Project title:	The Augmented Agronomist: Synthesis of Privacy-Preserving Neural Networks and Robotics to Assist Decision Support
Project number:	SF/TF 170
Project leader:	Marc Hanheide, University of Lincoln Georgios Leontidis, University of Aberdeen
Report:	Annual report, 2020
Previous report:	Annual report, 2019
Key staff:	-
Location of project:	University of Lincoln
Industry Representative:	Richard Harnden, Berry Gardens
Date project commenced:	30 November 2018

DISCLAIMER

While the Agriculture and Horticulture Development Board seeks to ensure that the information contained within this document is accurate at the time of printing, no warranty is given in respect thereof and, to the maximum extent permitted by law the Agriculture and Horticulture Development Board accepts no liability for loss, damage or injury howsoever caused (including that caused by negligence) or suffered directly or indirectly in relation to information and opinions contained in or omitted from this document.

All other trademarks, logos and brand names contained in this publication are the trademarks of their respective holders. No rights are granted without the prior written permission of the relevant owners.

AUTHENTICATION

We declare that this work was done under our supervision according to the procedures described herein and that the report represents a true and accurate record of the results obtained.

George Onoufriou

PhD student

University of Lincoln

Signature Date 2020-10-30

Report authorised by:

Marc Hanheide

Professor of Intelligent Robotics & Interactive Systems

University of Lincoln, Lincoln Centre for Autonomous Systems Research

Signature  Date 2020-10-30

Georgios Leontidis

Senior Lecturer in Computer Science

University of Aberdeen

Signature  Date 2020-10-30

CONTENTS

GROWER SUMMARY

Headline

To provide automated agronomy support for agronomists at scale using machine/ deep learning techniques for yield prediction, from high dimensional spatio-temporal data.

This approach will reduce costs whilst maximizing specialist human time in areas that require the most attention.

Background

This work on the augmented agronomist has been undertaken to help focus human time to the most vital areas, and act as an arm for agronomists to help locate problem areas in the crop, and improve yield prediction earlier than possible before. This system is also being created to improve trust, and security around the usually enigmatised deep learning models, and ensure data owner's privacy.

Summary

Over the course of this project we intend to complete the following key objectives:

- **Provide agronomists and agriculturalists with yield predictions.** This is the primary advantage provided by the augmented agronomist system which will provide alerts to the operator of deviations from forecasts, and highlight areas where predicted yield potential is not on target. This information will enable the operator to focus efforts in areas which require most attention in order to maximise yield potential.
- **Create an autonomous data collection system**. Hand collecting data at scale would be infeasible due to both time and cost investments being too high while also providing inconsistent results. We will develop a repeatable and autonomous data collection platform so that we can collect spacio-temporal data for yield consistently and at scale.
- **Create a data aggregation and utilization pipeline.** This pipeline will be designed to be able to handle distributed autonomous data collections, which is the most likely scenario faced in practice, such as multiple robots operating and feeding in their data simultaneously across multiple sites.
- **Deploy an agronomy assistive neural network to predict plant yield ahead of harvest.** Ultimately this project will culminate with several other concurrent projects to develop an autonomous data collection, and actuation platform (Thorvald), to collect, process, and act on the data.

Financial Benefits

According to Berry Gardens Growers (BGG) in 2018 they over estimated crop yield by 17.7% for 14 weeks of the 30-week growing period and underestimated the remaining 16 weeks by 10%, giving them an average absolute error (MAE) of 13.6% for the whole season. Underestimates cause surpluses, yield devalue, and subsequently costs by additional disposal of the yield. Additionally, over-estimates mean to meet demand, and contracts, growers will need to resort to expensive imported fruit, to cover the shortfall. In 2018 this cost BGG roughly 8 million pounds, whereas losses to the rest of the industry (70%) are estimated to cost 18 million pounds.

Current literature of deep learning enabled yield prediction expects an error (MAE) of roughly 15% (Konstantinos et al 2018; Maimaitijiang et al 2020). We can already match this and roughly the MAE of BGG using purely environmental data. We hope with the additional layers of image data, and more granular time series data that we can further improve upon this error, preventing further losses. We also hope to reduce the spread of inaccuracy, compared to purely human predictions, since human inaccuracy can vary wildly from person to person, and day to day, even if overall it gives a cumulative error of 13.6%.

Action Points

- There are no action points at this time.

SCIENCE SECTION

Introduction

Machine/ Deep learning is becoming a bigger and more important part of our daily lives through the rise of an ever-increasing quantity of available data. The use of machine learning in combination with user data is becoming increasingly widespread and impactful in everyday society performing tasks ranging from, natural language processing (Do et al. 2019), image recognition, diagnosis (Biswas et al. 2019), detection, classification (Fawaz et al. 2019), medical diagnosis (Anderson et al. 2019), self-driving cars (Huval et al. 2015), facial recognition (G¨uera and Delp 2018) among many other examples. However, one area with which deep learning has remained relatively underutilised is in agriculture, where the data is scarce. The existing research has relied on classical techniques, and relied on remote sensing datasets and to date has not taken advantage of the recent advances such as generative adversarial networks (GANs; Alvarez 2009; Chlingaryan, Sukkarieh, and Whelan 2018; Prasad et al. 2006) The primary reason why agriculture has not innovated in this area for so long is likely to be the lack of consistent data, but also the lack of willingness and trust of the growers/ agriculturalists to release data which could compromise their competitive advantage. Thus, if there is little to no data there can be little advancement with deep learning techniques, meaning we will likely have to collect our own data to find any meaningful relations between the features and targets with which to predict yield accurately and far enough ahead to facilitate timely and effective actions. We have access to a plethora of data collection possibilities including; through the RASberry project, a collaboration between University of Lincoln (UoL), Saga Robotics, and Berry Gardens Growers (BGG), funding autonomous strawberry data collection; and through members of the consortium which fund the Collaborative Training Partnership Fruit Crop Research (CTP-FCR) studentship programme. The involvement of Saga Robotics (SR) allows us to work on a common generic expandable robotic platform called Thorvald (Figure 1). Thorvald is an autonomous robot ready for use in many terrains and an ideal candidate platform to use for our own data collection and usage system thanks to its autonomy, funding, and available resources. The seasonality of strawberry crops mean that fruit is only available for harvest between late June to early October in our experimental system.

Figure 1: Thorvald robot adjacent to strawberry tabletop plantation at the University of Lincoln Riseholme campus. This Thorvald is equipped with 3 realsense cameras; RGB, depth, one Raspberry Pi Zero, one raspberry pi camera v2, one BME680; temperature, humidity, and air- quality sensor

Materials and methods

Initially we had to create a system to be able to attain our data before we could do anything with it. This should also serve as the basis of the augmented agronomist. In collaboration with the Lincoln Centre for Autonomous Systems (LCAS), and Saga Robotics who produce the Thorvald an autonomous data collection platform was developed.

- The creation of the data collection pipeline is a collaborative process in which this project developed security/ encryption, databases, and deep learning the autonomous data capture and worked closely with a specialist (Raymond Kirk, 3rd year CTP-FCR PhD student) in autonomous robotic control using the robot operating system (ROS), Thorvalds, and deep learning. Thus, some of the work reported herein is a result of this collaborative effort due to the cross disciplinary skill sets required. The work has been completed in several work packages: Data Collection; robotic control, pathing, orchestration, and data capture.
- Data Management; data aggregation/ pipelines, Deployment (dockerisation), MongoDB distributed replica sets and networking.
- Data Processing.
- User interface for feedback and control.
- Privacy preservation.

Results

Owing to the volume of work here, and to keep things concise this section will outline first the objective and then our findings and methods. There will primarily be a focus only in areas where this PhD is principally concerned:

- Data collection planning; Explore the production system (strawberries grown on tabletop), and determine what intricacies the system may have before data collection. An interesting aspect was the effect that insects have on the outcome/ yield such as the burrowing of wasps in the strawberries and pest damage more broadly. There initially appeared little way to be able to predict or account for sudden surges in yield loss due to wasp damage, but we found that wasps would only burrow into slightly or completely overripe strawberries (Figure 2). This is a direct consequence of poor picking, as any unpicked strawberries over ripen and attract wasps, which is something we could potentially see in the dataset, and could reasonably predict.

Figure 2: Wasp burrowing behaviour, eating unpicked overripe strawberries, which is not immediately apparent as it could easily be mistaken for a ripe strawberry

- Automatically traverse the strawberry tabletop. This is primarily thanks to the work of Raymond, that the robot traverses the strawberry tabletop safely and consistently.
- Automatically and repeatedly collect image data at set intervals down the row from the robot. As above. Collect data suitable for as many of our associated project needs as possible. For this project the key requirements were environmental factors such as temperature, humidity, strawberry images over time, traversal images, irrigation and many more.
- Create sensors appropriate for data collection. Now that we know what data we require we could decide upon what sensors and other constituents the system required to gather these. In

the case of this project a need for a sensor to collect temperature, and humidity resulted in use of a BME680 which is a small, cheap, I^2c gpio board that integrates well into small ARM boards such as the Raspberry Pi. In the absence of accessible GPIO pins on the Thorvald system a compartmentalized system for data collection was developed. The many iterations of this can be seen in figure 3. The sensors gradually grew in number, and became more efficient with regards to space, and power.

- Store the data locally on the robot platform due to difficulties transmitting large amounts of data wirelessly from inside the polytunnels. In previous projects data has been stored as files on the filesystem making them difficult to access, maintain, make consistent between multiple machines, and are unindexable/ unsearchable, resulting in a lot of wasted time. This project used MongoDB, which provides a common database to store all the data captures.
- Automatically collect intensive image and sensor data from select few plants in the row, every 15 minutes, day and night, using a stationary raspberry pi (Figure 4), resulting in images depicted in figure 5.
- Automated data syncing; Now that we have a database system we could use this for near-realtime data syncing. This makes it possible to do almost live updates, and predictions, for immediate problem identification, and response.

As a result of quite some monumental effort to create a system suitable for both the requirements of the funding body, and the near-realtime data provisioning of this PhD, we created an automated Thorvald data collection platform, gathering many different varieties of spatio-temporal data, while maintaining their consistency, availability, and fault tolerance securely. The data collected key features such as humidity, temperature, light intensity, RGB images of the strawberry plants and fruit from multiple angles, as well as depth, and positional information, to list a few. We have employed 3 different data collection methods, including an intensive stationary sensor array to capture images and sensor data every 15 minutes of a few select plants to model them more precisely. We have also collected continuous capture of images, and sensor data to create sensor-array-video. Lastly the main data collection uses 20 cm step sampling of the strawberry tabletop to model the complex intra-crop environmental changes, which the vast majority of other datasets are not granular enough to do. We can pinpoint exactly at which position the robot was during any given data point, what its orientation was, and all the associated sensor array readings.

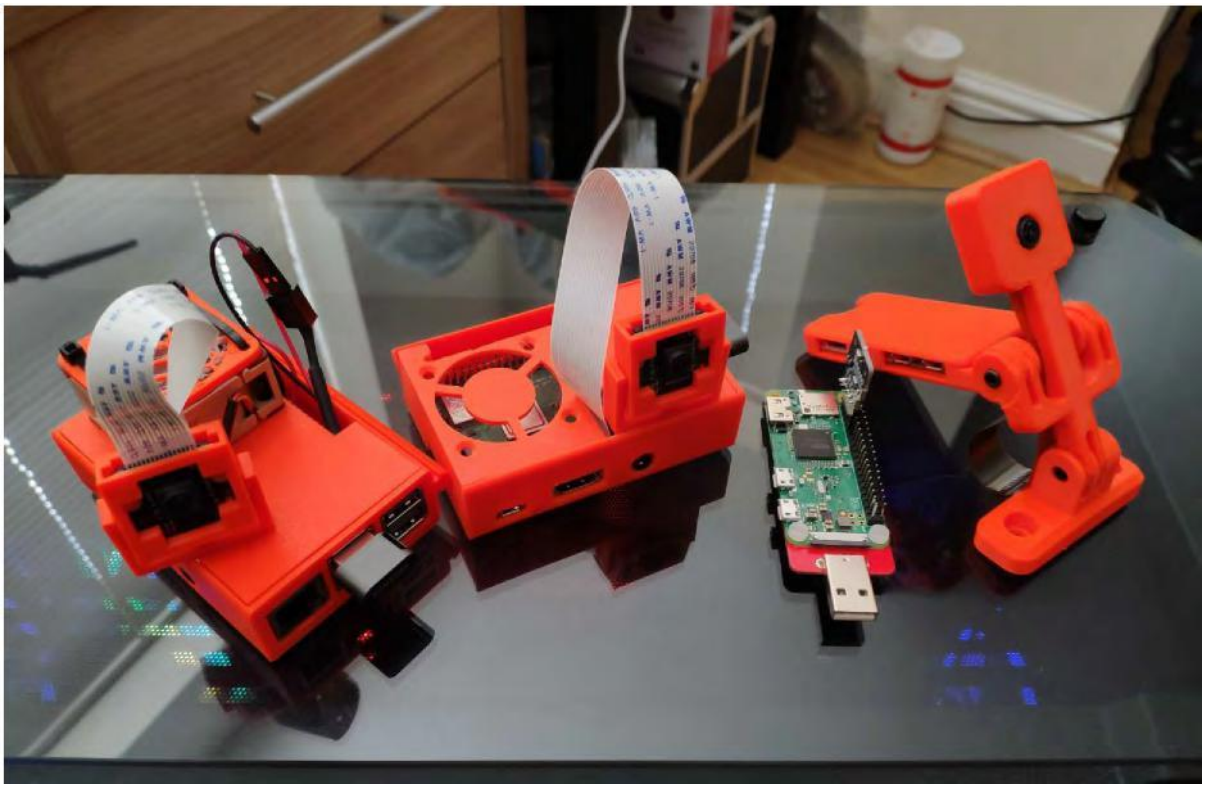

Figure 3: Plethora of cameras, and sensors created over the course of data acquisition stage. The majority of which had specially designed, and 3D printed protective polyethylene terephthalate glycol (PETG) casings which doubled as mounts to be able to attach to the Thorvald robots. PETG is also highly UV and weather resistant.

Figure 4: Raspberry pi stationary camera, used separately to the Thorvald robot, used to intensively monitor fewer strawberries over a longer period.

Figure 5: 4 Raspberry pi stationary camera images, with one enlarged. The latter 2 of the 4 depict drooping strawberry stems, from a lack of irrigation.

Data Management

Data management requires the development of a consistent, reliable, and available method to gather data from the pipeline developed above that is potentially distributed between not only different Thorvald robots, but also multiple data collection sites. This work package is summarised as below:

- Aggregate the data. One of the primary reasons for using MongoDB during data collection was an awareness of the distributed, sharding/ replica set functionality ingrained in MongoDB along with its ease of use and security when properly configured (Figure 6). MongoDB was found to be capable of automated data distribution such that each robots database would synchronise its contents with a larger more powerful network of MongoDB shards that was distributed using our Nemesyst framework (Onoufriou 2019a) (Figure 7).
- Back up the data and add redundancy. Simultaneously to aggregating the data from each of the robotic platforms these shards, and replica sets provided guaranteed redundancy

should any individual or even multiple datasets fail, with an automatic voting system to adjust which remaining replica set becomes the leader/ master. Using a logged database such as MongoDB on a simultaneously journaled file-system such as BTRFS or EXT4 meant that almost any data can be recovered that has been removed, deleted, corrupted or otherwise.

- Make the data accessible to others. Since many other projects are interdependent on this data, we implemented a key based user authentication mechanism, along with in place encryption, transport layer security, server authentication, replica set cross authentication and encryption.

Figure 6: MongoDB, and monitoring. This allows us to continuously be aware of the state of the replica sets, their load, who is the master/ primary, network traffic, etc. This is an invaluable resource to the management of such a system and will provide us with many useful tools to analyse the effects of our research on this distributed implementation.

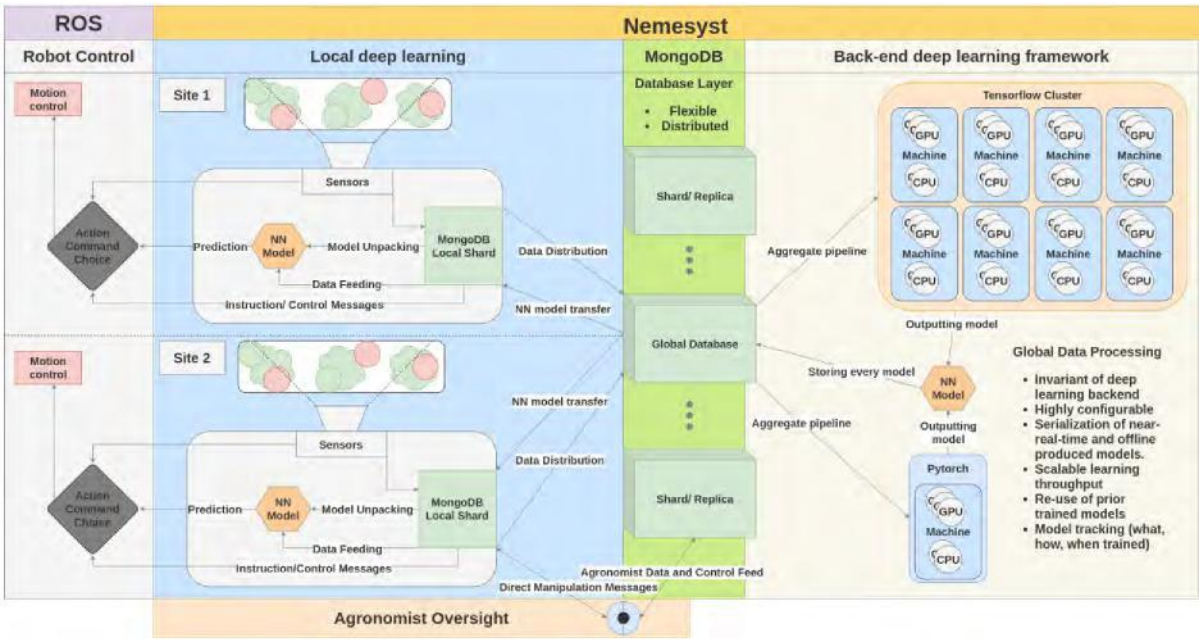

Figure 7: Original architecture for data capture, management, control, processing, and feedback.

Data Processing

Data that now resides in our databases is still raw, and unprocessed, thus we need to organise it to make it usable.

Firstly we “wrangled” the data, which here shall be the process of cleaning, normalising, and formatting the data, in such a way as to make it useable by neural networks. To this end we created automated wrangling scripts, which can operate continuously since our data can be streaming near-realtime at almost any point in time from the fields.

Broadly this process; turns categorical variables like cardinal wind direction into one hot encoded vector, scales data between 0-1 for consistency while also where neural networks operate and learn fastest, transform date and time to ISO standard datetime objects, among many other data wrangling steps.

Our first implementation primarily uses environmental factors (humidity, temperature, light intensity, etc), and yield (punnets) features to learn the relationships between the environment and its consequence on yield. The data was split randomly into training and test sets (70:30), but unfortunately due to a scarcity of yield values we could not use a further validation set as this would exasperate the effect of this too little data on training our models, leaving them under-trained.

Our second more comprehensive implementation which improves upon the first by using spacial data like images, along with environmental factors (just like the first), irrigation data, and yield is

now specifically in kg to be more consistent than punnets. We continue to use a 70:30 training test split, although we are now able to use cross validation with our increased yield data. This is still ongoing work, which we hope to have the results of soon.

User Interface

The augmented agronomist requires an easy-to-use user interface to allow agronomists to inform and direct the neural networks, such as in uncertain, or problematic scenarios. This way we can utilize the expertise of the agronomists to improve the neural networks, while also providing them with a tool with which to become alerted in places that require their attention, without having to do laborious checking. This will maximise their time in only key areas. This will also give us a good way to display graphs and other information to the agronomists, such that we can augment the agronomists. This is still ongoing to create the full application but an example is depicted in figure 8.

Figure 8: First example dashboard for uploading, previewing and infering from an uploaded dataset.

This user interface/ web app will also serve to centralise the algorithms, so they have one unified location, to make managing such a complex project easier, while also being a secure way to distribute, and collate information.

Privacy Preservation

We cannot talk in depth here, as this is ongoing work which is critically novel. However, we seek to ensure data collected cannot compromise the data owners, while still allowing data to be processed.

Results and Discussion

Thus far through the PhD we have collected, and aggregated in our first year pipeline:

- 35033 rows of weather data, or every 15 minutes for the year following 2019-01-01
- 23 records for number of punnets produced over the season

- 2200 intensive records of strawberry growth in the span of a week, including RGB, depth, infrared, location data, environmental data adjacent to the plant.
- 1503 (in progress annotation of each ripe and unripe strawberry) images every 20cm down each row, with the same features as above.
- 188 top down images of strawberries, with the same features as above.
- 188 bottom up images of strawberries with the same features as above.
- 5200 camera footage going down each row.

In our second year we collected and aggregated a much-improved collection of data automatically in spite of coronavirus:

- 360 records for the wet mass yield of the strawberries of the season, along with associated losses due to mechanical, deformation, or disease factors, on a per-row basis. (60.6 KB)
- 14,500 stationary camera timelapse images, showing growth over select plants intensively over the entire season. (32.4 GB)
- 110,700 images taken in sets of 3 as the robot traverses the tabletop, every 20 cm. (460 GB)
- 25,800 records for environmental factors such as temperature humidity, windspeed, etc recorded by the onsite weathervane every 15 minutes, for the whole year of 2020 up untill the end of the growing season. (5 MB)

Our initial implementation of 3 different neural networks have given us the following resulting MEAs:

Table 1: Original/ first neural networks model evaluation, where GRUs significantly outperform other recurrent neural networks on this data.

Neural Network Type	Mean Absolute Error (MAW) on Test Set
Recurrent Neural Network (RNN)	0.208
Long Short-Term Memory Network (LSTM)	0.293
Gated Recurrent Unit (GRU)	0.142

We are already working on expanding on this implementation with our new complete second year data (see: data processing), that is much more granular, consistent, and includes more features than previously requiring an overhaul of our previous techniques. We are also layering on image-based features, which we aim to reduce the MAE to sub 10%. Subsequently we intend to use deeper and much more complex neural networks such as Generative Adversarial Nets (GANs) to leverage this increased data quantity and quality.

Lastly we have also begun testing of our privacy preserving methods, and found them to be negligibly different in error compared to non-private models and data, although we are not prepared to disclose this in great detail publicly yet since it must be critically novel.

Conclusions

In summary we have created a completely new, distributed, and near-realtime autonomous data collection platform so that we can evaluate yield prediction at arbitrary scale. We have iterated on our existing database enabled deep learning framework (Nemesyst) with the generalised functionality missing that was required for this application, and published a paper on its application to a similar problem. We have created the necessary data wrangling and neural network predictors which we continue to improve. We are integrating all our work into an easy-to-use web application so neural networks can be directed by the agronomists easily. We can already match predictive accuracy of pre-existing papers while publishing our own conference paper, and are looking to leverage our dataset to surpass the MAE of 10%, to surpass human alone prediction consistently.

Knowledge and Technology Transfer

Towards furthering the science, we have:

- Further expanded our open source, and permissive, distributed deep learning framework for other to benefit from our work (DreamingRaven/nemesyst).
- Published a journal paper on Nemesyst under similar conditions (Onoufriou 2019a).
- Published a conference paper on our initial neural network implementation (Onoufriou 2020).

References

Alvarez, R (2009). “Predicting average regional yield and production of wheat in the Argentine Pampas by an artificial neural network approach”. In: European Journal of Agronomy 30.2, pp. 70–77.

Anderson, Rachel et al. (2019). “Evaluating deep learning techniques for dynamic contrast-enhanced MRI in the diagnosis of breast cancer”. In: Medical Imaging 2019: Computer-Aided Diagnosis. Vol. 10950. International Society for Optics and Photonics, p. 1095006.

Biswas, M et al. (2019). “State-of-the-art review on deep learning in medical imaging.” In: Frontiers in bioscience (Landmark edition) 24, pp. 392–426.

Chlingaryan, Anna, Salah Sukkarieh, and Brett Whelan (2018). “Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review”. In: Computers and electronics in agriculture 151, pp. 61–69.

Do, Hai Ha et al. (2019). “Deep learning for aspect-based sentiment analysis: a comparative review”. In: Expert Systems with Applications 118, pp. 272–299. Ershov, Mikhail (2015). Survey of Algebra. url: http://people.virginia. edu/~mve2x/3354 Spring2015/ (visited on 11/10/2019).

Fawaz, Hassan Ismail et al. (2019). “Deep learning for time series classification: a review”. In: Data Mining and Knowledge Discovery 33.4, pp. 917–963.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). Deep learning. MIT press. Goodfellow, Ian, Jean Pouget-Abadie, et al. (2014). “Generative adversarial nets”. In: Advances in neural information processing systems, pp. 2672–2680.

G¨uera, David and Edward J Delp (2018). “Deepfake video detection using recurrent neural networks”. In: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, pp. 1–6. Hayes.

Jamie et al. (2017). “LOGAN: evaluating privacy leakage of generative models using generative adversarial networks”. In: arXiv preprint arXiv:1705.07663. Huval, Brody et al. (2015). “An empirical evaluation of deep learning on highway driving”. In: arXiv preprint arXiv:1504.01716.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). “Deep learning”. In: nature 521.7553, pp. 436–444. 15 Liakos, Konstantinos G et al. (2018). “Machine learning in agriculture: A review”. In: Sensors 18.8, p. 2674.

Niedba la, Gniewko (2019). “Application of Artificial Neural Networks for MultiCriteria Yield Prediction of Winter Rapeseed”. In: Sustainability 11.2, p. 533.

Onoufriou, George (2019a). Nemesyst documentation. url: https://nemesyst. readthedocs.io (visited on 11/10/2019). — (2019b).

Nemesyst; Generalised and highly customisable, hybrid-parallelism, database based, deep learning framework. url: https : / / github . com / DreamingRaven/nemesyst (visited on 11/10/2019).

Onoufriou, George et al. (2019). “Nemesyst: A Hybrid Parallelism Deep LearningBased Framework Applied for Internet of Things Enabled Food Retailing Refrigeration Systems”. In: arXiv preprint arXiv:1906.01600.

Prasad, Anup K et al. (2006). “Crop yield estimation model for Iowa using remote sensing and surface parameters”. In: International Journal of Applied Earth Observation and Geoinformation 8.1, pp. 26–33.

Rahnemoonfar, Maryam and Clay Sheppard (2017). “Real-time yield estimation based on deep learning”. In: Autonomous Air and Ground Sensing Systems for Agricultural Optimization and Phenotyping II. Vol. 10218. International Society for Optics and Photonics, p. 1021809.

Wang, Anna X et al. (2018). “Deep transfer learning for crop yield prediction with remote sensing data”. In: Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies. ACM, p. 50. You, J iaxuan et al. (2017). “Deep gaussian process for crop yield prediction based on remote sensing data”. In: Thirty-First AAAI Conference on Artificial Intelligence.

Liakos, Konstantinos G., Patrizia Busato, Dimitrios Moshou, Simon Pearson, and Dionysis Bochtis. "Machine learning in agriculture: A review." *Sensors* 18, no. 8 (2018): 2674.

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

Onoufriou, G., Hanheide, M., Leontidis, G., (2020). The Augmented Agronomist Pipeline and Time Series Forecasting. UKRAS20 Conference: “Robots into the real world” Proceedings, 117119.

