Student thesis series INES nr 664

Crop yield predictions using sequential UAV imagery and deep learning

Daniël Zegeling

2024

Department of
Physical Geography and Ecosystem Science

Lund University

Sölvegatan 12

S-223 62 Lund

Sweden

Daniël Zegeling (2024).

Crop yield predictions using sequential UAV imagery and deep learning

Master degree thesis, 30 credits in *Physical Geography and Ecosystem Science*
Department of Physical Geography and Ecosystem Science, Lund University

Level: Master of Science (MSc)

Course duration: *January* 2024 until *June* 2024

Disclaimer

This document describes work undertaken as part of a program of study at the University of Lund. All views and opinions expressed herein remain the sole responsibility of the author, and do not necessarily represent those of the institute.

Crop yield predictions using sequential UAV imagery and deep learning

Daniël Zegeling

Master thesis, 30 credits, in Physical Geography and Ecosystem Science

Supervisor 1 Per-Ola Olsson
Department of Physical Geography and Ecosystem Sciences, Lund University

Supervisor 2 Rachid Oucheikh
Department of Physical Geography and Ecosystem Sciences, Lund University

Exam committee:
Examiner 1 Torbern Tagesson
Department of Physical Geography and Ecosystem Sciences, Lund University

Examiner 2 Xueying Li
Department of Physical Geography and Ecosystem Sciences, Lund University

Acknowledgements

I want to thank my supervisors, Per-Ola Olsson and Rachid Oucheikh, for all their help throughout the project. I would also like to thank Shangharsha Thapa for his assistance with the UAV images. Furthermore, I am grateful to my friends and roommates here in Lund; their company was much appreciated during the thesis. Finally, I would like to thank everyone in the Netherlands—my friends, family, and girlfriend—for all their support.

Abstract

This study investigates the use of sequential UAV (Unmanned Airborne Vehicle) imagery and deep learning for crop yield predictions. Accurate crop yield predictions are crucial for mitigating food shortages and making informed agricultural decisions. This research uses different sequence lengths of UAV images across five wavelength bands to model winter wheat, spring wheat, and barley crop yield in South Sweden. The images were processed and calibrated to reflectance values, providing high-resolution data. CNN-LSTM (Convolutional Neural Network and Long Short-Term Memory) models were used to leverage the data's spatial and temporal dimensions. Models were trained on data from 2022 and combined data from 2022 and 2023 to explore the general applicability of the model. The study aimed to understand better how the accuracy of crop yield predictions evolves throughout the growing season. It explored the effects of varying sequence lengths on the final prediction accuracy and whether adding images to the sequence improves the accuracy. Additionally, the research tested the models' performance on barley and spring wheat to assess their generalisability to other cereals. Results indicate that prediction accuracy improves significantly as the growing season progresses, with the highest accuracy observed closer to the harvest date. However, extending the sequence length of UAV data did not consistently enhance model performance. The study also revealed that models specifically tuned to winter wheat did not perform well when applied to other crops, highlighting the need for crop-specific model training. The research contributes valuable insights into optimising UAV and deep learning technologies for agricultural applications, emphasising the need for precise and targeted data collection strategies. Such advancements are essential for improving yield predictions and aiding farmers and policymakers in making timely and informed decisions to enhance food security and sustainability.

Keywords*: crop yield predictions, machine learning, deep learning, UAV imagery, CNN-LSTM, sequential data*

Table of Contents

Abbreviations

ANN	Artificial Neural Network
CNN	Convolutional Neural Network
DL	Deep Learning
DN	Digital Number
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
ML	Machine Learning
NDVI	Normalized Difference Vegetation Index
NIR	Near-Infrared
RE	Red Edge
RF	Random Forest
RGB	Red, Green and Blue
RNN	Recurrent Neural Network
SVM	Support Vector Machine
UAV	Unmanned Aerial Vehicle
VI	Vegetation Index

Tables and Figures

1. Introduction

Ending world hunger is one of the most significant problems (Grochowska, 2014) of the 21st century. The United Nations aims to end world hunger by 2030 (FAO, 2020) and includes it in the Sustainable Development Goals as Goal 2: Zero Hunger (UN, 2017). Over 800 million people yearly suffer from hunger, and 2 billion lack adequate food access. Malnutrition due to inadequate food access is a worldwide problem and a particular challenge in Asia and Africa (FAO, 2020). Eradicating world hunger is vital for a sustainable future, and great efforts have been made in the past to achieve this goal. Ending world hunger requires sustainable and resilient global food systems (Ingram, 2011).

However, recent pandemics and the effects of climate change have put our global food systems under pressure (FAO, 2020). It is estimated that an additional 77 million people will be exposed to hunger risks by 2050 due to climate change (Janssens et al., 2020). Our food systems are vulnerable to changes in climate, which poses a significant risk to food security (Gregory et al., 2005). Agricultural shortfall is seen as one of the drivers for Global Catastrophic Risk and could potentially increase world hunger and inequality (Cernev & Fenner, 2020). Climate change has already been found to affect crop yield in some regions of the world negatively, and it is projected to reduce global crop yield by 3% to 12% by 2050 and up to 24% by 2100 (Kogo et al., 2021; Wing et al., 2021; Guntukula, 2020).

The risk posed to our food systems calls for adaptative strategies in agriculture and decision-making (Anderson et al., 2020). Crop monitoring using remote-sensed data has become popular for providing timely information on crop health and productivity and is essential for agricultural planning (Karthikeyan et al., 2020). It can help farmers make well-informed decisions regarding pesticides, fertilisers or irrigation (Abbas et al., 2020). Crop yield predictions can aid farmers and policymakers in financial and management decisions (Elavasaran & Vincent, 2020). Early season crop yield predictions are crucial for policymakers to react timely to national food shortages (Rashid et al., 2021).

Crop yield predictions have become increasingly popular in recent years, partially due to the emergence of improved Machine Learning (ML) algorithms (Van Klompenburg et al., 2020). Traditional crop yield predictions were mainly made using mathematical, mechanistic models. The crop-specific models use various input data concerning soil-specific variables and climate data (Bali & Singla, 2022). However, due to the non-linearity in agricultural systems, high-accuracy crop yield predictions can be complex to achieve with purely mechanistic models. ML can better capture these non-linear relationships and create high-accuracy predictive models (Elavasaran & Vincent, 2020). Currently, Deep Learning (DL) algorithms are becoming increasingly popular. DL is a branch of ML and can find complex patterns within data.

Traditionally, crop yield predictions are based on satellite-based remote sensing. However, in addition to increasing DL use for predictions, unmanned airborne vehicles (UAVs) have also rapidly developed. In recent years, UAVs have become cheaper and more advanced, carrying low-weight multispectral cameras. While satellite imagery can be course-grained and have lower temporal resolution, UAVs can be flown on-demand and provide cost-efficient data with high spatial, temporal, and spectral resolution (Honkavaara et al., 2013). UAV data is often used with Machine Learning to accurately predict crop yield (Nevavuori et al., 2020; Bian et al., 2022; Maimaitijiang et al., 2020; Fei et al., 2023).

Crop yield prediction using satellite imagery and DL has been done before in Sweden, but only a limited number of studies were found (Broms et al., 2023; Bouras et al., 2023). Though numerous studies combining DL and UAV data have been done before (Arroyo et al., 2017; Shammi et al., 2024), it has not been performed much in Europe. Winter wheat is a commonly used crop in Sweden, but crop yield predictions for this crop have only been done in China and the USA before (Han et al., 2020; Wang et al., 2020). By testing crop yield modelling in Sweden, its performance can be analysed in a different context. Commonly, different DL models are applied to find the one best suited to the specific research area (Van Klompenburg, 2020). Finally, more information about the accuracy of crop yield predictions during different periods of the growing season should be available to aid farmers and policymakers in crop yield monitoring and decision-making, as there is a general lack of knowledge on the accuracy throughout the entire growing season of a crop.

1.1. Research Aims

This research will focus on the accuracy of crop yield predictions throughout the growing season using a deep learning network, a CNN-LSTM. The main aim is to understand better how the accuracy of crop yield predictions develops over the growing season. Additionally, the efficacy of single UAV runs will be compared against sequences of UAV data to assess the prediction quality. The yield prediction models will be constructed for winter wheat and tested for other cereal crops to test their effectiveness on crop yield predictions in a broader context and to find if the models can be generalised to other crops.

1. *How accurately can winter wheat crop yield be predicted with sequential UAV data?*

2. *Does the accuracy of crop yield predictions increase later in the growing season?*

3. *Using sequences of bi-weekly UAV data, does the accuracy of crop yield predictions keep increasing with the length of the sequence?*

4. *Can deep learning models tuned for a specific crop be applied to similar crop species?*

2. Background

2.1. Crops

This research includes three cereals: winter wheat, spring wheat and barley. The model will be trained on winter wheat. Winter wheat is a type of wheat which is sown before winter. The wheat has to endure lower temperatures and possible snowfall during the cold months before it starts sprouting in spring (Crofts, 1989). Spring wheat, on the other hand, is sown in spring. Barley is similar to winter wheat as it is also sown before winter. All three crops are harvested in August.

2.1.1. Differences in Yield

Winter wheat has been found to outperform spring wheat in terms of average crop yield consistently. Koppel & Ingver (2008) found a difference of 2 t ha^{-1} in favour of winter wheat across different sub-species. Entz and Fowler (1991) stated an average difference of 26% in yield between the two. Stofkopf et al. (1974) found higher differences between barley and winter wheat, with an average yield difference of 40%. Yield variability not only exists between crops but also within fields and between years. Yield values fluctuate and are affected by weather patterns. Crops sown in spring are more vulnerable to weather extremes than their winter varieties. In southern Sweden, extreme weather events are increasing due to climate change, partially causing these fluctuations (Sjulgård et al., 2023).

2.1.2. Differences in Growth

During the crop growth the cereals will pass through several growth stages. Early sown crops, such as winter wheat, will pass through these stages earlier than a spring crop. Anthesis (when the flower opens and is functional) is reached 23 days earlier in winter wheat than for spring wheat, and spring wheat's grain-filling period is shorter (Ozturk et al., 2006). These slight differences in growth stages cause differences in the structure of a plant when comparing winter crops to spring crops at one moment in time. This can lead to differences in spectral reflectance between the two crop types throughout the growing season (Kuester & Spengler, 2018).

2.2. Crop Yield Predictions

2.2.1. Statistical Modelling of Crop Yield

Modern crop yield predictions have been around for well over five decades. The 1970s saw the emergence of prediction models, some of which are still used today. These models are focused on statistical regression, using environmental input variables combined with historical patterns to predict crop yield (Hanuschak, 2013). The statistical models use large amounts of historical data to find patterns in crop growth and environmental variables. Inputs such as temperature and precipitation are tracked and compared to history. The models have been considerably improved in the past decades, adding more input variables and advanced statistical methods (Basso & Liu, 2019).

Statistical crop yield models do have certain shortcomings. The models are based on historical data and assume that past relationships will hold in the future. However, this assumption of stationarity does not always hold (Lobell & Burke, 2010), and due to climate change, environmental variables, such as temperature and precipitation, are moving out of the familiar range. It is uncertain that crop yield will stick to historical relations when moving into unknown territory.

2.2.2. Remote Sensing for Crop Yield Predictions

Remote sensing, the collection of data from a distance, e.g. by satellite or UAV, has been used in agriculture and crop yield prediction for decades. Remote sensing can provide accurate data over large areas and with a fair temporal resolution (Atzberger, 2013). Some of the first applications of remote sensing in the field of crop yield predictions were by providing data on the physical characteristics of the farmlands. Remote sensing has been used to map land use and soils on a large scale (Sishodia et al., 2020). The environmental variables mapped with remote sensing could be used as input for the statistical models. Satellite sensors providing meteorological information also saw a surge in the 60s, which could be used as an input for the models.

Advancements in remote sensing technology have made it possible to follow crop growth with high temporal resolution instead of relying on statistical models and environmental variables to model the eventual yield. The quality of sensors and data availability have improved tremendously in the last decades (Rogan & Chen, 2004). Remote sensing is used more frequently in agriculture than ever (Weiss et al., 2020). Using satellite images, crop yield predictions can be made across vast areas (Bolton & Friedl, 2013). Satellite images provide spectral data, out of which spectral indices correlating with vegetation traits can be calculated. For example, the Normalized Difference Vegetation Index (NDVI) is closely related to the Leaf Area Index and the fraction of Absorbed Photosynthetically Active Radiation (Baret & Guyot, 2013). These vegetation indices can also be used to predict crop yield (Bolton & Friedl, 2013).

Though satellite imagery is rapidly improving, it does not have the sufficient spatial resolution required for precision agriculture. UAVs can provide this kind of resolution for farmers and other stakeholders. Developing cheaper and more advanced UAVs has led to a surge in UAV-based studies around precision agriculture and crop yield predictions in the last decade (Maes & Steppe, 2018). Due to the small size of the area covered by UAVs, they may be unsuitable for larger-scale research because they cannot cover larger areas, but they excel on smaller scales because of the low cost and high resolution (Kasampalis, 2018). Farmers can fly UAVs on demand, monitor their crops and make quick management decisions for their crop health (Tsouros et al., 2019). For researchers, UAVs provide unprecedented data quality.

2.3. Machine Learning for Crop Yield Prediction

Not only the input data for crop yield prediction has changed significantly in recent years. With increasing computational strength, crop yield modelling is leaning further away from the traditional crop yield prediction methods and into ML. ML used for agricultural practices has increased exponentially recently (Benos et al., 2021). ML uses large datasets and high computer strength to find patterns within the data. UAVs can provide extensive and detailed datasets with much information on crop fields. ML is ideal for discerning patterns in these large amounts of data. ML can also discern non-linear patterns far better than previous crop yield models (Chlingaryan et al., 2018) and is more adaptable to scenarios with a changing climate.

ML models can be trained in two different ways: supervised and unsupervised learning. In supervised learning, the inputs and outputs are labelled, and the model tries to find the best way to reach the output based on the given inputs. Unsupervised learning is done without labelling the data. This gives the ML methods more freedom to look for patterns within the data (Benos et al., 2021). Supervised learning is more useful when the user wants a specific output. Models can be trained to produce the desired output and find patterns in the data to reach the desired outcome.

Van Klompenburg et al. (2020) have reviewed the available literature on crop yield prediction, finding the most used ML algorithms to be Neural Networks, Linear Regression, Random Forest (RF) and Support Vector Machine (SVM). This finding was supported in a review by Benos et al. (2021), which found the most used algorithms to be Artificial Neural Networks (ANN), Ensemble Learning, SVM and Regression. Benos et al. (2021) showed that ANNs were the most accurate of the popular methods, followed by Ensemble learning, SVM and then Decision Trees and Regression. A Decision Tree model is a simplified version of the RF, consisting of only one decision tree. The RF model consists of an ensemble of Decision Trees. The most accurate model, ANN, is an overarching term for a branch of deep learning ML algorithms.

2.3.1. Deep Learning

Recently, deep learning has become increasingly popular in agricultural research. Deep learning is a sub-branch of Machine Learning that emerged first in 2006. Deep learning (DL) methods are mainly known for stacking multiple processing layers on top of each other. Non-linear processes occur in each layer, and their output is then passed on to the next layer (Vargas et al., 2017). DL algorithms excel in handling raw image data and finding patterns in images by applying these different non-linear functions (LeCun et al., 2015). DL has found many valuable applications in image processing, medicine and biometrics (Vargas et al., 2017). DL algorithms' potential has also been noticed in the agricultural field and is now the most applied method for crop yield predictions (Van Klompenburg et al., 2020). Convolutional Neural Networks (CNN) are the most popular and can also be used in combination with other DL methods, such as Long Short-Term Memory (Sun et al., 2019).

2.4. Models for Crop Yield Prediction

The previous section aimed to provide a general overview of the developments in crop yield prediction in the last decades. This chapter will focus more on the specifics of crop yield modelling. The inputs and processes used throughout the research will be elaborated on and given context by collecting information from previous research. The focus will be on papers that have used a combination of UAV imagery and DL.

2.4.1. Input Features

2.4.1.1. Spectral Reflectance

The primary input in yield forecasting is spectral reflectance or indices derived from spectral reflectance. Spectral reflectance is the percentage of the incoming light reflected from the earth's surface at different wavelengths. Sensors carried by UAVs and satellites measure reflectance in wavelength bands, capturing light over different wavelengths and registering one digital value per pixel. The different wavelength bands used in this report can be seen in Figure 1. The wavelength bands in this figure are fairly broad; the sensors on the UAV capture a smaller part of this spectrum. The reflectance of a particular surface will differ over different wavelengths. Different materials will reflect light differently: water absorbs most light in the infrared wavelength bands, soil reflects more in the mid-range of infrared, and vegetation reflects most in the near-infrared (Shahi et al., 2022). The spectral signatures (reflectance over different wavelengths) of these three different surface types can be seen in Figure 1. Spectral reflectance in the visible and near-infrared (NIR) region (400-2500nm) is especially useful in evaluating soil and crop (Scotford & Miller, 2005). The most commonly used sensors on UAVs are RGB (red, green, blue), multispectral, hyperspectral and thermal. RGB sensors are used most because of their low cost (Shahi et al., 2022). However, multispectral cameras are better suited for agricultural purposes because they have bands in the NIR and Red Edge (RE), which help estimate crop yield (Bian et al., 2022).

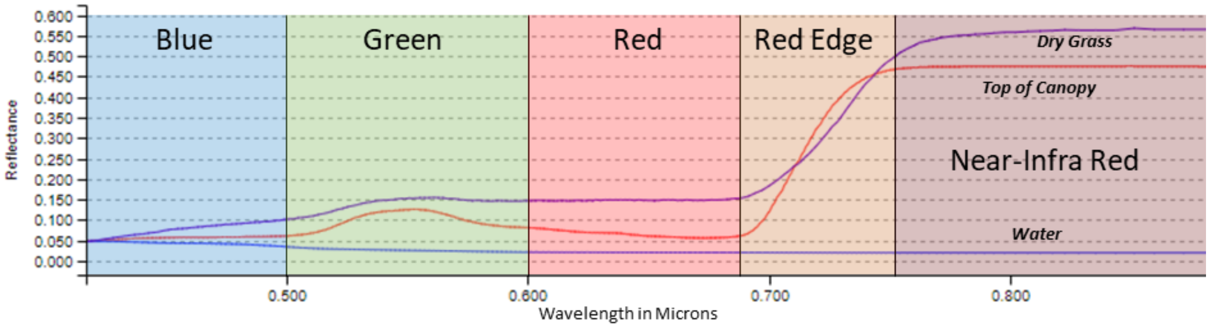


Figure 1. *Wavelength bands and spectral signatures of dry grass, healthy vegetation and water. (made in USGS Spectral Characteristics viewer)*

2.4.1.2. Vegetation Indices

Plants reflect light across different wavelengths in another way than soil or water does (Figure 1). A Vegetation Index (VI) combines different wavelength bands to reflect biophysical properties (Bian et al., 2022). VIs are often used to aid ML algorithms in finding patterns within the data (Bendig et al., 2015; Fei et al., 2023; Li et al., 2022; Maimaitijiang et al., 2020). Many different VIs have been developed, reflecting different aspects of the physical characteristics of the vegetation. Especially NDVI is among the most used VIs containing multi-spectral data (Huang et al., 2021). They focus on enhancing the ratio between infrared light, red edge and visible light. Less healthy or dense vegetation reflects less infrared light and will have a lower ratio (Tsouros et al., 2019).

2.4.2. Deep Learning Models

2.4.2.1. Convolutional Neural Network

A convolutional neural network will be used in this research. A CNN mimics a biologically inspired neuron network, like the ANN. It consists of many connected layers to form a neural network (O'Shea & Nash, 2015). A CNN differs from an ANN because it is not fully connected and uses grid-like inputs such as images and spatial data. CNNs consist of different types of layers such as: convolutional, non-linearity, pooling and fully-connected layers.

The convolutional layer slides kernels across the images to find latent patterns. This essentially means that a CNN find patterns more regionally instead of a fully connected image. The kernel operations find patterns and structure in the image; this can lead to precise edge detection and pattern recognition (Albawi et al., 2017). Padding can be applied throughout these kernel operations to maintain the same image dimensions. This process adds rows of zeros around the image to ensure the dimensions stay the same when sliding the kernel.

The non-linearity layer is also known as the activation function. The activation function applies a function to the feature map output from the convolutional layer. The activation function cuts off specific outputs and limits the output's generated values. The activation function breaks up the otherwise linear output of the convolutional layer and assists the CNN in mimicking the training data (Li et al., 2021). The pooling layer aims to downsize the spatial dimension of the input data and reduce complexity. The most common type of pooling is max-size. Here, a kernel of a specific size shifts over the image and selects the cell with the highest value; this is then kept in the size-reduced image. The highest value is selected to maintain the most important piece of information. Finally, the fully-connected layer sits at the end of the model architecture. It connects with all the neurons before to connect the entire model.

2.4.2.2. Long short-term memory

Long short-term memory (LSTM) is a type of Recurrent Neural Network (RNN). RNN is a form of DL that differs from the usual feed-forward neural network in that it can handle data sequences. Compared to other ML algorithms, the main difference of an RNN is that it uses a type of memory to help make predictions. The network remembers past values to aid in the making of predictions. The ordinary RNN can overflow with information if the time dimension gets too great. Hochreiter and Schmidhuber (1997) created the LSTM in 1997 to solve this problem.

The LSTM model uses a form of internal memory to store relevant information for long periods of time. It uses a memory block with input, output and forget gates to transfer relevant information in and out of the memory (Staudemeyer & Morris, 2019). The memory gates learn how much information they should let in or out of the memory based on the relevance of the information. The structure of an LSTM is shown in Figure 2. The processes within the model are shown in equations 1-6.

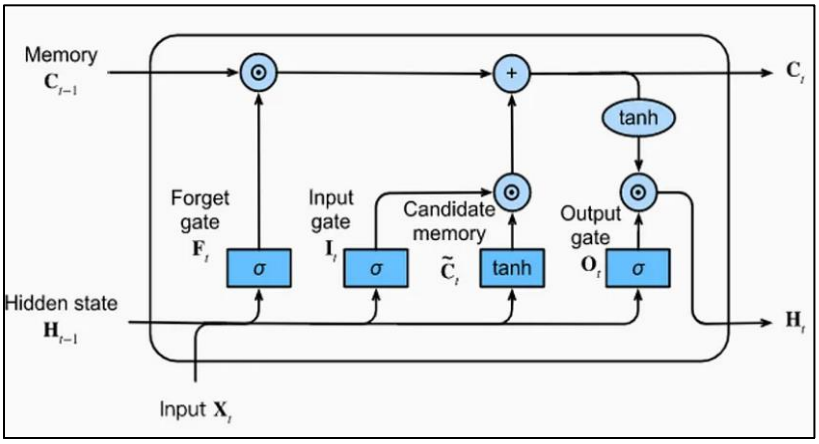


Figure 2. *The cell structure of an Long Short-Term Memory (LSTM) network. Showing its inputs outputs and internal structure. .*

$$i_t = \sigma\,(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \qquad (1)$$

$$f_t = \sigma\,(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \qquad (2)$$

$$g_t = tanh\,(W_{xg}x_t + W_{hg}h_{t-1} + b_g) \qquad (3)$$

$$o_t = \sigma\,(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \qquad (4)$$

$$c_t = f_t * c_{t-1} + i_t *g_t \qquad (5)$$

$$h_t = o_t * tanh\,(c_t) \qquad (6)$$

Where : x_t is the input at time step t , h_{t-1} is the previous hidden state, c_{t-1} is the previous cell state (memory), i_t , f_t, g_t , and o_t are the input, forget, cell, and output gates, respectively. σ is the sigmoid activation function. W and b are the weight matrices and bias vectors for each gate.

2.4.2.3. CNN-LSTM

A combination model of a CNN and LSTM combines the strength of a CNN to handle spatial data and the power of the LSTM to handle sequential data. A CNN-LSTM model has been used successfully for crop yield predictions before (Nevavuori et al., 2021; Sun et al., 2019). In the architecture of a CNN-LSTM model, the CNN typically operates as a feature extractor, transforming raw input data into a set of high-level feature maps. These feature maps encode spatial information and are then passed on to the LSTM component, which processes them sequentially to capture temporal dependencies and patterns. One common approach is to remove the final linear layer of the CNN, enabling the feature maps to serve as direct input to the LSTM, thus preserving spatial information throughout the sequential processing.

The training strategy for CNN-LSTM models can vary depending on the task and dataset characteristics. In some cases, the entire CNN-LSTM architecture is trained jointly as a single integrated model, allowing the network to learn hierarchical representations of spatial and temporal features end-to-end. Alternatively, the CNN and LSTM components can be trained separately, with pre-trained CNN weights sometimes used as initialisations for the feature extraction part. This modular training approach offers flexibility and allows for fine-tuning or transfer learning on specific tasks or domains.

3. Data and Study Area

3.1. Study Area

This study used data from five fields throughout Skåne, Sweden and of the years 2022 and 2023. Four fields are located in Alnarp, in the west part of Skåne. The final field is in the south-east of Skåne, in Löderup (Figure 3). Three of the fields were sowed with winter wheat. The other fields consisted of barley and spring wheat. Two fields were harvested in 2023, the other three in 2022. Table 1 contains an overview of the five fields and their sizes.

Table 1. *Overview of the five fields, their crop type, location, harvest year and size.*

ID	Crop Type	Location	Harvest Year	Size (Ha)
1	Barley	Alnarp	2022	1.34
2	Winter Wheat	Alnarp	2022	1.42
3	Winter Wheat	Alnarp	2022	1.34
4	Winter Wheat	Alnarp	2023	1.17
5	Spring Wheat	Löderup	2023	1.32

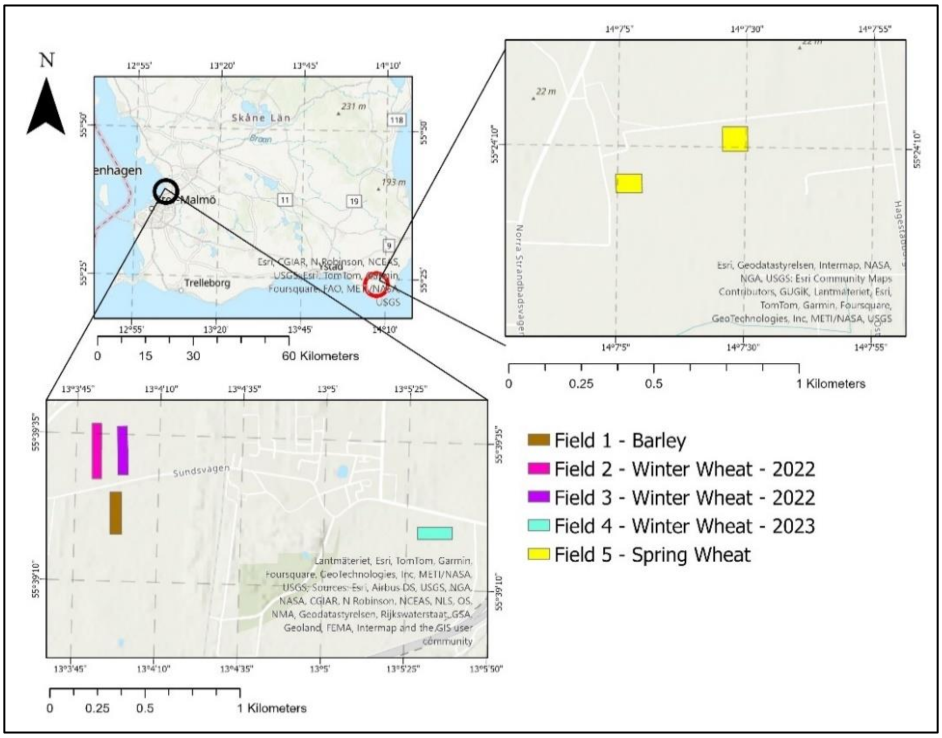


Figure 3. The location of the fields within the province of Skåne (top left). The fields are located in Alnarp (bottom left) and Löderup on the top right.

Skåne, located in the southernmost part of Sweden, experiences a temperate oceanic climate. This region is characterised by mild winters and cool summers, with relatively high precipitation evenly distributed throughout the year. The climate is influenced by its proximity to the sea, which moderates temperature extremes, creating favourable conditions for agriculture. In winter, the average minimum temperature ranges from -2°C to 1°C, while in summer, the average maximum temperature ranges from 20°C to 23°C. Skåne receives approximately 600-700 mm of precipitation annually. The growing season here is relatively long compared to other parts of Sweden.

3.2. Yield Data

Yield data was collected by a combine harvester. The harvester rides over the field and weighs the harvested crop automatically, which is then stored in a point format at a specific time interval. The combine harvester is equipped with GNSS, which records the location of the point. The Alnarp fields were harvested with the same combine harvester, recording crop yield every 5 seconds. The harvester has a swath width of 5 meters and records the roll and pitch of the vehicle next to the location and crop yield. Field 5 (Löderup) was harvested with a different machine, collecting point data every second. The machine has a swath width of 7.3 meters and measures vehicle speed and heading next to the timestamps and crop yield. The amount of points collected per field and their density can be seen in Table 2. The distribution of crop yield values of the five different fields is visible in Figure 4.

Table 2. *Statistics for collected yield point data. Showing the amount of points collected, the density of the points and the distance between collected points.*

ID	Number of Points	Point Density (points ha^{-1})	Average recording interval (m)
1	11,191	232.1	8
2	3,267	166.7	8
3	5,989	217.0	8
4	7,417	213.7	8
5	19,596	1,646.7	1

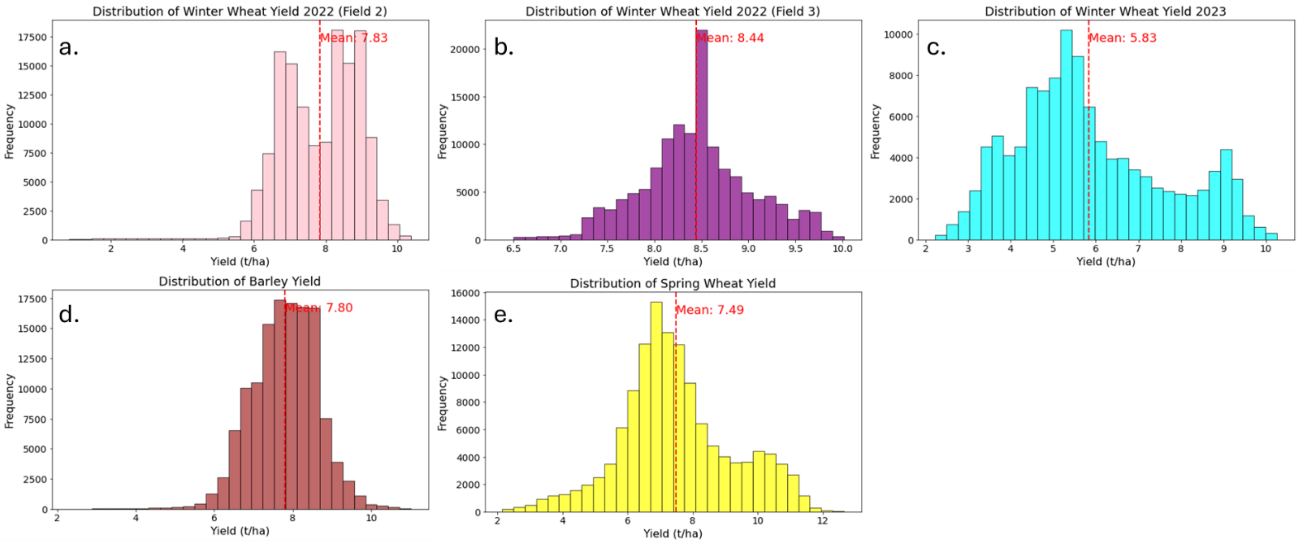


Figure 4. *Distribution of crop yield values of the collected point data for the five different fields used. Units are in t ha^{-1} and the red line indicates the mean yield for the field.*

3.3. UAV Data

UAV images were collected at different time intervals throughout the growing season of the crops in the years 2022 and 2023. All images were captured using a DJI P4 Multispectral UAV (Company). The UAV carries a camera with six sensors, five monochrome sensors for multi-spectral imaging and one RGB sensor. The ranges of the monochrome sensors are shown in Table 3. Each sensor has 2.08MP and a field of view of 62.7°. The flying height was set at 40 meters, which gives a nominal ground sample distance of 2.1 cm, and there was a 75% front and side overlap of the images. All flights were done fully automated in a grid. Flights were mainly done between 10:00 and 14:00 on clear days to avoid low solar elevation angles. The UAV was connected to the RTK-GNSS service Swedish Positioning Service (SWEPOS), giving it a centimetre accuracy. The timing of the flights varied per field. Flights started earliest in April and finished latest in August. The exact weeks of the flights per field and their sequence number are given in Table 4.

Table 3. *Wavelength band ranges of the drone sensors in nanometres.*

Sensor	Range
Blue	450 nm ± 16 nm
Green	560 nm ± 16 nm
Red	650 nm ± 16 nm
Red Edge	730 nm ± 16 nm
Near-infrared	840 nm ± 26 nm

Table 4. *Weeks of captured images for all five fields, including their sequence number in brackets.*

Field	**Week Number**						
	15	17	18	21	23	26	28
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1 – Barley	x			x	x	x	x
2 – Winter Wheat - 2022	x	x	x	x	x	x	x
3 – Winter Wheat – 2022	x	x	x	x	x	x	x
4 – Winter Wheat - 2023		x	x	x	x	x	x
5 – Spring Wheat					x	x	x

4. Methods

Several steps were undertaken to predict crop yield using UAV data and deep learning. First, the data was pre-processed to remove any flaws and noise, ensuring it was in a suitable format for the deep learning models. After pre-processing, the UAV and yield data were sorted into sequences and matched based on location. The next step involved splitting the data into training and validation datasets. Following three experimental setups, the training data was used to build multiple full-sequence and part-sequence models. These models were tested on the validation data to evaluate their accuracy. The validation data included samples for testing on the crop it was trained on, winter wheat, as well as barley and spring wheat. An overview of the methods can be found in Figure 5.

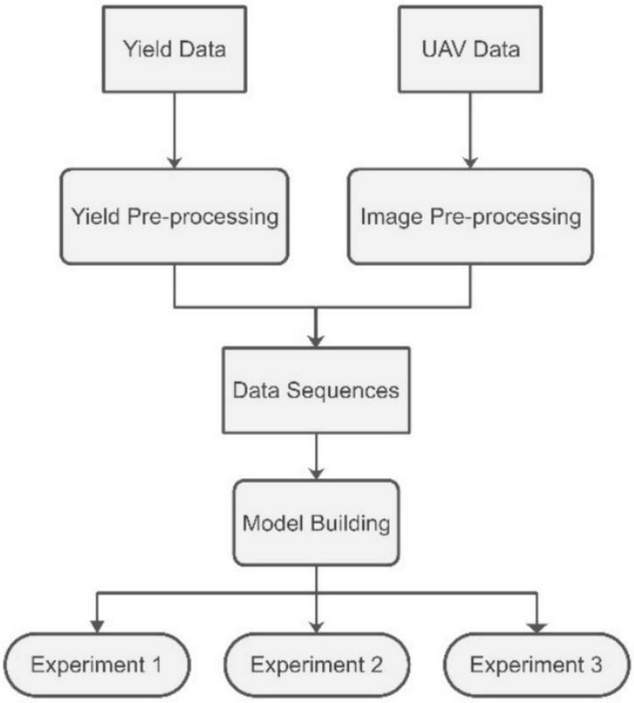


Figure 5. *Simplified diagram of the overall workflow of the methods.*

4.1. Pre-processing

The input data had to be pre-processed before being used as input into a DL model. The yield data and the UAV images needed pre-processing to prepare them as a suitable input. The workflow of the pre-processing can be seen in Figure 6.

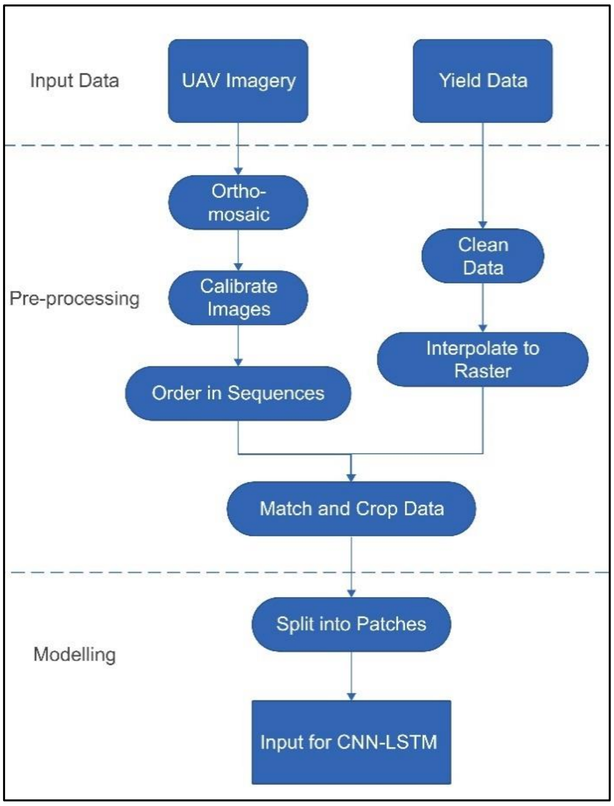


Figure 6. *Detailed workflow diagram of data pre-processing.*

4.1.1. Yield Data

Data was collected by a combine harvester, which recorded point data with GPS coordinates, a timestamp, yield data and rotation information. Harvesters can be prone to erroneous yield measurements, and careful pre-processing of yield data is required to obtain reliable results (Lyle et al., 2014). Criteria for the data cleaning were set following a paper by Hunt et al. (2019). Erroneous measurements can arise due to the harvester mechanics and GPSS measurements when these criteria are not matched. Speed and acceleration were calculated using the GPS locations and timestamps. The rotation speed was based on the roll and pitch per second. The criteria used to eliminate data are summarised in Figure 7. Data points collected outside of these criteria are likely to have larger errors. The local mean was calculated using the yield values of the three closest points. Additionally, a buffer of 15 meters was applied around the edge of the field because yield values can be significantly lower on the edges. The results of this initial cleaning can be seen in Figure 8.

- Turning angle > 0.6 rad per 30 seconds
- Speed > $8kmh^{-1}$ or < $2kmh^{-1}$
- Acceleration > $0.05kmh^{-1}$
- Yield outside global mean +- 2.5 standard deviation
- Yield outside local mean +- 2.5 standard deviation

Figure 7. *Overview of the six criteria used for cleaning the yield data points.*

After the pre-processing, the yield data was turned into a raster. The yield point data was interpolated using bilinear interpolation to a field size of 1m. Afterwards, the raster was cropped using manual set field boundaries. Finally, the raster was resampled using bilinear interpolation to a pixel size of 0.04m to match the UAV data. This led to a crop yield raster like in Figure 9.

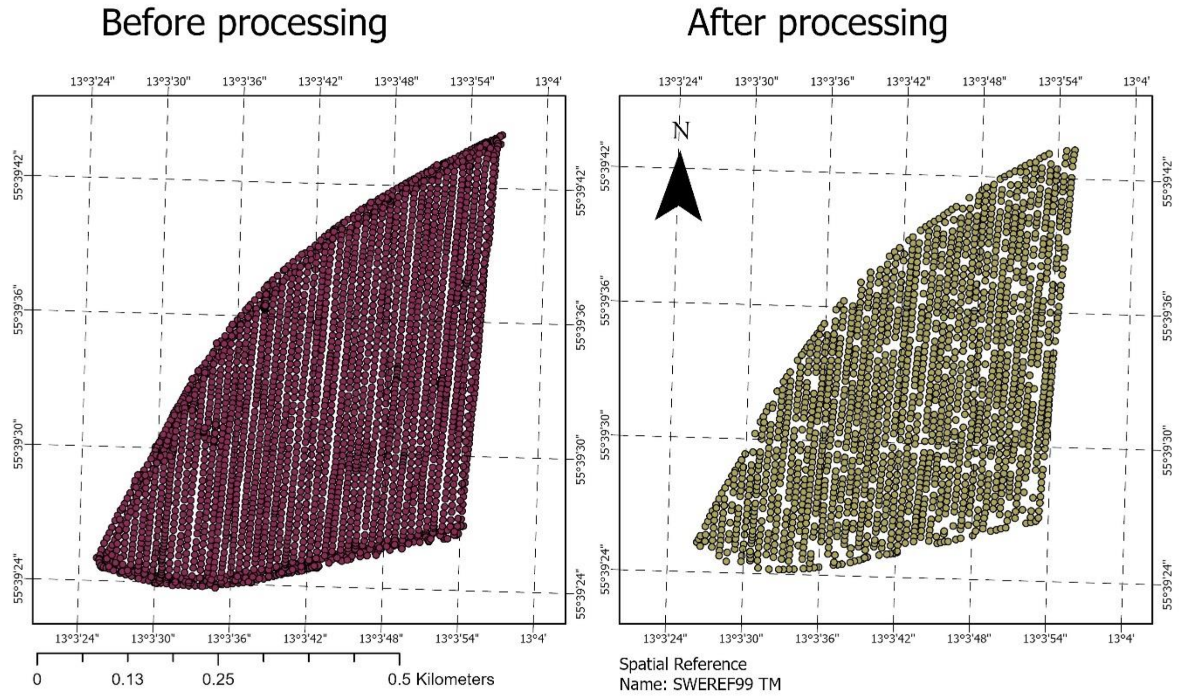


Figure 8. Overview of yield point data of field 2 (winter wheat) before and after the data cleaning processing steps.

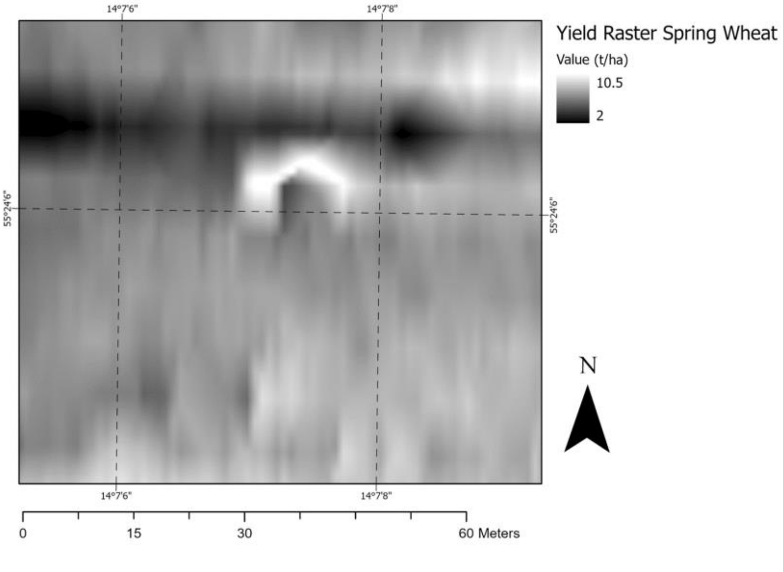


Figure 9. Example image of spring wheat yield raster (4cm resolution) after interpolation. Greyscale indicates the crop yield per pixel in t ha^{-1}

4.1.2. UAV Data

Each UAV flight over the agricultural fields involved capturing multiple overlapping images to facilitate the creation of orthophotos. Overlapping images are essential for orthophoto generation as they allow for the correction of perspective distortions inherent in individual aerial images. While single aerial images capture objects at the edges from an angle, orthophotos are produced by identifying and matching control points across overlapping images. This process eliminates the 'edge' effect, resulting in a composite image that appears to have been captured from an infinite distance directly overhead, providing a consistent nadir view throughout the entire image.

The raw UAV imagery consists of Digital Numbers (DN) calibrated for exposure compensation and irradiance normalisation following Olsson et al. (2021). This data needed to be calibrated from DN to actual surface reflectance values. This calibration is essential for accurate reflectance measurements. DN values can vary without calibration due to differences in sensor sensitivity, lighting conditions, and atmospheric effects. Calibration normalises these variations, allowing consistent comparisons across flights, dates, and fields. Calibrated data is necessary to ensure the findings are based on accurate and repeatable measurements. This is critical for the credibility and reproducibility of scientific studies.

The empirical line calibration method was employed to convert DN values to surface reflectance, assuming a linear relationship between these values (Olsson et al., 2021). Reflectance panels with known reflectance values for specific wavelength bands serve as references to establish this relationship. Three different reflectance panels were used, with varying reflectance values per wavelength band, as shown in Table 5. The central region of each panel was manually selected from the images, and the mean DN for these panels was extracted (Figure 10). These mean DN values were then paired with the panel's known reflectance values to formulate the empirical line equations. These equations were subsequently applied to all images to convert DN values to reflectance values.

Table 5. *Reflectance values per wavelength band for the used reflectance panels.*

Band	**9%**	**23%**	**44%**
Blue	0.072	0.217	0.383
Green	0.068	0.219	0.453
Red	0.077	0.223	0.439
RedEdge	0.085	0.235	0.467
NIR	0.100	0.252	0.497

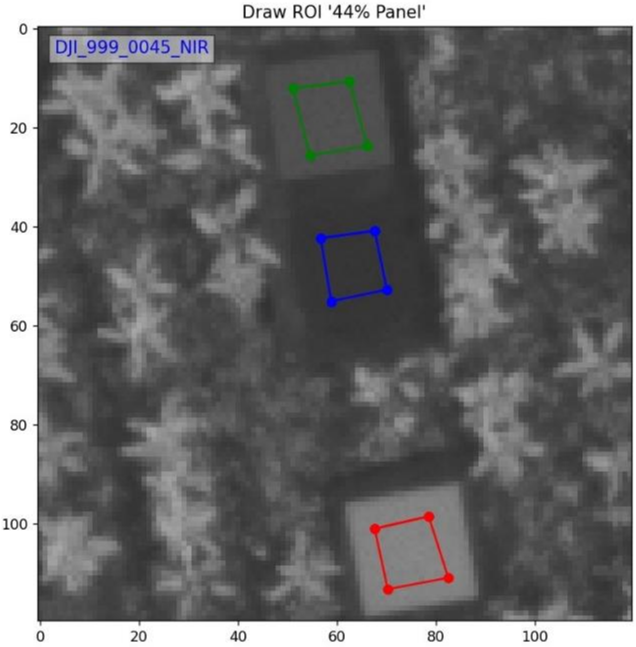


Figure 10*. Exemplary image of drawn ROIs used in the image calibration.*

Two main challenges arose during calibration: panel saturation and changing light conditions during the flight. Panel saturation occurred on the majority of the monochrome sensors. To establish a reliable linear relationship between DN values and reflectance, three panels are preferred. However, during panel saturation, one or more panels reach their maximum DN value of 255, making that panel unusable. This happened for the 44% panel on nearly all flights for the Red Edge and NIR bands. The calibration was still done with two panels, which is slightly less reliable. For the RGB monochrome sensors, more than one of the panels was saturated and rendered unusable. Instead, the RGB sensor was used to provide RGB data for all flights. Here, one panel was often unusable and the empirical line method was done with only two reference points.

The second main challenge, changing light conditions during the flight, was partially solved using images of the reflectance panels from before and after each flight. By comparing DN values before and after, it is possible to understand the light conditions throughout the flight better. Irradiance normalisation was already applied on the individual images following Olsson et al. (2021). However, changing light conditions can impact the mean DN value of the reflectance panels before and after the flight. For orthomosaics with significant differences before and after the flight, the solar irradiance data throughout the flight was compared with the solar irradiance data at the time of the calibration images. The orthomosaics with closer solar irradiance to the majority of the flight were chosen for calibration.

After calibrating, all orthomosaics were cropped to manually create field borders to match the yield data perfectly. All orthomosaics were resampled to ensure all pixels perfectly overlayed each other. Additionally, the orthomosaics had to be organised in sequences. The date stamp in the orthomosaics names was used to sort the orthomosaics of a field into order. Figures 11 and 12 show the average RGB values for all fields of the years 2022 and 2023, respectively. The empirical line method led to a negative reflectance value for some of the data. Figure 13 shows a winter wheat field in June 2023 before and after orthomosaics calibration, cropping and resampling.

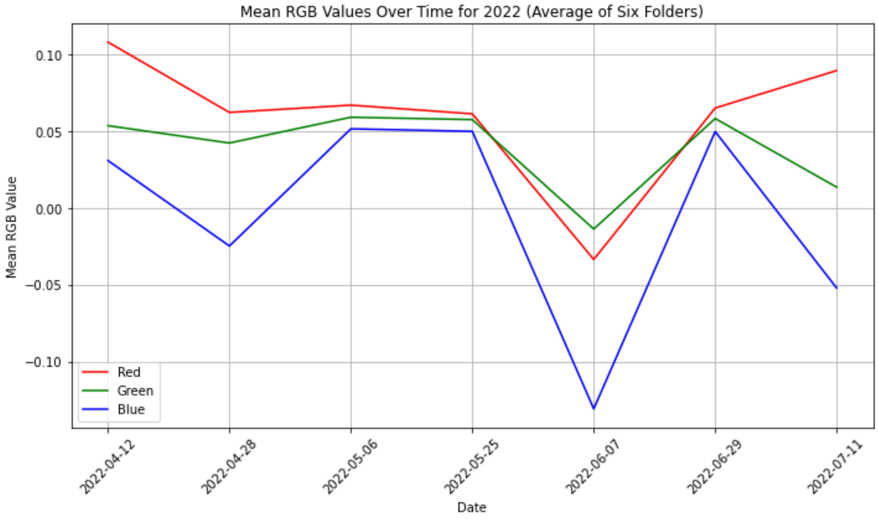


Figure 11. *Mean RGB reflectance values for calibrated fields in 2022 over time.*

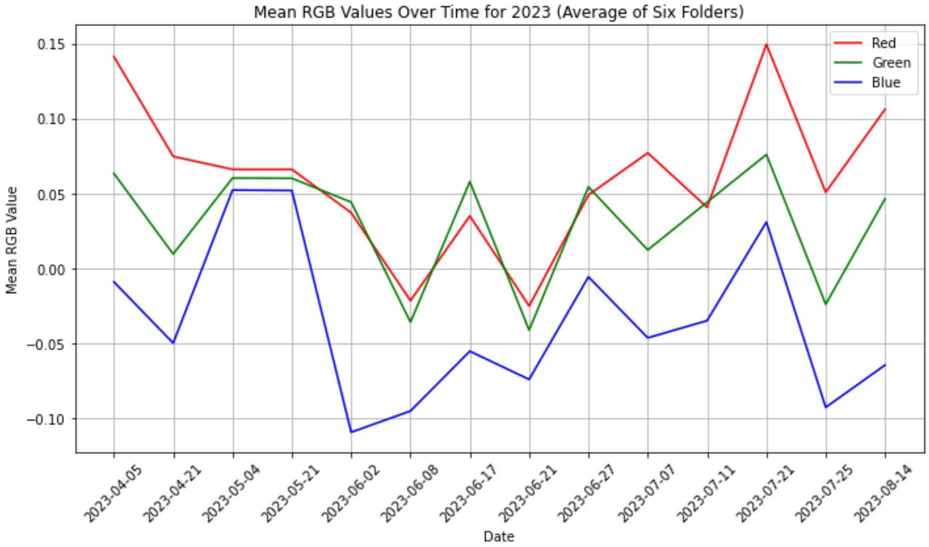


Figure 12. *Mean RGB reflectance values for calibrated fields in 2023 over time.*

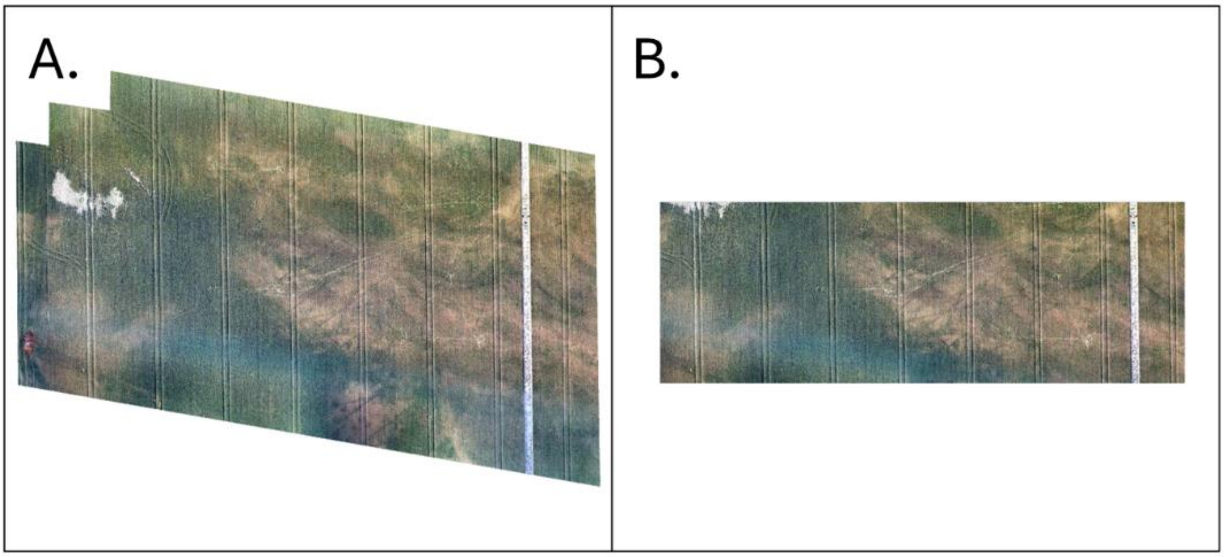


Figure 13. *Winter wheat field in 2023 (field 4) before (a) and after (b) all pre-processing steps.*

4.2. Model Architecture and Training

4.2.1. CNN-LSTM

The used model is a sequential model pairing a CNN model and an LSTM. The CNN and LSTM were tested separately and then combined into one single model. The sequential model consists of a CNN convoluting multiple layers and then passing these to the LSTM, where a dense layer will finally give a regression output. The model used time-distributed layers to ensure that the time dimension remained intact through the convolutional layers. These keep the time dimension separate to pass the data into the LSTM as a sequence. Hyperparameter tuning was done to find the optimal hyperparameters for the model. The tuned hyperparameters will be further discussed in 4.2.3.

A max-pooling layer followed each convolutional layer to decrease the size of the images passed through to the following layers and keep the computational time down. The convolutional layers used the ReLU (Rectified Linear Unit) activation function. This function is often used in machine learning and is found to improve training in neural networks. It returns positive values as normal but returns negative values as 0. The convolutional layers also received 'same' padding. By padding the borders of the images, the size does not decrease when kernels are applied to the pixels. The models were trained on the mean squared error loss metric. Additionally, the mean absolute error was given for the validation data. The selection of model architecture was based on the value of the mean absolute error of the validation data.

4.2.2. Patch Creation

The original images were too large and few to train a neural network, and the crop variability across a whole field can be substantial. The large size of the images caused the CNN to find patterns across the entire image, whilst most patterns for crop yield are down on the plant level. To decrease the model's complexity and increase the training material, the original images were split into patches. Each patch of images was trained on the mean crop yield for that patch. Working in patches also allows to model intra-field variability, as the field will be split into smaller sections. Different patch resolution sizes were tested during hyperparameter training to obtain the best model results. Patch sizes of 8, 16, and 32 pixels were tested (table 6).

4.2.3. Hyperparameter tuning

Hyperparameter tuning was applied to the image size, CNN layers, CNN units, and CNN kernel size (Table 6). Hyperparameter tuning was done based on the Mean Absolute Error (MAE). First, hyperparameter values were tested ad hoc to determine the range of values for the hyperparameter tuning. The results of which can be found in Appendix A1. During these tests, the epoch was noted from where the model no longer improved in accuracy. After these tests, the number of epochs was set at 8 for the hyperparameter tuning. Furthermore, three different values for learning rate were tested for the same model architecture. A learning rate of 0.001 was set for the hyperparameter tuning.

All values included in the hyperparameter tuning are noted in Table 6. The hyperparameter tuning was split into multiple parts to decrease the number of possible combinations. Different combinations of hyperparameters were tested using the Keras random search module. The CNN parameters were first used to compare the effect of the image size. Based on these results, the ideal architecture of the CNN was determined. This was then used to tune the ideal architecture for the LSTM. The effects of the max pooling layer were also tested and turned out in favour of applying max pooling. The results for all model runs can be found in Appendix A.

Table 6. *Tested hyperparameters for the CNN part of the CNN-LSTM model. Tested values indicate the tested possibilities for that specific hyperparameter.*

Hyperparameter	**Tested values**			
Patch size	8	16	32	
CNN Layers	1	2	3	4
CNN Units	32	64	96	128
Kernel size CNN	3	4	5	6

The architecture of the LSTM part of the model was decided by trying four different combinations of layers and unit size (Table 7). The number of units in the following layer was always lower than in the previous layer. Additionally, the effects of a dropout layer were tested for the most basic architecture.

Table 7. *Tested hyperparameter combinations for the LSTM part of the CNN-LSTM model.*

#	**Units layer 1**	**Units layer 2**	**Units layer 3**	**Dropout**
1	32	-	-	-
2	32	-	-	0.1
3	64	-	-	-
4	64	32	-	-
5	128	64	32	-

4.2.4. Training and testing data

The model was trained on two fields with winter wheat data, fields 2 and 3. These fields were both harvested in 2022. The fields were split into 80% training and 20% validating data. The percentage of training data was chosen to be high due to the presence of winter wheat data from other fields, which could be used for further testing. The other winter wheat field, field 4, was harvested in 2023 and kept separate during the first part of training. By doing this, the model's effectiveness in predicting across different years could be tested. Fields 1 (barley), 5 and 6 (spring wheat) were also kept apart for the testing phase.

4.3. Experimental Design

To test model performance and help answer the research questions, an experimental design consisting of three different experiments was set up: independent testing of a CNN and an LSTM, full sequence testing, and part sequence testing. Experiment 1, independent CNN and LSTM, was done to assess the overall performance of the CNN-LSTM model and gain insight into possible improved model accuracy by combining the two models. Experiment 2, full sequence testing, uses all available data and helps assess the reliability of part sequence testing compared to a baseline. Additionally, it could help evaluate the performance of models trained on an entire growing season of a specific crop type on a different crop type. Finally, experiment 3, part sequence testing, uses shorter sequences of 2, 3 or 4 images. It aids in exploring the effects of sequence length, the effects of image proximity to harvest data, and its usefulness in predicting yield for different cereals.

For experiment 1, the models were trained on data from 2022 and only tested on 2022, this was done to keep the total amount of model runs low and make it easier to compare to results from the other experiments. Experiments 2 and 3 were trained on data from 2022 and on a dataset consisting of 2022 and 2023. Both experiments were tested on all crops with sufficient data. The overview of all experiments and final model runs per experiment can be found in Figure 14.

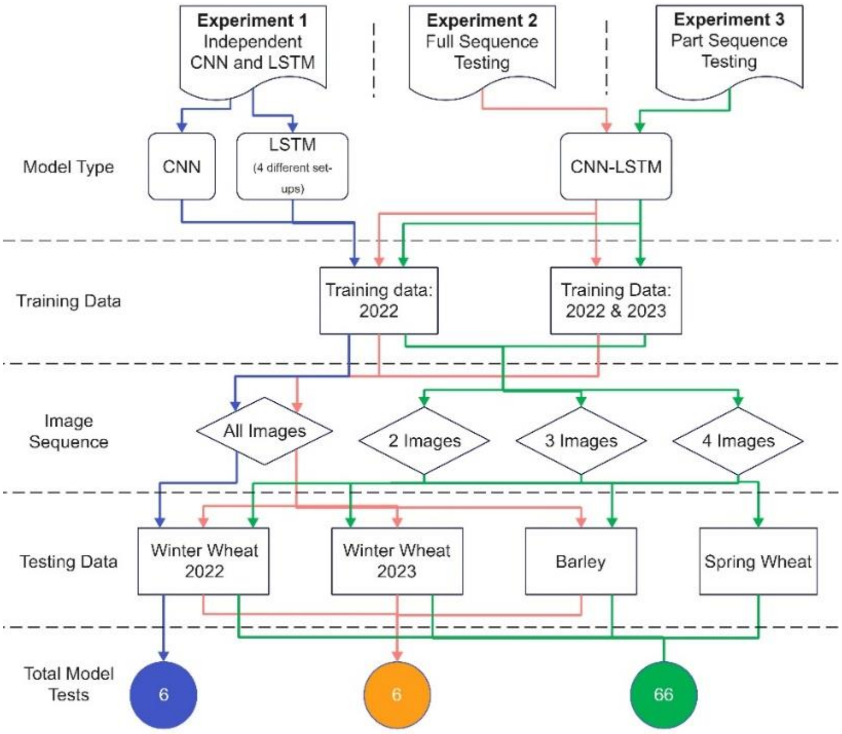


Figure 14. Flow diagram of model tests for the three experiments. Blue lines indicate experiment 1, orange lines indicate experiment 2 and green lines indicate experiment 3. Different configurations for the tests are indicates by the path of the coloured lines.

4.3.1. Experiment 1: Independent evaluation of LSTM and CNN

The LSTM model was first tested separately to assess its capabilities in finding patterns across the temporal aspect of the UAV data. Five different runs were done on a basic LSTM model. These runs varied in the grain size of the image data used. The first run modelled crop yield on single pixels. The other four runs aggregated the pixels by sliding windows of different sizes and computing the mean (Table 8). This window caused image size to decrease but also decreased outliers in the image data. Model run specifications can be seen in Table 8. Due to time constraints, the models were all tested on field 3 with only multispectral data. All the models were run for five epochs, and the validation data size was 20%. The LSTM had one layer consisting of 32 units. The input shape was (8,2), with 8 being the time dimension of the input and 2 the RE and NIR bands. Model compiling was done using the Adam Optimiser, and loss was based on the mean squared error.

A stand-alone CNN was also tested before the two models were combined. The model consists of two 2D convolutional layers with 32 and 64 units and a window of (3,3). A MaxPooling layer with a (2,2) window is between the convolutional layers. The batch size is 32, and the other model specifics are equal to the LSTM model runs. The CNN model was trained on the same winter wheat data from 2022 as the LSTM model runs.

Table 8. *Configurations for the 5 different LSTM model tests.*
Variations in windows size and batch size are indicated in the table.

LSTM Run	Window size	Batch size
Model 1	1×1	64
Model 2	2×2	32
Model 3	4×4	32
Model 4	8×8	16
Model 5	16×16	8

4.3.2. Experiment 2: Full sequence testing

A CNN-LSTM model was built using the model architecture following the hyperparameter tuning. The full sequence was used to estimate the capabilities of predicting crop yield using the 7 orthophoto-long sequences. Not all fields had the full 7-image long sequence. For these, a model was trained on a shorter sequence, omitting the vacant slots. This was the case for the winter wheat of 2023 and the barley field (2022). The spring wheat field only contained three valid timesteps and was used only in part sequence testing.

The model was initially trained on winter wheat fields 2 and 3, which contained data from the same cereal, winter wheat, and from the same year, 2022. By separating the other fields, an estimation can be done for how well the model performs on a different cereal type and on different years when using the full sequence.

4.3.3. Experiment 3: Part sequence testing

After the initial full sequence testing, the model was trained for different sequence lengths. By doing this, it is possible to estimate the model's accuracy over the growing season. By testing the model at different phases of the growing season, more information was provided on when it is possible to make a valid estimate of the eventual crop yield. Additionally, varying the sequence length may help estimate the efficiency of full sequences instead of only part sequences and could indicate the optimal time to capture images for yield predictions.

The part sequence lengths range from 2-image sequences to 4-image sequences. They were applied on all fields where possible. The sequences were applied with a stride of 1, causing a 7-image sequence to have 6-part sequence measurements. This was done to display an improvement in measurement accuracy throughout the crop's growing season. The same model architecture, derived from the hyperparameter tuning, was applied to each part sequence. The model was trained for a specific time period and then applied to the same time period on other fields. The spring wheat fields did not have image data for the first four timesteps, causing fewer part sequence testing than other fields.

4.3.4. Two-Year Training Data

Following the results of the full sequence and part sequence testing, another set-up of training data was used to test model performance. Instead of training the model only for 1 year, winter wheat data from 2 years was used. Field 1 was used as a representation of winter wheat in 2022, and 70% of field 4 in 2023 was used. More heterogeneity is introduced into the training data by including these different years. The average yield in 2023 was over 2 t ha^{-1} lower than in 2022 and had a completely different distribution (Figure 4). This variance makes the model better at dealing with different ranges of yield values but does not allow for unbiased testing of the model in different years. Nevertheless, by training it on two different years, the model should better at predicting crop yield, regardless of the year they were harvested.

4.3.5. Evaluation Metrics

Three values were calculated for all models to estimate model quality: Mean Absolute Error (MAE), weight ratio and slope. The MAE indicates the quality of each patch prediction. It is absolute because negative and positive values would cancel each other out. Secondly, the 'weight ratio' is used. For this metric, the complete yield from the field is used to calculate the predicted value and the test data. Using a ratio of predicted and actual yield, it is possible to compare the metric for different fields (equation 7).

$$Weight\ ratio = \frac{Total\ Predicted\ Yield}{Total\ Actual\ Yield} \qquad (7)$$

Finally, the slope is calculated by the line fitted through all predicted points. The ideal slope is 1, where every predicted value matches the yield's actual value. Lower values closer to 0 indicate that the model overpredicts for lower yield values and underpredicts for higher yield values. The range of model predictions is lower, with a lower slope. A negative slope indicates a very low accuracy model performance, it predicts actual low yield values to be higher than the actual high yield values.

5. Results

5.1. Model Architecture

During hyperparameter tuning, the results for 8×8 patches were consistently better than for 16×16 or 32×32 patches, frequently scoring a mean absolute error below 0.4. (See Appendix B for an overview of all tested hyperparameter combinations). Thus, a patch size of 8×8 was used for all further experiments. The number of layers of the model did not greatly impact the model's performance, with more layered models scoring only slightly higher than single-layer models. An architecture with three layers was chosen, with units and kernel size decreasing for further layers. An overview of the CNN used in the CNN-LSTM is shown in Figure 15.

The best LSTM architecture was a simple 1 layer model with 32 or 64 units (Appendix A). A single-layer model with 32 units was chosen to reduce model complexity and required computational power. Additionally, dropout did not impact the model's accuracy but was chosen to be kept in the structure. Following these results, the final model architecture was set to the one in Figure 15.

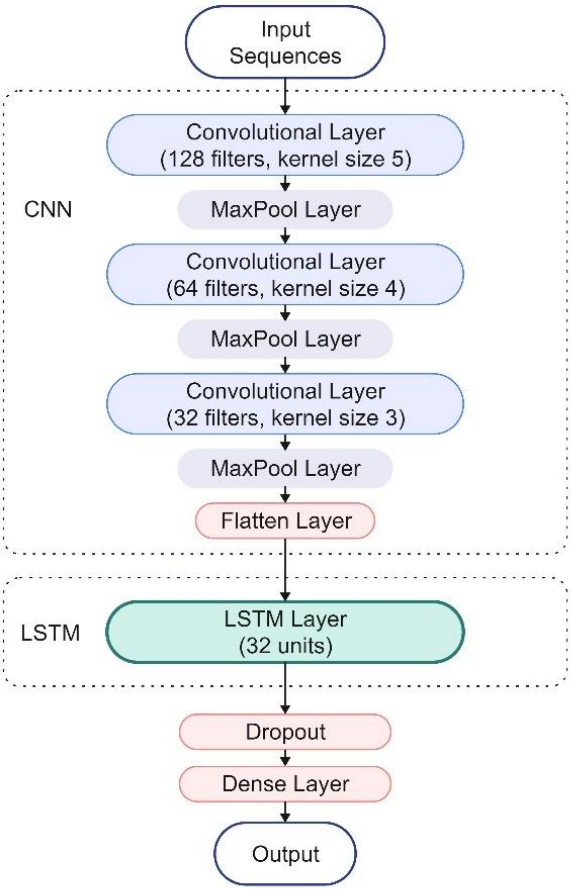

Input Sequences
CNN
Convolutional Layer (128 filters, kernel size 5)
MaxPool Layer
Convolutional Layer (64 filters, kernel size 4)
MaxPool Layer
Convolutional Layer (32 filters, kernel size 3)
MaxPool Layer
Flatten Layer
LSTM
LSTM Layer (32 units)
Dropout
Dense Layer
Output

Figure 15. *CNN-LSTM model architecture after hyperparameter tuning. Figure indicates the eventual amount of layers (3) and the different filters amounts, kernel sizes and LSTM units used.*

5.2. Experiment 1: LSTM and CNN

When testing the LSTM and CNN models separately, the CNN without the temporal aspect performed better than all LSTM models (Table 9). The grain size had a clear effect on the performance of the LSTM. The single-pixel model performed worst of all iterations. The 2×2 window caused considerable improvements in model performance, leading to a mean absolute error of 0.67. However, further window size increases only led to worse model performance.

Table 9. *LSTM model run results.*

Model	MAE	Weight Ratio	Slope
LSTM 1×1	0.77	0.99	0.74
LSTM 2×2	0.67	1.00	0.79
LSTM 4×4	0.65	1.01	0.77
LSTM 8×8	0.74	1.01	0.67
LSTM 16×16	0.86	0.96	0.55
CNN	0.60	1.00	1.06

5.3. Experiment 2: Full sequence testing

The full sequence model trained on 2022 achieved a MAE of 0.33, a weight ratio of 1.01 and a slope of 0.62. The predicted values have been plotted against the actual values in Figure 16. The blue trendline shows the quality of the model prediction against the 'ideal' black line. A weight ratio of 1.01 indicates a near-perfect estimation of eventual yield. A slope value of 0.62 shows a clear positive prediction trend but is imperfect.

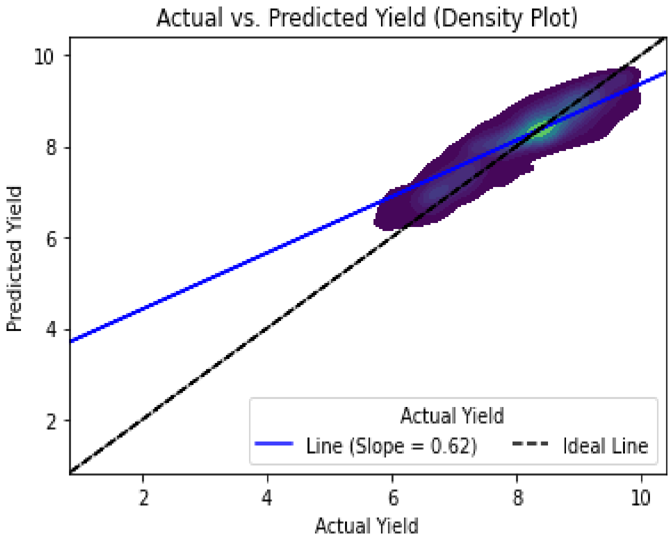


Figure 16. *Distribution graph showing the predicted values vs the actual values of winter wheat in 2022, on which the model was trained. Blue line indicates the fitted line through all points, the black line indicates the 'perfect' model.*

The same full sequence model, applied to the winter wheat field in 2023, has an MAE of 2.90, a weight ratio of 1.49, and a slope of 0.10 (Figure 17). A low slope value indicates a flawed model fit for this different year. The predicted values range from only 7.5 to 9.5, whilst the actual values have a much more comprehensive range. The MAE and weight ratio also show a bad model quality, being almost 50% off of the actual total yield.

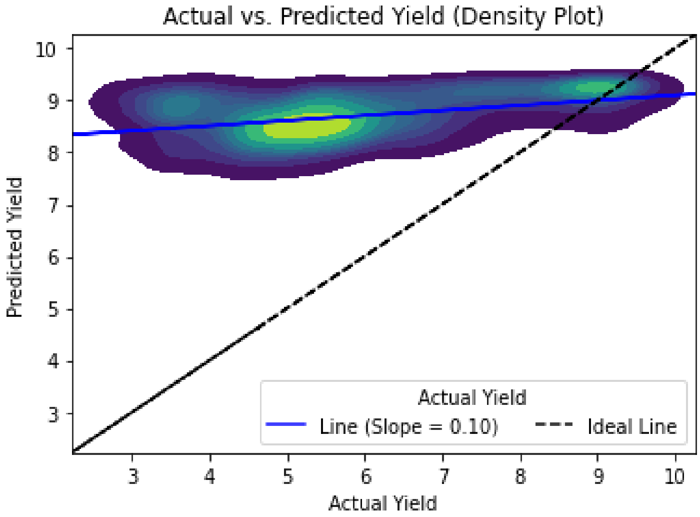


Figure 17. *Distribution graph showing the predicted values vs. winter wheat's actual values in 2023. Blue line indicates the fitted line through all points, the black line indicates the 'perfect' model.*

The full sequence model performed slightly better on the barley field in 2022. The MAE is 1.03, the weight ratio is 0.88, and the slope is 0.23 (Figure 18). The model underestimates the eventual crop yield of the barley field. However, the model slope does have a slight positive trend, indicating the model finds a relationship in the image data of the barley, like for the winter wheat in 2022.

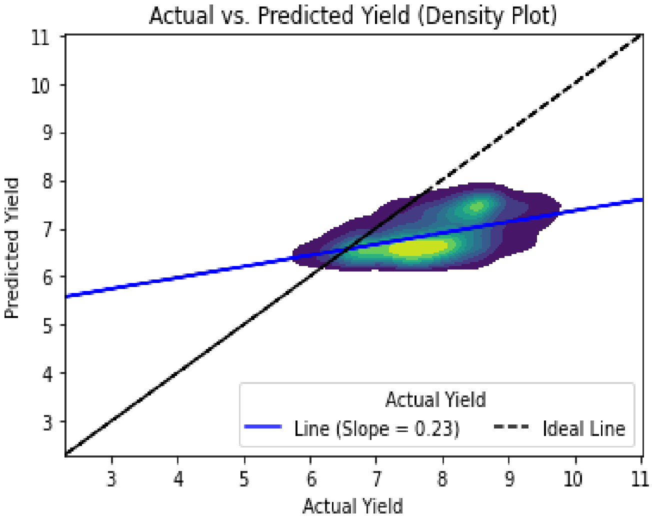


Figure 18. *Distribution of predicted vs actual yield for the full sequence model when tested on barley. Blue line indicates the fitted line through all points, the black line indicates the 'perfect' model.*

5.3.1. Full Sequence: Two-Year Training Data

The full sequence model trained on a mixed 2022 and 2023 dataset achieves a MAE of 0.42, a weight ratio of 0.99 and a slope of 0.82. The plot showing the predicted yield vs the actual yield can be seen in Figure 19. The model performs better on the two-year training data than the full sequence model trained on only 2022. It has a clear positive trend and a near-perfect weight ratio.

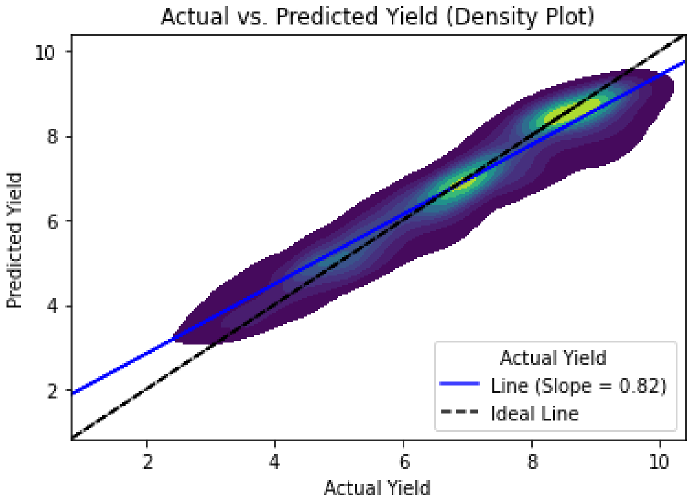


Figure 19. *Distribution graph of the full sequence model trained on multiple years on its training dataset. Blue line indicates the fitted line through all points, the black line indicates the 'perfect' model.*

The model performs very differently on the adjacent field in 2022, on which it was not trained. A MAE of 1.19, a weight ratio of 0.87 and a slope of -0.05 was achieved (Figure 20), indicating a low model accuracy. No clear trend is visible in the predicted data, and the model consistently underpredicts the actual yield.

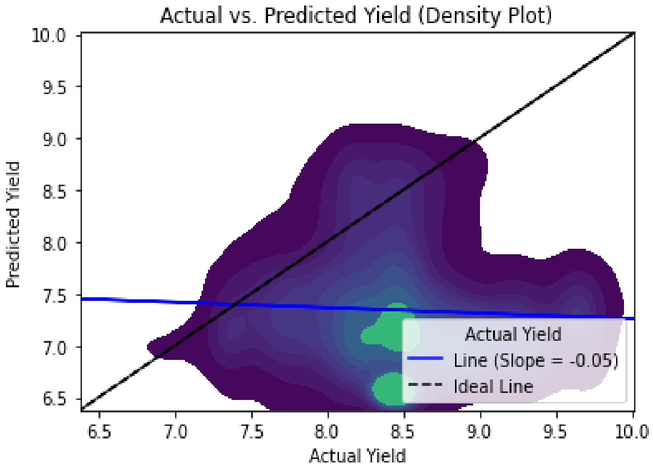


Figure 20. *Distribution graph of the performance of the two-year training data full sequence model on validation data from 2022. Blue line indicates the fitted line through all points, the black line indicates the 'perfect' model.*

Twenty per cent of the patches of winter wheat in 2023 were kept separate during the model training for validation. When applying the full sequence model on this part of the validation data, it performs better than for the validation winter wheat of 2022, with an MAE of 0.94, a weight ratio of 0.98, and a slope of 0.53 (figure 21). The model overpredicts for lower yield values and underpredicts for higher yield values but shows decent predictions all-round.

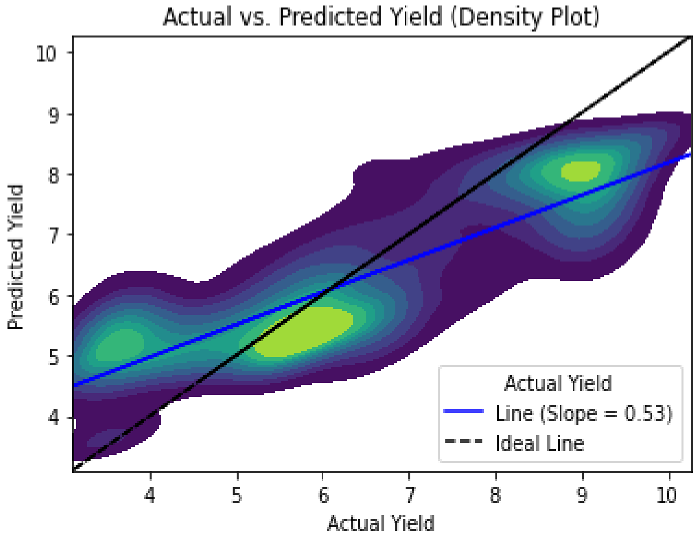
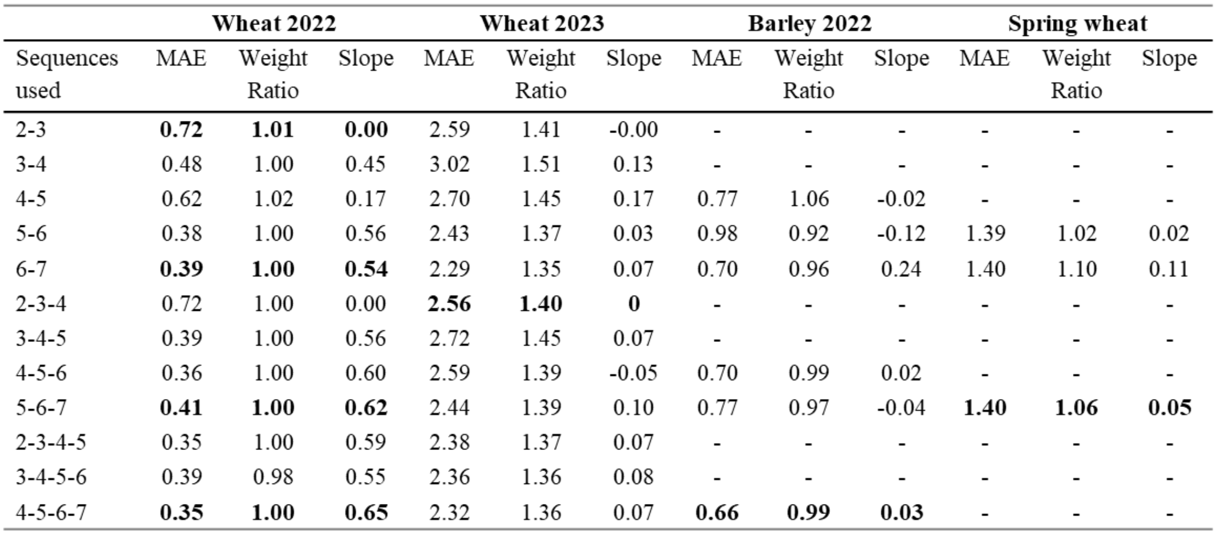


Figure 21. *Distribution graph of the two-year training data full sequence model on validation dataset from 2023. Blue line indicates the fitted line through all points, the black line indicates the 'perfect' model.*

5.4. Experiment 3: Part sequence testing

All part sequence model results of the original training data set are shown in Table 10. Model results are highest for winter wheat in 2022, on which it was trained. The part sequence models are not as accurate as their full sequence counterparts for fields other than winter wheat 2022. The weight ratio of spring wheat is high, even though the model does not accurately predict individual patch yields. Model accuracy of winter wheat in 2022 is better for part sequence models closer to the harvesting date of the crop. The weight ratio is near perfect for all model predictions for 2022 winter wheat.

Table 10. *Performance metrics for all part sequence models trained on winter wheat data from 2022. Noteworthy values were highlighted.*

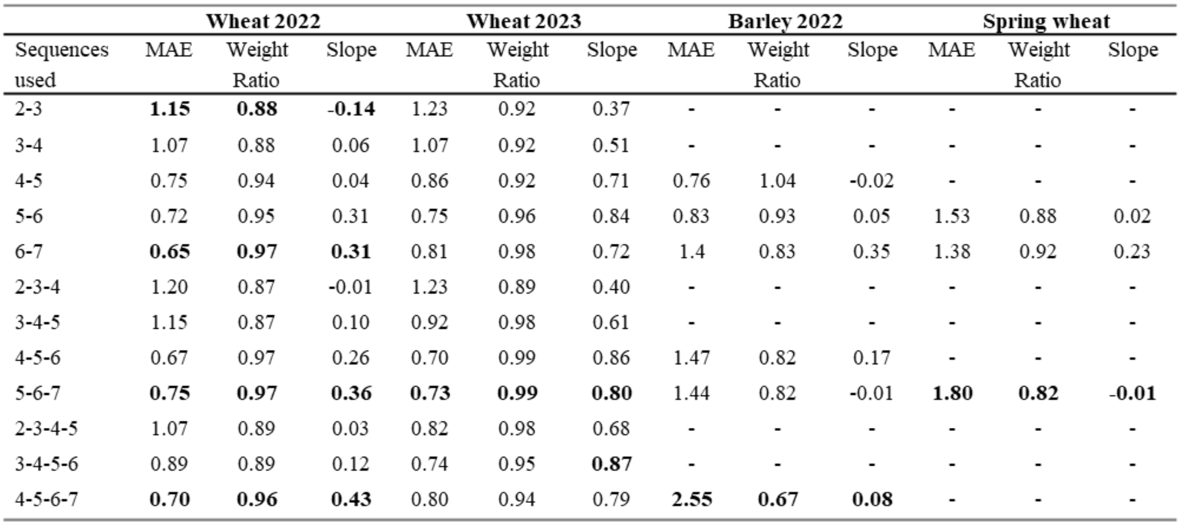


	Wheat 2022			Wheat 2023			Barley 2022			Spring wheat		
Sequences used	MAE	Weight Ratio	Slope	MAE	Weight Ratio	Slope	MAE	Weight Ratio	Slope	MAE	Weight Ratio	Slope
2-3	**0.72**	**1.01**	**0.00**	2.59	1.41	-0.00	-	-	-	-	-	-
3-4	0.48	1.00	0.45	3.02	1.51	0.13	-	-	-	-	-	-
4-5	0.62	1.02	0.17	2.70	1.45	0.17	0.77	1.06	-0.02	-	-	-
5-6	0.38	1.00	0.56	2.43	1.37	0.03	0.98	0.92	-0.12	1.39	1.02	0.02
6-7	**0.39**	**1.00**	**0.54**	2.29	1.35	0.07	0.70	0.96	0.24	1.40	1.10	0.11
2-3-4	0.72	1.00	0.00	**2.56**	**1.40**	**0**	-	-	-	-	-	-
3-4-5	0.39	1.00	0.56	2.72	1.45	0.07	-	-	-	-	-	-
4-5-6	0.36	1.00	0.60	2.59	1.39	-0.05	0.70	0.99	0.02	-	-	-
5-6-7	**0.41**	**1.00**	**0.62**	2.44	1.39	0.10	0.77	0.97	-0.04	**1.40**	**1.06**	**0.05**
2-3-4-5	0.35	1.00	0.59	2.38	1.37	0.07	-	-	-	-	-	-
3-4-5-6	0.39	0.98	0.55	2.36	1.36	0.08	-	-	-	-	-	-
4-5-6-7	**0.35**	**1.00**	**0.65**	2.32	1.36	0.07	**0.66**	**0.99**	**0.03**	-	-	-

5.4.1. Part Sequence: Two-Year Training Data

Table 11 displays all part sequence model results on the two-year training dataset. The weight ratio scores above 0.82 for all part sequence models. However, the slope and MAE do not score highly on most part sequence models. Some higher slope values were found on the winter wheat of both 2022 and 2023, indicating a better relationship between actual and predicted values. The MAE and weight ratio are relatively good for barley and spring wheat. However, the model accomplishes this by predicting the average value of the whole field instead of predicting particular values. Slope, weight ratio and MAE for winter wheat improve closer to the harvesting date, but sequence length does not have as much of an impact. **Table 11.** Performance metrics for all part sequence models trained on winter wheat data from 2022 and 2023. Noteworthy values were highlighted.

Table 11. *Performance metrics for all part sequence models trained on winter wheat data from 2022. Noteworthy values were highlighted.*



	Wheat 2022			Wheat 2023			Barley 2022			Spring wheat		
Sequences used	MAE	Weight Ratio	Slope	MAE	Weight Ratio	Slope	MAE	Weight Ratio	Slope	MAE	Weight Ratio	Slope
2-3	**1.15**	**0.88**	**-0.14**	1.23	0.92	0.37	-	-	-	-	-	-
3-4	1.07	0.88	0.06	1.07	0.92	0.51	-	-	-	-	-	-
4-5	0.75	0.94	0.04	0.86	0.92	0.71	0.76	1.04	-0.02	-	-	-
5-6	0.72	0.95	0.31	0.75	0.96	0.84	0.83	0.93	0.05	1.53	0.88	0.02
6-7	**0.65**	**0.97**	**0.31**	0.81	0.98	0.72	1.4	0.83	0.35	1.38	0.92	0.23
2-3-4	1.20	0.87	-0.01	1.23	0.89	0.40	-	-	-	-	-	-
3-4-5	1.15	0.87	0.10	0.92	0.98	0.61	-	-	-	-	-	-
4-5-6	0.67	0.97	0.26	0.70	0.99	0.86	1.47	0.82	0.17	-	-	-
5-6-7	**0.75**	**0.97**	**0.36**	**0.73**	**0.99**	**0.80**	1.44	0.82	-0.01	**1.80**	**0.82**	**-0.01**
2-3-4-5	1.07	0.89	0.03	0.82	0.98	0.68	-	-	-	-	-	-
3-4-5-6	0.89	0.89	0.12	0.74	0.95	**0.87**	-	-	-	-	-	-
4-5-6-7	**0.70**	**0.96**	**0.43**	0.80	0.94	0.79	**2.55**	**0.67**	**0.08**	-	-	-

The figures below show the distribution of yield predictions against the actual yield values for different part sequence model tests. Not all model tests have a distribution graph; some noteworthy results have been selected to display the overall distribution. Figure 22 shows the distribution for barley using a 4-5-6 sequence model and a 5-6-7 sequence model. The slope of the 4-5-6 model shows a better model fit than the final 3-part sequence model.

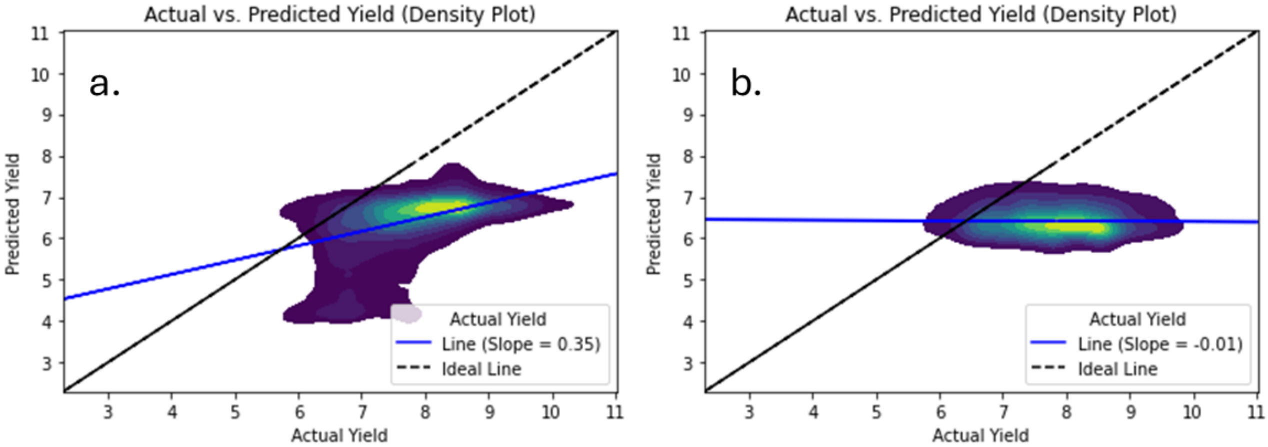


Figure 22. *Distribution graph of model performance of three sequence models 4-5-6 (a) and 5-6-7 (b) on barley. Blue line indicates the fitted line through all points, the black line indicates the 'perfect' model.*

Figure 23 shows the distribution of winter wheat field 3, on which the model was not trained. Figure 23a shows the distribution of part sequence model 2-3, which is similar in shape and slope to the results of the full sequence model for this field. The model shows a much better fit for part sequence 6-7 (Figure 23b). The predicted values are much closer to the actual values, and the slope is closer to 1. The part sequence model using sequences 4-7, Figure 23c, shows a similar distribution to Figure 23b but has more underpredicted values, lowering the intercept of the fitted line.

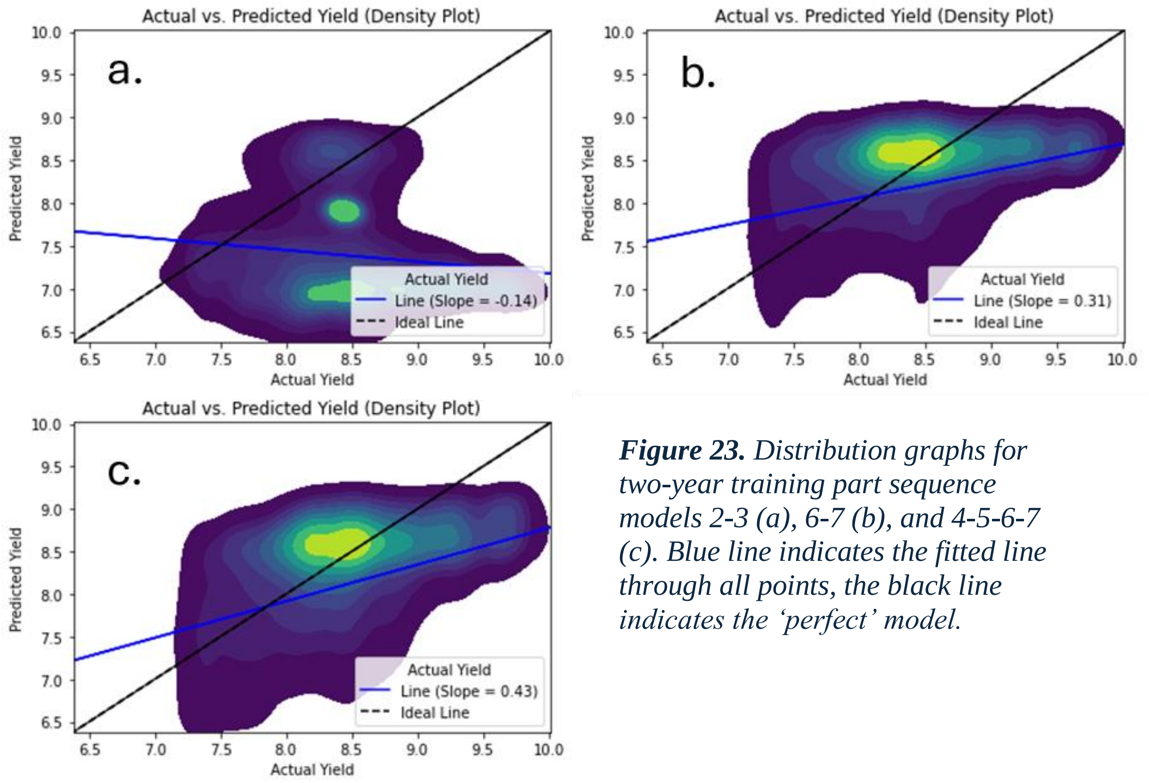


Figure 23. *Distribution graphs for two-year training part sequence models 2-3 (a), 6-7 (b), and 4-5-6-7 (c). Blue line indicates the fitted line through all points, the black line indicates the 'perfect' model.*

Figure 24 shows the distribution graph for the spring wheat field's 6-7 part sequence model. The fitted line has a slope of 0.23 and shows overprediction for lower values and underprediction for higher values.

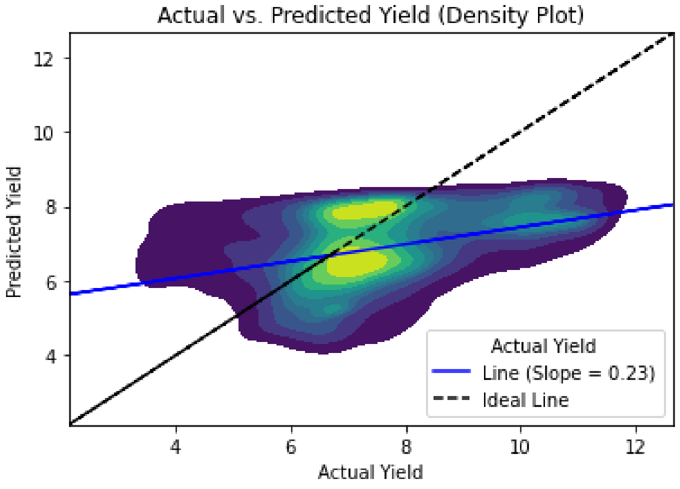


Figure 24. *Distribution graph of predicted values of the two-year training part sequence model 6-7 on spring wheat. Blue line indicates the fitted line through all points, the black line indicates the 'perfect' model.*

5.4.2. Performance of part sequence models over time

The figures below were created to better display the performance of the different part sequence models. Figure 25 shows the development of the three model criteria for 2-image part sequence testing using two years of training data. The MAE steadily decreases for the validation data of both 2022 and 2023. The slope and weight ratio also improve for sequences using later timesteps for both 2022 and 2023.

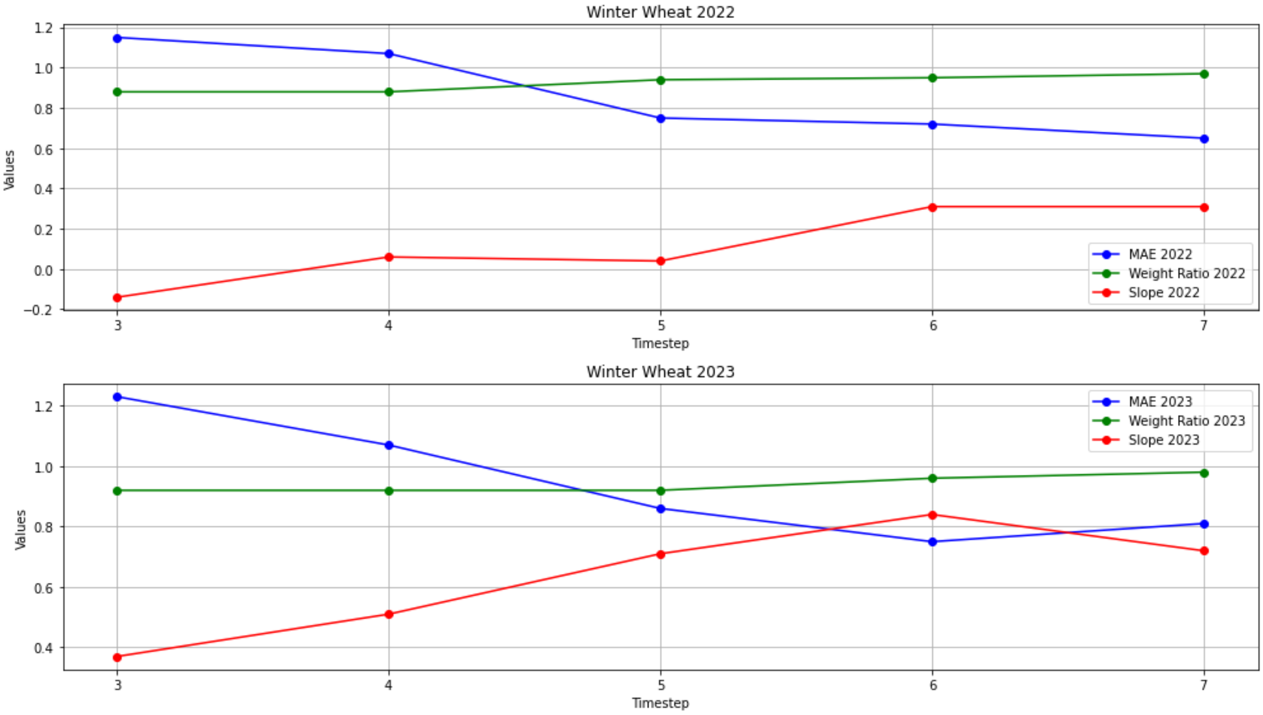


Figure 25. *Model performance metrics per timestep for 2-image sequence models. The timestep on the x-axis indicates the final timestep in the two-part sequence. Performance on the validation data from 2022 is on the top, and 2023 is on the bottom.*

Figure 26 shows the slope and MAE over time for different lengths of sequences. The slope was left out of the image to avoid crowding. No clear differences in model performances can be seen in the image.

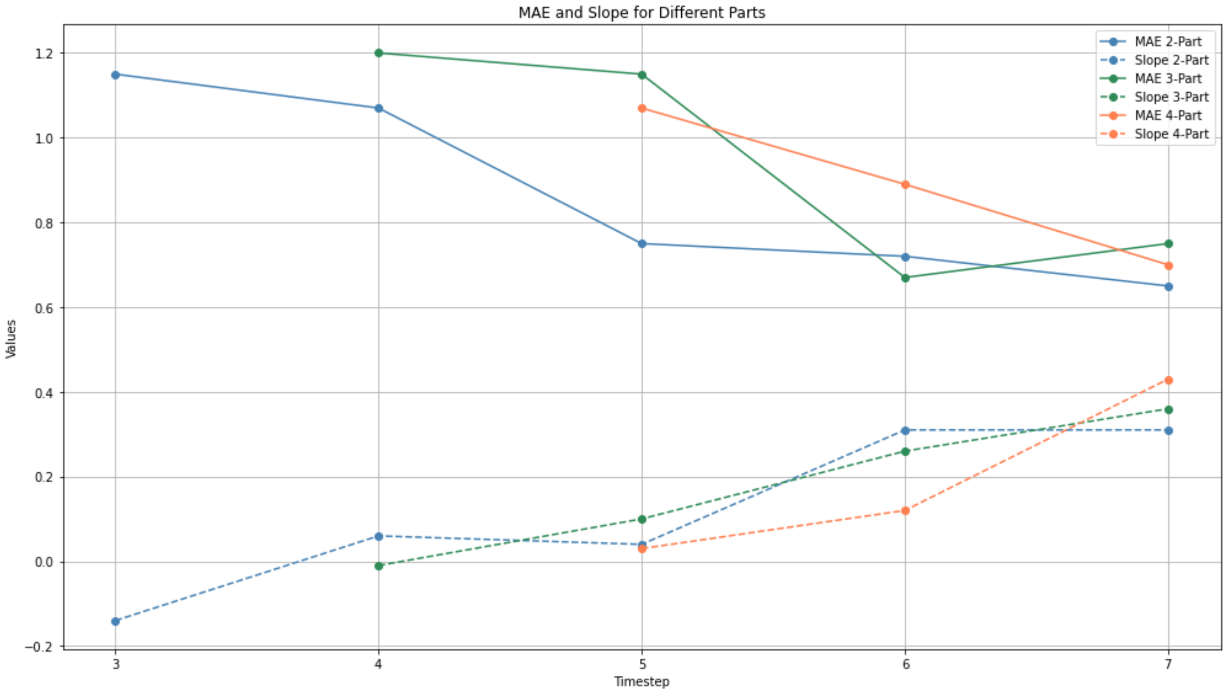


***Figure 26**. Evaluation metrics MAE and Slope for different sequence lengths over time. The x-axis indicates the last timestep of the sequence.*

6. Discussion

6.1. Prediction accuracy on winter wheat

The initial model performances on the validation data showed promising results. The model performed well on its validation data and found clear patterns in the data to make accurate predictions. The best model performance on the validation data was found for the full sequence modelling, reaching an MAE similar to Nevavuori et al. (2020). This full sequence allows the model to include the full temporal domain with all available data, achieving the highest accuracy. Some apparent differences were found in the part sequence modelling on the validation data.

When applying the initial models on the same crop but for a different year, accuracies were lower. The full sequence model, which was expected to perform best, deviates by nearly 50% in total yield for the field. Moreover, the mean absolute error is off 30% of the mean yield, showing that the model could not be applied to other years with reasonable accuracy. The predictions are off for the full sequence and part sequence modelling, with all mean absolute errors ranging from 2.32 t ha^{-1} to 3.02 t ha^{-1}. Figure 17, in full sequence modelling, shows what is going wrong in the model prediction. The model predicts a crop yield consistently between 7 t ha^{-1} and 10 t ha^{-1}, overestimating the yield in 2023. The trendline does show that the prediction values are lower when the actual yield is lower but does not come close to the actual values.

To understand why this is happening, looking at the yield distributions in Figure 4 is essential. Fields 2 and 3, on which the models were trained, have a different yield distribution than those harvested in 2023, where average yield was consistently lower. Because of this, the models are trained to predict values only within this range, causing problems in model training (Paul et al., 2021). However, the full sequence prediction for 2023 still shows a correct trendline, which leads to believe it recognises when lower yields are expected. Following these results, the models were also trained on two years of training data to help mitigate the issue of training data distribution.

The models trained on a training dataset consisting of data from both 2022 and 2023 performed better on winter wheat than models trained on only 2022. On the data it was not trained on, it achieved weight ratios close to 1 and a relatively good linear trend in predictions. The slope and MAE improve closer to the harvesting date, and sequence length also positively influences winter wheat modelling. The model performance is less accurate on barley and spring wheat. All evaluation metrics indicate an inaccurate model performance.

The difference in model predictions between the initial training dataset with only data from 2022 and the two-year training dataset with data from 2022 and 2023 displays some problems with the robustness of the modelling. Crop yield depends on climatic factors such as rain and temperature, which vary yearly (Wang et al., 2016). 2023 was a harsher year for agriculture than 2022, affecting crop yields. Ceglar et al. (2016) have highlighted these issues of inter-annual crop yield variability due to meteorological drivers. Figure 4 shows the yield distribution of the different fields, where the winter wheat in 2023 has a very different distribution than in 2022. Models must be trained on a broader range of years, including different distributions of yield values, to be more generally applicable to different years. For years where not droughts, but flooding is a problem, can still lead to unpredictable yield values with the current training dataset.

The CNN-LSTM model outperformed the separate LSTM and CNN model runs in experiment 1 (Table 9). The CNN-LSTM model uses the temporal and spatial domain, allowing it to explore more available patterns in the data (Yan et al., 2021). The LSTM model accuracy decreased for increasing grain size (Table 9), but the spatial features the CNN extracts from the larger images used, 8×8 pixels, made up for this decreased performance. The LSTM likely performed better with lower grain size because of the increased training data. This was lower for the single-pixel model as there might be too much noise in the data to make predictions accurately.

6.2. Developments in accuracy throughout the growing season

To answer RQ2, ‘Does the accuracy of crop yield predictions increase further in the growing season?’ the three evaluation metrics must be evaluated for sequences at different times. Figure 24 displays the evaluation metrics for the 2-sequence models of winter wheat in 2022 and 2023 for the models trained on 2022 and 2023. All model evaluation metrics improve for sequences closer to the harvesting date. For all fields where reasonable model accuracy was achieved, such as winter wheat in 2022 and 2023, the best part sequence model results are achieved with sequences closer to the harvesting date (Table 11). In earlier stages of the growing season, the crop still has to go through its entire development, and two similar patches could still end up with entirely different yield values. The further along the growing seasons, the closer the relationship between the reflectance of the plant and its eventual yield.

This becomes especially clear for the 2-image sequential models trained on the renewed training data when applied to the winter wheat field in 2022, which it was not trained on. Figure 23a shows that the model cannot accurately predict the final crop yield. The prediction accuracy improves with every sequence until the harvesting date (figure 25). The final distribution plot of sequence 6-7 in Figure 23c shows improvements in model accuracy and is only 3% off the field's total yield. These same improvements in model accuracy hold for 3 and 4-image length sequences, starting with a weight ratio under 0.9 for the sequences furthest away but approaching a slope of 1 for the closest sequences.

Moreover, using only the earliest growing season sequences, from April and May, consistently produces poor modelling results. On the initial training dataset, the MAE and weight ratio for winter wheat in 2022 is low, mainly due to the models being trained on this data. However, the models still struggle to find a clear pattern in the data. This becomes especially clear looking at its slope value in Table 11. Here, the model tries to minimise the MAE mainly by choosing a value close to the mean of the dataset, regardless of the data fed into the model. This is a clear example of why the calculated slope is essential in assessing modelling accuracy; it displays a lack of actual understanding of the model.

6.3. Effects of sequence length

To answer RQ3, ‘Using a sequence of bi-weekly UAV data, does the accuracy of crop yield predictions keep increasing when adding to the sequence?’ a comparison must be made between the model accuracies for different sequence lengths. This comparison can be difficult under the assumption that models using sequences closer to the harvesting date provide more accurate results. The comparison must be made by adding ‘historical’ sequential data to negate this. Does the model accuracy increase by providing a historical record of what the crop looked like at different phases of its growing stage? Figure 26 shows the development of the model slope and MAE for different sequence lengths. The x-axis displays the number of the final sequence used in the model.

Adding images to a sequence does not improve the model's performance. The different sequence length models all increase in performance for later sequences, again conforming to the hypothesis made for RQ2, suggesting that timing is more important than sequence length. The evaluation metrics vary greatly per timestep, making it difficult to draw any conclusions about the effectiveness of increased sequencing. Moreover, increasing sequence length to the full sequence modelling does not positively affect the model results. This suggests that the most recent reflectance values are more important than the reflection changes over time.

Figure 20's full sequence model for winter wheat in 2022 does not accurately predict crop yield. The part sequence models using only sequences closer to the harvesting date seem to outperform this full sequence model by using data that is better representative of the final yield. Figure 23c shows the distribution of the 2-part sequence model, which predicts the final yield much more accurately than the full sequence model. Using imagery captured earlier in the growing seasons can add unnecessary ‘noise’ to the modelling process. The CNN-LSTM model attaches too much value to these earlier sequences, whilst they may not be very representative of the final yield. Smaller sequences using newer data only use the most relevant data for predicting crop yield and can outperform the full sequence models. Additionally, shorter sequences require less computational power than longer sequences. They are trained more quickly and can make quicker predictions, which can be especially useful on larger datasets.

6.4. Prediction accuracy on other cereals

RQ4, '*Can deep learning models tuned for a specific crop be applied to similar crop species?* can be answered by the evaluation metrics in Table 10 and Table 11 for barley and spring wheat. Unfortunately, fewer barley and spring wheat sequences were available with a correct match at the time of captured images. This makes a full sequence comparison more complex, and only a few part sequence models can be run. In the initial model set-up, using training data from 2022, the model performance on barley and spring wheat seems accurate at first glance. The MAE and weight ratio scores both indicate accurate performance. However, figure 18 shows a different explanation. The model estimates all values within the yield distribution of the fields on which it was trained. The model estimates within this range for barley and spring wheat, whose total yields coincide with spring wheat in 2022. The slope suggests that the model does not do well in predicting specific yields, which can also be seen in the figure.

Model performance for models trained on the two-year dataset decreases for spring wheat and barley. The models now make predictions in a more extensive range of values but still do not accurately predict yield for different cereals. This decrease in performance compared to the first training dataset is explained by this broader range of prediction values, but its inability to predict yield has a different underlying cause. Accurately modelling crop yield for different cereals using models trained on winter wheat is only possible if the cereals behave similarly spectrally. If cereals look different from one another, the models cannot accurately predict the eventual crop yield.

Figure 27 shows the three crops at the beginning of July, at step 7, the end of the image sequence. These RGB images already display apparent differences in colour between the three cereals. Table 12 displays the mean RGB reflectance of the three different cereals for the final two images in the sequence at the end of June and the beginning of July. These differences present in mean reflectance likely cause the decreased performance of the model. Especially in timestep seven, the other cereals are not similar to winter wheat. This can be seen within the model performance in Table 12 as well. Part sequence model 5-6 performs better on barley than 6-7, containing timestep 7. The mean reflectance values deviate more from the winter wheat, decreasing model performance. This decrease in model performance can be explained by the fact that the full-grown crop looks different for barley and spring wheat than for winter wheat, which can also be seen in Figure 27.

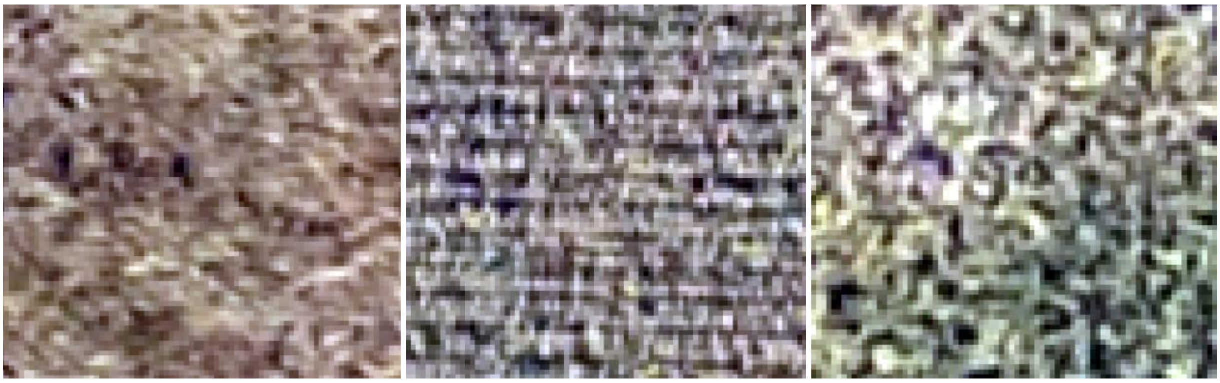


Figure 27. *RGB images displaying the visible colours for barley (left), spring wheat (middle), and winter wheat (right) on the 21st of June.*

6.5. Image calibration

The images collected by the drone were of high spatial resolution but had some evident problems, which put doubts on their reliability. All images were radiometrically calibrated using mean DN values from the reflectance panels. However, at least 1 of the panels was often saturated, leading to calibration only being done with two panels. Furthermore, the 44% reflectance panel was sometimes included in the radiometric calibration, although it should have been discarded. This led to miscalculations for the reflectance. This situation occurred when the DN values were below 255 for the 44% panel but should have been far higher, following a linear relationship. This can be seen in Figures 10 and 11, where, at many timesteps, the mean reflection is negative, which is impossible. A clear example are the images taken on 06-07-2022, with all values far below zero.

A CNN-LSTM can still find patterns in negative values, and if all images were calibrated using the same reflectance panels, this would not be an issue. However, because all images are taken at different times and calibrations are done separately each time, this leads to inconsistency in the data. By including the 44% panels the line plotted becomes too steep, which leads to negative values. These mistakes in calibration decrease model performance and the overall reliability of the data. However, most of the data was calibrated correctly and not all model error should be attributed to the effects of the calibration process. The multispectral calibration was done separately and did not include negative values. Table 12 shows a clear difference in the multispectral bands between the three cereal types.

Table 12. *Average reflection values of three fields for timestep 6 and 7 for all RGB and multispectral bands.*

Field	**Wavelength band and Timestep**									
	Red		Green		Blue		RE		NIR	
	6	7	6	7	6	7	6	7	6	7
Winter Wheat Field 2	6.7	8.4	5.9	1.5	5.3	-3.8	10.0	12.0	33.8	27.4
Spring Wheat	7.6	4.5	7.6	5.9	4.6	-0.1	10.6	9.3	22.5	26.6
Barley	6.5	9.7	5.8	-0.2	4.7	-8.7	12.0	16.5	36.6	36.4

6.6. Limitations

First, assessing modelling accuracy from a CNN-LSTM can be very complex. The models used have an architecture that can be summarised in a diagram, but they are much more of a ‘black box’ model, where an input goes in, and a particular output is achieved without knowing what happens. The relationships the models find within the data can be incredibly complex, and this makes it difficult to assess how well a model performs. This research used three metrics to evaluate model performance: MAE, weight ratio, and slope. None of these metrics are appropriate when viewed separately.

From a practical perspective, the weight ratio is the most critical metric. The eventual goal of modelling crop yield is to predict the yield of a whole field. If the weight ratio is 1, it could be said that it is a ‘perfect’ model, but it does not tell anything about what is happening within it. To approximate model performance, the slope was used, where a slope of 1 is ideal. A line with a slope of 0 cannot accurately predict singular patches but might still achieve a near-perfect weight ratio. Using the slope makes it possible to estimate how well the model predicts across different yield values, which helps understand how a model would perform in a different setting. However, even the distribution graphs with plotted lines only show the model's results, not the underlying processes. It is challenging to reason why the models are predicting in the way that they are.

Additionally, the weight ratio is indirectly the parameter being optimised in building the model, explaining the near-perfect weight ratios for the winter wheat of 2022 in Table 10. Minimising the MAE separately for each patch creates a balance between negative and positive mean errors. This can lead to a weight ratio of 1, regardless of the size of the MAE.

The crop yield used for predictions originally consisted of point data taken at 5-meter intervals. The eventual patch size used for training the model had a resolution of 8×8 pixels, which is 40×40cm. The crop yield point data was resampled to match the resolution of the imagery. However, resampling from 5×5m to 40×40cm leads to considerable uncertainty in the accuracy of the crop yield. On average, the crop yield would still be accurate. However, this uncertainty on the smaller patch scale also introduces inaccuracy into the modelling process, whereas accurate modelling requires accurate yield data collection (Doraiswamy et al., 2003). Experimentation was done with larger patch sizes, leading to lower model accuracy. This is likely due to insufficient training data caused by increasing the patch sizes, which lowers the amount batches available for training. Sufficient training data is required for a CNN-LSTM to be properly trained (Kamilaris & Prenafeta-Boldú, 2018).

A final limitation is found in the data used for the modelling. The crop yield data and UAV imagery had a limited overlap, rendering most crop yield data unusable. Furthermore, images must be captured at similar dates for different fields for sequence modelling. This was a problem in the modelling, as barley and spring wheat imagery were taken only later in the growing season. This only made a part of the images usable; for barley, five orthophotos were used, and for spring wheat, only three were available at similar dates as the winter wheat. This made it impossible to perform full sequence modelling; only three-part sequence models were used on spring wheat. Furthermore, remote sensing data generally comes with some uncertainty (Gahegan & Ehlers, 2000). Environmental conditions and sensor noise introduce uncertainty, and the empirical line function used for calibration is an approximation for calculating reflection values and is far from a perfect method.

6.7. Future applications and considerations

The crop yield modelling process has shown multiple positive and negative takeaways for future applications of crop yield modelling. The models performed reasonably well for winter wheat crop prediction. However, the difference in model performance for the two training datasets shows the importance of covering a more extensive range of values. Training a model on only one year of data makes the model unpredictable in its application in different years. Moreover, the model trained on winter wheat data from 2022 and 2023 will likely not perform as well when applied to different years. To produce a more robust model, input data from a broader range of years must be used. Even if the model is trained on multiple years of input data, it will remain sensitive to climatic conditions. Part sequence models using data from early in the growing season cannot predict extreme events such as flooding or droughts, which will negatively impact the eventual crop yield.

The models used in this research were trained on winter wheat in Sweden. The models are trained on recognising the spatial and spectral properties of the cereal. Because of this, the model could also be applied to different areas outside of Sweden if appropriately trained. The spectral properties of winter wheat will be similar worldwide, and the model should be able to make accurate predictions. However, it is essential to note the soil used to grow the cereal. Different soils have different spectral signatures, which negatively affect crop yield predictions. Moreover, different countries have different management practices, affecting the final crop yield during different stages of the growing season. For example, fertiliser or irrigation practices can influence crop yield, making the models less universally applicable.

Besides training the model with data from more than 1 or 2 years, the modelling output's resolution should also be changed. In this research, the crop yield had to be down-sampled to match the image resolution better and to provide sufficient training samples for the CNN-LSTM. Ideally, higher detail crop yield data would be used for modelling purposes. High-detail imagery has excellent benefits because it can capture details in the structure of the plant. However, a model can only capture this detail if it is matched by yield resolution. Otherwise, the yield values it predicts will never match the actual yield collected from that area. Downsampling orthomosaics that match the yield data could also have its benefits, but this would require a much larger amount of data than used in this study. Additionally, this would disregard the spatial patterns crops have on a higher resolution.

Finally, this research worked with only seven timesteps during the growing season. Imagery captured later in the growing season has been shown to improve model accuracy, but there were time gaps between each drone flight. Modelling crop yield with a more continuous series of UAV images might better display the development of prediction accuracy over time. This lack of data is especially true for spring wheat and barley, where little data was available. The modelling of these different cereals using winter wheat-based models did not look promising, but this might have been due to a lack of available data.

7. Conclusion

This research focused on the prediction of crop yield of cereals using sequential UAV imagery and DL. UAV images were made of five different fields using five different wavelength bands. These images were combined into orthophotos and then calibrated to reflection values. A CNN-LSTM model was used to use both the spatial and temporal dimensions of this series of UAV data. Different models were trained to investigate the effects of sequence length and timing of the captured imagery on model performance. Additionally, the models were trained on winter wheat and tested on barley and spring wheat to explore their applicability to different cereals.

The models initially performed well on the training dataset but performed poorly on winter wheat data from a different year. By varying the training dataset, the overall model performance increased and performed reasonably well in predicting winter wheat crop yield. The accuracy of crop yield predictions increased further in the growing season. Part sequence models using data from April and May struggled to predict crop yield accurately and were outperformed by models using later data. The model accuracy did not improve when adding images to the sequence. This finding indicates that longer sequences are not necessarily better and can, at times, hinder model performance instead of improving upon it. Finally, the models did not perform well when applied to different cereals. Barley and spring wheat predictions were not accurate when using models trained on winter wheat. Overall, new insights have been found on the influence of sequence length and prediction accuracy throughout the growing season. The most accurate predictions will be made closer to the harvest date, though accurate predictions can also be made earlier.

8. References

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

Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017, August). Understanding of a convolutional neural network. In *2017 international conference on engineering and technology (ICET)* (pp. 1-6). Ieee.

Anderson, R., Bayer, P. E., & Edwards, D. (2020). Climate change and the need for agricultural adaptation. *Current opinion in plant biology*, *56*, 197-202.

Arroyo, J. A., Gomez-Castaneda, C., Ruiz, E., de Cote, E. M., Gavi, F., & Sucar, L. E. (2017, March). UAV technology and machine learning techniques applied to the yield improvement in precision agriculture. In 2017 IEEE Mexican humanitarian technology conference (MHTC) (pp. 137-143). IEEE.

Atzberger, C. (2013). Advances in remote sensing of agriculture: Context description, existing operational monitoring systems and major information needs. *Remote sensing*, *5*(2), 949-981.

Bali, N., & Singla, A. (2022). Emerging trends in machine learning to predict crop yield and study its influential factors: A survey. Archives of computational methods in engineering, 1-18.

Baret, F., & Guyot, G. (1991). Potentials and limits of vegetation indices for LAI and APAR assessment. *Remote sensing of environment*, *35*(2-3), 161-173.

Basso, B., & Liu, L. (2019). Seasonal crop yield forecast: Methods, applications, and accuracies. *advances in agronomy*, *154*, 201-255.

Bendig, J., Yu, K., Aasen, H., Bolten, A., Bennertz, S., Broscheit, J., ... & Bareth, G. (2015). Combining UAV-based plant height from crop surface models, visible, and near infrared vegetation indices for biomass monitoring in barley. *International Journal of Applied Earth Observation and Geoinformation*, *39*, 79-87.

Benos, L., Tagarakis, A. C., Dolias, G., Berruto, R., Kateris, D., & Bochtis, D. (2021). Machine learning in agriculture: A comprehensive updated review. *Sensors*, *21*(11), 3758.

Bian, C., Shi, H., Wu, S., Zhang, K., Wei, M., Zhao, Y., ... & Chen, S. (2022). Prediction of field-scale wheat yield using machine learning method and multi-spectral UAV data. *Remote Sensing*, *14*(6), 1474.

Bolton, D. K., & Friedl, M. A. (2013). Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. *Agricultural and forest meteorology*, *173*, 74-84.

Bouras, E. H., Olsson, P. O., Thapa, S., Díaz, J. M., Albertsson, J., & Eklundh, L. (2023). Wheat Yield Estimation at High Spatial Resolution through the Assimilation of Sentinel-2 Data into a Crop Growth Model. *Remote Sensing*, *15*(18), 4425.

Broms, C., Nilsson, M., Oxenstierna, A., Sopasakis, A., & Åström, K. (2023). Combined analysis of satellite and ground data for winter wheat yield forecasting. *Smart Agricultural Technology*, *3*, 100107.

Ceglar, A., Toreti, A., Lecerf, R., Van der Velde, M., & Dentener, F. (2016). Impact of meteorological drivers on regional inter-annual crop yield variability in France. *Agricultural and forest meteorology*, *216*, 58-67.

Cernev, T., & Fenner, R. (2020). The importance of achieving foundational Sustainable Development Goals in reducing global risk. *Futures*, *115*, 102492.

Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and electronics in agriculture*, *151*, 61-69.

Crofts, H. J. (1989). On defining a winter wheat. *Euphytica*, *44*, 225-234.

Daniels, L., Eeckhout, E., Wieme, J., Dejaegher, Y., Audenaert, K., & Maes, W. H. (2023). Identifying the Optimal Radiometric Calibration Method for UAV-Based Multispectral Imaging. *Remote Sensing*, *15*(11), 2909.

Doraiswamy, P. C., Moulin, S., Cook, P. W., & Stern, A. (2003). Crop yield assessment from remote sensing. *Photogrammetric engineering & remote sensing*, *69*(6), 665-674.

E. Honkavaara, H. Saari, J. Kaivosoja, I. Pölönen, T. Hakala, P. Litkey, J. Mäkynen, L. Pesonen. Processing and assessment of spectrometric, stereoscopic imagery collected using a lightweight UAV spectral camera for precision agriculture. Remote Sens., 5 (2013), pp. 5006-5039

Elavarasan, D., & Vincent, P. D. (2020). Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications. *IEEE access*, *8*, 86886-86901.

Entz, M. H., & Fowler, D. B. (1991). Agronomic performance of winter versus spring wheat. *Agronomy Journal*, *83*(3), 527-532.

FAO, IFAD, UNICEF, WFP and WHO. 2021. The State of Food Security and Nutrition in the World 2021. Transforming food systems for food security, improved nutrition and affordable healthy diets for all . Rome, FAO.

Fei, S., Hassan, M. A., Xiao, Y., Su, X., Chen, Z., Cheng, Q., ... & Ma, Y. (2023). UAV-based multi-sensor data fusion and machine learning algorithm for yield prediction in wheat. *Precision Agriculture*, *24*(1), 187-212.

Gahegan, M., & Ehlers, M. (2000). A framework for the modelling of uncertainty between remote sensing and geographic information systems. *ISPRS journal of photogrammetry and remote sensing*, *55*(3), 176-188.

Gonzalez, Carmen G. "SDG 2: End Hunger, Achieve food security and improved nutrition, and promote sustainable agriculture." (2022).

Gregory P.J, Ingram J.S.I and Brklacich M 2005Climate change and food securityPhil. Trans. R. Soc. B3602139–2148

Grochowska, R. (2014). Specificity of food security concept as a wicked problem. *Journal of Agricultural Science and Technology B*, *4*(2014), 823-831.

Guntukula, R. (2020). Assessing the impact of climate change on Indian agriculture: Evidence from major crop yields. *Journal of Public Affairs*, *20*(1), e2040. http://doi.org/10.1098/rstb.2005.1745

Han, J., Zhang, Z., Cao, J., Luo, Y., Zhang, L., Li, Z., & Zhang, J. (2020). Prediction of winter wheat yield based on multi-source data and machine learning in China. Remote Sensing, 12(2), 236.

Hanuschak Sr, G. A. (2013). Timely and accurate crop yield forecasting and estimation: History and initial gap analysis. In *The first Scientific Advisory Committee Meeting, Global Strategy; Food and Agriculture Organization of the United Nations: Rome, Italy* (Vol. 198).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Huang, S., Tang, L., Hupy, J. P., Wang, Y., & Shao, G. (2021). A commentary review on the use of normalized difference vegetation index (NDVI) in the era of popular remote sensing. Journal of Forestry Research, 32(1), 1-6.

Hunt, M. L., Blackburn, G. A., Carrasco, L., Redhead, J. W., & Rowland, C. S. (2019). High resolution wheat yield mapping using Sentinel-2. *Remote Sensing of Environment, 233*, 111410.

Ingram, J. (2011). A food systems approach to researching food security and its interactions with global environmental change. Food security, 3, 417-431.

Janssens, C., Havlík, P., Krisztin, T., Baker, J., Frank, S., Hasegawa, T., ... & Maertens, M. (2020). Global hunger and climate change adaptation through international trade. Nature Climate Change, 10(9), 829-835.

Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). A review of the use of convolutional neural networks in agriculture. *The Journal of Agricultural Science*, *156*(3), 312-322.

Karthikeyan, L., Chawla, I., & Mishra, A. K. (2020). A review of remote sensing applications in agriculture for food security: Crop growth and yield, irrigation, and crop losses. Journal of Hydrology, 586, 124905.

Kasampalis, D. A., Alexandridis, T. K., Deva, C., Challinor, A., Moshou, D., & Zalidis, G. (2018). Contribution of remote sensing on crop models: a review. *Journal of Imaging*, *4*(4), 52.

Kogo, B. K., Kumar, L., & Koech, R. (2021). Climate change and variability in Kenya: a review of impacts on agriculture and food security. *Environment, Development and Sustainability*, *23*, 23-43.

Koppel, R., & Ingver, A. (2008). A comparison of the yield and quality traits of winter and spring wheat. *Latvian Journal of Agronomy*, *11*, 83-89.

Kuester, T., & Spengler, D. (2018). Structural and spectral analysis of cereal canopy reflectance and reflectance anisotropy. *Remote Sensing*, *10*(11), 1767.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

Li, Z., Chen, Z., Cheng, Q., Duan, F., Sui, R., Huang, X., & Xu, H. (2022). UAV-based hyperspectral and ensemble machine learning for predicting yield in winter wheat. *Agronomy*, *12*(1), 202.

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*.

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

Lyle, G., Bryan, B. A., & Ostendorf, B. (2014). Post-processing methods to eliminate erroneous grain yield measurements: review and directions for future development. *Precision agriculture*, *15*, 377-402.

Maes, W. H., & Steppe, K. (2019). Perspectives for remote sensing with unmanned aerial vehicles in precision agriculture. *Trends in plant science*, *24*(2), 152-164.

Maimaitijiang, M., Sagan, V., Sidike, P., Hartling, S., Esposito, F., & Fritschi, F. B. (2020). Soybean yield prediction from UAV using multimodal data fusion and deep learning. *Remote sensing of environment*, *237*, 111599.

Nevavuori, P., Narra, N., Linna, P., & Lipping, T. (2020). Crop yield prediction using multitemporal UAV data and spatio-temporal deep learning models. *Remote sensing*, *12*(23), 4000.

Olsson, P. O., Vivekar, A., Adler, K., Garcia Millan, V. E., Koc, A., Alamrani, M., & Eklundh, L. (2021). Radiometric correction of multispectral uas images: Evaluating the accuracy of the parrot sequoia camera and sunshine sensor. *Remote Sensing*, *13*(4), 577.

O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.

Ozturk, A., Caglar, O., & Bulut, S. A. N. C. A. R. (2006). Growth and yield response of facultative wheat to winter sowing, freezing sowing and spring sowing at different seeding rates. *Journal of Agronomy and Crop Science*, *192*(1), 10-16.

Paul, M., Ganguli, S., & Dziugaite, G. K. (2021). Deep learning on a data diet: Finding important examples early in training. *Advances in Neural Information Processing Systems*, *34*, 20596-20607.

Rashid, M., Bari, B. S., Yusup, Y., Kamaruddin, M. A., & Khan, N. (2021). A comprehensive review of crop yield prediction using machine learning approaches with special emphasis on palm oil yield prediction. *IEEE access*, *9*, 63406-63439.

Rogan, J., & Chen, D. (2004). Remote sensing technology for mapping and monitoring land-cover and land-use change. Progress in planning, 61(4), 301-325.

Scotford, I. M., & Miller, P. C. H. (2005). Applications of spectral reflectance techniques in northern European cereal production: a review. *Biosystems engineering*, *90*(3), 235-250.

Shahi, T. B., Xu, C. Y., Neupane, A., & Guo, W. (2022). Machine learning methods for precision agriculture with UAV imagery: a review. *Electronic Research Archive*, *30*(12), 4277-4317.

Shammi, S. A., Huang, Y., Feng, G., Tewolde, H., Zhang, X., Jenkins, J., & Shankle, M. (2024). Application of UAV Multispectral Imaging to Monitor Soybean Growth with Yield Prediction through Machine Learning. Agronomy, 14(4), 672.

Sishodia, R. P., Ray, R. L., & Singh, S. K. (2020). Applications of remote sensing in precision agriculture: A review. *Remote Sensing*, *12*(19), 3136.

Sjulgård, H., Keller, T., Garland, G., & Colombi, T. (2023). Relationships between weather and yield anomalies vary with crop type and latitude in Sweden. *Agricultural Systems*, *211*, 103757.

Staudemeyer, R. C., & Morris, E. R. (2019). Understanding LSTM--a tutorial into long short-term memory recurrent neural networks. *arXiv preprint arXiv:1909.09586*.

Stoskopf, N. C., Nathaniel, R. K., & Reinbergs, E. (1974). Comparison of Spring Wheat and Barley with Winter Wheat: Yield Components in Ontario 1. *Agronomy Journal*, *66*(6), 748-750.

Sun, J., Di, L., Sun, Z., Shen, Y., & Lai, Z. (2019). County-level soybean yield prediction using deep CNN-LSTM model. *Sensors*, *19*(20), 4363.

Tsouros, D. C., Bibi, S., & Sarigiannidis, P. G. (2019). A review on UAV-based applications for precision agriculture. *Information*, *10*(11), 349.

United Nations. (2017). *The sustainable development goals*. United Nations Publications.

Van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, *177*, 105709.

Vargas, R., Mosavi, A., & Ruiz, R. (2017). Deep learning: a review.

Wang, R., Bowling, L. C., & Cherkauer, K. A. (2016). Estimation of the effects of climate variability on crop yield in the Midwest USA. Agricultural and Forest Meteorology, 216, 141-156.

Wang, Y., Zhang, Z., Feng, L., Du, Q., & Runge, T. (2020). Combining multi-source data and machine learning approaches to predict winter wheat yield in the conterminous United States. Remote Sensing, 12(8), 1232.

Weiss, M., Jacob, F., & Duveiller, G. (2020). Remote sensing for agricultural applications: A meta-review. *Remote sensing of environment*, *236*, 111402.

Wing, I. S., De Cian, E., & Mistry, M. N. (2021). Global vulnerability of crop yields to climate change. *Journal of Environmental Economics and Management*, *109*, 102462.

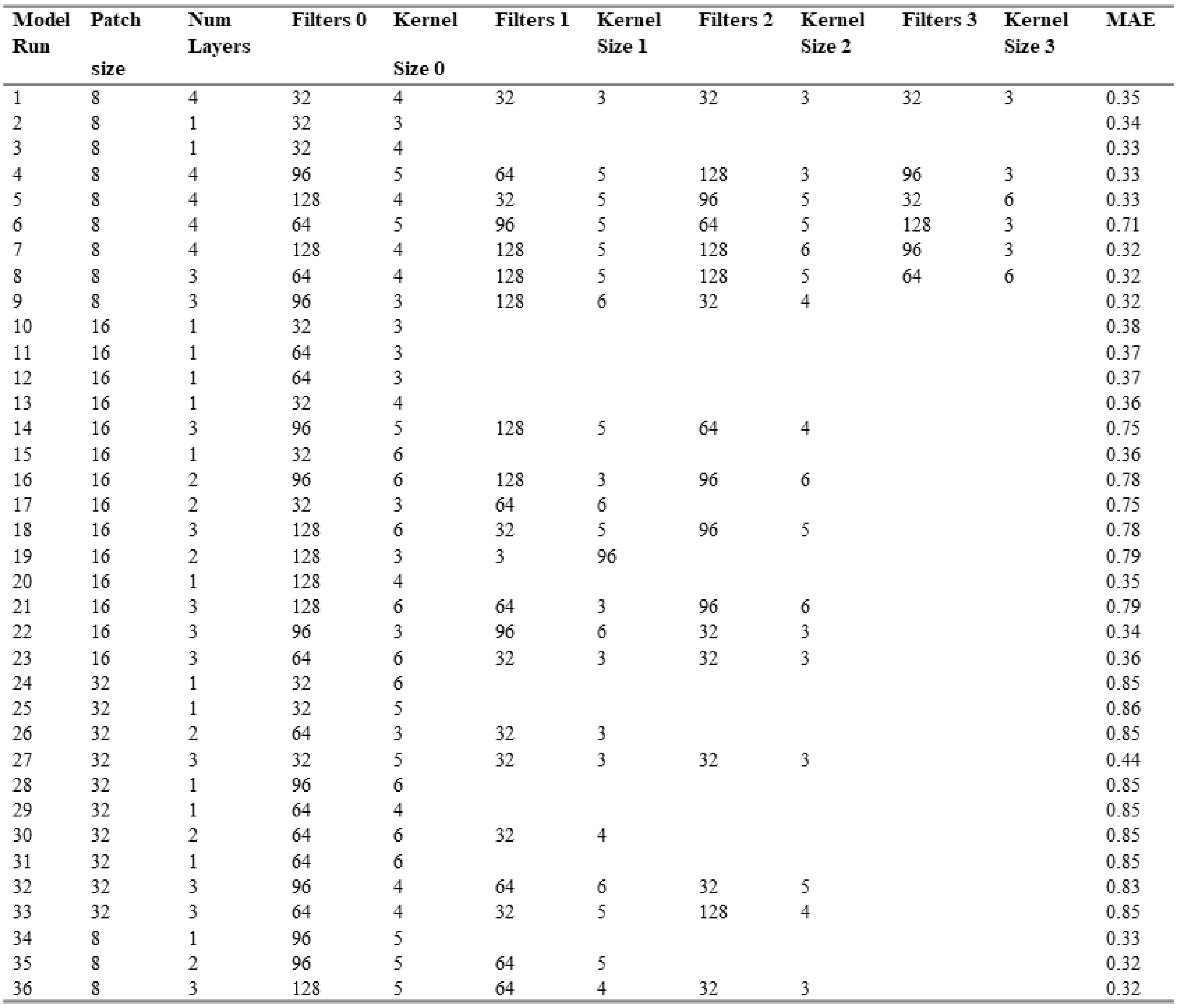
Yan, R., Liao, J., Yang, J., Sun, W., Nong, M., & Li, F. (2021). Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering. *Expert Systems with Applications*, *169*, 114513.

Appendices

Appendix A. The results of the first hyperparameter exploration. The table shows the different hyperparameters tested, the achieved MAE, and the number of epochs until a stable result was reached.

Model Run	CNN Layers	Filters	LSTM Layers	LSTM Units	Learning Rate	Epochs to stable	MAE
1	2	32, 32	1	32	0.001	5	0.48
2	5	128, 64, 64, 32, 32	1	32	0.001	7	0.49
3	1	32	1	32	0.001	4	0.45
4	1	32	6	32	0.001	4	0.59
5	1	32	3	32	0.001	7	0.47
6	1	32	1	256	0.001	6	0.46
7	1	32	2	128, 32	0.001	7	0.45
8	1	32	3	256, 128, 28	0.001	7	0.47
9	1	32	1	32	0.01	4	0.67
10	1	32	1	32	0.0001	10	0.48

Appendix B. The results of the hyperparameter tuning for the CNN architecture. Results were based on MAE.



Model Run	Patch size	Num Layers	Filters 0	Kernel Size 0	Filters 1	Kernel Size 1	Filters 2	Kernel Size 2	Filters 3	Kernel Size 3	MAE
1	8	4	32	4	32	3	32	3	32	3	0.35
2	8	1	32	3							0.34
3	8	1	32	4							0.33
4	8	4	96	5	64	5	128	3	96	3	0.33
5	8	4	128	4	32	5	96	5	32	6	0.33
6	8	4	64	5	96	5	64	5	128	3	0.71
7	8	4	128	4	128	5	128	6	96	3	0.32
8	8	3	64	4	128	5	128	5	64	6	0.32
9	8	3	96	3	128	6	32	4			0.32
10	16	1	32	3							0.38
11	16	1	64	3							0.37
12	16	1	64	3							0.37
13	16	1	32	4							0.36
14	16	3	96	5	128	5	64	4			0.75
15	16	1	32	6							0.36
16	16	2	96	6	128	3	96	6			0.78
17	16	2	32	3	64	6					0.75
18	16	3	128	6	32	5	96	5			0.78
19	16	2	128	3	3	96					0.79
20	16	1	128	4							0.35
21	16	3	128	6	64	3	96	6			0.79
22	16	3	96	3	96	6	32	3			0.34
23	16	3	64	6	32	3	32	3			0.36
24	32	1	32	6							0.85
25	32	1	32	5							0.86
26	32	2	64	3	32	3					0.85
27	32	3	32	5	32	3	32	3			0.44
28	32	1	96	6							0.85
29	32	1	64	4							0.85
30	32	2	64	6	32	4					0.85
31	32	1	64	6							0.85
32	32	3	96	4	64	6	32	5			0.83
33	32	3	64	4	32	5	128	4			0.85
34	8	1	96	5							0.33
35	8	2	96	5	64	5					0.32
36	8	3	128	5	64	4	32	3			0.32

Appendix C. The results of the hyperparameter testing for the LSTM architecture. The LSTM was trained with the CNN structure from trial 36 (see Appendix B).

Model Run	LSTM layers	Units 1	Units 2	Units 3	Additional	MAE
1	1	32				0.32
2	1	32			No max pooling layers	0.73
3	2	64	32			0.32
4	3	128	64	32		0.72
5	1	32			Dropout = 0.1	0.32
6	1	64				0.32