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.

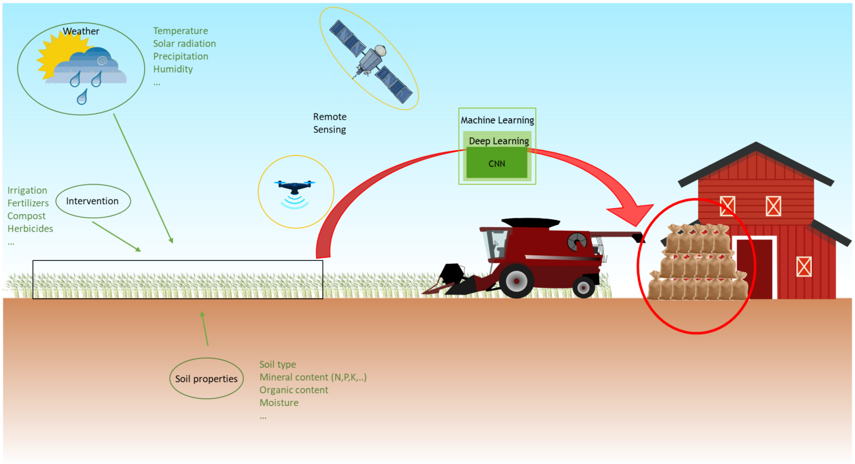

Figure 1. Illustration of crop production as a system with big data input and yield as the output.

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 characteristic 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.

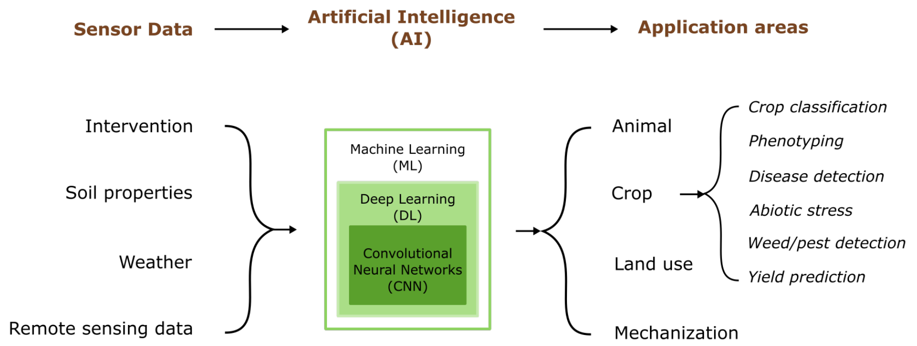

Figure 2. Application areas of DL in agriculture. In this study we focus on crop production, identifying six specific problems that can be targeted using DL models.

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 ($t = \{02{:}00, 05{:}00, 08{:}00, 11{:}00, 14{:}00, 17{:}00, 20{:}00, 23{:}00\}$):

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

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

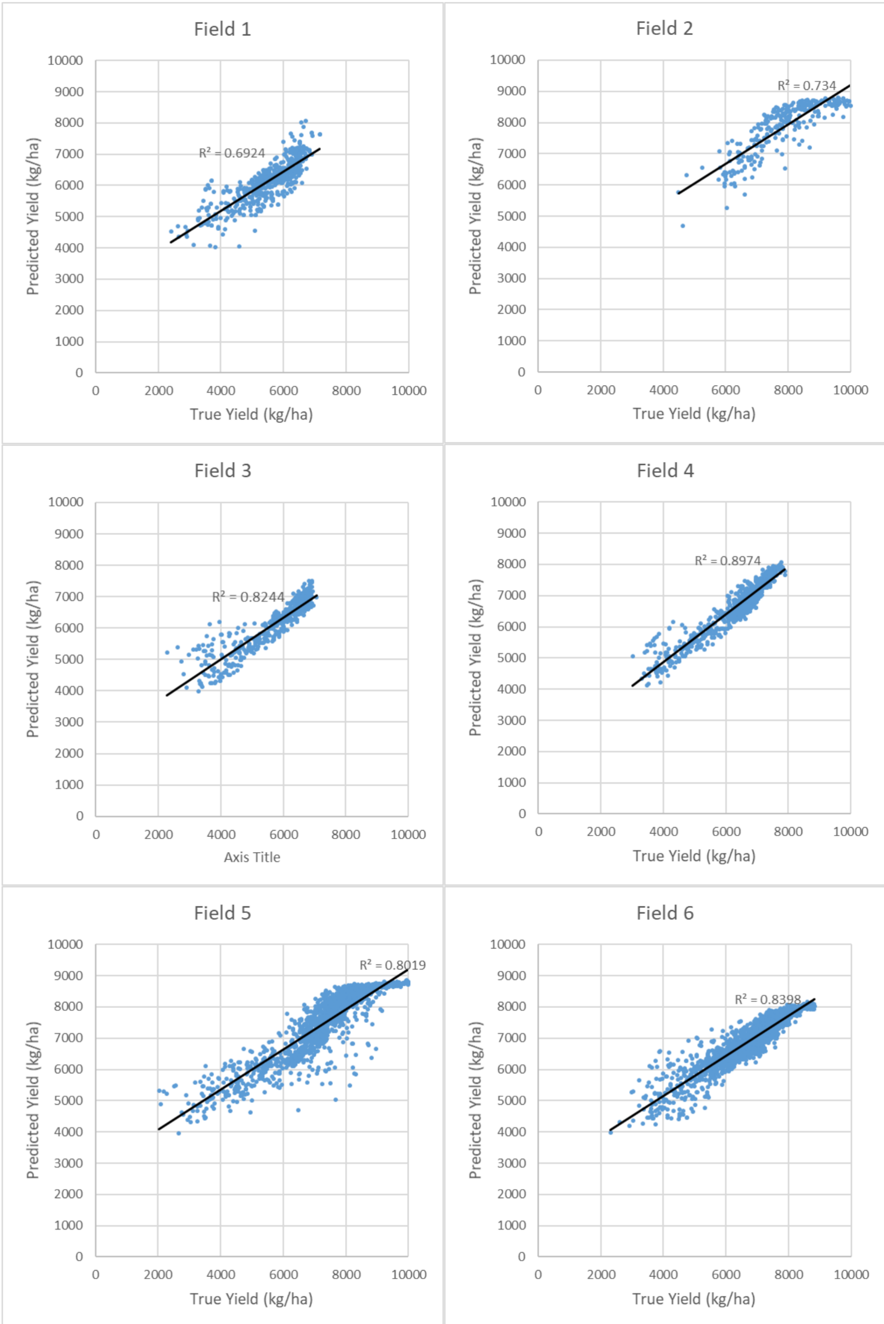

Figure 3. Correlations between the true and predicted yield for each of the 6 fields included in the study.

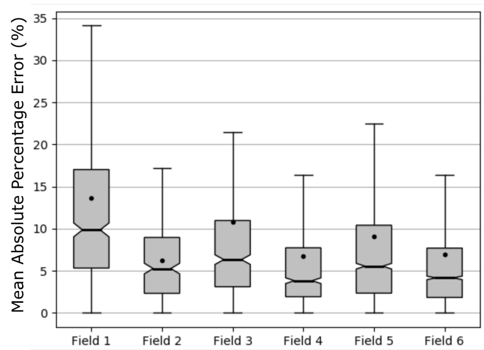

Figure 4. Boxplots of percentage error between true yield and predicted yield for each field.

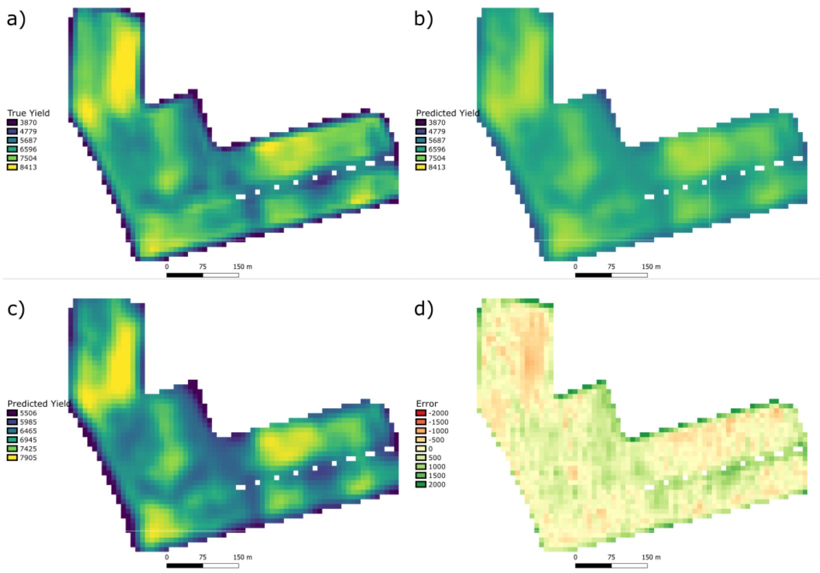

Figure 5. Visualisation of the true and predicted yield of Field 6. a) The spatial distribution of yield as recorded by the yield sensor on the combine harvester. b) Yield predicted by the CNN model. c) Predicted yield with colour-scale adjusted to min-max range. d) Error between predicted and true yield.

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.

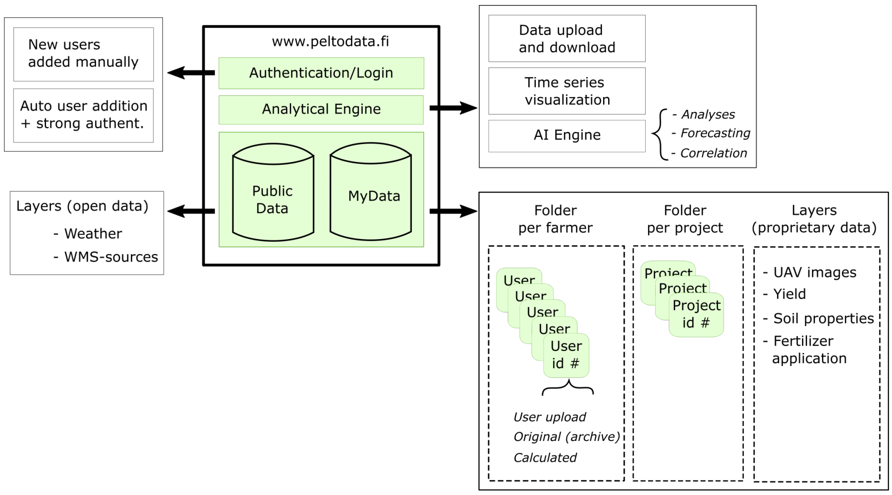

Figure 6. Structure of the Oskari based decision support platform for farmers.

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.