Project title:	The Augmented Agronomist
Project number:	2155898, BB/S507453/1
Project leader:	George Onoufriou, University of Lincoln
Report:	Annual report, 2021-10
Previous report:	Annual report, 2020-10
Key staff:	Georgios Leontidis, University of Lincoln Marc Hanheide, University of Lincoln Mark Else, National Inst of Agricultural Botany
Location of project:	University of Lincoln Brayford Pool, Lincoln, LN6 7TS, England
Industry Representative:	Richard Harnden, Berry Gardens Growers Ltd, Unit 20 Wares Farm ME17 4BA
Date project commenced:	2018-11-30

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.

[The results and conclusions in this report are based on an investigation conducted over a one-year period. The conditions under which the experiments were carried out and the results have been reported in detail and with accuracy. However, because of the biological nature of the work it must be borne in mind that different circumstances and conditions could produce different results. Therefore, care must be taken with interpretation of the results, especially if they are used as the basis for commercial product recommendations.]

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 Candidate

University of Lincoln

Signature ..... Date .....

Report authorised by:

Georgios Leontidis

Associate Professor in Machine Learning

University of Aberdeen/ University of Lincoln

Signature ..... Date .....

Marc Hanheide

Professor of Intelligent Robotics and Interactive Systems

University of Lincoln

Signature ..... Date .....

Mark Else

Plant and Fruit Physiologist

National Institute of Agricultural Botany, East Malling Research

Signature ..... Date .....

CONTENTS

GROWER SUMMARY

Headline

Using *machine learning/* artificial intelligence and (most likely) *existing data sources* describing environmental conditions like temperature and humidity, we can *accurately and completely privately predict strawberry yield weeks in advance*. This predictive performance ranges from 8% to 22% depending on the difficulty of the specific scenario being predicted for. This *performance improves as more data becomes available* which is necessary to be able to use more *advanced neural networks*. We produce these predictions without decrypting the input data at all, and returning a merely transformed cyphertext privately.

Background

Inaccurately forecasting fruit yields like strawberries directly causes issues such as fruit waste, and insufficient production to meet contractual obligations. Fruit waste through under-prediction leaves a surplus of fruit which then either has to be sold at a reduced rate or has to be destroyed. Over prediction leads to the opposite problem of having insufficient fruit to meet contractual obligations such as those agreed weeks in advance to retailers. This often means alternative supply has to be sought often from abroad to cover this shortfall, since conditions that likely caused this shortfall are likely to affect geographically adjacent producers in the same country, further exasperating the additional costs. Currently yield forecasting varies wildly from grower to grower which in recent years has been, at worst ±50% of the actual yield.

Summary

Deep learning, a subset of machine learning and artificial intelligence is a state-of-the-art predictive technology where there is data available to train/ use it. Often the more *good* data available the better the predictive performance becomes. However data availability is often scarce. This project sets to do two things primarily; To show how deep learning can provide an invaluable accurate predictive service to both growers and agronomists. To lower barriers to data sharing to enable even the most privacy concerned growers to share data to gain a benefit from these state-of-the-art neural networks, using state-of-the-art quantum resistant cryptography like fully homomorphic encryption. This way we can provide accurate predictions using likely already existing data sources, improve on food waste and costs, and maintain complete privacy for the growers and their methods. We exemplify this on strawberries but fundamentally is agnostic of this specific application. If there is data that describes a relationship well, then a neural network can be taught to predict it, privately.

Financial Benefits

It is extremely difficult to quantify the specific financial benefits of such a predictive system without specific context with which to apply. However in the strawberry domain we have been provided estimates at 26M/year losses across the UK industry due to the aforementioned over and under-prediction of strawberry yields, even with agronomists who usually perform within ±17%.

The costs of improving this yield prediction, and thus reducing losses amounts to expertise, as the hardware and data necessary are likely already existing. In the likely scenario that machine learning expertise does not exist in-house to operate these models, then external expertise is necessary. This could take the form of hiring a machine-learning engineer to setup on-site data processing. Or hiring a third party for continued deep learning as a service. However at this time only we are likely to be able to provide encrypted deep learning as a service due to its high barriers to entry, if privacy is of primary concern.

Action Points

The primary action points to be able to exploit deep learning/ machine learning are the existence of data and expertise. Even if the intention is not to utilise this specific system data is invaluable, and should be recorded wherever possible to be able to utilise it in the future. The primary factors that relate to good usable data, are consistency, representativeness, and granularity. Collecting good consistent data usually means automation, to have sensors on site collecting environmental conditions such as temperature, light intensity, humidity continuously without/ as few gaps/ breaks as possible, over as long a period as possible, preferably years.

For data to be representative then it should be directly related to the outcomes being predicted, in the case of strawberries that usually means locality. The sensors should collect data as representative of what is experienced by the strawberries as possible. MET office data for example is usually too distant and not representative of the climate the berries have specifically experienced and been affected by. Granularity is effectively the descriptiveness of the data, having more than one sensor in multiple locations makes the data more granular to the climate conditions especially across larger sites. The regularity of the sensor taking readings also improves granularity as it helps describe the environment over time in more detail, giving machines more information with which to base their predictions on.

SCIENCE SECTION

Materials and methods

Using our existing poly-tunnel facilities at the University of Lincoln Riseholme-campus we collect our own data during the 2021 growing season of strawberry yields, environmental, and spacial data. We use two 20m long, 5 row wide strawberry tabletop in each poly-tunnel. We grow two varieties of strawberry, Katerina in one poly-tunnel for all 5 rows, and Zara in the other poly-tunnel for the remaining 5 rows. Over the season we collect irrigation/soil based data variables; time, moisture, soil temperature, run-off, power, dripper input, and various other triggers for the irrigation system. We collect environmental based data; time, barometer, temperature, humidity, dew point, wind-speed, uv-index and dosage, wind-direction, precipitation, solar radiation, and redundant other variables for average calculation. Lastly we collect spacial based data; automatic robot imaging of the strawberry crop every 20 cm, stationary time-lapse footage of a section of crop in the Zara poly-tunnel in 15 minutes intervals, spacial location data of strawberries, quantity, time, and location of strawberry yield by weight.

We stream and store this data into MongoDB databases mostly automatically with respective backups to allow easy access and recovery. Now that we have the ground truth we can begin data wrangling of each respective category of data. Primarily we transform the data into machine friendly numeric values, and one-hot encoding, while normalising between 0 and 1 using a linear rescale to enable the machine learning to learn faster and more repeatedly.

This year we focused on privacy preservation of our neural networks, thus we created the necessary tools towards this end. Namely Python-FHEz our encrypted deep learning library.

This library uses the state-of-the-art Residue Number System (RNS) modified Cheon, Kim, Kim, Song (CKKS) or RNS-CKKS scheme provided by Microsoft Simple Encrypted Arithmetic Library (MS-SEAL). This Cryptosystem allows for computation of fixed-point precision arithmetic on fully homomorphically encrypted cyphertexts directly, without the need to decrypt the data, thus ensuring privacy. We rebind this C++ based cryptosystem into python using PyBind-11, and create our own graph based representation, traversal, computation, and automatic parametrisation methods.

This year due to using encrypted deep learning we chose to create our graphs to be a 1 dimensional (1D) convolutional neural network (CNN). To begin with we used a single layer 1D-CNN followed by a dense layer to combine and rescale the CNN output. We married this dense network with a mean squared error (MSE) loss calculation circuit to be able to back-propagate the loss and train the network. We implemented the adaptive-moment (adam)

algorithm for the back-propagation update phase since it is the most favoured in the most cases. We implemented FHE compatible variants of certain neural network components such as Sigmoid and Rectified Linear Unit (ReLU) activation functions.

Now that the graph/ neural network is complete we feed in the cleaned and normalised data, in batches of 3 to minimise the RAM usage since cyphertexts are extremely large. We split the data, 80-20 training-testing respectively, by time so that the neural network during training will only see a subset of the data that it has never interacted with and has not see any future sequences of for the purposes of testing. This 80% used for training is further split by 80-20 to create a 20% validation set for the purposes of model selection. Create example sequences by taking 3 weeks of environmental data out from historic yield values, and taking a further 3 weeks of sequences prior, as what would have been the currently known state of the fruit and its environment. We use a skipping-window to create these sequences directly from the database. The sequence shape once they have been extracted and batched is (3, 1, 7560, 15) when fed into the neural network. We then variably add any weather forecasting as part of the input sequence or forego it entirely to create assumption based yield predictions. However performance is significantly improved with prediction-to-output weather conditions/ forecasting. This is something we will be exploring more as the final section of the PhD and yield prediction performance.

Results

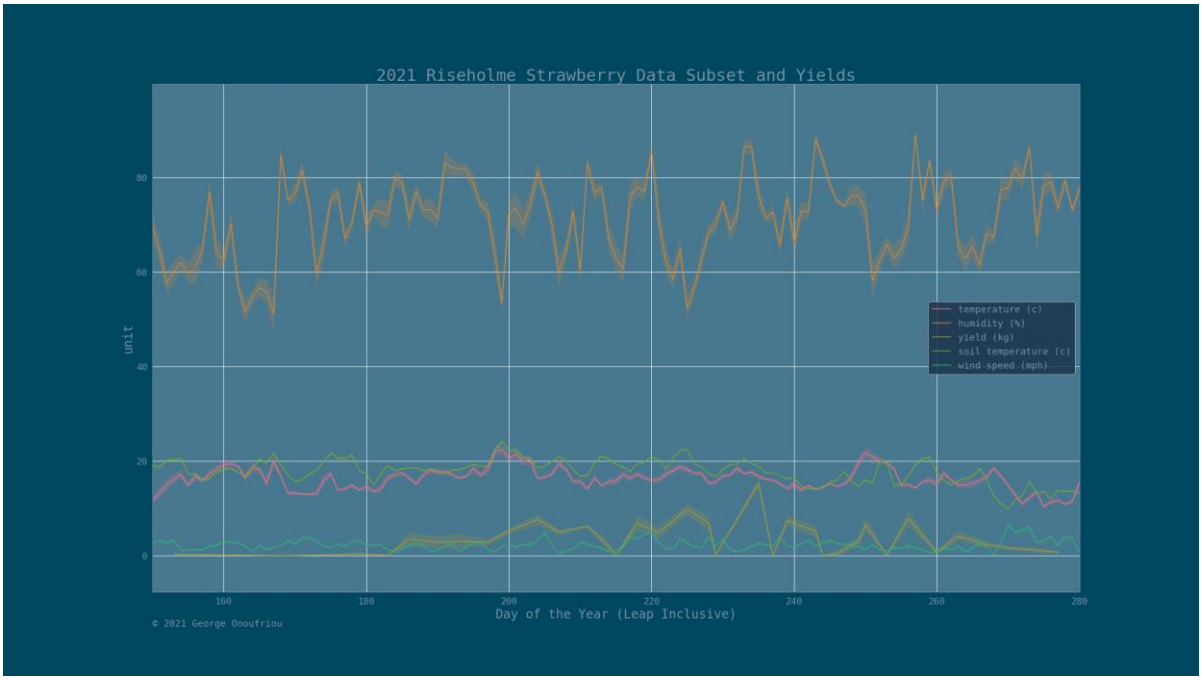

Figure: 1. Focused strawberry line graph showing a subset of key data in the growing season.

Figure 1 shows the a subset of the key features related to the strawberries and the strawberries immediate environment. This line graph also shows the variance in any single day through error margins. As can be seen over the growing season the output of the strawberries behaves quite erratically, along with many of the other features being relatively unstable, in-particular humidity.

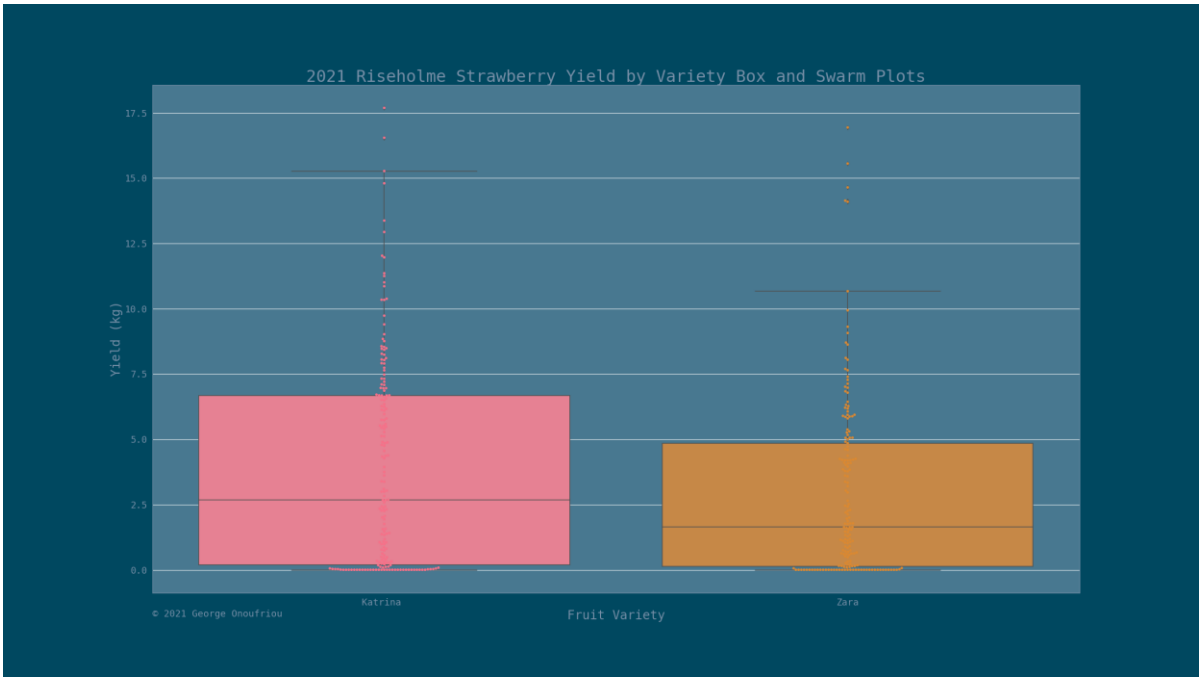

Figure: 2. Strawberry yield production by variety, box and overlaid swarm plot.

Figure 2 shows the overall production of berries by weight from both varieties in the 2021 season. This graph shows the berry weight in kg, and how the average, 75th percentile and maximum yields are all claimed by the Katerina variety.

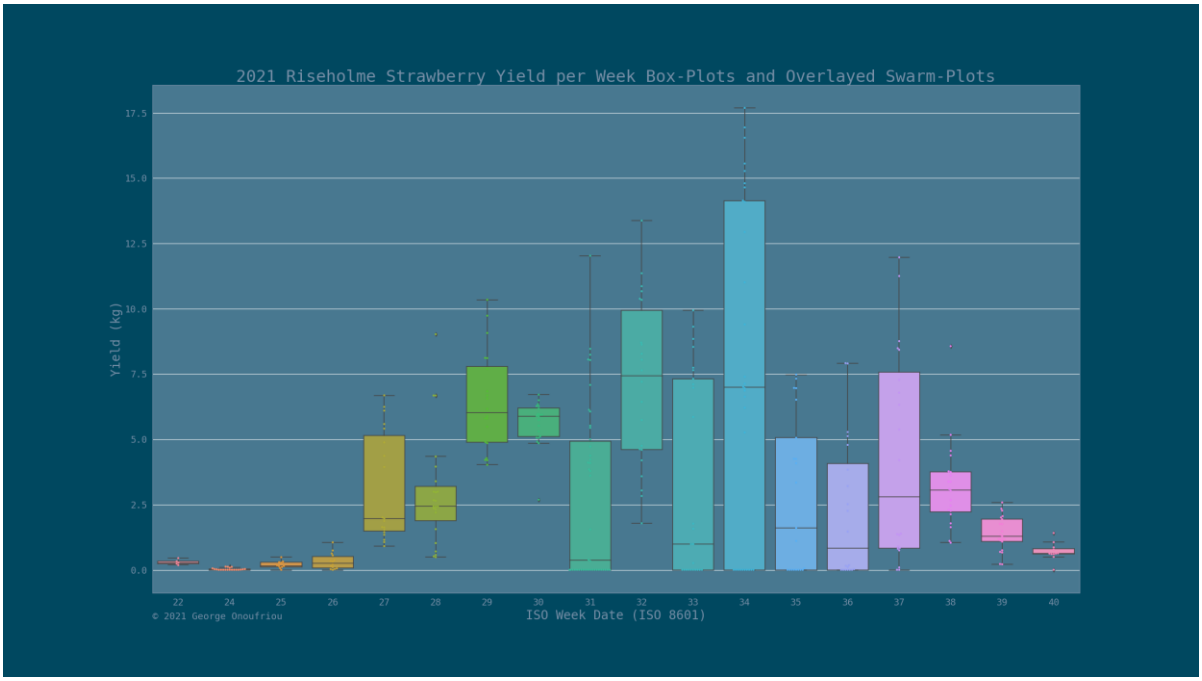

Figure: 3. Weekly view of 2021 strawberry yields over the growing season with overlaid swarm plot.

Figure 3 shows the individual growing weeks in the 2021 season, and the quantity of strawberries outputted by the individual picks by each individual row that week. However since there are two picks a week that are related to each other, the harvesting effect is not displayed in this graph.

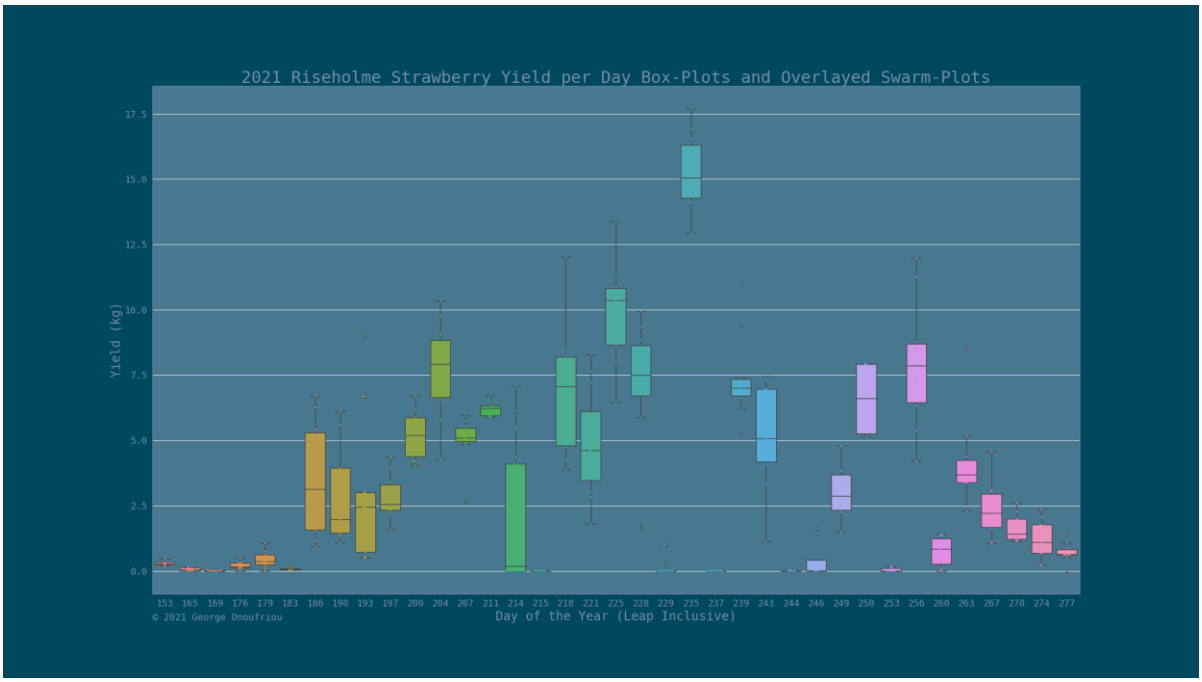

Figure: 4. Daily view of 2021 strawberry yields over the growing season with overlaid swarm plot.

Figure 4 shows the daily yields produced by each row as swarm dots, with boxplots to show the mean, quartiles, and range for all picks that day. This graph clearly shows the

consequences to the harvesting effect, and how adjacent picks especially large picks, have consequences on the following pick(s).

Days Ahead	Mean Absolute Percentage Error (MAPE)	Perfect Weather Forecasting
7	8	1
7	8	0
14	12	1
14	14	0
21	16	1
21	22	0

Table: 1. 2020 growing seasons 1D-CNN performance given interim or excluding interim weather.

Table 1 displays out convolutional networks mean absolute percentage error (MAPE)/ performance by forecasting gaps (7, 14, 21), and weather forecasting availability. This table shows that predicting within 1 week ahead, a performance of 8% is achieved. For 2 weeks, 12-14%, and for 3 weeks 16-22% as the cone of uncertainty increases non-linearly the further ahead the prediction is made for. These results are one growing-season behind due to the timing of theses report falling midway through a growing season.

Discussion

As can be seen in Table 1 our results are promising, and exceed our previous years results which were around 16% for two weeks ahead. We believe this is largely due to much better and cleaner data from improving our data collection techniques in the year of 2020. The data is only set to improve for 2021, and we will be able to merge the two to begin cross growing season predictions now that the data is satisfactory. We also noted that we may want to move to a scheduled/ sequenced output approach rather than predicting singular points, since the schedule clearly affects the resulting yield due to the harvesting effect shown between Figure 3 and 4. This would open more possibilities for schedule generation not just yield prediction, so that the neural network will be able to tell you when to optimally pick the fruits. Figure 2 seems to show that Katerina is outperforming Zara in total output yield, but we cannot be sure if this is a consequence to slight variations in climate and conditions between the two polytunnels. To confirm this we should likely rotate the strawberry varieties to ensure we can identify any seasonal systematic errors due to micro-conditions/ variances between our tunnels. There is still plenty of room for improvement in our results but now that we have sufficient multi-year data we should be able to squash the error somewhat due to being able to see a whole previous season.

Conclusions

Overall this year we showed how fully homomorphic encryption together with deep learning to produce a whole new field of encrypted deep learning can be successful and garner adequate results. We have found some improvements that can be made for this new 2021 season and its data processing, to more accurately predict strawberry yields ahead of time, and to a set schedule for optimisation. There still remains an objective to be achieved with regards to certainty metrics but this is a relatively easy objective to meet in preparation for finalising the project.

Knowledge and Technology Transfer

- University of Lincoln **Big Data Guest Lecture** on Artificial Intelligence and containerisation (2021-11-16)
- **IEEE-TPAMI** Journal EDLaaS; Fully Homomorphic Encryption Over Neural Network Graphs (submitted 2021-10-26)
- **CTP autumn event** (2021-11-03)
- **LAR** 2nd Mini-Conference (2021-07)
- **Internet of Food Things Virtual Workshop**: How Technology Can Facilitate Data Sharing In The Agri-Food Sector (2021-03-01)
- University of Lincoln School of Computer Science: **Research Seminars** (2021-02-10)
- **AHDB PhD Crops Conference** (2021-01-25)
- **New Scientist Live** 2020 (2020-11-28)

Glossary

- FHE; Fully Homomorphic Encryption
- DL; Deep Learning
- NN; Neural Network
- RNS; Residue Number System
- CKKS; Cheon, Kim, Kim, Song
- ReLU; Rectified Linear Unit
- Adam; Adaptive Moment
- CNN; Convolutional Neural Network

- ANN; Artificial Neural Network/ Dense Network

References

References should be included in the report and should be written in the following standard format:

Atkey, P.T. and Nichols, R. (1983). Surface structure of *Agaricus bisporus* by scanning electron microscopy. Mushroom Journal 129:334-335.

Grogan, H. (1997). Examination of the efficacy of two novel fungicides against *Dactylium dendroides*. Horticultural Development Company Annual report for project M 22.

Appendices

