Petteri Nevavuori

Feasibility of remote sensing based deep learning in crop yield prediction

The series, ISBN and ISSN numbers are added on this page in the library.

Dedicated to my wife Heli, our kids and my saviour,
Jesus Christ.

PREFACE/ACKNOWLEDGEMENTS

Preface or acknowledgements here.

ABSTRACT

In this dissertation the applicability of novel machine learning methods with remote sensing data were studied in the context of agricultural decision support systems in smart farming. The main focus is in the utilization of high resolution unmanned aerial vehicle (UAV) data to perform in-season crop yield estimation with spatial and spatiotemporal deep learning model architectures in Finnish coastal habitat. While open-access satellite data has already been utilized in crop related modelling, such as crop type classification and yield prediction, intra-field scale prediction with smaller fields common to Nordic countries requires images with higher resolution than what is currently available from the open-access satellite systems. In addition to using UAV remote sensing data, various combinations of crop field related sensor data, data from open-access sources and satellite data are evaluated. Data quality is also an important aspect with remote sensing data, with high altitude satellite-based earth observation suffering from occasional obstructions by cloud canopy. A decision tree model is employed to estimate cloud coverage by using UAV data as cloudless groudn truth. In this disseration it is shown that with high-resolution UAV data, crop yield prediction with convolutional neural networks (CNNs) is feasible and produces results accurate enough for performing corrective farming actions in-season. Using UAV data time series not only improves the modelling performance (post-season prediction) with high-resolution UAV RGB data but also improves the predictive capabilities (in-season prediction). Furthermore, using various data sources for crop yield prediction in addition to UAV RGB data is shown to improve the predictive capabilities of the model. In summary, the use of deep learning techniques can be seen to improve the smart farming decision support pipeline by providing performant and reliable decision engines.

CONTENTS

List of Figures

List of Tables

ORIGINAL PUBLICATIONS

Publication I P. Nevavuori, N. Narra and T. Lipping. Crop yield prediction with deep convolutional neural networks. *Computers and Electronics in Agriculture* 163.June (2019). DOI: 10.1016/j.compag.2019.104859.

Publication II N. Narra, P. Nevavuori, P. Linna and T. Lipping. A Data Driven Approach to Decision Support in Farming. *Information Modelling and Knowledge Bases XXXI*. Vol. 321. 2020. DOI: 10.3233/FAIA200014.

Publication III P. Nevavuori, T. Lipping, N. Narra and P. Linna. Assessment of Cloud Cover in Sentinel-2 Data Using Random Forest Classifier. *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2020, 4661–4664. DOI: 10.1109/IGARSS39084.2020.9323683.

Publication IV P. Nevavuori, N. Narra, P. Linna and T. Lipping. Crop Yield Prediction Using Multitemporal UAV Data and Spatio-Temporal Deep Learning Models. *Remote Sensing* 12.23 (2020). DOI: 10.3390/rs12234000.

Publication V P. Nevavuori, N. Narra, P. Linna and T. Lipping. Assessment of Crop Yield Prediction Capabilities of CNN usign Multisource Data. *New Developments and Environmental Applications of Drones - Proceedings of FinDrones 2020*. 2021. DOI: 10.1007/978-3-030-77860-6.

1 INTRODUCTION

This doctoral dissertation studies the applicability of novel machine learning methods with remote sensing data in the context of agricultural decision support systems (DSS) in precision agriculture [3] and smart farming [79]. Farmers have practiced precision agriculture for ages to optimize yield productions of their fields. Sources of intra-field variability were deduced by noting and exchanging annual observations and experimenting with interventions. However, both the observations and the conclusions drawn have been more or less based on intuition, rather than on objective data. From this emerges the need for data-driven decision making, i.e. smart farming, to aid the farmers in choosing the best actions to take to optimize crop cultivation [32]. The application of novel deep learning techniques has been on an increasing trend for the past few years in smart farming and precision agriculture application domains [41]. One of the key reasons for this progression is the abundant availability of sensor based data in terms of ground-based soil sensors, and low-altitude unmanned aerial vehicles (UAV) and high-altitude satellite systems [93]. Another factor is the open-access availability of other enviromental data, such as weather and land survey data. Thus, the use of remote sensing data to extract information with machine learning models for data-driven decision making has become more common. Especially, the number of studies using deep learning techniques to perform agriculture related modelling tasks has steadily increased [31].

Remote sensing data relevant to smart farming tends to be predominantly spatial in nature. This stems from the objects of interest - fields, forests and plots of land. Conventionally, open-access remote sensing data has been acquired from nationally operated multispectral satellite sources, such as Sentinel-2 (ESA, Paris, France) or Landsat 8 (USGS, Reston, Virginia, USA). Satellite data, while spatial, is also temporal due to regular and frequent overflights over land and sea surfaces. Commercially available UAVs have also been utilized [56]. While some UAVs come pre-fitted with quality RGB sensors, some systems are designed as platforms to which then desired

sensor technology is to be mounted. Due to the altitude from which the data is acquired, satellite and UAV data differ greatly in spatial resolution. This is illustrated in Figure 1.1, where (a) is an orthomosaic of UAV iamges of a field and (b) is the corresponding image as captured by Sentinel-2 satellite at approximately the same time. While the pre-fitted RGB cameras of UAVs allow data capture resolutions well below 1 m/px, open-access satellite data is available at resolutions starting from 10 m/px (Sentinel-2). These data, satellite and UAV, are readily in image-like spatial format. Other field-related observational data, such as data from soil sensors or soil samplings, are often interpolated over plots of interest to generate image-like data in the form of spatial rasters.

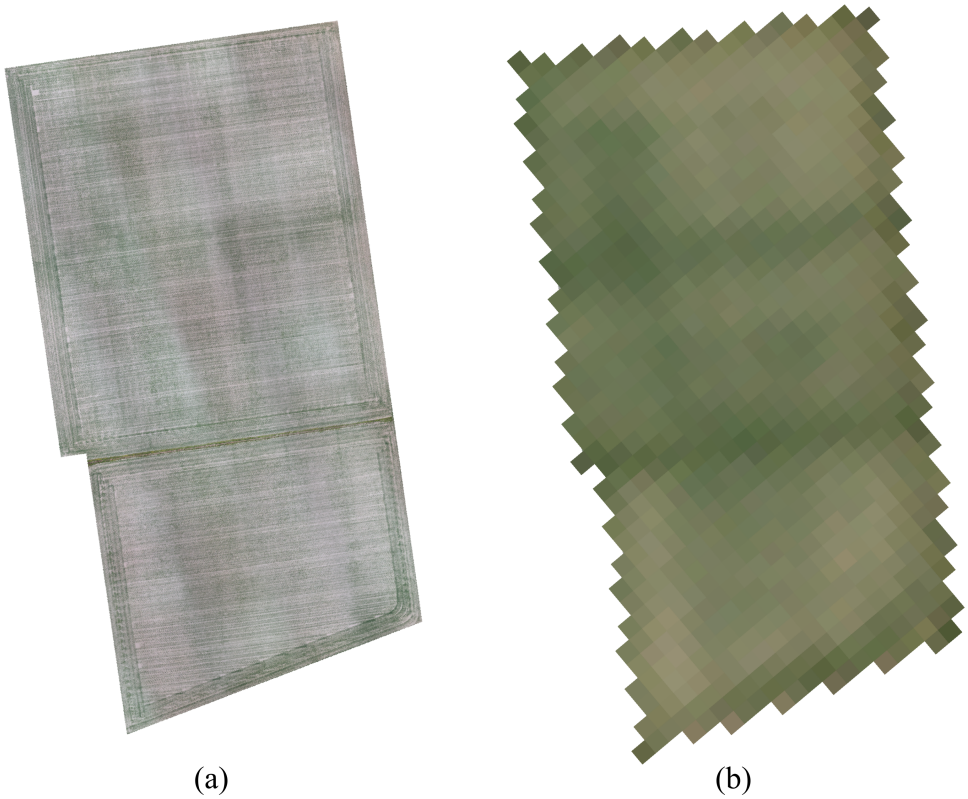

Figure 1.1 Images of a field from week 24 of 2018 from (a) UAV and (b) Sentinel-2.

The form of input data directly effects the selection of suitable data-based modelling techniques. Convolutional neural networks (CNN) [47, 48], a subset of neural network based deep learning techniques, excel with spatial data related tasks. These tasks include object recognition, imgae classification and image-based regression. Re-

cently, multiple studies have been conducted with CNNs in the context of agriculture and smart farming [30]. The use of sequential models capable of extracting temporal features is also relevant with remote sensing data. Long short-term memory (LSTM) networks [17, 24], an implementation of recurrent neural networks (RNN) [62], have been shown to perform well in modelling tasks involving sequential data [29]. The LSTMs have to be coupled with CNNs to perform spatiotemporal modelling. Another way to tap into spatiotemporal data is to use three dimensional CNNs, where two dimensions are used for single point-in-time spatial inputs and the third dimension as the dimension of change between distinct spatial inputs [86].

1.1 Research questions

In the context of using field related remotely and manually gathered data, the research questions of this study are:

> **RQ1.** Can intra-field yield variability be reliably predicted using deep learning models based on high resolution remote sensing data from the early phase of the growth season?
>
> **RQ2.** Which data sources add value to high resolution yield prediction with deep learning models?

RQ1 is heavily centered around data based modelling with field related data. Excelling at complex decision making with fuzzy problems, humans are ill-equipped to derive causal and correlational relationships, whether linear or non-linear, from larger bodies of raw numerical data. Spatial data, such as RGB images of a field, consist of thousands of data points with multiple values associated to a single point. Spatial deep learning models, on the other hand, have been specifically developed to perform input-output mapping with spatial data. Due to the nature of these models, they require black-box optimization techniques to find the optimal combination of various hyperparameters. Hyperparameters are values, that have an effect on the training and the capabilities of the model. These values are, for example, the learning rate coefficient of the model's optimizing algorithm or the number of neurons, a calculation unit, within a layer of the layered deep learning architecture. Succsefully attaining the first objective requires also proper handling of input and target data samples. The data has to be both ingestable by the models, while the model's results

have to be meaningful and interpretable by us humans. An additional key aspect is the usability of the models in commercial production environments. Having to do with usage and adoption, the usability of the models as a part of a bigger DSS has to be evaluated.

Generally, deep learning models benefit from feature-rich data. Being non-linear and layered models, they are optimized during training to find the most effective combinations of input and hidden features built from the input data to accomplish the performance goals. Data, however, incurs a resource cost on the modelling process. Firstly, the data acquisition has an effect on the overall feasibility of the modelling. UAV data, for example, requires manual operation in Finland due to legislation and regulations. Secondly, the contribution to model performance is not equal between distinct data sources. Yet another aspect of data is its quality, which itself might affect the general performance of the model and the system the model is used in. Thus, the data used in the modelling has to be evaluated both in terms of feasibility and usability (RQ2).

For years, the number of farmers has been on a decline in Finland. With rather static number of field plots, the farms get bigger and are thus in need of better farm and process management tools. Manual, semi-automated and automated data acquisition from various operational areas require data processing automation to provide actionable items in actionable time frame. Thus, this study is an attempt to answer the question whether data based modelling is beneficial for farm management and process optimization.

1.2 Publications and author's contribution

The publications selected for this dissertation fall into three categories. The first category is about novel intra-field crop yield prediction model development. Publications [I] and [IV] belong to this category. The second category is related to data evaluation assessment. The publications belonging to this category are [III] and [V]. The last category is about the context in which crop yield modelling is performed, decision support systems for agriculture. Publication [II] belongs to this last category. For the publications in the first and the second category, the author did the majority of the work. In these publications, the author alone was responsible for accumulating, preprocessing and preparing the data from various sources. The author

carried out the work of developing, implementing and training the models presented in the publications. Model performance evaluation and comparison to the state-of-the-art research was also conducted by the author. In those publications the author did not partake, however, in manual data acquisition, such as operating the UAVs during the growing season. The author of this dissertation was also responsible for writing the majority of text in these publications. In the publication [II] category, the work of the author was utilized in the study. The model architecture, code and results of [I] were utilized as a case study in the report.

Intra-field crop yield prediction model [I] [IV]

Performing crop yield predictions from RGB image data requires using models capable of ingesting spatial data and deriving salient features from them. As part of the Mikä Data project carried out in the Data Analytics and Optimization research group of the Pori unit of Tampere University, several fields were imaged during the growing seasons of 2017-2019. UAV-based orthomosaic images of crop fields have the data in a resolution high enough to allow for extracting image frames of fixed dimensions. The images of these fields were used to train models to perform frame-based crop yield prediction with single point-in-time [I] as well as time series [IV] image data. Throughout this study, point-in-time is used as an expression to distinquish between temporally distinct inputs from temporal sequences of multiple inputs. The point-in-time model is based on a CNN, with its depth and configuration tuned to perform mapping of RGB image frames of crop fields to geolocationally matched yield data collected from yield mapping sensors during harvest. The time series model is evaluated from a selection of spatiotemporal deep learning model architectures: a CNN-LSTM, a convolutional LSTM and a 3D-CNN. The best performing model architecture for mapping time series of RGB image frames of crop fields to corresponding crop yield data was the 3D-CNN. While crop related modelling has been performed on larger scales such as county-scale in USA [78] and China [27] and at country-scale in Europe and Africa [65], field-scale UAV-based crop yield estimation for intra-field predictions is a novel contribution to the best knowledge of the author.

Remote sensing data evaluation [III] [V]

In addition to performing crop yield estimation with UAV remote sensing data acquired manually, the use of crop field related sensor data, remotely and locally collected, is a topic of interest in the context of decision support in farming. As with any data, quality is one of the key interests. High altitude satellite-based earth observation suffers from occasional obstructions by cloud canopy. While Sentinel-2 data products contain pre-calculated information about the possible presence of cloud cover, there's still work to do on the detection accuracy [10]. Using UAV RGB image data as ground truth for cloudless data of crop fields, a random forest ensemble decision tree was trained in [III] to perform pixel-wise cloudiness classification of Sentinel-2 data. Normalized difference vegetation index (NDVI) was calculated for UAV RGB and Sentinel-2 true color RGB data and the difference used as an indicator for building the pixel-wise ground truth labels.

Another active area of research is combining data from multiple input sources to perform remote sensing data based modelling [18]. In [V], field-wise UAV RGB data was complemented with data from Sentinel-2 satellites, manually collected soil samplings, soil's electrical conductivity, weather data and topographical data. A CNN model configuration from [I] was then used as the baseline, as the performance had already been demonstrated with UAV RGB data. In addition to training a baseline RGB-only model, several input data configurations were tested and evaluated to see which combination of input data sources would provide the best performance.

Decision support system for farming [II]

While developing machine and deep learning methods have recently been an active research area [41], the research and development of user friendly decision support system platforms is crucial to deployment, and thus adoption, of developed models. In [V] a basis for such a platform was laid, with the focus being on the persistence and visualization of multisource spatial data of crop fields. Crop yield prediction models form the artificial intelligence (AI) engine of the open-source Oskari-based (*www.oskari.org*, MIT & EUPL licensed) agricultural data management and viewing platform, generating refined predicted data for deriving actionable decisions during the growing season.

2 DATA-BASED SMART FARMING

The objectives of this thesis stem from the farmers' need to derive data-based farming decisions from data measured of their fields. While aggregated field-level data provides general guidelines, the actions and interventions are performed at the intra-field scale. The decisions also have to be made within an actionable time frame during the growing season. However, the data alone is not enough. As unmanned aerial system (UAS) overflights can be utilized to provide frequent image snapshots of distinct states of fields and crop growth, predicting an outcome from these data is a difficult task humans. What is needed is an automated decision engine based on data-based machine learning techniques, capable of performing intra-field predictions using current state of crop development. Furthermore, this decision engine should be integrated into a holistic farming decision support system (DSS) to fully utilize the capabilities of modern sensors, connectivity and automatic data processing. This enables the farmers to make more informed decisions on what actions to take and in which parts of distinct fields.

In this chapter we will first review the relevant background and the current state-of-the-art smart farming and data sources in the context of crop yield prediction. While smart farming encompasses a broader farming context, from soil and water management to utilizing modern technology to optimize farming processes, we will constrain the discussion to the context of crop field management and crop yield estimation.

The chapter is constructed as follows. In the first section the current studies of data-driven smart farming are reviewed. This is to gain a proper view of the application context for machine learning models, which are discussed in Chapter 3. After that, data from distinct sources and the use thereof in agriculture-related modeling tasks are reviewed. Remote sensing is of particular interest, as it has been an active research area for a several years already. Other data sources, such as soil and weather data, are also discussed. In addition to reviewing relevant studies, we will also de-

scribe the data utilized in the studies related to this dissertation. In the last section of this chapter the modelling task of crop yield prediction is reviewed.

2.1 Precision agriculture and smart farming

The technologization towards the modern age farm has been a steady process, ongoing for several centuries. The first steps in this process were taken during the 18th century with important gradual developments in crop rotation and selective breeding techniques. After the World Wars, the farms were quickly mechanized and farming processes started to get more industrialized. Manual labor and animal work force wre replaced with more effective machinery. As digital computation resources became more common via mainframe architectures starting at late 1960s, software products were adopted as common tools for agronomical counselling instutions and, thus, farming management practices. The introduction of the internet and developments in telecommunication, sensor and computer technologies enabled the farms gain increasingly detailed grasp of different areas of crop farming. The introduction of digital computation first transformed the data handling and computation processes of agricultural experts and advisors, starting with punch hole cards and progressing towards software applications [80].

The developments in sensors, information technology (IT) systems and the general adoption of digital farm management and decision support systems have further driven the transformation towards what is known as precision agriculture. Precision agriculture is seen to encompass location-based technologies, processes and management concepts to better account for the intra-field variability for increased gains. While precision agriculture is focused mainly on farming operations in the field, smart farming extends the combination of physical sensors, IT systems and low latency connectivity to a holistic and automated farm management framework. This view is shared between multiple studies. Sundmaeker et al. [79] position precision agriculture within smart farming as do Wolfert et al. [93] and Tantalaki et al. [82]. While Rose and Chilvers use the terms more interchangeably, their use of the term smart farming implies a larger framework, encompassing precision agriculture as technology and sensor oriented subarea [61].

Being conceptual frameworks, both precision agriculture and smart farming have experienced developments via advancements in distinct technological areas. This is

reflected in recent studies. As discussed by Klerkx et al. in their review of digital agriculture, technologies such as precision farming, internet of things (IoT), machine learning (ML), deep learning (DL) and robotics have been a focus in an increasing number of agriculture related studies [40]. In a recent review of machine learning (ML) based crop yield prediction, Van Klompenburg et al. observed an increase in publications utilizing novel data-based modelling concepts starting from 2013 [41]. Similar observation is made in a review of the use of deep learning (see Chapter 3) in agriculture by Tantalaki et al. [82]. They observe a monotonic increase of 249 % on the average the number of annually published agriculture related deep learning focused studies between 2016 and 2019.

2.1.1 Decision support systems for agriculture

The concepts of smart farming and digitalized agriculture are among the most relevant topics in the agricultural research domain. The key elements in smart farming revolve around data collection and utilization [40], data-based decision making [32], interconnectivity of cyber-physical systems [101], automation of farming processes [101] and improved management of farm processes [82].

One of the core elements of smart farming is data collection. Small and interconnected sensors, being more generally labelled as IoT-sensors, are utilized in tandem with sensors installed on farming equipment and machinery to produce a multi-source data stream from the farm. Data accumulated over time paints a holistic picture of the farm and its operations. Novel AI-related techniques further facilitates data-based decision making via insight extraction and estimation. This enables the farmers to base their decisions on measured data in timely and accurate manner [79]. Moreover, the developments in soil sensors planted in crop fields enable the farmers to remotely monitor their fields, which in turn allows them to make more informed decisions on actions to take [82]. Being a subject closely related to the IoT, performing data aggregation and analysis on-site via edge computing is another projected direction for agricultural cyber-physical systems [101].

Sensors, data and insights require effective management systems. A holisting agricultural management system addresses a farm's needs on multiple levels, such as accounting, traceability and on-farm process management. The management systems are also required to connect the farm to its stakeholders, such as consumers, pub-

lic authorities and actors in the food value chain [82]. With the developments of the IT-sector in general, farm management solutions have also shifted from locally installed software to cloud-based services [101]. This change further opens up new possibilities for data-based decision making [61]. Especially, resource-intensive modelling techniques are easier to employ with dedicated servers. The adoption of smart farming practices makes the farm effectively a producer and manager of goods and operations related data. Being part of a larger agricultural ecosystem, the data generated on-farm is seen to benefit other instances, such as actors in the logistic chain and counselling institutions [32].

When smart farming is viewed as a holistic operating framework, equipment and independent systems add formidably to the complexity of the whole. There is a true need to further develop the integration of sensors, equipment, monitoring and management systems [79]. This calls for cooperation of business actors operating in the domain of smart farming, with IT operations being a focus of the development due to integrations. With working integrations, the benefits of accurate and timely automation can be reaped [101].

Several commercial decision support systems exist in the domain of agriculture. As the products are generally suites of modular and specialized applications, we will review the products only generally. Minun Maatilani (Mtech Digital Solutions Oy, Vantaa, Finland) provides the farmers with web-based applications for cattle and crop farm operations planning, accounting and management. There are explicit modules available for smart farming, which include features for managing cropping plans, creating and exporting fertilization tasks for machinery, importing of UAV data and yield maps. Satellite data is utilized to provide timely views of fields. Next Farming (FarmFacts Gmbh, Pfarrkirchen, Germany) has applications for crop and fertilization planning, fleet management, creation and management of prescription tasks for machinery. Users can import information about their fields, such as biomass, soil and yield maps. The software suite includes smart farming services such as UAV management, seeding and fertilication optimization and supplying geographic information system (GIS) data. 365FarmNet (365FarmNet Gmbh, Berlin, Germany) contains applications for farm management, crop cultivation and herd management. Via partner applications the suite provides the users with satellite-based field monitoring, crop, seed and fertilizer planning and fertilization optimization. MyEasyFarm (MyEasyFarm, Bezannes, France) contains applications for plant and plot manage-

ment, task management, imported data analysis (soil, yield, etc.) and task monitoring.

2.1.2 Crop yield prediction

Crop yield prediction, the primary focus of this study, is deemed one of the most challenging problems in the realm of smart farming, the latter encompassing a large variety of sub-tasks and smaller goals. Predictive yield modelling is seen to help farmers pinpoint problem areas in their fields [75], guide management decisions and reduce business risk [13] and provide vital information for the food supply chain [104]. As discussed by Triantafyllou et al. [87], crop and plant yield estimation is crucial when the goal is to optimize field-wise yields in cost-effective and proactive manner. In their study of a holistic remote sensing system architecture, predictive models are positioned adjacent to data analysis, information management and data processing modules within what they call "managament layer". Management layer provides managament logic to the applications operated by the users, farmers or agricultural experts.

According to Ünal in their review of deep learning method utilization in the context of smart farming, yield estimation is one of the most common agriculture related keywords present in the reviewes 120 studies [89]. The output, the harvested crop yield, is affected by a variety of environmental, crop-related and farmer-induced factors. Data-based modelling techniques, namely deep learning models, excel with such multivariate and non-linear data [97]. In their review of machine learning based crop yield prediction, van Klompenburg et al. [41] observe that the data sources often present in crop yield prediction studies include soil and crop information, climatological data, information about the nutrients and actions taken by the farmer.

In addition to gathering data from multiple sources, it is also necessary to collect data across multiple years. As Filippi et al. discuss, having the data cover larger time spans (*temporal coverage*) is deemed more important than having the field related data span larger areas (*spatial coverage*) [13]. A key aspect to using crop yield prediction in a smart farming DSS is to enable the farmer to decide on actionable items. Predicting the intra-field variability allows identifying underfperfoming areas in the fields [82]. With the increase of spatial resolution in predictions, the goals of precision agriculture are also easier to attain by focusing on distinct problem areas instead of

treating the whole field in uniform manner.

2.2 Data sources

Remote sensing has played a significant role in advancing crop field monitoring during the past decades and is considered one of the most important technolgies for precision agriculture and smart farming [88]. According to Khanal et al., the publicly accessible high-altitude satellite systems, such as Sentinel (ESA, Paris, France) and Landsat (USGS, Reston, Virginia, USA), have been a major catalyst in propelling remote sensing based agricultural research forward [38]. Other key factor in this progression has been the developments in computation and storage capabilities of such data. While high altitude monitoring is good for observing larger areas, low-altitude unmanned aerial vehicles (UAV) and unmanned aerial systems (UAS) are used to capture information in greater detail. According to Ünal et al. in their review of deep learning in smart farming, the use of UAVs in recent agricultural deep learning studies is so prevalent that their use can be considered an integral part of the smart farming framework [89].

Agricultural data is known to be heterogeneous [32]. According to Wolfert et al., this stems from the heterogeneity of the means of data accumulation, which include various remote sensing platforms, ground-based sensors and human-inputted data [93]. Another source of data heterogeneity are the objects of data measurement, i.e. the environment, machinery and operational records. In a recent review of the use of multisource and multitemporal data in remote sensing, Ghamisi et al. conclude that the increased availability of data from multiple sources accompanied with advances in computational tools has a positive effect on data-based modelling, increasing the efficiency and performance of the models [18]. Their review focuses solely on studies utilizing high and low altitude remote sensing platforms and their sensors. The sensor types include those of visible light RGB, multispectral, hyperspectral and laser imaging, detection and ranging, hereafter called lidar as per [18]. In a review of big data practices in agriculture, Kamilaris et al. observe that multiple data-based modelling studies in the domain of agriculture also utilize data from other sources [32]. These sources include weather stations, geospatial data, soil sensors, historical datasets and records kept by organizations, institutions and governments.

2.2.1 Low-altitude unmanned aerial vehicles

UAVs have been utilized for the past decade in multiple studies related to remote sensing, data-based modelling and agriculture. Recently published reviews show that the number of UAV-related studies has substantially grown. Therefore it is more beneficial to perform a metareview on recent reviews focused on low-altitude remote sensing and its applications.

To preface the review of UAV usage in the context of remote sensing and crop yield estimation in agriculture, it is necessary to note that UAVs utilized in studies are mainly just aerial platforms to which the sensors are mounted. This is in contrast to several commercially available UAVs with integrated RGB cameras. Generally, there are five types of sensors present in the recent studies: visual RGB, multispectral, hyperspectral, thermal and lidar sensors [55, 88, 96]. As implied by the name, visual RGB sensors capture the red, green and blue bands of the visible light spectrum in the 400-700 nm wavelength range [96]. Multispectral sensors usually add one to several additional channels from select wavelengths in the near-infrared (NIR) wavelength region of 780-2500 nm. Hyperspectral sensors are used to capture a continuous spectral range from visible to NIR wavelengths [96]. Thermal sensors measure the infrared radiation in the 3-8 μm wavelength region [55]. Compared to the sensors mentioned above, lidar is an active sensor, emitting the signal and measuring its reflection from various surfaces [38, 96]. Visual RGB sensors are generally the easiest to operate and cheapest to acquire. Multispectral and hyperspectral sensors need often be acquired and mounted separately and they cost considerably more than RGB sensors. Thermal and lidar sensors are among the most expensive UAV-mountable sensors [88].

Khanal et al. reviewed accomplishments, limitations and opportunities of remote sensing in agriculture [38]. Searching for studies related to remote sensing and agriculture, they discovered 3679 studies during the 20-year period from 2000 to 2019. The number of UAV related studies, according to their research, started to increase after 2013. The annual numbers rose from a handful at beginning of the considered period to well over a hundred UAV related studies published in 2019. Focusing on recent and major references, they review the applications of remote sensing in precision agriculture. They observe that UAVs have been utilized in the following applications:

- topographical mapping (1/3)
- tile drainage locationing (2/5)
- soil moisture and temperature mapping (3/8)
- crop emergence and density monitoring (5/5)
- nitrogen stress monitoring (1/3)
- crop disease monitoring (3/8)
- weed identification and classification (3/4)
- yield prediction (2/4).

The numbers after the items indicate the number of UAV related references reported out of all reported references for an application. Overall, they observe UAV related studies making up 16.3 % of the studies considering remote sensing in agriculture during 2015-2019. The majority of the studies they reviewed focused on satellite sources. Recently, however, there has been an increase in studies utilizing UAV-based data to perform data analysis and data-based modelling with high resolution data. In the studies they selected for closer inspection, the UAVs were equipped with visual, multispectral and thermal sensors for various applications. In their view, UAV platforms provide a reasonable means to gather high-frequency and high-resolution remote sensing data with. Citing US prices, they report UAV data collection to cost approximately 9.9$/ha. They also point out that operating UAVs is constrained by weather conditions, limited flight time and payload.

Touros et al. conducted a review on UAV-based applications for precision agriculture [88]. They reviewed 100 research papers published between years 2017 and 2019. According to them, UAVs can be used to produce high to ultra-high resolution images of crop fields by varying the flying height. They observe that UAVs are utilized in the following applications:

- crop growth monitoring (65.6 % of studies)
- weed mapping (12.5 % of studies)
- crop health monitoring (6.3 % of studies)
- crop irrigation management (5.2 % of studies).

While other applications were observed in addition to these, the four formed the majority (89.6 %). Limited to these application contexts, four distinct categories of

sensors were observed. These were multispectral (56.0 %), RGB (33.6 %), thermal (6.0 %) and hyperspectral (4.4 %). They conclude that the use of various vegetation indices derived from multispectral data are the most effective in crop parameter monitoring. Overall, they observed more than 30 distinct crop species among the reviewed studies. For this dissertation, crop grotwh monitoring as an application context is of the greatest interest, while crop yield prediction is considered a part of it in the review. RGB and multispectral sensors are reported being the most utilized for this application. Machine learning methods were observed as being able to exploit data from all sensor types, both separately and conjoined.

Xie and Yang reviewed the current state-of-the-art of UAV-mounted sensor utilization in plant phenotypic trait monitoring and estimation [96]. Main phenotypic traits include plant yield, biomass, height, leaf area index, chlorophyll content and nitrogen content. Overall, they observed 18 different plant varieties as the targets for UAV-based sensing in their review. Concluding from plant yield estimation focused studies, they suggest using RGB and multispectral sensors with UAVs. Biomass, height and leaf area index are also treated as proxy variables for plant yield. Biomass estimation was performed mainly with RGB and multispectral sensor data. Lidar was observed as the dominant sensor type with canopy height estimation. Leaf area index was mostly estimated using various vegetation indices derived from multispectral data with some studies resorting to RGB sensors as well. In conclusion, they observe that RGB and multispectral sensors are used predominantly in plant related monitoring and estimation studies. This is attributed to lower sensor costs, sensor lightness and the ease of data collection and analysis. Multispectral data is seen, however, to be crucial for some crop related monitoring and modelling contexts where vegetation indices based on he infrared part of the spectrum are utilized.

Messina and Modica reviewed the current state of the art of UAV thermal imagery and its applications [55]. Thermal sensors detecting the infrared radiation are used mainly to monitor ground surface temperature. It has been observed to be a rapid response variable to plant growth, yield estimation and stress factor evaluation. Compared to other sensor types, such as RGB and multispectral, operating thermal sensors requires more care. Environmental variables, such as humidity, clouds, dust and time of day, can impede the data acquisition process. Calibration of sensors and measuring environmental variables near the imaged objects is strongly suggested to perform corrections during data processing. The most utilized applications for UAV-

mounted thermal sensors observed in their review were the following:

- water stress detection and monitoring (23 studies)
- phenotyping (5 studies)
- yield estimation (4 studies).

2.2.2 High-altitude satellite systems

Remote sensing studies conducted with free and commercial satellite data have been common for longer than comparable studies with UAVs. For several years already, satellite data has been considered a core data source in the smart farming framework [93]. Some of the often utilized satellite systems with their specifications are given in Table 2.1, but it is to be noted that there exists much larger number of past and presently operational satellite missions. For reference, see the database of satellite missions at [69].

Table 2.1 Some of the commonly referenced satellite systems present in remote sensing and agriculture related studies.

Satellite	Spatial Resolution [m/px]	Revisit Time [days]	Number of Satellites	Spectral Channels	Spectral Range [μm]	Launch Year	Open Access
Landsat-7 [45]	15-60	16	1	8	0.441-12.36	1999	Yes
Landsat-8 [46]	15-60	16	1	11	0.435-12.51	2013	Yes
Sentinel-2 [73]	10-60	5	2	13	0.426-2.377	2015	Yes
WorldView 2 [94]	0.31-1.84	1.1	1	9	0.450-2.365	2009	No
WorldView 3 [95]	0.31-1.24	<1 to 4.5	1	29	0.450-2.365	2014	No
PlanetScope [58]	2.7-3.2	1	140	4	0.455-0.860	2016	No
Gaofen 1 [15]	2-16	4	1	5	0.450-0.900	2013	Yes
Gaofen 2 [16]	0.81-3.24	5-69	1	5	0.450-0.900	2014	No

Since the launches of higher resolution satellite systems, such as Landsat 8 in 2013 and Sentinel-2 in 2015, and the opening of their data, the usage of data from remote sensing satellites in various application domains has become more feasible. As Chivasa et al. discuss, a review of maize yield estimation applications based on remote sensing, coarse-resolution satellite data was largely unusable with smaller sized fields on the African continent [8]. The values in a pixel corresponding to a field would effectively always be contaminated with data unrelated to the field. Further-

more, to estimate a yield produced by a spatially irregularly shaped field requires data at high enough resolution to constrain the field data within reasonable borders.

Khanal et al. calculate that 64 % of the 3679 remote sensing and agriculture related studies published in and after the year 2000 utilize satellite-based data [38]. They also observe satellite data-based studies being more prevalent than UAV utilizing studies in the years from 2000 to 2010. According to their research focused on selected studies, satellite data have been utilized in the following agriculture related applications:

- tile drainage locationing (1/5)
- soil moisture and temperature mapping (3/8)
- nitrogen stress monitoring (1/3)
- crop disease monitoring (1/8)
- weed identification and classification (1/4)
- yield prediction (1/4)
- grain quality assessment (1/3)
- crop residue assessment (3/4).

Numbers in brackets indicate the satellite data utilization counts in all papers related to the particular application context. The numbers suggest that satellite-based studies are in the minority when compared to UAV studies. This, however, might be attributable to the authors of the review as they seem to put more focus on high resolution studies. UAVs and mid-altitude manned aircrafts are better at producing high resolution data. Regarding economics, medium resolution satellite data is largely open access and free to use. High-resolution satellite data is reported to cost from 1.28 USD/km^2 (5 m/px resolution) to 25 USD/km^2 /0.5 m/px resolution). Compared to UAVs at 9.9 USD/ha, the price with commercial satellites is cheaper for larger areas. Smaller areas require economic evaluation case-by-case, as minimum order size is enforced with commercial high-resolution satellite data.

In another recent study, Karthikeyan et al. review remote sensing applications regarding crop growth, irrigation and crop losses [35]. Focusing on the international and global scale, they assess uses of current operational satellite systems in performing large-scale data acquisition for monitoring and modelling of crop growth, losses

and irrigation. While they affirm that data gathered on site with UAVs and sensors is more efficient in smaller scale, they view satellites as unrivalled in continuous monitoring of larger areas. Regarding crop growth, they observe that the multispectral and hyperspectral instuments in satellite platforms enable the use of various vegetation indices relevant to crop assessment. To effectively utilize vegetation indices, the utilized satellite systems are required to have at least mediocre spatial resolution. Similar to [8], they acknowledge the problem of pixel value contamination for agricultural use with too coarse resolutions. For irrigation monitoring, they observe visible, infrared and microwave sensors being utilized. Recently, data fusion has also been utilized in generating yearly irrigation maps dating decades to the past. In these studies satellite data were complemented with other data, including weather, soil and topographical information. While they assess several application contexts, they conclude that higher resolution is often needed.

2.2.3 Weather data

Optical sensing is of crucial importance when performing spatial modelling in the context of crop yield prediction. While sensing crop growth stages is helpful, gathering data about the environment is mandatory to distinquish the effects of a crop type's phenological factors from external factors. In a study of a holistic remote sensing monitoring system, Triantafyllou et al. position weather data logging on par in terms of importance with other sensors installed and planted on site [87]. Reported weather related environmental factors include wind speed and direction, atmospheric pressure, light intensity, solar radiation and rainfall. In addition to specifically installed sensors, nationally collected weather data and forecasts have also been used [32].

Sun et al. conducted a multisource soybean yield prediction study at US county scale [78]. In addition to remote sensing and yield data, they utilized historical daily weather data accumulated in the Google Earth Engine [20]. The weather data, namely precipitation and atmospheric pressure, were utilized as rasters with 1 km/px ground sample distance. Analyzing their results, they attribute some of the lowest soybean yields partially to extreme weather. However, they note that singling out the effects of external factors on yield is complex. Their conclusion is that weather data accompanied with remote sensing data form a sufficient data set with

which to predict soybean yields using coarse resolutions.

In a study of maize growth stage prediction, Yue et al. utilized a county level meteorological data set as the predictor data in China [100]. The weather data consisted of daily aggregates for humidity, atmospheric pressure, temperature, precipitation, wind speed and sunlight amount measured from a single weather station. The temporal range of the data is reported being from 1981 to 2017. The weather data were temporally aligned with the maize growth data to facilitate timely estimation of the maize growth stage from meterological data only. Using days of growth as the predicted value, they report average absolute error of 1.06 days.

Wolanin et al. utilized time series of remote sensing and weather obsevations to estimate crop yields in the Indian Wheat Belt [92]. They utilized daily aggregates of temperature, precipitation, water vapour deficit, short-wave radiationa and day length information. In addition they utilized vegetation indices calculated from remote sensing data. They trained their models with data from multiple years to potentially observe the effects of environmental factors on the crop yield. Their conclusion is that while vegetation indices capture the effects of environment and render weather data somewhat redundant in their modelling approach, analysis of the model's utilization of meterological features provides insights into other study areas, such as crop breeding.

2.2.4 Soil data

Being the base of the growth for crops, soil and its composition plays a major role in how the plants grow and produce grains. As Tantalaki et al. show in a review of novel data-based applications in precision agriculture, soil and its features are commonly the target of modelling [82]. However, studies have also been conducted where soil and ground related data are used as predictor values.

In a review of machine learning and crop yield prediction, van Klompenburg et al. observed soil type and soil maps being utilized often in recent data-based modelling studies in the context of agriculture [41]. Invidual spatial soil features included soil type, pH, cation exhange capacity and location. Soil related features were, overall, observed to be the most prevalent group of data features present in the reviewed studies. These feautures were observed as predictors of crop yield 54 times, while second most popular group, solar information, saw 39 uses as predictor values in

similar setting. Soil information was also utilized both as predictor and predicted values in the reviewed studies.

Filippi et al. collected a multisource data set to estimate crop yields [13]. Soil related features included soil electrical conductivity, potassium, uranium, thorium, clay and sand content. These acquired data were processed to a resolution of 10 m/px. Other data sources included remotely sensed vegetation indices as well as received and forecasted precipitation. Regarding the use of soil data, they conclude that soil maps and geophysical data were not as significant predictors as initially assumed. However, they observed correlations between soil and ground related predictor values and point out that this might actually mask their combined significance.

Khanal et al. utilize soil related features in their study of machine learning based intra-field corn yield and soil feature estimation [37]. Using a single field for their study, the soil was sampled at intervals of one acre or 0.40 ha. From the samplings, the ratios for soil organic matter, potassium and magnesium were extracted. Cation exhange capacity and pH were also measured. These features were, however, treated as target values. The inputs for estimation consisted of high resolution multispectral (<1 m) images and digital elevation model data. Inputs were spatially aligned with corresponding soil samples, forming the soil-related input-target data set. In their study the authors compared statistical, linear and non-linear models. Spatial models, such as CNNs, were not, however, taken into comparison.

2.2.5 Lidar and topographical maps

As already mentioned as one of the sensors mountable to UAVs, lidar is often utilized when remotely sensed elevation information is acquired. As pointed out by Khanal et al., topographical features affect preseason farming management decisions, impacting field's water economy and soil quality [38]. Another common application context are tree and forest related studies [67].

In a review of multisource and multitemporal remote sensing data fusion, Ghamisi et al. point out multiple studies in which lidar data has been utilized [18]. In raw form a lidar sensor produces multidimensional point cloud of data, which contains information about locations and altitudes of the points. They observe that lidar is oftern accompanied with a separate hyperspectral sensor. One of the main reasons for this is that, lidar generally lacks spectral information often necessary. This is true

for scene classification, for example.

While lidar sensors provide exceptional accuracy when a digital elevation model (DEM) is required, other approaches exist for mapping the topography of the target area of interest. Recent advances in UAV-based photogrammetry, i.e. modelling structure from images taken from different angles, provide an alternative approach to map intra-field topographical variability [38]. Namely, the advances in UAV-based photogrammetry have enabled producing DEMs from considerably cheaper and lightweight RGB sensors. These methods, however, lack canopy penetration when compared to lidar [54].

2.2.6 Yield maps

Crop yield estimation is an important topic in the context of smart farming and precision agriculture. Correctly estimating crop yield mid-season enables the farmer to proactively focus on problem sectors of their fields. This can lead to increased profits via increased yields and cost savings due to the ability to focus on distinct areas instead of performing uniform treatments. Traditional approach to measuring the crop yield from a field consists mainly of weighing harvested grains and calculating an average for a field. To facilitate supporting intra-field decision making, combine harvesters can be equipped with yield monitoring systems. There exists various methods of measuring the harvested yield. These methods include optical measurement and kinetic mass flow sensors. Additionally, yield monitoring systems utilize global navigation satellite system (GNSS) to assign location information to the measurements. Accurate yield maps are necessary to model intra-field yield variability [38].

As shown by van Klompenburg et al. in their review of 50 machine learning based crop yield prediction studies, performing yield estimation with input data from various sources is a current and developing research topic [41]. The use of spatial, i.e. geolocated yield information at the intra-field scale, crop yield data is becoming common. While the authors of the review do not examine the formats of used data w.r.t. spatially arranged yield targets, the notable presence of CNN architectures (36.4 %) indicates presence of spatial input-target pairs as training samples. This is in contrast to crop yield estimation studies, where crop yield information is aggregated over larger areas, such as counties [51, 78, 91].

Filippi et al. utilize 10 m/px resolution yield maps as target data in their study of crop yield estimation using multi-layered and multi-farm data with machine learning methods [13]. Yield information was initially generated by combine harvesters equipped with yield mapping sensors and was then processed to generate yield maps for the study. In addition to using yield maps as targets, yield maps from preceding years are also used as inputs with which the predictions are made in addition to other inputs. These other inputs include soil, satellite and weather data, all of them represented in spatial format in resolutions from 10 m/px up to 5 km/px.

Khanal et al. [37] performed soil variable and corn yield prediction at intra-field scale, utilizing combine harvester generated spatial yield maps as one of the target values. The authors utilized the size of the harvester head and the travelled distance of the combine harvester between each logged yield point to assign input pixels (multispectral data, various indices and DEM data) to certain yield values. Input and yield data were, thus, utilized in point-wise rather than spatial format w.r.t. modelling.

Similarly, Zhao et al. [104] utilized yield maps produced by combine harvesters as the target values for predicting wheat yields from raw and processed Sentinel-2 data. They derived various vegetation indices from multitemporal Sentinel-2 multispectral data, which were then utilized in a linear and multivariate time series model to estimate yields. While the input data was utilized as points, albeit initially spatial, the point-wise models were utilized to estimate yield maps from within-season satellite data .

3 SPATIOTEMPORAL DEEP LEARNING IN AGRICULTURE

Deep learning refers to models composed of multiple layers. Generally, a model is viewed as deep if it has at least an input layer, one hidden layer and an output layer. The term neural, on the other hand, refers to the fact that originally the operation principle of artificial neural networks was taken from that of the brain, containing neurons as its basic building blocks. As discussed by Tantalaki et al., the increasing volume of agricultural data from multiple sources calls for modelling techniques with an ability to perform automatic feature weighing and selection with complex and heterogeneous data [82]. Being non-linear and data-based, deep learning models have been recently more and more the modelling technique of choice in several application contexts.

The intention of the preceding chapter was to provide a broad overview of the data sources and their prevalence in the realm of agricultural data-based modelling. The goal of this chapter is to give the reader an overview on the tasks and problems in smart farming where deep learning structures have been successfully used and to provide enough background to understand the selection of particular models in the studies included in this thesis as well as their application contexts. With the publications of this dissertation focusing on spatial data, the discussion will be limited to spatial and spatiotemporal applications. In addition to considering recent relevant reviews, this chapter will also delve more into indvidual studies in terms of methods, application contexts and attained performance.

Thus, this chapter is constructed as follows. The first section is dedicated to reviewing studies focusing on deep learning and smart farming in general. The focus of the section is to build a contextual foundation of how deep learning has been recently utilized in agricultural context. After that, the following section and its subsections are dedicated to distinct model architectures. For each architecture, brief introduc-

tion is given on the operating principles. These introductions are then followed by reviews of distinct studies to provide the reader with an understanding of the possibilities and possible limitations of each architecture.

3.1 Deep learning in agriculture

The use of deep learning techniques in agriculture and agriculture-related remote sensing applications has gained a lot of attention recently. According to several reviews, the number of deep learning studies in the mentioned context has increased dramatically since 2015. According to a review of deep learning techniques in agriculture by Kamilaris and Prenafeta-Boldú, the number of deep learning related studies in the context of agriculture were virtually non-existent prior to year 2015 [31]. In a review of crop studies focusing on crop yield prediction using machine learning, the yearly distribution of studies is heavily focused on past two years [41]. Similar observation is made also in [89], where 76 out of 120 reviewed papers were published in 2019.

In a review conducted by Kamilaris and Prenafeta-Boldú, 40 deep learning and agriculture related studies were examined [31]. The authors identified 16 distinct applications for deep learning, including crop or weed detection (8), plant or crop type classification (4), plant recognition (4), fruit counting (4) and crop yield estimation (2). Out of the selected studies, 30 studies utilized computer vision based algorithms in some form. These algorithms include various custom-defined and pre-trained convolutional neural networks (CNN). Other algorithms present in the studies include long short-term memory networks (LSTM), autoencoders and a hybrid CNN-LSTM. They observe that, in addition to performance increases attained with the use of deep learning techniques, the need to pre-engineer independent predictor features is mainly eliminated. The models are generally seen as performant, albeit the training times are observed to be generally higher than with traditional machine learning methods. The need of large data sets is seen as a considerable drawback. Another data-induced limitation is the training data set's limited expressiveness of the underlying data producing phenomenon. They conclude that with image-like data, deep learning offers performant and reliable modelling techniques.

Tantalaki et al. also discuss the role of neural networks and deep learning in their review of data-driven decision making in agriculture [82]. They attribute the in-

creased use of deep learning techniques in agriculture partially to the models' ability to handle complex and non-linear agricultural problems. Their review is focused more on the developments of machine learning in the agricultural domain up until recent times. Neural network related techniques are separated to simpler artificial neural networks (ANN) and more complex image-based deep learning techniques. Out of 29 studies published between the years 1995 and 2018, they observe 15 studies utilizing ANNs and two deep learning related studies. ANNs have been utilized in crop, soil, weed, disease and weather related applications. Crop related studies, where both ANN and deep learning techniques were utilized, include yield estimation, type classification and feature estimation. Their general observation is that, with the developments in both data-based modelling techniques, IT infrastructure and data generation processes, deep learning is a prominent direction for data-based modelling.

In a review focusing crop yield prediction using machine learning, van Klompenburg et al. found 50 related studies starting from 2008 [41]. Out of them, 30 studies utilized deep learning in some form. In those studies, a total of 33 various deep learning architectures were present. The architectures included CNN, LSTM and deep neural networks (DNN), with the CNN being the most common at 11 occurrences and the latter being present 7 times. Spatiotemporal architectures were also observed in some studies, including three-dimensional (3D) CNNs and CNN-LSTM hybrid models. In addition to deep learning models, several traditional machine learning algorithms were also used. These include linear regression and ensemble models, such as decision tree based random forest models. These models, as discussed by the authors, are often used as benchmark models for their deep learning counterparts.

3.2 Performance metrics to evaluate yield prediction

Crop-related deep learning research in agriculture is still a developing field which is starkly illustrated by the variety of performance metrics used across various studies. Kamilaris and Prenafeta-Boldú identified 16 different performance metrics in their review of 40 agriculture-related deep learning studies [31]. The metric usage varies according to the modelling task (e.g. classification or regression) and formulation of the modelling problem (object recognition) [30].

Even with a specific task, i.e. crop yield prediction, at least 11 different perfor-

mance metrics were observed [41]. The most popular metric with 40 % usage according to [41] was the root mean square error (RMSE). While easy to calculate from the data, the interpretation of the metric is reliant on the knowledge of the value range and scale of true targets. This is true with other similar metrics, such as mean absolute error (MAE) and mean square error (MSE). Lower scores generally indicate better performance with these metrics. There are at least three key factors affecting the variability of the performance metrics. The first has to do with data preprocessing and, more specifically, data scaling where the use of absolute values produces different performance values from scaled values. The second factor is the environment with changes in annual weather patterns (local variability) and studies performed on similar crops but in different climates and soil characteristics (global variability). The third factor introducing variability in performance metrics are the crop varieties. While a modelling method would be comparable to an other method in terms of architecture and design, crop yields are different between different crop types (e.g. corn versus wheat). Xie and Yang identified at least 18 different types of crops while listing only a portion of all cereal crops [96]. A selection of crop types and their average values across continents are given in Table 3.1 to give a sense of scale for RMSE, MAE and MSE performance value interpretations.

Table 3.1 Average crop yields of 2018 by crop type and continent. Values obtained from *Our world in data* service's crop yields data explorer [60] and are given in tonnes per hectare.

Crop	Africa	Asia	Australia	Europe	North America	South America
Wheat	2.86	3.38	1.92	4.00	3.21	2.98
Barley	1.70	2.11	2.24	3.55	3.67	3.80
Rice	1.53	3.54	2.02	4.23	7.59	4.45
Maize	2.04	5.37	7.34	7.54	11.77	5.26

One of the most popular yield prediction performance metrics is R^2, the coefficient of determination, which was used in 26 % out of all crop yield regression studies reviewed in [41]. The general defition of the metric is

$$\begin{aligned} SS_{\text{tot}} &= \sum (y_i - \mu_y)^2 \\ SS_{\text{res}} &= \sum (y_i - \hat{y}_i)^2 \\ R^2 &= 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}} \end{aligned} \tag{3.1}$$

where y_i is a target value, μ_y is the average of all target values and $\hat{y}_i$ is a prediction corresponding to a target value y_i. The metric represents the degree of variance explained in the target value given the input features in a regression model. R^2 values range typically from 0 to 1 with 1 indicating a perfectly fit model and a score of 0.5 is interpretable as model being able to account for 50 % of variance w.r.t. the target values. Being proportion-based metric it is scale-invariant and thus produces comparable values as long as the compared data sets describe similar phenomena.

Another utilized scale-invariant metric is the mean absolute percentage error (MAPE). It is defined as

$$\text{MAPE} = \frac{\sum_i^n \frac{|y_i - \hat{y}_i|}{y_i}}{n} \tag{3.2}$$

and the value represents the average proportional deviation from true target values. As with RMSE, MSE and MAE, lower score indicates better performance. However, while MAPE is scale-invariant, the threshold of good performance depends on the data and requires domain knowledge for drawing deeper conclusions.

3.3 Spatial and temporal deep learning architectures

Data is one of the most crucial factors to take into account when selecting model architectures. Real-world spatial data requires architectures with capability to extract non-linear spatial features. Sequential data requires models capable of modelling change in the dimension of time, or change in general. Time series of spatial data require architectures with both of the aforementioned capabilities.

Deep learning has been utilized in multiple smart farming related applications as shown in the previous section. In this section the focus is narrowed to discuss the model architectures relevant to spatial and spatiotemporal modelling. The publications selected for this dissertation mainly utilize spatial data. Publications [I-II] utilize UAV-based remote sensing data as spatial inputs. An input conforms to a distinct point in time. Publication [IV] utilizes time series of UAV-based spatial, i.e. spatiotemporal, data. While publication [V] makes use of temporally distinct spatial data, the distinction to publications [I-II] is in the use of multiple data sources. The only publication where spatial or spatiotemporal modelling is not utilized is [III], where the data, albeit initially spatial, is utilized in row-like manner.

3.3.1 Convolutional neural networks

Convolutional neural networks, often referred to as CNNs, have solidified their place in modeling tasks where the input data is either spatial or spatially representable [42, 81]. The main component of the model is the convolution operation, where a set of trainable kernels (or filters) is applied to the input data resulting in a set of spatial features describing the data. The model learns basic features in the first layers and composite features of these basic features at further layers [103]. To help the model better learn these features, batch normalization can be applied to the inputs [26]. The final output of a plain CNN is a set of feature maps. Depending on the use case, these can be either directly utilized or, for example, flattened and fed to a fully connected (FC) layer for regression or classification purposes. Traditional use case is to employ a CNN to extract spatial features from two-dimensional (2D) inputs. While spatial data normally contains multiple channels, the 2D kernels are applied to each channel, separately. CNNs can also be utilized in non-spatial manner when input data is tabular, i.e. row-like. The convolution operation is then applied one-dimensionally (1D), with the kernels operating on adjacent values in the row as defined by the kernel size.

As shown in multiple recent reviews of deep learning method utilization in agriculture and smart farming, CNNs constitute the majority of modelling approaches [31, 41]. Kamilars and Prenafeta-Boldú conducted a review of studies utilizing CNNs in the domain of agriculture and published between the years 2014 and 2017 [30]. They selected a total of 23 studies for closer inspection. They observe that CNNs were utilized mainly for classification tasks related to weed identification and fruit counting. Yield prediction is also observed as a major task. The studies divided evenly between using pre-trained models (12 studies) and custom-developed and trained architectures (11 studies).

In a study of CNN-based yield response modelling to crop management activities in the US, Barbosa et al. developed four different CNN architectures to predict corn yield using medium-resolution multisource input data [2]. The input data consisted of rates for nitrogen and seed, elevation maps, soil's electroconductivity and 3 m/px resolution commercial satellite data. The data used in the study was not multitemporal, meaning that each input represented a single point in time. Three of the models were based on 2D convolutions, differing on how the input data were

introduced to the model and how the model features were combined to produce the final output. Last of the models was a 3D CNN, which will be discussed in later in Section 3.3.5. In the study the authors tested several traditional ML models and each of the CNN architectures. The best performance was attained with a CNN extracting spatial features from distinct spatial input channels. They reported best performance of 0.70 root mean squared error (RMSE). The RMSE is a proportion of the standard deviation of crop yields, translating to 1140 kilograms per hectare (kg/ha) on average when unscaled.

Tedesco-Oliveira et al. utilized several existing CNN-based architectures to classify cotton bolls and predict cotton yields from high-resolution images taken manually at hand-held heights [83]. One of the employed models was a two-stage Faster R-CNN [59], which first proposes areas of interest and then identifies target objects from the areas. They also employed a CNN-based Single Shot Multibox Detection (SSD)[53] algorithm designed for recognition of multiple objects from images. Last employed was a lighter version of the SSD algorithm optimized for mobile devices, MobileNetV2 [68]. The Faster R-CNN architecture performed the best, having average recall of 0.66 and average precision of 0.59. Using spatial models as automatic boll counters, a linear regression model was trained to predict yield from automatically acquired image data. The yield prediction model utilizing CNN's outputs attained 17.86 % mean absolute percentage error (MAPE).

Yang et al. developed a CNN acrhitecture to separately utilize RGB and vegetation indices derived from multispectral data for predicting rice grain yield and ripening stages from high-resolution UAV-based images [98]. For the RGB data, the CNN consisted of five distinct convolutional layers. The CNN using vegetation indices had three convolutional layers. Their approach to yield estimation was to transform the regression problem to a classification problem by assigning a class label to a distinct yield range. While they report fluctuations in performance, assigning the cause to the variability in the input data, they report their best model attained 20.4 % MAPE and 0.585 for the coefficient of determination (R^2).

Kang et al. evaluate several machine learning and deep learning algorithms for maize yield prediction in the US with various data aggregated to county scale [33]. They utilized various low-resolution spatial data sources in addition to governmental and institutional sources for the environment. The number of input variables was between 58 and 891 depending on the experimental variable selection scheme. While a

CNN was utilized, they used it for tabular data, performing 1D convolutions. With tabular data, the CNN achieved 10.1 % MAPE, while the best performing model, a gradient-boosted decision tree called XGBoost [7], achieved 9.1 % MAPE. Regarding CNN's performance, the authors acknowledge that the architecture is designed to be used with spatial data while their data were non-spatial, albeit containing data from spatial sources pre-aggregation.

3.3.2 Long short-term memory networks

The Long Short-Term Memory (LSTM) networks, originally introduced in [24], have been widely utilized in sequence modeling tasks [70]. LSTMs belong to the deep learning architecture family of recurrent neural networks (RNN). LSTM generally operates with vector-like inputs, which include tabular data and vector outputs from other models. There are two general concepts to the LSTM that help it in learning temporal features from the data. The first is the concept of memory, introduced as the cell state. The other is the concept of gates, effectively trainable FC layers, manipulating this cell state in response to the new inputs from the data and past outputs of the model. To handle sequences of data, the model loops over the sequences altering its cell (C) and hidden (H) states in the process using the combination of learned parameters in the gates and non-linear activations when combining the gate outputs. LSTMs can also be employed in bidirectional and stacked form. Bidirectional LSTMs train an additional model in comparison to the unidirectional LSTM. One LSTM reads the input from start of the sequence to end ($t_0 \rightarrow t_n$), while the other reads the input from end to start ($t_n \rightarrow t_0$). The outputs of these two parallel models are then combined as a final sequence of temporal features [71]. When LSTMs are stacked, the first LSTM operates on the input sequence and subsequent LSTMs operate on sequences of feature vectors produced by preceding models. Bidirectionality helps the model learn features from both sides of input sequences, while stacking helps in learning higher level temporal features [21].

Jiang et al. developed a phenology-based LSTM for estimating corn yields in the US at county scale [28]. Using phenological data derived from remote sensing, crop information, meteorological and topographical sources, they aggregated tabular data to predict county-wise corn yields. The data was temporal, consisting of five time steps corresponding to distinct maize growth stages during the growing season. To

predict the corn yields, they developed a stacked LSTM architecture with two layers. They observed that with longer sequences the LSTM model achieved better results in predicting the crop yield, rising from 0.45 R^2 with single year of observations to 0.76 R^2 with ten years of observations. Inversely, the RMSE decreased from 1450 kg/ha to 870 kg/ha using one and ten years of observations, respectively. Best results were achieved with multiple years of training data, using 10 preceding years of data to predict the 11th year.

Kang et al. also evaluated the effectivity of an LSTM in their assessment of machine learning methods for maize yield prediction in the US [33]. The outline of the study is described in the subsection 3.3.1. With the sequential LSTM model, they report 9.1 % MAPE and approximately 15.0 bushels per acre (bu/ac) RMSE at best. As the data of the study was processed to be tabular, the sequential model was, at best, on par with the generally best performing model, the XGBoost. The authors however note that the overall performance of the LSTM was only slightly better than what the CNN attained.

Lin et al. utilized weekly aggregates of meterological factors to predict crop yield anomalies with a LSTM-based model [51]. Their LSTM architecture consisted of three stacked LSTMs, followed by an attention mechanism. As the study area, the US Corn Belt region, is vast, the authors have divided the inputs to spatially distinct regions. Corn yield estimates are generated separately for each region in the final FC layers of the model. The LSTM-based model attained 820 kg/ha RMSE and 0.76 R^2 score. The authors also trained two traditional machine learning models, a random forest and a lasso regression model. Their respective performance metrics were 1050 kg/ha RMSE (0.60 R^2) and 1140 kg/ha RMSE (0.53 R^2).

Schwalbert et al. studied the use of LSTMs in satellite data based soybean yield estimation in southern Brazil [72]. They utilized data from multiple sources, including low-resolution MODIS satellites, weather data and soybean yield information. Data was acquired for multiple years, from 2002 to 2016. The temporally sequential data was arranged to tabular form, aggregated according to municipalities by averaging. Time step between samples in a sequence was 8 days with sequence start at mid-October. They trained three models in total: an LSTM and two benchmark models, ordinary least squares (OLS) linear regression and random forest. As input sequences were grown longer by shifting the sequence end from January 16 to March 5, the LSTM outperformed both baseline models with the best performance

at 320 kg/ha RMSE. However, with shortest input sequence lengths both of the baseline models exhibited better performance. With the shortest sequence, the OLS model attained 530 kg/ha RMSE and the random forest 570 kg/ha RMSE, while the LSTM achieved 680 kg/ha RMSE. The range of observed soybean yields was from 200 kg/ha to 4200 kg/ha.

3.3.3 Hybrid CNN-LSTM

Due to the spatial nature of remote sensing data closely related to the task of crop yield estimation, the LSTMs (and its variants) are often coupled with spatial feature extracting CNNs. The hybrid CNN-LSTM is a composite model consisting of a spatial feature extractor or transformer, i.e., a pretrained CNN, and a temporal model, the LSTM [66]. The ability to perform temporal modelling with spatial data is often necessary with, for example, multitemporal remote sensing data. The general idea is to both gain the ability to utilize spatial data and perform sequential modeling with LSTM networks. Instead of feeding the final outputs of a CNN to an FC layer for regression or classification purposes, the CNN output is fed as an input to a sequential LSTM model. The final regression or classification result is produced from the features outputted by the LSTM.

Khaki et al. built and trained a hybrid CNN-LSTM model to predict corn and soybean yields in the US Corn Belt area [36]. In their model the spatial inputs, namely soil and weather data at 1 km/px resolution, were first processed by CNNs to extract vectors of high-level spatial features. These features were then fed to the LSTM alongside management and previous year's yield data. They used yearly data from 1980 to 2015 to train the models. The model attained yield prediction average correlation coefficients of 87.3 % and 86.2 % for corn and soybean, respectively. Comparing to the benchmark models, the results were, on average, 17.4 and 25.9 %-units higher for corn and soybean yield prediction, respectively again. The authors observe that the CNN-LSTM model is able to efficiently perform feature selection from large feature space. They also observe that the model generalizes well to unseen samples.

Yaramasu et al. utilized a CNN-LSTM architecture to extract spatiotemporal features from nationally generated crop type maps in the US [99]. The extracted features were then fed to a decoder to reconstruct an estimate crop type map for an

upcoming year. The spatial feature extracting CNN of the spatiotemporal encoder was a pre-trained VGG11 [76]. Their input data consisted of medium resolution, 30 m/px, spatial crop type classification maps spanning the US continent. Using year as the time step, they trained the model to predict a crop map for a 512×512 px or 236 km^2 area based on crop type changes in the preceding years. They achieved average overall accuracy of 77 %.

Yue et al. also developed spatiotemporal encoder-decoder architectures to predict progressions of meterological factors, such as daily precipitation [100]. Encoder-decoder architecture first learns to compress the data and the feature-wise interactions to a high-level vector or matrix representation, and the decoder then is used to recreate the desired target from this encoded output [90]. The encoder and decoder are jointly trained to facilitate extraction of robust high-level features. Daily data was utilized from a year to predict the progression of distinct meterological variables for the year following the input data. These predictions were then further utilized to estimate maize growth stages. While the focus of their study was a convolutional LSTM based encoder-decoder architecture, they also trained a CNN-LSTM based encoder-decoder for comparative purposes. Their input data consisted of tabular meteorological data, meaning the convolutional operations were 1D. With other models also trained, of the encoder-decoders the CNN-LSTM was able to surpass the best performing convolutional LSTM in daily cumulative precipitation prediction with MAE of 3.33 mm against 3.88 mm, respectively.

Rustowicz et al. built a multisource CNN-LSTM architecture to classify crop types in Ghana, South Sudan and Germany from satellite data [65]. Using high-resolution PlanetScope and medium-resolution Sentinel-1 and Sentinel-2 satellite data, they build an architecture capable of utilizing time series of spatial data from each satellite source. The model consist, thus, of three CNN-LSTMs for each input source. The outputs of these models are then concatenated to generate the final model output, crop type classification result. The authors also built and trained a spatiotemporal 3D-CNN model, discussed in Section 3.3.5. A random forest model was trained as a baseline model. The CNN-LSTM performed the best with majority of crop types in Germany and Ghana, attaining 95.8 % and 59.9 % respective overall accuracies. For South Sudan, the baseline RF performed the best with majority of crops with average accuracy of 88.7 % while the CNN-LSTM model attained 82.6 % accuracy.

Sun et al. designed and trained a hybrid CNN-LSTM architecture to predict

county-level soybean yields in the US [78]. The input data used was multiyear and multisource, consisting of data from MODIS satellite system, weather information, crop yield statistics and county boundary information. A sequence of yearly spatial inputs consisted of 34 time steps with 8 day span between steps. The CNN extracts spatial features from each spatial input of the whole input sequence. The sequence of spatial inputs is thus transformed to a sequence of high-level feature vectors. The sequence of vectors is then the input for the LSTM. In addition to using full sequences, the authors examine the model's prediction performance with shorter sequences from the beginning of a season. The CNN-LSTM attained 0.78 R^2 with full sequence and 0.74 R^2 using ten 8-day time steps from the beginning of the season.

3.3.4 Convolutional LSTM

Convolutional LSTM [74] is a model combining the features of convolutional and sequential models into a single architecture, using convolutional layers (convolution with pooling etc.) as the LSTM's gate functions. This makes it possible to feed the sequential model the spatial data directly. Akin to how convolutional networks learn, the gates learn to utilize the convolutional kernels to provide the best set of spatial features when building and modifying the cell state C. Thus, contrary to the CNN-LSTM, no pre-extraction of spatial features is required. While LSTM has been predominantly utilized in agriculture related sequential modelling tasks [31, 41], convolutional recurrent architectures have also been employed with a gated recurrent unit (GRU) [9]. The general difference between LSTM and GRU is the number of gates and, thus, trainable parameters. Convolutional layers are present in both with their convolutional recurrent variants.

Yue et al. studied of maize growth stage prediction with encoder-decoder model architectures and meterological data, focusing on a convolutional LSTM based model [100]. The setting of the study is described in Section 3.3.3. They authors decided to utilize 1D convolutions in the convolutional LSTM's input gate functions to extract features depicting complex interactions within the meteorological data. The authors report the convolutional LSTM encoder-decoder architecture performing the best in majority of meteorological factor estimation cases: 2.60 ^{o}C MAE and 3.46 ^{o}C RMSE for average temperature, 3.88 ^{o}C MAE and 10.50 ^{o}C RMSE for daily cumulative precipitation and 3.45 ^{o}C MAE and 4.17 ^{o}C RMSE for daily sunshine duration. The

convolutional LSTM achieved the closest average predicted-to-real ratio of 0.91. The next best average ratio of 0.75 was attained by a gated recurrent unit (GRU) based model.

Rußwurm and Körner sought to utilize the temporal nature of spatial earth observation satellite data in pixel-wise crop type multi-class classification by designing bidirectional convolutional recurrent models [63]. Compared models utilized either LSTM or GRU as the recurrent model. Using Sentinel-2 data interpolated to 10 m/px, they gathered data from a 4300 km^2 area in southern Germany, subdividing it to smaller 15 km^2 areas. Time was encoded to the data by introducing the information of the year and day-of-year of the satellite fly-over. The authors observed that the performance of both models were so similar, that they reported only the performance of the GRU-based convolutional recurrent model. They report the model attaining a 89.6 % average classification with data from 2016 and 2017. They conclude that their model is able to attain state-of-the-art accuracy without common satellite data preprocessing, such as atmospheric correction or cloud identification.

Russwurm and Körner further investigate the robustness of a single layer convolutional LSTM to clouds in satellite time series data [64]. The data is from the same area and overall similar to [63]. Training the model with similar data, they performed ablation experiments with varying degrees of cloud coverage. They observe that the convolutional LSTM is indifferent towards occasional cloud coverage. With images sorted according to their cloudiness from no clouds to up to half-cloudy, the model was with all cases capable at achieving between 90 % and 93 % accuracy. The accuracy was determined with regards to crop type classification. The authors conclude that clouds are effectively noise in temporal data and the convolutional LSTM is able to account for that.

Ienco et al. develop a multisource CNN and convolutional GRU based model to classify land cover from multitemporal Sentinel-1 and Sentinel-2 data [25]. The model consists of two identical architectures for distinct satellite sources. An architecture contains a CNN and a convolutional GRU, both producing a high-level feature vector of set length. The vectors are concatenated and then fed to a FC layer for classification. The satellite time series data was acquired of two distinct sites in Reunion Islands and Burkina Faso. Using only the convolutional GRU models, they attained 88.2 % accuracy on pixel-wise land type classification. Comparatively, using the CNN models only produced 87.7 % accuracy. The full model attained 89.9 %

accuracy.

Liu et al. utilized a bidirectional convolutional LSTM with hyperspectral data to perform spectral-spatial land cover classification [52]. Using three distinct satellite-based hyperspectral image data sets, the authors utilize the sequential nature of the convolutional LSTM to learn inter-spectral high-level spatial features from the input data. They compared the model to several other architectures, including CNN, LSTM, 3D-CNN, and CNN-LSTM. Best results on 16-class land cover classification were achieved with bidirectional convolutional LSTM, 3D-CNN and CNN-LSTM with respective average accuracies being 97.1 %, 95.2 % and 94.5 % over all data sets.

3.3.5 Three-dimensional CNN

As initially reported by [86], 3D-CNNs performed remarkably well in modeling tasks involving spatiotemporal data. Being CNNs, the 3D-CNNs utilize all same architectural features as more commonly used convolutional models. What's different is their use of convolution in the depth dimension, searching for robust features across sequences of input data in addition to spatial features extracted from the individual images. The sequential nature of input data is not limited to time, but can also be, for example, hyperspectral multi-layer point-in-time data with the aim of finding high-level inter-channel features [50].

Barbosa et al. developed and trained multiple models to predict crop yield from spatially formatted crop management data [2]. The outline of the study has been elaborated in Section 3.3.1. One of the models was based on 3D-CNN architecture, consisting of a single 3D convolutional layer coupled with two FC layers. They utilized the 3D-CNN to extract inter-channel high-level features from the inputs. The 3D-CNN model was observed to perform on-par with majority of other CNN-based solutions, attaining 0.73 RMSE, which translates to 1190 kg/ha when unscaled.

Terliksiz and Altylar studied the use of a 3D-CNN architecture to predict soybean yields from spatiotemporal data at county-scale in the US [84]. The input data was acquired annually from 2003 to 2016 from the spatially low-resolution MODIS satellites at 8-day intervals between each data sample. A single sequence consisted of 24 subframes cropped from the initial, larger area. The model implemented by the authors consists of two initial 2D CNN layers, followed by six 3D convolutions and two FC layers for single value prediction. Their model attained an average 4.42

bu/ac RMSE with various land cover ratio ablations in input frames. Comparing to other similar studies, their result is, at minimum, 0.90 bu/ac RMSE lower. They authors discuss, however, that the use of within-county smaller frames to predict a county-wide average yield can be misleading.

In their study of crop type classfication in Germany, South Sudan and Ghana, Rustowicz et al. also built and trained a 3D-CNN encoder-decoder architecture to classify pixels in sequences of frames extracted from satellite data [65]. The study is described in Section 3.3.3. The 3D-CNN model performed close to the CNN-LSTM model, attaining 95.2 %, 60.9 % and 85.3 % overall accuracies with Germany, Ghana and South Sudan data sets, respectively. While CNN-LSTM performed better with Germany by 0.6 %-units, the 3D-CNN had better overall accuracy with Ghana and South Sudan by 1.0 %-units and 2.7 %-units, respectively. With ablation studies, the best setting with 3D-CNN outperforms the best CNN-LSTM setting 1.3 %-units with South Sudan data set in terms of accuracy, while CNN-LSTM attains 2.2 %-units higher accuracy with Ghana data set.

Ji et al. developed a 3D-CNN model to perform crop type classification with spatiotemporal data [27]. They use data from 4 m/px resolution Gaofen 1 and 15 m/px Gaofen 2 satellite systems, acquired for several months from 2014 to 2016. The model constisted of three 3D convolution layers followed by two FC layers. The 3D-CNN outperforms other methods in pixel-wise classification, including a CNN and several traditional machine learning methods. The average overall accuracy of the model is 94.9 %. Two closest contender models, a CNN and a support vector machine (SVM), attain 91.8 % and 91.9 % average overall accuracies, respectively.

Li, Zhang and Shen studied the use of 3D-CNN models in spectral-spatial land cover classification with hyperspectral data [50]. The authors use three hyperspectral data sets from Italy, Botswana and India. A two-layer 3D-CNN architecturethe is built and trained to perform pixel-wise classification of remotely sensed scenes. The model is compared against models developed in other studies and similar settings, including a stacked autoencoder, deep belief network and a CNN. The 3D-CNN attained 99.3 % overall accuracy. With notable differences (>2 %-units) in prediction accuracies only observed with the Indian data set, the authors notice the 3D-CNN having the lowest misclassification ratio out of the compared models. While spatial model were observed to perform the best overall, the authors attribute the best performance of the 3D-CNN to its ability to learn salient inter-spectral features from

the input data.

4 CROP YIELD PREDICTION WITH DEEP LEARNING

Having established a solid background of crop yield prediction with deep learning models, it is time to review the contributions of the publications selected for this dissertation. The objective of this dissertation is to probe and answer the research questions outlined in Chapter 1. RQ1 and RQ2 provide a suitable line of division for the selected publications. The feasiblity of spatial and spatiotemporal deep learning models in-season yield prediction with high resolution remote sensing data is explicitly studied in publications [I] and [IV]. In both, the models are designed, trained and evaluated from scratch to perform intra-field crop yield prediction with UAV-based data, within a growing season. Furthermore, [II] uses the model of [I] in a case study to frame the use of such models in a farming DSS. Two remaining publications, [III] and [V], focus on data sources. In [V], the effects of additional data sources were evaluated by comparing crop yield estimation performance with a static model architecture from [I]. Publication [III], while taking a distinct approach in terms of the modelling technique used, focuses on the reliability of satellite-based remote sensing data.

This chapter focuses on the data, methods and results of the selected publications, leaving discussion of the results and methods to Chapter 5. The chapter consists of two main sections and is constructed as follows. In the first section, the developments and evaluations of intra-field crop yield prediction models are described. The section begins by looking at CNN-based yield prediction with distinct point-in-time frames presented in [I]. A frame is a sub-area extracted from larger images. Next, the evaluation of the usability of the best model of [I] in the farming DSS context is presented in [II]. Lastly, the development and evaluation of multiple spatiotemporal deep learning models in [IV] to perform crop yield prediction using UAV-based data is described. The second section focuses on data evaluation. The general outline

of [V] is first described, where the effects of varying the input data source configurations on crop yield prediction performance were studied. The process of using machine learning to evaluate satellite data reliability with regards to cloud canopy [III] is described last.

4.1 Intra-field crop yield prediction

4.1.1 Single input to single target

In working towards an effective in-season crop yield predictor model for the northern climate, the effort in [I] was to develop a CNN based deep learning framework using UAV-acquired multispectral data. RGB and NDVI images were fed as input data. The best perfofming CNN configuration in terms of architectural composition and hyperparameters (parameters defining the training setup, was iteratively developed via a tuning process).

The nine crop fields selected for this study are located in the vicinity of the city of Pori (61°29'6.5"N, 21°47'50.7"E). The total area of the fields was approximately 90 ha. The main crops grown in the fields were wheat and malting barley, however the model was trained over the fields without making a distinction between the crop type. Details of the fields, crops, imaging dates and corresponding growth phases are listed in Table 4.1. Thermal times for each crop variety are taken from a [44]. Sowing dates and imaging dates are used to calculate the growth phase as a fraction of the total thermal time for the crop variety. Images with dates prior to 1st of July form the early data set and the remaining images the late one.

Multispectral data was acquired from these fields during the growing season of 2017. The data was collected with a single Airinov Solo 3DR (Parrot Drone SAS, Paris, France) UAV equipped with NIR-capable SEQUIOA (Parrot Drone SAS, Paris, France) sensor. The images of individual spectral bands were stitched together to form complete orthogonal RGB and NDVI rasters of distinct fields using the Pix4D software.

The harvest yield data was acquired during September 2017 using two distinct setups attached to the harvesters: Trimble (Sunnyvale, California, USA) CFX 750 and John Deere (Moline, Illinois, USA) Greenstar 1. The data was initially in vector data point format. The points were first filtered according to [85] to preserve only

Table 4.1 Details of crops and their varieties sown in each of the 9 fields in 2017 (reproduced from [I]).

Field #	Size (ha)	Mean yield (kg/ha)	Crop (*Variety*)	Thermal time	Sowing date	Imaging date	Growth phase
1	5.96	5098	Wheat (*Zebra*)	1052	10 May	17 Aug	83 %
2	10.26	6054	Barley (*Trekker*)	979.7	16 May	8 Jun	15 %
						27 Jul	64 %
3	2.97	8971	Barley (*Trekker*)	979.7	17 May	8 Jun	15 %
						27 Jul	64 %
4	13.05	4673	Barley (*RGT Planet*)	982.2	15 May	6 Jul	42 %
5	4.66	6482	Barley (*Propino*)	981.4	15 May	15 Jun	22 %
6	7.29	6884	Barley (*Propino*)	981.4	15 May	15 Jun	22 %
7	10.92	7568	Barley (*Harbinger*)	976.3	24 May	6 Jul	36 %
8	15.28	7585	Barley (*Trekker*)	979.7	18 May	1 Jun	10 %
						13 Jul	49 %
9	18.86	6991	Wheat (*KWS Solanus*)	1065	13 May	15 Jun	21 %
						6 Jul	72 %

points corresponding to harvester speed between 2 and 7 km/h and yield between 1500 and 15000 kg/ha. Then the field-wise data was rasterized by interpolation using an exponential point-wise inverse distance algorithm.

The field-wise image data was then processed using a sliding window to extract geolocationally matched pairs of input RGB and NDVI image frames and frame-wise averaged crop yields as targets of predefined size from all the fields. The step of the applied sliding window was chosen to be 10 m according to the resolution of Sentinel-2 satellite data considering the possibility of using satellite data as an additional input to the network in future studies. In other words, frames having sides longer than 10 m share data with adjacent frames due to overlapping. The resolution of the UAV data was 0.3125 m/px or 32 px per 10 m. Square image frames with side lengths of 10m, 20 m and 40 m were considered. The number of extracted frames was approximately 15200 for each frame dimension. All nine fields were first split into overlapping data frames of sizes 10 m, 20 m and 40 m. A dedicated holdout test data set was then built from 15 % of shuffled data frames; this data was never presented to the model during training. The remaining 85 % of data frames were then used for training the models with k-fold cross validation. After the training phase of each model was completed, the test errors were calculated using the holdout test

data set to validate the performance of the trained model. This process is illustrated in Figure 4.1.

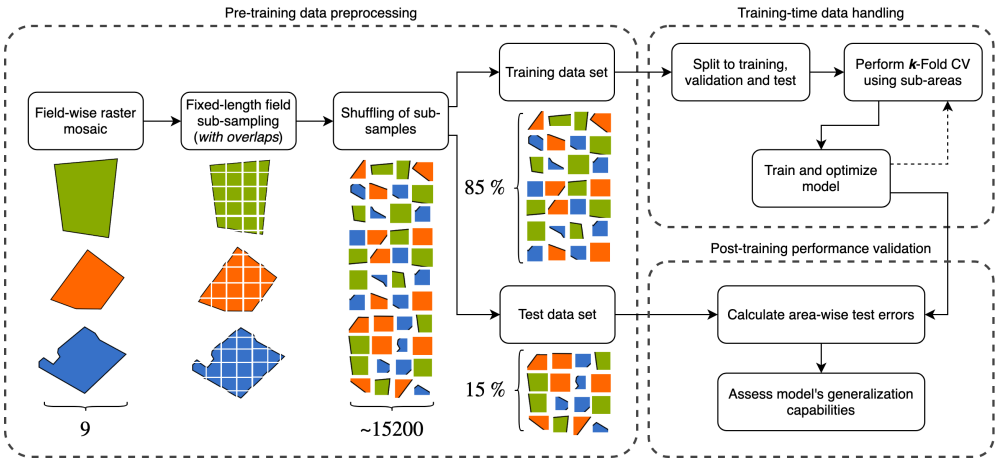

Figure 4.1 The process of data preparation prior to and during training (reproduced from [I]).

The basic architecture of the implemented CNN model follows closely the one reported in [43]. The general topology of the network is depicted in Figure 4.2. The model was implemented using the PyTorch framework [57]. The model's inputs can be single-band or multi-band images (B) with varying dimensions (D). The network has at least two convolutional layers accompanied with two fully connected layers. The depth of the network is controlled by the number of intermediary convolutional layers. The last convolutional layer has 128 kernels while the intermediary layers have 64 kernels. Max pooling is applied only in the first and last convolutional layers so that the size of the data representation stays consistent when network depth is varied. The model uses non-overlapping pooling windows with pooling window size of 5 and a pooling stride matching the pooling window size. Pooling is applied only in the first and the last convolutional layers. Rectified linear units (ReLU) [22] are used for layer-wise non-linear activation functions. This way our network is also scalable with respect to the number of layers. Two FC layers with 1024 neurons per layer are used to produce the final output from the CNN outputs.

Finding the optimal configuration of any deep learning network is an iterative process, where the model's parameters are initialized and tuned multiple times. The best training algorithm was evaluated among three options: Stochastic Gradient Descent with momentum (SGD-momentum) [6], RMSprop [23] and Adadelta [102]

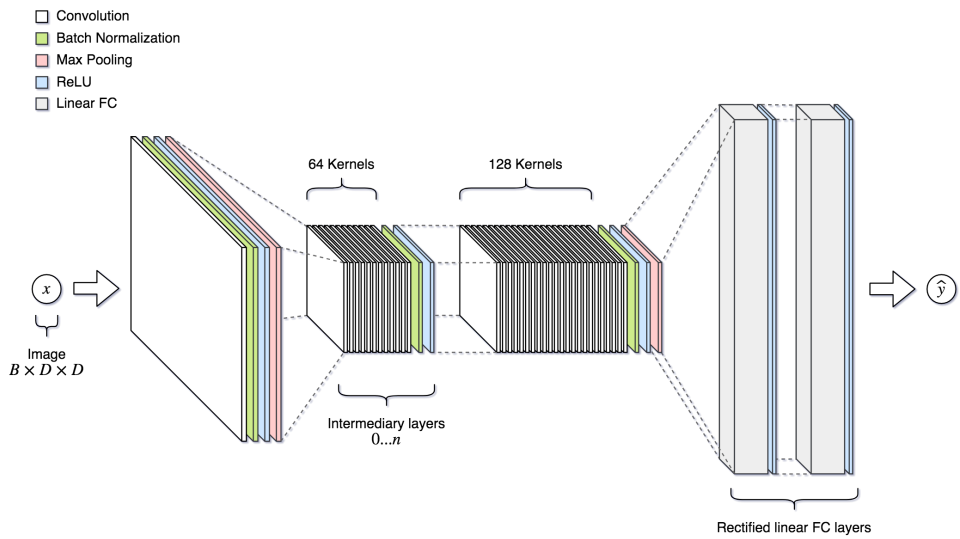

Figure 4.2 The overall topology of the implemented CNN (reproduced from [I]).

as suggested in [19] and present in [34]. Out of these, Adadelta performed the best. Optimizer parameters, such as learning rate, coefficient of using past gradients during backpropagation and weight decay, were tuned using random search [4]. Other parameters were tuned by testing various values from predefined selections. These parameters were

- input sample batch size (from 2^5 to 2^{10})
- the number of intermediate convolutional layers (from 4 to 12)
- input data type (NDVI or RGB)
- frame side length (10m 20m or 40m)
- early stopping patience (from 10 to 50).

The lowest test set error of 484 kg/ha MAE and 8.8 % MAPE was achieved using RGB data from the beginning of the growing season, i.e. pre-June images. R^2 score was 0.857. The best model consisted of six convolutional layers followed by two fully connected layers, using weight decay regularization with the coefficient being 10^{-3} and early stopping with patience of 50. The optimizer was also tuned, with the optimal values for learning rate and the coefficient adjusting the effect of past iterations' error corrections being 8×10^{-3} and 0.58, respectively. The results show that the lowest test errors were achieved with the largest frame side length of 40 m.

The best performing model was also utilized in a case study of six fields containing mostly barley, together accounting for 54.2 ha of land area and located near the city of Pori [II]. Imaging and yield acquisition methods were identical to [I], with some overlaps in the field-wise data. The crop yield prediction results indicate a consistent pattern of overestimating low yields and underestimating high yields. This is shown in Figure 4.3, where all values are absolute and in kg/ha. Over the six fields, the model attained 0.798 R^2, on average. Field-wise MAPE boxplots are depicted in Figure 4.4.

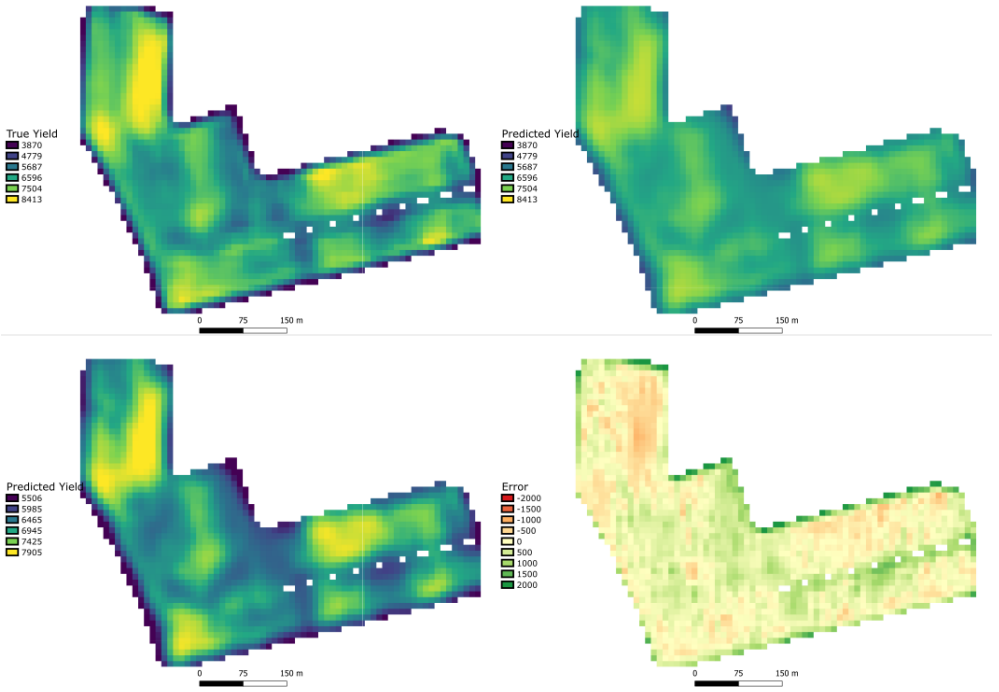

Figure 4.3 Visualisation of the true and predicted yield of a field (reproduced from [II]). Images of true and predicted yields in the top row share a similar scale. Bottom left image is scaled to predicted values only. Bottom right image depicts the error between true and predicted yield. Units are in kg/ha.

The model was also designed to be utilized as an AI engine in an Oskari-based (*www.oskari.org*, MIT & EUPL licensed) geospatial data mapping and farming decision support system. Through the web portal farmers can access their personal, authenticated accounts, upload data for visualization and call on AI based analytical tools for decision support.

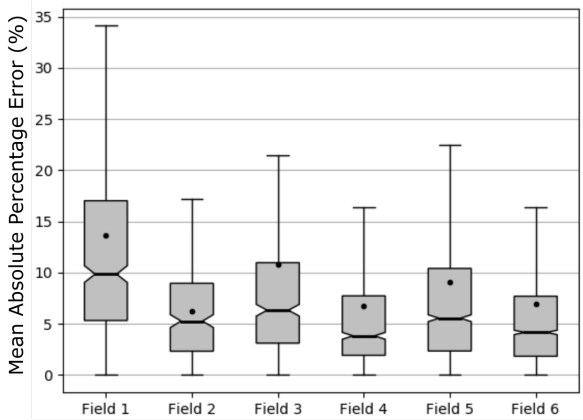

Figure 4.4 Boxplots of percentage error between true yield and predicted yield for each field (reproduced from [II]).

4.1.2 Sequence of inputs to single target

In [IV] we examined the effect of time, as an additional feature, on intra-field yield prediction. Especially, capabilities of deep learning time series models utilizing UAV remote sensing time series data as their inputs were focused on. The objectives were two-fold: to see if the performance of the point-in-time model of [I] could be surpassed using spatiotemporal deep learning model architectures and to see which spatiotemporal architecture would perform better in the same task. The usability of spatiotemporal models was evaluated in two settings, end-of-season (full sequence) and in-season (limited sequence) prediction. Three model architectures were designed, trained and evaluated: a CNN-LSTM [66], a convolutional LSTM [74] and a 3D-CNN [86]. These models utilize the properties of CNNs and LSTM networks them to perform spatiotemporal modelling. The main contribution of the study was to perform time series based intra-field yield prediction with multi-temporal data collected during the growing season with UAVs.

Nine crop fields totaling to approximately 85 ha and having wheat, barley and oats as the crop varieties, were included in the study. The field-wise data was acquired during year 2018 in the proximity of Pori, Finland (61°29'6.5"N, 21°47'50.7"E). Specific information about the fields is given in Table 4.2. The acquisition of input and target data was similar to [I].

Images of the fields were acquired with a SEQUIOA (Parrot Drone SAS, Paris,

Table 4.2 The fields selected for the multitemporal study in the proximity of Pori, Finland (reproduced from [IV]).

Field #	Size (ha)	Mean yield (kg/ha)	Crop (*Variety*)	Thermal time	Sowing date
1	11.11	4349.1	Wheat (*Mistral*)	1290.3	13 May
2	7.59	5157.6	Wheat (*Mistral*)	1316.8	14 May
3	11.77	5534.3	Barley (*Zebra*)	1179.9	12 May
4	11.08	3727.5	Barley (*Zebra*)	1181.3	11 May
5	7.88	4166.9	Barley (*RGT Planet*)	1127.6	16 May
6	13.05	4227.9	Barley (*RGT Planet*)	1117.1	19 May
7	7.61	6668.5	Oats (*Ringsaker*)	1223.4	17 May
8	7.77	5788.2	Barley (*Harbringer*)	1136.1	21 May
9	7.24	6166.0	Oats (*Ringsaker*)	1216.4	18 May

France) multispectral camera mounted on a Airinov Solo 3DR (Parrot Drone SAS, Paris, France) UAV on a weekly basis for 15 consecutive weeks. To encode passing of time for the temporal models, weather data was acquired from the open interface provided by the Finnish Meteorological Institute for Pori area. Being a common way to express crop growth phase, the cumulative temperature was utilized as the temporal feature in the input data. Temporally varying but spatially constant cumulative temperature was added as an additional layer in conjunction with the RGB layers to have the data contain necessary information for temporal feature learning. The targets data, crop yields, were acquired during the harvest of each field. The harvesters were equipped with either a Trimble Navigation (Sunnyvale, California, USA) CFX 750 or John Deere (Moline, Illinois, USA) Greenstar 1 yield mapping sensor systems, which produce a cloud of geolocated points with multivariate information about the harvest for each point in vector format.

The fields were split into smaller overlapping frames of size 40×40 m with a lateral and vertical step of 10 m. Sequences of frames of fixed width and height were extracted from sequences of field plot images and corresponding weather data as the input data. The input frames were then geolocationally paired with corresponding yield data to form input-target pairs. A total of 2586 sequences, 15 geolocationally

matching frame rasters per sequence, were extracted from the data. Lastly, data was shuffled and split to training and test sets with a 70%/30% ratio, respectively. The general process of generating the frames is depicted in Figure 4.5.

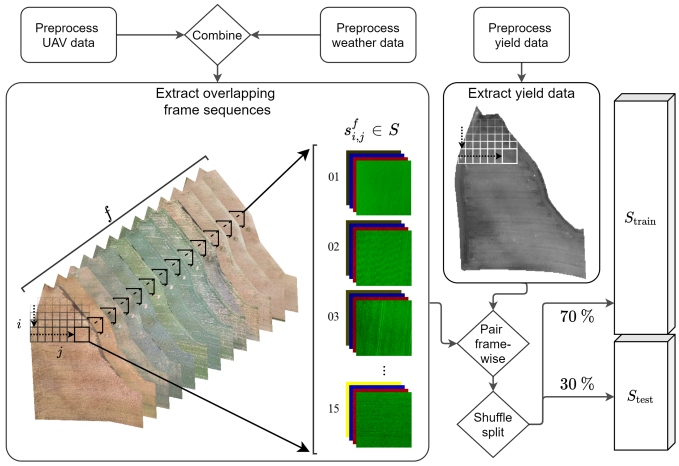

Figure 4.5 Input frame sequence and target average yield extraction process (reproduced from [IV]).

All models we trained with random search [4]. For the CNN-LSTM, the CNN of the model was first trained separately with distinct frames, i.e. point-in-time data. Training the model from scratch was required due to changes in input channel count. It was trained according to the best results of [I], using Adadelta [102] as the optimizer. For the spatiotemporal models, Adam [39] was used as the optimizing algorithm for each model architecture akin to [65], [99] and [52]. The spatiotemporal models were trained with frame sequences. A total of 950 models were trained, with 300 for each spatiotemporal model and 50 for the CNN of the CNN-LSTM.

In the first phase the models were trained to perform end-of-season predictions with full length frame sequences. The trained models were evaluated with a hold-out test set and the results are given in Table 4.3. The number of trainable parameters indicate the model complexity and the best values are in bold. Best performance was achieved with the 3D-CNN architecture.

In-season prediction performance was evaluated with the best performing 3D-CNN model configuration and using data from an actionable time frame from the beginning of the growing season. Multiple input data configuration were tested, forming varying sequences of 3 to 5 frames from five first weeks of imaging (weeks

Table 4.3 The end-of-season prediction performance metrics of the best spatiotemporal models (reproduced from [IV]).

Model	Test RMSE (kg/ha)	Test MAE (kg/ha)	Test MAPE (%)	Test R^2 -	Trainable parameters
Pretrained CNN	692.8	472.7	10.95	0.780	2.72×10^6
CNN-LSTM	456.1	329.5	7.97	0.905	2.94×10^6
ConvLSTM	1190.3	926.9	22.47	0.349	9.03×10^5
3D-CNN	**289.5**	**219.9**	**5.51**	**0.962**	7.48×10^6

21 to 25 of 2018). Overall, the best perfoming in-season sequence cconfiguration in terms of MAE was the four-week-long sequence taken from the beginning of the season (weeks 21 to 24) with 292.8 kg/ha MAE, 7.17 % MAPE and 0.929 R^2. Visualized prediction results are illustrated in Figure 4.6 with a 10 meter step between predicted points.

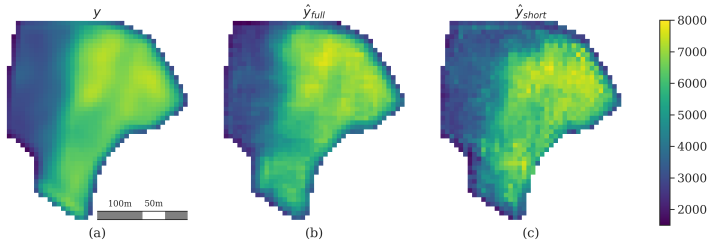

Figure 4.6 Frame-based 3D-CNN model performances against true yield data (reproduced from [IV]).

4.2 Remote sensing data evaluation

4.2.1 Additional input sources

In [V] the effects of additional field-related spatial or spatial-like data on the intra-field crop yield prediction capabilities were studied. The model architecture was taken from [I] and a baseline was trained with RGB data from the earlier half of the growing season of 2018 (weeks 21 to 26). The objective of the study was to assess crop yield prediction capabilities with the best CNN model composition from [I] by

varying the input data configurations with additional data. Additional data sources included data from the following sources: local weather stations, soil samplings, soil sensors and Sentinel-2 satellite system. Disregarding the changing number of input channels, the architectural and optimizer related hyperparameters were not changed to better isolate the effects of different input configurations on the yield estimation performance.

Four crop fields were selected for data acquisition in the vicinity of Pori, Finland (61°29'6.5"N, 21°47'50.7"E) for the growing season of 2018. The field information is provided in Table 4.4. The multisource input data for the fields consists of UAV-based RGB images, multispectral Sentinel-2 [12] satellite data, sparsely collected and analyzed soil samplings, machine-collected soil information, topography information and local weather station data.

Table 4.4 The fields selected for multisource study in the proximity of Pori, Finland (reproduced from [V]).

Field #	Size (ha)	Mean yield (kg/ha)	Crop (*Variety*)	Thermal time (°C)	Sowing date
1	7.59	5157.6	Wheat (*Mistral*)	1316.8	14 May
2	11.77	5534.3	Barley (*Zebra*)	1179.9	12 May
3	7.88	4166.9	Barley (*RGT Planet*)	1127.6	16 May
4	7.24	6166.0	Oats (*Ringsaker*)	1216.4	18 May

General information about the original data sources are given in Table 4.5. UAV data was acquired with weekly overfligths of each field using a SEQUIOA (Parrot Drone SAS, Paris, France) multispectral camera mounted on a Airinov Solo 3DR (Parrot Drone SAS, Paris, France) UAV. The Sentinel-2 satellite data for the fields was acquired from the Copernicus Open Access Hub (European Space Agency, Paris, France), date-matched with UAV-images. Soil samples were collected manually during November 2018 from the fields with 50 m steps by ProAgria, an agronomic counseling instution, and sent to a Eurofins (Eurofins Viljavuuspalvelu, Mikkeli, Finland) laboratory for further analysis. An MSP3 soil scanner (Veris Technologies, Salina, Kansas, USA) was used to map the fields at depths of 0-30 cm and 30-90 cm. The measurements were performed during April and May of 2019. Lidar-based topographical information was acquired from the open-access data portal of the National Land Survey of Finland. Weather data was collected with two separately located Van-

tage Pro2 (Davis Instruments, Hayward, California, USA) weather stations. Yield data was acquired during the harvest of 2018 with yield mapping sensor devices attached to the harvesters, either with a CFX 750 (Trimble Navigation, Sunnyvale, California, USA) or Greenstar 1 (John Deere, Molinde, Illinois, USA).

Table 4.5 General information of data sources and their original formats (reproduced from [V]).

Source	Type	Resolution/Step	Multitemporal	Channels
UAV	Raster	0.3125 m/px	Yes	3
Sentinel-2	Raster	[10,20,60] m/px	Yes	19
Soil samples	Vector	50 m	No	8
Veris MSP3	Vector	20 m	No	6
Topography	Vector	2 m	No	1
Weather	Tabular	-	Yes	2
Yield	Vector	Varying	No	1

All inputs were harmonized to the spatial resolution of the RGB data, 0.3125 m/px by interpolating coarser data sources with GDAL utility's `gdal_grid` program with `invdist:power=3:smoothing=20` interpolation algorithm. After that, overlapping frames were extracted from the data for each week, resulting in a total of 16375 frames. As the number of unique fields was low, maximizing the sample variability the model sees during training was necessary. The data was divided to distinct training, validation and test sets according to the UAV image acquisition week and shuffled to eliminate spatial autocorrelation in subsequent samples due to overlapping frame extraction.

The last step of data processing was to build the data sets for different data source configurations. Four distinct configurations were considered:

- *RGB Only*, which uses UAV RGB data only
- *No S2*, which uses UAV, soil, Veris MSP3, topography and weather data
- *S2 Raw*, which adds Sentinel-2 raw wavelength band data to *No S2*
- *S2 Full*, which adds calculated Sentinel-2 Level-2A product layers to *S2 Raw*.

Ten models were trained for each configuration to account for random initialization of the models inner parameters (weights) and the best models for each configuration were considered. The performance with larger number of fields using UAV RGB data has already been extensively studied in [I] and [IV]. Thus, training a model

with only UAV RGB data provides a studied baseline to which models trained with additional data can be compared. The baseline model using UAV RGB data only attained 1055.7 kg/ha test RMSE, 18.2% test MAPE and 0.343 test R^2. The best performing data configuration was *S2 Full* with 364.1 kg/ha test RMSE, 5.18% test MAPE and 0.922 test R^2 using all 39 layers of input data for each extracted frame. Compared to the baseline *RGB Only* model, the *S2 Full* attained 65.6% lower RMSE, 67.3% lower MAE, 71.5% better MAPE and 0.579 higher R^2 with the test set. Generally every model with multisource inputs performed better than the baseline model. This is shown in Table 4.6.

Table 4.6 The relative performance of the models trained with distinct multisource input data configurations to the baseline *RGB Only* model (reproduced from [V]).

Data Setting	Relative change from *RGB Only*			
	Test RMSE	**Test MAE**	**Test MAPE**	**Test R^2**
No S2	-15.5%	-17.2%	-18.7%	+0.188
S2 Raw	-56.3%	-59.4%	-61.9%	+0.532
S2 Full	**-65.6%**	**-67.3%**	**-71.5%**	**+0.579**

4.2.2 Satellite data reliability

Data from the Sentinel-2 satellites are intensively used for various applications such as land use and vegetation mapping or crop monitoring. Depending on climate conditions in the region of interest, one of the main obstacles in using the data for practical monitoring purposes is cloud coverage. Currently, the cloud mask of the Sentinel data is available in the form of the Level 1C product containing vector layers of dense and cirrus clouds. Also, the percentage of cloudy pixels (dense and cirrus) in the mask are provided. The Level 2A product further processes the Level 1C data to obtain the Scene Classification layer with cloud and cirrus probability values at 60 m spatial resolution. According to Coluzzi et.al. [10], caution has to be taken when using the provided cloud masks and improved cloud detection algorithms are welcome.

Therefore, a random forest classifier was trained to assess cloud cover in Sentinel-2 data in [III], using data acquired from crop fields by UAVs as ground truth for cloudless data. For cloudless multispectral ground truth data, ten crop fields were selected for imaging during 2018 and 2019 in the vicinity of Pori, Finland (61°29'N, 21°48'E).

The fields were imaged approximately weekly with two distinct drones both years, using 3DR Solo (Parrot Drone SAS, Paris, France) for the year 2018 and Disco-Pro AG (Parrot Drone SAS, Paris, France) for 2019. The drones were equipped with similar SEQUIOA (Parrot Drone SAS, Paris, France) multispectral cameras. Half of the fields had wheat (*Zebra/Mistral*), three had barley (*Harbringer/RGT Planet*) and two remaining had oats (*Ringsaker*) as the cultivated crop. The total area of the selected fields was approximately 93 ha. The drone images were downsampled to match the highest resolution available in Sentinel-2 images, 10 m/px. In total, 288 images of distinct crop field images constituted the complete data set.

However, comparing absolute values across bands for two different sensors and imaging platforms proved out to be difficult, as the data would require scaling to an unkown global maximum for Sentinel-2. Thus, using the NDVI values calculated from both data sources (UAV and Sentinel-2) was deemed appropriate due to the index providing normalized and thus comparable data between distinct imaging systems.

To facilitate data based modelling in a supervised setting, target values are required. Due to UAV flight altitude of 150 m eters, Sentinel-2 data can be seen as being cloudless when the NDVI values for a field are similar to the UAV based values as possible. Thus, the task of classification is that of classifying Sentinel-2 data either similar or dissimilar to the UAV data. The similarity for a single pixel-corresponding area is determined by

$$sim_{(s,d)} = \begin{cases} 1, & |s-d| \leq threshold \\ 0, & \text{otherwise} \end{cases} \tag{4.1}$$

where s and d are spatially and temporally aligned NDVI pixels for a field from the satellite and drone sources respectively. Similarity indicates that Sentinel-2 data is cloudless, while dissimilarity indicates cloudiness. The threshold had to be determined via empirical analysis. The task of determining the threshold for labeling is a task of balancing between (1) capturing as much similarities while (2) still excluding as many dissimilarities as possible. Using Students t-test, a total of 15 statistically similar ($p = 0.01$) week-aligned NVDI image pairs were found. Usign similar images, the threshold of similarity was empirically determined by comparing the ratio of pixels deemed similar produced by various thresholds with Equation 4.1. The threshold of 0.075 absolute difference in NDVI was selected. A single image pair

with the calculated similarity map is shown in Fig. 4.7. The first two figures depict the NDVI maps from corresponding sources. The third figure shows the absolute difference between the aligned Sentinel-2 and drone NDVI values. The fourth figure shows the thresholded absolute difference, indicating areas where the NDVI images are similar enough.

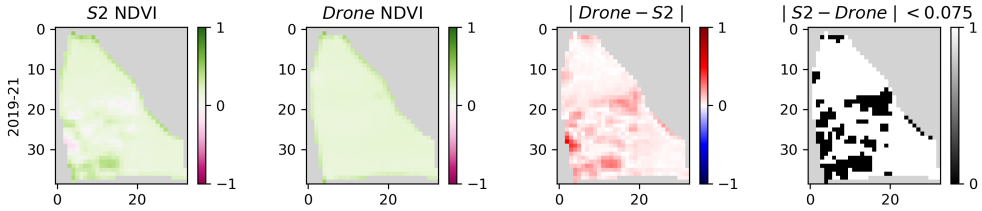

Figure 4.7 A visualization of a single week-aligned Sentinel-2 and drone NDVI image pair with the absolute difference and the similarity map (reproduced from [III]).

Table 4.7 The confusion matrix of similarity label predictions (reproduced from [III]).

Pred/True	0	1
0	TP 23237	FP 2580
1	FN 1807	TN 36037

Table 4.8 Similarity estimates with hold out test data (reproduced from [III]).

	$y = 0$			$y = 1$		
	Mean	Std	Median	Mean	Std	Median
Model	**0.07**	0.25	0.00	0.93	0.26	1.00
$\text{CLDPRB}_{\text{SIM}}$	0.45	0.45	0.26	**0.97**	0.14	1.00
SCL_{SIM}	0.28	0.45	0.00	0.95	0.22	1.00
Samples		38617			25044	

The thresholded binary value maps constitute the target data for pixel-wise binary classification, while Sentinel-2 data was used as inputs. A total of 381972 input-target samples (pixels) were extracted from the source data. The samples were then shuffled and split into training and test data sets with 190986 and 63661 samples, correspondingly. Due to using data in tabular manner, where an input pixel contains several values and spatial dependencies are not modelled, a decision tree based

random forest was deemed an appropriate model to use. The confusion matrix of model predictions against true labels with test data is shown in Table 4.7.

The comparison of sample-wise similarity estimations between the trained model and Sentinel-2 data products are given in Table 4.8. The estimates are given both for when the true target value was 0 (satellite differed from drone) and when it was 1 (satellite similar to drone). For cloudless Sentinel-2 data, the model performed close to existing cloudiness estimates provided with the data products. For cloudy data, the model performed significantly better.

5 CONCLUSIONS AND DISCUSSION

Information relevant for decision making in agriculture can be extracted from heterogeneous remote sensing, environmental and intervention-derived data by means of machine learning. With advancements in computational technologies, the development and training of non-linear multilayer algorithms has become feasible. These methods are commonly referred to as deep learning. Probably the most widely used deep learning structure is that of CNNs, proved to be superior in a variety of image analysis tasks. Other common structure is the RNN network, which is used for modelling sequences of data. A common property of the deep learning structures is that training of the models is performed based on data, i.e., no predefined and precalculated feature vector is needed. This, however, implies that extensive data sets are required for training the models and the operation principles of the models are usually not revealed. Figure 5.1 depicts some application areas of deep learning in agriculture.

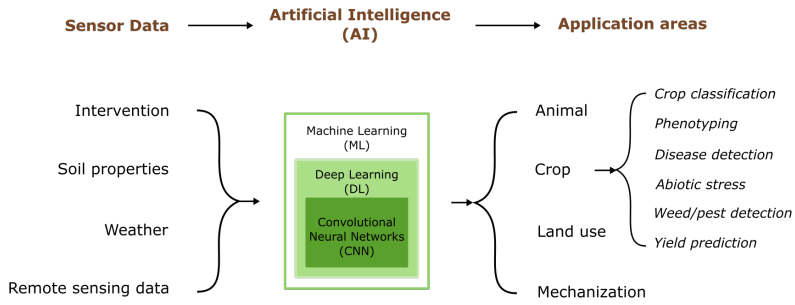

Figure 5.1 Application areas of DL in agriculture.

Remote sensing data can be acquired from satellites such as ESA's Sentinel-2, for example. The problem with the satellite data is that if there is a cloud cover during the overflight of the satellite, no useful data are obtained. The spatial resolution of Sentinel-2 imagery is at best 10 m, which is enough for many applications but too low to allow using texture-based information in the images. Using UAVs for data ac-

quisition offers better spatial resolution, the data acquisition time can be selected by the user and the data can be acquired also in cloudy conditions. Spectral wavelengths can be selected by using appropriate camera; UAV-mountable RGB-NIR cameras are available at affordable price. The drawback is that the UAV has to be operated locally and managing the data and extracting relevant information requires highly specialized skills.

5.1 Deep learning and intra-field yield prediction

The studies described in the publications [I], [II] and [IV] seek to predict yield at the intra-field scale using UAV based images in order to estimate yield variance within the field. This is in contrast to studies utilizing satellite-based, medium to low resolution data and predicting for considerably larger areas at lower spatial resolution. Models at intra-field scale offer the individual farmer the possibility of in-season monitoring of crop, which enables decision support systems for interventions necessary to achieve higher yields.

Publication [I] is an important first step towards establishing a combined model for wheat and barley yield prediction in the Finnish continental subarctic climate. The long summer growing days in this region present a unique profile of temperature and photoperiod, justifying a region-specific deep learning model for these crops. By collecting high-resolution, namely 0.31×0.31 m/px, data using commercial off-the-shelf UAV and camera packages, the attention was focused on a spatial scale that enables us to predict intra-field yield variation within the context of individual farm crop monitoring. Considering that [I] models the yield based only on images, the resulting prediction error of 484 kg/ha test MAE, 8.8 % test MAPE and 0.857 R^2-score is promising. The results of [I] indicate that the CNN models are capable of reasonable accurate yield estimates based on RGB images. This suggests that multiple spectral bands increase the information content in comparison to the condensed NDVI raster. From the results of [V] it can be suggested that complementing the RGB data with NIR channel might further enhance the prediction capabilities of the CNN model. Additionally, NIR-based vegetation indices could have improved modelling performance even more as discussed in [104]. Intra-field crop yield prediction based on multispectral UAV data based is, thus, a subject for a future study.

As further examined in [II], the case study with the CNN from [I] revealed some

limitations of the model in yield prediction. The model underestimated/overestimated the yield in the regions of high/low yield values, respectively. Another limitation is related to yield data pre-processing. In some cases the polygons of yield data overlap causing errors in yield density maps.

In [IV] the feasibility of using spatiotemporal deep learning architectures in modelling crop yield at the intra-field scale was evaluated using high-resolution UAV data. With full sequence modelling, a 3D-CNN based architecture performed the best with 218.9 kg/ha test MAE, 5.51% test MAPE and 0.962 test R^2-score. Compared to [I] using just a point-in-time single frame predictor with 484.3 kg/ha MAE and 8.8% MAPE, the modelling performance was improved by 265.4 kg/ha MAE (54.8% improvement) and 3.29% MAPE (37.4% improvement) with time series inputs. With a shorter sequence the 3D-CNN model attained 292.8 kg/ha test MAE, 7.17% test MAPE and 0.929 test R^2-score. As weather information was utilized in [IV] at city scale, the accuracy of the growth phase could be further improved with specifically located weather stations. Weather stations located in the approximate vicinity of the fields under scrutiny could provide better and more accurate measurements of the local temperatures and other climatological variables and thus might help the model produce even better predictions when sequences are involved. This was corrected in [V] by using two distinct weather stations located near the studied fields.

These results with point-in-time and multitemporal models are competitive in light of recent yield prediction studies. Sun et al. [77] utilized UAVs to gather hyperspectral data of potato tuber growth at the resolution of 2.5 cm/px. They utilized traditional ML methods, such as linear models and decision trees, to perform tuber yield estimation using individual data points gathered in-season at the intra-field scale, achieving 0.63 R^2-score for the tuber yield prediction accuracy with a Ridge regression. Lee at al. [49] used an UAV to collect multispectral data from wheat and corn fields to estimate intra-field crop nitrogen content using linear regression and point samples - spatial features were not utilized. They fit multiple linear models to wheat and corn and attained 0.872 R^2-score on average. Fu et al. [14] performed wheat leaf area index and grain yield estimation with various vegetation indices derived from point-in-time multispectral UAV data using multiple machine learning methods, neural networks included. The highest performance they attained was 0.78 R^2-score with a random forest model. However, they fed the input data as point samples. Performing county-scale soybean yield prediction, [78]

used a CNN, an LSTM and a composite CNN-LSTM to model soybean yield with in-season satellite data. They achieved an average 0.78 R^2-score with the spatiotemporal CNN-LSTM model. [98] utilized RGB and multispectral data acquired with a UAV from rice fields in China to predict rice yields with a composite CNN model on field block scale. Feeding the multisource data to distinct, parallelized CNNs, they report a rice yield prediction performance of 0.50 R^2 and 26.6% MAPE.

In their study, Sun et al. [77] used input data with resolutions from 500 × 500 m/px to 1 × 1 km/px. Rustowicz et al. [65] performed crop type classification in Europe and Africa with multi-temporal satellit data at resolutions from 3 × 3 m/px to 10 × 10 m/px. They attained F1 scores 91.4 for the CNN-ConvLSTM and 90.0 for the 3D-CNN, averaged over crop types in their Germany data set. Yaramasu et al. [99] performed pre-season crop type mapping for the area of Nebraska, US, employing a CNN-ConvLSTM to extract spatiotemporal features from multi-temporal multi-satellite composite data set. Using prior years of crop type related data to predict a map of crop types, they attained an average accuracy of 77% across all crop types in their data. The data was processed to a resolution of 30 × 30 m/px. Ji et al. [27] utilized a 3D-CNN to classify crop types from multi-temporal satellite data gathered from an area within China, acquiring a classification accuracy of 98.9% with the model. Their input data resolutions were from 4 × 4 m/px to 15 × 15 m/px. Borra-Serrano et al. [5] performed weekly UAV image collections in a controlled field experiment with soybeans, performing seed yield prediction with multiple linear models fit to the multi-temporal data. Thus, spatiotemporal modelling with novel techniques was not performed. With seed yield prediction, they achieved 0.501 adjusted R^2 score. The resolution of their data was 1.25 × 1.25 cm/px.

In [I] and [II] it is shown that with high-resolution UAV data, crop yield prediction with CNNs is feasible and produces results accurate enough for performing corrective farming actions in-season. In [IV] it is shown that adding time as an additional feature not only improves the modelling performance with high-resolution UAV RGB data but also improves the predictive capabilities. Additionally, using weekly UAV data gathered during the first month provides enough data for the model to build an accurately predicted yield map from which to draw further conclusions. The use of both high-resolution point-in-time and multitemporal remote sensing data is beneficial in crop yield modelling and prediction with deep learning. Furthermore, the easy accessibility of commercially available UAVs with mounted

RGB sensors enables image data acquisition in higher resolutions compared to satellites. This in turn opens up the possibilities to perform modelling and predictions at intra-field scale. As shown in the publications [I], [II] and [IV], the use of UAV-based data and proper spatiotemporal deep learning techniques is an enabler of more sophisticated decision support systems in the domain of agriculture.

5.2 Multisource input data assessment

In [V], the effects of using input data from multiple sources on the task of spatial crop yield prediction were evaluated. The performance with larger number of fields using UAV RGB data had already been extensively studied in [I] and [IV]. Thus, training a model with only UAV RGB data provides a studied baseline to which models trained with additional data can be compared against. The best performing data configuration was *S2 Full* with 364.1 kg/ha test RMSE, 5.18% test MAPE and 0.922 test R^2 using all 39 layers of input data for each extracted frame (see Section 3 of publication [V]). Compared to the baseline *RGB Only* model, the *S2 Full* attained 65.6% lower RMSE, 67.3% lower MAE, 71.5% better MAPE and 0.579 higher R^2 with the test set. Generally every model with multisource inputs performed better than the baseline model.The study indicates that increasing the number of input data sources increases the performance of intra-field crop yield prediction. To draw definite conclusions on the most optimal configuration of input data sources more data is required. With more representative data, generalizable conclusions are more warranted. The relative improvement compared to baseline of using UAV RGB only as the input data were notable. Consolidating UAV RGB data with soil information and ground topology data already somewhat improves the prediction performance, while largest performance gains were gained from using Sentinel-S2 in addition to UAV RGB, soil sampling, Veris MSP3 soil scanner, weather and topography data. As the data in [V] focuses on a single growing season, the generalization of a multisource crop yield prediction model with multiple years of data is a subject for a future study.

The study of publication [III] indicates that the random forest model outperforms the Sentinel-2 CLDPRB and SCL data layers in detecting cloudy areas ($y = 0$). For non-cloudy areas the detection accuracy was slightly higher for the Sentinel products. The developed method was found to improve the usability of Sentinel data in crop monitoring. By visual inspection it was observed that in many cases when

the Sentinel-2 products indicated the whole crop field to be cloud-covered, there were still significant areas of almost clear skies. The proposed algorithm proved capable in detecting these areas with considerable accuracy. The classification results are further usable in different applications. Firstly, commercial applications routinely utilize satellite data based NDVI maps, which greatly benefit from accurate estimations of pixel-wise cloud canopy. Another application is in preprocessing satellite data used as inputs for crop yield estimation.

5.3 Limitations

A limitation to our crop yield estimation studies is the use of aggregated crop type data collected from various fields. Using a single model to predict for wheat, barley and oats prohibits both the inference and the performance analysis of the model on a per-crop basis. Additionally, the remote sensing data based modelling approach doesn't take into account any existing crop growth models. Those could well be utilized to further provide better performance, akin to what has been done in [5].

Models trained at large regional scales rarely extrapolate to finer scales, though efforts have been taken to develop scalable models [11]. A good strategy of dividing the data into training, validation and testing sets on field basis would be required to prove that models are capable to generalize. This raises an important discussion point regarding how the frames were extracted in publications [I], [II], [IV] and [V], especially considering the overlap of data across adjacent frames. The frames were randomly allocated to training and test sets. Another important point related to input sample independence is the invariability of data from distinct sources. This is issue is specific to [V], where input samples contain both temporally and spatially invariable data. Temporally invariable data includes soil samplings, Veris MSP3 soil scanner data and topographical maps. Weather data, on the other hand, is spatially constant. While the cumulative temperatures and rain sums do change in time, the time-specific weather layers are effectively rasters with constant values. Whether the input data is split spatially or temporally to training and test sets, there is a case to be made that some data might be present in similar form in all split data sets simultaneously. With traditional machine learning, such as linear models, data which is not independent and identically distributed would require extensive prior work to find feature-wise couplings. As pointed out in [78], neural networks learn these

couplings implicitly from the data. Thus, the input layers are not handled in solitude one by one but are always utilized in the context of other data present in an input sample. The context includes spatial, temporal and inter-channel dimensions. Therefore, the test data as a combination of inputs can be considered distinct from training and validation sets. This is further reinforced by the results of [V]. The performance gains with UAV RGB data combined with temporally invariant soil information and ground data is trumped by the performance gains of data configurations using Sentinel-S2 data as additional inputs. This would suggest that the combination of the inputs matters more than presence of distinct, invariant data in training, validation and test sets.

Regarding multisource data in the context of smart farming and crop yield estimation, data itself is an evolving research topic. The use of multisource inputs in remote sensing, while focusing on multispectral data acquired from satellite systems orbiting the globe, has been extensively reviewed in [18]. The use of multispectral data from UAVs and the prediction architectures thereof is also a developing topic [55]. Another topic related to spatial data is that of autocorrelation [1]. To address autocorrelation of spatial frames in a future study, the inclusion of pixel-wise location information, as suggested in [1], should be sufficient to inform the deep learning model whether data similarity is due to proximity or some other factor or combination of them.

Several issues should be considered with cloud cover classification in [III]. Firstly, when training the random forest classifier, the thresholded absolute difference between the Sentinel-2 and drone data was used as the ground truth. While it can be argued that the main cause of this difference is cloudiness, there may also be other factors involved such as shadows or differences in irradiance. The satellite and drone imagery were not necessarily acquired during the same time of the day or same day of the week, although best time-matching pairs were looked for when selecting the data. In some cases a couple of days may cause significant changes in the crop development. Another limitation comes from using the NDVI data layers for ground truth assessment. While the NDVI index contains significant information for vegetation monitoring and is probably a good choice when assessing cloud cover in crop fields, its use reduces the generalizability of the results to other land cover types.

5.4 Conclusions

In summary, it is shown in [I] and [II] that with high-resolution UAV data, crop yield prediction with CNNs is feasible and produces results accurate enough for performing corrective farming actions in-season. In [IV] it is shown that adding time as an additional feature (time series data) not only improves the modelling performance (post-season prediction) with high-resolution UAV RGB data but also improves the predictive capabilities (in-season prediction). Additionally, using weekly UAV data gathered during the first month provides enough data for the model to build an accurately predicted yield map from which to draw further conclusions. The use of both high-resolution point-in-time and multitemporal remote sensing data is beneficial in crop yield modelling and prediction with deep learning. Furthermore, the easy accessibility of commercially available UAVs with mounted RGB sensors enables image data acquisition in higher resolutions compared to satellites. This in turn opens up the possibilities to perform modelling and predictions at intra-field scale. As shown in the publications [I], [II] and [IV], the use of UAV-based data and proper spatiotemporal deep learning techniques is an enabler of more sophisticated decision support systems in the domain of agriculture. Furthermore, the study presented in [V] shows that using various data sources for crop yield prediction in addition to UAV RGB data improves the predictive capabilities of the model. Referring to the use of specialized equipment in data acquisition, limiting the data sources to those common to majority of fields would, however, ensure better generalization capabilities for the models. Finally, as shown in [III], data based modelling can also be employed to perform quality assurance for satellite based data used in yield estimation and other relevant application contexts.

REFERENCES

[1] A. Amgalan, L. R. Mujica-Parodi and S. S. Skiena. Fast Spatial Autocorrelation. *Biometrics* 30.4 (2020), 729. ISSN: 0006341X. DOI: 10.2307/2529248. arXiv: 2010.08676.

[2] A. Barbosa, R. Trevisan, N. Hovakimyan and N. F. Martin. Modeling yield response to crop management using convolutional neural networks. *Computers and Electronics in Agriculture* 170 (2020). ISSN: 01681699. DOI: 10.1016/j.compag.2019.105197.

[3] J. C. Bell, C. A. Butler and J. A. Thompson. Soil-Terrain Modeling for Site-Specific Agricultural Management. eng. *Site-Specific Management for Agricultural Systems*. Madison, WI, USA: American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, 1995, 209–227.

[4] J. Bergstra and Y. Bengio. Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research* 13 (2012), 281–305. ISSN: 1532-4435. DOI: 10.1162/153244303322533223. arXiv: 1504.05070.

[5] I. Borra-Serrano, T. D. Swaef, P. Quataert, J. Aper, A. Saleem, W. Saeys, B. Somers, I. Roldán-Ruiz and P. Lootens. Closing the phenotyping gap: High resolution UAV time series for soybean growth analysis provides objective data from field trials. *Remote Sensing* 12.10 (2020), 1–19. ISSN: 20724292. DOI: 10.3390/rs12101644.

[6] L. Bottou. On-line Learning in Neural Networks. Ed. by D. Saad. New York, NY, USA: Cambridge University Press, 1998. Chap. On-line Le, 9–42.

[7] T. Chen and C. Guestrin. XGBoost: A Scalable Tree Boosting System. *arXiv* (2016), 1–6. ISSN: 0146-4833. DOI: 10.1145/2939672.2939785. arXiv: 1603.02754.

[8] W. Chivasa, O. Mutanga and C. Biradar. Application of remote sensing in estimating maize grain yield in heterogeneous african agricultural landscapes: A review. *International Journal of Remote Sensing* 38.23 (2017), 6816–6845. ISSN: 13665901. DOI: 10.1080/01431161.2017.1365390.

[9] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2014), 1724–1734. ISSN: 09205691. DOI: 10.3115/v1/D14-1179. arXiv: 1406.1078.

[10] R. Coluzzi, V. Imbrenda, L. Maria and S. Tiziana. A first assessment of the Sentinel-2 Level 1-C cloud mask product to support informed surface analyses. *Remote Sensing of Environment* 217 (2018), 426–443. DOI: 10.1016/j.rse.2018.08.009.

[11] R. J. Donohue, R. A. Lawes, G. Mata, D. Gobbett and J. Ouzman. Towards a national, remote-sensing-based model for predicting field-scale crop yield. *Field Crops Research* 227 (2018), 79–90. ISSN: 0378-4290.

[12] *ESA: Sentinel-2*. https://sentinel.esa.int/web/sentinel/missions/sentinel-2. (Visited on 04/12/2018).

[13] P. Filippi, E. J. Jones, N. S. Wimalathunge, P. D. Somarathna, L. E. Pozza, S. U. Ugbaje, T. G. Jephcott, S. E. Paterson, B. M. Whelan and T. F. Bishop. An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning. *Precision Agriculture* 20.5 (2019), 1015–1029. ISSN: 15731618. DOI: 10.1007/s11119-018-09628-4.

[14] Z. Fu, J. Jiang, Y. Gao, B. Krienke, M. Wang, K. Zhong, Q. Cao, Y. Tian, Y. Zhu, W. Cao and X. Liu. Wheat growth monitoring and yield estimation based on multi-rotor unmanned aerial vehicle. *Remote Sensing* 12.3 (2020). ISSN: 20724292. DOI: 10.3390/rs12030508.

[15] *Gaofen-1*. https://directory.eoportal.org/web/eoportal/satellite-missions/g/gaofen-1. (Visited on 03/02/2021).

[16] *Gaofen-2*. https://directory.eoportal.org/web/eoportal/satellite-missions/g/gaofen-2. (Visited on 03/02/2021).

[17] F. Gers and J. Schmidhuber. Recurrent nets that time and count. *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*. IEEE, 2000, 189–194 vol.3. DOI: 10.1109/IJCNN.2000.861302. arXiv: arXiv:1011.1669v3.

[18] P. Ghamisi, B. Rasti, N. Yokoya, Q. Wang, B. Hofle, L. Bruzzone, F. Bovolo, M. Chi, K. Anders, R. Gloaguen, P. M. Atkinson and J. A. Benediktsson. Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art. *IEEE Geoscience and Remote Sensing Magazine* 7.1 (2019), 6–39. ISSN: 21686831. DOI: 10.1109/MGRS.2018.2890023.

[19] I. Goodfellow, Y. Bengio and A. Courville. *Deep Learning*. MIT Press, 2016.

[20] *Google Earth Engine*. https://developers.google.com/earth-engine/. (Visited on 03/05/2021).

[21] A. Graves. Generating Sequences With Recurrent Neural Networks. (2013), 1–43. ISSN: 18792782. DOI: 10.1145/2661829.2661935. arXiv: 1308.0850.

[22] K. He, X. Zhang, S. Ren and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *Proceedings of the IEEE International Conference on Computer Vision* 2015 Inter (2015), 1026–1034. ISSN: 15505499. arXiv: 1502.01852.

[23] G. Hinton, N. Srivastava and K. Swersky. *Neural Networks for Machine Learning Lecture 6a: Overview of minibatch gradient descent*. 2014.

[24] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation* 9.8 (1997), 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. arXiv: 1206.2944.

[25] D. Ienco, R. Interdonato, R. Gaetano and D. Ho Tong Minh. Combining Sentinel-1 and Sentinel-2 Satellite Image Time Series for land cover mapping via a multi-source deep learning architecture. *ISPRS Journal of Photogrammetry and Remote Sensing* 158 (2019), 11–22. ISSN: 09242716. DOI: 10.1016/j.isprsjprs.2019.09.016.

[26] S. Ioffe and C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. (2015). ISSN: 0717-6163. DOI: 10.1007/s13398-014-0173-7.2. arXiv: 1502.03167.

[27] S. Ji, C. Zhang, A. Xu, Y. Shi and Y. Duan. 3D Convolutional Neural Networks for Crop Classification with Multi-Temporal Remote Sensing Images. *Remote Sensing* 10.2 (2018), 75. ISSN: 2072-4292. DOI: 10.3390/rs10010075.

[28] H. Jiang, H. Hu, R. Zhong, J. Xu, J. Xu, J. Huang, S. Wang, Y. Ying and T. Lin. A deep learning approach to conflating heterogeneous geospatial data for corn yield estimation: A case study of the US Corn Belt at the county level. *Global Change Biology* 26.3 (2020), 1754–1766. ISSN: 13652486. DOI: 10.1111/gcb.14885.

[29] R. Jozefowicz, W. Zaremba and I. Sutskever. An empirical exploration of Recurrent Network architectures. *32nd International Conference on Machine Learning, ICML 2015*. Vol. 3. 2015, 2332–2340.

[30] A. Kamilaris and F. X. Prenafeta-Boldú. A review of the use of convolutional neural networks in agriculture. *Journal of Agricultural Science* June (2018), 1–11. ISSN: 14695146. DOI: 10.1017/S0021859618000436.

[31] A. Kamilaris and F. X. Prenafeta-Boldú. Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture* 147 (2018), 70–90. ISSN: 0168-1699. DOI: doi.org/10.1016/j.compag.2018.02.016.

[32] A. Kamilaris, A. Kartakoullis and F. X. Prenafeta-Boldú. A review on the practice of big data analysis in agriculture. *Computers and Electronics in Agriculture* 143 (2017), 23–37. ISSN: 0168-1699. DOI: 10.1016/J.COMPAG.2017.09.037.

[33] Y. Kang, M. Ozdogan, X. Zhu, Z. Ye, C. Hain and M. Anderson. Comparative assessment of environmental variables and machine learning algorithms for maize yield prediction in the US Midwest. *Environmental Research Letters* 15.6 (2020). ISSN: 17489326. DOI: 10.1088/1748-9326/ab7df9.

[34] A. Karpathy and L. Fei-Fei. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.4 (2017), 664–676. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2016.2598339.

[35] L. Karthikeyan, I. Chawla and A. K. Mishra. A review of remote sensing applications in agriculture for food security: Crop growth and yield, irrigation, and crop losses. *Journal of Hydrology* 586.March (2020), 124905. ISSN: 00221694. DOI: 10.1016/j.jhydrol.2020.124905.

[36] S. Khaki, L. Wang and S. V. Archontoulis. A CNN-RNN Framework for Crop Yield Prediction. *Frontiers in Plant Science* 10.January (2020), 1–14. ISSN: 1664462X. DOI: 10.3389/fpls.2019.01750. arXiv: 1911.09045.

[37] S. Khanal, J. Fulton, A. Klopfenstein, N. Douridas and S. Shearer. Integration of high resolution remotely sensed data and machine learning techniques for spatial prediction of soil properties and corn yield. *Computers and Electronics in Agriculture* 153.August (2018), 213–225. ISSN: 01681699. DOI: 10.1016/j.compag.2018.07.016.

[38] S. Khanal, K. KC, J. P. Fulton, S. Shearer and E. Ozkan. Remote Sensing in Agriculture—Accomplishments, Limitations, and Opportunities. *Remote Sensing* 12.22 (2020), 3783. ISSN: 2072-4292. DOI: 10.3390/rs12223783.

[39] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *Iclr* (2014), 1–15. ISSN: 09252312. DOI: http://doi.acm.org.ezproxy.lib.ucf.edu/10.1145/1830483.1830503. arXiv: 1412.6980.

[40] L. Klerkx, E. Jakku and P. Labarthe. A review of social science on digital agriculture, smart farming and agriculture 4.0: New contributions and a future research agenda. *NJAS - Wageningen Journal of Life Sciences* 90-91.October (2019), 100315. ISSN: 22121307. DOI: 10.1016/j.njas.2019.100315.

[41] T. van Klompenburg, A. Kassahun and C. Catal. Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture* 177 (2020), 105709. ISSN: 01681699. DOI: 10.1016/j.compag.2020.105709.

[42] A. Krizhevsky, I. Sutskever and G. E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM* 60.6 (2017), 84–90. ISSN: 00010782. DOI: 10.1145/3065386. arXiv: 1102.0183.

[43] A. Krizhevsky, I. Sutskever and G. E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM* 60.6 (2017), 84–90. ISSN: 00010782. DOI: 10.1145/3065386. arXiv: 1102.0183.

[44] A. Laine, M. Högnäsbacka, M. Niskanen, K. Ohralahti, L. Jauhiainen, J. Kaseva and H. Nikander. Virallisten lajikekokeiden tulokset 2009-2016. (2017), 262.

[45] *Landsat 7*. https://www.usgs.gov/core-science-systems/nli/landsat/landsat-7. (Visited on 03/02/2021).

[46] *Landsat 8*. https://www.usgs.gov/core-science-systems/nli/landsat/landsat-8. (Visited on 03/02/2021).

[47] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard and L. D. Jackel. *Backpropagation Applied to Handwritten Zip Code Recognition*. eng. 1989.

[48] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner. Gradient Based Learning Applied to Document Recognition. *Proceedings of the IEEE* 86.11 (1998), 2278–2324. ISSN: 00189219. DOI: 10.1109/5.726791. arXiv: 1102.0183.

[49] H. Lee, J. Wang and B. Leblon. Intra-Field Canopy Nitrogen Retrieval from Unmanned Aerial Vehicle Imagery for Wheat and Corn Fields. *Canadian Journal of Remote Sensing* 46.4 (2020), 1–19. ISSN: 17127971. DOI: 10.1080/07038992.2020.1788384.

[50] Y. Li, H. Zhang and Q. Shen. Spectral–Spatial Classification of Hyperspectral Imagery with 3D Convolutional Neural Network. *Remote Sensing* 9.1 (2017), 67. ISSN: 2072-4292. DOI: 10.3390/rs9010067.

[51] T. Lin, R. Zhong, Y. Wang, J. Xu, H. Jiang, J. Xu, Y. Ying, L. Rodriguez, K. C. Ting and H. Li. DeepCropNet: a deep spatial-temporal learning framework for county-level corn yield estimation. *Environmental Research Letters* 15.3 (2020). ISSN: 17489326. DOI: 10.1088/1748-9326/ab66cb.

[52] Q. Liu, F. Zhou, R. Hang and X. Yuan. Bidirectional-Convolutional LSTM Based Spectral-Spatial Feature Learning for Hyperspectral Image Classification. *Remote Sensing* 9.12 (2017), 1330. ISSN: 2072-4292.

[53] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu and A. C. Berg. SSD: Single Shot MultiBox Detector. *Eccv*. Ed. by B. Leibe, J. Matas, N. Sebe and M. Welling. Vol. 9905. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016, 21–37. DOI: 10.1007/978-3-319-46448-0_2.

[54] M. Maimaitijiang, V. Sagan, P. Sidike, A. M. Daloye, H. Erkbol and F. B. Fritschi. Crop monitoring using satellite/UAV data fusion and machine learning. *Remote Sensing* 12.9 (2020). ISSN: 20724292. DOI: 10.3390/RS12091357.

[55] G. Messina and G. Modica. Applications of UAV thermal imagery in precision agriculture: State of the art and future research outlook. *Remote Sensing* 12.9 (2020). ISSN: 20724292. DOI: 10.3390/RS12091491.

[56] R. Näsi, N. Viljanen, J. Kaivosoja, T. Hakala, M. Pandžić, L. Markelin and E. Honkavaara. Assessment of Various Remote Sensing Technologies in Biomass and Nitrogen Content Estimation Using an Agricultural Test Field. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-3/W3 (2017), 137–141. ISSN: 2194-9034. DOI: 10.5194/isprs-archives-XLII-3-W3-137-2017.

[57] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga and A. Lerer. Automatic differentiation in PyTorch. *NIPS-W*. 2017.

[58] *PlanetScope*. https://earth.esa.int/eogateway/missions/planetscope. (Visited on 03/02/2021).

[59] S. Ren, K. He, R. Girshick and J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.6 (2017), 1137–1149. ISSN: 01628828. DOI: 10.1109/TPAMI.2016.2577031. arXiv: 1506.01497.

[60] H. Ritchie and M. Roser. *Our World in Data: Crop Yields*. https://ourworldindata.org/crop-yields. 2013. (Visited on 06/08/2021).

[61] D. C. Rose and J. Chilvers. Agriculture 4.0: Broadening Responsible Innovation in an Era of Smart Farming. *Frontiers in Sustainable Food Systems* 2.December (2018), 1–7. ISSN: 2571581X. DOI: 10.3389/fsufs.2018.00087.

[62] D. E. Rumelhart, G. E. Hinton and R. J. Williams. Learning representations by back-propagating errors. eng. *Nature (London)* 323.6088 (1986), 533–536. ISSN: 0028-0836.

[63] M. Rußwurm and M. K rner. Multi-temporal land cover classification with sequential recurrent encoders. *ISPRS International Journal of Geo-Information* 7.4 (2018), 129. ISSN: 22209964. DOI: 10.3390/ijgi7040129. arXiv: 1802.02080.

[64] M. Rußwurm and M. Körner. Convolutional LSTMs for Cloud-Robust Segmentation of Remote Sensing Imagery. (2018). arXiv: 1811.02471.

[65] R. Rustowicz, R. Cheong, L. Wang, S. Ermon, M. Burke and D. Lobell. Semantic Segmentation of Crop Type in Africa: A Novel Dataset and Analysis of Deep Learning Methods. *CVPR Workshops*. 2019, 75–82.

[66] T. N. Sainath, O. Vinyals, A. Senior and H. Sak. Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. Vol. 2015-Augus. Institute of Electrical and Electronics Engineers Inc., 2015, 4580–4584.

[67] A. Salmivaara, S. Launiainen, J. Perttunen, P. Nevalainen, J. Pohjankukka, J. Ala-Ilomäki, M. Sirén, A. Laurén, S. Tuominen, J. Uusitalo, T. Pahikkala, J. Heikkonen and L. Finér. Towards dynamic forest trafficability prediction using open spatial data, hydrological modelling and sensor technology. *Forestry: An International Journal of Forest Research* 93.5 (2020), 662–674. ISSN: 0015-752X. DOI: 10.1093/forestry/cpaa010.

[68] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L. C. Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2018), 4510–4520. ISSN: 10636919. DOI: 10.1109/CVPR.2018.00474. arXiv: 1801.04381.

[69] *Satellite Missions Directory*. https://directory.eoportal.org/web/eoportal/satellite-missions. (Visited on 03/02/2021).

[70] J. Schmidhuber. Deep Learning in Neural Networks: An Overview. *Neural Networks* 61 (2014), 85–117. ISSN: 08936080. DOI: 10.1016/j.neunet.2014.09.003. arXiv: 1404.7828.

[71] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45.11 (1997), 2673–2681. ISSN: 1053587X. DOI: 10.1109/78.650093.

[72] R. A. Schwalbert, T. Amado, G. Corassa, L. P. Pott, P. V. Prasad and I. A. Ciampitti. Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. *Agricultural and Forest Meteorology* 284.December 2019 (2020), 107886. ISSN: 01681923. DOI: 10.1016/j.agrformet.2019.107886.

[73] *Sentinel-2 MSI*. https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-2-msi. (Visited on 03/02/2021).

[74] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-k. Wong and W.-c. Woo. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. (2015). ISSN: 10495258. arXiv: 1506.04214.

[75] S. Shidnal, M. V. Latte and A. Kapoor. Crop yield prediction: two-tiered machine learning model approach. *International Journal of Information Technology (Singapore)* (2019), 1–9. ISSN: 25112112. DOI: 10.1007/s41870-019-00375-x.

[76] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. 2015. arXiv: 1409.1556v6.

[77] C. Sun, L. Feng, Z. Zhang, Y. Ma, T. Crosby, M. Naber and Y. Wang. Prediction of end-of-season tuber yield and tuber set in potatoes using in-season uav-based hyperspectral imagery and machine learning. *Sensors (Switzerland)* 20.18 (2020), 1–13. ISSN: 14248220. DOI: 10.3390/s20185293.

[78] J. Sun, L. Di, Z. Sun, Y. Shen and Z. Lai. County-Level Soybean Yield Prediction Using Deep CNN-LSTM Model. *Sensors* 19.20 (2019), 4363. ISSN: 1424-8220. DOI: 10.3390/s19204363.

[79] H. Sundmaeker, C. N. Verdouw, J. Wolfert and L. Perez Freire. Internet of Food and Farm 2020. *Digitising the Industry*. 2016.

[80] J. Syväjärvi. *Reikäkorteista digiaikaan: Maatalouden Laskentakeskus Oy 30 vuotta, tietojenkäsittelyä 58 vuotta*. Suomen Maatalouden Laskentakeskus Oy, 2016, 105587–105609.

[81] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich. Going deeper with convolutions. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 07-12-June (2015), 1–9. ISSN: 10636919. DOI: 10.1109/CVPR.2015.7298594. arXiv: 1409.4842.

[82] N. Tantalaki, S. Souravlas and M. Roumeliotis. Data-Driven Decision Making in Precision Agriculture: The Rise of Big Data in Agricultural Systems. *Journal*

of Agricultural and Food Information 20.4 (2019), 344–380. ISSN: 15404722. DOI: 10.1080/10496505.2019.1638264.

[83] D. Tedesco-Oliveira, R. Pereira da Silva, W. Maldonado and C. Zerbato. Convolutional neural networks in predicting cotton yield from images of commercial fields. *Computers and Electronics in Agriculture* 171 (2020), 105307. ISSN: 01681699. DOI: 10.1016/j.compag.2020.105307.

[84] A. S. Terliksiz and D. T. Altylar. Use of deep neural networks for crop yield prediction: A case study of soybean yield in lauderdale county, Alabama, USA. *2019 8th International Conference on Agro-Geoinformatics, Agro-Geoinformatics 2019* (2019), 2019–2022.

[85] J. Tiusanen. Aineiston käsittely ja muotoilu. *Käytännön Maamies* (2017).

[86] D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri. Learning spatiotemporal features with 3D convolutional networks. *Proceedings of the IEEE International Conference on Computer Vision* 2015 Inter (2015), 4489–4497. ISSN: 15505499. arXiv: 1412.0767.

[87] A. Triantafyllou, P. Sarigiannidis and S. Bibi. Precision agriculture: A remote sensing monitoring system architecture. *Information (Switzerland)* 10.11 (2019). ISSN: 20782489. DOI: 10.3390/info10110348.

[88] D. C. Tsouros, S. Bibi and P. G. Sarigiannidis. A review on UAV-based applications for precision agriculture. *Information (Switzerland)* 10.11 (2019). ISSN: 20782489. DOI: 10.3390/info10110349.

[89] Z. Unal. Smart Farming Becomes even Smarter with Deep Learning - A Bibliographical Analysis. *IEEE Access* 8 (2020), 105587–105609. ISSN: 21693536. DOI: 10.1109/ACCESS.2020.3000175.

[90] P. Vincent, H. Larochelle, Y. Bengio and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. *Proceedings of the 25th international conference on Machine learning - ICML '08* (2008), 1096–1103. ISSN: 1605582050. DOI: 10.1145/1390156.1390294.

[91] X. Wang, J. Huang, Q. Feng and D. Yin. Winter wheat yield prediction at county level and uncertainty analysis in main wheat-producing regions of China with deep learning approaches. *Remote Sensing* 12.11 (2020). ISSN: 20724292. DOI: 10.3390/rs12111744.

[92] A. Wolanin, G. Mateo-García, G. Camps-Valls, L. Gómez-Chova, M. Meroni, G. Duveiller, Y. Liangzhi and L. Guanter. Estimating and understanding crop yields with explainable deep learning in the Indian Wheat Belt. *Environmental Research Letters* 15.2 (2020), 024019. ISSN: 1748-9326. DOI: 10.1088/1748-9326/ab68ac.

[93] S. Wolfert, L. Ge, C. Verdouw and M. J. Bogaardt. Big Data in Smart Farming – A review. *Agricultural Systems* 153 (2017), 69–80. ISSN: 0308521X. DOI: 10.1016/j.agsy.2017.01.023.

[94] *WorldView-2*. https://directory.eoportal.org/web/eoportal/satellite-missions/v-w-x-y-z/worldview-2. (Visited on 03/02/2021).

[95] *WorldView-3*. https://directory.eoportal.org/web/eoportal/satellite-missions/v-w-x-y-z/worldview-3. (Visited on 03/02/2021).

[96] C. Xie and C. Yang. A review on plant high-throughput phenotyping traits using UAV-based sensors. *Computers and Electronics in Agriculture* 178.October 2019 (2020), 105731. ISSN: 01681699. DOI: 10.1016/j.compag.2020.105731.

[97] X. Xu, P. Gao, X. Zhu, W. Guo, J. Ding, C. Li, M. Zhu and X. Wu. Design of an integrated climatic assessment indicator (ICAI) for wheat production: A case study in Jiangsu Province, China. *Ecological Indicators* 101.July 2018 (2019), 943–953. ISSN: 1470160X. DOI: 10.1016/j.ecolind.2019.01.059.

[98] Q. Yang, L. Shi, J. Han, Y. Zha and P. Zhu. Deep convolutional neural networks for rice grain yield estimation at the ripening stage using UAV-based remotely sensed images. *Field Crops Research* 235.August 2018 (2019), 142–153. ISSN: 03784290. DOI: 10.1016/j.fcr.2019.02.022.

[99] R. Yaramasu, V. Bandaru and K. Pnvr. Pre-season crop type mapping using deep neural networks. *Computers and Electronics in Agriculture* 176 (2020), 105664. ISSN: 01681699. DOI: 10.1016/j.compag.2020.105664.

[100] Y. Yue, J. H. Li, L. F. Fan, L. L. Zhang, P. F. Zhao, Q. Zhou, N. Wang, Z. Y. Wang, L. Huang and X. H. Dong. Prediction of maize growth stages based on deep learning. *Computers and Electronics in Agriculture* 172 (2020), 105351. ISSN: 01681699. DOI: 10.1016/j.compag.2020.105351.

[101] M. A. Zamora-Izquierdo, J. Santa, J. A. Martínez, V. Martínez and A. F. Skarmeta. Smart farming IoT platform based on edge and cloud computing. *Biosystems Engineering* 177 (2019), 4–17. ISSN: 15375110. DOI: 10.1016/j.biosystemseng.2018.10.014.

[102] M. D. Zeiler. ADADELTA: An Adaptive Learning Rate Method. (2012). ISSN: 09252312. DOI: http://doi.acm.org.ezproxy.lib.ucf.edu/10.1145/1830483.1830503. arXiv: 1212.5701.

[103] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8689 LNCS.PART 1 (2014), 818–833. ISSN: 16113349. arXiv: 1311.2901.

[104] Y. Zhao, A. B. Potgieter, M. Zhang, B. Wu and G. L. Hammer. Predicting Wheat Yield at the Field Scale by Combining High-Resolution Sentinel-2 Satellite Imagery and Crop Modelling. *Remote Sensing* 12.6 (2020), 1024. ISSN: 2072-4292. DOI: 10.3390/rs12061024.

PUBLICATIONS

PUBLICATION
I

Crop yield prediction with deep convolutional neural networks
P. Nevavuori, N. Narra and T. Lipping

Computers and Electronics in Agriculture 163.June (2019)
DOI: 10.1016/j.compag.2019.104859

Computers and Electronics in Agriculture 163 (2019) 104859

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

Crop yield prediction with deep convolutional neural networks

Petteri Nevavuori[b,*], Nathaniel Narra[a], Tarmo Lipping[a]

[a] Tampere University of Technology, Finland
[b] Mtech Digital Solutions Oy, Finland

ARTICLE INFO

Keywords:
Crop yield prediction
Convolutional neural network
Wheat
Barley
UAV
Multispectral
NDVI
Growth phase

ABSTRACT

Using remote sensing and UAVs in smart farming is gaining momentum worldwide. The main objectives are crop and weed detection, biomass evaluation and yield prediction. Evaluating machine learning methods for remote sensing based yield prediction requires availability of yield mapping devices, which are still not very common among farmers. In this study Convolutional Neural Networks (CNNs) – a deep learning methodology showing outstanding performance in image classification tasks – are applied to build a model for crop yield prediction based on NDVI and RGB data acquired from UAVs. The effect of various aspects of the CNN such as selection of the training algorithm, depth of the network, regularization strategy, and tuning of the hyperparameters on the prediction efficiency are tested. Using the Adadelta training algorithm, L^2 regularization with early stopping and a CNN with 6 convolutional layers, mean absolute error (MAE) in yield prediction of 484.3 kg/ha and mean absolute percentage error (MAPE) of 8.8% was achieved for data acquired during the early period of the growth season (i.e., in June of 2017, growth phase < 25%) with RGB data. When using data acquired later in July and August of 2017 (growth phase > 25%), MAE of 624.3 kg/ha (MAPE: 12.6%) was obtained. Significantly, the CNN architecture performed better with RGB data than the NDVI data.

1. Introduction

Development-minded farmers have practiced what is now known as precision agriculture long before the dawn of the computing age. They were able to deduce sources of field variability and the actions to take for trying to secure an enhanced level of crop yields. The farmers accomplished this by taking notes of their fields during growing seasons and harvest time operations and tried to figure out the best actions for the year to come based on the accumulated knowledge and experience. However, as studied by Wolfert et al. (2017), the increase in data-producing devices and sensors has been an on-going trend in agriculture having enabled the farmers to shift towards data-driven decision-making. This is commonly called smart farming. A comprehensive review of various objectives and techniques used in smart farming can be found in Kamilaris et al. (2017).

An important trend in smart farming is the use of remote sensing to facilitate the extraction of information relevant for data-driven decisions (Miyoshi et al., 2017; Matikainen et al., 2017). Remote sensing data can be acquired from satellites such as ESA's Sentinel-2A, for example. The problem with the satellite data is that if there is a cloud cover during the overflight of the satellite, no useful data are obtained. The spatial resolution of Sentinel imagery is at best 10 m, which is enough for many applications but too low to allow using texture-based information in the images. Satellite data contains predefined wavelength bands from both the visible and the Near Infrared (NIR) spectral regions. In satellite-borne sensors, designed keeping in mind agricultural applications, the spectral bands are optimized for the calculation of relevant indices such as the Normalized Difference Vegetation Index (NDVI), for example. The spatial and temporal resolution of satellite data will improve in years to come, however, cloud cover will remain an obstacle, especially in northern climate.

Using Unmanned Aerial Vehicles (UAVs), or drones, for data acquisition offers better spatial resolution, the data acquisition time can be selected by the user and the data can be acquired also in cloudy conditions. Spectral wavelengths can be selected by using appropriate camera; UAV-mountable RGB-NIR cameras are available at affordable price. The drawback is that the UAV has to be operated locally and managing the data and extracting relevant information requires highly specialized skills. As the variety of UAVs and UAV-mountable sensors is high compared to satellite-borne sensors, analysis frameworks and services based on UAV-borne data are not yet equally developed. In Näsi et al. (2017), extraction of information related to the biomass and nitrogen content of vegetation (barley and grass) in test fields using various modalities of remote sensing data (satellite/aircraft/drone using RGB/multispectral/hyperspectral sensors) has been considered.

Information relevant for decision making in agriculture can be

* Corresponding author.
E-mail addresses: petteri.nevavuori@mtech.fi (P. Nevavuori), nathaniel.narra@tuni.fi (N. Narra), tarmo.lipping@tuni.fi (T. Lipping).

https://doi.org/10.1016/j.compag.2019.104859
Received 17 April 2019; Accepted 13 June 2019

extracted from remote sensing data by means of machine learning. Traditional machine learning techniques involve feature extraction as an initial stage. Based on the features, different tasks such as crop classification, weed detection or yield prediction can be addressed. In Ruß (2009) several traditional machine learning techniques have been applied to the task of yield prediction. It is, however, often difficult to find optimal features and the ability of the traditional methods to learn from the data is limited. With advancements in computational technology, the development and training of novel multilayer algorithms has become feasible. These methods are commonly referred to as deep learning. Among the various deep learning paradigms, Convolutional Neural Networks (CNNs) have proved especially efficient in image classification and analysis. In case of CNNs no features need to be precalculated as the feature extraction operation is performed by the convolutional layers of the network and optimal features are obtained in the course of training. Due to this kind of structure, CNNs require large amounts of training data to converge. The advantage of CNNs compared to traditional machine learning methods in crop yield prediction is discussed, for example, in You et al. (2017). CNNs have been successfully applied to crop classification (Chunjing et al., 2017) and weed detection (Sa et al., 2017; Milioto et al., 2017).

In working towards an effective in-season crop yield predictor model for the northern climate, our effort in this preliminary study is to develop a CNN based deep learning framework using UAV-acquired multispectral data. RGB and NDVI images, representing patches of wheat and barley fields, are fed as input data to a CNN and training is performed to tune the network parameters. In addition to testing the usefulness of deep learning models for crop yield prediction in general, we also experiment with various setups and training schemes of the CNN model. Training a deep learning network is typically an iterative process as there is a substantial number of cross-related parameters to tune. We first select the most promising training algorithm from three candidates (see Section 3.1) and determine the optimal number of convolutional layers of the CNN. After that, we optimize the performance of the network in terms of regularization and parameters of the training algorithm. The optimized framework is evaluated using two types of input data (RGB and NDVI) and three patch sizes (10, 20 and 40 m).

2. Data and methods

2.1. Data acquisition

The nine crop fields selected for this study are located in the vicinity of the city of Pori (61°29′6.5″N, 21°47′50.7″E). The total area of the fields was approximately 90 ha. The main crops grown in the fields were wheat and malting barley, however the model was trained over the fields without making a distinction between the crop type.

Multispectral data were acquired from these fields during the growing season of 2017 (i.e., from June to August; see Table 1). The data were collected with a single Airinov Solo 3DR UAV equipped with Parrot's NIR-capable SEQUIOA-sensor. The images of individual spectral bands were stitched together to form complete orthogonal RGB and NDVI rasters of distinct fields using the Pix4D software.

The UAV data were organized into two sets according to the time of data acquisition to see if the phase of the growing season had an effect on predicting the yield from the input image. Growing phase here is defined as the percentage of total thermal time on the day of imaging. Thermal time for each day was calculated as the magnitude of daily average temperature above 5 °C. The temperature readings were downloaded from the Finnish Meteorological Institute. Beginning of July 2017 was chosen as the separating time point between the two data sets as the UAV data dispersed equally enough around that date. The data sets containing images only prior to July 2017 were labeled as *early* (growth phase < 25% of the total thermal time) and the remaining data as *late* (growth phase > 25% of the total thermal time). Details of the fields, crops, imaging dates and corresponding growth phases are listed in Table 1.

The field-wise image data were then processed using a sliding window to extract geolocationally matched pairs of input image frames (UAV data) and targets (yield data) of predefined size from all the fields. The step of the applied sliding window was chosen to be 10 m according to the resolution of Sentinel-2A satellite data considering the possibility of using satellite data as an additional input to the network in future studies. Image frames of sizes 10 × 10 m, 20 × 20 m and 40 × 40 m were considered. The resolution of the UAV data was 0.3125 m or 32 pixels per 10 m. The overall number of extracted frames according to crop fields is given in Table 2. The individual data frames were treated as independent inputs fed to the CNN models. The process of data preparation prior to and during training is illustrated in Fig. 1.

The harvest yield data was acquired during September 2017 using two distinct setups attached to the harvesters: Trimble CFX 750 and John Deere Greenstar 1. As the yield measurement devices produce an irregular set of data points with multiple attributes, the data had to be processed to be handled as rasters of field-wise yield from the viewpoint of the trainable network. The data points were first filtered according to (Tiusanen, 2017) to preserve only points corresponding to harvester speed between 2 and 7 km/h and yield between 1500 and 15,000 kg/ha. The filtering and generation of rasterizable vector files was done using the FarmWorks software. The field-wise vector data files were then rasterized by interpolating them using an exponential point-wise inverse distance algorithm. Yield values constitute targets the model is trying to predict during the training of the CNNs. Thus, yield values were also extracted using sliding windows similar to the UAV images to have geolocationally matching pairs of inputs and targets. Yield values were then averaged over the analysis window to obtain scalar target values. The histograms and statistics of yield values for point data as well as window-averaged data using three sample area window sizes (10 m, 20 m and 40 m) over all crop fields are given in Fig. 2. As can be expected, the larger the window, the more concentrated the yield values are around the mean.

For clarity, we also visualize several NDVI and RGB input images of the largest sample area window size (40 m) with their corresponding yields in Fig. 3 with the color bar corresponding to yield image value range. The images with similar identifiers are from the same location. However, the target for the network will be the mean of the yield values over the analysis window corresponding to the input area. It is also important to note that the network was trained separately for RGB and NDVI input images so that the possible misalignment between the two image sources does not affect prediction results. This kind of approach enables us to evaluate which one of the two input sources, RGB or NDVI, gives better prediction results.

2.2. Building the convolutional neural network

Convolutional neural networks, or CNNs, are deep learning models specialized in handling grid-like data. Such data can be images or rows of multi-column data. *Deep learning* refers to models composed of multiple layers. Generally, a model is viewed as deep if it has at least an input layer, one hidden layer and an output layer. The term *neural* on the other hand refers to the fact that originally the operation principle of artificial neural networks was taken from that of the brain, containing neurons as its basic building blocks. Compared to traditional feedforward neural networks, CNNs possess some special features making them extremely efficient in finding salient features within the data. Some of these features are:

1. exploitation of the convolution operation
2. post-convolution pooling
3. specific non-linear activation functions.

In the following we provide a brief description of these elements

Table 1
Details of crops and their varieties sown in each of the 9 fields in 2017. Thermal times for each crop variety are taken from a report published by Laine et al. (2017). Sowing dates and imaging dates are used to calculate the growth phase as a fraction of the total thermal time for the crop variety. Images with dates prior to 1st of July form the early data set and the remaining images the late one.

Field #	Size (ha)	Mean yield (kg/ha)	Crop (*Variety*)	Thermal time	Sowing date	Imaging date	Growth phase
1	5.96	5098	Wheat (*Zebra*)	1052	10 May	17 Aug	83%
2	10.26	6054	Barley (*Trekker*)	979.7	16 May	8 Jun	15%
						27 Jul	64%
3	2.97	8971	Barley (*Trekker*)	979.7	17 May	8 Jun	15%
						27 Jul	64%
4	13.05	4673	Barley (*RGT Planet*)	982.2	15 May	6 Jul	42%
5	4.66	6482	Barley (*Propino*)	981.4	15 May	15 Jun	22%
6	7.29	6884	Barley (*Propino*)	981.4	15 May	15 Jun	22%
7	10.92	7568	Barley (*Harbinger*)	976.3	24 May	6 Jul	36%
8	15.28	7585	Barley (*Trekker*)	979.7	18 May	1 Jun	10%
						13 Jul	49%
9	18.86	6991	Wheat (*KWS Solanus*)	1065	13 May	15 Jun	21%
						6 Jul	72%

Table 2
Number of data frames extracted from each field using frame sizes of 10 m, 20 m and 40 m. The number of frames decreases slightly with increasing frame size due to field edge effects.

Field #	10 × 10 m data frames	20 × 20 m data frames	40 × 40 m data frames	Mean data frame count
1	761	745	735	*747*
2	1102	1159	1150	*1137*
3	783	731	691	*735*
4	1494	1486	1454	*1478*
5	610	586	590	*595*
6	942	931	916	*930*
7	1240	1247	1224	*1237*
8	3736	3786	3812	*3778*
9	4556	4548	4520	*4541*
∑	**15224**	**15219**	**15092**	*15178*

with additional information on other key elements of CNNs such as batch normalization and regularization. We evaluated various setups of these CNN elements to find the best-performing algorithm and assess its performance in crop yield prediction.

2.2.1. Convolution operation

The convolution operation is the first of multiple transformations performed in a convolutional layer of CNNs. Generally, the convolution operation can be described as calculating the sum of products between a set of input values and values of a convolutional *kernel*, also called a *filter*. In CNN, the kernel values are trained to find optimal features from the point of view of the task to be solved (in our case, predicting crop yield). The operating principle of the kernel is depicted in Fig. 4 and the position of convolutional layers in the overall structure of the CNN used in this study can be seen from Fig. 7.

2.2.2. Batch normalization

While not a requirement for CNNs, the state-of-the-art is to apply batch normalization (Ioffe and Szegedy, 2015) as a constituent of deep learning model layers. Batch normalization is an optimization strategy

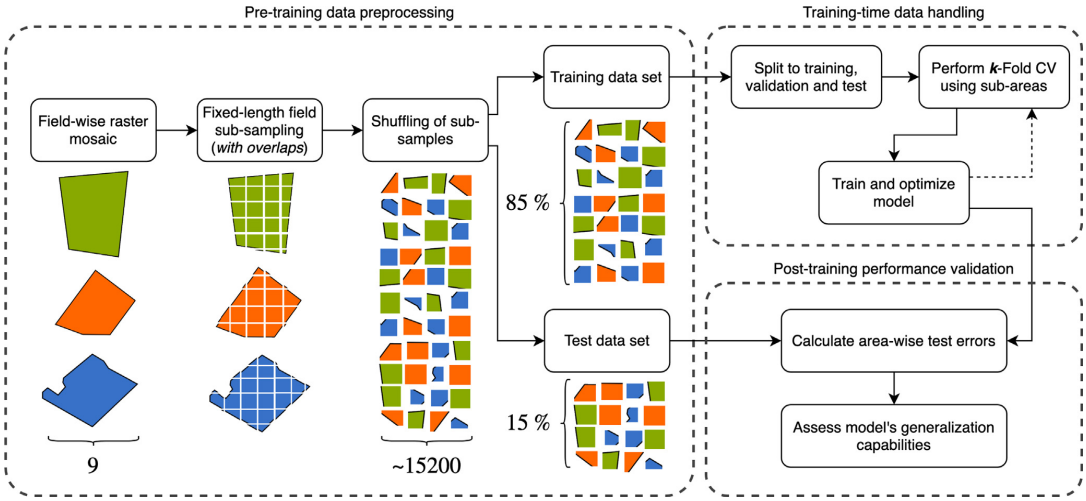

Fig. 1. All nine fields were first split to overlapping data frames of sizes 10 m, 20 m and 40 m. A dedicated holdout test data set was then built from 15% of shuffled data frames; these data were never presented to the model during training. The remaining 85% of data frames were then used for training the models with *k*-Fold Cross Validation. After the training phase of each model was completed, the test errors were calculated using the holdout test data set to validate the performance of the trained model.

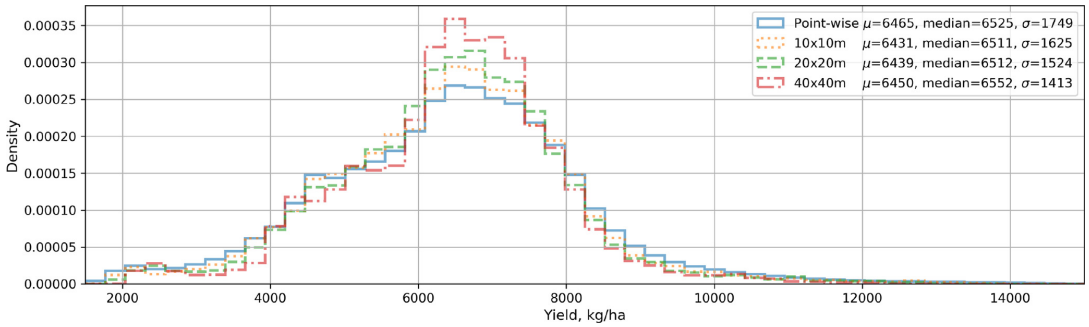

Fig. 2. Histograms and statistics of point-wise and window-averaged yield data. The histograms are normalized to probability densities to make point-wise graphs align with sliding window histograms count-wise. While sliding windows contain no-data points near field edges, only points containing data were taken into account.

for training deep models more efficiently. Batch refers to a subset of training data used for updating the model parameters (including kernel values) at a single iteration, albeit the term mini-batch is generally used to distinguish the whole data set (batch) from it's subset (mini-batch). It has been shown that normalizing the network layers for each batch (or mini-batch) of data stabilizes the learning, allowing to use higher learning rates and thus resulting in faster learning (Goodfellow et al., 2016). There are different implementations of batch normalization; the implementation used in the CNN of this study follows Eq. (1), where x is a mini-batch of activations, ϵ a non-significant constant to prevent numerical underflow, γ is the momentum and b is a layer-wise bias:

$$y = \frac{x - \mu_x}{\sigma_x + \epsilon} * \gamma + b. \tag{1}$$

2.2.3. Max pooling

The convolution operation is usually followed by pooling. Pooling means grouping of adjacent values using a selected aggregation function, which in our case was taking the maximum (hence max pooling) over the neighboring values within a predefined window. The step size of moving this window along the feature map is called *stride*. Pooling effectively diminishes the input image dimensions making the detected features more coarse and thus more robust to small variations (Goodfellow et al., 2016). The amount of dimension reduction is controlled by the stride parameter. The stride dictates how many applications of the pooling window are performed. An example of max pooling is given in Fig. 5 and the position of pooling in the overall structure of the CNN used in this study can be seen from Fig. 7.

2.2.4. Rectified linear units

A key element in any neural network is the layer-wise activation function of the neurons. A variety of activation functions have been designed, but the use of the rectified linear function in the activation units is the current standard for CNNs (He et al., 2015; Goodfellow et al., 2016). Activation units employing rectified linear functions are commonly referred to as ReLUs. The operating principle of this activation function is to allow only positive inputs to proceed linearly and is depicted in Fig. 6. We too use ReLUs as the activation functions in both the convolutional as well as the fully connected layers (see Fig. 7).

2.2.5. Fully connected layers

The convolutional layers of a CNN extract salient features from input images, i.e., factors with highest descriptive power regarding the data producing process. To utilize the learned features in a regression or a classification task, they have to be successfully mapped to a target value. This is performed typically by adding fully connected (FC) layers after the convolutional layers. The term *fully connected* refers to the principle that in these layers, each neuron (or unit) of the previous layer has a connection to each unit of the layer in question. Increasing the number of FC layers increases the capacity of the network to learn the mapping between the features and the target. It also increases the burden of optimization, as in FC layers the number of connections grows exponentially with the number of layers.

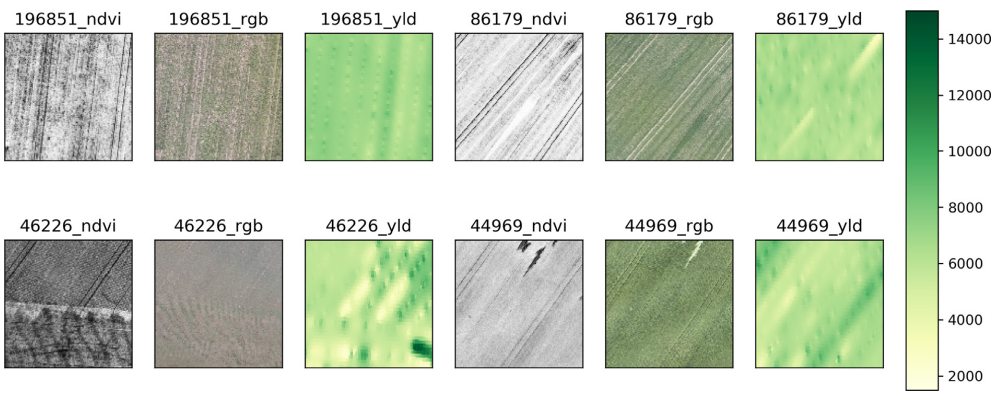

Fig. 3. Visualizations of NDVI and RGB input images and yield targets. The identification numbers above the images denote the distinct area from which the images were extracted.

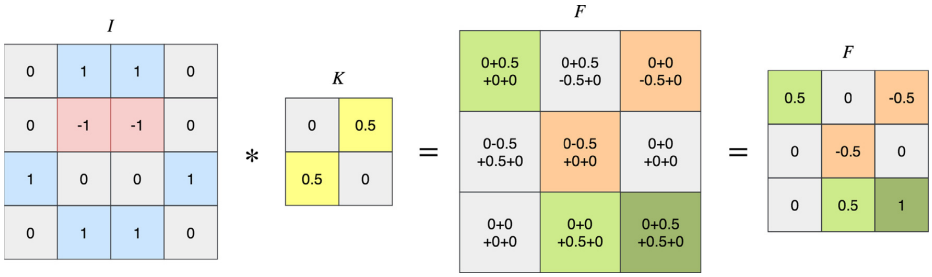

Fig. 4. The kernel K is applied to the input image I in a sliding window fashion. With each application, a sum of element-wise products is calculated and stored. After the kernel has been applied to the whole image, a complete feature map F is produced. A feature map indicates the result of detecting a kernel-specific feature in the input image.

2.2.6. Regularization strategies

Increasing the depth of a deep learning model allows it to learn more complex functions. This is also known as increased model's capacity (Goodfellow et al., 2016). When a model's capacity increases, it becomes more prone to overfitting to the training data in which case its ability to generalize (and, therefore, its performance on test data) deteriorates. This can be avoided with regularization, which effectively reduces the model's capacity diminishing the gap between training and test errors. Regularization is a comprehensive term for methods in machine learning that are used to lower the test error without focusing on training error.

In our model we make use of two distinct regularization strategies. First of the two is the L^2-penalty, also known as the weight decay. It diminishes the model's layer-wise parameters with each training iteration. When applied in conjunction with training by error back-propagation, the most relevant of the model's parameters retain their magnitude while non-relevant ones diminish. The second implemented regularization strategy is called early stopping. It is a robust meta-algorithm integrated into the training process to halt the training after n non-improving iterations. The hyperparameter n is called *patience* (Goodfellow et al., 2016).

2.2.7. Overall architecture

The basic architecture of the CNN implemented in this study follows closely the one reported by Krizhevsky et al. (2017). Their model performed extremely well in ImageNet Large Scale Visual Recognition Competition (Russakovsky et al., 2015) attaining top classification results in multiple categories. The general topology of our network is depicted in Fig. 7. The network was implemented using the PyTorch framework (Paszke et al., 2017). In our network we use non-overlapping pooling windows with pooling window size of 5 and a pooling stride matching the pooling window size. We also include the pooling function only in the first and the last convolutional layer. The reason for this is that at the lowest (i.e., in the case of 10 m ground resolution) our image size is 32 × 32 pixels and too many pooling operations would cause the data representation to collapse. This way our network is also scalable with respect to the number of layers. Regardless of the number of source image bands, our convolutional layers contain 64 kernels except for the last layer containing 128 kernels. Krizhevsky et al. (2017) incorporated two FC layers to the model with 2048 neurons per layer. We used similar number of layers with half the width, i.e., 1024 neurons per layer.

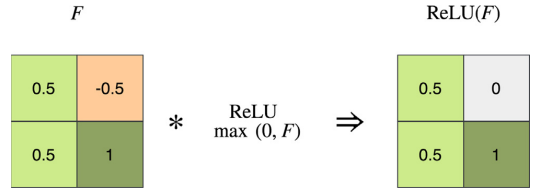

Fig. 6. An illustration of the effect of applying the rectified linear activation function to a pooled feature map.

2.3. Optimizing the network

Finding the optimal configuration of any deep learning network is an iterative process, where the model's parameters are initialized and tuned multiple times. The goal is to find a set of model's parameters (weights, biases, etc.) and hyperparameters (learning rate, optimizer coefficients, etc.) that in conjunction produce the best performance. The output of the iterative process is a single model usually performing best when compared to other models produced within the process. We used absolute error between the network output and the target value (i.e., crop yield values) as the performance measure. In machine learning, the best performing model is considered to be the one that generalizes well to previously unseen data. To measure the generalization performance across training instances, we extracted and reserved a subset of data as a holdout test set. This test data set was used outside of the training loop to ensure that the model never learned from it. With the rest of the data we performed k-fold cross validation using three folds per epoch. An epoch is a single complete iteration over the full training data set consisting of windowed image samples of all 9

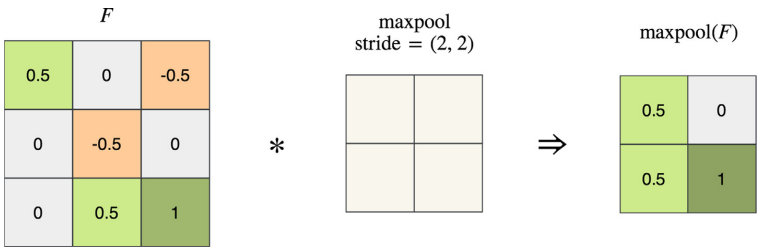

Fig. 5. An example of a simple application of max pooling, where the pooling is applied to a feature map F with pooling window size of 2 × 2 and a stride equaling the kernel size.

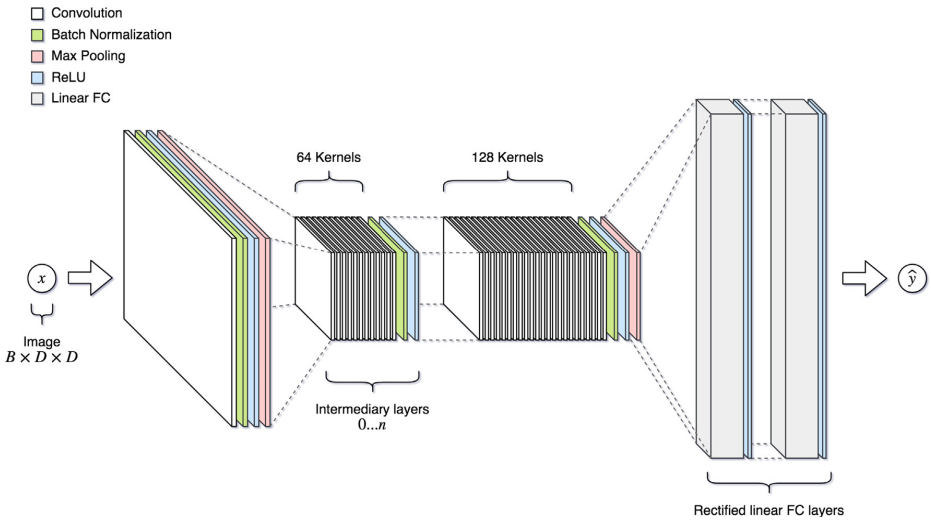

Fig. 7. The overall topology of the implemented CNN. Network's inputs can be single-band or multi-band images (B) with varying dimensions (D). The network has at least two convolutional layers accompanied with two fully connected layers. The depth of the network is controlled by the number of intermediary convolutional layers. The last convolutional layer has 128 kernels while the intermediary layers have 64 kernels. Max pooling is applied only in the first and last convolutional layers so that the size of the data representation stays consistent when network depth is varied.

fields.

The best training algorithm was evaluated among three options: Stochastic Gradient Descent with momentum (SGD-momentum) (Bottou, 1998), RMSprop (Hinton et al., 2014) and Adadelta (Zeiler, 2012). These training algorithms are suggested in Goodfellow et al. (2016) and they are also among the ones compared in Karpathy and Fei-Fei (2017). In a preliminary test, the three algorithms were tested for convergence by training the network for three epochs. Training was performed for each of the three data window sizes and each of the four sets of input data. The batch size was varied from 2^5 to 2^{10}. The worst performing algorithm was excluded and a second test performed on the remaining two by fixing the batch size to 128 (2^7) and training for 50 epochs, a number consistent across the training of almost every model.

The effect of the depth of the network on the performance was evaluated by training models with 4, 6, 8, 10 and 12 convolutional layers over 50 epochs per training session. The training was conducted for the NDVI and RGB images from early and late data sets and with all three input image dimensions using the previously selected training algorithm. At this stage, the best performing combination of - network depth, image type (NDVI or RGB) and window size - was selected based on error performance over the test data.

In the next step, the chosen training algorithm's hyperparameters (i.e., the learning rate and the past iterations' error correction adjustment) were tuned. In order to evaluate performance, benchmark models were created by initializing a model for each of the four data sets (i.e., early and late, RGB and NDVI). The hyperparameter values were searched over a coarse grid for values producing lowest test errors, followed by a more refined random search in the vicinity of the coarse minimum. Sensitivity of the network performance to initial values of the CNN parameters was also assessed.

In the last step, the hyperparameter combinations producing the best performance were used to test and tune the effect of regularization algorithms. Tuning of the weight decay coefficient (L^2 regularization) for early and late data sets was performed by searching over a coarse grid of values followed by refined search. Subsequently, the effect of early stopping was tested using values 10, 20, 30, 40 and 50 for the patience parameter (see Section 2.2.6).

3. Results

We measure the performance of the CNN by *mean absolute error*, i.e., the mean absolute difference between the true yield value and the CNN output (predicted value). This can also be called *loss*. We consider two different errors: the training error, obtained for the same data the network is trained with, and the test error, obtained for the data set aside for testing. The former one indicates how well the model is able to fit to the data, i.e., what is its capacity, while the latter one indicates how well the network is able to generalize to unseen data samples.

3.1. Selection of the training algorithm

Of the three training algorithms – Adadelta, SGD-momentum and RMSprop – the RMSprop showed poor convergence and was therefore ruled out from subsequent tests. Between the two remaining algorithms, Adadelta outperformed SGD-momentum and was chosen as the training algorithm for further experiments (see Table 3).

3.2. Depth of the network

In Fig. 8 the test and training errors for the three window sizes and for various networks depths are shown for the RGB data of earlier growth phase. The largest window (size 40 × 40 m) produced lowest test errors in majority of cases regardless of the network depth. The colored areas indicate gaps between training (lower bound of the area) and test (upper bound of the area) errors, also referred to as

Table 3
Lowest mean absolute test errors (kg/ha) observed among the three data window size configurations (10 m, 20 m and 40 m) with 50 epochs of training and a batch size of 128 samples for each source image type. Adadelta performed best with almost every source image configuration.

Optimizer	NDVI early	NDVI late	RGB early	RGB late
SGD with Momentum	1751.2	1183.7	1231.5	985.0
Adadelta	842.8	1165.1	836.2	989.5

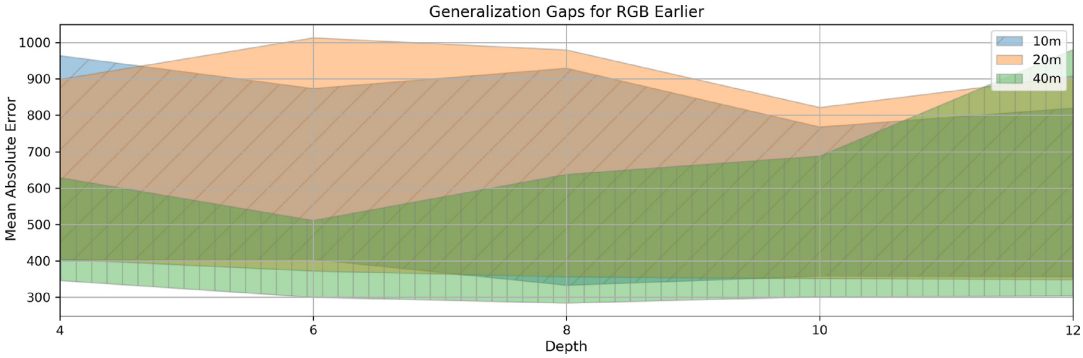

Fig. 8. The generalization gaps with early RGB images. The generalization gap is depicted as the difference between the training and the test errors. It shows how close the test error is to the effective capacity of the model, the training error. The lowest test errors (upper bound of the area) were achieved rather consistently with the source image window size of 40 × 40 m.

generalization gaps. The lowest test and training error combination is obtained with 6 convolutional layers. Also, the 40 m window and network depth of 6 convolutional layers result in the narrowest generalization gap.

3.3. Optimization of Adadelta hyperparameters

The hyperparameters of the chosen training algorithm, Adadelta, were tuned by considering the effects of the adaptive learning rate and the coefficient adjusting the effect of past iterations' error corrections (in the form of squared gradients) on learning. The latter is effectively similar to momentum, defining the magnitude by which the past affects the current learning process. The previous experiments were performed using default values for these hyperparameters, i.e., 1.0 for the learning rate, 0.9 for the coefficient for computing a running average of squared gradients, and 0 for the weight decay (see Table 4).

The initial grid search was conducted with hyperparameter values similar to those found in the original Adadelta research paper with an epoch limit of 50 compared to the original study's 6 epochs (Zeiler, 2012). For the early RGB data set, the optimal values were approximately 8×10^{-3} for the learning rate and 0.58 for the coefficient adjusting the effect of past iterations' error. For the late data set the respective values were 10^{-4} and 0.9. The effect of hyperparameter tuning on the performance of the network can be seen from the results in Table 4.

Table 4
The total improvement in test error compared to the benchmark model when using regularization and optimization of training algorithm hyperparameters. The benchmark models were trained with the early and late RGB image data with default parameters. Window size was 40 × 40 m. Errors are reported as mean absolute error (MAE) and mean absolute percentage error (MAPE). The best results are formatted in bold.

	RGB early		RGB late	
	MAE [kg/ha]	MAPE	MAE [kg/ha]	MAPE
Benchmark learning rate: 1.0 past err. coeff.: 0.9 weight decay: 0 patience: ∞	997.8	18.3%	1021.5	19.5%
with **Optimized Adadelta params.** learning rate (early/late): 0.008/0.0001 past err. coeff. (early/late): 0.58/0.9	546.2	9.6%	**624.3**	**11.4%**
and with **L²-regularization** weight decay: 0.001	558.4	9.4%	700.4	13.1%
and with **Early Stopping** patience: 50	**484.3**	**8.8%**	680.4	12.6%

3.4. Optimization of the regularization parameters

The CNN models using optimal hyperparameters for the Adadelta training algorithm were trained next with early and late RGB data sets of 40 × 40 m window to determine the effect of regularization on the prediction error and to tune the regularization parameters. The tuning of weight decay coefficient with grid search first and zoomed-in random search after that resulted in the optimal coefficient value of 10^{-3} for both data sets. The optimal patience values were around 50, again for both data sets. It was observed that the increase in patience increased the training time significantly. The selected patience value allowed the models for both data sets to converge in approximately 250 epochs. The effect of using the L^2-regularization alone and combined with early stopping can be seen from Table 4.

4. Discussion and conclusions

This study presents a training paradigm of a CNN based deep learning model for predicting wheat and barley yield. The results indicate that the best performing model can predict within-field yield with a mean absolute error of 484 kg/ha (MAPE: 8.8%) based only on RGB images in the early stages of growth (< 25% total thermal time). The model for RGB images at later growth stage returned higher error values (MAE: 680 kg/ha; MAPE: 12.6%). In searching for optimal performance, the input data window size (10 m, 20 m, 40 m), the data acquisition time (early vs. late) and data modality (RGB vs. NDVI) were varied. The 9 fields included in the study were imaged by a camera mounted to UAV and together taken as a source of > 10,000 input image frames covering a total area of 90 hectares. Network depth (i.e., the number of convolutional layers), the training algorithm and its hyperparameters as well as the CNN regularization scheme were also optimized. The lowest error was achieved using a network consisting 6 convolutional layers followed by two fully connected layers regularized with L^2-regularization coefficient of 10^{-3} and early stopping patience of 50. The optimizer was also tuned for the optimal value of the learning rate (8×10^{-3}) and the coefficient adjusting the effect of past iterations' error corrections (0.58). The results show that the lowest test errors were achieved with the largest data window size tested (40 m).

The training of any neural network is always influenced by the

combined randomness resulting from how the data is shuffled between cross validation folds, the optimization process and other factors. This in turn means that, while discrete error metrics produce a ranking across hyperparameter setups, slight variations between test errors can be attributed to the random nature of the optimization process as a whole. We optimized distinct models for early and late RGB data sets. The best performing model used RGB images from the early growing season and benefited from regularization. The model using the late RGB images didn't gain from added regularization, as the best performance was achieved during the tuning of the training algorithm (see Table 4).

In yield prediction, the shift from using traditional regression methods (Ruß, 2009) towards artificial neural network based methods (Chlingaryan et al., 2018) has resulted in improved performance (Jiang et al., 2004; Kaul et al., 2005). Among these ANN based studies, those using remote sensing image data to train their prediction models have achieved low prediction errors (≈ 5%). These models are specific to the crop types whose images they are trained with (e.g., soybean, wheat, rice). Jiang et al. (2004) working with satellite images reported an average relative winter wheat prediction error of 3.5%. The indices used for training the model were: NDVI, surface temperature, absorbed photosynthesis active radiation, water stress index and 10-year average crop yield. Bose et al. (2016) employed spiking neural networks to estimate winter wheat yield from satellite based NDVI images at the region level, achieving a best average relative error of 4.35%. In their recent work, You et al. (2017) leveraged advanced hybrid machine learning algorithms to achieve very low soybean yield prediction errors (3.19–5.65%) using only satellite images.

A commonality among these studies is the use of satellite imagery and large spatial scales of their analyses (region or county level predictions). Our study, in contrast, seeks to perform predictions at the intra-field scale using UAV based images in order to spatially analyze yield within the field. In one of the earliest studies on this topic, Davis and Wilkinson (2006) used satellite imagery of wheat crop (visible, infrared and radar) and an ANN model showing promising results (error slightly above 10%) for a single field (≈ 36 ha). Khanal et al. (2018) employed various machine learning algorithms (including neural networks) and aircraft based multi-spectral images to predict corn yield on a single field (17.5 ha). A few studies have applied ANN's for classifying crops (Rebetez et al., 2016) and yield (Pantazi et al., 2016) at the intra-field scale. However, rather than classifying within yield categories this study aims at quantitative predictions. Models at intra-field scale would offer the individual farmer the possibility of in-season monitoring of crop, which would enable decision support systems for interventions necessary to achieve higher yields. Models trained at large regional scales rarely extrapolate to finer scales, though efforts are underway to develop scalable models (Donohue et al., 2018). The methodology introduced by You et al. (2017) shows great potential and as authors claim its scalability, it would certainly be of interest in testing at the intra-field scale.

One important aspect of remote sensing based yield prediction has been finding image channels or indices containing the most discriminating features necessary for analysis (Panda et al., 2010). Consequently, the finding in this study that the RGB images perform better than NDVI, assumes significance and aligns with the study for estimating biomass and crop height (Näsi et al., 2017). This indicates that multiple spectral bands increase the information content in comparison to the condensed NDVI image. From a utility perspective, RGB cameras are cheaper with most commercially available UAVs already fitted with decent cameras able to produce images of high resolution. Models that can perform well without the need for expensive specialized equipment will make the analyses accessible to an individual farmer.

The relationship between crop yield and its environment is non-linear and may not be sufficiently contained in the features captured by images. As shown by the studies reporting low prediction error levels, by adding multi/hyper spectral data, temporal image data, soil and environmental features in the feature matrix, it is possible to constrain the resulting model error effectively. Considering that this study models the yield based only on images, the resulting prediction error of 8.8% is promising. Additionally, collection of multi-year yield maps from sensor-equipped harvesters would add valuable information to act as ground truth. More than 90 hectares of fields were mapped in this study (2017 season). In 2018 a similar set of data has been acquired while the data acquisition will be continued in 2019. This valuable database will serve to further train, tune and verify the current model for greater accuracy. An additional limitation of this study is that only minimal preprocessing was applied to the source data. Developing automated error correction methods for data preprocessing would be another important task when developing remote sensing based crop yield models. Careful artifact rejection and preprocessing would probably benefit the modeling considerably.

In conclusion, this study is an important step towards establishing a combined model for wheat and barley yield prediction in the Finnish continental subarctic climate. The long summer growing days in this region presents a unique profile of temperature and photoperiod, justifying a region specific deep learning model for these crops. By collecting data using commercial off-the-shelf UAV and camera packages, we focus our attention on a spatial scale that enables us to predict intra-field yield distribution within the context of individual farm crop monitoring. The results indicate that the CNN models are capable of reasonable accurate yield estimates based on RGB images. It is worth noting that the CNN architecture seemed to be performing better with RGB images than NDVI images. In the future, the developed model will be trained on a larger set of features (climate and soil) along with time series image data to tune the trained model for accuracy.

Acknowledgments

We would like to give special acknowledgments to Mtech Digital Solutions Oy for partly funding this research. We also want to thank the MIKÄ DATA project's research group of Tampere University of Technology for providing the data and additional knowledge required to use the data appropriately. A special thanks to Mikko Hakojärvi from Mtech Digital Solutions Oy for providing insight into the agricultural knowledge domain.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.compag.2019.104859.

References

Bose, P., Kasabov, N.K., Bruzzone, L., Hartono, R.N., 2016. Spiking neural networks for crop yield estimation based on spatiotemporal analysis of image time series. IEEE Trans. Geosci. Remote Sens. 54 (11), 6563–6573.

Bottou, L., 1998. On-line Learning in Neural Networks. Cambridge University Press, New York, NY, USA Ch. On-line Le, pp. 9–42.

Chlingaryan, A., Sukkarieh, S., Whelan, B., 2018. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: a review. Comput. Electron. Agric. 151 (November 2017), 61–69.

Chunjing, Y., Yueyao, Z., Yaxuan, Z., Liu, H., 2017. Application of convolutional neural network in classification of high resolution agricultural remote sensing images. ISPRS – Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. XLII-2/W7, 989–992.

Davis, I.C., Wilkinson, G.G., 2006. Crop yield prediction using multipolarization radar and multitemporal visible/infrared imagery. In: Proc.SPIE 6359, 6359–6359 – 12.

Donohue, R.J., Lawes, R.A., Mata, G., Gobbett, D., Ouzman, J., 2018. Towards a national, remote-sensing-based model for predicting field-scale crop yield. Field Crops Res. 227, 79–90.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision 2015 Inter, pp. 1026–1034.

Hinton, G., Srivastava, N., Swersky, K., 2014. Neural Networks for Machine Learning Lecture 6a: Overview of minibatch gradient descent.

Ioffe, S., Szegedy, C., 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.

Jiang, D., Yang, X., Clinton, N., Wang, N., 2004. An artificial neural network model for

estimating crop yields using remotely sensed information. Int. J. Remote Sens. 25 (9), 1723–1732.
Kamilaris, A., Kartakoullis, A., Prenafeta-Boldú, F.X., 2017. A review on the practice of big data analysis in agriculture. Comput. Electron. Agric. 143, 23–37.
Karpathy, A., Fei-Fei, L., 2017. Deep visual-semantic alignments for generating image descriptions. IEEE Trans. Pattern Anal. Mach. Intell. 39 (4), 664–676.
Kaul, M., Hill, R.L., Walthall, C., 2005. Artificial neural networks for corn and soybean yield prediction. Agric. Syst. 85 (1), 1–18.
Khanal, S., Fulton, J., Klopfenstein, A., Douridas, N., Shearer, S., 2018. Integration of high resolution remotely sensed data and machine learning techniques for spatial prediction of soil properties and corn yield. Comput. Electron. Agric. 153 (August), 213–225.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2017. ImageNet classification with deep convolutional neural networks. Commun. ACM 60 (6), 84–90.
Laine, A., Högnäsbacka, M., Niskanen, M., Ohralahti, K., Jauhiainen, L., Kaseva, J., Nikander, H., 2017. Virallisten lajikekokeiden tulokset 2009-2016, 262.
Matikainen, L., Karila, K., Hyyppä, J., Puttonen, E., Litkey, P., Ahokas, E., 2017. Feasibility of multispectral airborne laser scanning for land cover classification, road mapping and map updating. ISPRS - Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. XLII-3/W3, 119–122.
Milioto, A., Lottes, P., Stachniss, C., 2017. Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in CNNs. Adv. Intell. Syst. Comput. 531, 105–121.
Miyoshi, G.T., Imai, N.N., de Moraes, M.V.A., Tommaselli, A.M.G., Näsi, R., 2017. Time series of images to improve tree species classification. ISPRS – Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. XLII-3/W3, 123–128.
Näsi, R., Viljanen, N., Kaivosoja, J., Hakala, T., Pandžić, M., Markelin, L., Honkavaara, E., 2017. Assessment of various remote sensing technologies in biomass and nitrogen content estimation using an agricultural test field. ISPRS – Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. XLII-3/W3, 137–141.
Panda, S.S., Ames, D.P., Panigrahi, S., 2010. Application of vegetation indices for agricultural crop yield prediction using neural network techniques. Remote Sens. 2 (3), 673–696.
Pantazi, X.E., Moshou, D., Alexandridis, T., Whetton, R.L., Mouazen, A.M., 2016. Wheat yield prediction using machine learning and advanced sensing techniques. Comput. Electron. Agric. 121, 57–65.
Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A., 2017. Automatic differentiation in PyTorch. In: NIPS-W.
Rebetez, J., Satizábal, H.F., Mota, M., Noll, D., Büchi, L., Wendling, M., Cannelle, B., Pérez-Uribe, A., Burgos, S., 2016. Augmenting a convolutional neural network with local histograms - a case study in crop classification from high-resolution uav imagery. In: 24th European Symposium on Artificial Neural Networks, ESANN 2016, Bruges, Belgium, April 27–29, 2016.
Ruß, G., 2009. Data mining of agricultural yield data: a comparison of regression models. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 5633 LNAI, 24–37.
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet large scale visual recognition challenge. Int. J. Comput. Vision 115 (3), 211–252.
Sa, I., Chen, Z., Popovic, M., Khanna, R., Liebisch, F., Nieto, J., Siegwart, R., 2017. weedNet: Dense Semantic Weed Classification Using Multispectral Images and MAV for Smart Farming.
Tiusanen, J., 2017. Aineiston käsittely ja muotoilu. Käytännön Maamies.
Wolfert, S., Ge, L., Verdouw, C., Bogaardt, M.J., 2017. Big data in smart farming – a review. Agric. Syst. 153, 69–80.
You, J., Li, X., Low, M., Lobell, D., Ermon, S., 2017. Deep Gaussian process for crop yield prediction based on remote sensing data. In: 31th AAAI Conference on Artificial Intelligence, pp. 4559–4565.
Zeiler, M.D., 2012. ADADELTA: An Adaptive Learning Rate Method. undefined.

PUBLICATION
II

A Data Driven Approach to Decision Support in Farming
N. Narra, P. Nevavuori, P. Linna and T. Lipping

Information Modelling and Knowledge Bases XXXI 2020
DOI: 10.3233/FAIA200014

A Data Driven Approach to Decision Support in Farming

Nathaniel NARRA [a,1], Petteri NEVAVUORI [a,b], Petri LINNA [a] and Tarmo LIPPING [a]
[a] *Computing Sciences Unit, Tampere University, Finland*
[b] *Mtech Digital Solutions Oy*

Abstract. Precision Agriculture and Smart Farming are increasingly important concepts in agriculture. While the first is mainly related to crop production, the latter is more general, which also involves the carbon capture capacity of crop fields (Carbon Farming), as well as optimization of the farming costs taking into account the dynamics of market prices. In this paper we present our recent work in building a web-based decision support system for farmers to help them comply with these trends and requirements. The system is based on the Oskari platform, developed in Finland for the visualization and analysis of geospatial data. Our main focus so far has been in developing tools for Big Data and Deep Learning based modelling which will form the analytical engine of the decision support platform. We first give an overview on the various applications of deep learning in crop production. We also present our recent results on within-field crop yield prediction using a Convolutional Neural Network (CNN) model. The model is based on multispectral data acquired using UAVs during the growth season. The results indicate that both the crop yield and the prediction error have significant within-field variance, emphasizing the importance of developing field-wise modelling tools as a part of a decision support platform for farmers. Finally, we present the general architecture of the overall decision support platform currently under development.

Keywords. Smart farming, crop yield prediction, decision support, deep learning

1. Introduction

For ages, farmers have made notes on their farming activities to undertake proper actions to increase the productivity of their fields. The means and extent of these actions have changed in time - instead of digging ditches using spades, whole fields can be levelled by modern powerful machinery and fertilizers and pesticides are used to increase the yield. However, much of the decision-making regarding these modern means of cultivation is still done by intuition. At the same time, increasingly strict environmental regulations concerning farming and competition on the crop market forces farmers to optimize their cultivation activities to the limits. This optimization has multiple targets such as yield, carbon capture, environmental requirements, market prices etc. As in many other industries, data-driven modelling of production and developing model-based decision support systems has become an active area of research and development in agriculture [1].

[1]Corresponding Author: Pohjoisranta 11A, Tampere University, Pori, Finland; E-mail: nathaniel.narra@tuni.fi.

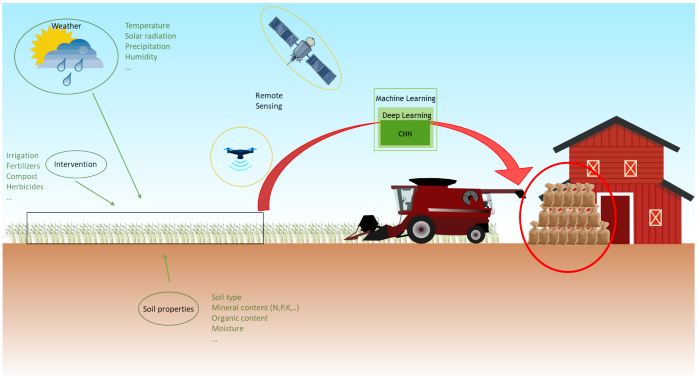

Figure 1. Illustration of crop production as a system with big data input and yield as the output.

Crop production can be viewed in engineering parlance as a system with input and output (see Figure 1). Climate, soil and other biotic and abiotic factors that have a bearing on plant growth (i.e. system dynamics) can be considered as input. These also include interventions conducted to either stimulate plant productivity or mitigate factors detrimental to productivity. With increasing accessibility in terms of affordability, ease of use and technical reliability, Internet of Things (IoT) and remote sensing technologies have enabled high amounts of data to be collected from crop fields. These data can either represent the input factors directly or constitute an indirect representation of the effects of these factors on the system (i.e. crop). Multi- or hyperspectral remote sensing is a common example of the latter type of data. Having all these data available, often in real time, opens up new avenues for studying the contribution of various factors to the yield (i.e., the output of the system).

The collected data comes in vast amounts and its analysis involves high computational cost that often preclude traditional analytical methods. Also, yearly variation in various factors such as climate, for example, suggests that the analytical tools used for decision support in agriculture have to be capable of learning from the environmental conditions. For example, in addition to training a model for crop yield prediction, it is also important to learn how the model needs to be adapted to the changes in the environment. Recent developments in Artificial Intelligence (AI) and, more specifically, in Machine Learning (ML) have produced promising new models for extracting information from large heterogeneous data sets. These methods have been extensively applied to study various aspects of agriculture [2][3][4].

A common term for the recent advancements in ML is Deep Learning (DL). DL refers to Neural Network type structures containing multiple computational layers with often thousands (or even millions) of parameters to be adapted in the training phase. Probably the most widely used deep learning structure is that of Convolutional Neural Networks (CNNs), proved to be superior in a variety of image analysis tasks. Other common structures include Long Short-Term Memory (LSTM) networks used for modelling sequences of data such as text, for example, and Generative Adversarial Networks (GANs), designed especially for generating new data based on certain features charac-

teristic to the training data set. A common property of the deep learning structures is that training of the models is performed based on data, i.e., no predefined and pre-calculated feature vector is needed. This, however, implies that extensive data sets are required for training the models and the operation principles of the models are usually not revealed.

In this study we present our recent work in designing a decision support platform for farmers. A central component of the platform is its analytical engine, involving machine learning models for various phenomena. Before presenting the general structure of the platform we present an overview on recently developed applications of deep learning in agriculture. We also present a case study on the development of crop yield prediction model using CNNs.

2. Applications of Deep Learning in Agriculture: Overview

The developments in DL algorithms, and importantly the deployment of numeric software tools to implement them, have resulted in a surge in their applications. Agriculture has been the domain of some such applications, indicated by a sharp increase in the number of publications applying DL methods to different areas of agriculture. In one of the earliest works, Kamilaris & Prenafeta-Boldú [5] review 47 published studies and recognize 16 topical areas. They further concentrate on CNN, a specific framework within DL, and review agriculture related studies using this methodology [6]. For the purpose of this study, we chose to focus on the literature considering crop production in open fields and related issues, thus excluding topics such as greenhouse farming, land-use classification, animal husbandry and fruit/orchard plantations (see Figure 2). This selection of scope was due to our ongoing work on crop yield monitoring of mainly wheat and barley fields in Finland.

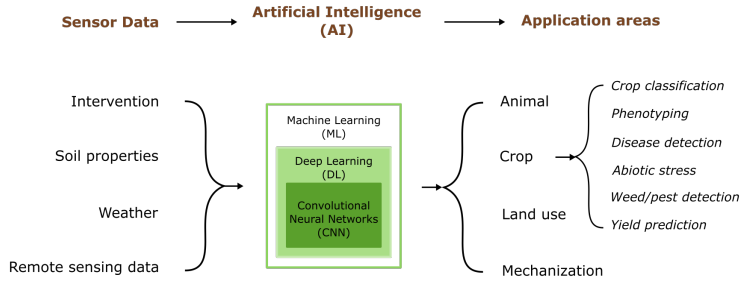

Figure 2. Application areas of DL in agriculture. In this study we focus on crop production, identifying six specific problems that can be targeted using DL models.

2.1. Crop recognition/classification

Crop recognition and classification using DL algorithms is generally relevant when the objective is to ascertain crop coverage over a large region (covering multitude of farms) based only on remote sensing images. The task can be to detect a single crop or a set of

crops. CNN based DL models have performed well in comparison to other ML methods, reaching very high classification accuracy ($> 85\%$) [7][8]. Most studies addressing this task use satellite data, but UAV imagery can also be used [9].

2.2. Phenotyping

Crop development can be assessed by quantifying the quality, structure or biomass productivity of the plants in a series of developmental stages. Ascertaining these phenological stages of plants can be important in precision agriculture for monitoring crop condition. This has implications for timing of harvest, pest control, yield prediction, farm monitoring and disaster warning. Various measures of performance can be used such as leaf counting, growth stage classification or plant maturity (age regression). Image based DL approaches have been shown to be superior to analyses based on hand crafted features [10][11]. In a recent article, Mochida et al. present an overview of various image based phenotyping studies that employ ML techniques [12].

2.3. Disease Detection

Disease, due to biotic stressors, of crops is a prime topic for testing the efficacy of DL methods in monitoring crop health. DL methods have show significant potential in improving the speed, accuracy and reliability in early detection of diseases [13]. Golhani et al. have presented an excellent review of neural network based approaches to disease detection using hyperspectral images [14]. Among the studies they review, a couple of CNN based studies performed especially well. Such studies tend to require higher resolution images and thus are most suitable for UAV based imagery. Though hyperspectral cameras are expensive currently, with falling costs they have the potential to be employed as an essential farm monitoring tool in the near future.

2.4. Abiotic Stress Detection

Abiotic stress is often unavoidable, especially in open-field cropping, and monitoring their expression in plants is important in mitigating their detrimental effect on crop productivity [13]. The stressors can be, for example, herbicide damage, water excess/deficiency, temperature extremes, nutrition deficiency. Using DL to detect and classify stress states, has resulted in superior performance in comparison to traditional regression methods.

2.5. Weed Detection

As with disease, weeds and pests can also reduce crop productivity significantly. This is essentially a task of identifying weeds and discriminating them from the crop by using detection/classification strategies. Early detection is of importance, which can be effectively accomplished using high resolution data able to capture the weeds at early stages of growth. Thus aerial and terrestrial autonomous vehicle based remote sensing systems are ideally suited for data collection. DL frameworks applied to UAV imagery have shown good results in accurately detecting and delineating specific weeds among crops [15][16]. However, this is a very challenging task and highly dependent on the specific context of the crop type and weed type. Visual similarity of the crop and weed or occlusion of the weeds in images can significantly complicate the analysis procedure.

2.6. Yield Prediction

All efforts in crop monitoring ultimately seek to improve crop productivity, i.e., yield. Earliest attempts at harnessing the potential of DL methods in predicting yield were made with encouraging results ($> 80\%$)[17]. Panda et al., used neural networks with multiple vegetation indices to predict corn yield with high accuracy ($83.5\% - 96\%$) [18]. Typically the output of the prediction model is in terms of yield classes (i.e. high, medium or low). Elavarasan et al. in their review of ML studies in yield prediction include studies with DL based yield prediction [19]. One of the interesting studies conducted by You et al (2017) used a combination of CNN and LSTM networks to predict soybean yield at regional level with very high accuracy [20]. Their method has the potential for scaling down to intra-field yield prediction.

3. Case Study: Prediction of Yield of Wheat and Barley Fields in Satakunta, Finland

DL models represent the data that they are trained on. As the growth of crops depends on climate and sunlight conditions, the variation of these conditions in time and space will potentially pose a challenge for a universal model. Thus there is a need for training models specific to regional conditions. Keeping this in mind, an effort was made to test the feasibility of using CNN models to predict wheat and barley yield grown in the Finnish continental subarctic climate.

3.1. Materials

Six fields, located in the Satakunta region of Finland near the city of Pori, were selected for this study. They vary in size, together accounting for 54.2 ha of land area. The data acquisition was conducted during the 2017 growing season. Image data were acquired using a UAV (Airinov Solo 3DR) with a multispectral camera (SEQUIOA, Parrot) mounted to it. Images were acquired in the early stage of crop growth, within 25% of the total thermal time of the respective crop variety. Pertinent details about the test fields are provided in Table 1. Crop yield data was collected in September 2017 using two sensor systems (Trimble CFX 750 and John Deere Greenstar 1) mounted to combine harvesters. Growth phase was determined by calculating the cumulative daily thermal time commencing from the date of sowing for each field. Thermal time for each day was calculated using Eq.(1), based on the daily mean temperature calculated at specific times ($t = \{02{:}00, 05{:}00, 08{:}00, 11{:}00, 14{:}00, 17{:}00, 20{:}00, 23{:}00\}$):

$$Th_t = \max\left(\left(\frac{1}{8}\sum T_t\right), 5\right) - 5 \tag{1}$$

3.2. Methods

The yield data from the harvester mounted sensors were contained in *shape* files (a file format for vector type geospatial data). The yield information is represented by polygons with an attribute describing the yield (in kg) collected over the area of the polygon.

Table 1. Details of crop fields and crop varieties in the 6 test fields. Thermal time for each crop variety is the total thermal time to crop maturity. The data to calculate the thermal time is taken from [21]. Sowing dates and imaging dates are used to calculate the growth phase as a fraction of the total thermal time for each particular crop variety.

Field #	**Size (ha)**	**Crop: (*Variety*)**	**Thermal time**	**Sowing date**	**Imaging date**	**Growth phase**
1	5.14	Barley:*Trekker*	979.7	16 May	8 Jun	15 %
2	2.97	Barley:*Trekker*	979.7	17 May	8 Jun	15 %
3	4.66	Barley:*Propino*	981.4	15 May	15 Jun	22 %
4	7.29	Barley:*Propino*	981.4	15 May	15 Jun	22 %
5	15.28	Barley:*Trekker*	979.7	18 May	1 Jun	10 %
6	18.86	Wheat:*KWS Solanus*	1065	13 May	15 Jun	21 %

These were converted to point data (polygon centroids) attributed with the yield density (kg/ha). This point data was then interpolated and rasterized to serve as the ground truth in training the DL model. The FarmWorks software tool was used in preprocessing the yield data.

The high resolution ($0.31 \times 0.31m$) images collected using the multispectral camera were compiled as mosaics using the Pix4D software tool and masked with the shape of respective fields. Two types of data sets were constituted from the measurements – 3-band RGB images and 1-band Normalized Difference Vegetation Index (NDVI) data.

A CNN model was constructed using the PyTorch [22] software library and refined through iterative tuning of relevant parameters such as: network depth (i.e., the number of convolutional layers of the CNN), the weights of the training algorithm, the hyperparameters of the training algorithm and the parameters of the regularization method. Additionally, three different image frame sizes ($10m, 20m$ and $40m$) were tested to determine the best image size to be fed to the CNN model. After all tests were performed, the best performance was observed with $40m \times 40m$ RGB image frames fed to a CNN network with 6 convolutional layers using the Adadelta training algorithm (learning rate = 0.008, past iterations' error adjustment coefficient = 0.58) with L2 regularization (weight decay = 0.001) and early stopping (patience = 50).

The CNN takes three $40m \times 40m$ image frames (1 per channel in RGB) and outputs a single density value (predicted yield). The resulting point data is georeferenced, representing the yield density predicted over the area of the image frame. In order to observe the capacity of the model to represent the spatial distribution of yield within a field, the point data was rasterized to visualize the predicted yield as a composite image of a single field.

3.3. Results

The ability of the CNN model to represent the yield distribution for each field is illustrated by the scatter plots in Figure 3. While the trend lines have similar slopes for each of the 6 fields, the data indicate a consistent pattern of overestimating low yields and underestimating high yields. In order to illustrate the prediction error relative to the magnitude of the yield, the mean absolute percentage error (MAPE) for each field is presented in (Figure 4). It can be seen that among the 6 fields the average percentage error is within 6% – 14%, with corresponding medians within 4% – 10%. The largest field (#6: 18.86 ha) was chosen to illustrate the ability of the model to follow the spatial yield

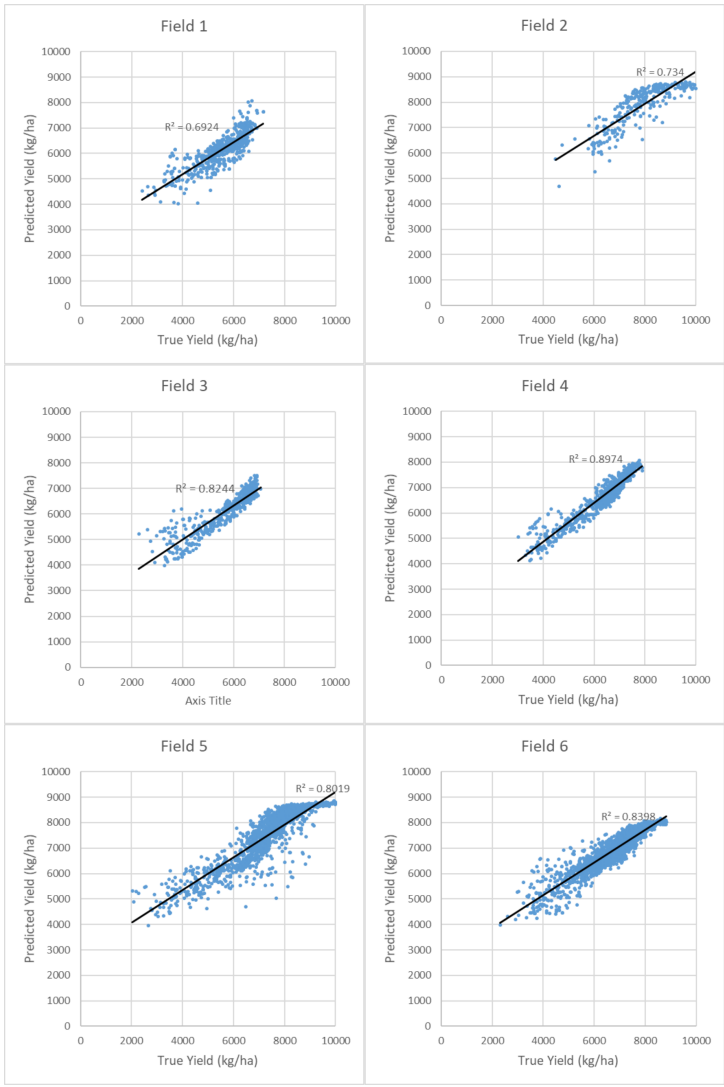

Figure 3. Correlations between the true and predicted yield for each of the 6 fields included in the study.

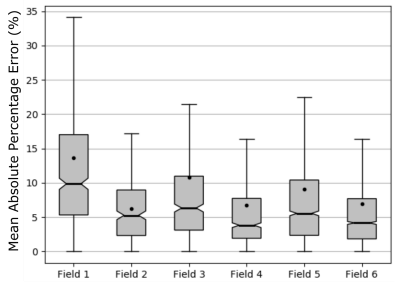

Figure 4. Boxplots of percentage error between true yield and predicted yield for each field.

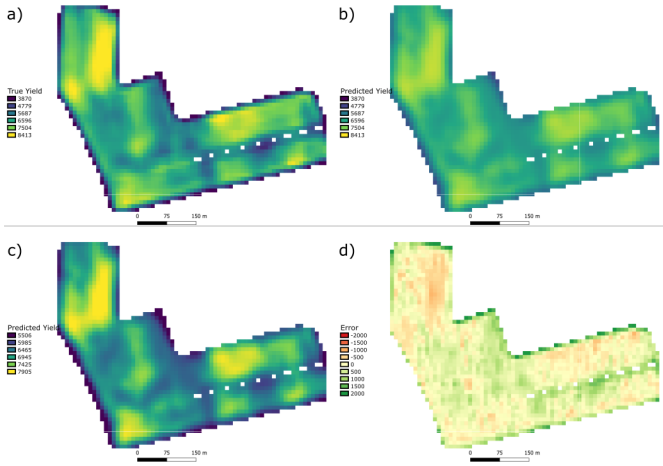

Figure 5. Visualisation of the true and predicted yield of Field 6. a) The spatial distribution of yield as recorded by the yield sensor on the combine harvester. b) Yield predicted by the CNN model. c) Predicted yield with colour-scale adjusted to min-max range. d) Error between predicted and true yield.

distribution within a field (Figure 5). The raster of the predicted yield when viewed as a colour map (Figure 5c) clearly illustrates the capability of the model to predict the spatial variations of the true yield. However, Figure 5b illustrates that the model is only capable of representing a limited range of values of the true yield. Figure 5d shows that the errors (calculated by subtracting true yield from predicted yield) are mostly in the high and low end of the range of values of true yield; thus the model over-predicts in low yield regions and under-predicts in high yield regions.

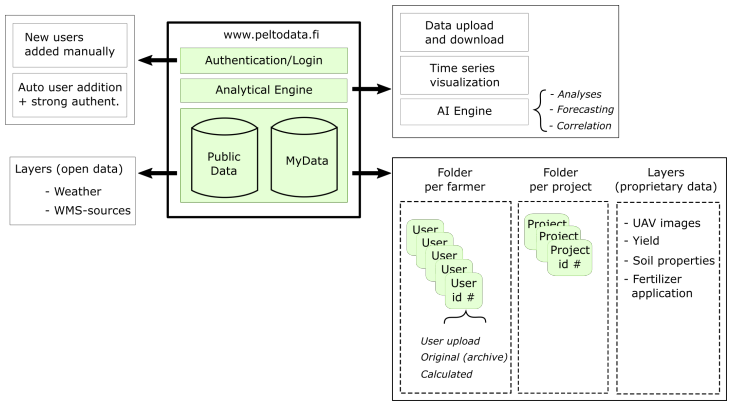

Figure 6. Structure of the Oskari based decision support platform for farmers.

4. Decision Support for Farmers: the Oskari Platform

A project was initiated with the goal of implementing a web based data repository as well as an analysis and decision support platform tailored to farmers' needs. The project explored various services and platforms and evaluated suitability based on their capability to handle access rights of farmers to their uploaded/transferred data on an individual basis. At the same time, emphasis was also placed on using open-source frameworks and contributing towards open data [23]. Consequently, mapping existing solutions revealed about 150 different platforms, though this search was not exhaustive. Overwhelming majority of the platforms were found to be either paid-for services and/or closed-source software and were therefore considered unsuitable. Among the few suitable platforms, Oskari was chosen. Oskari is an open-source (*www.oskari.org*; licence: MIT & EUPL) tool for web mapping applications using distributed spatial data infrastructure like Geoserver running as the back-end. Its front-end allows data management and custom visualisation based on an HTTP server and Java servlet extension.

The Oskari service for this project was implemented such that it can be accessed through the *peltodata.fi* domain. The architecture envisaged at this stage is illustrated in Figure 6. Through the web portal farmers can access their personal, authenticated accounts, upload data for visualization and call on AI based analytical tools for decision support. The technology to implement the analytical tools in the web based service environment has not yet been decided; the most promising options are the Shiny environment using the R language or the Python environment which has best support for DL models. The models implemented so far, including the CNN based yield prediction model, have run on a separate computer cluster.

The Data available to the farmers includes open source and proprietary data. Examples of open source data include weather data, satellite image data and land drainage maps, for example. Farm specific harvester yield maps, UAV based remote sensing image data (multispectral) and soil nutrient content maps are examples of proprietary data whose access is restricted and controlled by their owners.

5. Discussion and Conclusions

Application of ML and DL methods to agricultural (big) data has gained a lot of attention recently. The variety of problems addressed using these methods is wide, ranging from fruit counting to political decision making. In this paper we have focused on decision support for farmers cultivating open-field crops. Even in this restricted scope, there are various tasks that could be addressed by ML and DL as indicated in section 2.

As a case study, we have presented a CNN based yield prediction model, implemented and evaluated using UAV-acquired multispectral data from 6 crop fields in Satakunta, Finland. The results of the case study indicate that it can, with decent accuracy, model crop yield based on data acquired in the early phase of the growth season. Significantly, the model is capable of predicting within-field patterns of yield variation with good similarity to true yield. Training DL algorithms requires large amount of data. Kamilaris & Prenafeta-Boldú [5] lists some of the openly available datasets for training and possibly benchmarking the models. Also, training of the model and tuning of the model parameters is of high computational complexity and therefore cannot be performed on-line as a part of a decision support platform. Once trained, using the model for yield prediction is computationally relatively inexpensive. It remains to be studied how well the tuning of the algorithm can be generalized to the data from other areas and/or acquired in different years. Learning the effects of climate and other environmental conditions on the model efficiency is a long term research pursuit as data from different regions and weather conditions needs to be acquired and analyzed. Also, employing time sequences of data possibly using the LSTM DL networks would be a promising research area.

The presented case study revealed some limitations of the CNN model in yield prediction. The model underestimated/overestimated the yield in the regions of high/low yield values, respectively. The reason for this kind of behavior needs to be investigated. Another limitation is related to yield data pre-processing. In some cases the polygons of yield data overlap causing errors in yield density maps. This limitation will be addressed in the future by more careful pre-processing flow.

In its current form, the peltodata.fi portal aims to provide a few key services to the local farming community. Currently, the farmers can explore the harvester yield distribution, soil properties maps, UAV multispectral images among other open source maps. The farmers can also avail of the analyses such as predicted yield. The collaborating farmers will be involved in the development of the service to serve their needs most appropriately. With regards to the platform, Oskari has been adopted by several municipalities and government agencies in Finland, thereby forming a considerable user base. This has resulted in a core group of Oskari developers monitoring the trends and customer requirements to develop appropriate solutions. In addition, there are a lot of interesting data interfaces available, for example, from the Spatineo Director. (https://directory.spatineo.com/). An important aspect to be implemented in the future is the capability of data trading or download/port to smart devices for intervention (e.g. application of fertilizer, weedicide and irrigation).

References

[1] R. Rupnik, M. Kukar, P. Vraar, D. Koir, D. Pevec, and Z. Bosni, "AgroDSS: A decision support system for agriculture and farming," *Computers and Electronics in Agriculture*, 2018.

[2] D. I. Patrício and R. Rieder, "Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review," *Computers and Electronics in Agriculture*, vol. 153, pp. 69–81, 2018.

[3] T. U. Rehman, M. S. Mahmud, Y. K. Chang, J. Jin, and J. Shin, "Current and future applications of statistical machine learning algorithms for agricultural machine vision systems," *Computers and Electronics in Agriculture*, vol. 156, pp. 585–605, jan 2019.

[4] A. Chlingaryan, S. Sukkarieh, and B. Whelan, "Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review," *Computers and Electronics in Agriculture*, vol. 151, no. November 2017, pp. 61–69, 2018.

[5] A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," *Computers and Electronics in Agriculture*, vol. 147, pp. 70–90, 2018.

[6] ——, "A review of the use of convolutional neural networks in agriculture," *Journal of Agricultural Science*, no. June, pp. 1–11, 2018.

[7] M. Dyrmann, H. Karstoft, and H. S. Midtiby, "Plant species classification using deep convolutional neural network," *Biosystems Engineering*, vol. 151, pp. 72–80, nov 2016.

[8] S. Ji, C. Zhang, A. Xu, Y. Shi, and Y. Duan, "3d convolutional neural networks for crop classification with multi-temporal remote sensing images," *Remote Sensing*, vol. 10, no. 2, p. 75, jan 2018.

[9] J. Rebetez, H. F. Satizábal, M. Mota, D. Noll, L. Büchi, M. Wendling, B. Cannelle, A. Pérez-Uribe, and S. Burgos, "Augmenting a convolutional neural network with local histograms - a case study in crop classification from high-resolution uav imagery," in *24th European Symposium on Artificial Neural Networks, ESANN 2016, Bruges, Belgium, April 27-29, 2016*, 2016.

[10] H. Yalcin, "Plant phenology recognition using deep learning: Deep-pheno," *2017 6th International Conference on Agro-Geoinformatics*, pp. 1–5, 2017.

[11] J. R. Ubbens and I. Stavness, "Deep plant phenomics: A deep learning platform for complex plant phenotyping tasks," *Frontiers in Plant Science*, vol. 8, p. 1190, 2017.

[12] K. Mochida, S. Koda, K. Inoue, T. Hirayama, S. Tanaka, R. Nishii, and F. Melgani, "Computer vision-based phenotyping for improvement of plant productivity: a machine learning perspective," *GigaScience*, vol. 8, no. 1, jan 2019.

[13] A. K. Singh, B. Ganapathysubramanian, S. Sarkar, and A. Singh, "Deep learning for plant stress phenotyping: Trends and future perspectives," *Trends in Plant Science*, vol. 23, no. 10, pp. 883–898, oct 2018.

[14] K. Golhani, S. K. Balasundram, G. Vadamalai, and B. Pradhan, "A review of neural networks in plant disease detection using hyperspectral data," *Information Processing in Agriculture*, vol. 5, no. 3, pp. 354–371, sep 2018.

[15] M. D. Bah, A. Hafiane, and R. Canals, "Deep learning with unsupervised data labeling for weed detection in line crops in UAV images," *Remote Sensing*, vol. 10, no. 11, 2018.

[16] H. Huang, J. Deng, Y. Lan, A. Yang, X. Deng, and L. Zhang, "A fully convolutional network for weed mapping of unmanned aerial vehicle (UAV) imagery," *PLOS ONE*, vol. 13, no. 4, pp. 1–19, 04 2018.

[17] I. C. Davis and G. G. Wilkinson, "Crop yield prediction using multipolarization radar and multitemporal visible/infrared imagery," *Proc.SPIE*, vol. 6359, pp. 6359 – 6359 – 12, 2006.

[18] S. S. Panda, D. P. Ames, and S. Panigrahi, "Application of vegetation indices for agricultural crop yield prediction using neural network techniques," *Remote Sensing*, vol. 2, no. 3, pp. 673–696, 2010.

[19] D. Elavarasan, D. R. Vincent, V. Sharma, A. Y. Zomaya, and K. Srinivasan, "Forecasting yield by integrating agrarian factors and machine learning models: A survey," *Computers and Electronics in Agriculture*, vol. 155, pp. 257–282, dec 2018.

[20] J. You, X. Li, M. Low, D. Lobell, and S. Ermon, "Deep Gaussian Process for Crop Yield Prediction Based on Remote Sensing Data," *31th AAAI Conference on Artificial Intelligence*, pp. 4559–4565, 2017.

[21] A. Laine, M. Högnäsbacka, M. Niskanen, K. Ohralahti, L. Jauhiainen, J. Kaseva, and H. Nikander, "Virallisten lajikekokeiden tulokset 2009-2016," p. 262, 2017.

[22] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *NIPS-W*, 2017.

[23] P. Linna, T. Mkinen, and K. Yrjnkoski, "Open data based value networks: Finnish examples of public events and agriculture," in *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, May 2017, pp. 1448–1453.

PUBLICATION

III

Assessment of Cloud Cover in Sentinel-2 Data Using Random Forest Classifier

P. Nevavuori, T. Lipping, N. Narra and P. Linna

*IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*2020, 4661–4664

DOI: 10.1109/IGARSS39084.2020.9323683

ASSESSMENT OF CLOUD COVER IN SENTINEL-2 DATA USING RANDOM FOREST CLASSIFIER

P. Nevavuori

Mtech Digital Solutions Oy
Vantaa, Finland

T. Lipping, N. Narra, P. Linna

Tampere University
Tampere, Finland

ABSTRACT

In this paper, a novel cloud coverage assessment method for the Sentinel-2 data is presented. The method is based on the Random Forest classifier and the target values used in the training process are obtained by comparing the NDVI indexes calculated from the satellite and the UAV data. The developed method is shown to outperform the Sentinel Cloud Probability Mask (CLDPRB) and Scene Classification (SCL) data layers in detecting cloudy areas.

***Index Terms*—** Cloud coverage, Random Forest classifier, Crop monitoring

1. INTRODUCTION

Data from the Sentinel satellites are intensively used for various applications such as land use and vegetation mapping or crop monitoring, for example. Depending on climate conditions in the region of interest, the main obstacle in using the data for practical monitoring purposes may be cloud coverage. This is especially restricting if the data should be acquired from a narrow time window corresponding, for example, to a certain growth phase of crops. The problem could be alleviated by more accurate and higher resolution cloud coverage assessment compared to that available by the product of the Sentinel data.

Currently the cloud mask of the Sentinel data is available in the form of the Level 1C product containing vector layers of dense and cirrus clouds. Also, the percentage of cloudy pixels (dense and cirrus) in the mask are provided. The Level 2A product further processes the Level 1C data to obtain the Scene Classification layer with cloud and cirrus probability values at 60 m spatial resolution. Calouzzi et.al. [1] assessed these products concluding that caution has to be taken when using the provided cloud masks and improved cloud detection algorithms are welcome. Recently, Baetens et.al. [2] compared three cloud mask calculation algorithms: MAJA (used in the Level 2A product), Sen2Cor (used by ESA) and FMask (used by USGS), using their Active Learning Cloud Detection (ALCD) method for producing reference cloud masks. Classification accuracy of about 90 % was obtained by MAJA and FMask while SenCor gave 84 % accuracy.

In this paper we train the random forest classifier to assess cloud cover in Sentinel-2 data. Our primary usage of the data is crop monitoring and yield prediction for decision support for farmers. Therefore, the classifier is trained using data acquired from crop fields by UAVs: as UAVs fly below the clouds and the data they produce is not affected by cloud cover (if properly corrected for changes in irradiance), the difference between the UAV and Sentinel data can be used as ground truth for cloud cover.

2. DATA

2.1. Drone Images

For cloudless multispectral ground truth data, ten crop fields were selected for imaging in the vicinity of Pori, Finland (61°29'N, 21°48'E) and were imaged as a part of the MIKA DATA project [3, 4]. The total area of the selected fields was approximately 93 ha. Half of the fields had wheat (*Zebra/Mistral*), three had barley (*Harbringer/RGT Planet*) and two remaining had oats (*Ringsaker*) as the cultivated crop. The fields were imaged during the growing season for years 2018 and 2019 from the time of sowing to the time of harvest. All fields were imaged weekly. Due to varying weather conditions and the proximity of an airport, the temporal allocation of imaging flights to within a fixed daily time range was not possible. The images were thus taken during day time.

The fields were imaged with two distinct drones, using 3DR Solo for the year 2018 and Parrot Disco-Pro AG for 2019. The drones were equipped with similar Parrot Sequioa multispectral cameras. Distinct images were collated for each field to build a complete image of a field using the Pix4D software. During the process of building the image mosaics, the band data were also automatically normalized in terms of radiance utilizing the information provided by the multispectral camera's irradiance sensor. Using the red and near-infrared (NIR) channels, the normalized difference vegetation index (NDVI) was then calculated from each field's multi-band mosaic. To use the drone data in conjunction with the Sentinel-2 data, the collated drone images were downsampled to match the highest resolution available in Sentinel-2 images, 10 m/px. The downsampling was done using `cubicspline` interpolation algorithm in the `gdalwarp` utility. Lastly, the images for each field were cut to proper shape with field block border data provided by Ruokavirasto (*Finnish Food Authority*) [5]. This resulted in a total of 288 distinct crop field images. The field-wise sizes, crop varieties, yearly image counts and average valid pixel counts per image are given in Table 1

The use of NDVI images calculated from drone data is discussed in Sec. 2.3. Next we will discuss the acquisition and processing of the Sentinel-2 satellite data.

2.2. Sentinel-2 Data

Sentinel-2 satellite images were selected as the source data for the study. The data provided by the dual satellite system are widely used in agriculture and is freely available. The satellite images processed to the Sentinel product Level-2A [6] were downloaded from Copernicus Open Access Hub [7]. The satellite data products were downloaded for the growing seasons of 2018 and 2019.

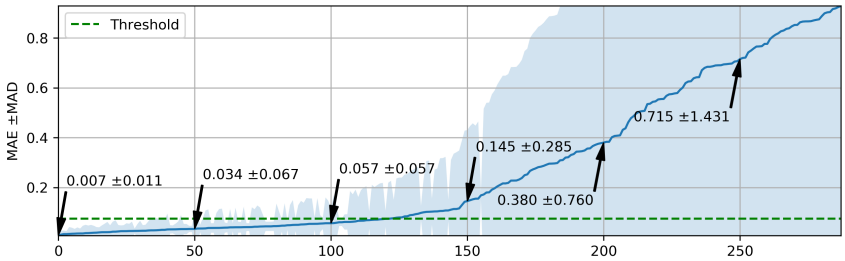

Fig. 1. The mean absolute errors (MAE) and mean absolute deviations (MAD) of week-aligned NDVI pairs in ascending order. The statistics are calculated over the pixels in the paired Sentinel-2 and drone NDVI images.

Table 1. Sizes, crops, image counts and average pixel counts of fields selected for drone imaging.

Field	Size, ha	Crop	Image Counts 2018	Image Counts 2019	Avg. Valid Px Per Image
1	11.08	Wheat	13	16	1065.5
2	8.24	Wheat	15	14	759.1
3	11.77	Wheat	13	16	1120.9
4	11.12	Wheat	15	16	1051.9
5	7.59	Wheat	15	16	705.2
6	7.61	Oats	12	15	739.8
7	7.24	Oats	13	15	681.9
8	7.77	Barley	13	15	1016.6
9	13.05	Barley	12	16	1251.3
10	7.95	Barley	12	16	715.5

The satellite data were selected with no limits on the estimated cloud coverage. The goal was to be able to find week-matching pairs for the drone data. The data was used as the training data for which information about the cloudless ground truth was available via drone data. The gathered data spanned initially the growing seasons of years 2018 and 2019. Part of the downloaded data was omitted during the process of week-matching Sentinel-2 data to Drone data. The satellite image data were cut to shape using field block borders already utilized with the drone data to ultimately generate image pairs of drone and satellite data aligned both temporally and geographically for distinct fields.

2.3. Target Data

Supervised machine learning requires the existence of *a priori* labeled data, the ground truth. With the aim of estimating cloud coverage in Sentinel-2 data in the spatial scale of crop fields, NDVI images gathered with drones at the altitudes well below clouds are considered as cloudless ground truth. This consideration is in relation to satellites flying at atmospheric altitudes. Comparing absolute values across bands for two different sensors and imaging platforms has proven to be difficult, as the data would require scaling to an unkown global maximum for Sentinel-2. However, the use of NDVI alleviates this problem by providing normalized and thus comparable data between distinct imaging systems.

Target data needs thus to be generated using the week-aligned NDVI data from both sources, the drone and the Sentinel-2 systems. Each spatially and temporally aligned satellite and drone NDVI image pair is compared pixel by pixel to determine whether the images are similar on the level of distinct pixels. A pixel correponds to an area of 10 $\times$ 10 meters. The similarity for a single pixel-corresponding area is determined by

$$sim_{(s,d)} = \begin{cases} 1, & |s-d| \leq threshold \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where s and d are spatially and temporally aligned pixels for a field from the satellite and drone sources respectively. The mean absolute errors (MAEs) of all week-aligned image pairs are depicted in Fig. 1. The determination of the threshold is discussed next.

To determine a proper absolute NDVI difference threshold for labeling Sentinel-2 pixels either similar or dissimilar to the drone pixels (see Eq. 1), the two data sources were compared using the Student t-test. The test was applied over the pixels in the images to compare whether the NDVI values in the images were statistically similar or not. A total of 15 statistically similar ($p = 0.01$) week-aligned image pairs were found. It is to be noted though, that the number of image pairs having MAE in close proximity to the similarity threshold was higher than just 15 (see Fig. 1).

The statistically similar data (15 image pairs) were then used to empirically determine the proper threshold for classifying NDVI differences in terms of pixel-wise similarity. The tested thresholds were selected from the proximity of upper end of the MAE for the statistically similar data samples as shown in Table 2.

Table 2. NDVI difference metrics for similar image pairs

Image pairs	15
Avg. Diff.	0.001 ±0.046
MAE	0.026 ±0.022
MSE	0.003 ±0.010
RMSE	0.046 ±0.092

In more general terms, the task of determining the threshold for labeling is a task of balancing between (1) capturing as much similarities while (2) still excluding as many dissimilarities as possible. To elaborate, labeling every pixel in the statistically similar images as similar would require increasing the absolute NDVI threshold to levels possibly having some pixels incorrectly labeled as similar. The ratios of pixels labelled as similar for each similar image pair with different thresholds is given in Table 3. In combination with visual

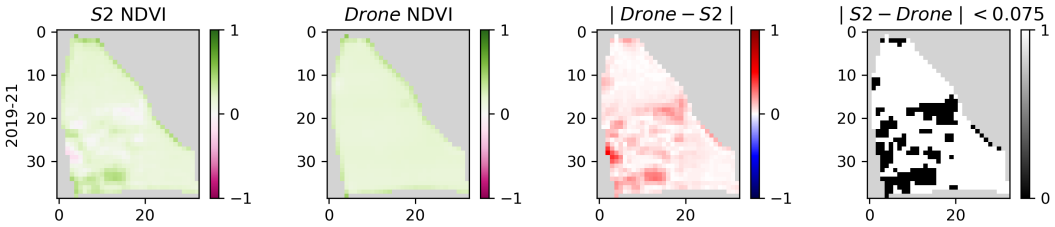

Fig. 2. A visualization of a single week-aligned Sentinel-2 and drone NDVI image pair with the absolute difference and the similarity map. The first two figures depict the NDVI maps from corresponding sources. The third figure shows the absolute difference between the aligned Sentinel-2 and drone NDVI values. The fourth figure shows the thresholded absolute difference, indicating areas where the NDVI images are similar enough.

evaluation, a threshold of 0.075 absolute NDVI difference was selected. A single image pair with the calculated similarity map is shown in Fig. 2.

Table 3. Similarity ratios with various thresholds

Threshold	Similarity
0.025	89.13%
0.050	94.40%
0.075	96.14%
0.100	97.13%

2.4. Building the Modeling Data Sets

After the generation of field and week specific similarity label maps, the data required only minor preprocessing. As the Sentinel-2 data products are delivered as separate files for distinct bands and layers, the satellite data were merged to construct multi-band images instead of multiple images of distinct bands. The following Sentinel-2 data were merged:

- *Sensor bands*: 1 to 8, 8A, 9, 11 and 12
- *Level-2A layers*: AOT, SCL, TCI, WVP and CLDPRB

The separately calculated NDVI data were also merged in conjunction with the alpha-channel generated during the processing of the data. As per machine learning best-pactices, the categorical values from the scene classification layer (SCL) needed to be separated to distinct binary raster layers according to the SCL classification labels, which is also known as transforming a multi-class representation to class-wise one-hot representation [8].

Thus, the final processed input data constituted 30 distinct layers of data for each pixel. The dataset was then created by extracting multi-band Sentinel-2 pixels as input samples and their spatially and temporally corresponding binary similarity label map pixels as target values. In other words, a single input sample was a $[1 \times 30]$ and its corresponding target sample a binary-valued $[1 \times 1]$ vector. A total of 381972 input-target samples (pixels) were extracted from the source data. The samples were then shuffled and split into training and test data sets with 190986 and 63661 samples, correspondingly. No scaling was applied due to the selected decision tree based model.

3. MODEL

Data based modeling with machine learning methods is in practice a tradeoff between model explainability and increased performance. While training an accurate model for classifying distinct Sentinel-2 pixels as similar or dissimilar to the cloudless ground truth data from drones is the primary goal while the explainability was deemed as an important objective to pursue as well. This is why an ensemble model called Random Forest from the decision tree algorithm family was selected. The ensemble model is able to model non-linear relationships, work with unscaled data and provide easily understandable explanations of decisions' causes [9]. The model implementation was part of the Python's `scikit-learn` framework [10].

Table 4. The confusion matrix of similarity label predictions.

Pred/True	**0**	**1**
0	TP 23237	FP 2580
1	FN 1807	TN 36037

4. RESULTS

The model was allowed to train 500 sub-trees, varying the tree structure and features used for each tree, using the training data set only. The performance of the model was then evaluated with the hold out test data set. The confusion matrix of model predictions against true labels is shown in Table 4. The precision of the model is

$$PPV = \frac{TP}{TP + FP} = 0.900, \tag{2}$$

where PPV stands for positive prediction value. The model's true positive rate, i.e., recall, is then

$$TPR = \frac{TP}{TP + FN} = 0.923. \tag{3}$$

The F_1-score, a statistical test accuracy measure for binary classification analysis is then calculated using Eqs. 2 and 3 by

$$F_1 = 2 * \frac{PPV * TPR}{PPV + TPR} = 0.911. \tag{4}$$

Another interesting metric is the negative prediction value

$$NPV = \frac{TN}{TN + FN} = 0.952, \tag{5}$$

which shows the model's precision in predicting dissimilarities. In conjunction with test data set result analysis, the model was also evaluated with distinct images from the original source data.

Due to Sentinel-2 satellite data being sensitive to changes and disturbances in atmospehric conditions, the cloud estimation information from the scene classification layer (SCL) and cloud probablity mask (CLDPRB) calculated in the Level-2A processing of the Sentinel-2 data can not be taken as definitive truth. They, however, form a proper baseline to which compare the trained model's performance against.

The model predictions are based on the similarities of Sentinel-2 and drone NDVI images, i.e. label 1 indicates predicted similarity. Taking a mean of a set of predicted values describes the mean predicted similarity for that set. The two cloudiness estimation masks in the Sentinel-2 data product are formulated differently.

As the name indicates, the CLDPRB mask contains pixel-wise probability values for the estimated degree of cloud coverage. The model-equivalent similarity measure would thus be

$$\text{CLDPRB}_{\text{SIM}} = 1 - \text{CLDPRB}, \tag{6}$$

where larger values imply increased degree of estimated similarity.

On the other hand, the SCL layer contains pixel-wise labels, with some labels indicating cloudiness (see [6]). To gain information about the SCL layer's model-equivalent similarity measure, the cloud-related label ratio

$$p_{cl} = \frac{\text{count}(\text{SCL}_{cl})}{\text{count}(\text{SCL})} \tag{7}$$

is first counted with the cl being a set of cloud-related class labels. The inverse

$$\text{SCL}_{\text{SIM}} = 1 - p_{cl} \tag{8}$$

can then be seen as the implied cloudless ratio for a set of samples. The comparison of sample-wise similarity estimations between the trained model and Sentinel-2 data products are given in Table 5. The estimates are given both for when the true target value was 0 (satellite differed from drone) and when it was 1 (satellite similar to drone).

Table 5. Similarity estimates with hold out test data.

	$y = 0$			$y = 1$		
	Mean	Std	Median	Mean	Std	Median
Model	**0.07**	0.25	0.00	0.93	0.26	1.00
$\text{CLDPRB}_{\text{SIM}}$	0.45	0.45	0.26	**0.97**	0.14	1.00
SCL_{SIM}	0.28	0.45	0.00	0.95	0.22	1.00
Samples		38617			25044	

5. DISCUSSION AND CONCLUSIONS

Our study indicates that the Random Forest model outperforms the Sentinel-2 CLDPRB and SCL data layers in detecting cloudy areas ($y = 0$). For non-cloudy areas the detection accuracy was slightly higher for the Sentinel products (see Table 5). Several issues should be considered, however, when comparing these results. Firstly, when training the Random Forest classifier, the thresholded absolute difference between the Sentinel-2 and drone data was used as the ground truth. While it can be argued that the main cause of this difference is cloudiness, there may also be other factors involved such as shadows or differences in irradiance. The satellite and drone imagery were not necessarily acquired during the same time of the day or same day of the week, although best time-matching pairs were looked for when selecting the data. In some cases a couple of days may cause significant changes in the crop development. Another limitation comes from using the NDVI data layers for ground truth assessment. While the NDVI index contains significant information for vegetation monitoring and is probably a good choice when assessing cloud cover in crop fields, its use reduces the generalizability of the results to other land cover types.

Despite the mentioned limitations, the developed method was found to improve the usability of Sentinel data in crop monitoring. By visual inspection it was observed that in many cases when the Sentinel-2 products indicated the whole crop field to be cloud-covered, there were still significant areas of almost clear skies. The proposed algorithm proved capable in detecting these areas with considerable accuracy.

6. REFERENCES

[1] Rosa Coluzzi, Vito Imbrenda, Lanfredi Maria, and Simoniello Tiziana, "A first assessment of the sentinel-2 level 1-c cloud mask product to support informed surface analyses," *Remote Sensing of Environment*, vol. 217, pp. 426–443, 09 2018.

[2] Louis Baetens, Camille Desjardins, and Olivier Hagolle, "Validation of copernicus sentinel-2 cloud masks obtained from maja, sen2cor, and fmask processors using reference cloud masks generated with a supervised active learning procedure," *Remote Sensing*, vol. 11, 02 2019.

[3] Petteri Nevavuori, Nathaniel Narra, and Tarmo Lipping, "Crop yield prediction with deep convolutional neural networks," *Computers and Electronics in Agriculture*, vol. 163, no. June, pp. 104859, 2019.

[4] Nathaniel Narra, Petteri Nevavuori, Petri Linna, and Tarmo Lipping, "A Data Driven Approach to Decision Support in Farming," in *Information Modelling and Knowledge Bases XXXI*, Ajantha Dahanayake, Janne Huiskonen, Yasushi Kiyoki, Bernhard Thalheim, Hannu Jaakkola, and Naofumi Yoshida, Eds., vol. 321, pp. 175 – 185. IOS Press, 2020.

[5] Ruokavirasto, "Peltolohkorekisteri," .

[6] ESA, "Level-2A Algorithm - Sentinel-2 MSI Technical Guide - Sentinel Online," .

[7] ESA, "Open Access Hub," .

[8] Danfeng Hong, Naoto Yokoya, Nan Ge, Jocelyn Chanussot, and Xiao Xiang Zhu, "Learnable manifold alignment (LeMA): A semi-supervised cross-modality learning framework for land cover and land use classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 147, pp. 193–205, jan 2019.

[9] Peter Flach, "Machine Learning: The Art and Science of Algorithms that Make Sense of Data," p. 409, 2012.

[10] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake Vanderplas, Arnaud Joly, Brian Holt, and Gaël Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," 2013.

PUBLICATION
IV

Crop Yield Prediction Using Multitemporal UAV Data and Spatio-Temporal Deep Learning Models

P. Nevavuori, N. Narra, P. Linna and T. Lipping

Remote Sensing 12.23 (2020)
DOI: 10.3390/rs12234000

remote sensing

MDPI

Article

Crop Yield Prediction Using Multitemporal UAV Data and Spatio-Temporal Deep Learning Models

Petteri Nevavuori [1,*], Nathaniel Narra [2], Petri Linna [2] and Tarmo Lipping [2]

[1] Mtech Digital Solutions Oy, 01301 Vantaa, Finland
[2] Faculty of Information Technology and Communication Sciences, Tampere University, 33014 Tampere, Finland; nathaniel.narra@tuni.fi (N.N.); petri.linna@tuni.fi (P.L.); tarmo.lipping@tuni.fi (T.L.)
* Correspondence: petteri.nevavuori@mtech.fi

Received: 4 November 2020; Accepted: 4 December 2020; Published: 7 December 2020

Abstract: Unmanned aerial vehicle (UAV) based remote sensing is gaining momentum worldwide in a variety of agricultural and environmental monitoring and modelling applications. At the same time, the increasing availability of yield monitoring devices in harvesters enables input-target mapping of in-season RGB and crop yield data in a resolution otherwise unattainable by openly availabe satellite sensor systems. Using time series UAV RGB and weather data collected from nine crop fields in Pori, Finland, we evaluated the feasibility of spatio-temporal deep learning architectures in crop yield time series modelling and prediction with RGB time series data. Using Convolutional Neural Networks (CNN) and Long-Short Term Memory (LSTM) networks as spatial and temporal base architectures, we developed and trained CNN-LSTM, convolutional LSTM and 3D-CNN architectures with full 15 week image frame sequences from the whole growing season of 2018. The best performing architecture, the 3D-CNN, was then evaluated with several shorter frame sequence configurations from the beginning of the season. With 3D-CNN, we were able to achieve 218.9 kg/ha mean absolute error (MAE) and 5.51% mean absolute percentage error (MAPE) performance with full length sequences. The best shorter length sequence performance with the same model was 292.8 kg/ha MAE and 7.17% MAPE with four weekly frames from the beginning of the season.

Keywords: crop yield prediction; UAV; spatio-temporal modelling; time series; deep learning; cnn-lstm; convolutional lstm; 3d-cnn

1. Introduction

The abundance of modern sensor and communication technology already present in production facilities and similar highly connected environments has also seeped into the realm of agriculture. Various globally, nationally and locally available data generating remote sensing systems are in place, providing relevant data for optimizing several agricultural outputs. On the global and national scale, satellite systems (Sentinel and Landsat missions, for example) provide temporally relevant spatial data about visible land surfaces. Nationally, there are various instruments in place to both track and predict climatological variables. Data for fields and relevant other entities is also gathered on a per-field basis by agricultural expert institutions. While satellite data is meaningful when monitoring large fields, smaller fields common to European countries, as an example, require higher resolution data. Human-operated unmanned aerial vehicles (UAV) play a key role in high resolution remote sensing in fields, that otherwise would wholly be covered by just tens or, at most, a couple hundreds of open-access satellite spatial data resolution pixels (10 $\times$ 10 m/px for Sentinel-2, for example). Also, utilizing modern sensors and global navigation satellite system (GNSS) tracking with agricultural machinery further adds detail to the pool of generated data. Modern data-based modeling techniques

also benefit from increased resolution of spatial data, as they are able to better learn the relevant features in performing a given task, e.g., intra-field yield prediction. Feeding this data to automated processing and decision making pipelines is a vital part of Smart Farming enabling Decision Support Systems [1].

In [2] we performed crop yield estimation with point-in-time spatial data, point-in-time estimation being contrary to time series regression. In this study we examined the effect of time, as an additional feature, on intra-field yield prediction. Especially, we focused on the capabilities of deep learning time series models utilizing UAV remote sensing time series data as their inputs. Firstly, we wanted to see if we could surpass the performance of the point-in-time model [2] by using spatio-temporal deep learning model architectures. Secondly, we wanted to see which spatio-temporal architecture would perform better in the same task. Lastly, we perform comparative evaluation of different sequence configurations to perform actionable crop yield predictions with data collected at the beginning of the growing season.

We utilize the properties of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks to perform spatio-temporal modelling. The CNN is briefly discussed in Section 2.3.1 and the LSTM in Section 2.3.2. The model architectures that we implemented are the following:

CNN-LSTM. CNN and LSTM networks can be utilized as separate but sequentially connected feature extractors, where the CNN first ingests spatial data and then provides extracted spatial feature data to the LSTM [3]. These models are hereafter referred to as CNN-LSTM are discussed in Section 2.3.3.

ConvLSTM. Convolutional learning properties can also be utilized differently. Models that utilize convolutional layers embedded into the LSTM architecture in a manner that eliminates the necessity to use spatial feature extraction prior to feeding the data to a sequential model are hereafter referred to as ConvLSTMs [4] and discussed in Section 2.3.4.

3D-CNN. A fully convolutional architecture can also be used to model sequential data. It is done by applying the convolution in the direction of time (or depth) in addition to width and height dimensions of spatial data [5]. Fully convolutional models utilizing the third dimensions for convolution are hereafter referred to as 3D-CNNs and discussed in Section 2.3.5.

1.1. Related Work

Regarding data similar or related to our study, recent crop-related studies utilize satellite-based data at scales larger than single fields. Ref. [6] performed county-scale soybean yield prediction with a CNN-LSTM architecture in parts of US. In addition to US national weather and yield data, they used time series satellite data from the MODIS satellite system. The data resolution was from 500×500 m/px to 1×1 km/px. Ref. [7] performed crop type segmentation of small holder farms in Germany, Ghana and South Sudan using data from Sentinel S1, S2 (10×10 m/px) and PlanetScope (3×3 m/px) satellite systems and time of year as an additional feature. Ref. [8] performed crop type mapping with a 30×30 m/px crop-specific annual land cover data combined from various satellite data sources for the area of Nebraska, US. Ref. [9] classified crop varieties from satellite time series data frames collected by the Chinese Gaofen missions with data resolutions from 4×4 m/px up to 15×15 m/px.

In the broader context of time series modelling with remote sensing data, several recent studies utilize spatio-temporal model architectures. The US county-scale soybean yield prediction by [6] was performed using a CNN-LSTM composite architecture, where a sequence of input frames was transformed into vectors of spatial features and then fed to an LSTM. Ref. [7] employed both a CNN-LSTM and a 3D-CNN model to perform crop type classification in Ghana and South Sudan, feeding multi-layer remote sensing time series data frames to the models. Ref. [8] built and trained a bidirectional ConvLSTM to predict crop maps from satellite data at the early stages of the growing season in Nebraska. While their main contribution was to affirm the feasibility of such model, they also

employed their model in a CNN-LSTM setting, using pre-trained CNN called VGG11 [10] to extract spatial features from sequences of past crop map images and then feeding these sets of spatial-like features further to the ConvLSTM. Ref. [9] used a 3D-CNN architecture in their crop type mapping study, feeding sequences of RGB image data from distinct areas to the network thus having the model learn both spatial and temporal features from the data. Ref. [11] built and trained a bidirectional ConvLSTM to automatically extract meaningful features from hyperspectral data consisting of several hundred bands for land cover pixel classification. They utilized the sequential modeling power of the ConvLSTM for feature extraction from individual images, feeding distinct bands to the model as if they were items in a sequence. Among other models, they also trained a 3D-CNN for the task to compare performance. Ref. [12] utilized a Gated Recurrent Unit (GRU) in building the convolutional recurrent model, i.e., ConvLSTM-like architecture. Their domain of application was in utilizing novel machine learning methodologies in performing land cover classification from satellite data. They employed their ConvLSTM-like model in parallel with a CNN to produce pixel-level land cover classification and report improved performance against widely utilized decision tree models for similar task. Ref. [13] employed the ConvLSTM in a encoder-decoder architecture to predict maize growth stage progression using several meteorological features using nationally collected meteorological data in China. The ConvLSTM was used as a feature extracting encoder while the decoder was an LSTM producing a desired output sequence. They modified the ConvLSTM to perform 1D convolutions on row-like data. A CNN-LSTM was also trained for comparative purposes. Ref. [14] compared the performance of 3D-CNNs against other deep learning architectures in the task of performing scene classification based on hyperspectral images. While the domain of their application is that of spectral-spatial and not spatio-temporal, they report 3D-CNN performing the best among other tested model compositions. Ref. [15] employ a wide array of CNN configurations to perform yield estimation using soil and nutrient information available pre-season arranged as spatial data. Most relevant to our study is their utilization of a 3D-CNN architecture which they use to ingest point-in-time data and learn salient features across varying input data rasters to estimate the crop yield.

1.2. Contribution

In contrast to studies performed at larger spatial scales, the main contribution of our study is to perform time series based intra-field yield prediction with multi-temporal data collected during the growing season with UAVs. In the context of using of remote sensing data in performing data-based modeling to aid in Smart Farming, we perform time series regression with remote sensing data, which is both collectable using commercially available UAVs and has spatial resolutions well below 1×1 m/px. We also use meteorological information, cumulative temperatures, to inform the models about change between weekly data. Our study builds on [2], introducing an extra variable to the modeling task, time, to see whether using time series data is more beneficial than using point-in-time data only. We develop, train and compare several spatio-temporal models to determine the most suitable model for intra-field yield modelling from a selection of models already utilized in the context of spatio-temporal modelling with remote sensing data. To also see if the spatio-temporal models can be used with a limited sequence of data from the beginning of the growing season, we evaluate the predictive capabilities and, thus, the usability, of the best performing model by feeding it time series data limited in this manner.

2. Materials and Methods

2.1. Data Acquisition

RGB images. Nine crop fields totaling to approximately 85 ha and having wheat, barley and oats as the crop varieties, were included in the study. The data was acquired during the year 2018 in the proximity of Pori, Finland (61°29′6.5″ N, 21°47′50.7″ E). Specific information about the fields is given in Table 1. The fields were imaged with a SEQUIOA (Parrot Drone SAS, Paris, France) multispectral

camera mounted on a Airinov Solo 3DR (Parrot Drone SAS, Paris, France) UAV from the average height fo 150 m using a minimum of three ground control points for each field and preflight color calibration. The imaging was done weekly, from week 21 to 35 and spanning 15 weeks in total. Due to weather conditions precluding UAV flight, gaps in data were present. For each field-specific set of images, a complete mosaic image of a field was constructed with Pix4D (Pix4D S.A., Prilly, Switzerland) software and cut to match the shape of the field boundaries. Radiometric correction was perofrmed using the illumination sensor of the Sequioa camera. The field image data was used as inputs to perform predictions with the considered models.

Table 1. The fields selected for the study in the proximity of Pori, Finland. The thermal time is calculated as the cumulative sum of temperature between the sowing and harvest dates. Mean yield has been calculated from processed yield sensor data for each field.

Field Number	Size (ha)	Mean Yield (kg/ha)	Crop (*Variety*)	Thermal Time	Sowing Date
1	11.11	4349.1	Wheat (*Mistral*)	1290.3	13 May
2	7.59	5157.6	Wheat (*Mistral*)	1316.8	14 May
3	11.77	5534.3	Barley (*Zebra*)	1179.9	12 May
4	11.08	3727.5	Barley (*Zebra*)	1181.3	11 May
5	7.88	4166.9	Barley (*RGT Planet*)	1127.6	16 May
6	13.05	4227.9	Barley (*RGT Planet*)	1117.1	19 May
7	7.61	6668.5	Oats (*Ringsaker*)	1223.4	17 May
8	7.77	5788.2	Barley (*Harbringer*)	1136.1	21 May
9	7.24	6166.0	Oats (*Ringsaker*)	1216.4	18 May

Weather data. The weather data was acquired from the open interface provided by the Finnish Meteorological Institute for Pori area. The thermal growing season started on 13th of April in 2018 and the cumulative temperature was calculated using that as the beginning date. As growth of crops is dictated by the accumulation of sunlight amongst other climatological, soil and nutrient variables, cumulative temperature was deemed robust enough indicator of interval between subsequent data collection days (instead of e.g., time in days). Being a common way to express crop growth phase, the cumulative temperature was utilized as a part of the input data to encode passing of time for the temporal models.

Yield data. As the target data, i.e., the data used as the ground truth for training the models, yield data was acquired during the harvest of each field. The harvesters were equipped with either a Trimble Navigation (Sunnyvale, California, USA) CFX 750 or John Deere (Moline, Illinois, USA) Greenstar 1 yield mapping sensor systems. The systems produce a cloud of geolocated points with multivariate information about the harvest for each point in vector format. This data was first accumulated field-wise and then filtered to contain data points where the yield was between 1500 and 15,000 kg/ha and the speed of the harvester was between 2 and 7 km/h [2]. Finally, the yield map rasters were generated by interpolating the vector points over each field.

2.2. Data Preprocessing

The RGB images taken with the UAV and the cumulative temperatures for imaging dates were utilized as the input data with which the predictions about yields were performed. As spatial models generally have a built in limitation of being able to utilize data with fixed dimensions only, data had to be clipped to smaller fixed dimension frames. As an intended side-effect, using smaller frames makes it possible to better model intra-field yield variability. Like in [2], the fields were split into smaller overlapping frames of size 40 × 40 m with a lateral and vertical step of 10 m. cumulative temperature was added as an additional layer in conjunction with the RGB-layers to have the data contain necessary information for temporal feature learning. The added layer contains constant values corresponding to the field and time of acquisition. The design choice of introducing this data as an additional layer was to have a single source of similarly constructed data for each model architecture.

During the extraction of frames we included every frame that had at least half of its data present at field edges into the final data set. The reasoning behind this was that the spatial models effectively learn filters that are applied over the spatial input data in a successive manner (see Section 2.3.1 for more). Thus, salient features are expected to be present in a frame albiet being just partial due to being located at a field's edge.

The data was also scaled to aid the models in their learning. All values were scaled to the range $[0, 1]$ using feature-wise maximum values as scalers. For the value ranges of unscaled input RGB data, the cumulative temperature calculated from the beginning of the thermal growing season and yield data, see Table 2. As the input data for this study was temporally sequential, the geolocationally matched frames were clipped across every image acquired at a different date for each field. Each sequence of frames was then coupled with geolocationally matching average yield.

Table 2. The value ranges of used input and target variables prior scaling.

Data	Min	Max	Mean	Std
RGB: R	105	254	186.0	19.5
RGB: G	72	243	154.3	18.8
RGB: B	58	223	126.7	18.9
Cumulative °C	388.6	2096	1192	545.0
Yield, kg/ha	1500	14,800	5287	1816

As the last step, the sequences of frames coupled with matching yield information were shuffled and split to training and hold out test data sets with 70%/30% ratio. The samples in the training set are used to optimize the model during the training. The test set is then utilized to evaluate model capabilities with previously unseen data, i.e., its generalization capabilities.

With the total number of generated sequences of frames being 2586, the training data set contained 1810 frame sequences (27,150 frames) and the test set 776 frame sequences (11,640 frames). The general process of generating the frames is depicted in Figure 1.

With the resolution of 0.325 px/m, a single spatial layer in the input data had the dimensions of 128 × 128 px. Using RGB-data with an additional layer constructed from the cumulative temperature conforming to the imaging date, a single frame of data consisted of four layers. With 15 frames, each frame corresponding to a particular week of the growing season, an input sequence of frames thus had the dimensions of $[15 \times 4 \times 128 \times 128]$.

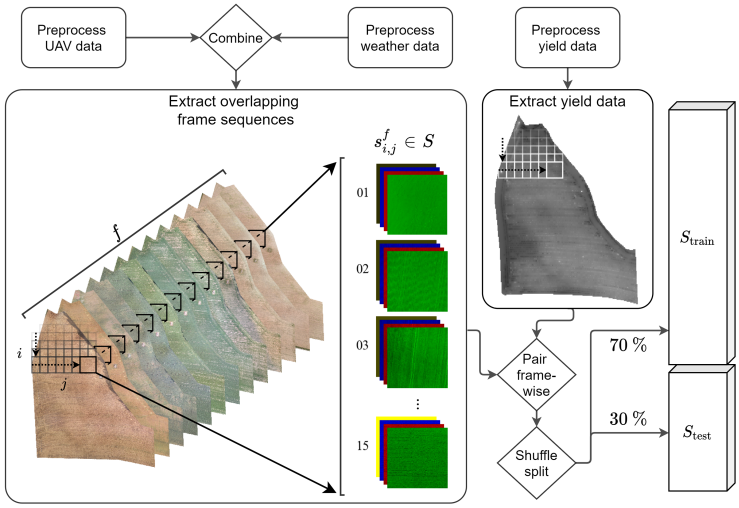

Figure 1. Input frame sequence and target average yield extraction process. Sequences of frames S of fixed width and height were extracted from cumulative temperature enhanced RGB image mosaic sequences as the input data, with f being a distinct field and $s^f_{i,j}$ an extracted sequence of frames from f. The four-layer YBRG, Y being the cumulative temperature, input frames were then geolocationally paired with corresponding yield data to form input-target pairs. Lastly, data was shuffled and split to training and test sets.

2.3. Model Architectures

2.3.1. Convolutional Neural Networks

Convolutional neural networks, often referred to as CNNs, have solidified their place in modeling tasks where the input data is either spatial or spatially representable [16,17]. The main component of the model is the convolution operation, where a set of trainable kernels (or filters) is applied to the input data resulting in a set of spatial features describing the data. For more in-depth explanation of the operations within a single convolution layer, like the application of convolution and pooling, see [2]. The model learns basic features in the first layers and composite features of these basic features at further layers [18]. To help the model better learn these features, batch normalization can be applied to the inputs [19]. The final output of a plain CNN is a set of feature maps. Depending on the use case, these can be either directly utilized or, for example, flattened and fed to a fully connected (FC) layer for regression or classification purposes.

2.3.2. Long Short-Term Memory Networks

The Long Short-Term Memory (LSTM) networks, originally introduced in [20], have been widely utilized in sequence modeling tasks [21]. There are two general concepts to the LSTM that help it in learning temporal features from the data. The first is the concept of memory, introduced as the cell state. The other is the concept of gates, effectively trainable FC layers, manipulating this cell state in response to the new inputs from the data and past outputs of the model. To handle sequences of data, the model loops over the sequences altering its cell (C) and hidden (H) states in the process using the combination of learned parameters in the gates and non-linear activations when combining the gate outputs. Following the Pytorch [22] implementation of LSTM, the following functions are computed:

$$
\begin{aligned}
g_t^i &= \sigma(W_x^I x_t + W_h^I h_{t-1}) \\
g_t^f &= \sigma(W_x^F x_t + W_h^F h_{t-1}) \\
g_t^c &= \tanh(W_x^C x_t + W_h^C h_{t-1}) \\
g_t^o &= \sigma(W_x^O x_t + W_h^O h_{t-1}) \\
C_t &= g_t^f \odot C_{t-1} + g_t^i \odot g_t^c \\
H_t &= O_t = g_t^o \odot \tanh(g_t^c)
\end{aligned}
\tag{1}
$$

where $g_t^{\{i,f,c,o\}}$ are the outputs of the input, forget, cell and output gates, respectively. The gates of the model contain its trainable parameters W. x_t denotes the external input and h_{t-1} the model's previous output. t denotes the current time step. C_t and H_t are the final computed cell and hidden states, respectively. The output O_t of the model is the last computed hidden state H_t. b are the bias factors and $\odot$ is the dot product. The general architecture of an LSTM is depicted in Figure 2.

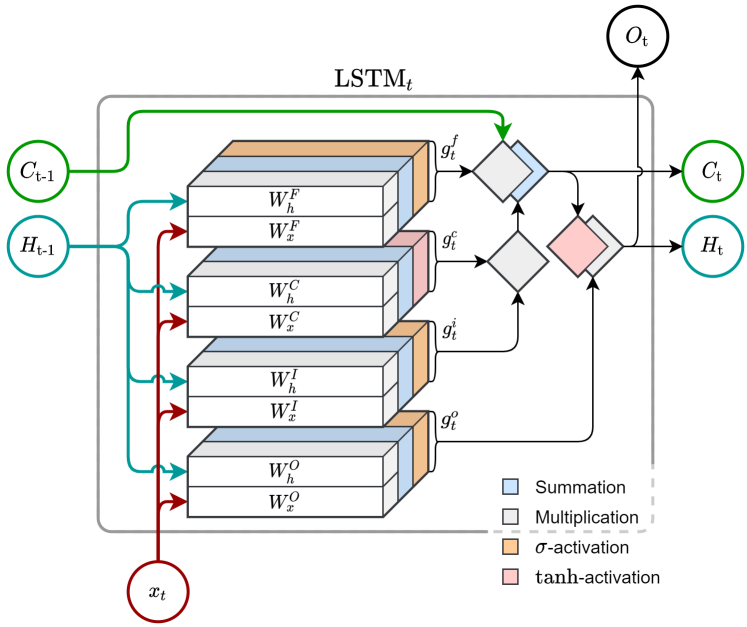

Figure 2. The inner architecture of an LSTM at a time step t. The model takes as its inputs the previous cell state C_{t-1} and hidden state H_{t-1} with the current item x_t of the input sequence. The H_{t-1} and x_t are then passed to forget (W^F), cell (W^C), input (W^I) and output (W^O) gates. These gates, effectively shallow FC layers, are responsible for determining what to keep from previous memory C_{t-1} accumulated from past experiences and what to incorporate to it as the current C_t. This is how the model is able to learn temporal features.

LSTMs can also be employed in bidirectional and stacked form. Bidirectional LSTMs train an additional model in comparison to the unidirectional LSTM presented in Figure 2. One LSTM reads the input from start of the sequence to end ($t_0 \rightarrow t_n$), while the other reads the input from end to start ($t_n \rightarrow t_0$). The outputs of these two parallel models are then combined as final temporal feature outputs [23]. When LSTMs are stacked, the first LSTM operates on the input sequence and subsequent LSTMs then operate on sequences of temporal feature outputs produced by preceding models. Bidirectionality helps

the model learn features from both sides of input sequences, while stacking helps in learning higher level temporal features [24].

2.3.3. CNN-LSTM

The CNN-LSTM is a composite model consisting of a spatial feature extractor or transformer, i.e., a pretrained CNN, and a temporal model, the LSTM [3]. The general idea is to both gain the ability to utilize spatial data and perform sequential modeling with LSTM networks.

The architecture of the pretrained CNN was implemented according to [2] with certain adaptations to have the model better serve as pre-trained spatial feature extractor of the composite CNN-LSTM. Firstly, the model was modified to accept four-band inputs. Secondly, the CNN layers were decoupled from the prediction-producing FC layers as a separate sub-module. Other than those two, the CNN consists of six convolutional layers with batch normalization in every layer and max pooling applied in the first and last layers. All convolution operations utilize 5×5 kernels with 128 kernels in the last operation and 64 in the ones preceding that. The convolutions are performed with zero padding to maintain constant dimensions. The maintaining of dimensions was initially implemented to allow adding an arbitrary number of in-between convolutional layers to the model without diminishing the intermediate hidden output dimensions to oblivion. The input max pooling uses a 8×8 and the output max pooling a 2×2 kernel. The output of the last convolutional layer is passed to a linear layer, squashing the hidden feature space to 256 features akin to [3]. These features are fed to the recurrent LSTM model. When pre-training the CNN only, the output of the squashing linear layer is fed to another linear layer producing the prediction outputs for error metric calculations. This outermost linear layer is omitted when the CNN is used as a part of the CNN-LSTM. The architecture of the spatial feature extracting CNN of the composite CNN-LSTM model is depicted in Figure 3.

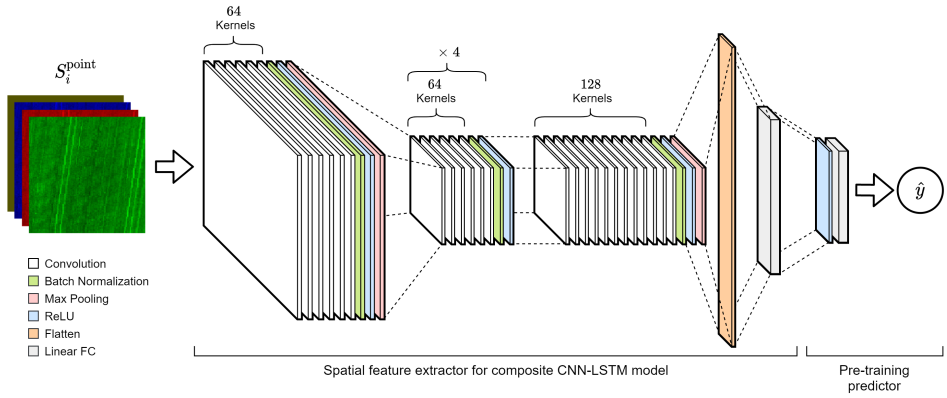

Figure 3. The spatial feature extracting CNN of the CNN-LSTM composite model, i.e., the pretrained CNN. The model is similar to the best performing model of [2]. Alterations in the FC layer composition had to be made to provide sufficient features for the LSTM utilizing the CNN as its input generator.

The temporal feature extracting part of the model is an LSTM, accepting sequences of spatial features as its inputs. During the hyperparameter optimization, we performed architectural experiments also with bidirectional and stacked (multi-layered) LSTMs. Generally, the option to use dropout [25], a regularization technique, is also part of the architectural implementation and that is also the case with Pytorch's LSTM implementation.

While the spatial feature extracting CNN could have been jointly trained with the LSTM, we chose to use a pre-trainend CNN to see whether the point-in-time spatial features could be utilized to perform sequential regression, similar to [10]. We selected this approach also to tie the composite model better

to the framework of the study by [2] by first training the CNN and then examining the ability of the LSTM to learn the temporal features in isolation from the CNN.

2.3.4. ConvLSTM

ConvLSTM [4] is a model combining the features of convolutional and sequential models into a single architecture, using convolutional layers (convolution with pooling etc.) as the LSTM's gate functions. This makes it possible to feed the sequential model the spatial data directly. Akin to how convolutional networks learn, the gates learn to utilize the convolutional kernels to provide the best set of spatial features when building and modifying the cell state C. Thus, contrary to the CNN-LSTM, no pre-extraction of spatial features for further spatial modelling is required. Our implementation of the recurrent architecture follows Equation (1).

From the point of architectural composition, the ConvLSTM is an LSTM at its essential core. In the ConvLSTM, using Figure 2 as a reference, the cell and hidden state altering gates $W^{\{F,C,I,O\}}$ have, however, been changed from conventional LSTM's shallow FC layers to shallow CNNs. To extract robust features from the input data, the gates W_x^* for inputs also employ a max pooling layer with a 3×3 kernel having padding and stride to halve input image dimensions after the first convolution. Due to the nature of CNNs learning spatial features in increasing complexity from layer-to-layer, we also allowed the model to utilize up to two convolutional layers for each W. Like with the CNN of the CNN-LSTM, we wanted to make sure that the intermediate feature map dimensions remaing unchanged, i.e., do not diminish as items in the sequences are processed. Thus, we used 32 convolutional kernels with 5×5 kernel shape and sufficient padding. The possibility to use batch normalization for inputs, stacking, bidirectionality and dropout were also implemented to find the best performing architectural composition.

2.3.5. 3D-CNN

As initially reported by [5], 3D-CNNs performed remarkably well in modeling tasks involving spatio-temporal data. Being CNNs, the 3D-CNNs utilize all same architectural features as more commonly used convolutional models. What's different is their use of convolution in the depth dimension, searching for robust features across sequences of input data in addition to spatial features extracted from the individual images. The sequential nature of input data is not limited to time, but can also be, for example, hyperspectral multi-layer point-in-time data with the aim of finding salient intra-band features [14]. The 3D-convolution is applied with a learnable three dimensional kernel, depicted in Figure 4. Kernel dimensions are in $[Z \times X \times Y]$ format, where Z denotes the time dimension.

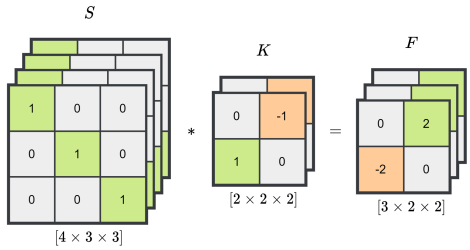

Figure 4. An illustration of 3D convolution. The 3D convolution operation effectively applies the kernel in one additional dimension cmompared to the normal convolution, the depth or z-axis. Like the input, the kernels are three dimensional. The dimensions of the feature map conform to how many times a kernel can wholly be applied to the input data along all three dimensions. With stride of one in each dimension, a $[2 \times 2 \times 2]$ kernel is applied on a $[4 \times 3 \times 3]$ input sequence of layers two times in x and y dimensions and thrice in z dimension, resulting in a $[3 \times 2 \times 2]$ feature map, its values being sums of products over distinct applications of K akin to 2D convolution.

The general architecture of the 3D-CNN we implemented conforms closely to how a CNN is generally constructed with the exception of using 3D instead of 2D convolutions. In the first layer the data is, however, grouped by layers as depicted in Figure 5. This is to have the model learn the spatio-temporal features of the data on a per-layer basis.

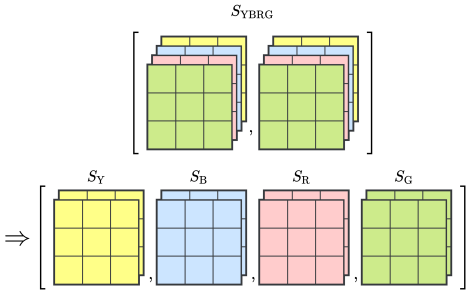

Figure 5. In the first layer of the 3D-CNN, the sequences of multi-layer input data are handled layer-wise. This helps the model first learn layer-wise spatio-temporal features, which are then composed as interlayer spatio-temporal features in the subsequent layers.

Our implementation follows the general CNN architecture of Figure 3. All layers prior to the last have the same number of kernels, the last having twice as many. We employ max pooling only in the first and last layers while the intermediate layers preserve intermediate feature map dimensions. The exact number of kernels is determined via hyperparameter tuning. As per [5], we perform convolutions with $[3 \times 3 \times 3]$ kernels with zero padding, having the pooling layers perform the diminishing of feature map dimensions. Mixing the depth-wise steps from [5] and spatial steps from [2], the first max pooling employs a $[1 \times 8 \times 8]$ kernel while the last a $[2 \times 2 \times 2]$ kernel. The kernels' strides equal respective kernel sizes, i.e., no overlap is applied. Like with the ConvLSTM, we the option to utilize batch normalization in every layer was also implemented for hyperparameter tuning purposes.

2.4. Training and Optimization

The process of training neural networks generally requires hyperparameter tuning. While model parameters, such as the layer-wise weights, are optimized during training in response to regression errors with the selected optimization algorithm, the hyperparameters are what dictate how the model is initialized and in what manner the optimization is applied. Examples of these hyperparameters include the learning rate and model depth. From available hyperparameter tuning methods we chose to use random search, in which a distribution is defined for each hyperparameter and then a value is randomly drawn for each distinct training [26].

We first performed the hyperparameter tuning for the pretrained CNN of the CNN-LSTM. Unlike sequential models, the pretrained CNN was fed single frames (i.e., point-in-time) drawn randomly from the set of all training data set frames. The goal was to have the model learn general spatial features for the whole growing season. Following [2], we used Adadelta [27] as the optimizer. Due to having input data consist of four distinct layers instead of only RGB layers, we performed tuning for the learning rate. Weight decay and the ρ coefficient were utilized from [2].

For the spatio-temporal models, we used Adam [28] as the optimizing algorithm for each model architecture akin to [7,8,11]. The spatio-temporal models were trained with frame sequences. Each model was tuned for LSTM and CNN architectural (where applicable) and optimizer hyperparamenters. The architectural and optimizer hyperparameters are given in Table 3. All hyperparameters were tuned simultaneously and not in sequential succession, meaning that the

hyperparameter values were drawn from their respective distributions for each hyperparameter at the start of a training.

Table 3. The model-specific hyperparameters and their distributions tuned during the random search. Square brackets indicate closed interval with lower and upper limit included, while curly brackets indicate a set from which a value was chosen from. The presence of $\vee$ indicates a boolean toggle with $p = 0.5$. The $\log_{10}$-uniform distribution used for the learning rate draws a value a from a float-uniform distribution in a given range to calculate 10^a.

Hyperparameter	Distribution	Pre-CNN	CNN-LSTM	ConvLSTM	3D-CNN
LSTM Architectural parameters					
LSTM layers	int-uniform	-	$[1, 3]$	$[1, 3]$	-
Dropout	float-uniform	-	$[0, 1] \vee 0$	$[0, 1] \vee 0$	-
Bidirectional	bool		$0 \vee 1$	$0 \vee 1$	-
CNN Architectural parameters					
CNN layers	int-uniform	-	-	$[1, 2]$	$[2, 5]$
Batch normalization	bool	-	-	$0 \vee 1$	$0 \vee 1$
Kernels	set	-	-	$\{32, 64, 128\}$	$\{32, 64, 128\}$
Optimizer parameters					
Learning rate	$\log_{10}$-uniform	$[-4, -1]$	$[-4, -2]$	$[-4, -2]$	$[-4, -2]$
L2-regularization	float-uniform	-	$[0, 1] \vee 0$	$[0, 1] \vee 0$	$[0, 1] \vee 0$

With each sequential model type we performed 300 distinct model training session, using random search for hyperparameter tuning and Skorch [29] as the training framework. For the pretraining of the CNN-LSTM's spatial feature extracting CNN, 50 models were trained to tune the learning rate due to the additional layer in inputs. The ρ-coefficient of the Adadelta algorithm and the weigth decay parameters were utilized from [2]. The total number of trained models was thus 950. The parameters of each model were initialized with *xavier*-uniform intialization [30]. During training, we utilized early stopping with patience for stagnant progress of 50 epochs. A single training iteration was allowed to continue for a maximum of 250 epochs. With continued training, where the best performing model parameter configuration is utilized as the starting point for a subsequent round of training, a model was allowed to be trained a maximum of 500 epochs. However, the use of early stopping was allowed to halt the training prior reaching that limit. Training was conducted with a separate training data set, having the training process utilize 5-fold cross-validation, where the training and validation batches are derived from the training data set. The final evaluation of a trained model was performed with the hold-out test data set. The models were trained in a distributed computation environment, utilizing Nvidia Tesla V100 Volta and Pascal architecture cloud GPUs.

3. Results

From the sets of trained models produced during the hyperparameter tuning process, the best performing models were singled out. During training we monitored the mean squared error (MSE) of the 5-fold cross validation. We also computed metrics for root mean square error (RMSE), mean absolute error (MAE) of unscaled targets, mean absolute percentage error (MAPE) and the coefficient of determination (R^2). The best performing model architecture was the 3D-CNN, expressing notably better performance with the best performing model than the rest of the trained architectures. The model performing worst was somewhat surprisingly the ConvLSTM, showing performance inferior even to the pretrained CNN trained with just point-in-time data. The performance metrics for the unscaled predicted and true target values for each model architecture are given as in Table 4.

Table 4. The performance metrics of the best-performing models resulting from model-specific hyperparameter tuning process with samples from the test set. The trained models were evaluated with a hold-out test set. Best performance was achieved with the 3D-CNN architecture. The number of trainable parameters indicate the model complexity. Best performance values are in bold text.

Model	Test RMSE (kg/ha)	Test MAE (kg/ha)	Test MAPE (%)	Test R^2 -	Trainable Parameters
Pretrained CNN	692.8	472.7	10.95	0.780	2.72×10^6
CNN-LSTM	456.1	329.5	7.97	0.905	2.94×10^6
ConvLSTM	1190.3	926.9	22.47	0.349	9.03×10^5
3D-CNN	**289.5**	**219.9**	**5.51**	**0.962**	7.48×10^6

The most consistently fitting sequential architecture was the CNN-LSTM in terms of test set performance with trained models. Other architectures produced occasional ill-fitted models with errors several magnitudes higher than their best performing counterparts. The RMSE percentiles depicting the general consistency in fitting for the spatio-temporal models are given in Table 5.

Table 5. The RMSE percentiles accross all trained spatio-temporal models. The RMSE percentiles indicate the consistency of a model architecture in generalizing to unseen samples with the training data. Out of the three, the CNN-LSTM was most consistent in how it was able to fit to the data and produce generalizable results. The training of other model architectures produced occasionally ill-fitted models. Best performance values are in bold text.

Model	Test RMSE (kg/ha)				
	Min	25%	50%	75%	Max
CNN-LSTM	456.1	**655.1**	**1475.6**	**1623.7**	$\mathbf{2.152 \times 10^3}$
ConvLSTM	1190.3	1477.8	1646.6	8750.2	1.334×10^6
3D-CNN	**289.5**	1355.4	1493.6	1649.0	1.926×10^6

Due to the training of the model architectures being a process of empirically evaluating randomly drawn hyperparameter sets, visualization of the hyperparameters against a performance metric further helps in understanding model fitting consistency. Out of the architectures, the CNN-LSTM and the 3D-CNN show similar behaviour in hyperparameter value distribution, the latter having a discernible dispersion in the values against the performance metric. ConvLSTM, as already stated, exhibits clearer sporadicity. The architecture-specific hyperparameter distributions plotted against the test RMSE are given in Figure 6.

The best performing configuration of hyperparameters dictating how a model is to be initialized and trained were sought by performing random search. In random search, each hyperparameter is assigned with a distribution, from which a value is drawn for each independent training of the model. The hyperparameters for the best performing models are given in Table 6.

In addition to performing comparative performance evaluation between the selected deep learning architectures with data sequences spanning the time from sowing to harvest, we also evaluated the performance of the best performing model configuration (architecture with hyperparameters) using data from an actionable time frame. In other words, we combined various configurations of input data sequences starting from image data acquisition dates closest to sowing (week 21) and ending at the midsummer (week 25). The following sequence configurations were built using the aforementioned time range:

- Weeks 21, 22, 23, 24, 25; five temporal frames.
- Weeks 21, 22, 23, 24; four temporal frames.
- Weeks 22, 23, 24, 25; four temporal frames.

- Weeks 21, 22, 23; three temporal frames.
- Weeks 23, 24, 25; three temporal frames.
- Weeks 21, 23, 25; three temporal frames.

We trained ten iterations of the best model configuration, the 3D-CNN, for each input sequence type to account for the effects of random model parameter initialization. The training was conducted as before, utilizing 5-fold cross-validation with the training data and testing the generalization capabilities with the hold-out test data, separately. The performance of these trained models with the test data are given in Table 7, where each row corresponds to a distinct configuration of input frame sequences. The best perfoming configuration in terms of RMSE and MAE is the four week long sequence taken from the beginning of the season (weeks 21 to 24). In terms of MAPE, the best performing configuration, however, consists of five weeks from the beginning of the season (weeks 21 to 25), although the difference to the four week sequence is small.

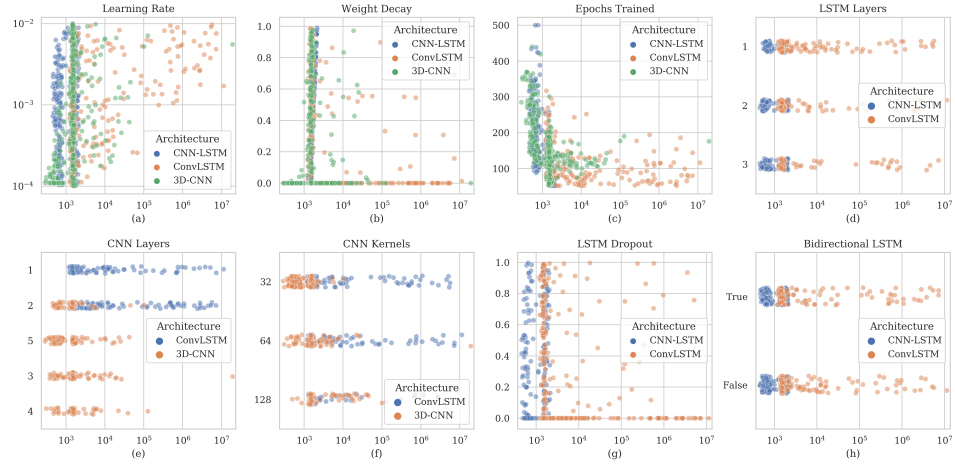

Figure 6. The architecture-specific distributions of hyperparameters against the test RMSE (x axis). (**a**,**b**) are the optimizer hypeparameters and (**c**) is the training length in epochs. (**d**,**g**,**h**) are the LSTM architectural hyperparameters, while (**e**,**f**) are the CNN architectural hyperparameters. (**d**,**e**,**f**,**h**) contain categorical values in y axis with values spread category-wise for easier observation of clustering. The rest of the sub-figures have a continuous y axis.

Table 6. The architecture specific hyperparameter values for the best performing models. The value types conform to the values given in Table 3. The feature extracting CNN of the CNN-LSTM was not tuned for hyperparameters as tuning results from previous study were utilized.

Hyperparameter	**Pre-CNN**	**CNN-LSTM**	**ConvLSTM**	**3D-CNN**
	LSTM Architectural parameters			
LSTM layers	-	2	2	-
Dropout	-	0.5027	0.9025	-
Bidirectional	-	0	1	-
	CNN Architectural parameters			
CNN layers	6 *	-	1	5
Batch normalization	Yes *	-	No	No
Kernels	128/64 *	-	32	32
	Optimizer parameters			
Learning rate	1.000×10^{-1}	7.224×10^{-4}	1.361×10^{-3}	1.094×10^{-4}
L2-regularization	0.9 *	0.0	0.0	0.0

* Values taken from [2].

Table 7. Retraining results of the best performing 3D-CNN configuration with various input sequence configurations from the test set. The input data was constructed from the first five imagings (weeks 21 to 25). The composition of weekly data was varied and variations evaluated by fitting the best performing 3D-CNN configuration to each variation. Best performance values are in bold text.

Weeks in Input Sequence	Test RMSE (kg/ha)	Test MAE (kg/ha)	Test MAPE (%)	Test R^2 -
21, 22, 23, 24, 25	413.8	320.6	**7.04**	0.921
21, 22, 23, 24	**393.9**	**292.8**	7.17	**0.929**
22, 23, 24, 25	439.3	343.0	7.90	0.911
21, 22, 23	543.5	421.4	10.02	0.864
23, 24, 25	425.0	326.6	8.25	0.917
21, 23, 25	478.1	369.3	8.72	0.895

Operating with single frames, the models can be used to construct predictions for whole fields. This is achieved by extracting frames from an image of the fields and feeding them as inputs to the model. Re-arranging the predictions to original field shape yields a map of frame-wise yield predictions. The performance of the best performing 3D-CNN configuration with both full length and shortened sequences is illustrated in Figure 7 with a 10 m step between predicted points. As the test set was constructed from frame sequences randomly taken from all extracted frame sequences, the illustrations contain frames from both the training and the test set.

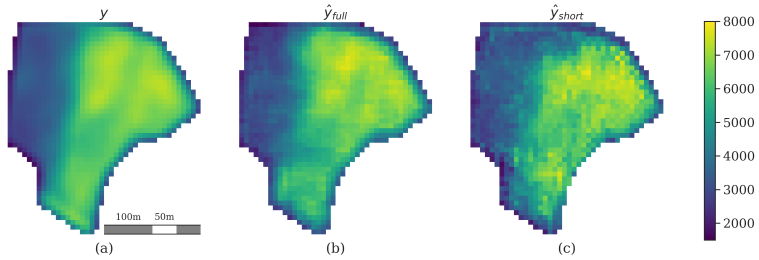

Figure 7. Frame-based 3D-CNN model performances against true yield data. (**a**) is the true yield map of the field. (**b**) is the modelled prediction, utilizing the full length frame sequences. (**c**) is then the actual in-season prediction utilizing four first frames of the weekly frame sequence. Units are absolute values in crop yield kg/ha. One pixel in the images corresponds to a 10 × 10 m area. Images are unsmoothed, represent the values as they were produced and contain samples from both training and test sets due to how the sets were constructed.

4. Discussion

In this study we evaluated the feasibility of using spatio-temporal deep learning architectures in modelling crop yield at the intra-field scale. Using sequences of UAV and weather data collected in the vicinity of Pori, Finland, during the growing season of 2018, we split the fields to geolocationally matched temporal sequences of frames of fixed width and height. We developed and trained three different model architectures: CNN-LSTM, ConvLSTM and 3D-CNN. We first determined the best performing architecture by performing hyperparameter tuning with complete temporal sequences of frames (15 time steps). With the best performing model architecture and hyperparameter configuration we then evaluated the predictive capabilities of the models by using a shorter temporal sequence of frames from the beginning of the growing season.

Of the architectures, the 3D-CNN performed the best in full sequence modelling. The best performing model consisted of five 3D-CNN layers using 32 kernels in the layers. Other architectural configurations are given in Section 2.3.5. The model attained 218.9 kg/ha test MAE, 5.51% test MAPE and 0.962 test R^2-score. Compared to the study presented in [2] using just a point-in-time single

frame predictor with 484.3 kg/ha MAE and 8.8% MAPE, the modelling performance was improved by 265.4 kg/ha MAE (54.8% improvement) and 3.29% MAPE (37.4% improvement). In terms of prediction performance with smaller sequences from the beginning of the season, the 3D-CNN performed the best using four first frames of the whole input sequences. With a shorter sequence the model attained 292.8 kg/ha test MAE, 7.17% test MAPE and 0.929 test R^2-score. The respective improvements to the best performing model presented in [2] were 191.5 MAE (39.5% improvement) and 1.63% MAPE (18.5% improvement).

Recent studies make use of UAVs in a variety of imaging applications. The use of UAVs has become more common, as shown in [31], a review of UAV thermal imagery applications in the domain of precision agriculture. One of the reasons is the increased need of performing classification and regression at scales more accurate than what is attainable by publicly available satellite data sources. However, the use of point-in-time UAV data is common. Ref. [32] utilized UAVs to gather hyperspectral data of potato tuber growth at the resolution of 2.5 cm/px. They utilized traditional ML methods, such as linear models and decision trees, to perform tuber yield estimation using individual data points gathered in-season at the intra-field scale, achieving 0.63 R^2-score for the tuber yield prediction accuracy with a Ridge regression. Ref. [33] used UAV to collect multispectral data from wheat and corn fields to estimate intra-field crop nitrogen content using linear regression and point samples—spatial features were not utilized. They fit multiple linear models to wheat and corn and attained 0.872 R^2-score on average. Ref. [34] performed wheat leaf area index and grain yield estimation with various vegetation indices derived from point-in-time multispectral UAV data using multiple machine learning methods, neural networks included. The highest performance they attained was 0.78 R^2-score with a Random Forest. However, they fed the input data as point samples.

Satellites perform frequent overflights over vast areas accross the globe. They are thus an ideal source of automatically generated multi-temporal remote sensing data [35]. This is one of the reasons, why spatio-temporal modelling is more notably present in the context of publicly available satellite data sources, contrary to UAV data requiring manual collection. Spatio-temporal models akin to the setting of our study have been utilized in various modelling tasks with remote sensing data in the domain of agriculture. Performing county-scale soybean yield prediction, ref. [6] used a CNN, an LSTM and a composite CNN-LSTM to model soybean yield with in-season satellite data. They achieved an average 0.78 R^2-score with the spatio-temporal CNN-LSTM model. Their input data resolutions were from 500 × 500 m/px to 1 × 1 km/px. Ref. [7] performed crop type classification in Europe and Africa with multi-temporal satellit data at resolutions from 3 × 3 m/px to 10 × 10 m/px. They attained F1 scores 91.4 for the CNN-ConvLSTM and 90.0 for the 3D-CNN, averaged over crop types in their Germany data set. Ref. [8] performed pre-season crop type mapping for the area of Nebraska, US, employing a CNN-ConvLSTM to extract spatio-temporal features from multi-temporal multi-satellite composite data set. Using prior years of crop type related data to predict a map of crop types, they attained an average accuracy of 77% across all crop types in their data. The data was processed to a resolution of 30 × 30 m/px. Ref. [9] utilized a 3D-CNN to classify crop types from multi-temporal satellite data gathered from an area within China, acquiring a classification accuracy of 98.9% with the model. Their input data resolutions were from 4 × 4 m/px to 15 × 15 m/px. Ref. [36] performed weekly UAV image collections in a controlled field experiment with soybeans, performing seed yield prediction with multiple linear models fit the multi-temporal data. Thus, spatio-temporal modelling with novel techniques was not performed. With seed yield prediction, they achieved 0.501 adjusted R^2 score. The resolution of their data was 1.25 × 1.25 cm/px.

In remote sensing, the multitemporal aspect of satellite sensor data the has been well studied. In their review of the applications of multisource and multitemporal data fusion in remote sensing, ref. [37] show how studies utilizing the temporal feature of satellite data are rather common. However, in terms of models and data usage settings, they only briefly mention how novel deep learning architectures have only recently been applied in this data domain. They cite that both data and methods, especially the latter, are still under development and a subject of further research. While

some studies have not found additional benefit in using multitemporal data [38], partly due to selected data utilization techiniques, others find benefit over using just point-in-time data [33,35–37].

Regarding the poor performance of the ConvLSTM in our study, the studies by [7,8] might provide some basis for understanding the phenomenon. In both studies, the ConvLSTM was preceded with an exclusively spatial feature extracting CNN model. The extracted feature maps were then fed to the ConvLSTM for temporal feature extraction. While in our study we experimented with multiple convolutional layers in the ConvLSTM model, it could very well be that using a pre-trained CNN akin to CNN-LSTM is required for the ConvLSTM as well. Model complexity is another way to look at this, as the 3D-CNN model was more complex compared to the ConvLSTM. This indicates that the effective capacity of the ConvLSTM might indeed be too low. Thus, increasing the effective capacity by either adding a spatial feature pre-extractor or increasing the gate-wise layer count could increase the performance of this model architecture in similar study setting.

As we utilized weather information at city-scale, the precision of change in the growth phase could be further improved with specifically located weather stations. Weather stations located in the approximate vicinity of the fields under scrutiny could provide better and more accurate measurements of the local temperatures and other climatological variables and thus might help the model produce even better predictions when sequences are involved. Using other data sources, such as soil information and topology maps, could also be further utilized to improve the predictive capabilities of the model. As growing season provides information about how the crops have concretely developed, the soil and topology maps provide more in terms of a prior that the UAV images are then used to further develop as new samples emerge.

A limitation to our study is the use of aggregated crop type data collected from various fields. Using a single model to predict for wheat, barley and oats prohibits both the inference with and the performance analysis of the model on a per-crop basis. Additionally, the remote sensing data based modelling approach doesn't take into account any existing crop growth models. Those could well be utilized to further provide better performance, akin to what has been done in [36], but this is outside the scope of our study. That being said, the modelling task of this study was not that of crop growth, but yield estimation with UAV remote sensing data.

5. Conclusions

Our study seeks to combine three increasingly common but yet seldom co-utilized concepts in the domain of crop yield estimation: the use of high resolution UAV image data, time series regression and novel spatio-temporal neural network architectures. It has already been shown that crop yield prediction with spatial neural networks, i.e., CNNs, is feasible and produces results accurate enough for performing actions in-season [2]. In this study, we show that adding time as an additional feature not only improves the modelling performance with UAV RGB data (see Table 4) but also improves the predictive capabilities (see Table 7). Furthermore, using weekly UAV data gathered during the first month provides enough data for the model to build an accurately predicted yield map from which to draw further conclusions.

To conclude, the use of multitemporal remote sensing data is not only common but also beneficial in crop yield modelling and prediction. Furthermore, the easy accessibility of commercially available UAVs with mounted RGB sensors enables image data acquisition in higher resolutions compared to satellites. This in turn opens up the possibilities to perform modelling and predictions at intra-field scale. As shown in our study, the use of UAV-based data and proper spatio-temporal deep learning techniques is an enabler of more sophisticated Decision Support Systems in the domain of agriculture.

Author Contributions: Conceptualization, P.N. and N.N.; methodology, P.N. and N.N.; software, P.N.; validation, P.N., N.N. and T.L.; formal analysis, P.N. and N.N.; investigation, P.N.; resources, P.N., N.N. and P.L.; data curation, P.N., N.N. and P.L.; writing—original draft preparation, P.N.; writing—review and editing, P.N., N.N. and T.L.; visualization, P.N.; supervision, N.N. and T.L.; project administration, P.N., N.N. and T.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by Mtech Digital Solution Oy, Vantaa, Finland.

Acknowledgments: We would like to thank Tampere University for providing the computational resources and MIKÄ DATA project for providing us with the data.

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

References

1. Narra, N.; Nevavuori, P.; Linna, P.; Lipping, T. A Data Driven Approach to Decision Support in Farming. In *Information Modelling and Knowledge Bases XXXI*; IOS Press: Amsterdam, The Netherlands, 2020; Volume 321, pp. 175–185, doi:10.3233/FAIA200014.
2. Nevavuori, P.; Narra, N.; Lipping, T. Crop yield prediction with deep convolutional neural networks. *Comput. Electron. Agric.* **2019**, *163*, 104859, doi:10.1016/j.compag.2019.104859.
3. Sainath, T.N.; Vinyals, O.; Senior, A.; Sak, H. Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia, 19–24 April 2015; pp. 4580–4584.
4. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.Y.; Wong, W.K.; Woo, W.C. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 802–810.
5. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3D convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
6. Sun, J.; Di, L.; Sun, Z.; Shen, Y.; Lai, Z. County-Level Soybean Yield Prediction Using Deep CNN-LSTM Model. *Sensors* **2019**, *19*, 4363, doi:10.3390/s19204363.
7. Rustowicz, R.; Cheong, R.; Wang, L.; Ermon, S.; Burke, M.; Lobell, D. Semantic Segmentation of Crop Type in Africa: A Novel Dataset and Analysis of Deep Learning Methods. In Proceedings of the CVPR Workshops, Long Beach, CA, USA, 16–20 June 2019; pp. 75–82.
8. Yaramasu, R.; Bandaru, V.; Pnvr, K. Pre-season crop type mapping using deep neural networks. *Comput. Electron. Agric.* **2020**, *176*, 105664, doi:10.1016/j.compag.2020.105664.
9. Ji, S.; Zhang, C.; Xu, A.; Shi, Y.; Duan, Y. 3D Convolutional Neural Networks for Crop Classification with Multi-Temporal Remote Sensing Images. *Remote Sens.* **2018**, *10*, 75, doi:10.3390/rs10010075.
10. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA, 7–9 May 2015.
11. Liu, Q.; Zhou, F.; Hang, R.; Yuan, X. Bidirectional-Convolutional LSTM Based Spectral-Spatial Feature Learning for Hyperspectral Image Classification. *Remote Sens.* **2017**, *9*, 1330, doi:10.3390/rs9121330.
12. Ienco, D.; Interdonato, R.; Gaetano, R.; Ho Tong Minh, D. Combining Sentinel-1 and Sentinel-2 Satellite Image Time Series for land cover mapping via a multi-source deep learning architecture. *ISPRS J. Photogramm. Remote Sens.* **2019**, *158*, 11–22, doi:10.1016/j.isprsjprs.2019.09.016.
13. Yue, Y.; Li, J.H.; Fan, L.F.; Zhang, L.L.; Zhao, P.F.; Zhou, Q.; Wang, N.; Wang, Z.Y.; Huang, L.; Dong, X.H. Prediction of maize growth stages based on deep learning. *Comput. Electron. Agric.* **2020**, *172*, 105351, doi:10.1016/j.compag.2020.105351.
14. Li, Y.; Zhang, H.; Shen, Q. Spectral–Spatial Classification of Hyperspectral Imagery with 3D Convolutional Neural Network. *Remote Sens.* **2017**, *9*, 67, doi:10.3390/rs9010067.
15. Barbosa, A.; Trevisan, R.; Hovakimyan, N.; Martin, N.F. Modeling yield response to crop management using convolutional neural networks. *Comput. Electron. Agric.* **2020**, *170*, doi:10.1016/j.compag.2019.105197.
16. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90, doi:10.1145/3065386.
17. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9, doi:10.1109/CVPR.2015.7298594.
18. Zeiler, M.D.; Fergus, R. Visualizing and Understanding Convolutional Networks. In *Computer Vision—ECCV 2014*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2014; Volume 8689, pp. 818–833.
19. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv* **2015**, arXiv:1502.03167.

20. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780, doi:10.1162/neco.1997.9.8.1735.
21. Schmidhuber, J. Deep Learning in Neural Networks: An Overview. *Neural Netw.* **2014**, *61*, 85–117, doi:10.1016/j.neunet.2014.09.003.
22. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in PyTorch. In Proceedings of the NIPS-W, Long Beach, CA, USA, 4–9 December 2017.
23. Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681, doi:10.1109/78.650093.
24. Graves, A. Generating Sequences with Recurrent Neural Networks. *arXiv* **2013**, arXiv:1308.0850, doi:10.1145/2661829.2661935.
25. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958, doi:10.1214/12-AOS1000.
26. Bergstra, J.; Bengio, Y. Random Search for Hyper-Parameter Optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305, doi:10.1162/153244303322533223.
27. Zeiler, M.D. ADADELTA: An Adaptive Learning Rate Method. *arXiv* **2012**, arXiv:1212.5701, doi:10.1145/1830483.1830503.
28. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *ICLR* **2014**, 1–15, doi:10.1145/1830483.1830503.
29. Tietz, M.; Fan, T.J.; Nouri, D.; Bossan, B.; Skorch Developers. Skorch: A Scikit-Learn Compatible Neural Network Library That Wraps PyTorch. 2017. Available online: https://skorch.readthedocs.io/ (accessed on 16 October 2020).
30. Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. *J. Mach. Learn. Res.* **2010**, *9*, 249–256.
31. Messina, G.; Modica, G. Applications of UAV thermal imagery in precision agriculture: State of the art and future research outlook. *Remote Sens.* **2020**, *12*, 1491, doi:10.3390/RS12091491.
32. Sun, C.; Feng, L.; Zhang, Z.; Ma, Y.; Crosby, T.; Naber, M.; Wang, Y. Prediction of end-of-season tuber yield and tuber set in potatoes using in-season uav-based hyperspectral imagery and machine learning. *Sensors* **2020**, *20*, 5293, doi:10.3390/s20185293.
33. Lee, H.; Wang, J.; Leblon, B. Intra-Field Canopy Nitrogen Retrieval from Unmanned Aerial Vehicle Imagery for Wheat and Corn Fields. *Can. J. Remote Sens.* **2020**, *46*, 454-472, doi:10.1080/07038992.2020.1788384.
34. Fu, Z.; Jiang, J.; Gao, Y.; Krienke, B.; Wang, M.; Zhong, K.; Cao, Q.; Tian, Y.; Zhu, Y.; Cao, W.; et al. Wheat growth monitoring and yield estimation based on multi-rotor unmanned aerial vehicle. *Remote Sens.* **2020**, *12*, 508, doi:10.3390/rs12030508.
35. Liu, S.; Marinelli, D.; Bruzzone, L.; Bovolo, F. A review of change detection in multitemporal hyperspectral images: Current techniques, applications, and challenges. *IEEE Geosci. Remote Sens. Mag.* **2019**, *7*, 140–158, doi:10.1109/MGRS.2019.2898520.
36. Borra-Serrano, I.; Swaef, T.D.; Quataert, P.; Aper, J.; Saleem, A.; Saeys, W.; Somers, B.; Roldán-Ruiz, I.; Lootens, P. Closing the phenotyping gap: High resolution UAV time series for soybean growth analysis provides objective data from field trials. *Remote Sens.* **2020**, *12*, 1644, doi:10.3390/rs12101644.
37. Ghamisi, P.; Rasti, B.; Yokoya, N.; Wang, Q.; Hofle, B.; Bruzzone, L.; Bovolo, F.; Chi, M.; Anders, K.; Gloaguen, R.; et al. Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art. *IEEE Geosci. Remote Sens. Mag.* **2019**, *7*, 6–39, doi:10.1109/MGRS.2018.2890023.
38. Hauglin, M.; Ørka, H.O. Discriminating between native norway spruce and invasive sitka spruce—A comparison of multitemporal Landsat 8 imagery, aerial images and airborne laser scanner data. *Remote Sens.* **2016**, *8*, 363, doi:10.3390/rs8050363.

PUBLICATION

V

Assessment of Crop Yield Prediction Capabilities of CNN usign Multisource Data

P. Nevavuori, N. Narra, P. Linna and T. Lipping

*New Developments and Environmental Applications of Drones - Proceedings of FinDrones 2020*2021

DOI: 10.1007/978-3-030-77860-6

Noname manuscript No.
(will be inserted by the editor)

Assessment of Crop Yield Prediction Capabilities of CNN using Multisource Data

Petteri Nevavuori · Nathaniel Narra · Petri Linna · Tarmo Lipping

Received: date / Accepted: date

Abstract The growing abundance of digitally available spatial, geological and climatological data opens up new opportunities for agricultural data based input-output modeling. In our study, we took a Convolutional Neural Network model previously developed on Unmanned Aerial Vehicle (UAV) image data only and set out to see whether additional inputs from multiple sources would improve the performance of the model. Using the model developed in a preceding study, we fed field-specific data from the following sources: near-infrared data from UAV overflights, Sentinel-2 multispectral data, weather data from locally installed Vantage Pro weather stations, topographical maps from National Land Survey of Finland, soil samplings and soil conductivity data gathered with a Veris MSP3 soil conducitivity probe. Either directly added or encoded as additional layers to the input data, we concluded that additional data helps the spatial point-in-time model learn better features, producing better fit models in the task of yield prediction. With data of four fields, the most significant performance improvements came from using all input data sources. We point out, however, that combining data of various spatial or temporal resolution (i.e., weather data, soil data and weekly acquired images, for example) might cause data leakage between the training and testing data sets when training the CNNs and, therefore, the improvement rate of adding additional data layers should be interpreted with caution.

Keywords crop yield prediction · CNN · multisource input · remote sensing · intra-field

P. Nevavuori
Mtech Digital Solutions Oy,
E-mail: petteri.nevavuori@mtech.fi

P. Linna · N. Narra · T. Lipping
Tampere University,
E-mail: {nathaniel.narra, petri.linna, tarmo.lipping}@tuni.fi

1 Introduction

The application of novel and performant deep learning techniques has seen an increasing trend in the last few years in the domain of Smart Farming and Precision Agriculture [6]. Multiple factors are at play: the abundance of open access satellite system spatial data, availability of commercial unmanned aerial vehicles (UAVs) mountable with external sensors, developments in the soil sensor and camera sensor technologies and the constant need to optimize the production of farms.

Convolutional Neural Networks (CNN), being a subset of deep learning, have been utilized in recent studies on crop yield prediction [6]. The spatial model architecture has been used in predicting cotton yield from RGB data taken at close proximity [15], cereal crop yield prediction from mid-altitude UAV RGB data [11], rice grain yield estimation [18] and crop yield prediction using multisource inputs on patch-scale [4]. In [11] we compared intra-field crop yield estimation performance with NDVI and RGB data from the earlier and later part of the growing season with a variety of CNN configurations. The focus of that study was to assess the generalization capability of a yield prediction model with UAV RGB data.

1.1 Objectives

In this study, we examine the effect of additional field-related spatial or spatial-like data on the intra-field crop yield prediction capabilities using data gathered from the earlier half of the growing season of 2018 (weeks 21 to 26). The objective of this study is to assess crop yield prediction capabilities with the best CNN model composition from [11] by varying the input data configuration. The focus of this study is to see whether additional data, such as weather data, soil and ground information and open-access Sentinel-S2 data would improve the point-in-time prediction performance compared to just using UAV-based RGB data. To limit the scope of the study, architectural and hyperparameter tuning of the CNN model is not addressed here to better isolate performance changes to data and the tuned out architectural and optimizer related hyperparameters were thus taken from [11].

2 Material and Methods

2.1 Data Acquisition

For this study, four crop fields were selected for data acquisition in the vicinity of Pori, Finland (61°29'6.5"N, 21°47'50.7"E) for the growing season of 2018. The field information is provided in Table 1. Following the conclusions of [11], only data from the earlier half of the growing season was considered for UAV and Sentinel-S2 data.

Table 1 The fields selected for the study in the proximity of Pori, Finland. The thermal time is calculated as the cumulative sum of temperature between the sowing and harvest dates. Mean yield has been calculated from processed yield sensor data for each field.

Field #	Size (ha)	Mean yield (kg/ha)	Crop (*Variety*)	Thermal time (°C)	Sowing date
1	7.59	5157.6	Wheat (*Mistral*)	1316.8	14 May
2	11.77	5534.3	Barley (*Zebra*)	1179.9	12 May
3	7.88	4166.9	Barley (*RGT Planet*)	1127.6	16 May
4	7.24	6166.0	Oats (*Ringsaker*)	1216.4	18 May

Table 2 General information of data sources and their original formats.

Source	Type	Resolution/Step	Multitemporal
UAV	Raster	0.3125 m/px	Yes
Sentinel-S2	Raster	[10,20,60] m/px	Yes
Soil samples	Vector	50 m	No
Veris MSP3	Vector	20 m	No
Topography	Vector	2 m	No
Weather	Tabular	-	Yes
Yield	Vector	Varying	No

The multisource input data for the fields consists of UAV-based RGB images, location data, multispectral Sentinel-2 [1] satellite data, sparsely collected and analyzed soil samplings, machine-collected soil information, topography information and local weather station data. General information about the original data sources are given in Table 2. Some of the data were collected during the growing season of 2018 either manually or automatically, while other data were acquired within one year time difference from the aforementioned season. A total of 39 layers constitute the input data sets, while a single layer, the crop yield, is used as the ground truth. These data are described next and the data layers are numbered for further reference.

2.1.1 UAV

It has already been demonstrated that UAV-based RGB data from the first half of the growing season works better than the data from the second half of the growing season and better than NIR only in crop yield prediction [11]. The UAV data of this study has also been used in [10]. The images were taken at average height of 150 meters with a minimum of three ground control points for geometric calibration. Color correction was performed pre-flight and illumination sensors were used for radiometric calibration. We selected UAV-based RGB data acquired for the first weeks after sowing (weeks 21 to 26 of 2018). Thus, every imaged field has five distinct UAV RGB rasters in the collected data set. The data were acquired with overfligths using a SEQUIOA (Parrot Drone SAS, Paris, France) multispectral camera mounted on a Airinov

Solo 3DR (Parrot Drone SAS, Paris, France) UAV. Field-wise orthomosaics were constructed with Pix4D (Pix4D S.A., Prilly, Switzerland) software. UAV data contains the following layers:

1. Red
2. Green
3. Blue

2.1.2 Sentinel-S2

The Sentinel-S2 satellite data for the fields was acquired from the Copernicus Open Access Hub (European Space Agency, Paris, France). The data were date-matched to UAV images during acquisition, prioritizing images where the algorithmically determined cloud probability was lowest. Thus, five Sentinel-S2 rasters with temporal spacing similar to the UAV data were selected for the data set. With the abbreviated names of product layers in brackets, the Level-2A Sentinel-S2 consists of the following layers:

4. Wavelength 0.443 μm (B01)
5. Wavelength 0.490 μm (B02)
6. Wavelength 0.560 μm (B03)
7. Wavelength 0.665 μm (B04)
8. Wavelength 0.705 μm (B05)
9. Wavelength 0.740 μm (B06)
10. Wavelength 0.783 μm (B07)
11. Wavelength 0.842 μm (B08)
12. Wavelength 0.865 μm (B8A)
13. Wavelength 0.945 μm (B09)
14. Wavelength 1.610 μm (B11)
15. Wavelength 2.190 μm (B12)
16. Aerosol optical thickness at 550 nm (AOT)
17. Scene classification layer (SCL)
18. Water vapour map (WVP)
19. Cloud probability (CLDPRB)
20. True color, red (TCIR)
21. True color, green (TCIG)
22. True color, blue (TCIB)

2.1.3 Soil samples

Soil samples were manually collected from the fields by ProAgria, an agronomic counseling instution, and sent to a Eurofins (Eurofins Viljavuuspalvelu, Mikkeli, Finland) laboratory for further analysis. Soil samples were collected with 50 m steps so that a single sample represented an area of 50 $\times$ 50 m. The samples were collected manually once during November 2018. Being point vectors, the data were rasterized with the `gdal_warp` program of the GDAL utility [17]. Soil sample data contains the following layers:

23. Calcium
24. Copper
25. Potassium
26. Magensium
27. Manganese
28. Phosphorus
29. Sulfur
30. Zink

2.1.4 Veris MSP3

To get a finer map of soil chracteristics, a MSP3 soil scanner (Veris Technologies, Salina, Kansas, USA) was used to map the fields at depths of 0-30 cm and 30-90 cm. The measurements were performed during April and May of 2019. The MSP3 measures the soil's electrical conductivity (EC), which is an indicator of soil compactness, wetness and soil type proportions. Additionally, the instrument measures the pH of the soil. Being irregularly spaced point data initially, data had to be rasterized from point vectors. The rasterization was done with the `gdal_warp` program of the GDAL utility [17]. Each field was measured once. Veris MSP3 data contains the following layers:

31. Shallow EC
32. Deeper EC
33. Ratio, (EC SH / EC DP)
34. Infra-red reflectance
35. Red reflectance
36. Soil pH

2.1.5 Topography

The National Land Survey of Finland conducts light detection and ranging (LiDAR) based elevation mappings on a regular basis in Finland. This data is openly available for anyone to download [2] and contains laser scanned point-cloud data with approximately one point per 2 m^2 [9]. The LiDAR data set was acquired for each of the four fields. The LiDAR data were converted from point-cloud data to spatial rasters using the ArcGIS (Esri, Redlands, California, USA) software. During the conversion, the data were interpolated to match UAV data data in terms of resolution. The topography data contains only the following layer:

37. Elevation information

2.1.6 Weather data

Weather data were collected with two separately located Vantage Pro2 (Davis Instruments, Hayward, California, USA) weather stations. As the fields constitute two distinct clusters, a weather station was placed in the immediate

vicinity of each field cluster. While the stations log multiple variables with a time resolution of just minutes, we utilized accumulated daily statistics and matched data to UAV acquisition dates. Thus, five weather data maps were constructed for each field spacing matching the dates of the UAV data. The weather data contains the following layers:

38. Cumulative temperature sum
39. Cumulative rain sum

2.1.7 Yield data

As the task of regression is that of supervised prediction, the training of the CNN model requires information about the ground truth, the target values. These were acquired during the harvest of 2018 via yield mapping sensor devices attached to the harvesters, either with a CFX 750 (Trimble Navigation, Sunnyvale, California, USA) or Greenstar 1 (John Deere, Molinde, Illinois, USA). CFX 750 utilizes optical sensors to measure yield throughput and moisture. Greenstar 1 utilizes a kinetic mass flow sensor to measure yield throughput and a separate moisture sensor. The yield maps generated by the mapping equipment were initially in the form of vector point-clouds. The irregularly spaced points were filtered prior rasterization to contain only points where the yield was between 1500 and 15000 kg/ha and the harvester speed between 2 and 7 km/h, following the yield pre-processing methodology of [11]. Rasterization was then done by interpolating the yield data to form a raster image.

2.2 Data Preprocessing

2.2.1 Interpolation

The first step after the acquisition of data was to harmonize the spatial resolution across multiple different sources. The UAV data were initially downsampled to 0.3125 m/px, or 32 pixels per 10 meters. This is to match the method of data processing in [11]. Main reasons are to limit the inputs to reasonable size and to have the input dimensions conform to a power of 2 for GPU-based computations. The coarser data, namely Sentinel-S2, soil samples, Veris MSP3, elevation and yield data, required upsampling via interpolation to match this resolution. The interpolation was done using the GDAL utility's `gdal_grid` program with `invdist:power=3:smoothing=20` interpolation algorithm. As with the input data, also the target crop yield data were interpolated to UAV matching resolution. Example results of interpolation are depicted in Figure 1.

2.2.2 Input Feature normalization

After interpolation, the next step was to normalize the data. While absolute values could also be directly used, scaling the input values close to the magni-

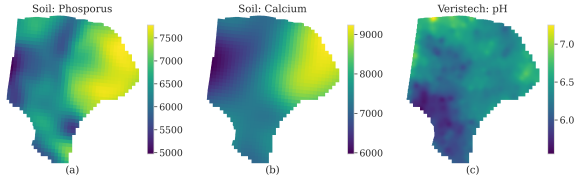

Fig. 1 Examples of input data interpolations on field-scale. (*a*) is the interpolated phosporus map, (*b*) the interpolated calcium content in the field and (*c*) the pH map as measured by the Veris MSP3 soil mapper.

tude of the model's parameters (i.e. weights) helps the model converge faster. Input layers were normalized using a function

$$d^{NORM} = (d - \mu_d)/(max(d) - min(d)), \quad d \in D \tag{1}$$

where d is a layer in the set of all layers D in the data set and d^{NORM} is the normalized layer. However, the target crop yield values were not scaled, akin to [11].

2.2.3 Frame separation

The CNNs require input data to have fixed dimensions. Low number of fields and the irregularities of field shapes led us to extract smaller, fixed dimension frames from the field data. Following [11], we extracted overlapping 40 $\times$ 40 m (128 $\times$ 128 px) frames with 10 m horizontal and vertical steps. Prior extraction, all input and target data from various input sources were aligned in terms of geolocation and resolution to ensure frame extraction from mathching areas. Frames containing half or more valid pixels were included in data, while those having less than half were discarded. This resulted in a total of 16375 input-target frames.

2.2.4 Data sets

Extracted samples were then divided into training, validation and test sets. Training and validation sets were utilized during the training, while the test set was set aside as the out-of-sample performance evaluation data set. As the number of unique fields was low, we wanted to maximize the sample variability the model sees during training. We first attempted to train the models with data separated on a per-field basis with two fields for training, one for validation and one for testing. This led to the model overfitting to the training data and poor generalization performance due to low training data set variability. Similarly low performance was achieved with splitting fields to separate

Table 3 Compositions of training, validation and test sets used to train and evaluate the models.

Data set	Weeks	Frames	Proportion
Training	21,23,25	7561	46.2%
Validation	24	2938	17.9%
Test	22, 26	5876	35.9%

training, validation and test sections. We thus decided then to divide the data temporally to distinct training, validation and test sets according to the UAV image acquisition week. The samples were then shuffled to eliminate spatial autocorrelation in subsequent samples due to overlapping frame extraction. Used weeks, sample counts and sample count proportions for separated sets are given in Table 3.

2.3 Model Architecture

Convolutional neural networks, CNNs, are a subset of spatial model architectures within the broader context of deep learning. CNNs excel in tasks, where the inputs fed to the model are either images or image-like data, i.e. spatial data [14, 7]. While the inner workings of the CNNs has already been well documented [11], we quickly review the operating principles of a CNN. The architecture operates with layers, like many of the deep learning architectures. Each layer is a combination of a convolution operation, which is often followed by a pooling operation. At the heart of the model are the trainable filters of the convolution operation, i.e. the kernels, which produce feature maps for further use.

In our study, we implement and use the best performing CNN architecture of [11]. The model consists of six convolutional layers, followed by two fully connected (FC) layers. Convolutional layers consist of 2D convolutions, batch normalization and non-linear activation with a rectified linear unit (ReLU). First and last convolutional layers also employ max pooling with 2×2 kernel to extract more robust features and reduce intermediate output data dimensions. First five convolutional layers operate with 64 5×5 kernels and the last convolutional layer with 128 5×5 kernels. The outputs of the last convolutional layer are then flattened to a single vector, which is then fed to two 1024 neuron FC layers, both having ReLU activation. Last FC layer outputs the final prediction result. The model was implemented with PyTorch [12] and trained with Skorch [16].

2.4 Training

To gauge the effects of multisource data on the crop yield prediction task with spatial inputs, we performed trainings with four different input data configurations. The data configurations and the input data sources included in

Table 4 The different data configurations used for training distinct models. *RGB Only* uses UAV RGB data only. *No S2* uses UAV, soil, Veris MSP3, topography and weather data. *S2 Raw* adds Sentinel-S2 raw wavelength band data to *No S2*. *S2 Full* adds calculated Sentinel-S2 Level-2A product layers to *S2 Raw*. An X indicates the inclusion of an input data source to a data configuration, while a dash indicates the exclusion.

Source	Channels	RGB Only	No S2	S2 Raw	S2 Full
UAV	1-3	X	X	X	X
Soil	23-30	-	X	X	X
Veris	31-36	-	X	X	X
Topo	37	-	X	X	X
Weather	38-39	-	X	X	X
S2 bands	4-15	-	-	X	X
S2 other	16-22	-	-	-	X
Band count		3	20	32	39

them are further given in Table 4. To elaborate, the derived data configurations were as follows:

- *RGB only.* As [11] was conducted with RGB data from UAVs only, we wanted to make baseline performance evaluation with UAV RGB data only. No other sources were included in this setting.
- *No S2.* Next we wanted to see the effects of soil and weather data on the predictive performance. We thus included all other sources of data (UAV, soil, Veris MSP3, topography and weather) but excluded the satellite data.
- *S2 Raw.* As Sentinel-S2 Level-2A products contain additional algorithmically generated layers, we wanted to see the effect of including just the raw wavelength bands with other input data sources.
- *S2 Full.* The last setting was to use all data acquired for this study.

Because data were distinct from data used in [11], we initialized and trained all models anew for each data configuration. To account for the effects of randomized network parameter initialization, we trained 10 models per data configuration, 40 trainings in total. We used Adadelta [19] as the optimizer, 0.58 for the learning rate, 0.001 for the weight decay and 0.9 for the Adadelta's ρ coefficient as those were the best performing hyperparameters in [11]. Similarly, we used early stopping with a patience of 50 stagnant epochs and continued the training once. The models were trained with Nvidia Tesla V100 Volta and Pascal architecture server GPUs in a distributed computation environment.

3 Results

The CNN models with distinct input data configurations were trained with data of four unique fields. The model architectures, hyperparameters and the training procedures were identical to [11]. As the aim of our study was to evaluate the effects of introducing multisource inputs to crop yield prediction, we trained spatial yield prediction models with four distinct data configurations. The data configurations are discussed in Section 2.4. As the training time loss

Table 5 The test set performance of the same CNN architecture and hyperparameter configuration with various data configurations. *RGB Only* is the baseline model. Out of the configurations, the model performed best with all input data layers (*S2 Full*).

Data Configuration	**Test RMSE (kg/ha)**	**Test MAE (kg/ha)**	**Test MAPE (%)**	**Test R^2 -**
RGB Only	1055.7	838.8	18.2	0.343
No S2	892.4	694.9	14.8	0.531
S2 Raw	461.0	340.9	6.94	0.875
S2 Full	**364.1**	**274.3**	**5.18**	**0.922**

function we used the mean squared error (MSE). Other loss metrics were also calculated, including the square root of the MSE (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE) and the coefficient of determination (R^2). These metrics were not, however, monitored and neither did they not influence model selection during training.

The baseline model using UAV RGB data only attained 1055.7 kg/ha test RMSE, 18.2% test MAPE and 0.343 test R^2. Out of all data configurations, the best performance of 364.1 kg/ha test RMSE, 5.18% test MAPE and 0.922 test R^2 was achieved using all input data presented in our study (*S2 Full*). The performance results for all data configurations with the held out test data set are given in Table 5.

To gain a better view into how the models train with distinct data predicted, we also examined the unseen test sample distributions of predicted values against ground truth values, the true crop yields. With the data, the baseline *RGB Only* model's predictions resemble a Gaussian distribution centered around the mean 5140 kg/ha of true yield values. As more inputs are introduced, the predicted distributions' shapes align with the true values more closely, expressing multi-modal peaks where the true values have them. The test set distributions are depicted in Figure 2.

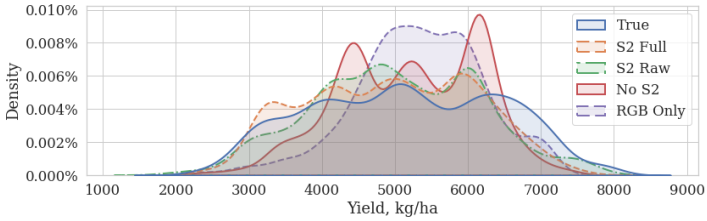

Fig. 2 Distributions of predictions against true yields with the holdout test set.

Table 6 The relative performance of the models trained with distinct multisource input data configurations to the baseline *RGB Only* model. For RMSE, MAE and MAPE negatiThe input data configurations are defined in Section 2.4.

Data Setting	Relative change from *RGB Only*			
	Test RMSE	**Test MAE**	**Test MAPE**	**Test R^2**
No S2	-15.5%	-17.2%	-18.7%	+0.188
S2 Raw	-56.3%	-59.4%	-61.9%	+0.532
S2 Full	**-65.6%**	**-67.3%**	**-71.5%**	**+0.579**

4 Discussion and Conclusions

In this study, we evaluated the effects of using input data from multiple sources on the task of spatial crop yield prediction. Using a CNN model architecture developed for UAV RGB inputs from [11], we introduced additional data from sources like soil samplings, Veris MSP3 soil scanner, topographical maps, weather stations and Sentinel-S2 satellites to the model. We trained ten models for each distinct input data configuration: (1) a *RGB Only* baseline model, (2) a *No S2* multisource model with satellite data excluded, (3) a *S2 Raw* multisource model with raw satellite band data included and (4) a *S2 Full* multisource model with all input data. Out of each set of ten trained models, we selected the models performing best. The model architecture and hyperparameters for the training were taken from [11] and left unchanged to constrain the variability in performance to data only. Only thing varying between model trainings, in addition to four distint input data source configurations, were the initialized model weights.

The performance with larger number of fields using UAV RGB data has already been extensively studied in our previous studies [11] and [10]. Thus, training a model with only UAV RGB data provides a studied baseline to which models trained with additional data can be compared against. The best performing data configuration was *S2 Full* with 364.1 kg/ha test RMSE, 5.18% test MAPE and 0.922 test R^2 using all 39 layers of input data for each extracted frame. Compared to the baseline *RGB Only* model, the *S2 Full* attained 65.6% lower RMSE, 67.3% lower MAE, 71.5% better MAPE and 0.579 higher R^2 with the test set. Generally every model with multisource inputs performed better than the baseline model. This is shown in Table 6.

Crop yield prediction with spatial data and spatial deep learning models has seen an increase in the past few years [6]. Having been studied with a variety of different architectures, from feed-forward networks to hybrid spatio-temporal models, studies have also been conducted with CNN as the main architecture. In [11], a single CNN model was developed to predict crop yields from fields with varying crop types (wheat, barley and oat) from UAV images collected of Finnish crop yields during 2017. Using smaller frames extracted from ortho-images, the best performance was 484.3 kg/ha MAE and 8.8% MAPE. Using soil nutrient data, seed rate, elevation maps, soil's electroconductivity and satellite data in USA, [4] trained a CNN to predict crop yields for nine fields. They report an average scaled MSE of 0.70 which translates to

1145 kg/ha. [18] utilized RGB and multispectral data acquired with a UAV from rice fields in China to predict rice yields with a composite CNN model on field block scale. Feeding the multisource data to distinct, parallelized CNNs, they report a rice yield prediction performance of 0.50 R^2 and 26.6% MAPE.

As we had sufficient data overlap across multiple input sources and the data were acquired from only four unique fields, objective multisource crop yield prediction performance evaluation requires more care in interpreting the results. Relative increase in performance from best performing UAV data utilizing *RGB only* model to the best *No S2* model with additional soil and weather data was notably small. Largest improvements were gained with the introduction of Sentinel-S2 data. Adding raw Sentinel-S2 bands to the RGB, soil and weather data increased the performance by 40.8% RMSE, 42.2% MAE, 43.2% MAPE and 0.344 R^2 from *No S2*. Thus, the increase in performance with Sentinel-S2 is considerably higher than what was achieved with adding soil, topography and weather data to UAV RGB data.

Data acquisition for remote sensing and multisource input data for smart farming is generally laborous and resource intensive. While satellite data is generated automatically, UAVs require semi-autonomous operation at best and the collection of soil data requires extensive on-site manual labour. With more data from a variety of sources a more extensive and representative study can be conducted.

Another limitation stems from differences in spatial and temporal dispersion of different input data sources. UAV, Sentinel-S2 and weather data vary temporally in the data we have used, whereas soil samplings, Veris MSP3 and topographical maps do not. As our data was split temporally to training, validation and test sets, the latter are present in all of these data sets. On the other hand, weather data varies only temporally and constitutes spatial rasters with constant values corresponding to the time of UAV imaging. This means that whether the data is split temporally or spatially, some layer or part of data is always present in training, validation and test sets. As [13] point out, deep learning models are able to implicitly learn linear and non-linear couplings from data with correlations. This means that the deep learning models learn sets of representative features from complex combinations of the inputs and not from single input values on solitude. Furthermore, the performance gains with UAV RGB data combined with temporally invariant soil and ground data is trumped by the performance gains of data configurations using Sentinel-S2 data as additional inputs. This would suggest that the combination of the inputs matters more than presence of distinct, invariant data in training, validation and test sets. However, the concrete effects of simultaneous layer-level data existence in training, validation and test data sets are presently unknown to us and, thus, a subject of future research.

Regarding multisource data in the context of smart farming and crop yield estimation, data itself is an evolving research topic. The use of multisource inputs in remote sensing, while focusing on multispectral data acquired from satellite systems orbiting the globe, has been extensively reviewed in [5]. The use of multispectral data from UAVs and the prediction architectures thereof

is also a developing topic [8]. Another topic related to spatial data is that of autocorrelation [3]. To address autocorrelation of spatial frames in a future study, the inclusion of pixel-wise location information, as suggested in [3], should be sufficient to inform the deep learning model whether data similarity is due to proximity or some other factor or combination of them.

In conclusion, our study indicates that increasing the number of input data sources increases the performance of intra-field crop yield prediction. To draw definite conclusions on the most optimal configuration of input data sources more data is required. With more representative data, generalizable conclusions are more warranted. As the data in this study focuses on a single rowing season, a future plan is to study the generalization of a multisource crop yield prediction model with multiple years of data. Yet in this study the relative increase from baseline of using UAV RGB only as the input data were notable. Consolidating UAV RGB data with soil and ground topology data already somewhat improves the prediction performance, while largest performance gains were gained from using Sentinel-S2 in addition to UAV RGB, soil sampling, Veris MSP3 soil scanner, weather and topography data.

Acknowledgements We would like to thank Mtech Digital Solutions Oy for partially funding this research, Tampere University for providing the computational resources and MIKÄ DATA project for providing us with data.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. ESA: Sentinel-2. URL https://sentinel.esa.int/web/sentinel/missions/sentinel-2
2. PaITuli - Spatial data for research and teaching. URL https://paituli.csc.fi/download.html
3. Amgalan, A., Mujica-Parodi, L.R., Skiena, S.S.: Fast Spatial Autocorrelation. Biometrics **30**(4), 729 (2020). DOI 10.2307/2529248. URL http://arxiv.org/abs/2010.08676 https://www.jstor.org/stable/2529248?origin=crossref
4. Barbosa, A., Marinho, T., Martin, N., Hovakimyan, N.: Multi-stream CNN for spatial resource allocation: A crop management application. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, vol. 2020-June, pp. 258–266 (2020). DOI 10.1109/CVPRW50498.2020.00037
5. Ghamisi, P., Rasti, B., Yokoya, N., Wang, Q., Hofle, B., Bruzzone, L., Bovolo, F., Chi, M., Anders, K., Gloaguen, R., Atkinson, P.M., Benediktsson, J.A.: Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art. IEEE Geoscience and Remote Sensing Magazine **7**(1), 6–39 (2019). DOI 10.1109/MGRS.2018.2890023
6. van Klompenburg, T., Kassahun, A., Catal, C.: Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture **177**, 105709 (2020). DOI 10.1016/j.compag.2020.105709. URL https://linkinghub.elsevier.com/retrieve/pii/S0168169920302301
7. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Communications of the ACM **60**(6), 84–90 (2017). DOI 10.1145/3065386. URL http://dl.acm.org/citation.cfm?doid=3098997.3065386

8. Messina, G., Modica, G.: Applications of UAV thermal imagery in precision agriculture: State of the art and future research outlook. Remote Sensing **12**(9) (2020). DOI 10.3390/RS12091491
9. National Land Survey of Finland: Laser scanning data. URL http://www.nic.funet.fi/index/geodata/mml/laserkeilaus/mml_laserkeilaus_2016_eng.pdf
10. Nevavuori, P., Narra, N., Linna, P., Lipping, T.: Crop Yield Prediction Using Multitemporal UAV Data and Spatio-Temporal Deep Learning Models. Remote Sensing **12**(23), 4000 (2020). DOI 10.3390/rs12234000. URL https://www.mdpi.com/2072-4292/12/23/4000
11. Nevavuori, P., Narra, N., Lipping, T.: Crop yield prediction with deep convolutional neural networks. Computers and Electronics in Agriculture **163**(June), 104859 (2019). DOI 10.1016/j.compag.2019.104859. URL https://linkinghub.elsevier.com/retrieve/pii/S0168169919306842
12. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in PyTorch. In: NIPS-W (2017)
13. Sun, Y., Guo, G., He, X., Liu, X.: Multi-level coupling network for Non-IID sequential recommendation. IEEE Access **7**(Iid), 186247–186259 (2019). DOI 10.1109/ACCESS.2019.2961182
14. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition **07-12-June**, 1–9 (2015). DOI 10.1109/CVPR.2015.7298594
15. Tedesco-Oliveira, D., Pereira da Silva, R., Maldonado, W., Zerbato, C.: Convolutional neural networks in predicting cotton yield from images of commercial fields. Computers and Electronics in Agriculture **171**, 105307 (2020). DOI 10.1016/j.compag.2020.105307. URL https://linkinghub.elsevier.com/retrieve/pii/S0168169919319878
16. Tietz, M., Fan, T.J., Nouri, D., Bossan, B., skorch Developers: skorch: A scikit-learn compatible neural network library that wraps PyTorch (2017). URL https://skorch.readthedocs.io/en/stable/
17. Warmerdam, F., Rouault, E.: GDAL — GDAL documentation (1998). URL https://gdal.org/
18. Yang, Q., Shi, L., Han, J., Zha, Y., Zhu, P.: Deep convolutional neural networks for rice grain yield estimation at the ripening stage using UAV-based remotely sensed images. Field Crops Research **235**(August 2018), 142–153 (2019). DOI 10.1016/j.fcr.2019.02.022. URL https://doi.org/10.1016/j.fcr.2019.02.022 https://linkinghub.elsevier.com/retrieve/pii/S037842901831390X
19. Zeiler, M.D.: ADADELTA: An Adaptive Learning Rate Method (2012). DOI http://doi.acm.org.ezproxy.lib.ucf.edu/10.1145/1830483.1830503. URL https://arxiv.org/pdf/1212.5701.pdf http://arxiv.org/abs/1212.5701