built will be able to be run locally easily. Running this on a virtual machine in Microsoft Azure, or even locally on Jupyterlab, will be easy and quick to do.

3.8 Shapley Values Implementation

The Python library for implementing Shapley values, SHAP (SHapley Additive exPlanations), does not integrate well with modern releases of Tensorflow. This means it is necessary to roll back the version of Tensorflow installed to version 1.15.0, compared to the current version at time of writing of 2.16.1. This rollback of Tensorflow doesn't affect any of the other modules used. To accommodate Tensorflow 1.15.0, it is also necessary to rollback Python to version 3.7.0, compared to the current version at time of writing of 3.14.0. Again, this rollback provides no significant problems as the other modules adapt well to the changing versions. NumPy is used to provide graphical analysis of the Shapley values, and the values will be printed to a CSV file as well.

3.9 Code

```
df = pd.read_csv('New Schema_1737558927900.csv', header=None, engine='python', skiprows=1)

num_features = df.shape[1] - 1

x = df.iloc[:, :-1].values
y = df.iloc[:, -1].values

num_classes = len(np.unique(y))
```

Figure 3.2

This reads the CSV file to a Pandas dataframe. It also skips the first row and final column, as they are categories and labels respectively. X is the input value matrix containing all the data the network will learn from, and y is the label vector. The ".values" converts it to a NumPy array. The final line finds all the unique values there are in y, and counts how many classes there are in total.

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2)

x_train = x_train.reshape(-1, num_features, 1)
x_test = x_test.reshape(-1, num_features, 1)

train_dataset = tf.data.Dataset.from_tensor_slices((x_train, y_train)).shuffle(1000).batch(32).prefetch(tf.data.experimental.AUTOTUNE)
test_dataset = tf.data.Dataset.from_tensor_slices((x_test, y_test)).batch(32).prefetch(tf.data.experimental.AUTOTUNE)
```

Figure 3.3

Next the training and test data are split, with 80% of it going for training and 20% being used to test it. A separate group of data will be used for evaluation to prevent overfitting. Then the input data for Tensorflow models that expect three-dimensional input is reshaped, to allow for one dimensional convolutional neural networking. A Tensorflow data pipeline for the training set is then created. The model trains in batches of 32, and "prefetch" allows Tensorflow to prepare the next batch while the current batch is still being processed. This improves the performance of the neural network, optimising it further. A similar data pipeline for the testing set is also created.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d_5 (Conv1D)            (None, 48, 64)            256
_________________________________________________________________
batch_normalization (BatchNo (None, 48, 64)            256
_________________________________________________________________
max_pooling1d_5 (MaxPooling1 (None, 24, 64)            0
_________________________________________________________________
conv1d_6 (Conv1D)            (None, 24, 128)           24704
_________________________________________________________________
batch_normalization_1 (Batch (None, 24, 128)           512
_________________________________________________________________
max_pooling1d_6 (MaxPooling1 (None, 12, 128)           0
_________________________________________________________________
dropout (Dropout)            (None, 12, 128)           0
_________________________________________________________________
conv1d_7 (Conv1D)            (None, 12, 256)           98560
_________________________________________________________________
batch_normalization_2 (Batch (None, 12, 256)           1024
_________________________________________________________________
global_average_pooling1d (Gl (None, 256)               0
_________________________________________________________________
dense_10 (Dense)             (None, 128)               32896
_________________________________________________________________
dropout_1 (Dropout)          (None, 128)               0
_________________________________________________________________
dense_11 (Dense)             (None, 5528)              713112
=================================================================
Total params: 871,320
Trainable params: 870,424
Non-trainable params: 896
```

Figure 3.4

This is the summary of the model. Two dropout layers of 0.3 and 0.5 respectively are used to help prevent overfitting. ReLU as the activation function is also used due to it providing non-linearity to help the model learn complex functions. Batch normalisation stabilises the outputs of the previous layers and accelerates the training. It helps prevent covariate shift and improves conver-

gence.

```
model = models.Sequential([
    layers.Input(shape=(num_features, 1)),
    layers.Conv1D(64, kernel_size=3, activation='relu', padding='same'),
    layers.BatchNormalization(),
    layers.MaxPooling1D(pool_size=2),
    layers.Conv1D(128, kernel_size=3, activation='relu', padding='same'),
    layers.BatchNormalization(),
    layers.MaxPooling1D(pool_size=2),
    layers.Dropout(0.3),
    layers.Conv1D(256, kernel_size=3, activation='relu', padding='same'),
    layers.BatchNormalization(),
    layers.GlobalAveragePooling1D(),
    layers.Dense(128, activation='relu'),
    layers.Dropout(0.5),
    layers.Dense(num_classes, activation='softmax')
])
```

Figure 3.5

Reducing the pool size to two means only the most important features are kept in the local nodes, which severely reduces the computational cost and increases spatial invariance. Three convolutional blocks are used, due to four using a lot more computational power without providing a notable increase in accuracy. This gets up to 256 filters for deeper feature extraction.

Global Average Pooling is used instead of a Flatten layer here to encourage the model to generalise better. It takes the average of each feature map, leading to a reduced number of parameters compared to Flatten. Softmax converts outputs into probabilities across the classes. The progressive filter growth allows for more advanced features to be extracted at each layer, leading to a deeper understanding of the data quickly.

```
model.compile(optimizer='adam', loss=rmsle, metrics=['mae'])

model.fit(train_dataset, epochs=100)
```

Figure 3.6

Here the model is compiled for training. The optimiser Adam is used. This is the traditionally "best" and most efficient optimiser for neural networks. Adam combines the ideas of two predecessing optimisers; Momentum and RMSProp. Momentum incorporates the use of past gradients to update weights, instead of only using current gradients. This helps escape local minima and generalises better. RMSProp, or Root Mean Square Propagation, increases the learning rate of each weight based on how big the gradient is. Combining these two properties, Adam works very efficiently out of the box for any neural network, only falling short on very large datasets (not applicable here) and when running very low on memory, in which case a super-efficient alternative like SGD (Stochastic Gradient Descent) may be more applicable.

Adam is widely considered the standard for neural networks and is a standard optimiser in applications like Tensorflow and PyTorch. RMSLE is used when

relative error is more important than the absolute error. This means a greater weight is given to under-predicting a large final value than being slightly off for a smaller value. MAE (Mean Absolute Error) is used as an additional metric because it uses actual units to calculate the average absolute error. This is mostly used as an addendum to the RMSLE value.

```
y_pred = model.predict(x_test)
y_pred_classes = np.argmax(y_pred, axis=1)

loss, accuracy = model.evaluate(x_test, y_test)
print(f"Test Loss: {loss:.4f}")
print(f"Test Accuracy: {accuracy:.4%}")
```

Figure 3.7

Here `y_pred` is used to hold the predictions of the growth coefficients in the test dataset. `y_pred_clipped` then converts `y_pred` to all positive values by removing all negative values. This is important for how RMSLE is calculated, as it can not handle negative values. This will be used later as an axis on a graph for the results.

Loss and MAE are variables obtained from running the model on the test data. Loss holds the RMSLE value, which measures how different two probability distributions are. These are the predicted probabilities versus the true labels given in the dataset. MAE gives the average amount that the crop coefficient prediction is off by. It is less sensitive to large outliers and cares less about the difference between overestimates and underestimates. MAE isn't relative and so is often easier to comprehend initially. The output values are here for debugging purposes. If the numbers are wildly off at a glance, then the code can be adjusted to try improve it.

```
x_train_shap = x_train.reshape(-1, num_features)
x_test_shap = x_test.reshape(-1, num_features)

background = x_train_shap[np.random.choice(x_train_shap.shape[0], 100, replace=False)]

explainer = shap.KernelExplainer(lambda x: model.predict(x.reshape(-1, num_features, 1)), background)

shap_values = explainer.shap_values(x_test_shap[:100])

shap_df = pd.DataFrame(shap_values[0], columns=[f'Feature {i+1}' for i in range(num_features)])

shap_df.to_csv('shap_values1.csv', index=False)
print("SHAP values saved to shap_values1.csv")

shap.summary_plot(shap_values[0], x_test_shap[:100])
```

Figure 3.8

This is where the Shapley values are set up. SHAP expects a 2D input, so `x_train` and `x_test` must be reshaped to accommodate this. Then a background dataset is created, which represents a reference point used to simulate what the

model would predict if certain features were missing. This starting point is the basis of Shapley calculations. It tests the model on different combinations of features to see how well the model predicts based on that, then fills in the missing features with the background dataset. It takes 100 samples from `x_train` for this, which should act as a decent distribution of the training data. 100 are used because `KernelExplainer` is a computationally expensive process, so increasing that number above 200 means the program takes an exponentially long time to compute. 100 seems like a decent middle ground. The first round of testing took about 30 minutes for the Shapley calculations.

The `KernelExplainer` is the main function for computing Shapley values. It creates multiple versions of the input, each with randomly masked features, then replaces the masked features with values from the background dataset and predicts them using the neural network. It then uses weighted linear regression to estimate each features contribution. `KernelExplainer` is a slow process but can be used with basically any model of AI, so again it is an industry standard. Other models are generally faster, such as `TreeExplainer` and `LinearExplainer`, but none are quite as versatile as `KernelExplainer`, so that is the one that has been chosen.

The Shapley values for the first 100 test samples are computed and saved to `shap_values`. The output is a list of arrays, one for each class. The Shapley values are then saved to a CSV file so they can be reviews and studied further in the analysis section. A SHAP summary graph is also drawn, which gives the top features contributing to the predictions, the level of impact of the feature and whether the impact was positive or negative. These results will be expanded upon with analysis of the CSV file, but this graph gives a good abstract to immediately identify the most important features.

4 Results

4.1 Efficiency

MODEL	PERCENTAGE OF TRAINING DATA USED	AVERAGE PREDICTION ACCURACY	MOST IMPORTANT FEATURE (SHAPLEY)
CNN (large dataset)	80%	96%	Fertiliser tonnage
CNN (Nottinghamshire dataset)	80%	91%	Vegetation Index
CNN (without vegetation index)	80%	87%	Fertiliser tonnage
CNN (Nottinghamshire without vegetation index)	80%	76%	Nitrate % in soil
CNN (no dimensions reduced)	90%	95%	Water stress
CNN (Nottinghamshire no dimensions reduced)	90%	91%	Vegetation Index
Traditional Formula	-	85%	Vegetation Index

Figure 4.1: Average Efficiency

Each model trained 100 times gives these outputs. The base CNN with a dimensionality reduction executed performed the best with a 96% average accuracy. Despite the Shapley values showing that fertiliser tonnage was the most important factor in determining the accuracy, the base CNN loses almost 10% average accuracy with an increase in deviation when the vegetation index is removed. This suggests that the average accuracy is mostly affected by a few major features, rather than being a product of analysis of a great number of features.

The Nottinghamshire dataset used for testing was smaller than the combined datasets of Lincolnshire, Derbyshire, and South Wales, but still performed better than the traditional formula as long as the vegetation index was included. In 2 of the 3 Nottinghamshire datasets used, the vegetation index was the most important factor in determining the overall precision.

4.2 40 Features Selection

The datasets that had dimensionality reductions performed were done so with Random Forests. This is due to Random Forests having the highest reduction ratio without performance loss according to Widmann[31]. The Random Forests dimensionality reduction was performed in Python with the RandomForestClassifier() function. The data sets performed only marginally worse without the reduction in dimensionality, with a slightly higher standard deviation.

The Random Forest algorithm found a sweet spot at 40 total features. Any features within the top 40 had a relevancy score upwards of one standard deviation above the mean, while anything outside of the top 40 had a low score. As seen in Table 2 in the appendix, each feature is distinct, with very little overlap. Any features left out of Table 2 were either closely related to a feature

that made it in, such as a group of fertiliser features being condensed down into "total fertiliser tonnage", or were mostly irrelevant to the growth, such as "soil electrical conductivity".

A spread of topics were covered by the selected features naturally, which indicates a robust dimensionality reduction.

4.3 Shapley Graphs

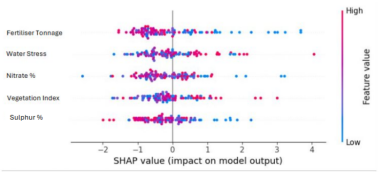

Figure 5.2: Large dataset

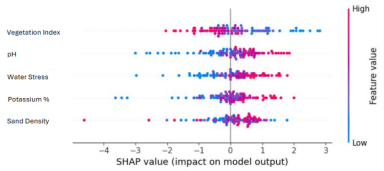

Figure 5.3: Nottinghamshire

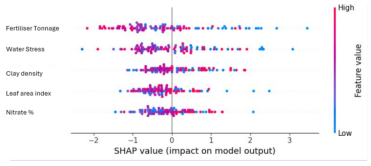

Figure 5.4: No vegetation index

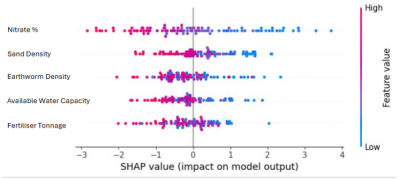

Figure 5.5: Nottinghamshire without vegetation index

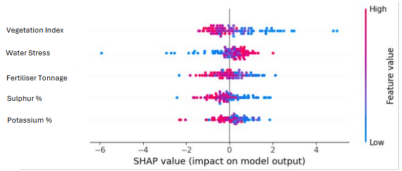

Figure 5.6: No dimensions reduced

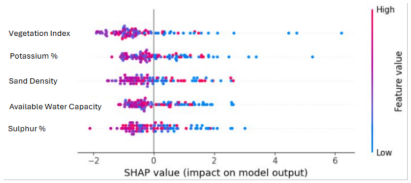

Figure 5.7: Nottinghamshire without dimensions reduced

Figure 5: SHAP value plots for different datasets, showing feature impact on model predictions

These Shapley graphs were generated in Python using the Shapley Kernel-Explainer and a Pandas dataframe. They take the five most important features as measured by the Shapley function and output how they have affected the final judgment on the CNN.

Figures 5.1, 5.3 and 5.5 use the large dataset from Lincolnshire, Derbyshire and South Wales. Figures 5.2, 5.4 and 5.6 use the Nottinghamshire dataset. Tests have been included for both datasets that feature the use of no vegetation indices, and without their dimensions reduced to feature the full 131 features. In the forms with the dimensions reduced, they use 40 features.

The red plot points represent features who have a high value, and blue points represent features who have a low value. These points exist on a colour gradient.

Where the points appear on the graph affect how exactly the model's output was influenced. The further left the dot appears, the more of a negative impact

the feature had.

For example, the Nottinghamshire dataset [Figure 5.2] has a large cluster of red dots in the positive for potassium %, while having a few blue dots strung along the negatives. This implies that having a high potassium content in the soil affects the growth coefficient, and as such the output of the crop, positively.

As some blue points stretch far to the left along the graph, that implies having a dangerously low amount of potassium in the soil will affect the calculations massively in the negative. However, only a few examples of such low figures were present, so it can be assumed that the soil in Nottinghamshire in general has a slightly above average potassium content, which lines up with what Cranfield University[29] determined, that to use the naturally acidic soil of Nottinghamshire farmers had to apply "periodic heavy dressings" of potassium.

5 Analysis

5.1 Dimensionality Reduction

The Random Forests dimensionality reduction performed did not greatly affect the results for either the large dataset or the Nottinghamshire dataset. The large dataset lost only a single percent of accuracy while the Nottinghamshire dataset wasn't any less precise at all. However, the Shapley graphs [Figure 5] show that the five most impactful features changed slightly for both datasets. The density of nitrates in the soil fell out of the top five features for the large dataset, falling to seventh place.

The reasoning for this would likely be that within the 91 features that were not used after the dimensions were reduced, a fifth of them related to nitrogen in the soil and rapeseed's ability to absorb nitrogen. Thus, with the dimensions reduced, the feature importance of just the "nitrate %" feature alone was spread amongst the other nitrogen-based features.

For the Nottinghamshire dataset, despite the accuracy not changing at all (although the standard deviation did increase slightly), pH fell completely out of relevancy for determining crop yield. Again, this is because some factors in the 91 unused features can roughly determine the pH of the soil without the need of an actual pH value. Increased phosphorus, iron and manganese values point to an increase in the acidity of the soil, while the presence of lime indicates that the soil was originally acidic but is now less so.

Large quantities of clay usually indicate a higher pH than seven, so either side of neutral on the pH scale can be roughly appropriated with knowledge of the nutrients and makeup of the ground. The increase of usage of the "sulphur %" feature aligns with the fall of the importance of the "pH" feature and the proliferation of the usage of relevant attributes that were originally disused.

Although the accuracy was almost unchanged after the reduction in dimension, the standard deviation was slightly reduced. This points to a dimensionality reduction being important not for accuracy, but for certainty and for a greatly reduced computational time.

The workload for the neural network was greatly lowered after the dimensions were reduced, while outputting almost the same accuracy (in fact, a slight increase) and being more confident in its results. This also shaved hours off the computation of the Shapley values, which only really matters if the usage of such an AI would further warrant the usage of Shapley computation or if it is only useful for this experiment.

This decrease in standard deviation after the dimensionality reduction also points to a lot of the 131 features in the dataset being little more than noise. This is further supported by the accuracy increasing slightly. While the increase in data gathered has generally proved useful for the original formula, this neural network has managed to specify which features are the most important, and as such data collection based around these highlighted features should be the main focus.

Expanding the range of features that are initially collected will not neces-

sarily make the predictions better, it may just add noise to the calculations and muddy the waters. The only features to accentuate themselves without the dimensionality reduction were those relating to the hydration of the soil and the crop. All water features were within the top 50% of features in terms of Shapley values and including them in the calculations led to an increase in accuracy across the board.

In Figure 5.1 it can be seen the features on the graph are spread out amongst the -2 to +4 range. This means some points greatly impacted the output of the model. For fertiliser tonnage, a few blue points drift far to the right; low levels of fertiliser indicating high yields of the crop. This seems counter-intuitive at first, but within the context of the rest of the plot may make sense. The general consensus of points is dotted slightly into the negative, around the -1 area. This, along with the majority of those points being high values of fertiliser tonnage, suggest the neural network believes the crops may have been over-fertilised. Thus, the low values that impact the CNN the most are actually what the CNN is recommending for the optimal fertilisation of the fields.

The points in Figure 5.1 are spread quite widely. However, after the dimensions that were reduced are added back in and the full dataset is used, it can be seen in Figure 5.5 that the groups of dots are a lot closer to neutral, mostly within the -2 to +2 range. It can be assumed this is due to the impact of the individual features being lessened as they are averaged across a larger number of total features. It has been determined that a lot of the total features don't provide much to the calculations at all, and as such are given a Shapley score close to 0. This is generally true for the Nottinghamshire dataset as well, although the Nottinghamshire without dimensions reduced graph, shown in Figure 5.6, does experience a few low-value outliers having a large positive effect on the network. These exceptions will be discussed later. In general though, the more dimensions added in to the dataset from which these AI are tested, the lower the range of Shapley values there will be.

It can be concluded from all of this that not only is a dimensionality reduction useful, it has little to no downside at all. It increases accuracy slightly, lowers the standard deviation meaning the answers are more precise, greatly reduces the computational cost of both the neural network and Shapley values and reveals which features of the soil and crops data collection efforts should be focussed on. Expanding data collection on the most important features will lead to even more accurate predictions with less noise. Along with the Shapley scores, it can also serve in grouping together related features that can indicate a trend within the crop that can be passed along as information to the farmer. These trends can sometimes be easier to identify in comparison to models tested on datasets that haven't had their dimensions reduced, however, so consider running both kinds of datasets to help discover these links. The dimensionality reduction should aim to cull at least half of the features, but it has been found that the optimal number was to end up with around 40. This gave a good spread of different areas which can be used to predict without too much overlap between them. Fewer can be considered if computational time is a factor, but anything under 20 starts dramatically falling in terms of accuracy.

5.2 Generalisation

The inclusion of the vegetation index is seen to be critical in maintaining a high accuracy when performing regressional predictions. As seen in Figure 5.1, utilising the large dataset on the base CNN model, almost 10% accuracy is lost when removing the vegetation index feature and nothing else. This, and the fact that the vegetation index is not even the most influential variable for the base dataset, implies the calculations are very top-heavy, so only a few main features are making up the bulk of the equations.

This reinforces the idea behind the analysis of the dimensionality reductions in section 5.1. Even with the reduction in features, some variables provide only small changes to the computation of the estimations. Shapley values of features not in the top five most important skew heavily towards the centre, implying that despite there being a wide range of values within those features, the other features within the calculations are being given a higher importance by the algorithm. As stated previously, using a number of features below 20 starts to dramatically decrease the average accuracy, so it can be determined that while some features may not actually have much importance according to the Shapley values, they still bear enough of an influence over the end result by offering diversification of features that they should still be factored in to future studies. 20 features offers a solid enough range of different factors without overlap, and allows the machine a greater chance to be able to adapt to generalise better. As proposed before, roughly 40 features should be present ideally, so the range of factors can be reassessed for plots of rapeseed in other areas.

Currently, from Figure 5.1, it can be seen that testing on purely the Nottinghamshire dataset offers a slightly reduced overall accuracy for the model. This implies that the model is somewhat failing to generalise properly to new data. It isn't drastic, but the failings of only including three different areas of rapeseed within the training dataset can already be seen. The Nottinghamshire dataset is decidedly the least similar of the four areas present, mostly due to it having a very low pH compared to the average, according to Cranfield University[29].

The algorithm did manage to identify the pH (and as such, the potassium) problems within the Nottinghamshire data, which shows it can generalise to a degree, but still other factors such as nitrate% and the vegetation index were generally considered more important. This led to a slight decrease in the accuracy. This would be more acceptable if the Nottinghamshire dataset was far removed in similarity to the datasets of South Wales, Lincolnshire and Derbyshire, but the truth is that all of these reside in the UK, which on a global scale offer a very similar set of results to each other.

Applying a model trained only on rapeseed grown in the UK to any other country in the world is likely going to lead to much worse generalisation problems. The fix for this is simple; increased data gathering efforts globally, or at least within the areas where this model is going to be applied. This is an obvious observation, but losing 5% accuracy by simply testing on Nottinghamshire leads this study to believe this will not generalise well globally, considering Nottinghamshire borders both Lincolnshire and Derbyshire. While Nottinghamshire is

quite different from those two due to its unique geography, it still represents the issue that outside the UK this model would fail to produce meaningful estimates, apart from areas that are very geographically similar, such as northern France and perhaps southwestern Germany.

As previously stated, the training resulted in the algorithm providing a high level of importance to only a few features. This fact in and of itself means generalisation will be harder to achieve, especially if data collection in other parts of the world differs from the methods used for the main dataset. The data for this experiment was collected quite extensively from the four provinces, and the satellite imagery used was substantial as well. Such tasks were performed quite easily in the UK, with a sizeable economy and high standard of living, with access to the world's foremost data collection technology. In poorer areas of the world, access to even satellite imagery of crops can be substandard, and data collection from local agencies will often be imperfect as well. Giving a dataset of a vastly different climate to the UK and gathered from worse data collection methods to this model would almost certainly lead to a very poor end result.

This leads to a solution for areas where data collection is a harder task. This model, while potentially not acclimatised to the new area, can still give an idea of the features that are most important to focus data collection efforts on. Many of the models prioritise the vegetation index, for example, so this suggests greater efforts should be observed in trying to obtain satellite imagery over other areas of observation. The data can then be re-fed into the model to be trained again and theoretically produce better results. In an area where data collection is decidedly more difficult, using a model such as this to prioritise where efforts to improve data collection occur is vital. This would naturally sort out the generalisation problem as well.

In relation to the original formula used by the Welsh Office, this current iteration of the formula is already far ahead in terms of generalisation. The original formula was built around and entirely localised around the South Wales valleys, the flats of Pembrokeshire and the southern third of Powys. Anything north of that within Wales failed to produce any meaningful results, and so can be assumed that it would fair just as bad outside of Wales. Granted, it achieved an 85% accuracy in South Wales, but that dropped to less than 45% within Mid Powys, roughly the exact centre of Wales. That's only a difference of between 80-160 kilometres. The fact the model can generalise to the East Midlands with a higher average accuracy than the baseline original formula is already a success.

5.3 Vegetation Index Necessity

With the vegetation index receiving such a large share of responsibility over the final estimates according to Figure 1.1, it can be concluded that satellite imagery is the most vital aspect of data collection for the prediction of plant growth. It ranked within the top five features for Shapley values with every dataset that included it in Figure 5, and for the Nottinghamshire base dataset

and the version with no dimensions reduced it was the single most important variable in determining the output.

However, this comes with a downside; using NDVI within this formula comes at the cost of making the formula way too narrow. NDVI is a poor predictor of crop growth by itself and is often specialised to the area and specific crop for which it is calculating the efficiency of growth. It requires the other factors to be present to boost them. It can be considered a multiplier for the other values' calculations, but alone will provide little chance at estimating a crop's growth coefficient. That is fine in the context of this research, but it harks back to the generalisation problem. Data collectors would need to refine the exact index values and calculations used to define them for every crop and every new piece of arable land. The researcher also has to be aware of the time of year and weather present when the satellite imagery is captured and will have to adjust their NDVI (or whatever index they choose to use) calculations for that. This ties back in to the problems faced in areas where access to this technology is limited; ideally, you would like data collected at the same time each year, with the same weather, for all the same areas that you are conducting the research on. Access to that information can prove costly.

Without access to a vegetation index score, it can be seen that the CNN on the base dataset still performs marginally better than the traditional formula, offering a 2% increase in accuracy. This would also come with the benefit of a slight increase in generalisability. This is a notable result, as the traditional formula nowadays includes a calculation for the vegetation index. As such, despite the vegetation index being given a notable Shapley value for every calculation it was included in, it is still possible to make an accurate prediction without it. This is what makes data collection techniques and crop estimating calculations in poorer areas of the world viable. Vegetation indices, while an excellent addition to a machine learning formula if present, are not fully necessary for a good estimation when other features have been measured efficiently.

87% estimation accuracy is considered "very good" by the data collection agency of the Welsh Office, and would already be at the threshold for considerations to replace the tradition formula. This is before the machine learning algorithm is specialised more towards Welsh soil, meaning it would likely adapt to the rocky soil of North Wales, the acidic soil of Mid Wales and the sandy soil of West Wales better than the current formula. Even with no vegetation indices, this neural network can outperform and out-generalise the existing formula. The NDVI calculations on top of this only compound the advantage the machine learning algorithm has. However, in Figure 5.1, it can be seen that the Nottinghamshire dataset without the vegetation indices performs significantly worse than the training dataset. This contributes to the idea that generalisation will still be a major problem without the vegetation indices, but including the indices gives an additional 15% accuracy. This suggests that when trying to adapt this formula for new, untrained land, the usage of vegetation indices may be vital for a good estimation, but for land that has already had extensive data collected on it that has been fed into training the machine, it is not as necessary.

When the vegetation index is removed from the dataset, it can be observed

that the most impactful features can change substantially. These features can often be easily identified as stand-ins for vegetation indices, as can be seen in Figures 5.1 and 5.3, where leaf area index, a static value based on the plant (in this case rapeseed), becomes the fourth most influential variable once the vegetation index is removed. This makes sense, as these features are related, due to the vegetation index being the measurement of the reflection of light off the leaves of a plant. However, other variables can be observed to move higher that are normally unrelated to vegetation indices. The density of clay becomes an important variable in Figure 5.3. Although this can lead to the result that clay density and vegetation index are directly linked, it is safer to conclude that the model simply tried to lean more on the characteristics of the soil for predictive quality after the vegetation index was removed. This then leads to a decrease in accuracy, due to the model using a starkly different calculation post-removal.

This can be seen to a larger degree with the Nottinghamshire dataset, where the vegetation index was the most influential feature. It can be seen that the five most influential features are completely shifted after the removal, as the machine tried to assemble a new system to calculate the estimates. Interestingly, it assigned many of the features a low Shapley score when the feature had a high value. This includes features that traditionally should never do this, including available water capacity. That should only provide a low Shapley score when the value for it is low or if the land was flooded, which it was not. This leads to the assumption that the machine is not calculating these predictions correctly and as such is leading to a stark drop in accuracy. 76% was the lowest accuracy obtained from the tests, so it can be seen that the machine struggled without the vegetation index with the new Nottinghamshire data. This, combined with the generalisation issues (and the stark soil-based differences Nottinghamshire offers compared to its neighbours), leads to the model underperforming against the traditional formula. Some of the variables featured in the top five of the Nottinghamshire dataset without the vegetation index are interesting, and not seen elsewhere, yet are likely not notable due to the lower overall accuracy the predictions using this dataset output. It is simply to be assumed that these features being given the importance they have been is due to miscalculations from the model.

5.4 Shapley Values

Fertiliser tonnage is represented within four of the six datasets used to test the CNN. Although this suggests that the level of fertiliser used has a direct influence on the result of the growth of the plant, it can be determined from Figures 5.1, 5.3, 5.4 and 5.5 that high fertiliser values negatively affected the result of the estimate. This conclusively reveals that the crops examined in the four areas were over fertilised. Very consistently, especially in both Figures 5.1 and 5.3, both of which pertain to the original large dataset, it can be seen that low levels of fertiliser added to the crops would greatly affect the impact on the model output.

Shapley analysis enables the identification of problems like this within the

actual maintenance and planting of the crops, offering a unique insight that can't be afforded with a simple formula, or even with machine learning by itself. This study proposes that the main significant advantage of a neural network is not the increase in accuracy, but the clarity that the analysis through forms such as Shapley values provides. The original formula had four main variables. One of those is a constant based on the crop itself and the recommended fertiliser tonnage (which is assumed to be followed), so that offers no insight. Another is the vegetation index, which also offers little insight by itself. The other two variables are based around the water intake and respiration of the plant. While these are useful, as can be seen from Figure 4, they are not of the highest importance in determining the growth coefficient. Water stress is the most important factor within Figure 5.5, and water in general appears in the top five of each dataset, but it can also be seen that factors not included in the original formula can easily be more important, and leaving them out for no good reason besides cost of data collection is unsatisfactory.

For example, the level of nitrates in the soil features heavily in three of the datasets. It is even the most influential feature in Figure 5.4. From these Shapley results, it can be observed within the large dataset that having a high amount of nitrates for the crops meant a slight advantage to their growth, while a low level would lead to a dramatic drop in their ability to produce. However, within Figure 5.4, which showcases the Nottinghamshire dataset without the use of the vegetation index, it can be seen that high levels of nitrates in the soil drastically reduce the growth coefficient predicted by the neural network, while low levels indicate the opposite. This can lead to a few conclusions that would be good starting points for an investigation into this area.

For Nottinghamshire, this discrepancy could be down to a data collection error. If the data was taken at a later date than when it was supposed to, the plants could already have processed through their critical growth stages and found no need for additional nitrates, so a high density area post-nitrate uptake would imply that the crops had too heavy of a nitrate intake, which would affect the growth negatively. If if it assumed this isn't an error with how the data was collected, a few other potential conclusions can be drawn. For one, this could imply that other factors enabling strong growth such as pH or water stress are at optimal levels, allowing the plant to take up and use nitrates in the soil a lot more efficiently, so it requires lower overall levels in the soil. It could also be concluded that the plants may be receiving non-soil sources of nitrogen, such as compost or manure. These additional factors were not included in the dataset for Nottinghamshire, so it can be assumed they were not present, but this could still warrant investigation. It can also be seen that while this low level of nitrate is not currently affecting crop yield, it may become a risk in future harvests, and so an eye should be kept on it regardless.

This analysis does not necessarily provide immediate answers, but it gives the data collectors and farmers an avenue in which investigations can be started. Having an idea of where to begin will mean galvanising an inquiry into these features becomes a lot more likely. With the traditional formula, no problem with the nitrate concentrations within the soil would be noticed, and would only

be discovered by farmer undertaking personal research on their land. With a more comprehensive and professional collection and analysis of data, it can be discovered where problems lie, and these facts can be passed back down the line to be remedied. This can lead to immediate improvement of the practices around maintaining the land, the crops, and the prevention of negative effects that could potentially inhibit the production of the harvest.

However, there needs to be awareness of the potential downsides to the use of Shapley values. While at high prediction accuracies they can effectively identify the most prominent features within the calculations, at low accuracies they will prove a lot less useful. If the neural network can not properly estimate the output of the plants, then the Shapley values will not be able to correctly assign how much a feature should be worth. It will still reveal what the machine is using as its main focus, but this will not offer valuable insight into the actual practical growth process. Instead, this can be used to help improve the model itself. Identifying what the model thinks is important is the first step to readjusting it to assign the importance into the proper areas. Ultimately this process should be used during the design phase of the model, as during testing Shapley values will not provide much understanding while the machine can't predict to a high degree of accuracy.

Shapley values also fall into a problem with computational costs. Within this study, they took by far the longest time to compute. They also required a hefty processor to run them, while the neural network itself was in comparison much more lightweight. It was not uncommon for the computation of the Shapley values to take over triple the time it took to train and test the AI model. As the datasets grew, the Shapley calculations took an exponentially longer amount of time. This could be a concern for researches without access to high level processing units, and very large datasets that need to be analysed. Dimensionality reductions greatly reduced these problems, but even at the minimum number of features recommended, a large enough dataset would require some time to calculate. This may be due to a fault with the Shapley value module used; it required a much older version of both SHAP and Python to integrate with Tensorflow Keras, as the integration package has not been updated since 2018. Newer versions of SHAP exist but have not yet been incorporated into Tensorflow. This problem can be circumvented by using another machine learning package for Python that does offer integration with SHAP.

5.5 Variance

Within Figure 5, a negative correlation between accuracy and variance of Shapley values can be observed. That is to say, the higher the overall accuracy of the predictions, the lower the range of Shapley values will be assigned to the features. This suggests that the model has learnt stable, generalisable patterns with the high accuracy results, and makes fewer highly variable decisions. This means the data is clean, and the model isn't overfitting, especially to noise. Conversely, the datasets analysed to have the lower accuracies have a great deal more Shapley variance. This suggests they are not confident in their predictions

and may be overfitting or underfitting, possibly both at the same time.

It would be overfitting by not learning to omit random noise and assigning great importance to it, vastly skewing the results. It may be underfitting by failing to actually model any meaningful interactions, and so combines the noise with regular data to form inconsistent patterns and failing to make a solid prediction. Naturally, these problems can stem from incomplete or too small a range of features, which removing the vegetation index would do. However, adding in the rest of the measured features without a dimensionality reduction would increase the noise-to-signal ratio too much, and the algorithm would fail to filter out the noise and start forming incorrect patterns. This explains the lower accuracy and higher Shapley variance in the altered datasets, Figures 5.3, 5.4, 5.5 and 5.6. Careful evaluation of the Shapley values allows the researcher to manually alter data as well to prevent or enhance the pattern finding abilities of the model.

All of this is to suggest that the accuracy results may be a bit misleading, as the high accuracy comes with a great deal of certainty and the lower accuracy comes with a greater deal of uncertainty. Thus, steps should be taken to keep the data as clean as possible, so that the algorithm can identify the relationships easier, and this will naturally lead to a much higher accuracy.

5.6 Comparison to Previous Literature

Other studies have developed algorithms to predict crop yield, but have fallen short in some areas. For example, this study by Abbasi et al.[32] produced an algorithm using SVMs and ANNs, predicted the yield of corn, then identified the most influential variables using SHAP. However, their approach was hampered by a lack of distinguishing features, and as such they had to develop over-arching features such as "Soil Fertility Index", which groups together variables like potassium density, sodium density and pH. Thus, due to so many important variables being grouped together, their Shapley values always scored the Soil Fertility Index as high. This leads to a model that may fail to generalise well, as it is standardised to American cornfields only and would fail to account for a soil patch that has a drastic lack of phosphorous, for example. The Shapley values identifying vaguely that the soil makeup is important does not help to recognise any problems that may be underlying. The CNN proposed in this research instead has the capability to identify much more specific features, only feeling the need to group them up if a correlation matrix and RF process deem it suitable. It allows farmers the knowledge of what exactly could be improved, instead of simply telling them the soil nutrition isn't good and the irrigation is a factor. The study by Abbasi et al. ended up with a high accuracy, but failed to properly identify the features that could be changed on the ground level by the farmer. Studies like this one, that don't use CNNs or other image processing models, fail to take into account the vegetation index, which consistently proves to be highly influential[2][10][33] when predicting crop yield.

Other papers[34][35] use older, simpler machine learning models, such as SVM and LR, which, while computationally cheaper, fail to truly identify trends

in large datasets and again fail to incorporate satellite data into their calculations. A CNN does not produce poor results from tabular data, while simultaneously ranking among the best algorithms for parsing visual data such as satellite imagery[36]. However, models utilising XGBoost do consistently rank high with accuracy[37][38], and feature lower computational power than a CNN. Further study comparing the performance of CNNs and XGBoost, using both vegetation indices and Shapley values, may be warranted. The CNN used in this study, combined especially with the Shapley values, can take a long time to compute, meaning the dimensionality reduction was necessary not just for cleaning the data, but to calculate the yields in an appropriate amount of time. XGBoost algorithms tend to perform a lot quicker than their neural network counterparts[39]. However, they do start breaking down in accuracy with large datasets, so a CNN striking a fine balance between number of cases and number of features will theoretically outperform XGBoost in a comprehensive study.

Other studies utilise more time sensitive data, such as weekly recordings, when calculating crop yields. The dataset in this study only had six recordings over three years, so its function as a time series is relatively null. Time series data is best handled with other algorithms, such as XGBoost, LSTMs or even RF[40]. This study did not delve into the time-sensitive nature of crop yield prediction, so a future study may wish to incorporate that if the dataset allows.

In terms of raw accuracy, the CNN proposed in this study performs admirably against others. This study by Torres-Tello and Ko[41] ended with an accuracy of 75% when utilising both a CNN and Shapley values to predict the crop growth of six different crops in an aeroponics environment. Another study, also by Torres-Tello et al.[42] achieved an 82% accuracy when predicting the yield of canola, a type of rapeseed. They used a 1D CNN model, with a focus on performance and size, so while their accuracy was not comparable to the CNN presented in this research, they shrank the computational requirements to only 6% the size of similar models. That study showed CNNs can work with comprehensive datasets and not be computationally expensive, but at the cost of some accuracy. In their case, after shrinking the CNNs cost requirements, it lost 5.7% accuracy. Determining how much accuracy to be lost is acceptable for increasing the speed of the algorithm is a balancing act to be studied in future research.

Other studies[43] that utilised non-machine learning based approaches to yield prediction also fell short in accuracy in comparison to this study. At a field level, an accuracy of 63% was achieved, while at a county level this rose to 83%, which is still short of the 96% achieved by the highly-cultivated dataset model used in this research. Non-machine learning based systems generally have a higher accuracy on a macro level, but fail to identify smaller patters and differences that neural networks specialise in finding. However, they are a lot easier to compute and study, so if only an overview of yield prediction over a large area is needed, then running a CNN with Shapley values seems unnecessary. However, for a deep dive into the individual features of a plant, the soil, irrigation or the weather, the model proposed in this research excels.

The use of Shapley values is largely novel within the field of crop yield

prediction, with only a few other studies[3][32][35][38] utilising them to any degree. Such studies do not consider the downsides to using Shapley values, such as computational cost and the extensive data cleansing required. Shapley value analysis is also hampered by the SHAP library in Python being difficult to integrate with other machine learning modules. The SHAP algorithm in this study was problematic to implement, yet the results from it proved conclusive and revealed aspects of the analysis that would be missed otherwise. Other studies fail to produce a deep dive into the exact variable that may influence the growth of a crop, and thus form correlations between that variable and other features.

Features that ranked highly in importance, as calculated by Shapley values, were roughly similar to the features rated highly in this research. For example, in the study by Pant et al.[35], the most influential features were the type of crop, pH, temperature and rainfall. In their study, rainfall was the only feature relating to water stress. Thus, apart from crop type, which wasn't a feature in this research due to it only focussing on rapeseed, the influential features in both studies were similar.

Similarly, in this study[44], which achieved an accuracy of 91% with a Random Forests algorithm, the highlighted features by the Shapley analysis were precipitation (again, the only water stress-based feature except for snow water), pH and temperature. Other research, like these two studies, failed to fully analyse the soil nutrient and texture makeup, which according to this study are quite influential in the outcome of the crop. This could be an issue with the data collection not being comprehensive enough. Expanding the range and depth of data collection will allow a more comprehensive and accurate study into the factors surrounding crop growth, and will similarly allow Shapley values to highlight what is necessary for farmers to consistently produce high yields.

6 Conclusion

In this study, a Convolutional Neural Network was used to predict the growth coefficient of winter rapeseed in the British counties of Nottinghamshire, Derbyshire, Lincolnshire and the South Wales area. This was to seek to replace the outdated formula used by the Welsh Office, aiming to achieve a higher average accuracy and generalisability, while also using Shapley values to give greater insight into the reasons behind the prediction. It has been determined that in ideal conditions, and with proper data collection techniques, the CNN could outperform the traditional formula in all metrics, yet still had generalisation problems. However, the traditional formula was so limited in scope that it could only provide accurate answers for South Wales, whereas the proposed model could generalise to other parts of the UK.

The Shapley value analysis revealed that the machine learning algorithm would routinely find patterns surrounding the water content of the fields and crop, and used the vegetation index to a great degree. This research determined these two areas of features to be the most influential while also identifying other key variables for pattern finding.

6.1 Open Issues

However, it has been discovered that dimensionality reductions are an important part of the data collection and refining process, as without them, the model performed worse than the traditional formula. A Random Forests classifier was used to cut the dimensions down from 131 features to 40 features, a reduction of 70%. For this data, is has been determined that it could be cut down further to 20 features if needed, but 40 gave a nice range of different fields of features and allowed the CNN to perform well while recognising trends. Using all features instead disrupted the pattern-seeking progress and led to overfitted results with a low accuracy. This study also determined that while leaving the vegetation index out will harm the accuracy, it is possible to still be more accurate than the initial formula without it, if only by a small margin. The vegetation index proved very useful for the determination of the growth coefficient for rapeseed, and should be used whenever possible for further study.

This study also determined there were still generalisability issues, seen with the testing on the Nottinghamshire dataset. The failure to adapt well to the change in average pH and potassium content led to a general lower accuracy measure across the board for all Nottinghamshire-based dataset tests. The limited scope for data collection will have warranted this, as only data from four areas of the UK, three of which reside in the East Midlands, was collected. This means generalisation to the rest of the UK will need to be performed on a case-by-case basis, and usage outside of the UK will produce uncertain results at best. However, this can be remedied.

6.2 Future Research

Future research in this subject should pay close attention to configuring the CNN (or perhaps replacing it, though past research[18][21] claims a CNN is best for this task) so as to increase its potential accuracy. Such research may change how the CNN is build, or how exactly the data collected is shaped and passed through the algorithm. How the vegetation index is calculated may be changed as well[8][9]. The choice of bands and exact index used may affect the end result, but special care must be taken to ensure they can be used effectively with the chosen crop[9][10].

To expand the study further, simply including a greater area globally and adapting the model so it will not underfit would naturally increase the generalisability. This would also serve to showcase the importance of certain features globally through the use of Shapley values. Identifying key variables in each geographic region in further studies will provide knowledge for the local area and help maintain the model's usefulness as it can adapt to newer areas more proficiently.

Further studies may also want to dig deeper into the data collection process, how it can be made cheaper[6] and more effective in poorer and particularly sandier areas. Sandy soil can present unique issues with how the water stress, vegetation index and soil density values are calculated, so special attention should be paid when collecting data from these kinds of areas. For poorer areas, making use of efficient and cheap equipment can help local researchers carry out these tests without the use of expensive and potentially logistically impractical equipment.

Appendix

Table 1: Full Feature Set (131 Variables)

Category	Features
Remote Sensing & Vegetation	NDVI, Canopy Cover Percentage, Fractional Vegetation Cover, Green Vegetation Index, Infrared Percentage Vegetation Index, Ratio Vegetation Index, Difference Vegetation Index, Leaf Area Index, Green Leaf Area Index, Canopy Height, Aboveground Biomass, Dry Matter Accumulation, Green Biomass, Chlorophyll Content
Plant Structure & Phenology	Plant Height, Stem Density, Branch Density, Plant Population Density, Emergence Rate, Survival Rate, Lodging Index, Canopy Closure Percentage, Senescence Percentage, Flowering Intensity, Flower Density, Pod Density, Seed Density, Reproductive Growth Ratio
Growth Timing	Days Since Sowing, Days Since Emergence, Days to Flowering, Days to Pod Formation, Days to Seed Fill, Days to Maturity, Flowering Duration, Pod Development Duration, Seed Filling Duration, Growing Degree Days, Cumulative Growing Degree Days, Vernalisation Days, Thermal Time Since Emergence, Thermal Time Since Flowering, Growth Stage Code, Crop Growth Rate, Relative Growth Rate
Soil Physical Properties	Soil Moisture Content, Root Zone Moisture, Soil Temperature, Soil pH, Soil Electrical Conductivity, Soil Organic Matter, Soil Organic Carbon, Soil Water Holding Capacity, Soil Compaction Index, Sand Density, Silt Density, Clay Density, Soil Texture Sand Fraction, Soil Texture Silt Fraction, Soil Texture Clay Fraction, Earthworm Density
Soil Nutrients	Available Nitrogen, Available Phosphorus, Available Potassium, Available Manganese, Available Sulphur, Nitrate Percentage, Potassium Percentage, Phosphorus Percentage, Sulphur Percentage, Soil Calcium Content, Soil Magnesium Content, Soil Iron Content

Category	Features
Weather & Climate	Mean Air Temperature, Maximum Air Temperature, Minimum Air Temperature, Diurnal Temperature Range, Daily Rainfall, Cumulative Rainfall, Rainfall Intensity, Rainfall Distribution Index, Relative Humidity, Vapour Pressure Deficit, Solar Radiation, Photosynthetically Active Radiation, Sunshine Duration, Cloud Cover Fraction, Wind Speed, Wind Direction, Atmospheric Pressure, Dew Point Temperature, Potential Evapotranspiration, Actual Evapotranspiration, Rate of Transpiration (ET0), Frost Days Count, Heat Stress Days Count, Cold Stress Days Count, Consecutive Dry Days, Consecutive Wet Days, Extreme Rainfall Events, Drought Severity Index
Fertiliser & Management	Total Nitrogen Applied, Total Phosphorus Applied, Total Potassium Applied, Total Sulphur Applied, Total Fertiliser Tonnage, Nitrogen Application Rate, Phosphorus Application Rate, Potassium Application Rate, Sulphur Application Rate, Fertiliser Application Frequency, Nitrogen Application Timing, Phosphorus Application Timing, Potassium Application Timing, Foliar Fertiliser Rate, Organic Fertiliser Amount, Nitrogen Use Efficiency, Phosphorus Use Efficiency, Potassium Use Efficiency
Water & Irrigation	Water Stress Index, Soil Water Deficit, Available Water Capacity, Irrigation Amount, Irrigation Frequency, Irrigation Duration, Irrigation Efficiency, Crop Water Use, Groundwater Depth
Biotic Stress	Disease Severity Index, Pest Population Density, Light Leaf Spot Incidence

Table 2: Reduced Feature Set (40 Variables)

Category	Features
Remote Sensing & Vegetation	NDVI, Canopy Cover Percentage, Leaf Area Index, Aboveground Biomass, Chlorophyll Content
Crop Phenology & Development	Days Since Sowing, Days Since Emergence, Days to Flowering, Days to Maturity, Growing Degree Days, Flowering Duration, Pod Development Duration, Seed Filling Duration, Growth Stage Code
Soil Properties	Soil Moisture Content, Soil Temperature, Soil pH, Soil Organic Carbon, Available Nitrogen, Available Phosphorus, Available Sulphur, Available Manganese, Sand Density, Clay Density, Earthworm Density, Nitrate Percentage, Potassium Percentage
Weather & Climate	Mean Air Temperature, Maximum Air Temperature, Minimum Air Temperature, Solar Radiation, Daily Rainfall, Potential Evapotranspiration, Heat Stress Days Count, Frost Days Count
Management & Stress Indicators	Total Fertiliser Tonnage, Water Stress Index, Disease Severity Index, Available Water Capacity

References

[1] United Nations, "Population — Global Issues." https://www.un.org/en/global-issues/population, 2025. Accessed June 18, 2025; covers world population trends and United Nations policy context.

[2] M. Rashid, B. S. Bari, Y. Yusup, M. A. Kamaruddin, and N. Khan, "A comprehensive review of crop yield prediction using machine learning approaches with special emphasis on palm oil yield prediction," *IEEE Access*, vol. 9, pp. 63406–63439, 2021.

[3] D. A.-L. Mariadass, E. G. Moung, M. M. Sufian, and A. Farzamnia, "Extreme gradient boosting (xgboost) regressor and shapley additive explanation for crop yield prediction in agriculture," in *2022 12th International Conference on Computer and Knowledge Engineering (ICCKE)*, pp. 219–224, 2022.

[4] Food and Agriculture Organization of the United Nations, "FAOSTAT: Crops and livestock products (QCL)." https://www.fao.org/faostat/en/data/QCL, 2024.

[5] Department for Environment, Food Rural Affairs, "Cereal and oilseed rape areas in england at 1 june 2023." https://www.gov.uk/government/statistics/cereal-and-oilseed-rape-areas-in-england/cereal-and-oilseed-rape-areas-in-england-at-1-june-2023, 2023.

[6] A. Dahane, R. Benameur, and B. Kechar, "An innovative smart and sustainable low-cost irrigation system for smallholder farmers' communities," in *2022 3rd International Conference on Embedded & Distributed Systems (EDiS)*, pp. 37–42, 2022.

[7] G. Davrazos, T. Panagiotakopoulos, S. Kotsiantis, and A. Kameas, "Iot-enabled crop recommendation in smart agriculture using machine learning," in *2023 14th International Conference on Information, Intelligence, Systems Applications (IISA)*, pp. 1–4, 2023.

[8] X. Yongqiu, W. Shenqiang, S. Pengfei, C. Xiaoqin, S. Jianlin, W. Hua, X. Zhihua, L. Xiaoming, Y. Guang, and Y. Xiaoyuan, "Ammonia emission patterns of typical planting systems in the middle and lower reaches of the yangtze river and key technologies for ammonia emission reduction," *Chinese Journal of Eco-Agriculture*, vol. 29, no. 12, pp. 1981–1989, 2021.

[9] G. R. Fae Ibrahim, A. Rasul, and H. Abdullah, "Sentinel-2 accurately estimated wheat yield in a semi-arid region compared with landsat8," *International Journal of Remote Sensing*, vol. 44, no. 13, pp. 4115–4136, 2023.

[10] J. Saha, Y. Khanna, J. Mukhopadhyay, and S. Aikat, "From supervised to unsupervised learning for land cover analysis of sentinel-2 multispectral images," in *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, pp. 1965–1968, 2020.

[11] S. M. Shawon, F. B. Ema, A. K. Mahi, F. L. Niha, and H. Zubair, "Crop yield prediction using machine learning: An extensive and systematic literature review," *Smart Agricultural Technology*, 2024.

[12] M. E. Sakka, M. Ivanovici, L. Chaâri, and J. Mothe, "A review of cnn applications in smart agriculture using multimodal data," *Sensors (Basel, Switzerland)*, vol. 25, 2025.

[13] S. Khaki, L. Wang, and S. Archontoulis, "A cnn-rnn framework for crop yield prediction," *Frontiers in Plant Science*, vol. 10, 2019.

[14] P. Nevavuori, N. G. Narra, and T. Lipping, "Crop yield prediction with deep convolutional neural networks," *Comput. Electron. Agric.*, vol. 163, 2019.

[15] Y. Chunjing, Z. Yueyao, Z. Yaxuan, and H. Liu, "Application of convolutional neural network in classification of high resolution agricultural remote sensing images," vol. 42, p. 989 – 992, 2017. Cited by: 58; All Open Access, Gold Open Access, Green Open Access.

[16] I. Sa, Z. Chen, M. Popović, R. Khanna, F. Liebisch, J. Nieto, and R. Siegwart, "weednet: Dense semantic weed classification using multispectral images and mav for smart farming," *IEEE robotics and automation letters*, vol. 3, no. 1, pp. 588–595, 2017.

[17] G. Ruß, "Data mining of agricultural yield data: A comparison of regression models," in *Advances in Data Mining. Applications and Theoretical Aspects* (P. Perner, ed.), (Berlin, Heidelberg), pp. 24–37, Springer Berlin Heidelberg, 2009.

[18] A. K. Srivastava, N. Safaei, S. Khaki, G. Lopez, W. Zeng, F. Ewert, T. Gaiser, and J. Rahimi, "Winter wheat yield prediction using convolutional neural networks from environmental and phenological data," *Scientific Reports*, vol. 12, no. 1, p. 3215, 2022.

[19] M. A. Jabed and M. A. Azmi Murad, "Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability," *Heliyon*, vol. 10, p. e40836, Dec. 2024.

[20] P. W. Khan and Y. Byun, "Genetic algorithm based optimized feature engineering and hybrid machine learning for effective energy consumption prediction," *IEEE Access*, vol. 8, pp. 96274–196286, 11 2020.

[21] K. Yamaguchi, "Intrinsic meaning of shapley values in regression," in *2020 11th International Conference on Awareness Science and Technology (iCAST)*, pp. 1–6, 2020.

[22] Asian Development Bank, "Shanxi integrated agricultural development project." Project Data Sheet and Project Completion Report, 2013–2017. Project No. 38662–013 in the People's Republic of China; includes credit and training to 66,000 farm households across 26 counties.

[23] United Nations, Department of Economic and Social Affairs, Population Division, "Population by Single Age – Both Sexes (XLSX)." Online dataset from World Population Prospects 2024, 2024. Downloaded from the UN WPP "Standard Projections" data portal.

[24] H. Wang, J.-e. Gao, C.-h. Zhao, X.-q. Xu, Y.-x. Zhang, and H. Shao, "The optimal allocation of water resources in baojixia up-tableland irrigation district based on the ecological basic flow of weihe river," in *2011 International Symposium on Water Resource and Environmental Protection*, vol. 2, pp. 1051–1054, 2011.

[25] Met Office, UK, "Monthly Precipitation Observations 1991–2020 (v1.1.0.0)." ArcGIS Hub (Climate Data Portal), 2021. 12km gridded HadUK precipitation data (mm) for 1991–2020, British National Grid.

[26] V. Heuzé, G. Tran, D. Sauvant, M. Lessire, and F. Lebas, "Rapeseed meal." https://www.feedipedia.org/node/52, 2020. Feedipedia, a programme by INRAE, CIRAD, AFZ and FAO. Last updated on July 23, 2020, 16:36.

[27] R. Snowdon, W. Lühs, and W. Friedt, *Genome Mapping and Molecular Breeding in Plants - Oilseeds*. Springer Science+Business Media, 2006.

[28] H. Akoglu, "User's guide to correlation coefficients," *Turk. J. Emerg. Med.*, vol. 18, pp. 91–93, Sept. 2018.

[29] C. University, "Landis – the land information system." https://www.landis.org.uk, 2024.

[30] C. Alewell, B. Ringeval, C. Ballabio, D. A. Robinson, P. Panagos, and P. Borrelli, "Global phosphorus shortage will be aggravated by soil erosion," *Nature Communications*, vol. 11, p. 4546, Sep 2020.

[31] M. Widmann, "Data dimensionality reduction series: Random forest." Data Science Blog, Oct. 2022. Explains the use of Random Forests for dimensionality reduction with high reduction ratio and minimal performance loss.

[32] M. Abbasi, P. Váz, J. Silva, and P. Martins, "Machine learning approaches for predicting maize biomass yield: Leveraging feature engineering and comprehensive data integration," *Sustainability*, vol. 17, no. 1, 2025.

[33] A. Khaliq, M. A. Musci, and M. Chiaberge, "Understanding effects of atmospheric variables on spectral vegetation indices derived from satellite based time series of multispectral images," in *2018 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pp. 1–4, 2018.

[34] D. Tamayo-Vera, M. Mesbah, Y. Zhang, and X. Wang, "Advanced machine learning for regional potato yield prediction: analysis of essential drivers," *npj Sustainable Agriculture*, vol. 3, p. 12, Mar. 2025.

[35] H. Pant, G. Joshi, B. Rawat, H. R. Goyal, Y. Joshi, and C. S. Bohra, "Comparative study of crop yield prediction using explainable ai and interpretable machine learning techniques," in *2025 Fifth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, pp. 1–7, 2025.

[36] A. Milioto, P. Lottes, and C. Stachniss, "Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in cnns," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2229–2235, 2018.

[37] L. B. Rananavare and S. Chitnis, "Crop yield prediction using satellite imagery," in *2023 7th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, pp. 1–6, 2023.

[38] Y. Khosravi and T. B. Ouarda, "A geographically weighted xgboost framework for pixel-level modeling of vegetation responses using multi-source earth observation data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 235, pp. 105–132, 2026.

[39] Ágnes Backhausz, E. Bognár, V. Csiszár, D. Tárkányi, and A. Zempléni, "Comparison of classical, xgboost and neural network methods for parameter estimation in epidemic processes on random graphs," *Franklin Open*, vol. 11, p. 100271, 2025.

[40] Y. Yan, Y. Wang, J. Li, J. Zhang, and X. Mo, "Crop yield time-series data prediction based on multiple hybrid machine learning models," *CoRR*, vol. abs/2502.10405, 2025.

[41] J. Torres-Tello and S.-B. Ko, "Interpretability of artificial intelligence models that use data fusion to predict yield in aeroponics," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, pp. 3331–3342, 2021.

[42] S. Valarezo-Plaza, I. J. T.-T. Member, K. D. Singh, S. J. Shirtliffe, M. I. S. Deivalakshmi, I. S.-B. K. S. Member, T. in ChiefDaniloDe StephanyValarezo-Plaza, and J.-T. StephanyValarezo-Plaza, "A novel optimized deep learning model for canola crop yield prediction on edge devices," *IEEE Transactions on AgriFood Electronics*, vol. 2, pp. 436–444, 2024.

[43] T. Su, Y. Jin, S. Wu, S. Ruan, H. Cao, H. Zhong, Y. Guo, S. Guo, H. Meng, and Y. Deng, "Study on regional rapeseed yield estimation based on data assimilation technology considering canopy photosynthesis and component succession characteristics," *Journal of Integrative Agriculture*, 2025.

[44] I. A. Cheema, M. Hanif, M. Sarwar, and M. I. Khan, "Ineterpretable machine learning for soybean yield prediction with shap-based insights," *The Journal of Animal and Plant Sciences*, 2026.