Skolkovo Institute of Science and Technology

DATA-DRIVEN MODELING OF PLANT GROWTH DYNAMICS IN CONTROLLED ENVIRONMENTS

Doctoral Thesis

by

DMITRY SHADRIN

DOCTORAL PROGRAM IN COMPUTATIONAL AND DATA SCIENCE AND ENGINEERING

Supervisor
Professor, Dr. Maxim Fedorov

Moscow – 2020

I hereby declare that the work presented in this thesis was carried out by myself at Skolkovo Institute of Science and Technology, Moscow, except where due acknowledgement is made, and has not been submitted for any other degree.

Candidate (Dmitry Shadrin)

Supervisor (Prof. Maxim Fedorov)

Data-driven modeling of plant growth dynamics in controlled environments

by

Dmitry Shadrin

Tuesday 27th October, 2020 17:28

Submitted to the Skoltech Center of Computational and Data Science and Engineering
on July 2020, in partial fulfillment of the requirements for the
Doctoral Program in Data Science

Abstract

Sustainable and energy-saving production of food are becoming increasingly important with the growth of the world population and global warming. Today food production demands for automation and optimization of growth dynamics and resources consumption. Thus, a reliable mathematical approach that can perform a comprehensive description of growth systems should be developed for solving this task. However, at present hardly any mathematical approach is universal, flexible, and robust enough to meet the demands and describe and predict plant growth dynamics in controlled conditions. The existing solutions are partial – though they involve a huge amount of empirical data and parameters, the proposed models of plant growth dynamics vary dramatically. Thus each particular design of a growth system and every sample of plant cultivation need these models to be adapted and optimized. Noteworthy, the systems under consideration should be treated especially carefully, as even a subtle change brought to the system might impair the applied model and thus destroy the predicting performance of the mathematical approach. Another factor to be taken into account is a rapid increase in the number of plant hybrids, which makes it a hopeless task to perform a comprehensive analysis of each plant hybrid and to develop a strict mathematical model describing its growth under different conditions.

In view of the aforementioned, the main scientific issues to be solved in the present thesis are (i) developing universal data-driven approaches for real-time precise description of growth conditions and (ii) improving the accuracy and robustness of the whole cycle of plant growth dynamics assessment and prediction in different controlled environments. In recent years there has been significant progress in the implementation of computer vision and machine learning technologies for precision agriculture in particular for plant phenotype and for growth dynamics prediction. However, the implementations of these technologies are partial. This thesis seeks to fill in the gap and contribute to improving and adapting current data-driven approaches for precision agriculture, as well as to achieving end-to-end implementations. To go in further detail, the present work includes the following steps:

- Developing and constructing novel automated artificial growth systems equipped with sensor systems and non-invasive machine vision plant growth monitoring systems.

- Collecting relevant and unique datasets in laboratory and industrial experimental setups, describing plant growth dynamics in different environments.

- Proposing new approaches, adapting, and improving different existing state-of-the art data-driven and hybrid modeling approaches for plant phenotype and growth dynamics prediction. Evaluation and comparison of used methods on the collected datasets. The main data-driven and hybrid methods that were adapted and implemented are: Kalman filtering, dynamic mode decomposition, merging 2D/3D data, convolutional (fully) neural networks, recurrent neural networks (benefits and drawbacks of all the enlisted methods are to be discussed in the thesis).

- Developing a novel computer vision based system for continuous seeds germination monitoring. The proposed system allows to automatically seed detection and quantification of germination rate.

- Plant health monitoring. Proposing a novel and practically useful approach enabling to find optimal spectral wavebands for early remote plant disease detection and classification. Testing the proposed approach on the own hyperspectral near-infrared dataset obtained in reflected spectra for apple tree diseases on different stages of development.

- Modeling of environmental parameters that have effect on plant growth dynamics in field conditions. Applying machine learning techniques to model the spatial distribution of highly variable environmental parameters and quality of growth conditions. Improving of the current state-of-the-art results of environmental parameters modeling. These problems have not been solved yet precisely due to the high complexity and non-linearity of parameter dependencies in open systems. Also, approaches developed for open systems can be easily transferred to artificial systems.

Overall, all the proposed and tested data-driven methods for plant phenotype and growth dynamics prediction showed high accuracy, universality, and significant level of automatization. The created experimental setups and collected unique datasets appear to be highly relevant and could be used in further investigations in this research area.

Publications

Main author

1. Dmitrii Shadrin, Andrey Somov, Tatiana Podladchikova, and Rupert Gerzer. Pervasive agriculture: Measuring and predicting plant growth using statistics and 2d/3d imaging. In *2018 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, pages 1–6. IEEE, 2018a

2. Dmitrii G Shadrin, Victor Kulikov, and Maxim Fedorov. Instance segmentation for assessment of plant growth dynamics in artificial soilless conditions. In *BMVC*, page 329, 2018b

3. Dmitrii Shadrin, Alexander Menshchikov, Dmitry Ermilov, and Andrey Somov. Designing future precision agriculture: Detection of seeds germination using artificial intelligence on a low-power embedded system. *IEEE Sensors Journal*, 19(23):11573–11582, 2019b

4. Dmitrii Shadrin, Artem Chashchin, George Ovchinnikov, and Andrey Somov. System identification-soilless growth of tomatoes. In *2019 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, pages 1–6. IEEE, 2019a

5. Dmitrii Shadrin, Alexander Menshchikov, Andrey Somov, Gerhild Bornemann, Jens Hauslage, and Maxim Fedorov. Enabling precision agriculture through embedded sensing with artificial intelligence. *IEEE Transactions on Instrumentation and Measurement*, 2019d

6. Dmitrii Shadrin, Mariia Pukalchik, Ekaterina Kovaleva, and Maxim Fedorov. Artificial intelligence models to predict acute phytotoxicity in petroleum contaminated soils. *Ecotoxicology and Environmental Safety*, 194:110410, 2020b

7. Dmitrii Shadrin, Mariia Pukalchik, Anastasia Uryasheva, Evgeny Tsykunov, Grigoriy Yashin, Nikita Rodichenko, and Dzmitry Tsetserukou. Hyper-spectral nir and mir data and optimal wavebands for detection of apple tree diseases. In *ICLR (CV4A)*, 2020c

8. Dmitrii Shadrin, Tatiana Podladchikova, George Ovchinnikov, Artem Pavlov, Maria Pukalchik, and Andrey Somov. Kalman filtering for accurate and fast plant growth dynamics assessment. In *2020 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*. IEEE, 2020a

Co-author

1. Andrey Somov, Dmitry Shadrin, Ilia Fastovets, Artyom Nikitin, Sergey Matveev, Oleksii Hrinchuk, et al. Pervasive agriculture: Iot-enabled greenhouse for plant growth control. *IEEE Pervasive Computing*, 17(4):65–75, 2018

2. Maria A Pukalchik, Alexandr M Katrutsa, Dmitry Shadrin, Vera A Terekhova, and Ivan V Oseledets. Machine learning methods for estimation the indicators of phosphogypsum influence in soil. *Journal of Soils and Sediments*, 19(5): 2265–2276, 2019

3. Artyom Nikitin, Ilia Fastovets, Dmitrii Shadrin, Mariia Pukalchik, and Ivan Oseledets. Bayesian optimization for seed germination. *Plant methods*, 15(1): 43, 2019

4. Maria Pukalchik, Dmitrii Shadrin, and Maxim Fedorov. Global trends and perspective development directions in precision agriculture. *APK Russia Journal (in Russian)*, 2018

5. Sergey Nesteruk, Dmitrii Shadrin, Vladislav Kovalenko, Antonio Rodríguez-Sánchez, and Andrey Somov. Plant growth prediction through intelligent embedded sensing. In *IEEE International Symposium on Industrial Electronics 2020, IEEE ISIE 2020 (Accepted)*, 2020

Dedicated to my parents.

Acknowledgments

First of all, I would like to thank my supervisor Prof. Maxim Fedorov. He has been a really great mentor and support me a lot. During our weekly meetings we had comprehensive and open scientific discussions which were really important for my Ph.D. research. He shared experience which helped with establishing my own view on how all things should work correctly in science and academia. I should also notice that it is hard to find supervisor who takes care of students as Maxim Fedorov does. Thanks Maxim for being so kind attentive and helpful to me and my research.

I am really grateful to Prof. Rupert Gerzer who gave me lots of opportunities for international collaborations, in particular, with German Aerospace Agency, where I conducted experiment on plant growth under supervision of Gerhild Bornemann and Jens Hauslage. This was a very nice and unique experience resulting in a big part of my thesis. Rupert gave me very useful suggestions for scientific career and I am very glad to hear them in time.

I would like to thank Andrey Somov, whose contribution to my Ph.D. research was great. He convinced me that I can write scientific paper, after that my first serious paper appeared and I started to recognize myself as a scientist. We worked together very productive and published a lot of research. Andrey opened for me lots of opportunities in research activities.

I am really happy to collaborate with Maria Pukalchik, who is leading the sustainability lab in CDISE (Skoltech). Together, we conducted research in the application of AI for environmental sustainability and published a series of papers. She motivated me and helped a lot technically, I can ask anytime regarding all questions that arose during preparing my Ph.D. thesis. Also, it was very productive to work with students in the sustainability lab. I was always in contact with Maria. It would have been impossible to finish Ph.D. research without her support.

I would like to thank Alexander Menshchikov for close and very effective collaboration. Together we conducted research and published several papers. He was very helpful in discussing technical details. He always supported me and inspired me for doing good research.

Let me also thank Ivan Oseledets for giving me such deep knowledge in math. He was always open for technical discussions and gave several principal suggestions for my Ph.D. research

Of course, I would like to thank all Skoltech professors who taught me during my study. All obtained knowledge is highly relevant not only for Ph.D research but also for future career

I really appreciate having a chance working with: Tatiana Podladchikova, with whom we developed Kalman filtering approach for plant growth dynamics prediction; Victor Kulikov, with whom I worked on Instance segmentation for plants; George Ovchinnikov, and Artem Chashchin helped me with dynamic mode decomposition approach; with Dmitry Ermilov we trained NNs for germination detection; with the help of Dzmitry Tsetserukou experiment of plant disease detection was conducted; with Artyom Nikitin we worked on Bayesian optimization of seeds germination and Gaussian process regression for environmental parameters assessment.

Finally, I am glad to have a chance to thank my parents and friends who gave me strong support during my life, so that I was able to go so far making dreams to be a scientist the reality and prepare this work. It would have been definitely impossible without them.

Contents

List of Figures

List of Tables

List of abbreviations

Used abbreviations

IoT - Internet of Things

ML - Machine Learinig

AI - Artificial Intelligence

CV - Computer Vision

ANN - Artificial Neural Network

CNN - Convolutional Neural Network

FCNN - Fully Convolutional Neural Network

RNN - Recurrent Neural Network

DMD - Dynamic Mode Decomposition

GP - Gaussian Process

SVR - Support Vector Regression

TPH - Total Petroleum Hydrocarbons

Chapter 1

Introduction

1.1 Thesis objectives and contribution

The present thesis focuses on several problems related to applying computer vision and machine learning approaches to solving problems in precision agriculture, including plant phenotype growth dynamics prediction and environmental conditions assessment. Contributions to these topics are briefly discussed in the following sections:

Artificial growth systems equipped with machine vision and sensors systems. Collecting of the relevant datasets.

At the first step two small-scale artificial growth systems were designed and constructed. The former has the capacity to grow up to 20 plants simultaneously in entirely artificial conditions. The environmental parameters of the system, such as nutrient solution and light spectra, were controllable; the system was equipped with a system of sensors to measure automatically all the basic growth conditions (pH, electrical conductivity (EC), temperature, humidity). To obtain sequences of images allowing the growth dynamics reconstruction and prediction, a robotic system with a mounted digital camera taking high resolution images was created. The system also includes an algorithm for an automatic leaves area (projection) calculation. All measurements were synchronized and organized as a structured database. To represent the state of the system in-situ monitoring web interface was developed. Also,

this experimental setup allows to obtain 3D scans of plants, which can be used for reconstructing 3D model of the plant during its growth.

The other system was similar to the first one, except that its growth capacity was 54 plants simultaneously, and nutrient solution could be directed to different sections of the system. The projected leaves area and environmental conditions were measured automatically in the same manner. The process of exploiting the systems showed that they appeared to be very sensitive and allowed us to obtain reliable data. Diurnal fluctuations of the leaves area projection and relatively low changes in the environmental conditions were captured with a good resolution. These systems, being artificial, allowed us to alter growth parameters in order to investigate the growth rate under different conditions.

Thus, it is possible to obtain relevant datasets of growth conditions and corresponding plant responses for building and testing huge variety of different models for reconstruction and prediction of the plant growth dynamics. Overall, the developed experimental setups can be used for setting up a variety of laboratory experiments aimed at investigating CV, ML, or other approaches for modeling plant growth dynamics and plant phenotype.

Data-driven techniques for growth dynamics modeling and plant phenotype

A novel approach for evaluating and predicting of plant growth dynamics using computer vision and machine learning approaches was proposed. This approach consists of the application of convolutional neural networks (CNN) for segmentation of plants allowing projected leaves area calculation. Also, CNNs were used for solving instance segmentation task allowing to track the growth of leaf. It was shown that it is possible to model the growth dynamics of each leaf. Recurrent neural networks (RNN) were used as the basis for the prediction of growth dynamics. End-to-end solutions that were able to collect and preprocess data that describe plant growth dynamics, extract features and predict plant growth dynamics was proposed and successfully tested in the developed experimental setups. The possibility to assess and predict biomass by merging 2D and 3D techniques was also shown. The main

feature of the proposed approach is that it allows to predict biomass using simple 2D cameras and correlations between biomass and leaves area. Proposed approaches showed their practical usefulness, prediction errors for different time horizons were low compared to existing techniques.

The main advantage of the proposed methodology is that it can be universally adapted to different plant species and environmental conditions. Also, this methodology can be easily implemented practically. Being installed, such a system requires minimal involvement of the user in the data processing and dynamics prediction procedure. The novelty of research is such end-to-end implementation of CV and ML techniques. Implementing the proposed data-driven approaches into the embedded systems opens up new vistas for building novel distributed monitoring and analysis systems in the greenhouse.

Hybrid modeling of plant growth dynamics

Since incorporating of precise measurement systems and training neural networks not always practically reasonable in the existing greenhouses, an alternative approach was proposed. It is based on more straightforward CV algorithms and the models widely used to describe and predict the state of dynamical systems. In particular, CV algorithms were used as feature extractors and Kalman filter was adopted for growth dynamics prediction. In addition, a novel application of dynamic mode decomposition (DMD), initially designed for describing flow dynamics, was proposed and tested. The approach under consideration suggests to use the features derived from the set of differential equations as a state vector and then to use the DMD approach to perform system identification. The proposed approaches for growth dynamics evaluation and prediction have several advantages: they are computationally fast and comparably accurate, and they need less data as compared to the above mentioned data-driven methods. However, one of the drawbacks of our approaches contrary to pure data-driven techniques, is that they require fine-tuning and adaptation before being applied to a particular plant species and greenhouse.

Computer vision for seeds germination

Germination rate is a crucial parameter that predefines the future growth of the plant. Currently, the monitoring of seed germination is largely performed manually. This procedure has several drawbacks. First, this method disturbs the germination system. Second, it is quite a time consuming process. Finally, manual measurements do not presuppose a continuous period of monitoring. Therefore, an automatic system detecting and quantifying the germination rate is of a high demand.

The present study describes a novel experimental setup equipped with CNN-based and CV methods for automatic germination rate assessment. The proposed methods were evaluated on the own obtained dataset and showed a high accuracy and the ability to monitor the germination process precisely and continuously. This intelligent system opens up endless possibilities for performing data-driven optimization of the germination process, which is state-of-the-art for precision agriculture.

Industrial experiment for obtaining the unique dataset, and plant biomass prediction

The thesis includes unique industrial experiment on plant growth dynamics prediction in greenhouses. The experiment was set up in a greenhouse; in its course image data were collected and biomass measurements were obtained for 540 plants. Environmental conditions were measured using a developed sensing system, while the images were taken by mounted 4 digital cameras. The image dataset was labeled by putting segmentation masks on the plants. Based on the labeled dataset several architectures of fully convolutional neural networks (FCNN) were trained. The trained FCNNs showed that it was possible to derive each mask of the plant and to calculate projected leaves area in the industrial environment. The sequence of images in couple with FCNN allowed us to derive plant growth dynamics in the greenhouse; biomass measurements showed the correlation between the biomass and the projected leaves area and were used to predict the biomass. The main advantage and uniqueness of the conducted experiment is that it was obtained comprehensive dataset and CV and ML algorithms were tested in industrial conditions showing high practical usefulness.

Hyper-spectral approach for disease detection

Detecting plant diseases and deviations in growth is a high priority for both field cultures and those grown in greenhouses, since these phenomena may result in dramatic yield losses. Currently, there is a great demand for a precise, non-invasive system that would enable detecting diseases at the initial stages and, thus, prevent the disease from spreading. The present thesis argues for a high potential of near-infrared spectrum region to detect diseases at early stages. A unique dataset was collected, which included several of the most common fungal apple trees diseases and reflected the spectrum of each disease in near-infrared spectra region correspondingly to the stage at which it was obtained. The resultant dataset was shared with the community. Basing on resultant spectra and the developed mathematical approach a novel methodology was proposed for deriving optimal wavebands. This methodology was proved to be effective for early detection of a particular disease. It should be noted as well that the developed methodology could be implemented to a wider range of plants, as fungal diseases are mainly of the same origin.

Data-driven modeling of environmental parameters

The present thesis presents and discusses the ways to use data-driven techniques in solving several topical environmental issues that closely relate to the precision of plant growth modeling. These issues either remained unsolved so far or solutions were based on old fashioned, inaccurate methods and required an exceeding number of assumptions to be made. Direct modeling gives inaccurate results due to soil factors have high non-linear and complex effects. Therefore, the existence of one general approach giving accurate results for a huge variety of soils is highly doubtful.

The first problem to be solved is soil phytotoxicity assessment and deriving the main factors that influence phytotoxicity and biological response. To study the way ML methods such as SVR and feed-forward NNs could be used in addressing these problems, we tested these approaches on two datasets. The first one contains of the soil samples contaminated by oil and the second contains of the soil samples mixed with phosphogypsum. For all the soil samples chemical, biological and toxicological properties were measured. The tested ML methods showed that an accurate

prediction of biological and toxicological parameters could be obtained for a huge variety of soils. Moreover, ML methods were able to derive key soil parameters and to quantify the effect of biological and toxicological responses.

Another problem that regularly appears in open systems (i.e. fields) is modeling a spatial distribution of environmental parameters. The present thesis proposes an approach to solve such a problem for environmental modelling. Kriging (Gaussian process), the method widely used for this purpose, was improved by implementing an automatic optimal kernel structure search based on Bayesian information criteria. This method have never been applied before in precision agriculture. It was tested on the dataset describing water quality measurements taken in different locations and showed a much better accuracy and robustness as compared to standard methods. Another advantage of the method is its automatic routine. It finds the best possible kernel structure, which reduces the human factor.

Noteworthy, the data-driven approaches developed and tested for open systems can be adapted for greenhouses and artificial systems to solve several problems, such as finding key factors in nutrient solution that affect growth dynamics or finding the spatial distribution of humidity and temperature in a greenhouse.

1.2 Thesis structure

The diagram in Fig. 1-1 illustrates the flow of information through the structure of the thesis.

Section 1.4 - **Motivation: Review of global trends in digital agriculture.** In this section the emerging trend for precision agriculture as a subject for investigations is described. The discussion is proved by statistical numbers reflected the dynamics of research specific publication activity and amount of funding during the last years. The review resulted in a publication in "APK Russia" Journal [Pukalchik et al., 2018] (co-author).

Chapter 2 - **Background**. Review of the background of different techniques that are used in the following chapters. More specifically, the section is focused

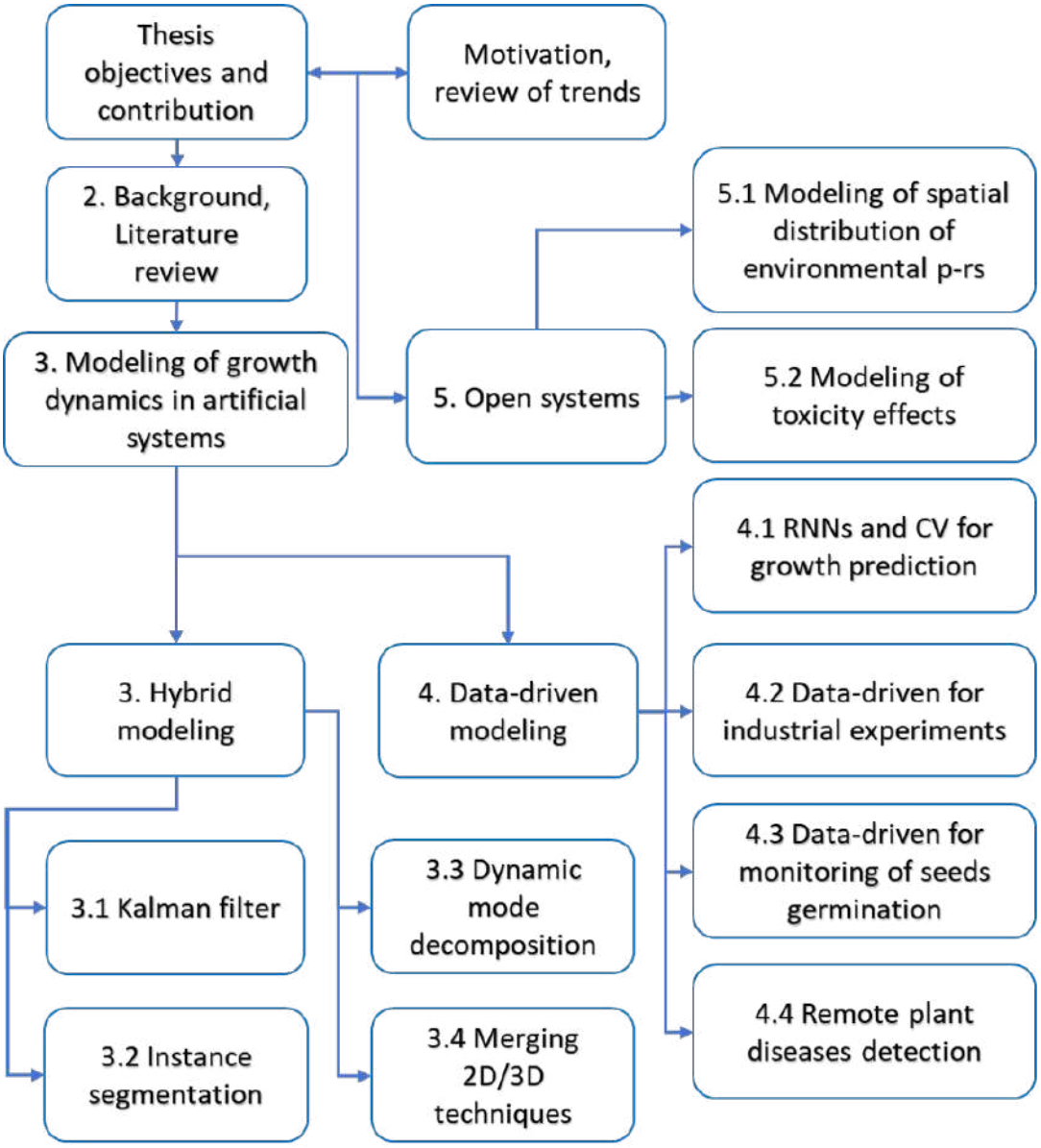


Figure 1-1: Thesis structure.

on the following branches: artificial growth systems, sensor systems in precision agriculture, imaging technologies for plant phenotype, deep learning methods for growth dynamics prediction, hybrid models for growth dynamics prediction, approaches for seeds germination monitoring, computer vision methods for diseases detection. The background also includes section, where it is shown the benefits and limitations of the differential-equations based modeling of plant growth dynamics. The discussion and conclusions supported by the performed modeling.

Chapter 3 - **Hybrid combination of methods for modeling of plant growth in controlled environments.**

In this chapter, small scale experimental setups and obtained relevant datasets were described. Using these datasets, it was shown the implementation of Kalman filtering approach, dynamic mode decomposition approach, instance segmentation coupled with simple models for plant growth dynamics prediction and plant phenotype. Finally, the approach for biomass prediction was proposed. This work resulted in five publications in conferences proceedings: three papers in IEEE International Instrumentation and Measurement Technology Conference (I2MTC) [Shadrin et al., 2018a, 2019a, 2020a], paper in British Machine Vision Conference proceedings (BMVC, CVPPP) [Shadrin et al., 2018b], and IEEE International Symposium on Industrial Electronics [Nesteruk et al., 2020].

Chapter 4 - **Data-driven enhancement for plant growth modeling in controlled environments.**

The first part of this chapter presents a wide range of possible ways to apply data-driven approaches. In particular, it describes the application of RNNs and CNNs to the plant growth dynamics prediction based on the data obtained in two experiments. The first was a small-scale experiment on growing tomato plants in artificial conditions conducted in DLR (Deutsches Zentrum für Luft- und Raumfahrt, German Aerospace Agency). The second was an industrial-scale experiment on growing cucumber plants in Michurinsk greenhouse. The results of the pure implementation of data-driven approaches were published in IEEE Transactions on Instrumentation and Measurement Journal [Shadrin et al., 2019d]. The sensor system that is used for the industrial experiment described in other published work [Somov et al., 2018] (co-author).

The second part of the chapter describes the developed intelligent computer vision based system for germination rate assessment and evaluates the performance of the system. It is concluded that the developed system shows a high accuracy and autonomy. This work was published in IEEE Sensors Journal

Shadrin et al. [2019b]. This system was used for optimization of germination rate and based on this research another paper was published in Plant Methods Journal [Nikitin et al., 2019] (co-author).

The last section describes the plant health monitoring approach based on near-infrared hyper-spectral data analysis. The procedure of obtaining of hyper-spectral dataset on apple tree diseases and finding optimal wavebands for disease detection using the proposed methodology is described. The proposed methodology showed that it is possible to detect different diseases in the initial stages when they can not be visually detected. This research was published in ICLR 2020 (CV4A) conference proceedings [Shadrin et al., 2020c].

Chapter 5 - **Data-driven modeling of environmental parameters for improvement plant growth prediction.**

This chapter is aimed at the description of proposed data-driven approaches for improvement of modeling universality and accuracy of the important environmental parameters that have a huge effect on plant growth dynamics. Improvement in the modeling of spatial distribution was proposed based on Gaussian process regression and automatical kernel structure search. This approach outperforms all existing methods which was demonstrated on modeling of water quality map. The results are under revision in Frontiers in Plant Science journal. By means of using ML methods accuracy and descriptive ability were also improved in predicting phytotoxicity effects of contaminants and in deriving the driving factors that make an effect on growth dynamics in case of using fertilizers. The results were published in Ecotoxicology and Environmental Safety [Shadrin et al., 2020b], and Journal of Soils and Sediments [Pukalchik et al., 2019] (co-author).

Chapter 6 - **Conclusion.** The concluding chapter of the thesis presents and discusses the obtained results.

1.3 Co-authorship statement

Chapter 1 and Chapter 2 are my own work on the introduction to the thesis and description of the background and related literature. Section 1.4 in Chapter 1 is joint work with Prof. Maria Pukalchik. Together we aggregated information about trends in publication activity and funding in precision agriculture. Maria is a primary author of the review paper, which is based on the obtained analysis.

Chapter 3 is mainly my own work. However, different colleagues contributed to different sections in this chapter.

I designed and assembled Two experimental setups were. I also wrote the code for receiving data from sensors and cameras and create and implement the algorithm for automatical projected leaves area calculation. All these pieces of code were synchronized by me properly. This enable to obtain high quality datasets. I collected several datasets including data from sensors, 2D data, 3D data that describe plant growth dynamics. Dr. Jens Hauslage helped with the critical suggestions for the improvement for the second experimental setup developed by me in DLR.

Kalman filtering for growth dynamics prediction was adapted together with Prof. Tatiana Podladchikova. I performed the implementation of the theoretical model in code, modeling and writing paper. Co-authors helped with formatting and proofreading.

Dynamic mode decomposition for growth dynamics prediction was a joint work with George Ovchinnikov and Artem Chashchin. George Ovchinnikov and I proposed the idea to extract features and prepossess dataset for application of DMD based on theoretical findings. Together with Artem Chashchin this method was implemented and growth dynamics were modeled. Paper was written mainly by me and Artem Chashchin.

Instance segmentation for plant phenotype was performed by me and Viktor Kulikov. Data labeling was performed by me. Data preprocessing and

training of CNNs for obtaining segmentation masks were done together. Data post-processing for tracking of each leaf and growth dynamics prediction was done by me. I mainly wrote the paper on the implementation of data-driven approaches for plant growth dynamics prediction.

Merging of 2D/3D approach for biomass prediction performed by me, from the idea and 2D/3D data collection to testing of prediction algorithms. The paper was written by me with the help of Prof. Andrey Somov.

Chapter 4 is mainly my work with contributions from my colleagues to different sections. Seeds germination system was proposed by me. I collected the dataset and preprocessed it. Together with Dmitry Ermilov, we trained CNNs for proposing regions with seeds. I developed and implemented the algorithm for quantification of the germination rate inside the proposed regions. Alexander Menshchikov was responsible for the embedding of all intelligent system. The major part of the paper was written by me, several sections were written by Alexander and Dmitry. Andrey Somov helped with the overall structure.

Training and evaluation of recurrent neural networks were performed by me on the own obtained data in the DLR's experiment. Alexander was responsible for the embedding of all intelligent system. I wrote the major part of the paper, the section about the embedded system was written by Alexander. Andrey Somov helped with the overall structure.

I proposed and developed the design of the industrial experiment which is aimed generally for biomass prediction. Together with Artyom Nikitin we deployed the sensor system and collected relevant data. I guided the process of labeling of the dataset based on which I trained fully convolutional neural networks for performing semantic segmentation. Alexander helped with the data prepossessing procedure. Together we worked on the publication preparation.

Methodology for processing of hyper-spectral data and finding optimal wavebands for early plant disease and was proposed, implemented and evaluated by me. Hyper-spectral data were collected mainly by me with the help of colleagues from the robotics lab, who are co-authors of the respective paper.

Paper was written mainly by myself.

Chapter 5 is a joint work with Maria Pukalchik. Maria, being the expert in the research field of open environments, helped to correctly state the problems and shared knowledge about state-of-the-art solutions. Development of improvements to the current state-of-the art based on ML approach and implementation of them in code was performed by me. We published several papers with Maria. She was responsible for the expertise, discussion, and problem statement. I was responsible for technical development and implementation of improvements to current approaches and demonstration of the results.

Chapter 6 is my own work on summarizing the whole thesis and outlining the main results

Research, discussed above, were conducted under the supervision of Maxim Fedorov, who helped to find the most relevant directions and suggested the general approaches for solving the particular problem.

1.4 Motivation: Review of global trends in digital agriculture

This review examines the world trends in research interest and financing of research projects for precision agriculture for the period from 2008 to 2017. Six priority research areas are identified: Artificial intelligence and data visualization, livestock, crop production, information systems, genetics and Earth sciences. The importance of interdisciplinary projects is increasing. This increase includes the application of methods for processing huge amounts of received data. The most influential scientific trend in precision agriculture today is the implementation of artificial intelligence systems and digital methods in data processing for various sectors of the agricultural industry.

The main objective of the agricultural industry is to achieve sustainable growth in agricultural production while reducing the consumption of energy and natural resources [Shatalina, 2017]. In the search for a solution to this problem, a development

direction has been formed, called "precision agriculture" or "precision farming", which involves the implementation of solutions for collecting, storing and processing data on the state of various objects, preparing recommendations for the best management of them [Kaloxylos et al., 2012]. The global market for precision agriculture is constantly growing, according to experts, by 2022 it can reach 25.4 billion dollars with an annual expected growth up to 12.6% [Orbis Research, 2018]. Automatic control systems for machines and equipment, sensors, remote sensing, integrated electronic communications and machine learning methods are widely used for collecting data on the state of land resources, planning irrigation measures, and distributing of fertilizer application [Kukar et al., 2019, Daccache et al., 2015, Nefedov, 2015]. The agricultural industry is rapidly transforming under the influence of nanotechnology, the capabilities of metagenomics for the selection of living organisms, manufacturers integrate production and marketing chains and adapt their products to the needs of a particular consumer. At the same time, digitalization tools are getting cheaper, and cloud technologies are designed to make their use massive and affordable for consumers. A number of reviews provide comprehensive information on the current capabilities and disadvantages of the well-known technological methods for various sectors of agriculture and animal husbandry, and factors affecting the success of the introduction of precision farming systems in Russia are indicated [Wolfert et al., 2017, Kiryushin, 2009, Yakushev V.P., 2014].

The challenge facing the industry and manufacturers of high technology products is to identify and support promising trends. Future key decisions for global markets are formed primarily at the stage of research. The purpose of this is to give new possibilities, which allows us to solve specific practical problems in agriculture in the future. The main channel for disseminating information about the research is not only the publications of researchers in scientific journals but also data on allocated funding for individual projects published in open sources. Therefore, it is possible to evaluate the trends that will determine the course of development of precision agriculture in the future, including through an analysis of global trends in financing of individual research projects on precision farming and precision agriculture.

Data sources

Information on the financing of research for the period from 2007 to 2017 is given according to the database (DB) Dimensions (see Table 1.1). The review deals with the financing of research primarily through foreign foundations and scientific organizations, which occupy a leading place in the open "grant market". These include foundations such as the European Commission (Belgium), the National Institute for Food and Agriculture (USA), the British Council for Biotechnology and Biological Sciences (UK), the Russian Science Foundation (Russia), as well as scientific organizations such as the French National Institute of Agriculture farms (France), Institute of Technology in Switzerland (Switzerland), Institute of Agricultural Chemistry and Agriculture (USA). The share of financial investments of these funds and institutions in the global grant market is small relative to global R&D expenditures, but fame and authority allow to consider them the most influential investors in intellectual innovation defining global trends.

The source of information for assessing the publication activity of researchers in the desired areas was the database of publications in English Web of Science Core Collection (hereinafter - the database Web of Science) and the database of publications in Russian eLibrary (hereinafter - the eLibrary database). A Russian publication is understood to mean a publication, the author (or at least one of the co-authors) of which indicated the Russian organization as affiliation. The review includes a comparison in the areas of research “precision farming” and “precision agriculture”, the specificity and frequency of occurrence of individual sub-areas of research in the world scientific literature according to the Web of Science database. The list of information resources used in the preparation of the article is presented in Table 1.1.

Table 1.1: Information resources used in review.

Data base	Type	URL
Dimensions	R&D Data base	https://www.dimensions.ai/
Web of Science	Abstract Data base	https://apps.webofknowledge.com
eLibrary	Abstract Data Base	https://elibrary.ru

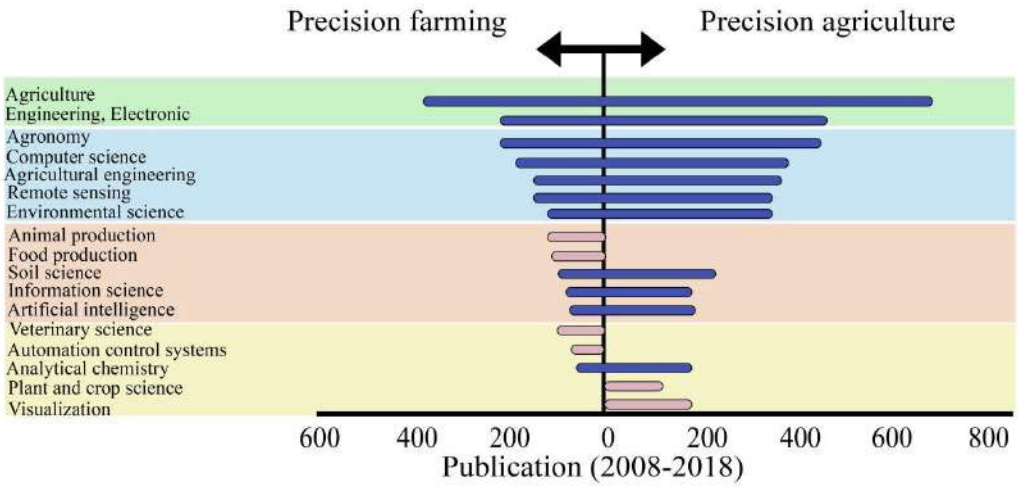


Figure 1-2: Frequency of the use of the terms precision agriculture and precision farming in publications from the Web of Science database from 2008 to 2018 indicating key research areas within the framework of topics.

Results of the review

The key element of the search and identification of works on a given topic is terminology. “Precision farming is an integrated agricultural production system based on the achievements of information technology, the use of automatic control and regulation systems for agricultural machinery and equipment, sensor technology and the general computerization of all agricultural management processes and aimed at optimizing agricultural technologies and stabilizing the productivity with minimal negative environmental impact” [Shpaara D., 2009]. In the worlds scientific literature, two types of key terms are used: “precision farming” and “precision agriculture”, the scope of which are overlapped in many applications. An analysis of the Web of Science database publications over the past 10 years shows that livestock, food, and veterinary medicine are included in the field of precision farming, whereas precision agriculture is characterized by such unique areas as crop production, plant visualization and phenotyping (see Fig. 1-2).

Comparison of publications in identical research areas (for example, agriculture, engineering, etc.) by using in search the keywords “precision farming” and “precision farming” show that among the most highly cited articles are native English speakers (USA, Great Britain, Australia) more often use the terminology "precision agricul-

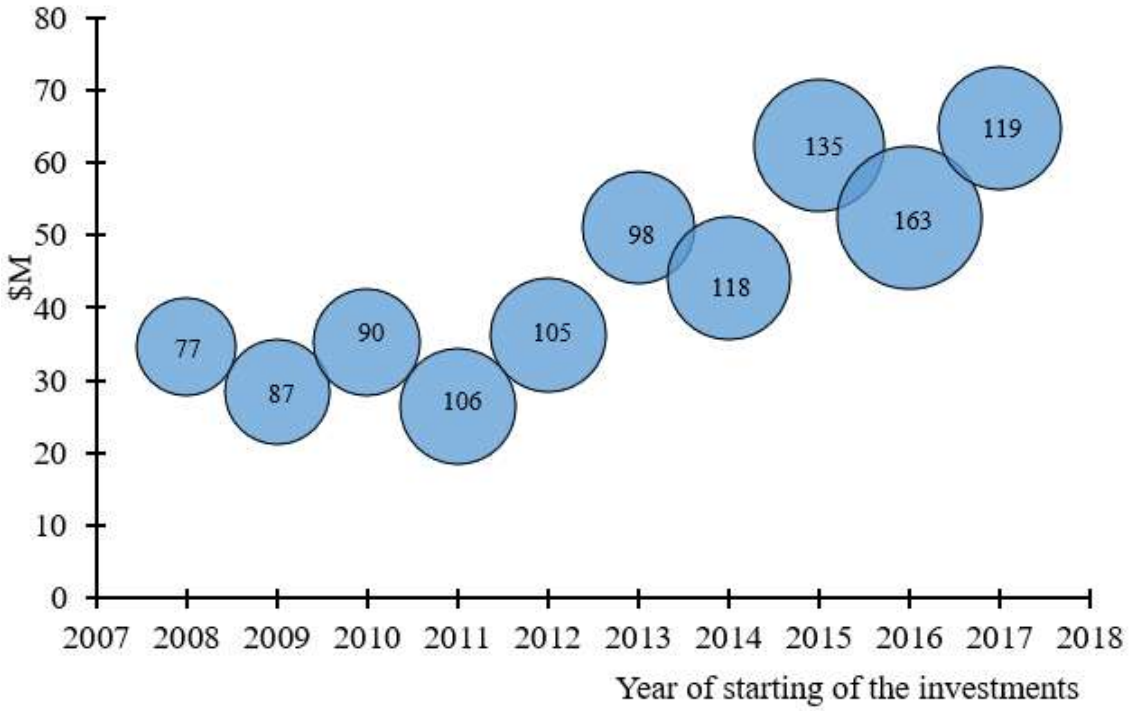


Figure 1-3: Trends in research financing in the field of precision farming and precision agriculture for the period 2008-2017 (The x-axis reflects the year the projects began financing; the y-axis - the amount of funds allocated according to the Dimensions database, million dollars per year; the circle size and the value inside it reflect the number of projects supported during one calendar year, units).

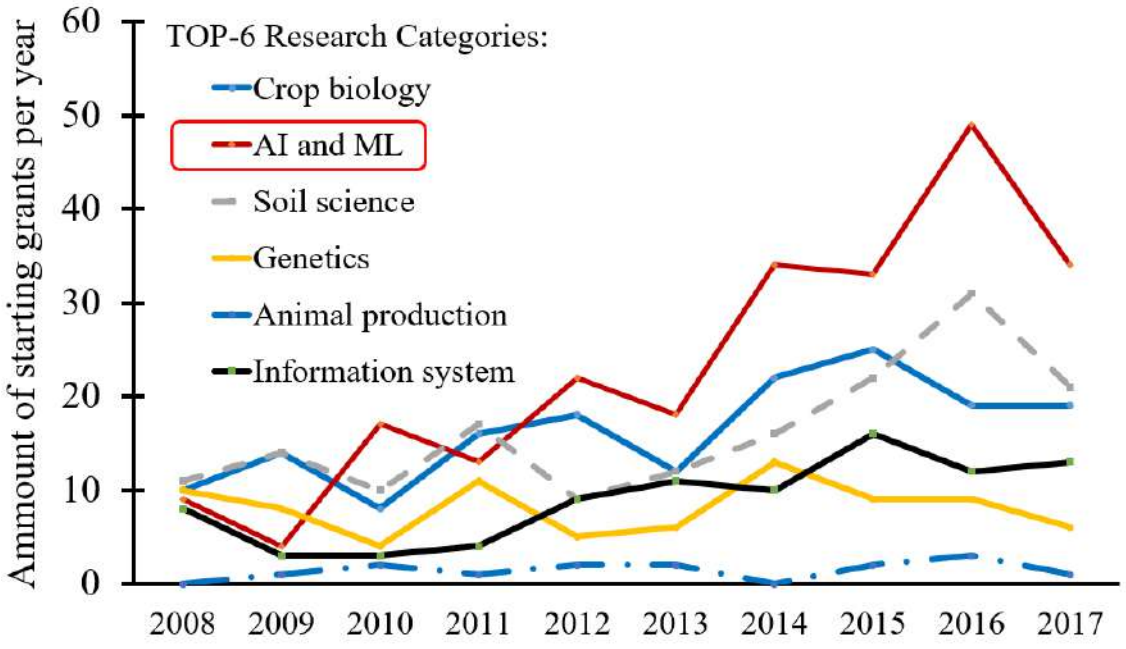
ture", while scientists from non-English-speaking countries (China, India, the EU, Iran, etc.) in similar cases use the term "precision farming". In this study, the search was conducted simultaneously using these two terms to obtain more relevant statistics. Today, precision farming and precision agriculture are in a phase of rapid growth. The trend of the last decade is the increasing interest to developments in this field of science, the total amount of allocated funds for research on the lines of various funds according to the Dimensions database has increased from 34$ million in 2008 to 68$ million by 2017 (see Fig. 1-3)

The study of the structure of supported grants, including an analysis of the average annual numbers of supported research projects for 2008-2017, is presented in Fig. 1-4a. It has been revealed that research in the world is being conducted or is being prepared for carrying out mainly in six areas, conditionally aggregated according to the semantic principle in the following tags: artificial intelligence and data visu-

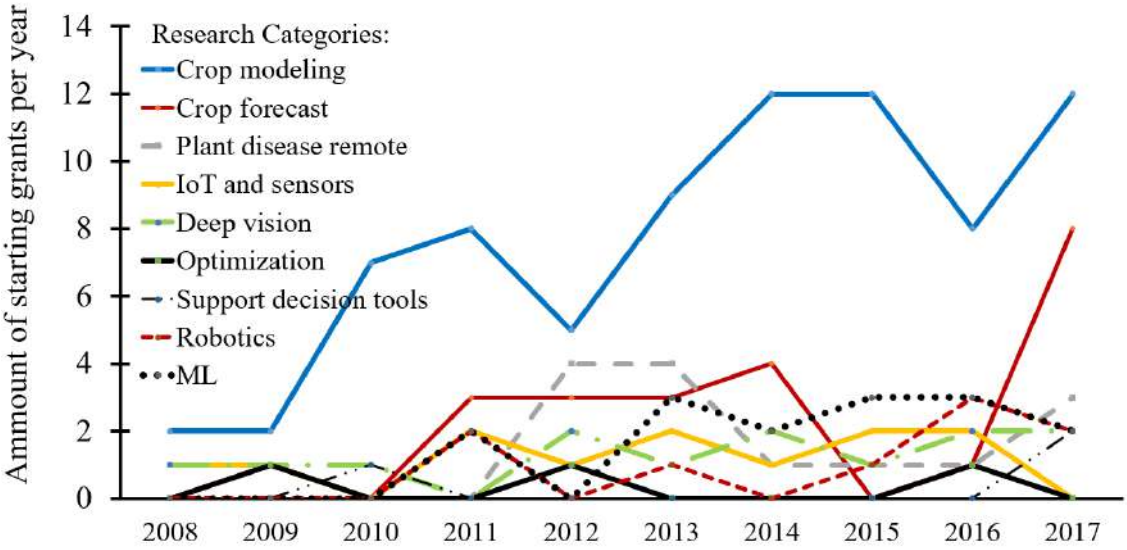
alization, animal husbandry, crop production, information systems, genetics, earth sciences (soil science). This list gives a real picture of the key growth points in the field of precision farming on a planetary scale. There is a constantly increasing interpenetration of related, previously separately developed, scientific fields. Research increasingly goes beyond the scope of one discipline, acquiring the properties of a "scientific composite", becoming in the full sense multidisciplinary.

The main direction of the world agriculture industry of the future, as expected, should be "Artificial intelligence and data visualization". The distribution of the topics of scientific researchers in this area is presented in Fig. 1-4b. These trends may be due to a significant breakthrough in artificial intelligence, robotics and engineering over the past decade. Because of the introduction of advanced computer vision systems, it has become possible to apply automation technologies not only in large agricultural holdings but also in individual farms and small farms. As a positive result of recent discoveries, we can expect progress in optimizing and predicting crop yields for various soil and climatic zones. Among the many supported projects, it is worth noting the Scientific and Research Workshop "Robots for Micro farms" (2017–2021), funded by the European Union as part of the Horizon 2020 program [Robotics Microfarms, 2018]. The ultimate goal of the developers is to create a digital platform for monitoring the status of crops using robotics and a hardware-software complex for supporting decision-making on farm management.

The research direction of remote detecting of plant diseases is actively developing. The prerequisites for this development were a series of qualitative breakthroughs in the field of data processing technologies, as well as multispectral cameras that have a good resolution (1010 × 1010 pixels or more). These tools allow to obtain extensive information on the status of crops in the field, and to identify the development of diseases in the early stages [Behmann et al., 2018, Kuska and Mahlein, 2018, Mahlein et al., 2012, Candiago et al., 2015]. The project "Improving the forecast of risks for precision agriculture: automated monitoring of the spread of pathogenic plants" (2014-2018), aims to monitor the spread of "rust" of wheat leaves caused by Phragmidium or Puccinia mushrooms in real-time [Robotics Microfarms, 2018]. Successful implementation of the project will undoubtedly lead to a significant re-



(a)



(b)

Figure 1-4: The number of supported grants and distribution by priority topics within the direction of precision agriculture and precision farming for the period 2008-2017 according to the Dimensions database (a). The number of supported grants and sub-topics of research in the field of "Artificial intelligence and data visualization" for precision farming and precision farming (b).

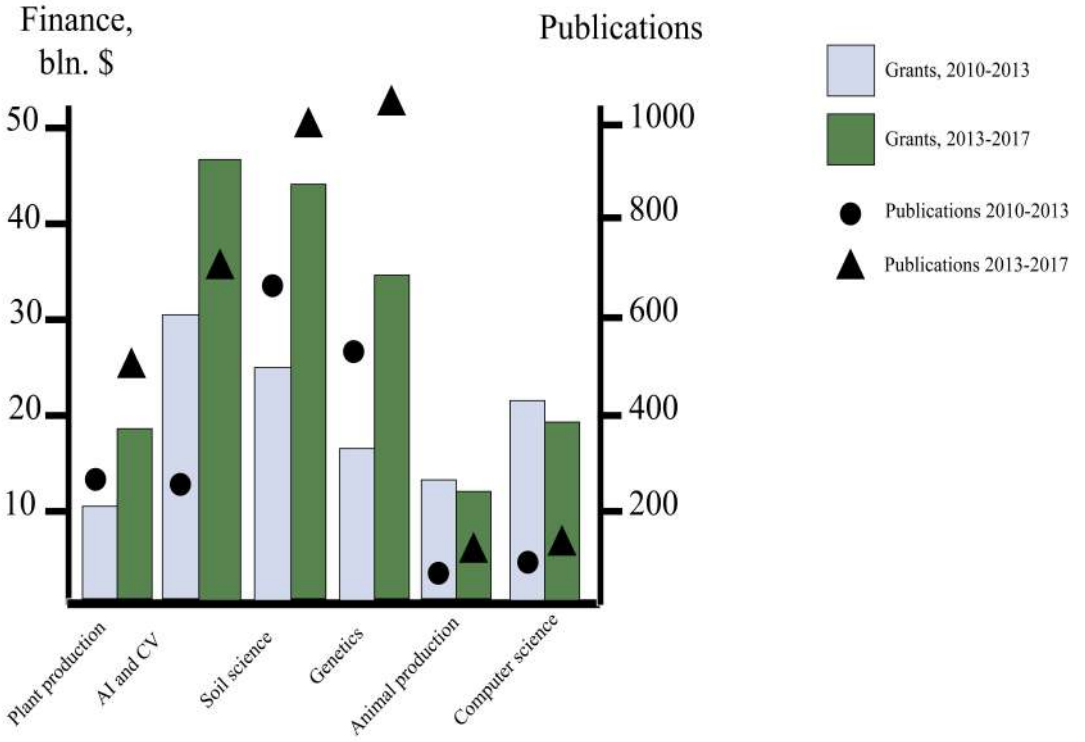


Figure 1-5: A comparative analysis of publication activity (according to the Web of Science database) and project financing of millions of dollars (according to the Dimensions database) in the field of precision agriculture and precision farming.

duction in losses for farmers in the UK associated with crop losses, reaching £ 50 million per year [Mitchell, 2014]. An important role will be assigned not to fight the consequences of plant diseases, but to prevent and early detect, and then, in the long run, to significantly optimize and reduce the consumption of fungicides.

The data were obtained using the Web of Science database show that the number of publications in key areas of research financing in the field of precision agriculture and precision farming is also constantly growing (Fig. 1-5). An increase in funding in the direction leads to a significant increase in publication activity.

Conclusions

The growing use of precision agriculture approaches creates a huge commercial market for the development of the agriculture industry around the world. Over the past ten years, leading countries have invested huge amounts and efforts to gain an advantage in this industry. Obviously, all of these technologies should be intensively

developed in particular the following areas:

1. Development of mathematical algorithms and artificial intelligence systems for automatic plant phenotype and growth optimization.

2. Development of recommendation systems for monitoring and controlling the quality of land resources, as well as livestock and crop production facilities.

Chapter 2

Background

2.1 Internet of Things approach for controlling of artificial growth systems

The constant increase in the Earth's population and the ongoing urbanization impose certain requirements on the amount of food that would suffice and satisfy the growing demands of cities and remote areas. At the same time, food production is limited by the season and the characteristics of each territory. These factors imply severe restrictions on food quantity and its availability. Another limiting factor is the degradation of soils and the inherent lack of adjusting the growing environmental conditions [Turner et al., 2016]. These factors pose an obstacle for many countries to securing a sufficient food production on their territory. Precision agriculture is a technological paradigm that seeks to optimize the outcomes of observing the agricultural system by means of automatizing the processes of observing, measuring and responding to every stage, while keeping the overall control of the growth system to secure its resource efficiency. Undoubtedly, precision agriculture opens up wide vistas for exploiting state-of-the-art technologies that have been successfully applied recently, e.g. remote sensing [Zhou et al., 2016], artificial intelligence [Lane et al., 2017], robotics [Chaudhury et al., 2015], sensor networks [Eugster et al., 2015] and Internet of Things (IoT) [Alavi et al., 2018]. These technologies seem to be a promising path to secure food safety, reduce negative anthropogenic impacts on the

environment, and, ensure the economic profit [Elijah et al., 2018, Taylor et al., 2013].

Artificial growing systems with the hydroponic environment, e.g. greenhouses, are effective enough in maintaining optimal resources consumption and thus in increasing the yields. These systems are flexible enough to adjust the growing parameters all over the year. It results in their efficient operation: they provide 10-12 times higher harvesting from 1 m^2 comparing to the field cultivation [Muñoz et al., 2007]. Originally, the problem of plant growth dynamics assessment in controlled artificial conditions was a crucial point in life support systems development for space and associated ground applications. Although coming from space technologies, developing artificial closed controlled systems for pervasive agriculture is in high demand nowadays, it is expected to guarantee food provision to meet the demand imposed by the increasing population of the world, including the people who live in remote areas or in harsh environments [Wark et al., 2007]. The successful production of vegetables in a typical greenhouse assumes the involvement of an experienced grower. The grower takes the best action after assessing all available factors having impact on the growing process. However, decisions, in this case, are grounded on the experience rather than on science. One of the problems, that should be solved is obtaining comprehensive data in real-time that describe the dynamics of plant growth as well as state of growing system. Accurate and reliable assessment of plant growth dynamics parameters is crucial for the future success of the whole growing system parameters optimization. Such a task may be performable if the Internet of Things approach is used. By means of the recent achievements in the Internet of Things related technologies precision agriculture can become a reality [Miorandi et al., 2012, Sasidharan et al., 2014, Taylor et al., 2013, Wark et al., 2007].

Indeed, high quality and comprehensive monitoring systems are required to perform optimal control [Somov et al., 2012]. Different types of such systems with improved performance have been proposed recently; most of them are typically based on wireless sensor network (WSN) and the Internet of Things (IoT) paradigm involving myriads of sensors in the monitoring process [Spirjakin et al., 2015, Bai et al., 2018, Ferentinos et al., 2017]. For example, the wireless sensor network paradigm application to control the climate and environment conditions in a greenhouse is

reported in [Pahuja et al., 2013b, Mirabella and Brischetto, 2011, Mendez et al., 2012]. Tiny sensors were deployed at different height to measure temperature and relative humidity. If threshold values are violated, actuators are activated to keep the predefined settings in the greenhouse. This approach relies on compact sensor nodes that can be deployed anywhere to perform low-power monitoring tasks and periodically send the obtained measurement to the user or the cloud via a wireless channel. Noteworthy, the sensor nodes can be deployed in difficult-to-access areas without cabling production. It makes them easy to be set up and debugged, and reduces the maintenance costs for monitoring the infrastructure. Many types of sensors were developed to ensure ubiquitous monitoring in greenhouses [Lachure et al., 2015]. The possibility for optimization of WSN systems by their synchronization making possible to distributed systems described in [Macii et al., 2009]. Deep learning based feature representation can also help for processing data and soft sensor development [Yao and Ge, 2018]. Though there is a significant progress in IoT research that described in [Mehra et al., 2018, Ibayashi et al., 2016, Shadrin et al., 2019, Siregar et al., 2017], the successful implementation and deployment of IoT solutions into existing greenhouse infrastructure is still fragmented [Somov et al., 2018].

The main benefit of greenhouses is that all their parameters including the water consumption, nutrients addition, light duty cycle, can be adjusted. Growing plants in artificial soilless system has several advantages over the traditional soil systems, since they guarantee almost a total control on the growing process at different stages [Rius-Ruiz et al., 2014, Cho et al., 2015, Jung et al., 2014, Andaluz et al., 2016]. There is a huge variety of existing designs of soilless systems based on hydroponics, aquaponics, or aeroponics approaches. All of them play a significant role in commercial food production [Lakkireddy et al., 2018]. Moreover, various systems – different in size and features – could be used in a range of application scenarios from industrial greenhouses to small-scale systems for scientific experiments [Resh, 2016, Billings, 2018]. A solid theoretical ground makes it the possible to control the greenhouses in order to provide them with optimal growing conditions. For instance, automated and smart solutions were proposed for running a greenhouse efficiently

using the sensor network paradigm and Internet of Things [Frighetto et al., 2019, Pahuja et al., 2013a, Somov et al., 2018]. Examples of wireless sensing and control systems for aeroponic and for the hydroponic artificial growing systems are described in [Kernahan, 2016] and [Ibayashi et al., 2016] respectively. The design of artificial soilless systems for the industry is becoming increasingly complicated; likewise, the productivity of these systems has been increased significantly in recent years. This became possible due to the wide implementation of optimization technologies in this industry. Such systems proposed in [Montero et al., 2017, Palencia et al., 2016, Putra and Yuliando, 2015, Silva, 2016, Kloas et al., 2015, Harun et al., 2015, Kozai et al., 2015, Kaneda et al., 2015].

2.2 Computer vision and machine learning for plant growth dynamics assessment and prediction

Image based technologies for plant phenotyping

The study of plant growth dynamics responses to the environment is a key component to improve the combination of image-based and dynamic controlled closed artificial systems. Knowing the plant structure and the possibility to study its functioning in an autonomous manner through image processing enable predictive analysis performing and creating recommendation models for growing plant in the best possible conditions under resources constraints [Fiorani and Schurr, 2013, Granier and Vile, 2014, Golzarian et al., 2011]. In-situ image analysis is currently a very popular and well developed method for monitoring and diagnosing large-scale crop fields aimed at optimizing resources consumption [Zhou et al., 2017, Aboutalebi et al., 2018, Duan et al., 2017].

Leaves area and structure is one of the most important characteristics that represents dynamics and wellness of the plant growth. It provides the basics for research in the following areas: plant phenotyping, plant physiology, and plant pathology [Scharr et al., 2016, An et al., 2016, Freschet et al., 2015]. It seems to be a promising way to apply the method to the plant growth dynamics optimization using the

rate of leaves area growth, since the rate fluctuation indicates nutrient and energy resources consumption [Medrano et al., 2015, Kang and Wang, 2017]. Leaves area can also be effectively used as an indicator of the total biomass accumulation in the plant; this parameter, in its turn, can be directly used for modeling the plant growth and for the assessment and optimization of the nutrient and energy resources consumption [Weraduwage et al., 2015]. The advanced tool for automatic leaves structure investigation, in particular, leaves vein analysis using imaging techniques is described in [Bühler et al., 2015]. This tool allows to segment veins and perform quantification analysis of their properties. However it is still challenging to develop the automatic, non-destructive, universal, scalable, robust and precise method for leaves area assessment and prediction. The major bottlenecks are high variety of species, huge amount and complexity of the underlying processes and stresses which have an effect on the output (i.e. leaves area) [Singh et al., 2016, Campbell et al., 2018].

In greenhouses and indoor farming image-based technologies have been being implemented since recently. The 2D approach is often used if the plant is characterized by large leaves and a simple structure. However, it typically relies on a complicated software to perform the analysis and suffers from leaf overlap and concavity [Li et al., 2014, Minervini et al., 2015, Ghanem et al., 2015]. 2D approach perfectly works for in-situ investigation of plant growing phenomics at the initial stage. Computer vision and machine learning based solutions do have their advantages not only in assessing the plant growth phenomics, but also in assessing fruit characteristics [Pouladzadeh et al., 2014]. Training machine learning algorithms allow for a deeper understanding of the dependencies in plant growth systems. At the same time, dynamics predictions can be based on the high-quality plant images and the data associated with the conditions of the plant growth. Quite a number of reviews describe the development and application of image-based technologies for analysing the plant structure and functioning. Generally, most plant phenotype CNN-based algorithms for object detection or segmentation are similar to the one described in [Ubbens and Stavness, 2017, Dai et al., 2015, 2016, Scharr et al., 2016, Gu et al., 2015].

Development and validation of computer vision algorithms require good quality annotated database. The most popular first comprehensive benchmark data that can be used for typical computer vision tasks was obtained by [Minervini et al., 2016, Cruz et al., 2016]. The previously obtained datasets that can serve for solving a smaller range of computer vision problems were described in [Silva et al., 2013, Nilsback and Zisserman, 2010]. High demand for benchmark data is reflected by the frequency of their usage. Thus, for example, an open-source datasets were used for evaluating the precise recurrent instance segmentation algorithm in which the end-to-end RNN architecture with an attention mechanism was proposed [Romera-Paredes and Torr, 2016, Ren and Zemel, 2017]. Open-source datasets were used for evaluating of the instance embedding approach where pixels of an object are encoded into vectors and clustered using the popular mean-shift algorithm [De Brabandere et al., 2017]. Since the process of an annotated dataset preparation is time-consuming and not always precise, (e.g leaf masking), computer-generated models or so-called synthetic plants can help to overcome this problem [Giuffrida et al., 2017, Ubbens et al., 2018]. Also, imaging systems may be costly, thus, affordable hardware and a software setup for plant phenotyping are currently in a great demand. One of the proposed systems with such features provides the possibility to count the leaves area using a robust machine learning algorithm [Minervini et al., 2017]. Another one can perform time resolved analyses of plant growth which is also essential for understanding growth phenotypes [Dhondt et al., 2014].

3D imaging

Another set of approaches is based on 3D imaging. This approach helps to capture the plant shape in three dimensions and study it. Undoubtedly, it may seem tempting to process 3D images of the plant growth as we can derive more information about the plant structure in comparison with 2D images, but the systems for receiving precise 3D imaging data are typically on several orders of magnitude more expensive than 2D imaging systems. Laser scanning is also used for plant digitization and has been successfully applied to forestry and statistical analysis of canopies [Paulus et al., 2014, Yang et al., 2013]. Its application is limited to extract-

ing single plant attributes due to these tasks are computationally intensive. The 3D scanning system for taking quick and accurate images is proposed in [Nguyen et al., 2016]. The approach involves two tilting cameras, the methods for camera calibration and background removal. Quite a similar approach, where the authors use a robotic arm equipped with a 3D imaging system for 3D plant growth measurement is proposed in [Chaudhury et al., 2015]. The drawback of the invention is that it requires much time for processing and data recording. A semiautomatic 3D imaging system for plant modeling is reported in [Quan et al., 2006]. The bottom line of this research is to combine reconstructed 3D points and the images for guarantying a more effective segmentation of the data into individual leaves. Although the 3D imaging approach is getting popular, the image acquisition for 3D reconstruction is typically carried out manually [Paulus et al., 2013]. A 3D phenotyping platform for laboratory experiments was successfully developed and a 3D dataset in couple with environmental information was obtained in one of the most recent works [Uchiyama et al., 2017]. Different approaches for 3D plant reconstruction are described in [Gibbs et al., 2017, Liang et al., 2013, Pound et al., 2014, Vázquez-Arellano et al., 2016, Chaudhury et al., 2015].

Machine learning for modeling of growth dynamics

In fact, societal concerns about food safety and the environmental impact resulted in the growing interest in the application of artificial intelligence in agriculture[Eli-Chukwu, 2019]. The recent advances in data science and machine learning for constrained devices are vital for making real-time inference procedures and prediction [Davies and Clinch, 2017b, Lane et al., 2017]. At the same time, it requires real data for providing high-quality inference to the user. Indeed, AI opens up wide opportunities for more accurate monitoring or optimization if based on computer vision (CV), deep learning (DL), and machine learning (ML) methods. For example, CV is vital for plant health monitoring, overall biomass assessment, and non-destructive measurement of the elements content in plants [Lin et al., 2013, Chaudhury et al., 2015, Mao et al., 2015].

When it comes to the prediction of plants growth dynamics, there is a lack of

robust universal models available for the quantitative prediction of plant biomass changing with time. Although there are mathematical models to be applied to direct simulation of plant growth (the so-called 'bottom-up' approach [Rodríguez et al., 2015]), most of them are based on solving the systems of differential equations and involve large numbers of (semi) empirical parameters. Therefore, they have to be adapted to each specific type of plants, as well as to the cultivation technique. This makes them sensitive to some hidden changes in the environmental and other conditions that are difficult to track [Vereecken et al., 2016]. In fact, technically, in the remote areas it is almost impossible to obtain all the necessary parameters for making a good quality predictive model to assess the plant growth dynamics based on the "bottom-up" approach.

The good examples of using deep learning for plant growth dynamics in particular, yield prediction in a greenhouse was made using RNNs based on the former yield and stem diameter values, as well as microclimate conditions [Alhnaity et al., 2019, Hochreiter and Schmidhuber, 1997]. Examples of predicting the greenhouse environmental conditions, such as CO_2 concentration, temperature, humidity based on recurrent neural network were proposed in [Jung et al., 2020]. The possibilities of using deep learning methods for solving challenges in precision agriculture are presented in a comprehensive review [Kamilaris and Prenafeta-Boldú, 2018]. The conclusions drawn by the authors were that in most cases deep learning outperformed the existing regression and classification models, as well as classical computer vision methods. The productivity of deep learning methods for dynamic optimization of water temperature is reported in [Yumeina et al., 2015]. Another type of machine learning methods that can be useful for precision agriculture is reinforcement learning. It can be used for modeling and optimizing the duty cycle of artificial light, and predicting the plant growth, where 2D imaging is often used to perform the leaves analysis and to make the associated inference [Somov et al., 2018, Rajendran et al.]. However, there is still a gap between modeling and experimentation: the reason is that many more experimentally collected data are required for modeling of agriculture-related scenarios [Rötter et al., 2018].

Non-destructive methods for assessment of seeds germination

Another task for agriculture is optimization seed germination process. It includes a number of interconnected processes that influence the optimal plant growth. The problem of seed germination modeling under different conditions was tackled in [Bello and Bradford, 2016]. However, at present the research in this area remains fragmented. One of the aspect of germination process, in particular the connection between the oxygen consumption and the germination rate of roots, as well as a presumable method to obtain such measurements was described in [Lee et al., 2017]. Several dynamic models of seeds germination are discussed in [Forcella et al., 2000, Bello and Bradford, 2016]. The computer vision system for monitoring the germination time course of a sunflower is described in [Ducournau et al., 2005]. The comprehensive study on the root growth response (temporal and spatial) to the limited nutrient availability based on modeling and experimental results is shown in [Postma et al., 2014]; this paper stated which of the parameters that are defining the root growth should be used to maximize crop production. The way to use k-nn to analyse the images of diverse germination phenotypes and to detect single seed germination in *miscanthus sinensis* was shown in [Awty-Carroll et al., 2018]. The study of low-level root phenomics was described in [Miyamoto et al., 2001]; the paper investigated hydraulic conductivity of rice roots. The use of 2D and 3D image-based technologies for in-depth examination of root-soil interactions is described in [Gregory et al., 2009]. The methodology for high precision three dimensional imaging of roots in soil based on the magnetic resonance imaging is presented in [van Dusschoten et al., 2016]. This technology allows non-destructive monitoring of root parameters, performing quantitative analysis and deriving spatial distribution of roots, which opens a wide range of possibilities to perform fundamental research on the root growth response for different environmental conditions and stresses. The comparison of magnetic resonance imaging with X-ray computed tomography for reconstruction of the three dimensional root structure in soil is presented in [Metzner et al., 2015]; the paper states that these methods have benefits for application on different scales. Although the proposed and evaluated methods are very precise, they are very expensive. A variety of methods that can be coupled with the image-

based sensing technologies for the roots phonemics are described in [Das et al., 2015]. However, there are no publicly available benchmark datasets for investigating the seeds germination process. Still, from the viewpoint of testing, it is vital to rely on a relevant benchmark dataset. Recently, such a tool for monitoring germination process and obtaining relevant datasets was developed by [Falk et al., 2020]. The main idea of this research is to use fully convolutional neural networks to perform segmentation of a germinated seed.

Plant diseases detection

The other important task for precision agriculture is plant diseases detection. Automatic systems for visual monitoring, along with the newest machine learning techniques, are very powerful tools for monitoring and assessing the quality and quantity of agricultural production. At the same time, these systems could help farmers to decrease their yield losses due to remote monitoring. The possible application of plant growth optimization by advanced non-invasive technologies for disease detection has been previously described in [Hanan, 2017, Park et al., 2011, Mahlein et al., 2012, Rumpf et al., 2010]. Nowadays, farmers are going to improve the overall quality and quantity of their harvest, to predict maturation rate, and to check the productivity of the plantings, while decreasing the workload [Fan et al., 2011]. Harvest losses because of pests and diseases are a major threat that costs billions of dollars every year [Rubatzky and Yamaguchi, 2012]. Meanwhile, most of the current approaches for monitoring and detecting diseases performed manualy, result in approximate assessment. Special tools moreover are time-consuming and require many of human resources. In addition, workers without special equipment can hardly cover and analyze all plantations or greenhouses due to their large sizes. This problem leads to a loss of important information about the epicenters of new plants diseases and their real diversity. In addition, it is challenging to build a comprehensive map of diseases that occurs in space and time. Diseases in a plant can occur quickly in real field environments: e.g., common fungal diseases such as apple scab can attack the plant in two weeks, thus, only real-time monitoring can help to detect diseases at early stages to apply fungicides promptly [Vanderplank, 2012, Sophie et al., 2010].

Computer vision systems go beyond human capabilities and help to evaluate long-term processes and events occurring in the whole electromagnetic spectrum. Among them, hyper-spectral systems provide the features that can be used as fingerprints of plant diseases at a certain wavelength. This finding can be used as a tool for developing new computer vision systems adapted for specific agriculture purposes. The importance of the fluorescence imaging systems and multispectral imaging for deriving features that are related to plant health and properties are described in [Fiorani et al., 2012]. It is also important to discriminate one disease from another one by using classification approaches jointly with image processing [Khan et al., 2019, Pantazi et al., 2019, Sarfraz, 2014, Patil and Kumar, 2011]. Nowadays, deep learning models for plant disease detection, that include long-short term memory neural networks (LSTM) coupled with convolutional neural networks, in particular for the detection of apple scab, are showing good results [Baranwal et al., 2019, Turkoglu et al., 2019, Ferentinos, 2018]. Neural networks have also shown their usefulness for plant disease detection based on the hyper-spectral data [Golhani et al., 2018]. There are several recently published datasets that allow training deep learning models, such as [Parraga-Alava et al., 2019, Nouri et al., 2018]. However, in the available near-infrared hyperspectral dataset, there are not so many obtained data and spectra for training ML algorithms, and also, their bandwidth are narrow [Nouri et al., 2018]. The use the hyper-spectral data can improve the accuracy if the specific spectra for diseases detection are known. In addition, it is not necessary to collect a huge dataset, as the features have already been extracted. This allows using simpler algorithms to detect plant diseases, which in turn enables a low-power implementation in embedded devices.

Embedded systems

Most of the data-driven approaches proposed in the literature are computationally heavy, so the actual infrastructure of greenhouses would hardly allow to run such systems directly because complex transmission systems and data processing systems are to be created. This leads us to the challenges associated with the system autonomous operation. In terms of an autonomous operation, it is the limited en-

ergy storage and high-power consumption that are of great concern. To address the above problem, a distributed low-power embedded solution with AI on board is required. An essential pre-requisite for solving this problem is artificial intelligence which is able to run on a low-power embedded system. This solution does not require local powerful data processing or complicated data transmission to the cloud which is restricted in the wireless sensor networks applied to the monitoring tasks in greenhouses [Pahuja et al., 2013a]. As it was mentioned, the bottleneck could be addressed by the application of edge computing paradigm [Shi et al., 2016]. It performs the data-intensive computation tasks onboard of nodes without involving massive data transmission to the remote server. The ultimate advantage of these systems is a significant reduction in the output data size. For example, instead of sending images to the cloud server (in the range of few MB), edge-computing systems generate the post-processed data, e.g. segmentation masks or the text files with coordinates, labels of objects on the image and quantity characteristics of interest (in the range of few KB) which are used for further system control. Also, such a distribution improves the reliability and prevents the data losses caused by the blackouts or hackers attacks to the server. Nowadays, edge computing is a very relevant topic, because of an intensive development of mobile platforms, IoT, robotics, embedded systems, wearables, etc. Edge computing has multiple advantages in comparison with cloud computing and fog computing due to the following technological limitations: privacy issues, dependency on the internet connection, delays associated with network latency, the number of possible clients depends on the servers computational capabilities [Marantos et al., 2018, Ignatov et al., 2018]. This is why the application of variety fitting approaches of ML algorithms to mobile and embedded platforms with a limited computational capacity is necessary also for overcoming the above-mentioned issues. There are many ways to fit the machine learning algorithm on board of edge computing platforms. They may include hardware acceleration (HA) by digital signal processor (DSP) [Codrescu et al., 2014] or graphical processing unit (GPU) [Latifi Oskouei et al., 2016]. The DSP HA is widely used in mobile platforms due to their high performance along with a low power consumption (even in comparison with CPUs and GPUs). The GPU HA im-

plies parallel computations split between CPU and GPU. It could be implemented with the following libraries: TensorFlow Mobile [TFM, 2019], Android neural network API (NNAPI) [NNA, 2019], RenderScrpit-based CNNdroid [Latifi Oskouei et al., 2016] and RSTensorFlow. The latter is a GPU-based accelerator of matrix operations, which makes it possible to accelerate matrix multiplication up to 3 times [Alzantot et al., 2017]. Furthermore, some studies show that RenderScript could be used even with CPUs imprecise computing modes to lower execution time of computationally-intensive models [Motamedi et al., 2019]. In addition, System-on-Chip (SoC) manufacturers propose SDKs compatible with their products only. They include SNPE by QUalcomm, HiAI platform by HiSilicon, NeuroPilot SDK by MediaTek, etc.

Bottlenecks of ML and CV application

There are several Bottlenecks of the CV application. ML and CV methods are essentially black boxes, so it is often impossible to properly interpret their results from the physical point of view [Minervini et al., 2015, Alhnaity et al., 2019]. The state-of-the-art DL methods that are suitable for performing phenotyping tasks are quite heavy from the computational point of view which significantly limits their application in real scenarios. A huge amount of data has to be obtained and annotated for training neural networks (NN) powering the DL approaches, which is usually a time-consuming task. Additionally, the computational infrastructure in some greenhouses is not powerful enough to run the NNs. This is why a method for plant growth dynamics prediction that requires less training data and is computationally efficient as well as accurate enough, is in a high demand now. One of the possible solutions is using computationally simple algorithms that rely on the extracted features of plant growth. Such algorithms can be based on the Kalman filters that have already demonstrated their efficiency in a wide range of applications, e.g. sensors networks, control, vehicle trajectory tracking, interferometry, radar tracking [Kalman, 1960, Zhou et al., 2018, Pletschen and Diepold, 2017, Xia et al., 2018, Nilsson et al., 2015, Kulikov and Kulikova, 2015]. However, they have limited applications in precise agriculture where they were exploited in the crop

phenology evaluation [Vicente-Guijalba et al., 2013], in improving the prediction performance of the deterministic model [Ruíz-García et al., 2013], or crop disease detection [Hamuda et al., 2018]. It was reported an approach based on Kalman filter to monitor biomass evolution in plant cells that are featured with very low volume of samples [Albiol et al., 1993]. From this perspective, using Kalman filters as a computational core of a plant dynamics prediction system is a promising approach in terms of accuracy and computational efficiency.

2.3 "Bottom-up" modeling of plant growth: review, examples

In this section the "Bottom-up" modeling are reported to show the advantages and disadvantages of this approach, basing on the theoretical models described in [Bessonov and Volpert, 2006]. One of the typical approaches to describe plant growth dynamics is modeling of kinetics. To develop the basic model of plant growth it is necessary to define the most important physical principles that underlying plant growth. To show such principles let us consider the growing plant in one dimension (vertical stem). The plant takes nutrients from the bottom (roots), then these nutrients are transported to the top of the plant, where metabolites are generated. Using these metabolites and nutrients, cells are fissioned and stem is growing. Light also plays a very important role in plant growth, but in the following modeling, it is supposed that the amount of photoactive radiation (PAR) is enough to support all processes in plants. Also, it is supposed that nutrients are not consumed by roots development. From the biological point of view it is assumed a constant and narrow width of the meristem (the tissue on the top that is responsible for cell division). After cell division, newly divided cells are becoming new meristem on the top, while the previous meristem layer is becoming common tissue of the plant.

Let L be the length (height) of the plant, which is much larger that the width. The concentration of nutrients in the point $x = 0$ is fixed (it is supposed that there is unlimited access to the nutrients and no growth of the roots). In the point $x = L(t)$, which represents the top of the plant, metabolites are creating and the

plant is growing. Increasing the length of the plant can be described using the following equation:

$$\frac{dL}{dt} = f(R), \tag{2.1}$$

where R is the concentration of metabolites. Cells that are responsible for transmission of the nutrients to the top of the plant are located in the internal part of the plant $0 < x < L(t)$. Nutrient concentration C in this part is depending on x and t. It is supposed that diffusion-advection equation can describe the process of transmitting of the nutrients to the top:

$$\frac{\partial C}{\partial t} + v\frac{\partial C}{\partial x} = d\frac{\partial^2 C}{\partial x^2}, \tag{2.2}$$

where the diffusion coefficient is d and v is the nutrients (fluid) transmission velocity. As the nutrient solution is assumed to be incompressible and to have uniform distribution in plant, v can be presented as:

$$v = \frac{dL}{dt}. \tag{2.3}$$

The following boundary conditions should be introduced into the Eq. (2.2) according to our assumptions:

$$x = 0 : C = 1; x = L(t) : d\frac{\partial C}{\partial x} = -g(R)C. \tag{2.4}$$

The boundary condition $d\frac{\partial C}{\partial x} = -g(R)C$ represents the nutrients flow from the internal part of plant to the top boundary. It depends on the nutrient concentration $C(L,t)$ and function $g(R)$ that is responsible for controlling the cell divi-

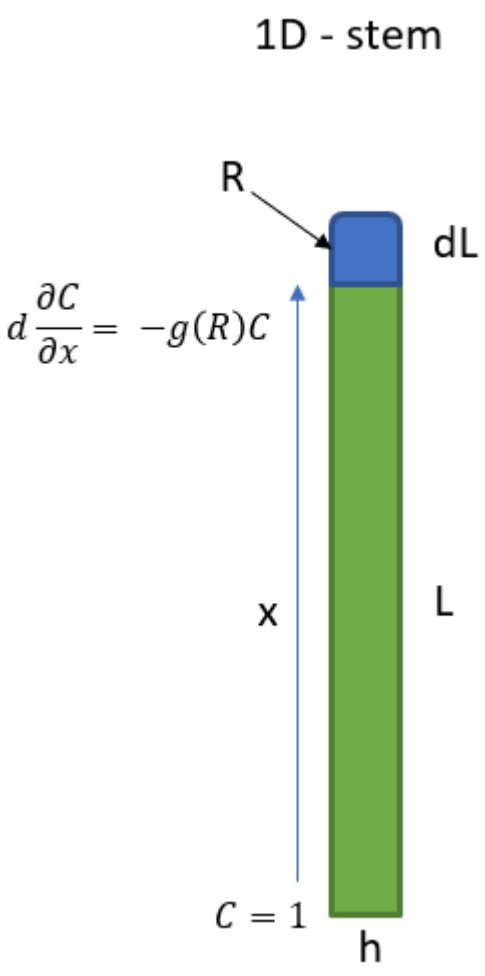


Figure 2-1: Graphical sketch of the solved mathematical problem.

sion.

The last equation that should be provided to complete the system of differential equations is that describing dynamics of changing of metabolites concentration R, Eq. (2.5):

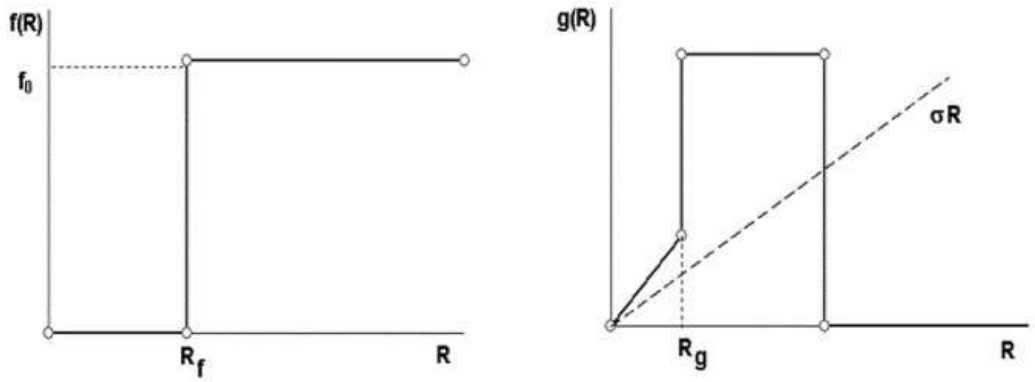
$$h\frac{dR}{dt} = g(R)C - \sigma R, \tag{2.5}$$

where $g(R)C$ represents the dynamics of the creation of metabolites, while σR represents the consumption by the divided cells, h is width of the plant. It was made the assumption that the rate of changing of metabolites concentration R depends on the metabolites concentration of R.

Overall, it was obtained a system of differential equations that allow modeling the growth dynamics of one dimensional plant (stem) with the assumptions. R is produced only by mitosis and defines the plant growth rate. There is a continuous nutrient exchange and converting on the top boundary. The generating of R depends on the certain function $g(R)$ and nutrient concentration C. The schematic representation of the developed mathematical model is presented in the Fig. 2-1.

It will be shown in the following examples of modeling that the shape of $g(R)$ has dramatic influence on the total result of modeling. Function $g(R)$ assumed to have piecewise shape as the process of generating R is assumed to be autocatalytic (see Fig. 2-2b), while f is the step function. It is supposed that there is no growth ($f = 0$) until R reaches certain threshold R_f and after that $f = const$ (see Fig. 2-2a). These assumptions have the biological basis as the process of generation of R can be self-accelerating.

Simulations. The simulations of the plant growth dynamics based on the model described above (1D case) were performed in the *Matlab* environment. First, it was investigated the influence of critical parameters on the result of simulation. There are two critical parameters in functions $f(R)$ and $g(R)$: R_f and R_g. If $R > R_f$ then plant grows, if $R > R_g$ then the generation of metabolites is accelerated. There are two different simulation scenarios for different relations between R_f and R_g. For simulations the following values were set constants $d = 0.001\frac{su^2}{tu}$ and $\sigma = 0.009\frac{su}{tu}$,



(a) (b)

Figure 2-2: Shapes of functions (a) $f(R)$ and (b) $g(R)$, Source: [Bessonov and Volpert, 2006].

$h = 0.001su$, $g_0 = 0.01\frac{su}{tu}$, $h_0 = 0.0031\frac{su}{tu}$, where su is space unit and tu is time unit.

Figure 2-3 shows an example of simulations in the case if $R_f > R_g$: $R_g = 0.01$, $R_f = 0.08$. It can be observed the linear growth and when the length approaches the maximum length, the growth stops. The concentration of the metabolites and nutrients at the apex monotonically decrease and tend to zero.

If $R_f < R_g$: $R_g = 0.01$, $R_f = 0.08$ (all the other parameters are the same as above) it can be observed the periodical growth (see Fig. 2-4). This happens because it is needed time to accumulate the metabolites and to exceed R_f to trigger the growth. Increasing the length of the plant the time period for accumulation also increases (see Fig. 2-4).

In order to investigate the sensitivity of the model to the parameters simulations were performed to derive dependencies of the maximum length on the various of the estimated parameters for the case of the periodic growth ($R_f < R_g$). The results about dependence of the maximum length on diffusion coefficient d, width h and parameter g_0 for various R_g values are presented in Fig. 2-5(a-c) correspondingly. All the other parameters remain constant and their values are the same as above. From these results it can be noticed the high sensitivity of the maximum length to the model parameters. For example, if $R_g = 0.1$ and the diffusion coefficient changes from 0.0005 to 0.0014 the final length is increasing from 2.17 to 6.72 and for the same $d = 0.0014$ if the R_g is double from 0.1 to 0.2, the final length is increasing from 3.45 to 6.72 (see Fig. 2-5a). The maximum length monotonically decreases

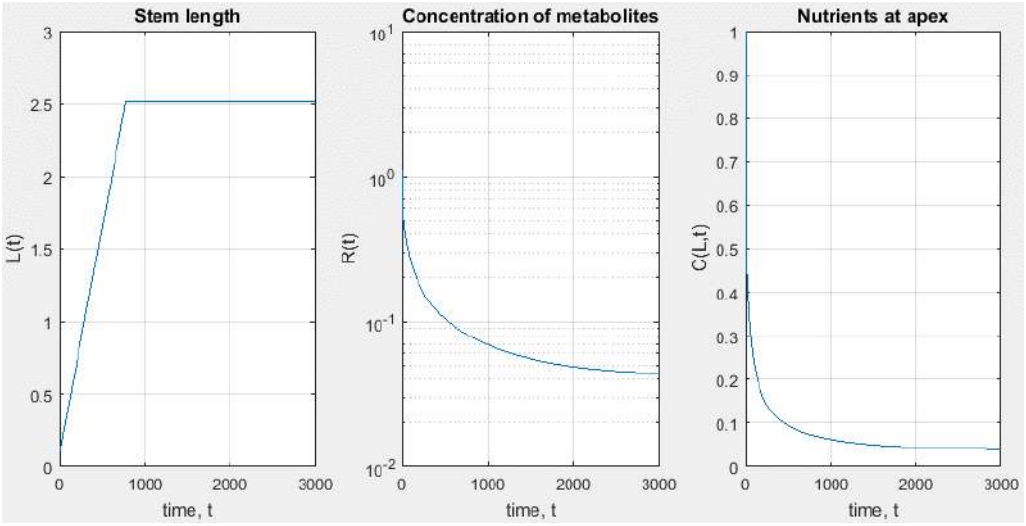


Figure 2-3: $R_f > R_g$, Linear growth of stem length with corresponding changing of the concentrations of metabolites and nutrients concentration at apex.

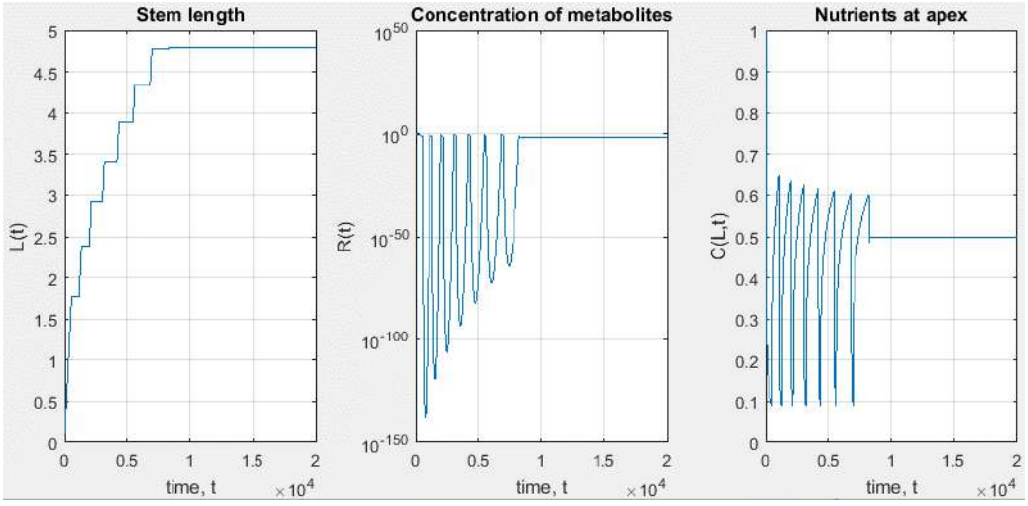
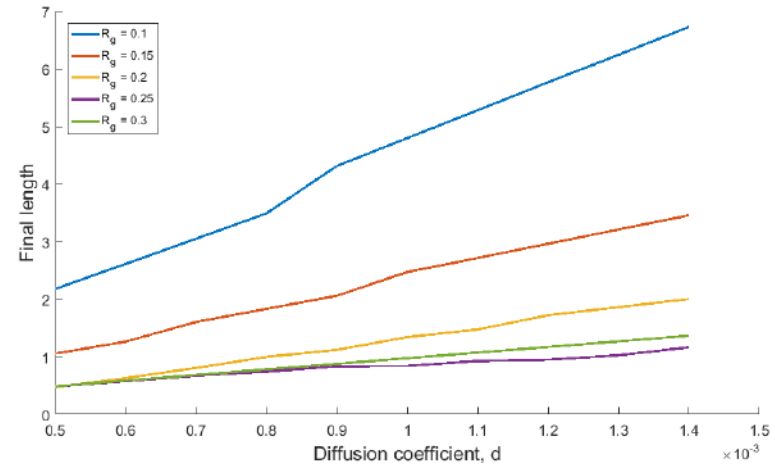


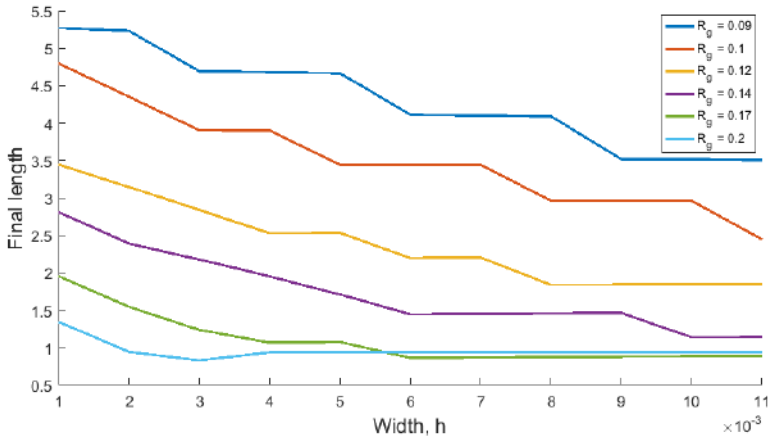
Figure 2-4: $R_f < R_g$, Periodic growth of stem length with corresponding changing of the concentrations of metabolites and nutrients concentration at apex.

with the increase of width h for almost all investigated values of the parameter R_g (see Fig. 2-5b). For example for $R_g = 0.1$, the maximum length halved from 4.79 to 2.44, while h is changing from 0.001 to 0.011. It can be observed the high sensitivity to the values of R_g for the same value of h. If $h = 0.001$, for $R_g = 0.1$ maximum length is 4.79 and for $R_g = 0.2$ the maximum length is 1.34. Finally, modeling showed the sensitivity of the maximum length to the small values of g_0. With the increase of g_0 from the 0.003 to 0.01 for $R_g = 0.1$, the maximum length increases from 3.64 to 5.17 (see Fig. 2-5c).

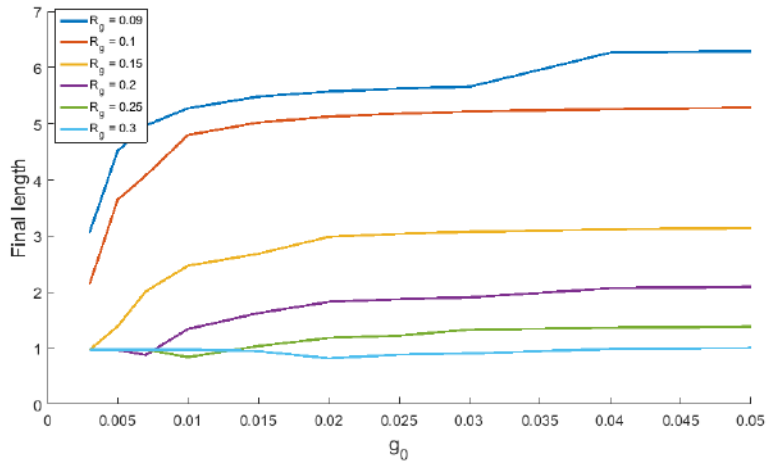
The performed modeling based on the "Bottom-up" approach which includes differential equations showed that there are lots of uncertainties and it is difficult to obtain a robust prediction even for 1D case. Here lots of assumptions were done on shapes of curves, critical values, parameters of the model. The complex processes of photosynthesis and respiration also were not included. In real life all these parameters are unknown. So it is hard assess them with the necessary precision for perform modeling that will give the meaningful result, and also they can vary for different plants. For 2D cases and plants with branching, it was also shown, that solutions are not always stable in [Bessonov and Volpert, 2006]. All these make this approach almost impossible to use for growth dynamics assessment of the real plants or it can be used for a narrow range of tasks. However, the power of this approach is in the possibility to do the "low-level" modeling and theoretically investigate any case and influence of any parameter on the growth dynamics.



(a)



(b)



(c)

Figure 2-5: Results on the modeling of the maximum length dependence on (a) diffusion coefficient d for various R_g; (b) width h for various R_g and (c) parameter g_0 for various R_g, in the case of periodic growth: $R_f < R_g$.

Chapter 3

Hybrid combination of methods for modeling of plant growth in controlled environments

3.1 Experimental setups, collection of relevant data from experiments

Experimental Setup

Artificial growth and monitoring systems. The experimental setup was designed and created based on the greenhouse hydroponic system. The main advantage of hydroponic systems over the open soil systems is the ability to control almost all the conditions influencing the plant's growth rate. In addition to that, the plant response time is much faster, since its roots are always in the direct contact with a nutrient liquid solution, so one can find and evaluate the impact of the individual chemical compounds on the growth process and invesigate the quantitative effect of each parameter on the plant growth. Also, it is possible overcome perception-action problems that typically occur when plants are grown in the soil. Moreover, the hydroponics allows for the exploration of the plant reaction on different environmental conditions or maintain the particular state for a long period of time. It creates the opportunity for carrying out a variety of experiments and collecting all the relevant

data about the plant growth dynamics. It also enables the optimization of the whole growth process. The experimental setup is shown in Fig. 3-1.

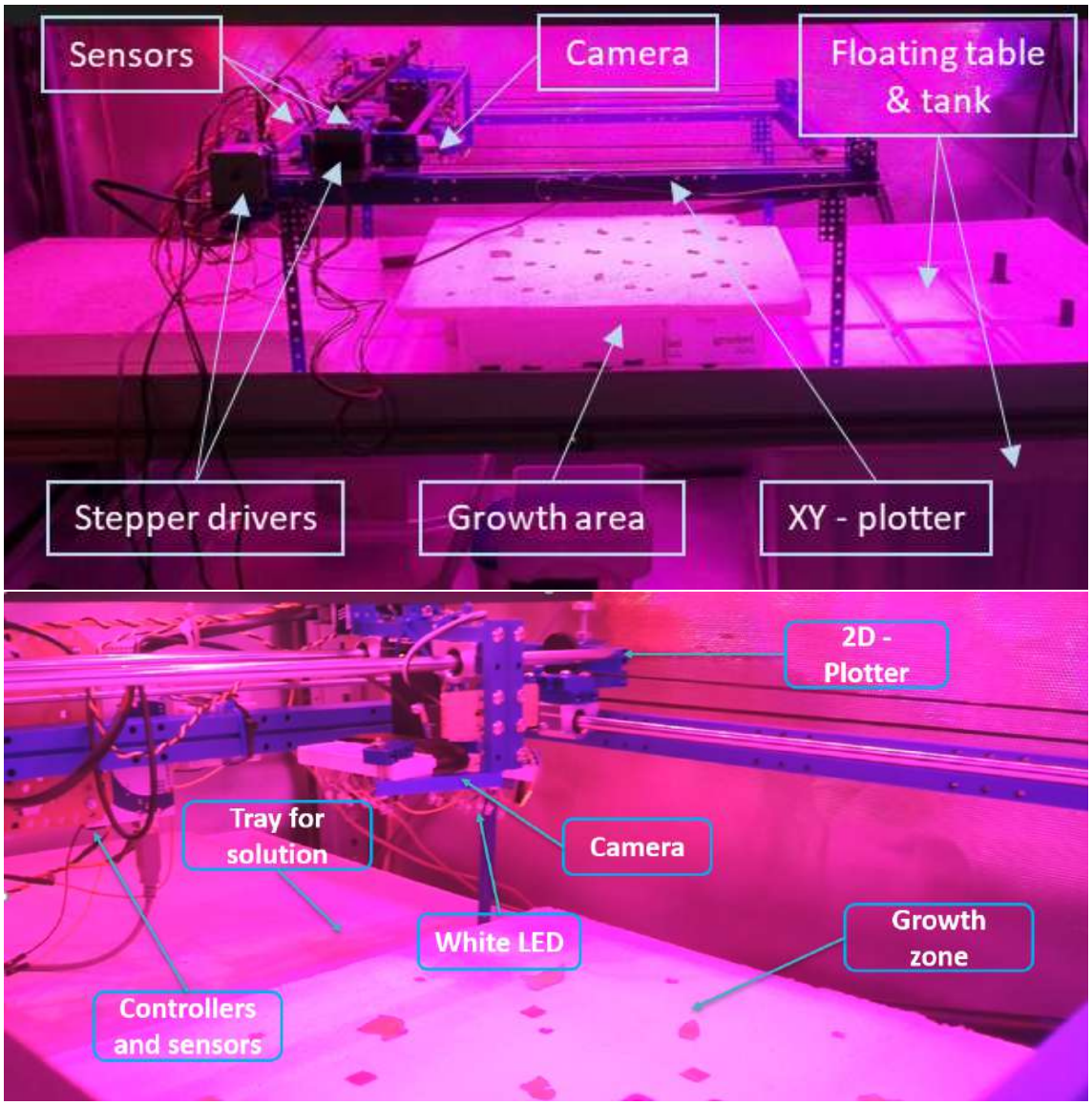


Figure 3-1: The experimental greenhouse hydroponic system enabling the monitoring of growth dynamics and system parameters in real-time.

The testbed consists of two subsystems: the first one is for growing plant (hydroponic system) and the second one is for monitoring growth dynamics and system parameters. Plants nutrition was provided by the constant feeding layer technology realized through recycling of a nutrient solution on a floating table. In the construction 1 cm feeding layer was provided by a 10 $Watt$ pump and a 50 l tank for satisfying requirement for such systems that it should be 1-2 full recycle of nutrient

Feature	Value/description
Max. amount of plants	20
Illumination	150 *Watt* multispectral LED
Feeding solution recycle	60 liter tank, 10 *Watt* pump and 1.5 *cm* of feeding layer
Substrate	0.65 liter rock wool blocks
Fertiliser	Flora NOVA produced by GHE

Table 3.1: Hydroponic growth system design summary

solution per one hour. Feeding solution was prepared by using the recommended recipe of popular commercial fertiliser concentrate *Flora NOVA* produced by company GHE. Each plant grew in a 0.65 l (10x10x6.5cm) rock wool substrate. The testbed possesses the ability to grow up to 20 small plants, e.g. dwarf tomatoes. The amount of simultaneously growing plants depends on a purpose of research. For identifying the nutrient uptake it makes sense to put into the system as much plants as possible: it will be easier for sensors to detect the dynamics of parameters changing, e.g. changing of pH, EC, temperature. In the case of imaging it is recommended to put plant sparse for avoiding overlapping of leaves. Typically, light emitting diodes (LED) are used as a light source in a small size artificial growth system. This choice is justified due to their much easier control comparing to other types of light source. According to best-known practices, it was decided to use blue/red diodes which are one of the most important for the photosynthesis process in ratio 1/4 and total power 150 $Watt$ equipped with a relay module for controlling the LEDs. For this particular system, the period from germination to the end of vegetation for lettuce is approximately one month, the same period for tomatoes. The summary of the system design showed in Table 3.1. The hydroponic system design was developed based on the best world practices for providing the optimal conditions for a plant growth [Jones Jr, 2016, Sanyé-Mengual et al., 2015].

Monitoring system The system for monitoring of growth dynamics and system parameters has a 2D-plotter (*MakeBlock*) with the mobile carrier on which an RGB digital camera *Logitech c920* (1980x1080) is mounted. The assembled system allows to automatically take the sets of images of each plant in high resolution in a pre-

defined time period. Also, the placement of camera directly above the plane where the plants are located makes perspective distortions minimal and avoids effect of shadowing. The environmental conditions that will be used for further modelling were measured automatically using the electrical conductivity (EC), temperature, pH, relative humidity (RH), and feeding solution flow rate sensors. All these sensors and digital cameras before starting the experiment were tuned and calibrated. The LED system is constructed as a part of the whole experimental setup and it can be controlled as well [Shadrin et al., 2018].

Data Acquisition

Fig. 3-2 shows the system architecture and relations among the testbed subsystems. By using the hydroponic system and automatic data acquisition system, data which describe the plant growth dynamics (2D images) and system parameters were recorded. Next, this data were sent to a database, then a server process this data, calculates the projected leaves area, and predicts the projected leaves area growth based on the selected model. Red lines indicate the semi-automatic control effectuated between two blocks. For example LED-duty cycle was controlled semi-automatically. LED illumination had a duty cycle 18h/6h (day/night) at the beginning of the experiment and by the end of the experiment, duty cycle was slightly decreased to 16h/8h (day/night). For collecting and processing data automatically, a custom software for a desktop PC and smartphone was developed (see Fig. 3-2). The developed software is flexible: the user can easily integrate new sensors in the monitoring system. Also, the proposed system showed its robustness to the power interruption: the XY-plotter is automatically re-calibrated in case of power shut up. The software was released via a custom made scripts developed in Python and C programming languages. A set of continuous experiments on the different types of plants aimed at growth dynamics assessment and collecting the relevant dataset for more than a month was successfully performed. For these experiments it was developed software for hardware control, e.g. stepper drivers on the XY plotter and controllers for LED, and software for data receiving and processing. Both pieces of software must be synchronized properly. During the experiments,

2D images of plants were taken every 30 minutes within approximately one month for each type of plant cultivation. In parallel, remaining system parameters were collected, they were measured automatically, organized and stored into a database. A high-precision algorithm for calculating the projected leaves area of plants was developed and successfully implemented. The developed algorithm relies on a reference point - the red square object (see Fig. 3-3) with known area for performing the calibration, i.e. calculation of a specific pixel size: area/pixel. Then the algorithm performs the calculation of a number of green pixels which belong to the plant. A pixel is identified as green in its RGB value if it is in the certain bounds which are set up before the experiment starts. A Similar approach was discussed in [Easlon and Bloom, 2014]. A white background was used to reduce noise, e.g. green pixels not belonging to the plant. For monitoring of the experiment online it a custom web-interface was developed.

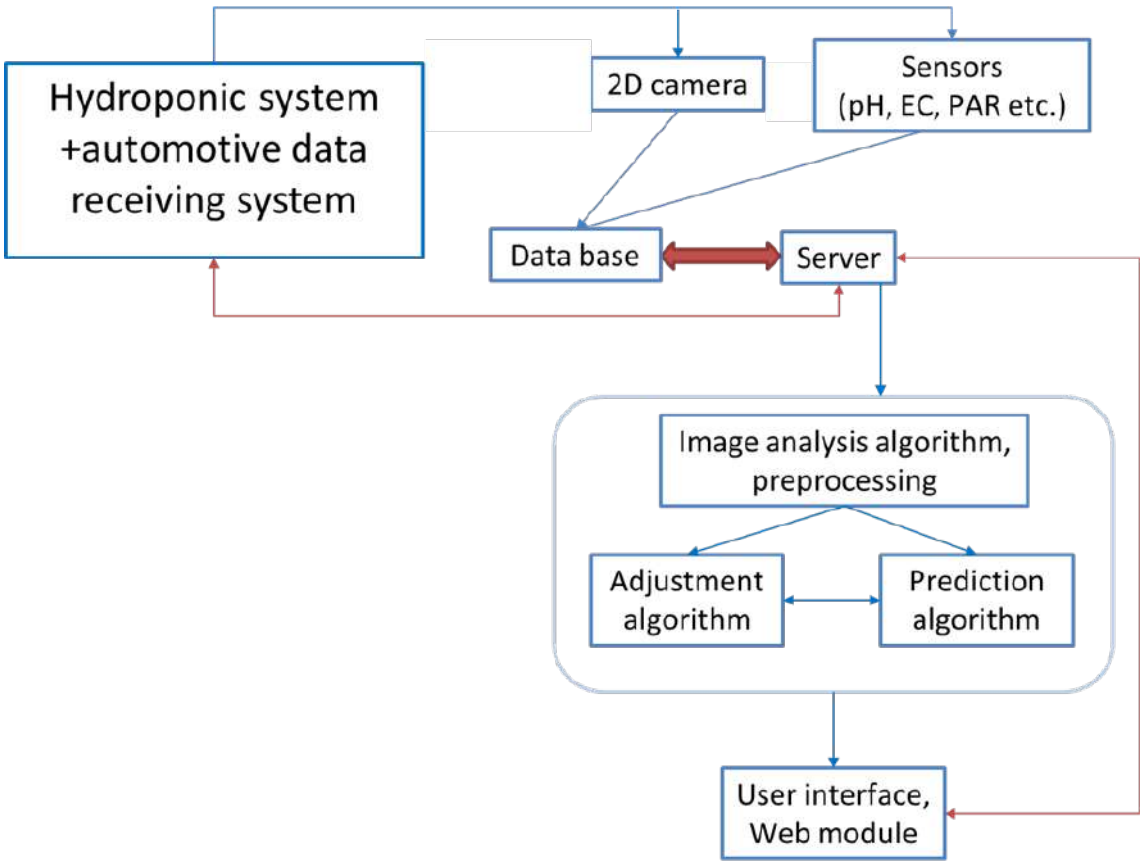


Figure 3-2: System architecture for 2D data acquisition and processing.

Totally, two small scale experiments were conducted for obtaining relevant data

about lettuce and tomato growth dynamics.

The first one was conducted within nearly one month during which four dwarf tomato plants 'MicroTina' [Scott et al., 2000] grew. It was decided not to use too many samples to avoid overlapping of plants and obtaining good quality images. During this experiment the data from the sensors and images were taken simultaneously every 30 minutes to be recorded into a database. The dataset of the time-sequenced top-down images of plant growth (3168 images) and growth conditions is available online https://github.com/DmitriiShadrin/TGD-Tomato-Growth-Dynamics.

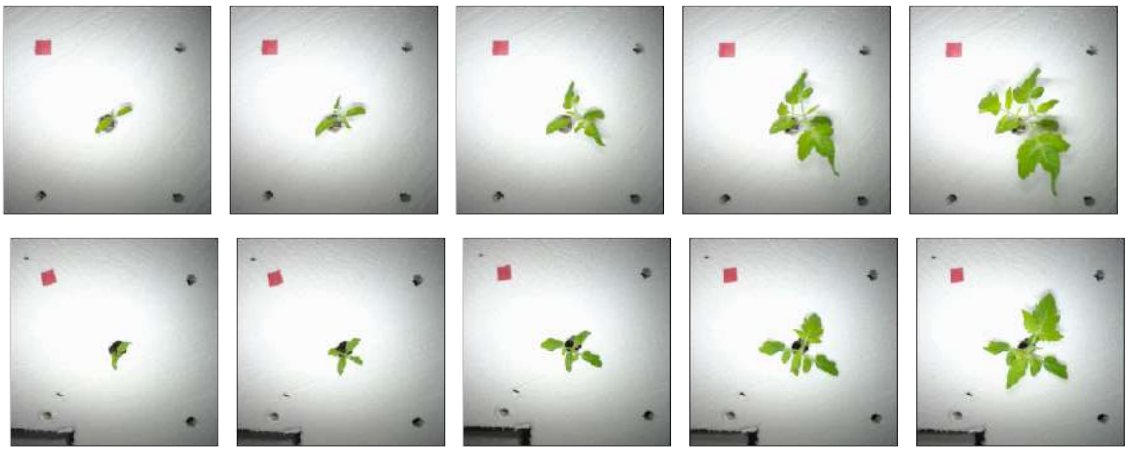


Figure 3-3: Examples of top-down tomato images received during the experiment.

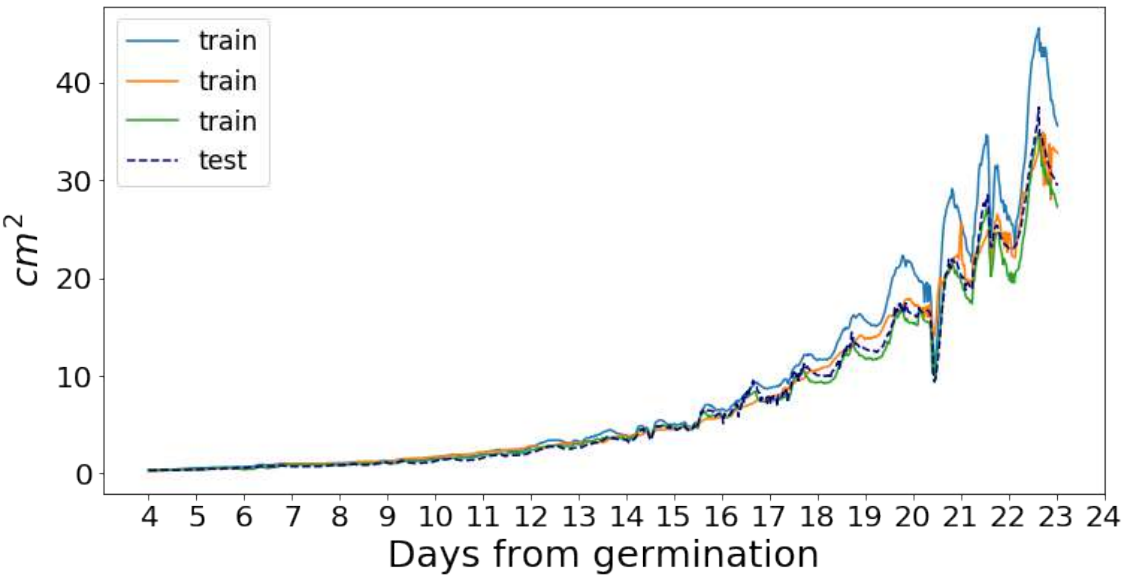


Figure 3-4: Projected leaves area calculations during the experiment.

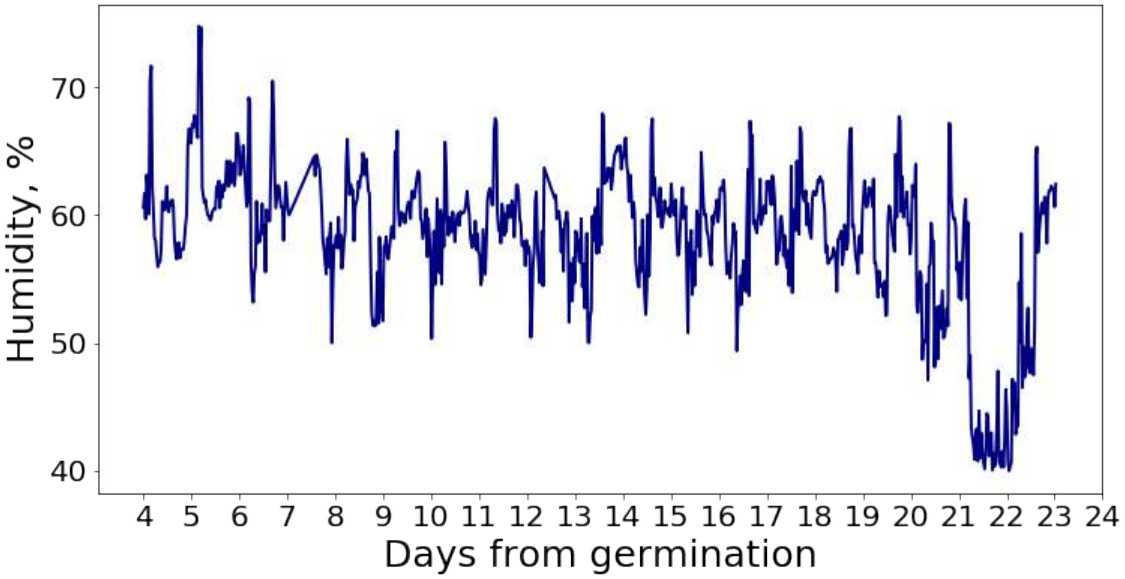


Figure 3-5: Measurements of humidity during the experiment.

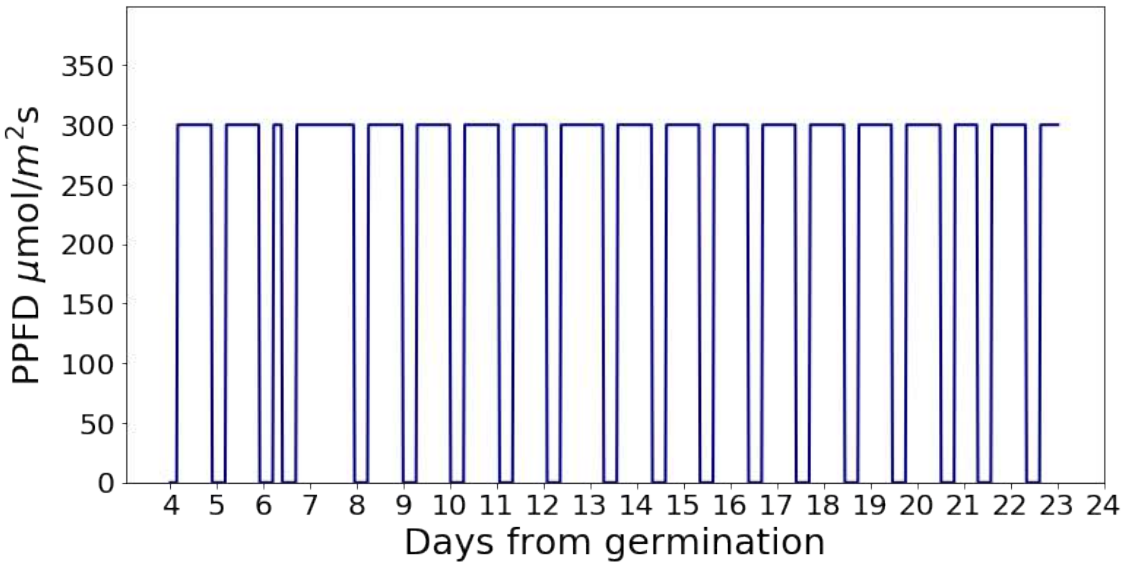


Figure 3-6: Illumination duty cycle (Photosynthetic Photon Flux Density).

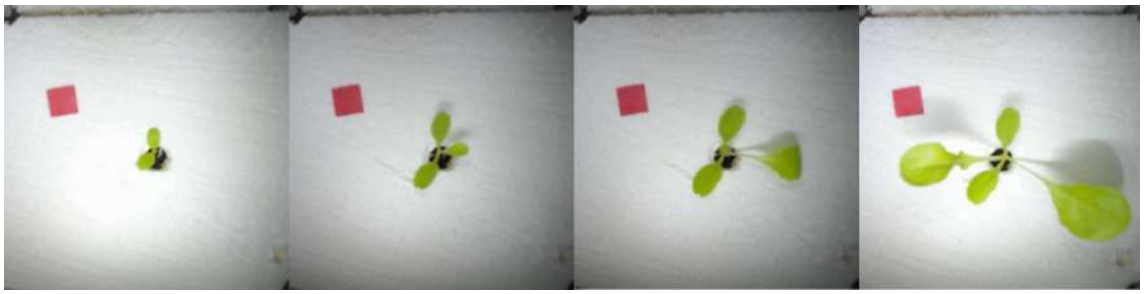
An example of raw top-down images for different tomatoes are shown in Fig. 3-3, where the time interval between two pictures from left to right is approximately three days. For each image the projected leaves area was calculated automatically using the procedure, discussed above. Totally, 1079 images were obtained for each tomato plant, and the same amount of data points from each sensor. For further growth dynamics assessment and system identification purposes, 792 first images of each plant and the same amount of the data points were used from each sensors. This was done because on the later stage of growth leaves started to overlap significantly and the projection of leaves area can be significantly different from the actual leaves area.

Fig. 3-4 presents the calculated projected leaves area for each one of the four plants. In the Fig. 3-4 it can be observed the diurnal fluctuations (oscillations) of projected leaves area. This happens because of the diurnal movements of the leaves. This effect was investigated in detail in [Kao and Forseth, 1992], where diurnal soybean leaves movement were investigated under different nitrogen and water availability. One of the outcomes of this study was that cosine of leaves surface (normal) to the horizontal surface varies dramatically during day and the amplitude of variation differs up to twice for different nutrition. Machine vision approach for detection of plant water stress is discussed in [Kacira et al., 2002], where as the feature of the stress detection was used the coefficient of relative variation of top-projected canopy area. In general, the diurnal regulation of plant growth one of the effects of which is leaves movements is described in [Nozue and Maloof, 2006]. The proposed in this section automatical measurements of projected leaves area allow assessing the stresses caused by nutrition, illumination and other environmental factors that influence growth dynamics by numerical analysis of the obtained fluctuations of the in projected leaves area.

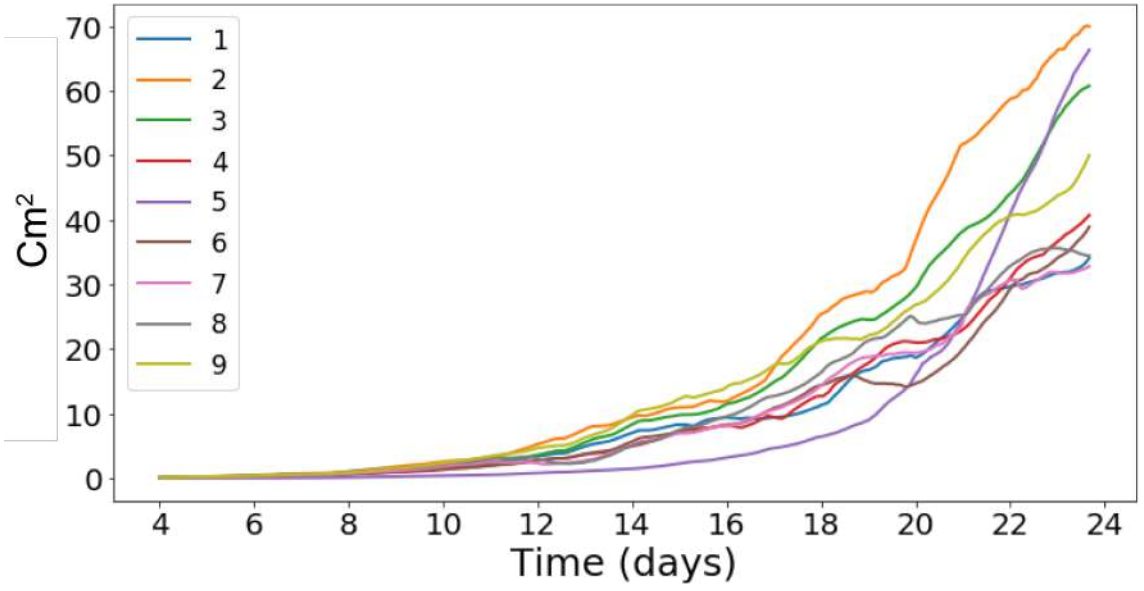
Two out of six recorded parameters are presented in Fig. 3-6 and Fig. 3-5 for showing the system performance during the experiment on tomato growth dynamics assessment. Figure 3-5 shows humidity during the experiment representing an example of the data obtained from the sensors. Light duty cycle was set most of the time during the experiment and was 16/8 hours reconstructing day/night cycle

except in some days. Fig. 3-6 shows the light duty cycle.

A similar experiment was conducted for lettuce growth dynamics assessment. Data from the sensors and digital camera were taken in the same time interval as for tomatoes in 30 minutes. The growth conditions were maintained at the optimal level through the experiment. Overall, a sequence of 7380 raw images was obtained (example in the Fig. 3-7a) along with the corresponding environmental conditions. Images were processed in-situ on the collection stage using the developed software. The projected leaves area of each plant is shown in Fig. 3-7b. Fluctuations as in the first experiment are not observed in this case as the calculated projected leaves area was smoothed. The experiment stopped when leaves started overlap, thus making impossible to estimate the actual leaves area from just 2D images properly.

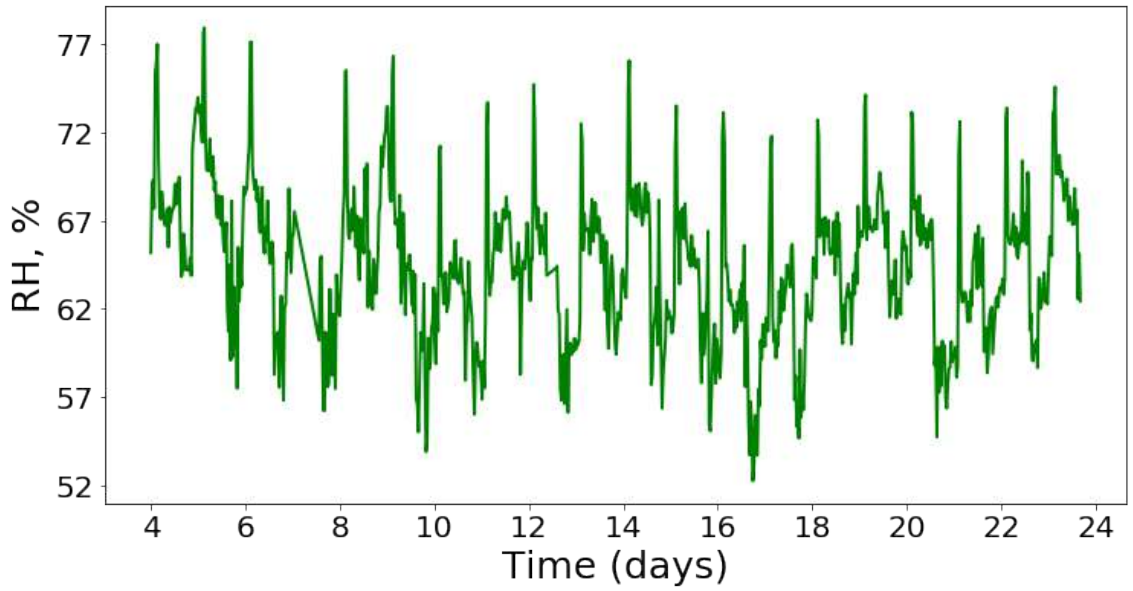


(a)

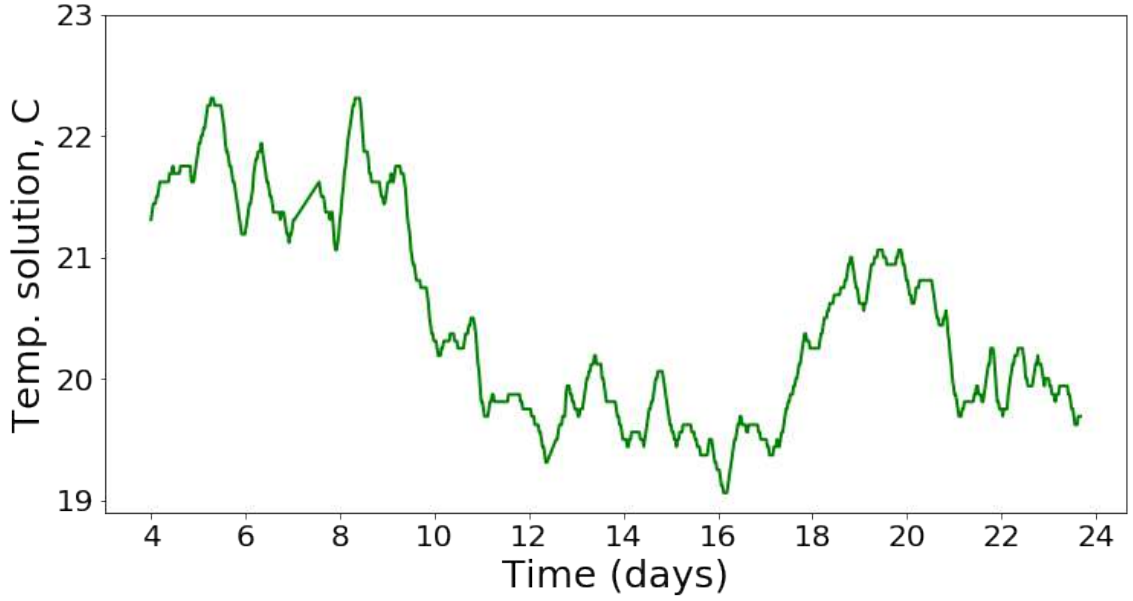


(b)

Figure 3-7: Examples of collected data on lettuce: (a) raw top-down images of lettuce, (b) calculated and pre-processed projected leaves area for 9 plants.



(a)



(b)

Figure 3-8: Examples of measurements from sensors, recorded during the experiment: (a) dynamics of relative humidity change during the experiment - maintained in the optimal range, (b) dynamics of feeding solution temperature change during the experiment - maintained in the optimal range.

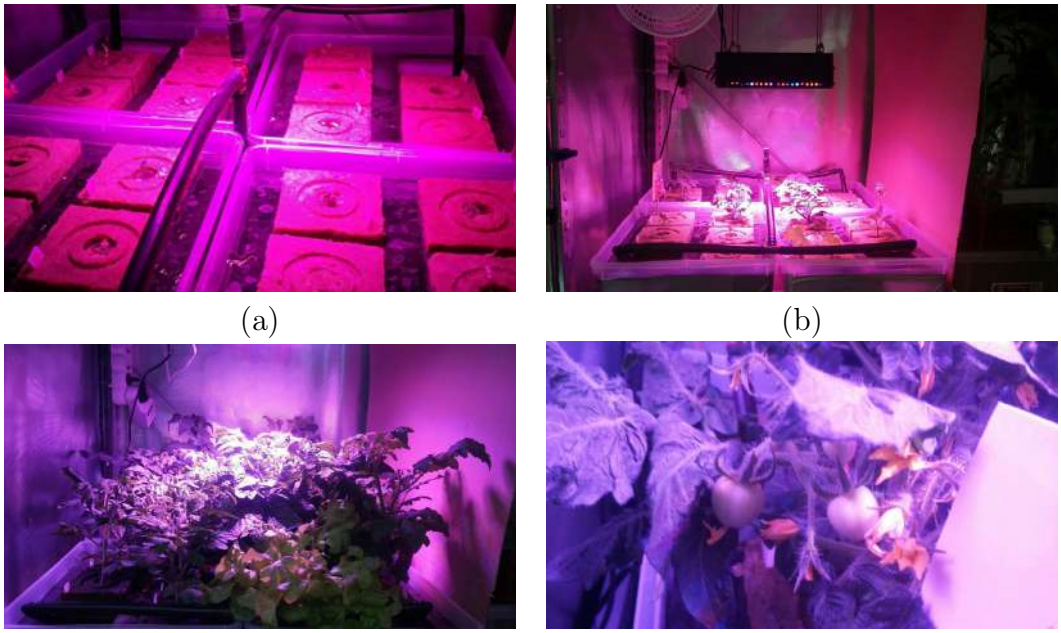
Figure 3-8a and Fig. 3-8b show examples of measured environmental parameters as well as the fact the growing system supports the stable plant growing. This point ensures that the obtained data can be used for model testing purposes. Also, it is beneficial that the amount and quality of data (images and data from sensors) are good enough for training machine learning algorithms.

3D data collection

For the purpose of biomass estimation and finding correlations between leaves area and plant biomass 3D data that describe different stages of plant growth was collected. It should be noticed that in the case of 3D experiment there is no projection of leaves area and the actual leaves area is measured A hydroponic system with a constant feeding layer was designed and assembled for control of ambient conditions. For data collection, a manual 3D scanner Artec Space Spyder was used. Below the features of the hydroponic system shown in Fig. 3-9 are summarized:

- System for growing 18 tomato plants,
- 180W multispectral LED light,
- 60 liter tank,
- 0.65 liter rock wool blocks as a substrate,
- 8 W pump (100 liter/h),
- 1.5 cm of feeding solution layer

Apart from the benefits of usage of the rock wool blocks as a substrate for plants cultivation described above it also gives an opportunity to inspect and perform 3D scanning of each plant in the experiment without interruption of the system operation and without damaging the plants. Fig. 3-9 shows the performance of the system and acceleration of the physiological processes in the plants in a hydroponic system of this type. In this experiment it took around one month from germination to the first flowering.



(a) (b)

(c) (d)

Figure 3-9: Hydroponic system where tomato plants were growing during 1 month and 1 week: (a) germination stage, (b) vegetation stage, and (c-d) flowering.

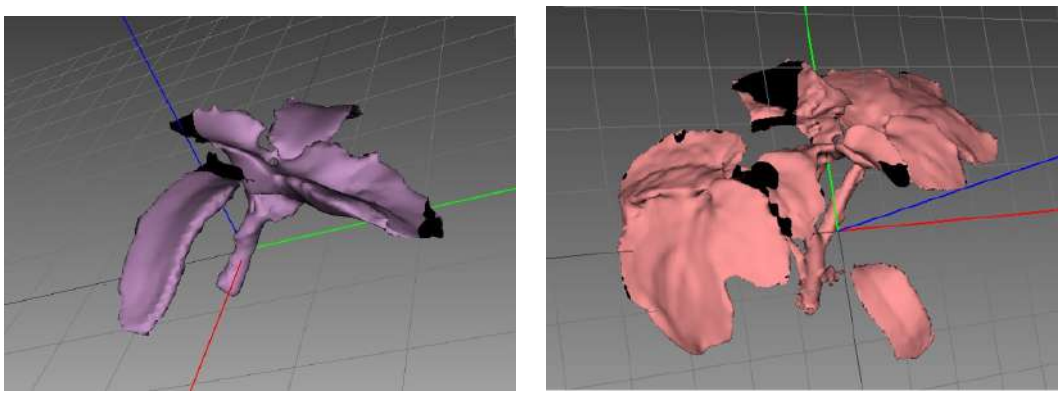


Figure 3-10: 3D images of the tomato plant in the beginning of vegetation lifetime.

For initial data acquisition, 18 tomato plants were used. These experimental samples were composed of two dwarf tomato sorts: Bonsai Micro (9 plants) - the same sort as that was used in 2D experiment and Bonsai (9 plants) were germinated in optimal conditions and then transplanted into the hydroponic system. The system conditions were monitored manually for the sake of controlling the allowable rates of feeding solution parameters (pH, temperature, humidity, electric conductivity). All of these parameters could be corrected if necessary. 3D images collection of plants was organized in the following way: first the plant was taken out from the hydroponic system, then it was put on to a rotating table, and scanned using the 3D scanner under green spectrum illumination conditions (because plants reflect this spectrum making the final 3D image more accurate). After receiving the 3D images, their preprocessing, and smoothing, the main parameters including the actual leaf areas and their biomass were calculated. Smoothing and preprocessing procedures were performed by commercial *Artec Studio* software which supports the scanner operation. The preprocessing helps to reduce the noise and to remove unnecessary parts of scanned image where the parameters were indicated by program recommendations with manual tuning. One of the main functions in the Artec studio and used in this work is "Fusion". This function enables the creation of a polygonal 3D model based on received clouds of points. Fig. 3-10 shows an example of a 3D image of the tomato in the beginning of vegetation lifetime. In total, 80 3D cloud of points of dwarf tomato plants were received and processed for the period from germination to the beginning of flowering. These 3D images represent the dynamics of plant growth and can be used for biomass (volume) assessment, based on leaves area (or projected leaves area) calculations. Also, 55 3D clouds out of 80 will be used for biomass (volume) dynamics assessment.

3.2 Kalman filter for simple models

Development of a state-space model of plant growth

In this section it will be presented the Kalman filtering approach for growth dynamics prediction based on the selected growth model. One of the models that is widely

used for estimating the crop growth dynamics and for estimating the population growth in biological systems is the Verhulst model Eq. (3.1) [Frighetto et al., 2019, Kalmykov and Kalmykov, 2015]. It is also known as logistic S-curve:

$$\frac{dS}{dt} = \mu S(1 - \frac{S}{S_{max}}), \tag{3.1}$$

where μ is growth rate ($\frac{1}{time\ step}$), S and S_{max} - current and maximum leaves area (projected) respectively in cm^2. Integration of (3.1) gives the following Eq. (3.2):

$$S(t) = S_{max}\frac{S_0 e^{\mu t}}{S_{max} + S_0(e^{\mu t} - 1)}. \tag{3.2}$$

The Verhulst model is widely applied for assessment of the dynamics of life systems. For example, Verhulst model was applied for spatio-temporal population control to management of aquatic plants [Frighetto et al., 2019] [Costa et al., 2003]. It should be noticed that the Verhulst model was used for assessment of the dynamics of the projected leaves area, not the leaves area themselves. In the experiments, it was obtained 2D top-down images of plants. Using these images, the projection of leaves area was calculated. Growing plants broadwise can have an effect on the values of the calculated projected leaves area. The projection of leaves area has limitations for the investigated plants because these plants are not able to grow infinitely broadwise. This effect was also observed experimentally. For the development of models data from the initial stages of growth will be used and is assumed that the plant has some maximum projected leaves area. Also, it will be shown that the obtained experimental data have a good fit to the Verhulst model.

The main aim of the following modelling procedure is to estimate the main growth parameters μ and S_{max} based on the current measurements of the projected leaves area. In general case parameters μ and S_{max} are time-dependent. hanging of these parameters are guided by the environmental conditions. In the following modeling it will be estimated the evolution of these parameters in time and compared the results with true (measured) values of them. Rearranging Eq. (3.2) leads to Eq.

(3.3):

$$S(t) = \frac{1}{(\frac{1}{S_0} - \frac{1}{S_{max}})(e^{-\mu t} - 1) + \frac{1}{S_0}}. \tag{3.3}$$

It should be noted that based on Eq. (3.3), $S(t)$ tends to S_{max} as t growth. On the one hand, if t is small then expression $(e^{-\mu t} - 1)$ in Eq. (3.3) is close to zero. It means that the model of growth is less sensitive to S_{max} and the estimation of S_{max} can be done for larger values t. On the other hand, for large values of t, $e^{-\mu t}$ (containing unknown parameter μ) tends to be zero and the contribution of this expression to the total value of $S(t)$ is low. It means that the measurements of the projected leaves area $S(t)$ will have more impact on the estimation of μ compared to S_{max} on the initial stage of growth.

Equation (3.2) that reflects the growth dynamics trends of plants was taken as a basis for the model creation at state-space. The state vector X_k at time k was defined as follows in Eq. (3.4):

$$X_k = \begin{bmatrix} \mu_k \\ S_{max,k} \end{bmatrix}, \tag{3.4}$$

where k is the $k - th$ state of the system. The future state of the system X_{k+1} is presented as a function of the current state in Eq. (3.5):

$$X_{k+1} = F_{k+1,k} X_k + W_{k+1}, \tag{3.5}$$

where $F_{k+1,k}$ is the transition matrix relating the state vectors X_k and X_{k+1} is given by Eq. (3.6):

$$F_{k+1,k} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{3.6}$$

The state noise is defined by Eq. (3.7) to take into account the unpredictable

variations of plant growth:

$$W_k = \begin{bmatrix} w_k^{\mu} \\ w_k^{S_{max}} \end{bmatrix}, \tag{3.7}$$

where w_k^{μ} and $w_k^{S_{max}}$ are the noises of growth rate and maximum projected leaves area, respectively, with zero mathematical expectation - $E[W_k] = 0$. Covariance matrix of the state noise is defined in Eq. (3.8):

$$E[W_k, W_k^T] = Q = \begin{bmatrix} \delta_{\mu}^2 & 0 \\ 0 & \delta_{S_{max}}^2 \end{bmatrix}, \tag{3.8}$$

where δ_{μ}^2 is the variance of parameter μ and $\delta_{S_{max}}^2$ is the variance of the parameter S_{max}. If μ and S_{max} are assumed to be constant, then $Q = \mathbf{0}$.

To estimate the state vector $\hat{X}_{k+1}$, the measurements of projected leaves area of plants were used. The measurement equation is defined in the following way, presented in Eq. (3.9):

$$\begin{aligned} z_k &= S_k + \nu_k = h(X_k) + \nu_k, \\ h(X_k) &= \frac{1}{(\frac{1}{S_0} - \frac{1}{S_{max}})(e^{-\mu k} - 1) + \frac{1}{S_0}}, \end{aligned} \tag{3.9}$$

where z_k is the measurement of projected leaves area, $h(X_k, k)$ is the nonlinear measurement function of the state vector X_k and ν_k is the measurement noise with variance $\delta_{\nu,k}^2$ and zero mathematical expectation. During the modelling procedure it can be assumed that the variance of the measurement noise depends on $S(t)$ and equals for example $\alpha S(t)$ because with larger projected leaves area, a greater measurement error occurs.

The recurrent algorithm of Kalman filter consists of two repeating procedures, namely extrapolation and filtration:

Extrapolation is performed to estimate the future state vector X_k using Eq.

(3.10)

$$\hat{X}_{k,k-1} = F_{k,k-1}\hat{X}_{k-1,k-1}, \tag{3.10}$$

where $\hat{X}_{k,k-1}$ is the the extrapolated estimate of the state vector X_k. The first subscript k denotes the time at which the extrapolation is made, while the second subscript $k-1$ represents the number of observations $z_1, z_2, ..., z_{k-1}$ used to obtain the extrapolated estimate $\hat{X}_{k,k-1}$.

The prediction error covariance matrix $P_{k,k-1}$ is given by Eq. (3.11):

$$P_{k,k-1} = F_{k,k-1}P_{k-1,k-1}F_{k,k-1}^T + Q, \tag{3.11}$$

Filtration equation incorporates a new observation for obtaining an improved estimate and is given by Eq. (3.12):

$$\hat{X}_{k,k} = \hat{X}_{k,k-1} + K_k(z_k - h(\hat{X}_{k,k-1}, k)), \tag{3.12}$$

where $\hat{X}_{k,k}$ is the filtered estimate of state vector X_k at time k using k available measurements.

The filter gain K_k and filtration error covariance matrix $P_{k,k}$ are calculated according to Eq. (3.13):

$$\begin{aligned} K_k &= P_{k,k-1}\tilde{H}_k^T(\tilde{H}_k P_{k,k-1}\tilde{H}_k^T + \delta_{\nu,k}^2)^{-1}, \\ P_{k,k} &= (I - K_k\tilde{H}_k)P_{k,k-1}, \end{aligned} \tag{3.13}$$

where $\tilde{H}_k$ is the derivative $h(X_k)$ as the dependence between the projected leaves area measurements z_k and the state vector X_k is nonlinear. $\tilde{H}_k$ is computed from calculating the partial derivatives of $h(X_k)$ with respect to the state vector X_k in

the point $\hat{X}_{k,k-1}$, where $\hat{X}_{k,k-1} = \begin{bmatrix} \hat{\mu}_{k,k-1} \\ \hat{S}_{max,k,k-1} \end{bmatrix}$:

$$\hat{H}_k = \frac{dh(X,k)}{dX^T}|_{\hat{X}_{k,k-1}} = \\ = \begin{bmatrix} k(\frac{1}{S_0} - \frac{1}{\hat{S}_{max,k,k-1}})e^{-\hat{\mu}_{k,k-1}k}\hat{S}^2_{k,k-1} \\ -\frac{(e^{-\hat{\mu}_{k,k-1}k}-1)\hat{S}^2_{k,k-1}}{\hat{S}^2_{max,k,k-1}} \end{bmatrix}^T. \tag{3.14}$$

In Eq. (3.14) as the extrapolated estimation of the projected leaves area should be used Eq. (3.15):

$$\hat{S}_{k,k-1} = \frac{1}{(\frac{1}{S_0} - \frac{1}{\hat{S}_{max,k,k-1}})(e^{-\hat{\mu}_{k,k-1}k} - 1) + \frac{1}{S_0}}. \tag{3.15}$$

It can be noticed from Eq. (3.14) that the expression $ke^{-\hat{\mu}_{k,k-1}k}$ at the numerator of the first element tends to zero with the increase of k since the exponential function decreases faster compared to the linear increase of k. This means that the first element of $\hat{H}_k$ defining the weight of the parameter μ in the projected leaves area measurements can be estimated more accurately for lower values of k. The opposite situation for the second element in Eq. (3.14) was detected where the numerator increases with the increase of k. Thus, more precise estimation of S_{max} can be achieved using larger values of k.

Results of modeling

Assessment of a plant growth dynamics based on simulated data. The main aim of the performed modelling using the simulated data is to show the robustness of the method and present the boundaries of the method application. The example of simulated leaves area growth (projected) based on Eq. (3.2) and measurements is shown in Fig. 3-13. For simulations it was assumed that the variance of the measurement error is $\delta_{\nu,k} = 0.01 * S_k$. The growth rate was assumed to be $\mu = 0.04$ and maximum projected leaves area $S_{max} = 1000\, cm^2$ for performing the

simulation. First, it was evaluated the method assuming the zero state variances: $\delta_{\mu}^2 = 0$ and $\delta_{S_{max}}^2 = 0$. The results of the reconstructed μ and S_{max} (for $Q = 0$) are shown in Fig. 3-11 and Fig. 3-12. From these figures, it can be noticed that the proposed method is characterised approximately by the same convergence rate as for the non-linear least squares for both the maximum projected leaves area and the growth rate. However, Kalman filtering is much faster as it is an iterative method. It takes 0.012 s for the simulation by using Kalman filtering with the parameters described earlier. For non-linear least squares execution time is 29.827 s.

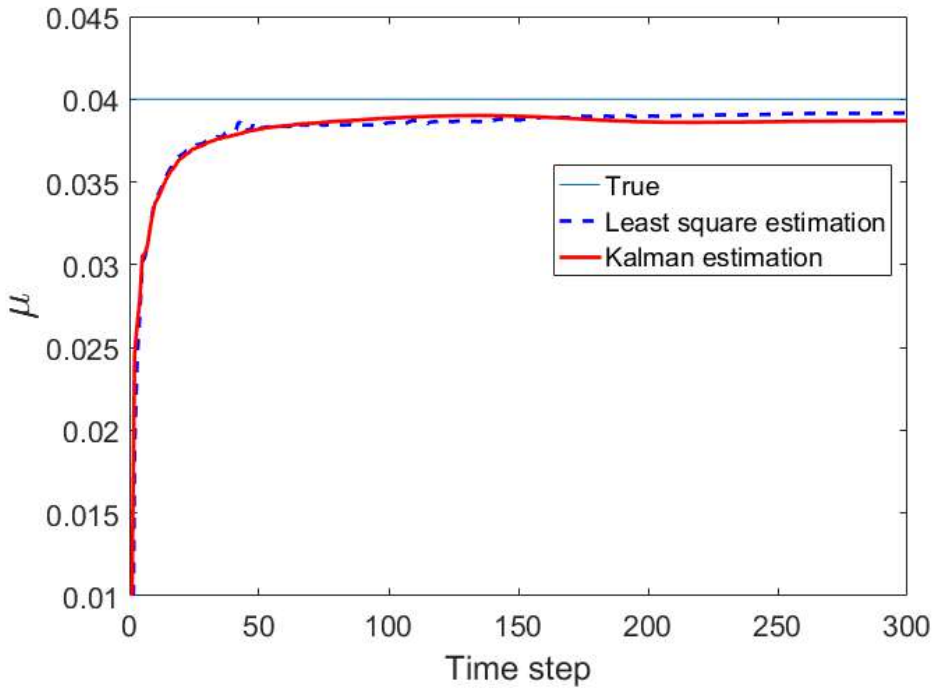


Figure 3-11: True, estimated by Kalman filter and by the non-linear least square dynamics of growth rate changing in time. Modeling was made with the assumption that $Q = 0$.

Next, this method was evaluated assuming the following state variances: $\delta_{\mu}^2 = 0.01$ and $\delta_{S_{max}}^2 = 10$. Fig. 3-13 presents the estimation of the projected leaves area based on the filtered estimations of μ and S_{max}. The results of the reconstructed μ and S_{max} are shown in Fig. 3-14 and Fig. 3-15. As it was discussed previously in this section, μ is better reconstructed on the initial stage of growth (first half of simulated data) while S_{max} is better estimated on a later stage of growth (see Fig. 3-14 and Fig. 3-15).

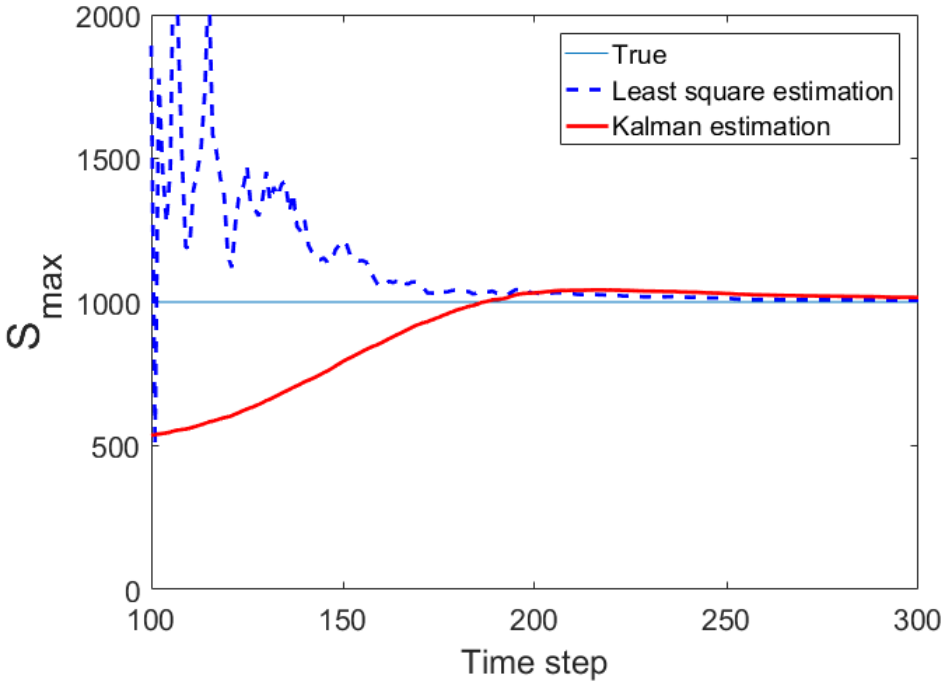


Figure 3-12: True, estimated by Kalman filter and by the non-linear least square dynamics of maximum projected leaves area changing in time. Modeling was made with the assumption that $Q = 0$.

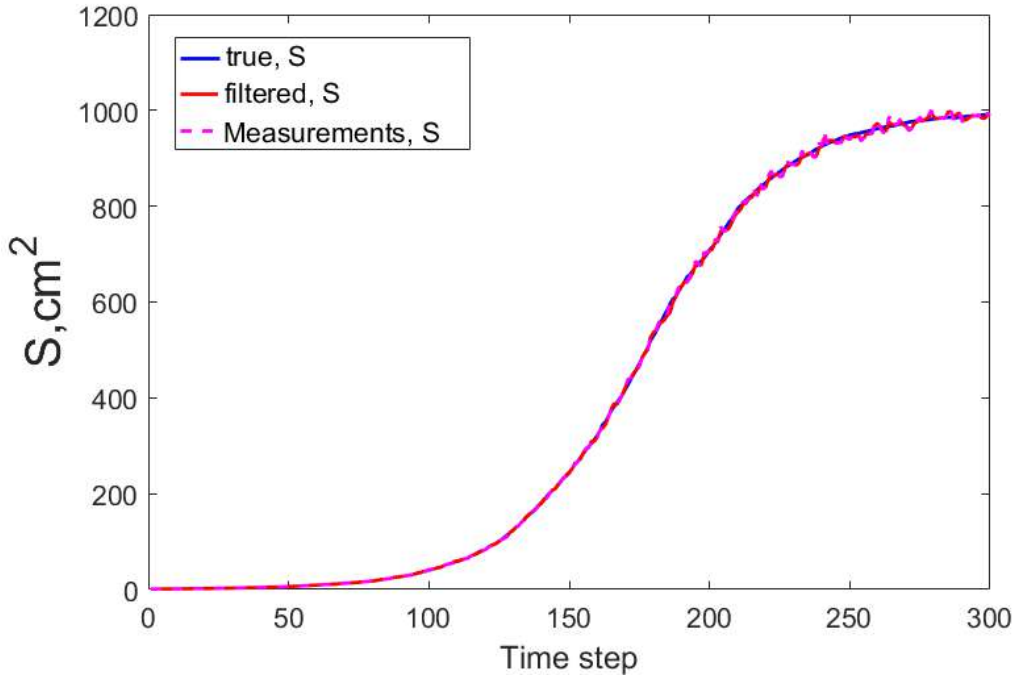


Figure 3-13: Projected leaves area: true, measurements and filtration. Modelling was made with the assumption that $Q \neq 0$.

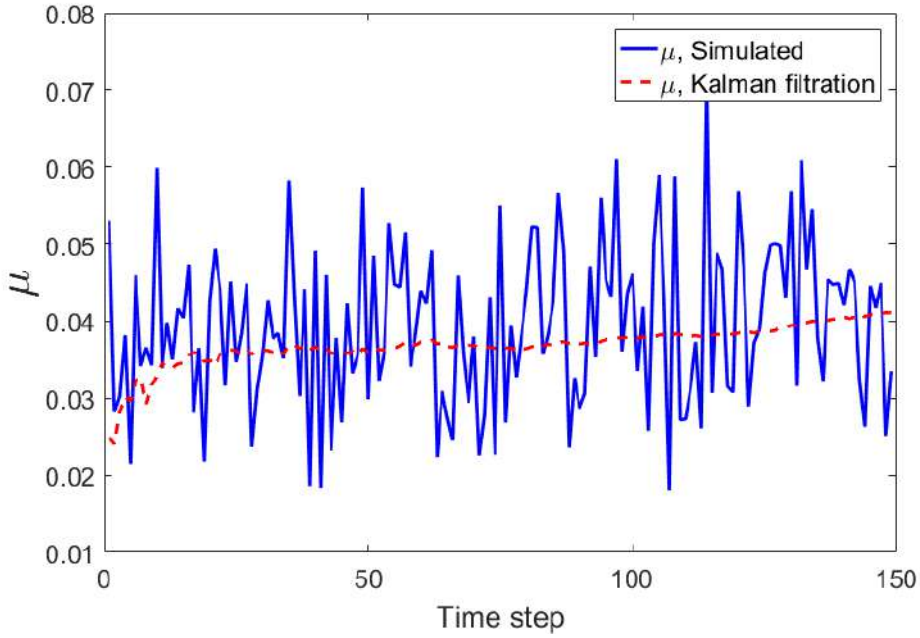


Figure 3-14: Simulated and estimated by Kalman filter dynamics of growth rate. Modeling was made with the assumption that $Q \neq 0$.

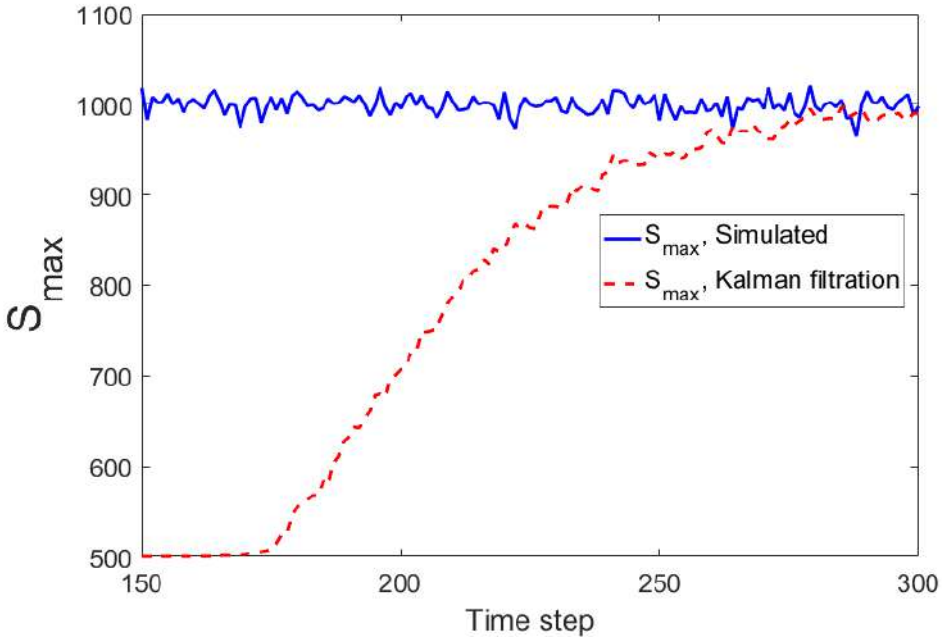


Figure 3-15: Simulated and estimated by Kalman filter dynamics of maximum projected leaves area. Modeling is made with the assumption that $Q \neq 0$.

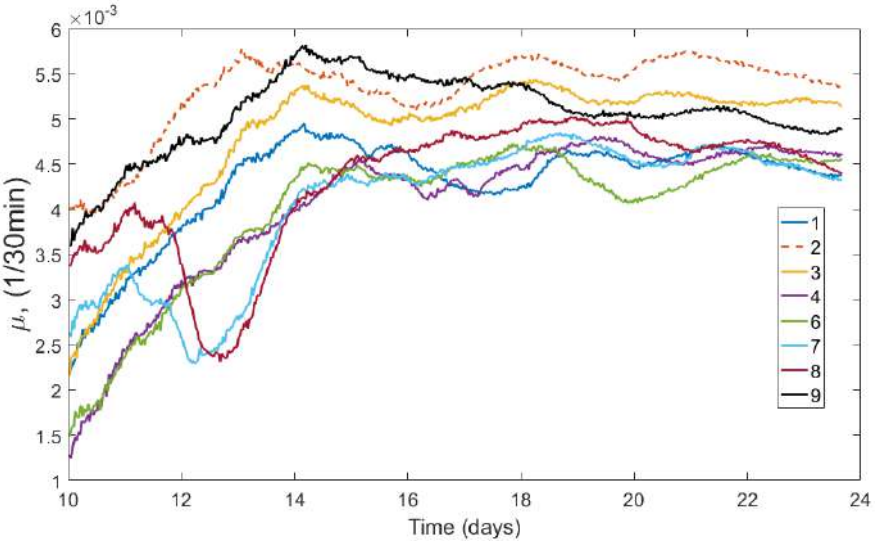


Figure 3-16: Plants growth rate dynamics estimation from the experimental data is shown for each plant (number 1-9, except for the 5-th). The 2-nd plant showed the maximum growth rate which matches the experimental data.

Assessment of plants growth dynamics based on collected data. For evaluating the proposed method on the experimental data the collected dataset on lettuce growth described in Section 3.1 was used. From the Fig. 3-7b it can be noticed that the obtained data represent the initial stage of plant growth. It means that based on the findings discussed previously it is feasible to estimate the growth rate for this period. The results of plant growth rate estimation (all plants except for the 5-th as it was an outlier as its germination period was twice longer) are presented in Fig. 3-16. According to Fig. 3-7b the visible growth of almost all plants started approximately from the 10-th day; the same is observed in Fig. 3-16. Also, as can be seen from Fig. 3-16 the 2-nd, the 3-rd, and the 9-th plants (decreasing order) have the maximum growth rate, which correctly reflect the experimental growth dynamics that is shown in Fig. 3-7b.

Conclusions

In this section, the Extended Kalman filter approach was adapted and implemented for the evaluation of the plants' growth dynamics. Validation of the proposed method was performed on the simulated and obtained on the custom made experimental setup. The results of the methods' validation showed its high accuracy and

potential for the application in precision agriculture and solving the optimization tasks. In particular, the benefit of the proposed method is the high computational efficiency which allows its usage on embedded devices. Also, this method is model-based which means that it is possible to include the additional parameters into the model and evaluate the effect of them on the growth dynamics.

3.3 Instance segmentation for high throughput plant phenotyping systems

Dataset annotation

Instance segmentation algorithms allows to receive more detailed information that describes plant growth dynamics. In particular, using the series of images of plants and performing instance segmentation of leaves for each image it is possible to assess projected area of each leaf, thus to reconstruct and model each leaf growth dynamics. For demonstration of this approach the same dataset (as for Kalman filtering) of top-down 2D images of lettuce growth was used (see description in the Section 3.1). The part of the raw dataset was labelled. This labelled dataset is suitable for testing semantic and instance segmentation algorithms based on FCNNs for phenotype and also gives possibilities to test other types of computer vision algorithms. This annotated dataset is publicly available: `https://github.com/DmitriiShadrin/PlantGrowthDynamics`. The dataset includes 4815 raw lettuce images for the period of 11 days growth after germination and 75 manually annotated image data. All these data have a time reference that gives the possibility to estimate plant growth dynamics. For these images, leaf masks and leaf bounding boxes were extracted manually by using online labelling tool *LabelMe* [Russell et al., 2008]. Figure 3-17 shows examples of images from the dataset with the corresponding leaf masks (bounding boxes were included in the dataset, but are not represented in Fig. 3-17 as the image will be confusing). In total, 356 leaf masks and bounding boxes were obtained. The dataset contains both: relatively simple annotated images - 62 with three instances, 47 with four instances, and complicated - eight images with

rich structure and 10 instances. By now 75 images were annotated as it was enough for training the instance segmentation model, but the amount of annotated images will be increased and added to the current publicly available dataset. A similar benchmark dataset was collected and annotated by [Scharr et al., 2016], and is very popular for conducting challenges, where competitors try to achieve the highest IoU (Intersection over Union) of different CV algorithms (in majority FCNNs). The annotated dataset on lettuce proposed in this section, compared to existing has one big advantage: the sequences of images with time reference ones with allow dynamics modeling after performing segmentation tasks.

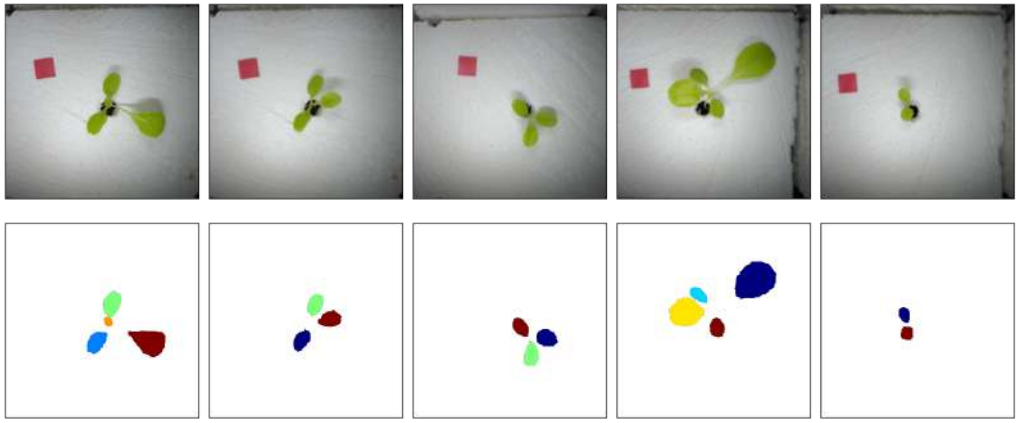


Figure 3-17: Examples of lettuce images at different growth stages with corresponding leaf masks; the pictures are taken from *manually annotated* data set.

Image processing

The estimation of individual leaf growth dynamics requires separation between leaf instances on the image. In order to solve this task it was used the Deep Coloring method [Kulikov et al., 2018]. Deep Coloring reduces instance segmentation to the task of pixel classification (coloring). The latter task can be accomplished using almost any of the recently developed deep convolutional architectures for semantic segmentation. In this work, U-net as semantic segmentation backbone was used [Ronneberger et al., 2015]. Simply speaking, this method enforces all pixels of the same object to take the same color, while also enforcing pixels belonging to different but adjacent object instances to take different colors. The example of output of this

method (performance on the test images) is depicted in Fig. 3-18 (top). A simple component analysis allows to extract individual leaves on the image.

To train an instance segmentation network, the annotated dataset described above was used. The training set was split into two parts: training and test set, 65 and ten images respectively. The training set was augmented with random crops, rotations, flips and scaling. Other training parameters were taken from [Kulikov et al., 2018]. The instance segmentation accuracy, achieved on the test set was 0.74 symmetric best dice coefficient (SBD) [Scharr et al., 2016]. This score is slightly worst than the score of this algotirhm on CVPPP A1 dataset, where the instance segmentation method achieves 0.80 SBD [Kulikov et al., 2018].

The instance segmentation algorithm produces labels for instances independently for each image and they may differ between sequential frames. In order to estimate the leaf growth dynamics, a post-processing step had been implemented. The post-processing includes tracking each label and making sure that each instance has the same index for the whole sequence. For each sequential pair of labeled images the linear assessment problem based on the inverted pairwise intersection over union between instances was solved [Munkres, 1957]. Linear assessment provides us correspondences between labels, the labels on the second frame are modified to match the labels from the first frame. To make this procedure more stable, over-segmented images were removed from each sequence.

Growth dynamics of projected leaves area was reconstructed according to the already used model (see Eq. (3.2)). S_{max} and S_0 are constant and set to $100cm^2$ and $0.5cm^2$ respectively, μ is the estimated parameter [van Eeuwijk et al., 2018].

Overall, it was reconstructed the growth dynamics for each leaf of plants in the dataset, described in Section 3.1 (lettuce) for the period of 11 days after germination. For each image, the instance segmentation network was applied along with detailed masks of individual leaves as shown in Fig. 3-18 (top). Each leaf instance was tracked thought all time sequence providing information about its size in pixels. In order to convert the size to real-world units, a calibration objects (red square of size $1 \times 1cm$) was used.

The observed growth dynamics is presented in Fig 3-18 (bottom), where the

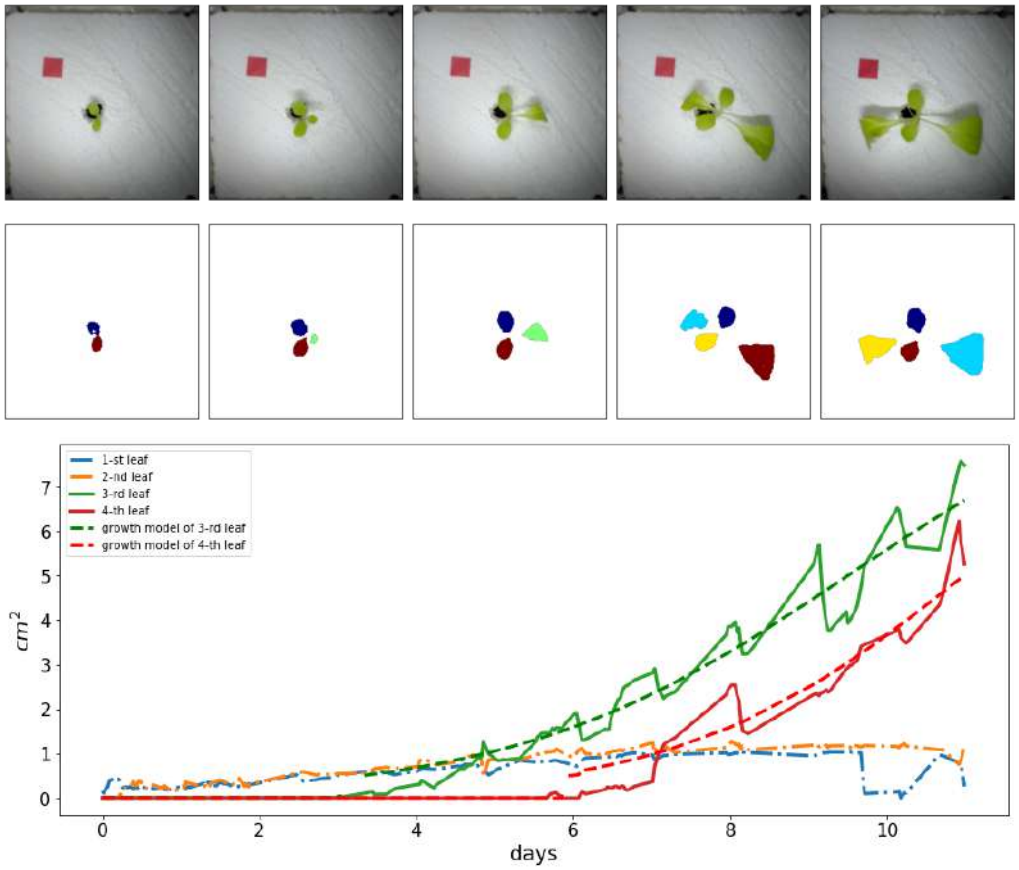


Figure 3-18: Results of dynamics reconstruction. Dotted lines depicted the fitted growth model for third and fourth leaves, based on the predicted segmentation masks that represent projected leaves area. Pictures above represents raw lettuce images with segmented leaf instances masks by *instance recognition*; the images approximately correspond to the graph time frame.

Plant sample	Growth rate of 3-rd leaf, 1/day	Growth rate of 4-th leaf, 1/day
1	0.43	0.48
2	0.47	0.52
3	0.47	0.55
4	0.44	0.59
5	0.47	-
6	0.40	0.48
7	0.40	0.48
8	-	-
9	0.43	0.67
mean ± std	0.438 ± 0.027	0.538 ± 0.066

Table 3.2: Growth rate estimation.

exponential growth of third and fourth leaves can be observed. The First and second leaves that appeared in the beginning grew up to $1cm^2$ and then their size remained stable for all investigated plants. This is happened due to physiological reasons. For almost all plants similar and feasible values of growth rate μ were received. Assessed growth rates are presented in Table 3.2. It was not possible to calculate such dynamics for the eighth lettuce sample due to its side location relative to the camera. For the fifth lettuce sample, the fourth leaf not appear till the 11-th day. It is important to notice that results of modeling (assessment of leaves growth rate) are in correspondence with the results obtained by Kalman filtering (see Section 3.2).

Conclusions

A novel high-throughput method for analysis and prediction of plant growth dynamics by a combination of modern computer vision and modelling techniques was developed and presented. This methodology was tested on the obtained datasets. The results of the tests show the possibility to make detailed reconstruction of dynamics of plant growth. In particular, the advantage of the proposed methodology that it is possible not only to perform high-throughput plant phenotyping, which is commonly used for quantitatively non-invasive monitoring of plant organs, but also for the assessment of the dynamics of plants organs growth. Such CV systems are very important for plant studies. One of the limitations of this method is that

method is based on the computationally complex FCNN. The other limitation of the methodology is that for a more complex background much larger annotated dataset than described above, is needed (starting from hundreds of images with thousands of instances) for achieving reasonable accuracy (IoU or SDB). However, this problem can be overcome by using the pre-trained FCNNs. Such FCNNs can be fine-tuned, which allows to reduce the amount of trained data. This approach can provide a background for development of systems for automatic optimization of plant growth in artificial conditions based real-time detailed plant growth dynamics. In addition, an annotated dataset that describes 11 days of growth under controlled conditions was obtained.

3.4 Dynamic mode decomposition for complex models

Development of the model

For calculation experiments, the modified DMD algorithm was used [Schmid, 2010]. This algorithm was originally designed to solve the problem of extracting features of dynamic systems (specifically of flow fields) based only on its snapshots. Series of images of plants, or features from these images and growth conditions parameters can be represented like snapshots and the DMD algorithm can be applied for them. Suppose that we have a series of snapshots $x_i = x(t_i) \in \mathbb{R}^N$, $i = 1, \ldots, M$, over time periods $t_1, t_2, \ldots, t_M$, where $t_{i+1} = t_i + \Delta t$. Also, suppose that there is a linear operator A that approximates the evolution of the system: $x_{i+1} \approx Ax_i$, $i = 1, \ldots, M-1$. If we denote $Y = \begin{bmatrix} x_2 & \ldots & x_M \end{bmatrix}$ and $X = \begin{bmatrix} x_1 & \ldots & x_{M-1} \end{bmatrix}$, then the previous result can be written as $Y \approx AX$.

The goal of the DMD algorithm is to find the eigen decomposition of such operator A, which in turn can be rewritten as $A = \underset{\tilde{A} \in \mathbb{R}^{N \times N}}{\operatorname{argmin}} \|Y - \tilde{A}X\|_F^2$. Later, a more generalised definition of DMD was given that allows to use it with data collected at irregular timings: $A = YX^+$, where $+$ is the Moore-Penrose pseudoinverse [Tu et al., 2013].

The actual version of the algorithm incorporates an additional control which is discussed in [Proctor et al., 2016]. The original matrix X is replaced with $\tilde{X} =$

$\begin{bmatrix} X \\ U \end{bmatrix} \in \mathbb{R}^{(N+K)\times M}$, where $U \in \mathbb{R}^{K\times M}$ is a matrix of K control variables measured at N timesteps. The matrix A now has the size $N \times (N + K)$ and the problem transforms to the search of $A = \underset{\tilde{A}\in\mathbb{R}^{N\times(N+K)}}{\operatorname{argmin}} \|Y - \tilde{A}\tilde{X}\|_F^2 = Y\tilde{X}^+$.

Due to linearity of DMD, the non-linear dependence of the data must be created because processes in plants are non-linear. The Fishman and Genard approach was used as a basis to create features, that could describe non-linear dynamics of plant growth [Fishman and Génard, 1998]. In the original paper, a fruit, namely a peach, was investigated. However, this approach can be extended to processes, that describe the entire dynamics of plant growth. First, it was considered the equation for modelling water balance in the plant. The rate of change of amount of water w in the plant is the algebraic difference of water inflow from substrate U_{in} and the transpiration from surface of the plant T_p [Fishman and Génard, 1998]:

$$\frac{dw}{dt} = U_{in} - T_p. \tag{3.16}$$

Transpiration T_p is assumed to be proportional to the surface area of plant S_p and to the difference between the relative humidity of the air-filled space within the plant H_p and the ambient air H_f:

$$T_p = \alpha S_p \rho (H_f - H_p), \tag{3.17}$$

where ρ is the permeation coefficient of the plant surface to water vapour. Coefficient α is defined as following:

$$\alpha = M_W P^*/RT, \tag{3.18}$$

where M_W is the molecular mass of water, P^* is the saturation vapour pressure, R is the gas constant and T is the absolute temperature.

Plant surface area S_p is assumed to be proportional to the plant mass m (except roots):

$$S_p = \beta m^\theta. \tag{3.19}$$

Here β and θ are an empirical constants. It should be noticed that the S_p values are mostly defined by leaves area of plant. Investigation of dependency between leaves area and the biomass in the Eq. (3.19) is needed to assess parameters β and θ can be performed by using data-driven approaches. This issues, will be discussed in the next sections and chapters.

The inflow of fertilizer U_{in} is proportional to differences between chemical potential of water in plant Φ_p and chemical potential of nutrients in substrate Φ_s:

$$U_{in} = \gamma(\Phi_p - \Phi_s), \tag{3.20}$$

where γ is empirical coefficient. The chemical potential of water in the plant Φ_p is the difference between turgor pressure P_p and osmiotic pressure π_p:

$$\Phi_p = P_p - \pi_p. \tag{3.21}$$

The osmotic pressure is:

$$\pi_p = RTn_s/w, \tag{3.22}$$

where R is the gas constant, T is the temperature, n_s is the number of moles of osmotically active solution, w is the water volume. To model turgor pressure P_p, it is necessary to model the plastic deformations of the leaves. The chemical potential of water in a substrate Φ_s is commonly measured in the experiments.

The state vector for DMD method was constructed based on theoretical findings discussed above. First, the following parameters were included into the state vector: Humidity, temperature of ambient air, temperature of nutrient solution, electrical conductivity (EC), recycling flow rate and e^{-pH} (based on Eq. (3.22), where π_p depends on nutrient concentration n_s, which is an exponent of -pH). These parameters are directly measured by sensors. The light duty cycle was also added to the state vector. The moving average over 3 samples to its values was applied. This was done for taking into account the fact that there is some delay in plant reaction to the illumination.

The state vector was supplemented by non-linear parameters. Taking into ac-

count the transpiration equation (3.17), the following parameter was constructed:

$$F_1 = S(T - T_{wet})/T, \tag{3.23}$$

where it was assumed that humidity $H \approx T - T_{wet}$ and $1/T$ was taken from α coefficient. T is the ambient temperature and T_{wet} is the temperature of fertilizer. Another set of parameters are related to inflow of fertilizer and approximates osmotic pressure π_p (Eq. 3.20):

$$\begin{aligned} F_2 &= e^{-pH} T_{wet}/S^{1.1}, \\ F_3 &= e^{-pH} T_{wet}/S^{1.2}, \end{aligned} \tag{3.24}$$

where e^{-pH} represents the concentrations of nutrients n_s, T_{wet} is the temperature of fertilizer. The total volume of the water in plant w is assumed to be proportional $S^{1.1}$ or $S^{1.2}$. The tugor pressure P_p is connected to plasticity effects. The only possible way to take it into account (using the measured parameters) is to include inverses of all parameters (except for light duty cycle).

In total, 19 parameters were proposed to construct the state vector. Parameters selection procedure for the state vector will be performed numerically, by comparing the prediction accuracy based on the selected subset of parameters. The abbreviations of these parameters are listed in Table 3.3.

Table 3.3: Some control parameters and their definitions.

Variable	**Description**
pH	pH - acidity of an aqueous solution
H	Relative humidity
T	Temperature of ambient air
T_{wet}	Temperature of nutrient solution
EC	Electrical conductivity of feeding solution
F	Recycling flow rate
L	Light duty cycle with applied moving average
F_1	$S(T - T_{wet})/T$
F_2	$e^{-pH} T_{wet}/S^{1.1}$
F_3	$e^{-pH} T_{wet}/S^{1.2}$

Searching for optimal subset of parameters

Next, the optimal subset of controlled parameters that could help to learn the dynamics of the system by using of DMD should be defined. The dataset that describes tomato growth dynamics will be used for modeling (see Section 3.1). In the Fig. 3-4 it can be noticed that after approximately 400 timesteps (which corresponds to approximately 12 days of observations) there are noticeable oscillations of the leaves area projection and its significant growth compared to the previous values. As it was discussed before, these oscillations happen due to the plant physiology, in particular, due to the diurnal processes occurring in plants. Thus, it makes sense to work with two datasets: the first includes 400 timesteps and the second includes 392. In addition, it was discovered that modeling the growth dynamics using the same parameters for the entire dataset, results in worse accuracy compared to the option when the dataset is split into two parts. For each dataset, the DMD algorithm with control will be applied to the values for the first three plants (train set) and predictions will be made for the fourth plant's values (test set) given the vector at the first timestep and the matrix A obtained from training procedure of DMD. Different combinations of parameters will be used for control. The optimal choice of set of parameters will be the combination that gives the smallest root mean squared error (3.25):

$$\varepsilon_{RMSE} = \sqrt{\frac{1}{N_{test}} \sum_{i=1}^{N_{test}} \|x_i - x_i^{pred}\|^2}. \tag{3.25}$$

The smallest error was obtained with the following parameters: H, T, T_{wet}, L, F_1, F_2, e^{-pH}, $1/F_1$, $1/F_2$, $1/F_3$ for the first part of the split dataset and T, T_{wet}, L, $1/F$ and $1/F_3$ for the second one. It should be noticed that the majority of the parameters are the same for each part of the split dataset. However, the weights for these parameters are different in each case and the impact of each parameter on the plant growth dynamics has a different effect. By using the reconstructed matrix A and the initial vector predictions of the growth dynamics (projected leaves area) for 400 steps ahead (12 days) with acceptable accuracy were obtained. Such big prediction horizon opens wide possibilities for optimizing of the system. During

the experiments it was noticed that it is even possible to reconstruct the diurnal oscillations.

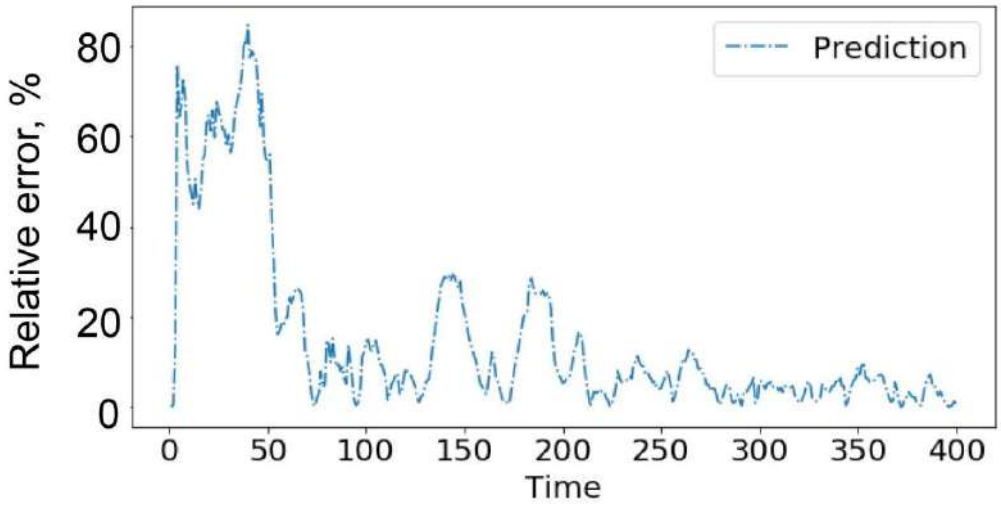


Figure 3-19: Relative errors for DMD prediction with control applied to the first part of the data.

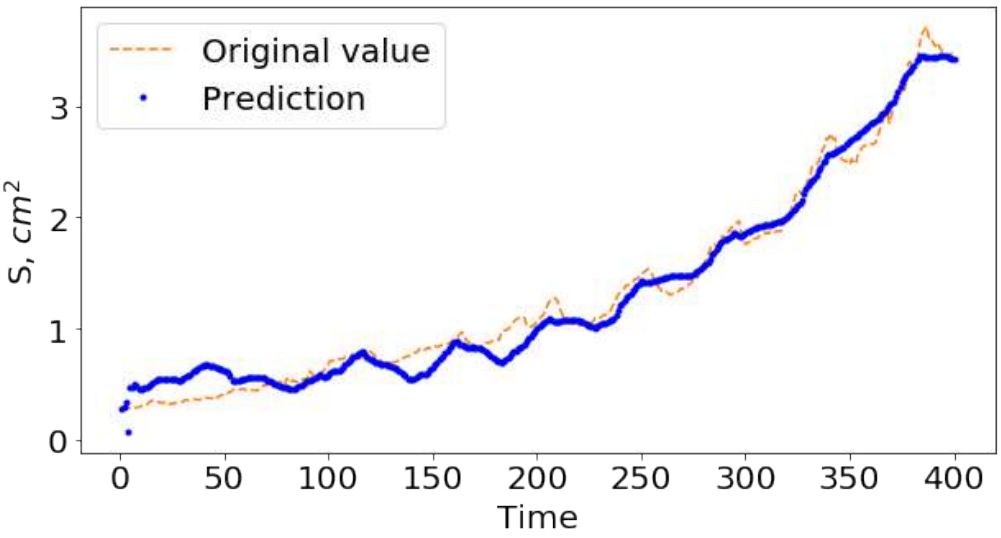


Figure 3-20: Prediction of the projected leaves area for DMD with control applied to the first part of the data.

Results of modeling by using DMD

Fig. 3-19 shows the relative error of DMD prediction of projected leaves area for the first part of the data. This relative error decreases in time and has fluctuations

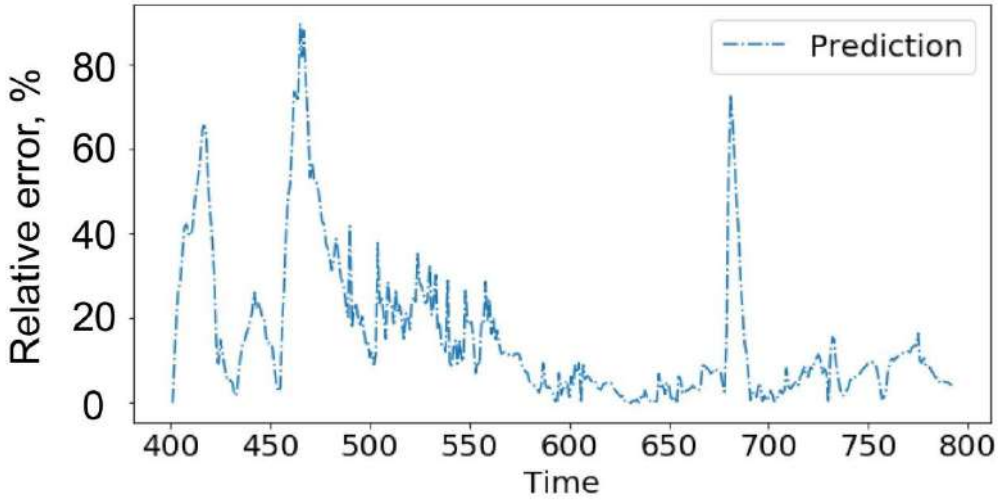


Figure 3-21: Relative errors for DMD prediction with control applied to the second part of the data.

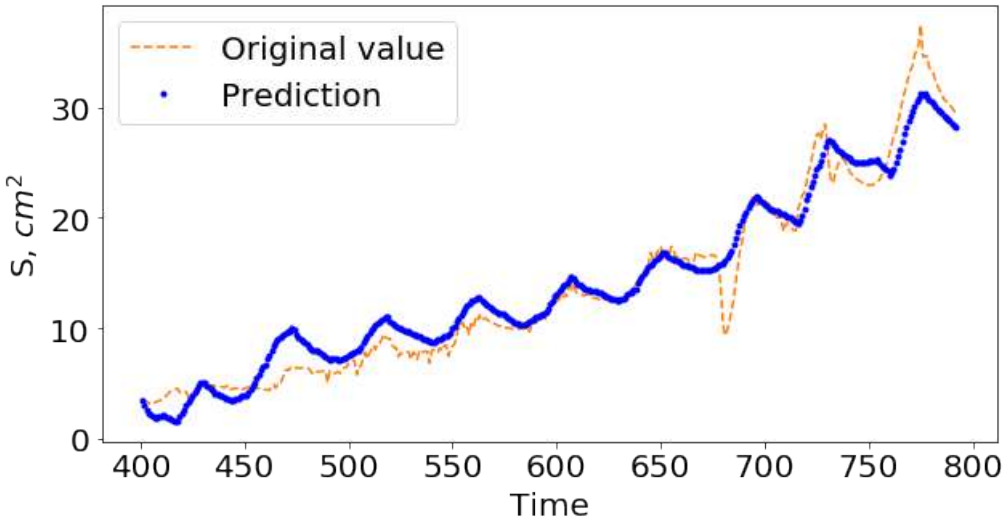


Figure 3-22: Prediction of projected leaves area for DMD with control applied to the second part of the data

around 10% that means reasonable accuracy. The average relative error for the whole time interval for the first part of the modelled data is 15%. Fig. 3-20 presents the originally measured projected leaves area of tomato growth over time and the modelled projected leaves area for the first part of the data. Fig. 3-22 demonstrates the measured and modelled projected leaves area for the second part of the data. The fluctuations of projection are observed in measurements and in the reconstructed model. Relative error (see Fig. 3-21) also decreases in time to its minimum of $1.1 \cdot 10^{-3}$% and then increases. The spike in Fig. 3-22 which leads to the spike in latter part of Fig. 3-21 happens due to the rapid and dramatic decrease in the pH value during the experiment. The average relative error for the whole time interval of the second part of the modelled data is 16%. Also, 4-fold cross-validation was performed for the data obtained from four plants. For the first part of split dataset the cross-validation error (as an average relative error) has the mean 23.9% and the variance 8.8%. For the second part of the split dataset the mean is 18.5% and the variance 8.6%. The result of the variance of the error is not completely representative for the data and the method used since it was applied only to the 4 folds cross-validation and the accuracy increases over the time on average. Overall, the main trend and oscillations can be accurately predicted using the proposed approach.

Conclusions

The promising approach for modeling of the plant growth dynamics using the data obtained from artificial growth system and parameters was demonstrated. This method was derived from growth model and data-driven method, namely, dynamic mode decomposition. The achieved results on reconstructing the plant growth dynamics and assessment of the system dynamics ensured its practical feasibility, robustness, and prediction accuracy proved by the average relative error that is 15%. Such a method can be widely and easily applied to other types of plants, because in the proposed method specifics of plant growth dynamics are taken into account by DMD algorithm that use as the input parameters derived from general equations. This means that it is not necessary to adopt equations and coefficients for each type

of plant for using the proposed approach.

3.5 Merging 2D/3D computer vision techniques for biomass growth assessment

As it was discussed previously, one of the main objective function for optimization of plant growth dynamics is biomass. However, it is very complicated to obtain direct measurements of plant biomass without interruption of growth process. It is also time consuming to do direct measurements of plant biomass. One of the possible ways to perform biomass assessment and prediction is usage of data-driven non-invasive image based techniques. Based on these approaches it will be possible to obtain relevant information about plant biomass with high time resolution which in turn will allow to perform fine control of objective function (biomass) and to solve optimization problems for growing plants in greenhouses. In this section it will be described a generic approach for predicting plant biomass growth dynamics using data-driven approaches. This approach is based on statistical analysis of the dependency between actual leaves area and biomass without performing multi parameter modelling of each particular plant. In the following modeling, sequences of 2D and 3D images (clouds of points) will be used along with associated image analysis. This allows merging benefits of both techniques: 2D and 3D imaging. In this section tomato growth dynamics will be investigated. Two datasets will be used: (1) dataset of sequence of 2D images of growing dwarf tomatoes and environmental parameter, that was used for DMD modeling (see Section 3.1 and Section 3.4), (2) dataset of 3D images of dwarf tomatoes on different stages of growth (cloud of points) for biomass estimation (see Section 3.1).

For modeling the plant biomass growth dynamics, the following key steps were identified:

- Collection of 2D and 3D data describing plant growth dynamics.
- Derivation of dependencies between the actual leaves area and biomass of the investigated plants based on 3D images (cloud of points).

- Modeling and prediction of projected leaves area based on 2D images and different modeling techniques.
- Reconstruction and prediction of the biomass based on derived dependencies between actual leaves area and biomass and predicted projected leaves area.

Following the methodology, for further reconstruction of the biomass using predicted projected leaves area it is important to find dependency between them. It should be noted that the term biomass is used in the meaning of plants' volume, because the total mass of the plant is determined mainly by water content, which in turn form the volume of the plant. Thus, knowledge about plant volume allows easily to reconstruct biomass. In Section 3.4 it was mentioned that the coefficients of Eq.(3.19) are identified empirically. One of the approaches is to make direct measurements of the biomass and leaves area, but during these, it is necessary to disturb the plant. Another approach that is proposed and will be used is taking 2D and 3D images, of plant during its growth. In this case the disturbing of plant is minimal. First, the Eq.(3.19) was rewritten in the form Eq. (3.26):

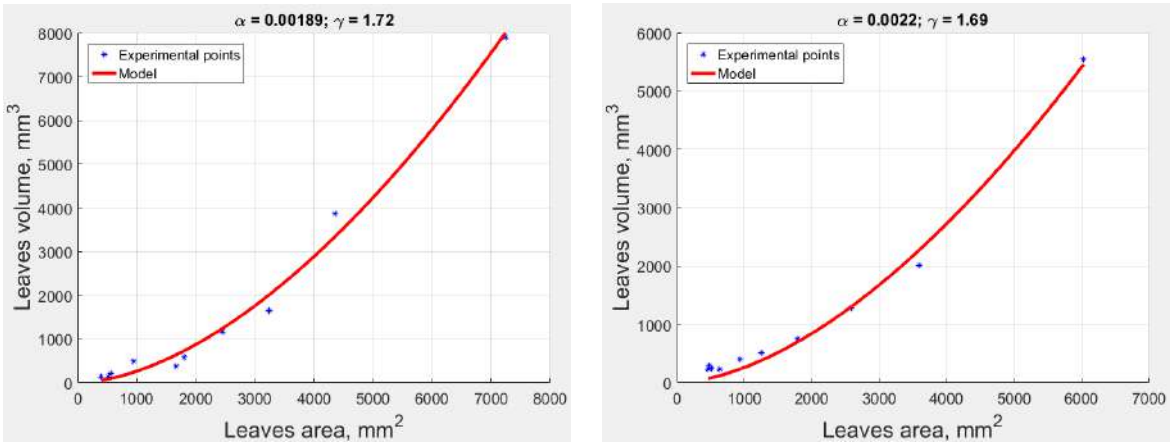
$$m = \alpha S^{\gamma}, \tag{3.26}$$

where according to Eq.(3.19) $\alpha = \frac{1}{\beta}^{\frac{1}{\theta}}$ and $\gamma = \frac{1}{\theta}$. Using the dataset described above, volume of plants leaves and corresponding actual leaves area were calculated using 3D clouds of points. In Fig. 3-23 the calculations of actual leaves area and volumes as well as the estimated parameters of Eq. (3.26) are shown for three plants (out of 18 plants in the experiment for which 3D measurements were obtained). In particular, these three plants were chosen for modeling because 3D data were collected very frequently, every day. For other plants in the experiment 3D data obtained once per 2-3 days. The summary of the obtained parameters Eq. (3.26) is presented in Table 3.4.

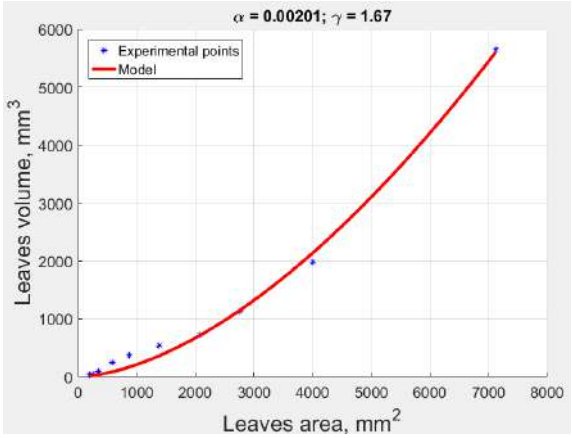
From Table 3.4 it can be noticed that estimated parameters for two similar sorts of dwarf tomatoes have close values. This means that it is possible to use them for modeling similar types of dwarf tomatoes. As it was discussed previously (see Section 3.1 and Section 3.4), a large dataset of 2D sequences of images of MicroTina dwarf

Table 3.4: Summary of the obtained parameters for actual leaves area and biomass dependency.

Sort	α	γ
Bonsai micro	0.0019	1.72
Bonsai micro	0.0022	1.69
Bonsai	0.0020	1.67



(a)



(b)

Figure 3-23: The relationships between leaves volume and actual actual leaves area for (a) Bonsai micro (two selected plants), (b) Bonsai (one selected plant) dwarf tomato sort.

tomatoes growth was created. MicroTina dwarf tomatoes have a similar structure as the Bonsai and Bonsai micro tomato sorts (only one difference is that it growth slower and finally it is more compact). Thus, it is possible to use the obtained coefficients for modeling and predicting the MicroTina tomatoes biomass based on 2D images. For all 4 tomato plants growth parameters μ and S_{max} of the growth model Eq. (3.2) were estimated based on the trend of projected leaves area growth calculated from the set of 2D images. The results are presented in the Table 3.5

Table 3.5: Growth parameters estimation of the Verhulst model for four dwarf tomato plants

	1	**2**	**3**	**4**
μ, 1/day	0.2806	0.2757	0.2705	0.2715
S_{max}, cm^2	191.3	110.4	120.3	154.5

From the Table 3.5 it can be noticed that estimations of growth rate for four plants is similar. This means good estimation of these parameters. For the tomatoes, data were obtained on the initial stage of vegetation period. Such precise assessment of the growth rate was observed and explained while modeling the initial stage of growth dynamics of lettuce in the Section 3.2. Figure 3-24 shows the Verhulst model fitting Eq. (3.2) applied to one of four tomato plants. The average relative error is 13%.

Using the fitted model (see Fig. 3-25a) it is possible to assess and to predict biomass (volume) (see Fig. 3-25b) based on α and γ parameters obtained previously. It should be noted that any suitable and precise method for modeling and predicting projected leaves area growth can be used as a basis for further biomass assessment. For example, Kalman filtering (see Section 3.2) or DMD (see Section 3.4) that rely on both model-based and data-driven techniques or pure data-driven approaches based on recurrent neural networks and fully convolutional neural networks (see Section 4) have already shown their accuracy and robustness for plant growth dynamics prediction.

Besides obtaining dependencies between biomass and leaves area using collected 3D data for biomass prediction it is also possible to reconstruct and investigate some physiological processes in plants. After plotting and approximating the trend of how

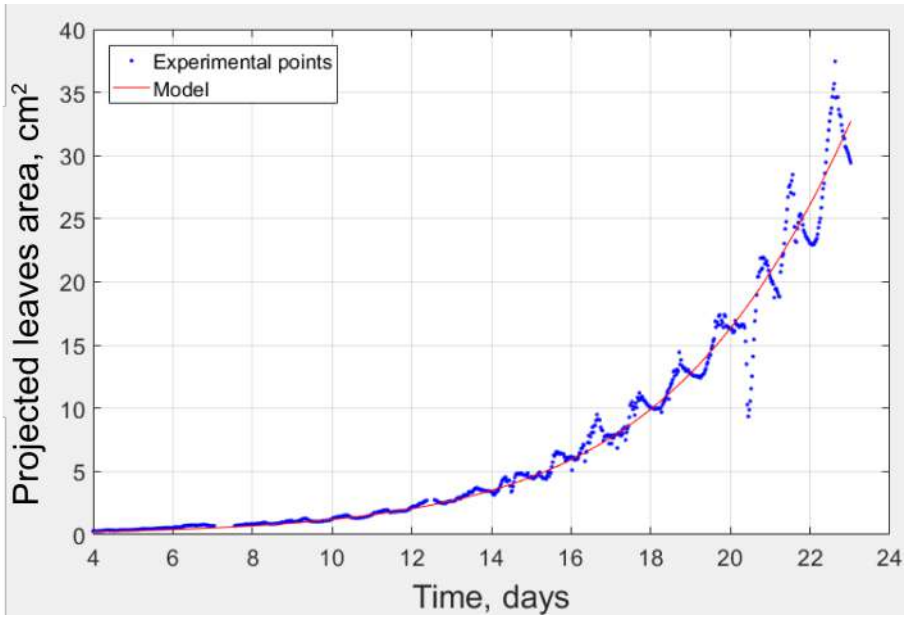
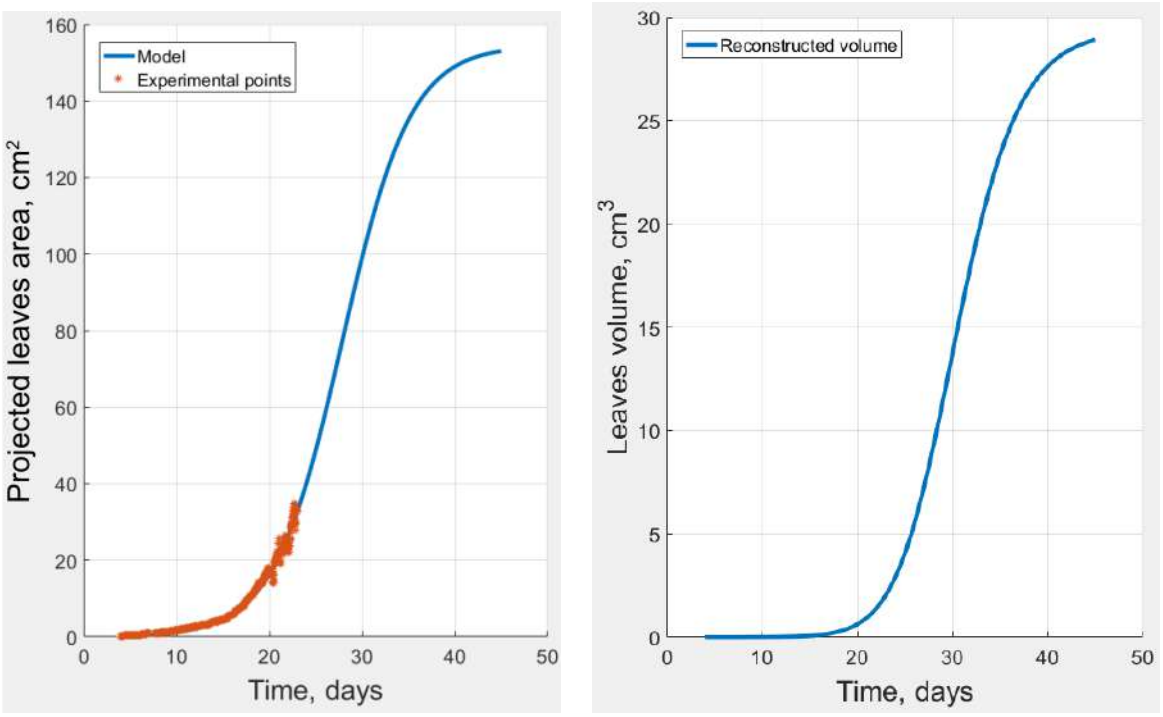


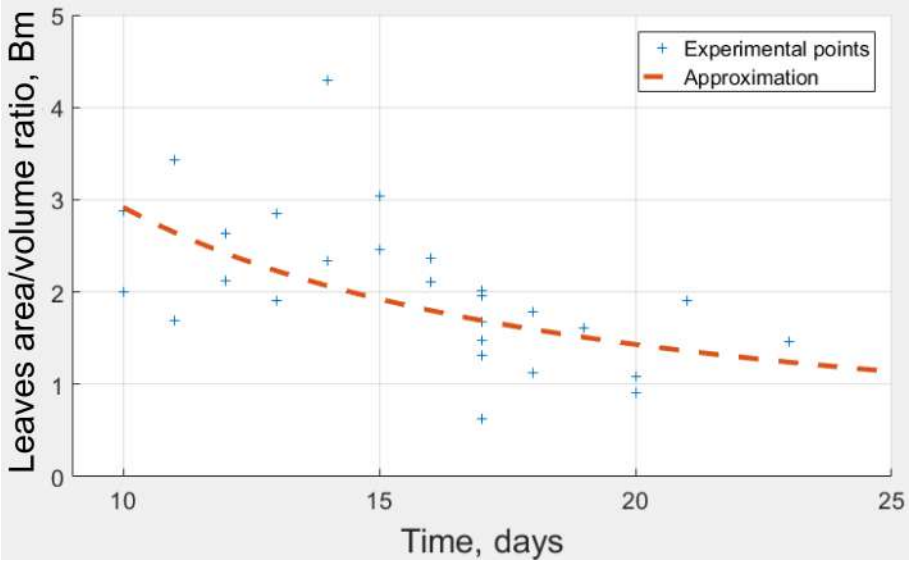
Figure 3-24: Example of Verhulst model fitting to experimental data on tomatoes growth.



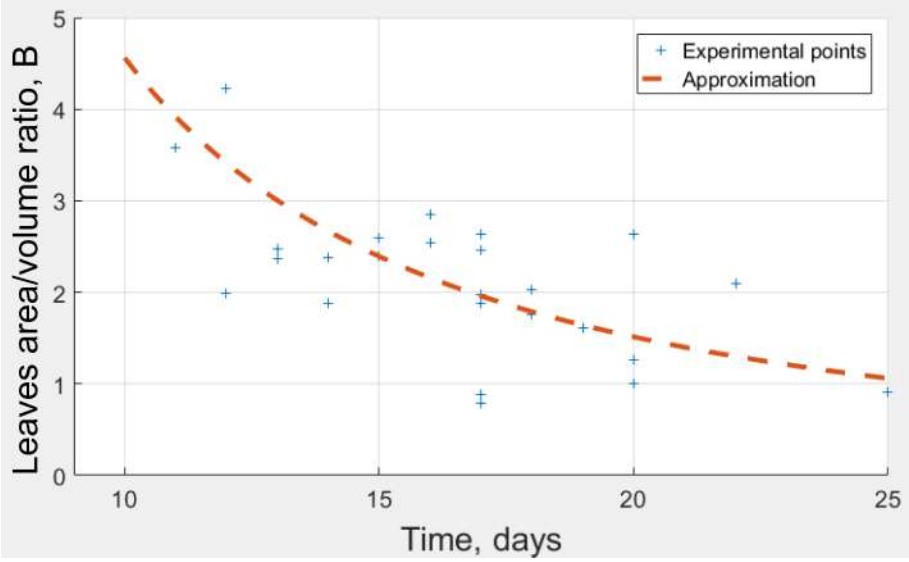
(a) (b)

Figure 3-25: Biomass (volume) prediction using predicted projected leaves area and obtained actual leaves area/biomass dependencies

the ratio between leaves area and biomass changes in time for several plants (power model) it was observed effect that this ratio decreases and tends to 1 (see Fig. 3-26). The standard deviations of the fitted models are 0.55 and 1.3 respectively. Similar findings were observed using model-based approaches in [Weraduwage et al., 2015].



(a)



(b)

Figure 3-26: Dynamics of the ratio of actual leaves area/volume changing in time for (a) Bonsai micro and (b) Bonsai tomato sorts.

Conclusions

In this section, the workflow and the possibilities of 2D/3D based approaches for non-invasive and robust plant biomass growth dynamics prediction were shown. One of the advantages of this method is that it is necessary to obtain 3D data only once for the particular type of plant. Then predictions of biomass can be done using only 2D images and this in turn opens wide possibilities for solving real-time biomass growth optimization problems in greenhouses. One of the limitations of this method is that it gives accurate results only for plants with a simple structure of for plants at the initial stage of growth when leaves are not overlapping. However, during the later stages of growth, it can be overcome by introducing the set of 2D cameras that have different angles of view on the plants. The other limitation is that there is a difference between the actual leaves area measured by 3D camera and projected leaves area, measured by 2D camera, but at the initial stage of growth this difference is insignificant. The benefit of the proposed method is that 2D cameras are cheap and easy to use. All these will give affordable actual information to farmers about the current and future status of plant growth dynamics. A combination of 2D and 3D techniques, as it was shown, also gives ample opportunities to perform a non-invasive investigation of physiological processes in plants.

3.6 Conclusions

In this chapter, the developed experimental setups for the collection of the relevant data was presented. Hybrid modeling approaches that allow to describe and predict plant growth dynamics were evaluated. Using the experimental setups several comprehensive and novel datasets that describe plant growth dynamics (2D and 3D images) and growth conditions (data from sensors) in artificial environments were obtained. These datasets were used for testing the following proposed methods:

- Extended Kalman filtering. Growth parameters such as maximum projected leaves area and growth rate were predicted on the modeled and experimental data. Computer vision methods were used for calculating projected leaves area. The main benefit of this method is the high computational efficiency

which allows its implementation in embedded devices. Also, there is a wide possibility to include additional parameters into a state vector and to evaluate the effect of them on growth dynamics.

- Instance segmentation. It was shown experimentally that using instance segmentation in couple with a simple growth model allows to predict the growth rate of each leaf of the plant, enabling the real-time detailed reconstruction of growth dynamics.

- Dynamic mode decomposition. By using the features derived from growth model that is based on differential equations along with dynamic mode decomposition, it was shown that it is possible to perform accurate prediction of plant growth dynamics. The main benefit of this method is that it is accurate, includes physical principles in modeling procedures and small amount of data points are needed to train DMD algorithm.

- The workflow for biomass prediction based on the merging of 2D and 3D approaches and simple models was shown and evaluated. Using these approaches, the dependence between actual leaves area and biomass was estimated and the prediction of biomass based on 2D images was performed.

Overall, the proposed methods showed good trade-off between complexity, universality and accuracy for plant growth dynamics assessment and prediction.

Chapter 4

Data-driven enhancement for plant growth modeling in controlled environments

4.1 Recurrent neural networks and computer vision for plant growth dynamics prediction

Recent advances in computational methods, machine learning and increase of computational power, together with the availability of sensors, enabled the collection and processing of enormous amounts of data [Mois et al., 2017, Davies and Clinch, 2017a]. This progress led to the development of data-driven modeling approaches, e.g. ANNs, possessing huge expressive power for high-dimensional data description and generalization for precision agriculture.

In the following section, the RNNs for prediction of the plant growth dynamics will be used. The RNN is a class of ANN where the nodes contain the feedback response and enable the storage of information about their internal state. One of RNNs attractive features is that they are potentially able to link previous information with the current state. The RNN can process the data that are represented as time dependent sequences by using the internal state information. A typical RNN may have a problem with the processing of long-term dependencies. To overcome

this problem the Long Short-Term Memory (LSTM) NNs were introduced as a special architecture of RNN capable of learning long-term dependencies [Hochreiter and Schmidhuber, 1997]. The key element of LSTM is a cell state which can be changed in the process of training. This feature is important for modeling the plant growth dynamics since the future dynamics of the plant growth is in strong relation with the previous states passed long time before. Recently, many applications for the LSTM NN architecture have appeared. [Stollenga et al., 2015]. However, the application of RNN (LSTM or GRU) in *precision agriculture* for crop yield prediction or plant growth dynamics description based on environmental growth conditions is a novel research direction [Chlingaryan et al., 2018]. The ubiquitous and efficient application of these models is essential. For tackling this problem, a low-power sensing solution able to run the models on board and functioning in a distributed manner was proposed and tested.

To perform plant growth modeling and dynamics prediction, an experimental setup based on the hydroponic approach allowing simultaneous plants growth in different conditions (nutrient solutions) was designed. This setup was equipped with an automatic image acquisition system and controlled LED illumination. Using this experimental setup, one month experiment was conducted. During this experiment, the sequences of raw images of plants growth were collected in fixed time intervals under different conditions. Using similar approach, discussed in the Section 3.2 projected leaves area were calculated simultaneously based on the obtained images. The data on the projected leaves area were used for training the NNs and other machine learning algorithms.

Experimental setup and measuring system

The designed and assembled experimental setup is shown in Fig. 4-1. In this system, it is possible to grow up to 54 small plants (e.g. dwarf tomatoes), and to automatically monitor their growth by cameras. In the experiment 48 dwarf tomatoes MicroTina were grown [Scott et al., 2000]. The tray for growing plants was separated into 6 isolated sections. In each of them, the plants were fed using different feeding solutions. Each of the sections contained 8 plants. They were fed

Figure 4-1: Experimental setup: growing (bottom) and data collection system (see video cameras on top).

manually and were grown in a 0.65-liter rock wool block. The tray was covered by the foam plastic so as to facilitate the post-processing images. The imaging system contained 6 high resolution cameras Logitech c920 mounted above each section on a regulated platform. The artificial illumination of the plants was provided with 150 Watt Neususs LED. Additional white LED illumination was added and switched on while the cameras took pictures. The automatically controlled day/night light duty cycle was 17/7 hours. It was slightly decreased by the end of the experiment to 15/9 hours due to plant physiology reasons.

At the beginning of the experiment, 70 seeds of MicroTina were germinated in a small rock wool cartridge in a separate tray under LED light with a 17-hour duty cycle (from 1 a.m. to 6 p.m.) at a 0.6 m. distance. The commercial nutrient solution in the amount of 1.3 l was used for germination of all seeds (Vostex concentration: 0.5 ml/l of water). A special type of solution for each section was prepared. There were five feeding solutions that contained the base feeding solution which is the output of the filtering system used in life support systems. To this "Base" solution special additives (P, K, Ca, Mg, Microelements) were added. The other one is the reference solution known as the "Hoagland" nutrient solution [Hoagland et al., 1950]. Then 48 of 70 seeds that had already been propagated, were put into the experimental setup. Each rock wool block from a particular section was watered with 0.5 l of

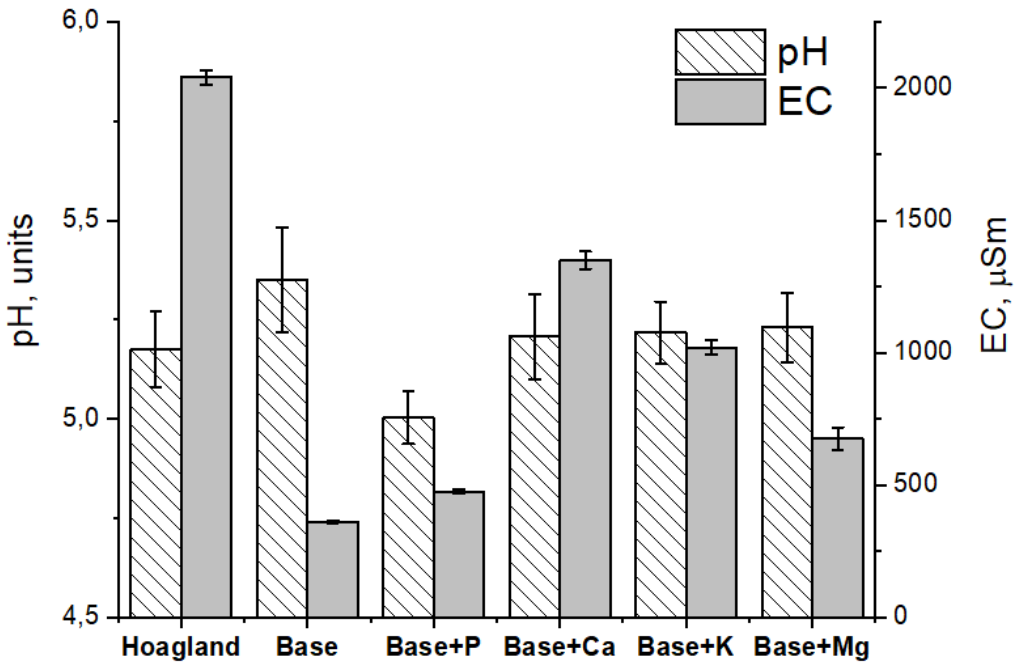


Figure 4-2: Nutrient solutions properties: pH and EC. Properties were measured for each type of nutrient solutions that were prepared four times during the experiment.

a solution prepared for a section, respectively (8 blocks-4 l of each solution were initially used). The experiment was conducted slightly longer than one month; the solutions, therefore, were prepared four times. Properties, such as pH and electrical conductivity (EC) of the prepared nutrient solutions, are presented in Fig. 4-2. EC and pH were measured with ± 0.5 μSm and ± 0.005 accuracies respectively. It is essential to prepare solutions with the same EC for ensuring similar conditions in each section of the plant growth. Deviations of the pH level are not crucial, since for each type of solution it should lie within a certain range. Every day at around 15:00 each plant was fed with 20 ml of nutrient solution.

Image data acquisition and elaboration

During this experiment images of each section with plants were automatically taken and recorded every 30 minutes. In total, 5514 raw images with a strict time dependency were collected. Examples of raw images are shown in the Fig. 4-3.

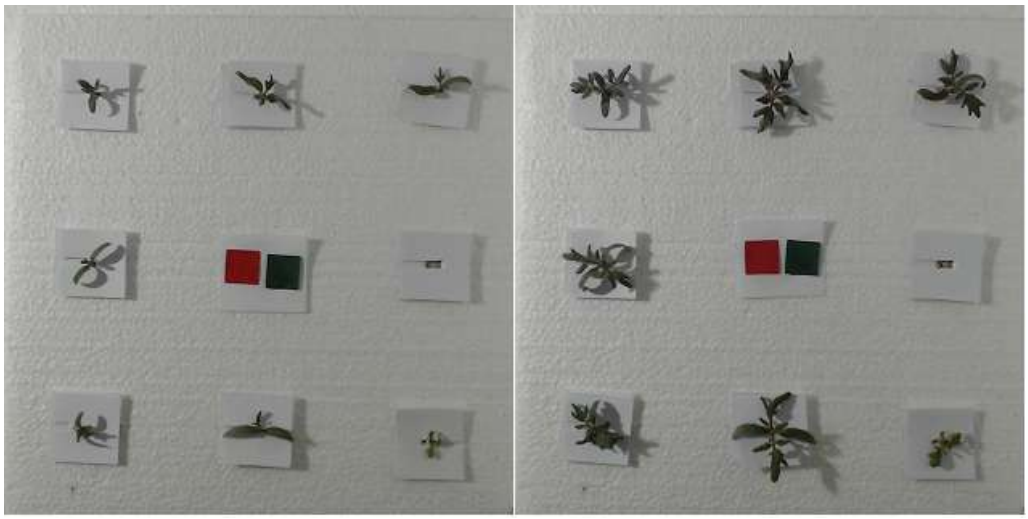
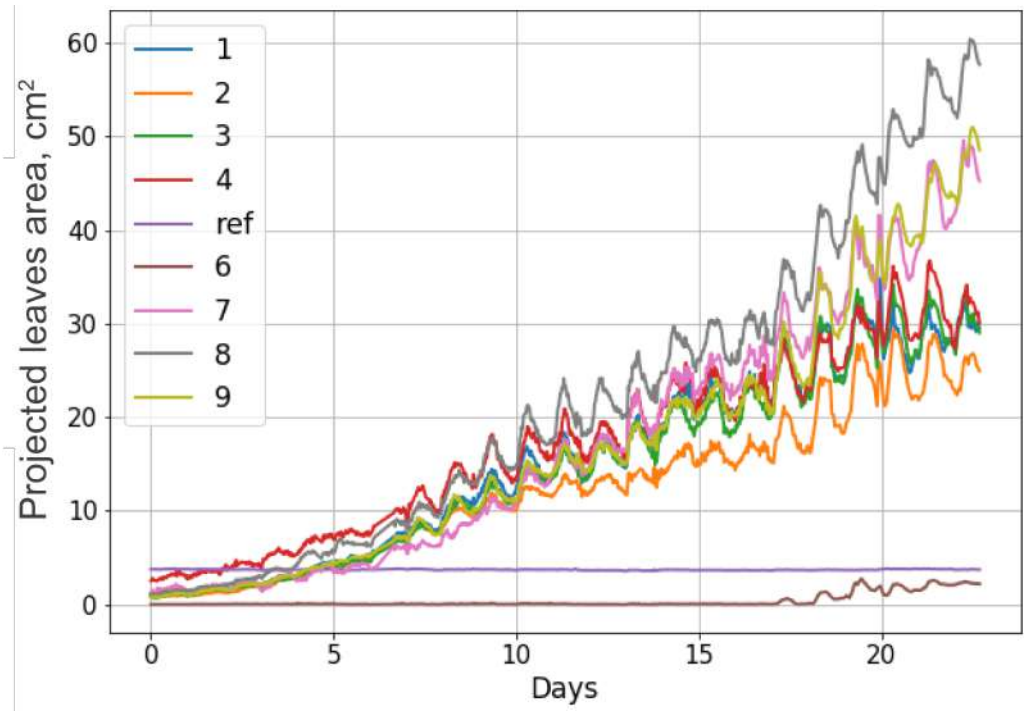
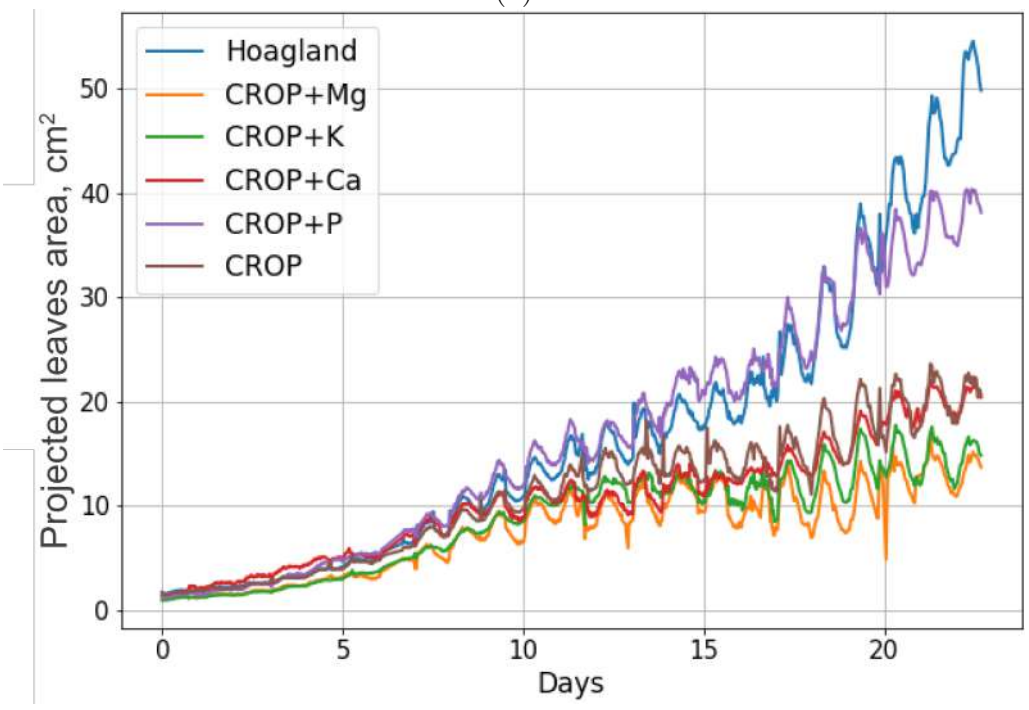


Figure 4-3: Example of top-down images obtained in the experiment. On the left: tomatoes on the 5-th day after germination. On the right: tomatoes on the 10-th day after germination

Prior to starting the experiment, cameras and the algorithm for projected leaves area calculation were calibrated. For this purpose, from the tomato leaf, it was cut out the part of the leaf of a certain area (0.5, 1, and 2 cm^2 leaf areas were tried out). Then, the projected leaves area was recalculated relatively to the label with a known area (red square with a known area) and set up the algorithm and lighting, trying to minimize the error in determining the projected leaves area. A similar calibration procedure was also performed for different parts of the camera field of view. During the main experiment, the same illumination as in the calibration procedure was used. The leaves area of the tomatoes was not measured directly during the experiment, but several control checks of the algorithm and the camera were carried out. Due to the fact that during these checks it is necessary to stop the experiment for about 5 minutes, these checks were not done frequently though. During checking, a part of leaf with manually measured area was placed in the camera field of view. Then the reference projected leaves area was estimated by taking two consecutive pictures of only tomatoes and tomatoes with the reference. Then the projected leaves area obtained from one picture with the reference was subtracted from projected leaves area obtained from another picture without the reference. As a result of these checks at various stages of the experiment, no serious deviations in the operation of the



(a)



(b)

Figure 4-4: (a) Example of projected leaves area calculation for dwarf tomato plants that were grown in section with "Base+P" feeding. (b) Average projected leaves area of plants in each growing section with corresponding feeding solution.

system and an estimate of the surface area of the foliage were found.

Figure 4-4a shows an example of the projected leaves area calculations for each tomato plant fed by the nutrient solution "Base + P", the average variance of the calculated projected leaves area is 5.4. Figure 4-4b shows the average projected leaves area for the plants in each section, the average variance of the projected leaves area is 21.2. From Fig. 4-4b it can be concluded which additive to the base solution is the best. In this case phosphorus additive has the best effect among others. It is highly important that it is possible to estimate the effect of different factors on plant growth dynamics by using simple cameras. This opens a wide possibility to drive plant growth process in the most useful direction. In totally, 44112 measurements of the leaves area projection were obtained. It should be noted that the estimation of the leaves area was done by measuring its maximum projection which, in general, may not be equal to the real leaves area. However, these measurements also can give us additional information about the hidden dynamics of the plant growth (such as diurnal fluctuations in relative location of leaves discussed before). This additional information can be included in the predictive model, making it more precise. As there is no need to use classical statistical methods, it is possible to directly use obtained data as an input to NN without estimation of errors.

Performance evaluation

Calculation experiments were done by using the data of 12 days and the LSTM model was trained for each section. The dataset was split into the training and test sets with 400 and 200 data points for each section, respectively.

The adam optimizer was used with hyper-parameters: $lr = 0.001$ (learning rate); $beta_1 = 0.9$ (exponential decay rate for the first moment); $beta_2 = 0.999$ (exponential decay rate for the second moment); $epsilon = 10^{-8}$. Hyper-parameters were taken similar to the most commonly used in practice [Kingma and Ba, 2014]. The mean squared error was used as a loss function. The hidden states were reset for each epoch of training. The network was trained for 20 epochs which is a reasonable amount for the stabilization of the loss function for most of the tested architectures with one hidden layer of LSTM. Different amount of points from the previous steps

for doing predictions and realized, that 3-4 data points set is enough for doing accurate predictions even for a 3 hour horizon. The train/test dataset was created in a following way. Each train/test data sample contained 13 points: a sequence of 3 points (projected leaves area) from the previous three steps and a sequence of 10 points for the next 10 steps. Dataset preparation process also included transformation to stationary time series and scaling to (-1:1) range. In this work, it was evaluated the root mean square error (RMSE) of predictions for time horizon from 30 minutes to 5 hours.

The results of leaves area (projection) prediction presented for three out of six possible options (6 different solutions were in the experiment) are shown in Fig. 4-5a, Fig. 4-5b and Fig. 4-5c. Projected leaves area in Fig. 4-5a, Figure 4-5b and Fig. 4-5c was taken as a sum of leaves area of all the plants in the section. Diurnal fluctuations were also predicted by the trained model. This result allows us to assess real leaves area by interpolation of the resulting curve on top. The results of this prediction demonstrate good fit to the ground truth. RMSE in Fig. 4-6 shows how the error changes with respect to the size of step prediction. It is worth noting that we have a good prediction even for the 5 hour prediction horizon. Since RMSE varies from 9 to 14 for different solutions for 5 hour prediction horizon and for the test dataset projected leaves area values varies from approximately 100 to 140, it was demonstrated that the obtained performance withstands the application requirements even for the 5 hour horizon (maximum that is required): 5-10% of relative error. For lower prediction horizon the relative error is less. There are several types of responses of the plants to the stress: long-term and short-term. The long-term response typically appears in up to several days. The short-term response to the stress is typically much more damaging and has a time lag of up to 5-6 hours [Roy, 2012, Acevedo et al., 1971]. Modeling the plant growth for this period is vital as it allows for predicting and preventing the effects of stress on the plant growth dynamics. It is expected that the accuracy of predictions capture the diurnal fluctuations of the plant growth dynamics (relative leaves locations) as it is one of the main and explicit short-term driving reason for changing of leaves area projection. Thus, the amplitude of fluctuations can serve as the lowest bound of

permissible prediction accuracy.

Comparative study on RNN architectures and methods

The ablation study was also performed with different amount of neurons and different architectures. Some of results that are calculated for the section with "Base+P" feeding solution presented in Table 4.1. As an example, for 2 LSTM units, there are 62 trainable parameters and the RMSE is 3.25 and 9.01 for 30 minutes and 3 hours prediction horizon, respectively. For 10 LSTM units, there are 670 trainable parameters (which is not acceptable as we have 400 data samples for training and the common rule is that the number of estimated parameters should be less than the train set size) and the RMSE is 3.19 and 8.72 for 30 minutes and 3 hours prediction horizon, respectively. If 2 layers of LSTM NN will be used, then in the case of 2 LSTM units on each layer RMSE will be 3.27 and 9.2 for 30 minutes and 3 hours prediction horizon, respectively, and in the case of 4 LSTM units on each layer the RMSE is 3.2 and 9.0 for 30 minutes and 3 hours prediction horizon, respectively. Based on these findings it was decided to use LSTM with 4 units as a compromise between the amount of parameters (178), accuracy (RMSE is 3.18 and 8.6 for 30 minutes and 3 hours prediction horizon, respectively) and complexity.

Table 4.1: Performance of different architectures of LSTM neural networks.

Amount of LSTM neurons	**Mean RMSE and std for 0.5/3 h prediction horizon**	**Amount of trainable p-rs**
2	$3.25 \pm 0.022/9.01 \pm 0.13$	62
10	$3.19 \pm 0.012/\ 8.72 \pm 0.21$	670
2/2	$3.27 \pm 0.002/9.2 \pm 0.041$	118
2/4	$3.2 \pm 0.004/8.9 \pm 0.045$	210
4	$3.18 \pm 0.014/8.6 \pm 0.042$	178

The performance of LSTM was also compared with the performance of 4 units Gated Recurrent Unit (GRU) RNN with 146 parameters. The RMSE is 3.21 and 8.7 for 30 minutes and 3 hours prediction horizon, respectively, for GRU RNN. It means that it is also possible to use other highly efficient types of RNNs. The same amount of epochs, i.e. 20, was used for training the RNN with GRU neu-

rons. The adam optimizer with the same parameters as for the LSTM RNN was used. Also, data preparation procedure was implemented for GRU RNN, including the transformation to the stationary time series, scaling, splitting to the train/test dataset, was equal to the tested LSTM RNN. Non-recurrent approaches were also tested. Prediction based on CNN over a window of data. 1D convolution with 16 filters and $kernel\ size = 2$ was tested. Then max pooling and dense layers with ReLU activation were put. The data prepossessing procedure was the same as for the LSTM RNN, however, 50 epochs are required for convergence and loss stabilization. The amount of parameters is 328 and accuracy (RMSE) is 3.71 and 9.7 for 30 minutes and 3 hours prediction horizon, respectively. This result is worse than for the LSTM architecture used in this work and more parameters have to be estimated. Increasing the amount of filters/layers leads to dramatic increase in the number of training parameters, while small number of training parameters is crucial for the tested dataset. For window size 10, $RMSE = 4.61$ (excellent result over all window sizes and approx. 40% worse than for LSTM), for window size 2, RMSE reached 7.91. Based on this research it was realized that LSTM showed its' robustness compared to simpler approaches.

It was investigated how newly obtained data during the experiment could improve the accuracy, and whether it is important to retrain the algorithm. For doing this, tests with several splitting options of the dataset were conducted. Four hundred data samples in the training set showed the acceptable accuracy of the predictions and further increase of training samples does not significantly improve the accuracy. It means that it is not necessary to retrain the NN during this time period for achieving more precise results, and this is highly important since the embedded sensing system with AI onboard works only with the pre-trained NNs - there is no need to periodically upload newly trained NN, which could lead to the decrease of system autonomy.

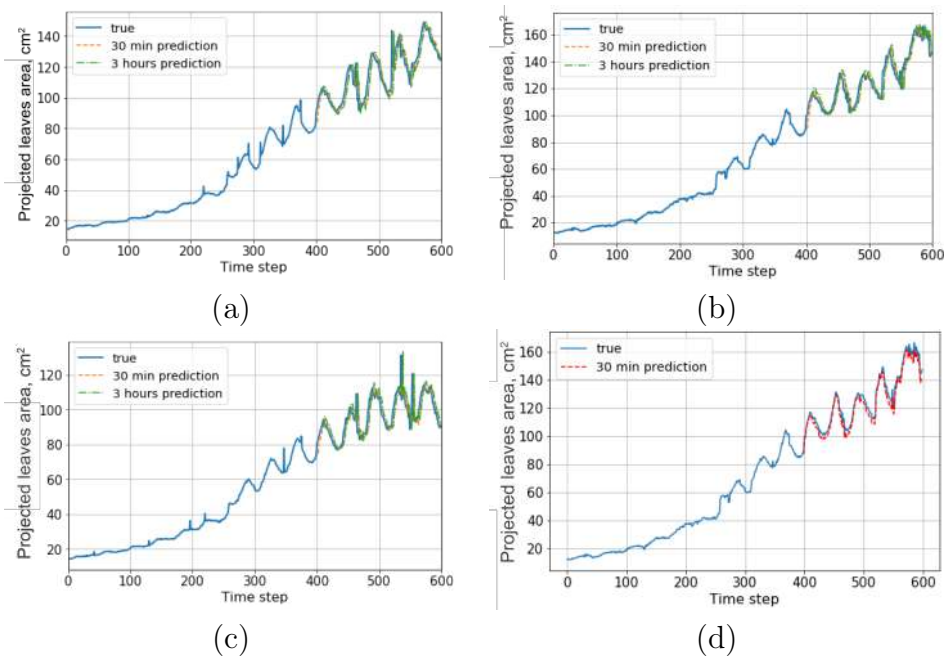


Figure 4-5: (a) Prediction of a projected leaves area for section that fed with "Hoagland" nutrient solution. (b) Prediction of a projected leaves area for section that fed with "Base + P" nutrient solution. (c) Prediction of a projected leaves area for section that fed with "Base + Ca" nutrient solution. (d) Prediction of a projected leaves area based on autoregression for section that fed with "Base + P" nutrient solution. Each time step in (a), (b), (d) represents 30 minutes.

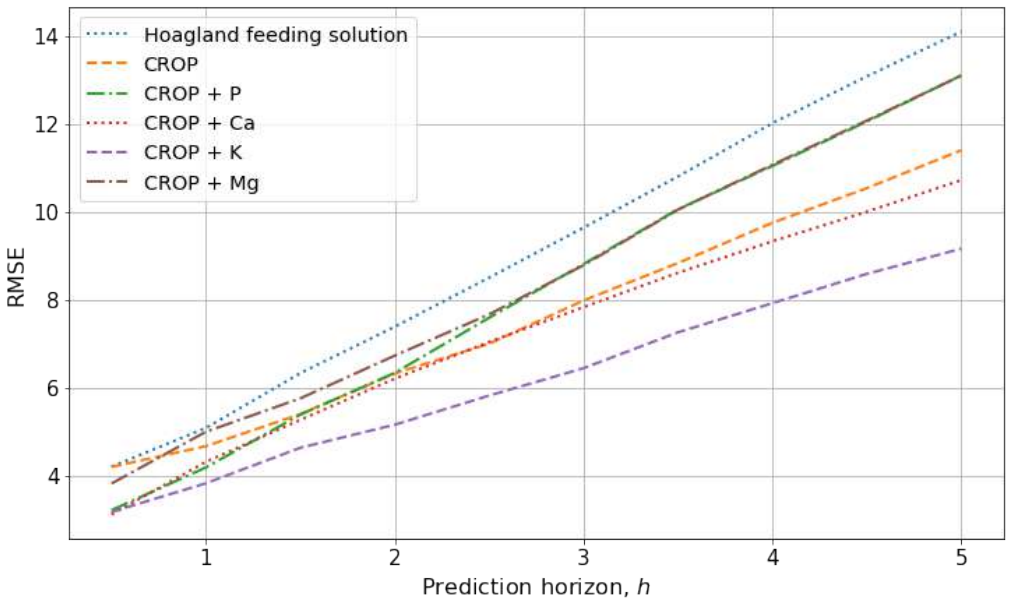


Figure 4-6: Dependence of root mean squared errors for prediction of projected leaves area on the number of prediction steps for all 6 solutions.

Implementation and performance evaluation of the developed ML models in low-power embedded systems

Embedded low-power systems allow to construct distributed autonomous sensing systems for precision agriculture. One of the attractive features of such systems in precision agriculture is that it is easy to cover huge growth areas for sensing without constructing complex communications because such systems not only receive sensing information, but also process it on-board, sending only useful processed information. The LSTM NN that showed the best performance in the Section 4.1 was first trained on the desktop and then implemented by converting to the binary graph on the embedded device. It was decided to use Raspberry Pi 3B in couple with the external Graphical Processing Unit (GPU) Intel Movidius based on *Myriad* processor for running NNs on it [Ionica and Gregg, 2015]a. The block diagram of the low-power prototype is presented in Fig. 4-7. One of the limitations of the external GPU is that it allows only to run trained NN architectures, it is impossible to train NNs on board. However, external GPU allows to run trained state-of-the-art deep neural networks (CNN, RNN), which could be enough to perform high-throughput plant phenotype.

Performance evaluation of the proposed low-power embedded system was done. First, LSTM was trained on the desktop, then this pre-trained model was deployed on the developed low-power system. To obtain statistics on the average prediction time and power consumption, the system was run iterative. Power consumption and prediction time were assessed for each iteration. Each iteration means that the system turns on, then receive the picture of plants, after that process it and reconstructs projected leaves area, sends this information to the LSTM NN and finally, the pre-trained NN make predictions. Overall this process takes 30 seconds in average, 3 seconds of which needs to make predictions. As obtaining images of plants and making predictions once every 30 minutes is enough, consuming 30 seconds for the whole prediction cycle is reasonable. The power bank with a capacity 2550 mAh was used during testing of the system. Mean power consumption was measured for each run of the loop. Using this battery, system can perform 8663 continuous predictions before the battery was fully discharged. The mean power

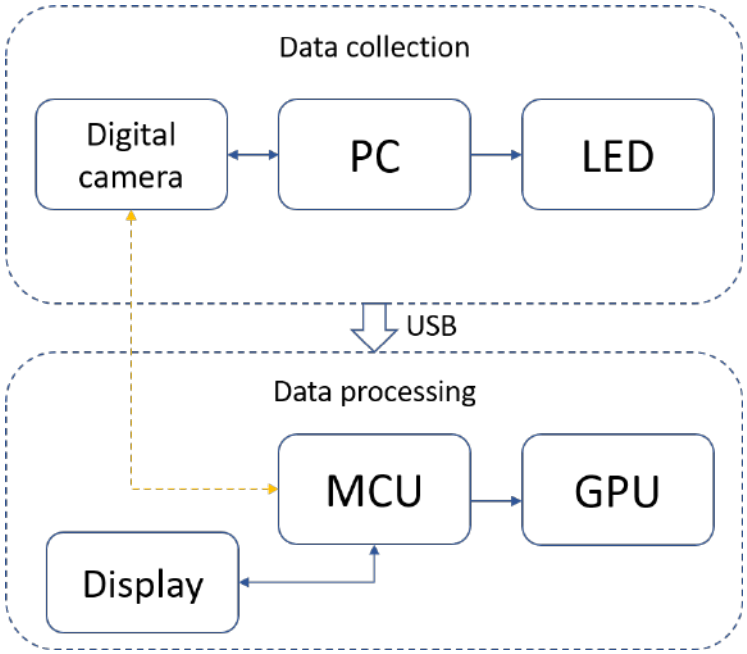


Figure 4-7: Block diagram of the proposed intelligent low-power sensing system for precision agriculture.

consumption among all iterations was 2.23 W. Since predictions are made once per 30 min, this means that 48 predictions per day should be done. Thus, the capacity of the tested battery is enough for system running up to six month.

Conclusions

In this section, the generic solution for predicting plant growth dynamics using data-driven approach was presented. The implementation of the predictive system (LSTM and CV) on the low-power embedded platform was also shown. Performance evaluation of the proposed solution has demonstrated that the developed AI architecture based on a recurrent neural network (RNN), in particular the Long-Short Term Memory network (LSTM) is characterized by reasonable precision for big horizon of prediction. The proposed solution can be used as an autonomous tool for continuous plant growth dynamics monitoring for up to 180 days. Together with an actuating capability, the proposed approach is promising for guarantying easy-to-deploy, generic, and robust optimization tool for the precision agriculture. The *Tomato Growth* dataset for training and testing procedures were collected by

the designed and assembled experimental setup coupled with the automatic imaging and processing system Shadrin et al. [2019c]. This dataset is publicly available for the research community. It can be used in a variety of computer vision tasks for the development and verification of new machine learning algorithms. For effectuating the growing process on the experimental setup a hydroponic system approach was used which ensures the optimal control of the plant nutrition and provides the opportunity to "drive" the growth system in a desirable way.

4.2 Computer vision in industrial scale experiments for growth dynamics assessment

The main idea of this experiment is to perform the industrial deployment of the AI approaches discussed and tested before, in the small scale laboratory artificial systems and to show their practical usefulness. This industrial experiment was conducted in a Michurinsk greenhouse where set of sensor nodes including digital cameras were deployed for collecting data describing plant growth dynamics at the initial stage. Digital cameras were used for collecting image datasets for performing semantic segmentation of plants and calculating projected leaves area. Cucumbers (*Cucumis sativus* L.) were used as a tested plant species. One of the features of the experiment is that it allows to build a comprehensive database for obtaining statistically reliable results. Also, direct measurements of biomass were performed sequentially on a specially developed grid to find correlations between biomass and projected leaves area and give us the possibility to perform in-situ biomass assessment and prediction using the methodology, described in the Section 3.5. The cucumber (*Cucumis sativus* L.) is one of the most produced crops in a greenhouse. Over the last years, several dynamic or simulations models have been proposed to predict the cucumber growth in greenhouse conditions Sun et al. [2012], Ramírez-Pérez et al. [2018]. Such models include a huge amount of different heterogeneous environmental conditions parameters, complicated mathematical models that describe crop growth and need to be tuned and adapted for each plant hybrid and growing system. Also, precise measurements of some of these parameters are pos-

sible only manually which is time consuming and ineffective. This means, that the efficiency of simulations can be low in cases when the farmers could not monitor all parameters as a routine. The incorporating IoT and computer vision systems can help to solve these issues, because they provide possibilities to monitor the plant's phenology changes in real time, and produce highly satisfactory forecasting of biomass (or other target parameters). The other important outcome of this research is that the proposed methodology is possible to use for fundamental research that aims at finding plant characteristics, dependencies and assessment the plant response to changes of the environmental parameters with high time resolution. This in turn opens wide possibilities for investigation of the hidden dynamics that was impossible to observe before, using standard techniques.

Methodology

This research aimed at the development of the intelligent computer-vision based system that provides the robust and accurate plant biomass growth dynamics prediction. As sensors, digital cameras generating images for further analysis and environmental sensors for monitoring specific anomalies were used. The methodology of the conducted research is the following:

- To deploy and tune set of sensors and digital cameras. The sensors and cameras were installed in the industrial greenhouse to collect the data describing the plant growth dynamics (images) and the environmental parameters.
- To set up one month experiment on cucumbers growth and collect relevant data from sensors and cameras as well as carry out the biomass measurements.
- To train and evaluate FCNN for performing the segmentation tasks and projected leaves area calculation. To calculate per-plant projected leaves area using sequences of the obtained images.
- To derive dependency between the projected leaves area and biomass.
- To estimate the parameters of projected leaves area growth model and perform predictions.

- To reconstruct and predict the biomass based on the assessed and predicted values of predicted leaves area.

Deployment

Plants and growing system. The cultivation of cucumber seedlings was carried out according to the low-volume hydroponic technology based on growing plants it rock wool substrate. Sowing for each seed was carried out in rock wool blocks 10×10×6.5 cm (*Grodan delta*). All rock wool blocks were preliminary dunk in a specially prepared nutrient solution (see Table 4.2). Vermiculite was sprinkled on top of the seeds to avoid additional evaporation. Totally, 496 plants were sowed and evenly distributed on the floating table in rock wool blocks for the experiment.

Table 4.2: The composition of the nutrient solution for initial saturation of cubes by fertilizer and further watering of plants.

Chemical component	NH^{4+}	K^{+}	Ca^{2+}	Mg^{2+}	$NO_3{}^{-}$	$S0_4{}^{2-}$	$H_2PO_4{}^{4-}$
$mmol/l$	1.25	6.75	4.5	3.0	16.75	2.5	1.25
Chemical component	Fe	Mn	Zn	B	Cu	Mo	
$mmol/l$	20.0	10.0	5.0	30.0	0.75	0.5	

The rock wool blocks were placed on one table, 8 m long, 1.82 m wide according to the experimental scheme (see Fig. 4-9). The rock wool blocks were saturated with a nutrient solution. The parameters of the fertilizer solution that was used for initial blocks saturation are electrical conductivity (EC) that was 1.50 mS/cm and pH was in range 5.3-5.5. Further watering of plants was carried out by the partial flooding method. Necessary watering time was determined by the weight of the rock wool block. The weight of a fully saturated rock wool cube 10×10×6.5 cm is 650-660 g. Watering was carried out by adding 350-370 g of nutrient solution. As seedlings grow, they need more elements of mineral nutrition; therefore, EC of feeding solution was slightly increased during the experiment (see Table A.1). However, when the seedlings formed four real leaves and a good root system, according to the technology of cultivation, they should be transplanted on rock wool slabs, which have a capacity of 16l of nutrient solution, 4l per plant (4 plants per slab). In our experiment

seedlings continued to grow in a $0.65l$ cube. So further watering using the common technology was impossible and the last few waterings were made with distillate water (see Table A.1). The biomass measurements during the experiment are shown in Fig. 4-8, where each point contains from 11 to 82 measurements of biomass. So, 480 measurements were done in total.

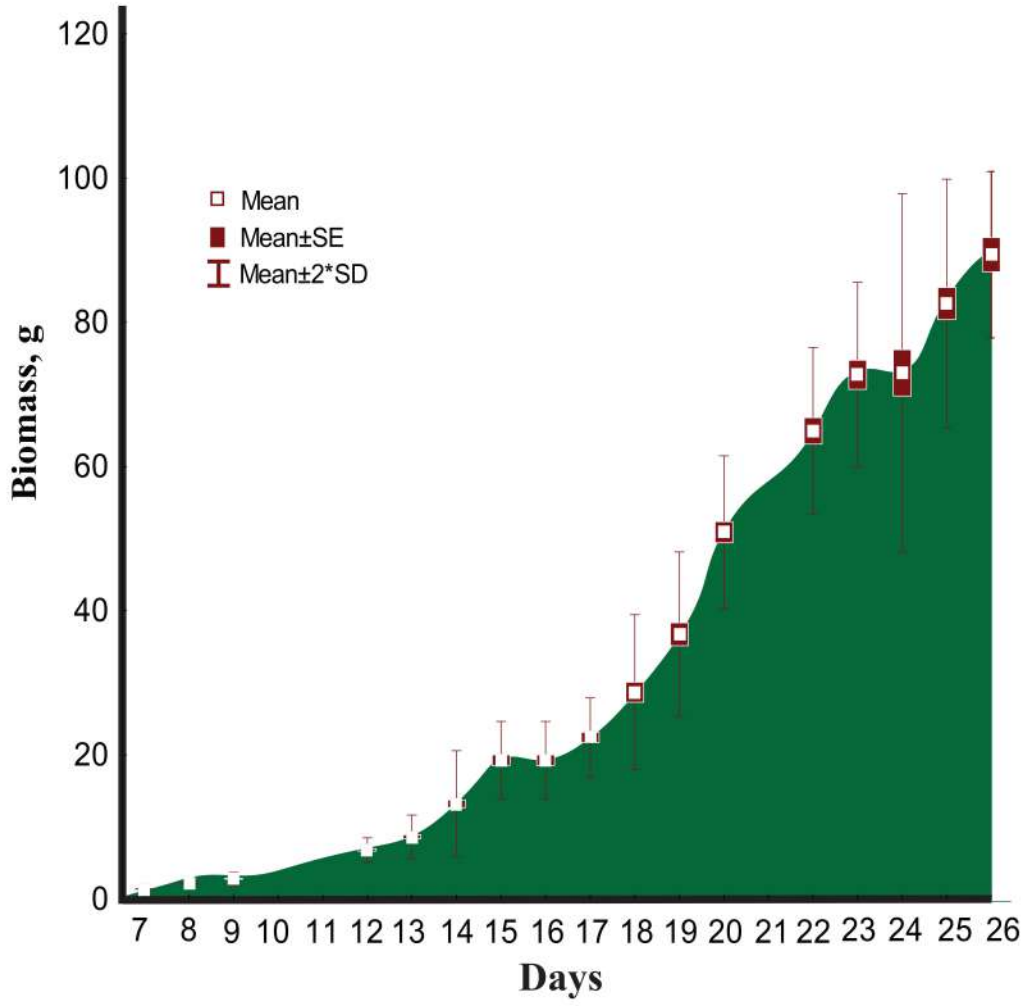


Figure 4-8: Biomass changing in time.

Conductivity measurements (EC), measured in mS/cm. Conductivity measurements were performed with a METTLER TOLEDO conductivity meter. Measurements of pH were carried out on a Sartorius PB-11 instrument. Measurements of changing EC and pH in the rock wool blocks parameters during the experiment are presented in the Table A.2. Solutions samples were taken out by a syringe from a rock wool block. Two samples from the middle of each one of the 3 zones were

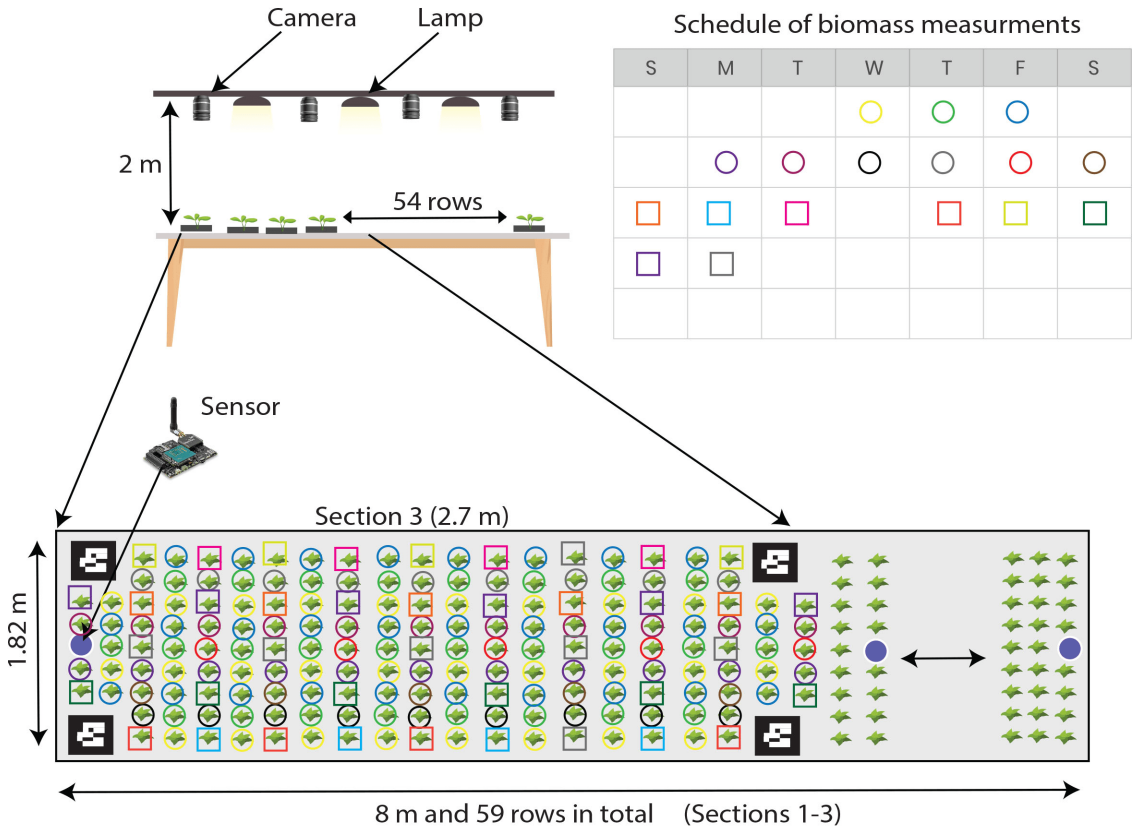


Figure 4-9: Industrial experiment scheme and biomass measurements schedule

taken. According to Table A.2 the following samples were taken: from zone I - samples 1 and 4, from zone II - samples 2 and 5 and from zone III - samples 3 and 6. These measurements ensure equal conditions for growing all plants. Such procedure reduces the deviations and makes the obtained dataset relevant and statistically correct.

Sensor system. The IoT system was deployed in the greenhouse with a total area of 5000 m^2 located in Michurinsk, the Tambov Region, Russia. The experiment was conducted in the special zone where plants were growing at the initial stage (720 m^2). In the deployment, WaspMote sensor nodes organized in a WSN were used. These devices are built around the low power ATmega microcontroller, communicate at 2.4 GHz and have mesh networking capability. The transmission power level was set up at level 3 (-0.77 dBm) and was chosen after the analysis of the Received Signal Strength Indicator (RSSI) only since the Link Quality Indicator (LQI) and Packet Delivery Rate (PDR) metrics are not available in the industrial-oriented

WaspMotes. As a power source, it was used a battery pack containing three 3.7 V Li-ion polymer cells connected in parallel with the output of 6600 mAh ensuring long-term operation for the devices. The sensor nodes are equipped with the following sensors: temperature, PAR, humidity and CO_2 concentration. Several sensor nodes per tray were deployed: one at the beginning and the second at the end of each tray (8 m x 1.82 m) with plants. The distance between the two neighbour sensor nodes is 3m. The goal of WSN nodes is to take measurements at the root level and send data every 5-10 minutes (depending on type of sensor) to a gateway. There are extra WSN nodes in each zone for measuring the ambient conditions of the greenhouse. The nodes collect the data and transmit them to the Libelium gateway which stores the data at the local Data Storage and sends them to the cloud for planning and modeling purposes.

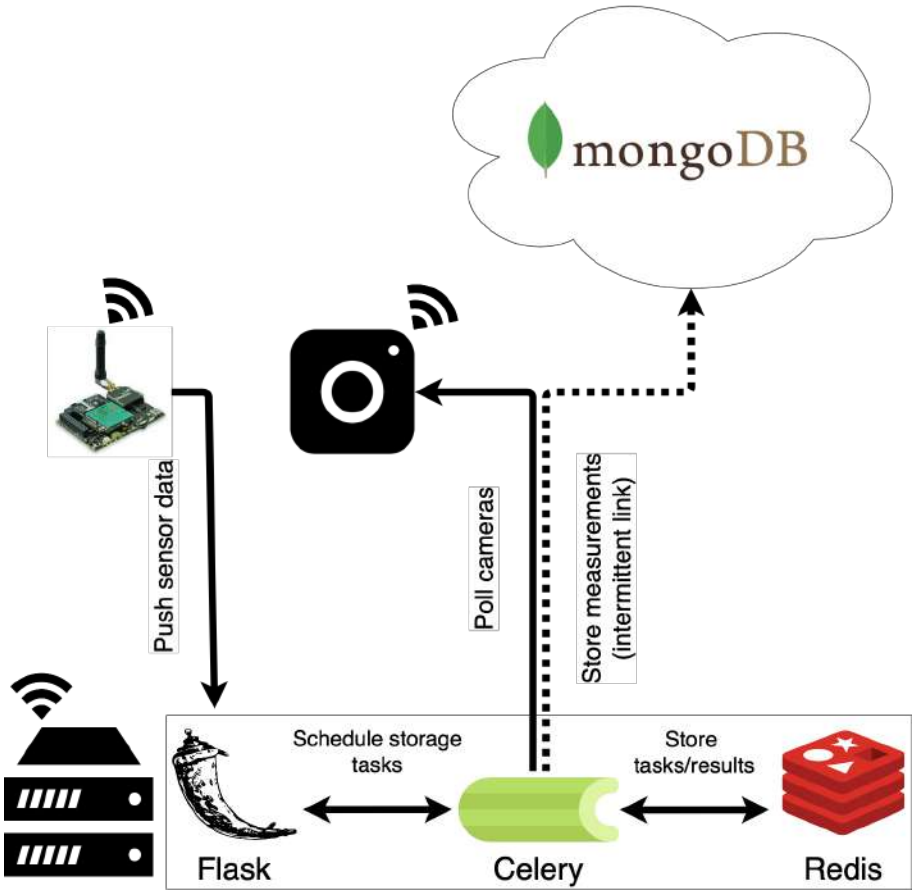


Figure 4-10: Sensor data collection and storage.

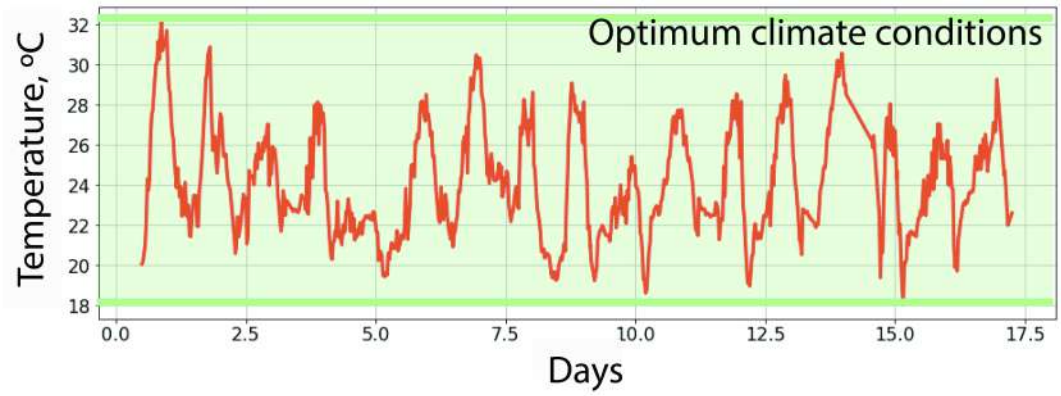


Figure 4-11: Temperature measurements.

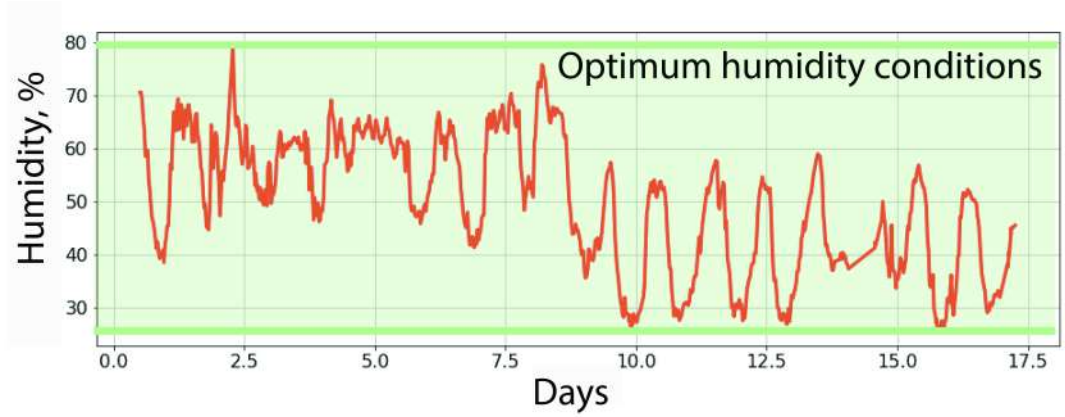


Figure 4-12: Humidity measurements.

Software and Data Storage. Schematic view of the data collection system is shown in Fig. 4-10. It consists of three main components: (i) A Flask-based HTTP server implemented in Python programming language; (ii) a distributed task queue Celery and in-memory Redis database as a message broker with persistence enabled; and (iii) a general-purpose schema-less Database Management System (DBMS) MongoDB allowing for the storage of unstructured data along with the arbitrary binary objects using GridFS. The HTTP API server allows for the collection of sensor data using push strategy with the sensor nodes sending measurements every 30 minutes. The MongoDB database was used to store the sensor measurements and camera images. It is hosted using a DBaaS service which was sometimes inaccessible due to the intermittent Internet connectivity of the deployment site. Therefore, it is crucial to persistently store the received data locally using Celery queue with Redis as a broker and task storage. When the Internet connection is restored the

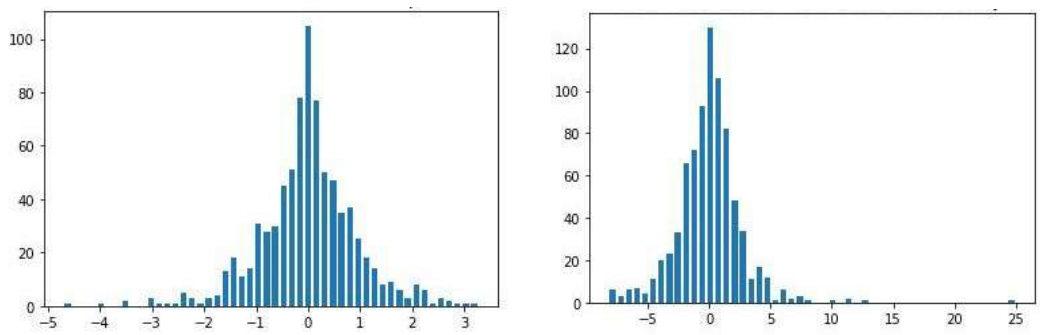


Figure 4-13: (a) Distribution of the first derivatives for temperature (b) Distribution of the first derivatives for humidity

data were sent to MongoDB. Additionally, Celery allows one to periodically (every 30 minutes) receive the image data from the cameras using a poll strategy. All of the software components were built using Docker containerization system, that is easy-to-deploy.

Some examples of obtained measurements of temperature and humidity corresponding to the experimental results described below are shown in Fig. 4-11 and Fig. 4-12. They ensure the permissible values of these environmental parameters during the experiment. For maintaining sustainable growth of the plants it is important to monitor not only the absolute values of the environmental conditions, but also the rate of changing (first derivatives). Rapid changes of the environmental parameters, even being in the optimal boundaries, may affect the growth dynamics. They can lead to plant development in a wrong way. Tracking these changes is only possible using distributed sensors that collect measurements with high time resolution. The analysis of the derivatives was performed using the obtained measurements. The results are shown in Fig. 4-13a and Fig. 4-13b. These results show that the environmental parameters changed smoothly what lead to a normal plant growth dynamics.

Image data collection and annotation. According to Fig. 4-9, 4 digital cameras with the resolution 1920x1080 were mounted 2 meters above the floating table. These cameras took 2 images sequentially every 30 minutes for 31 days. 2494 raw images were taken from each camera; 9976 top-down images were taken in total. After the data cleaning procedure and choosing the images only for the time in-

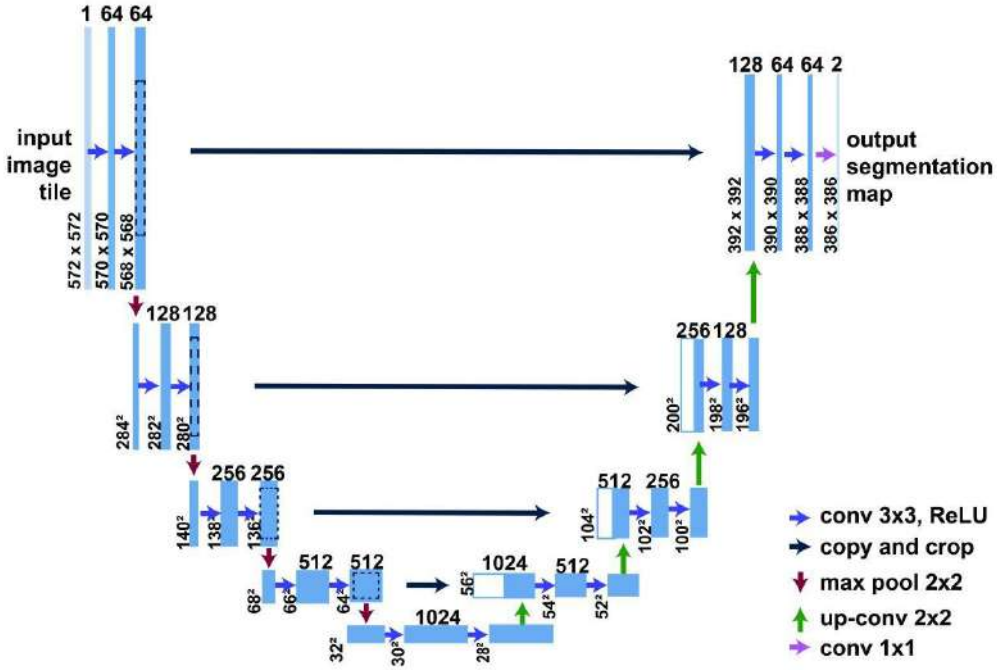


Figure 4-14: U-net architecture, Source: [Ronneberger et al., 2015].

terval that represents the active growth stage interval, 4 sequences of 975 images representing 25 days of observation for each camera were kept. These data were used for further investigation and assessment of growth dynamics. All images were flattened using the calibration images to avoid distortions. Totally, 248 images were annotated. A selection of 62 images out of 975 from each of 4 cameras for annotation purposes was done in the following way: from each day of observation 3 images at 9:00, 15:00 and 21:00 were kept. The annotation procedure consisted of putting the segmentation masks and bounding boxes for each plant in the image. Overall, 45389 instances for 248 images were obtained after the annotation procedure.

Growth assessment and modeling

FCNNs for leaves segmentation. The set of FCNNs, e.g. U-Net [Ronneberger et al., 2015], FCN8s, FCN16s [Long et al., 2015], was trained within the PyTorch framework and validated using the labeled dataset. The architecture of the U-Net is presented in the Fig. 4-14. U-Net consists of a contracting path that captures context and an expanding path that enables precise localization. It is effective for segmentation of the small datasets with excessive data augmentation and also it is

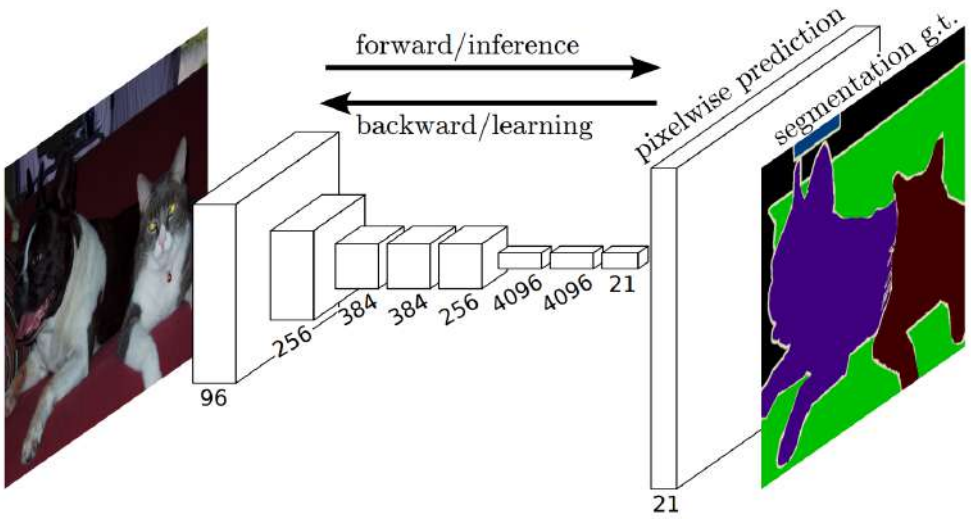


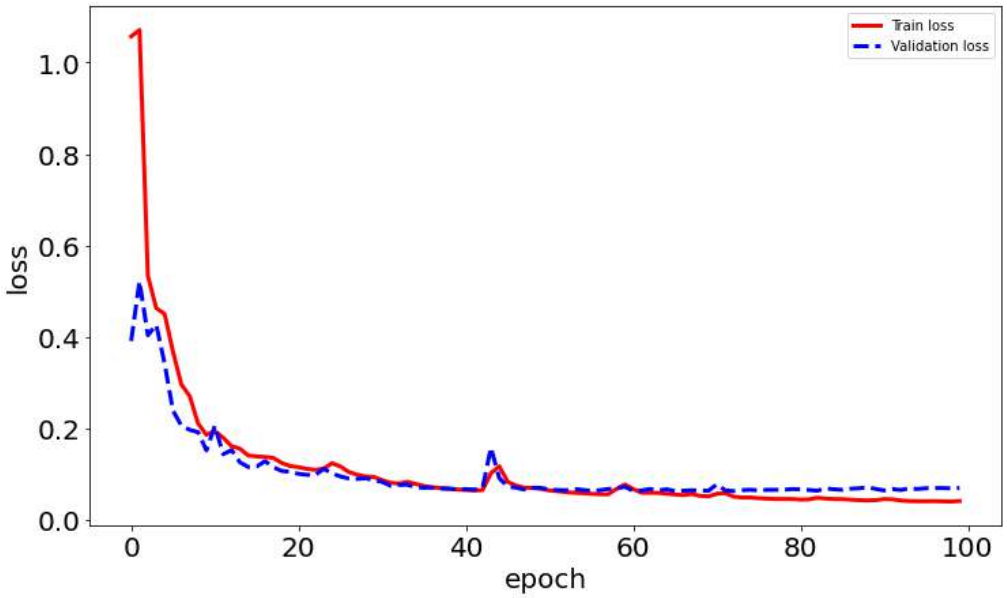
Figure 4-15: FCN architecture, Source: [Long et al., 2015].

effective for border segmentation which is important for plant segmentation. The contracting path consists of 3x3 unpadded convolutions, each followed by ReLU, max-pooling 2x2 with stride 2. The expansive path contains the upsampling of the feature map, followed by 3x3 convolutions. After upsampling, the resulting feature map concatenates with the corresponding feature map from the contracting path. Then it is followed by two 3x3 convolutions each followed by ReLU. Finally, 1x1 convolution is used for each one of the 64 components. After that, the pixelwise softmax over the whole feature map was calculated. The architecture of the typical FCN is presented on the Fig. 4-15. FCN's convolutional layers that include pooling and ReLU activations are followed by deconvolutional layers (or backwards convolutions) to upsample the intermediate tensors so that they match the width and height of the original input image.

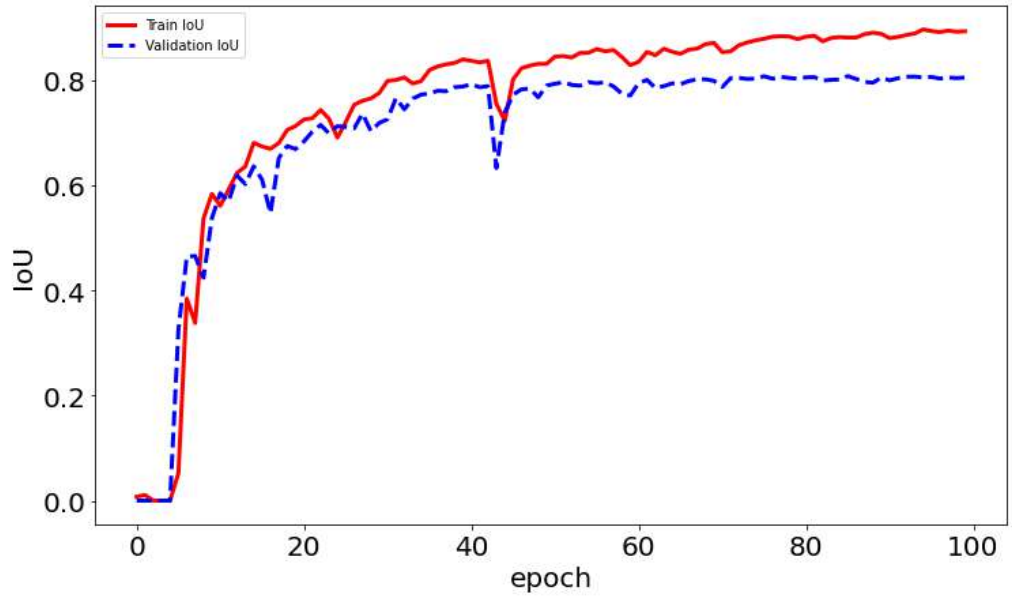
Out of 62 images from each camera 50 were used for training and 12 for validation. The other 913 images from each camera were kept as test data. To assess the quality of the trained model the average IoU between the predicted masks and the ground truth masks for validation data was evaluated using the following formula:

$$\frac{\sum_{\forall\{w_g,w_p\}} \mathrm{IoU}(w_g, w_p)}{\#\{w_p\}}, \tag{4.1}$$

where



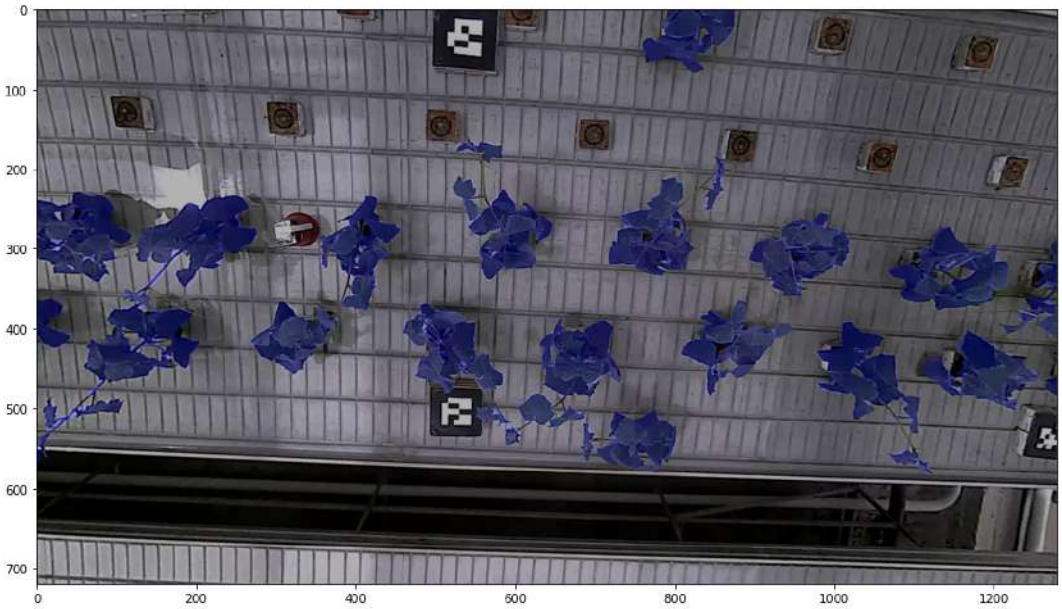
(a)



(b)

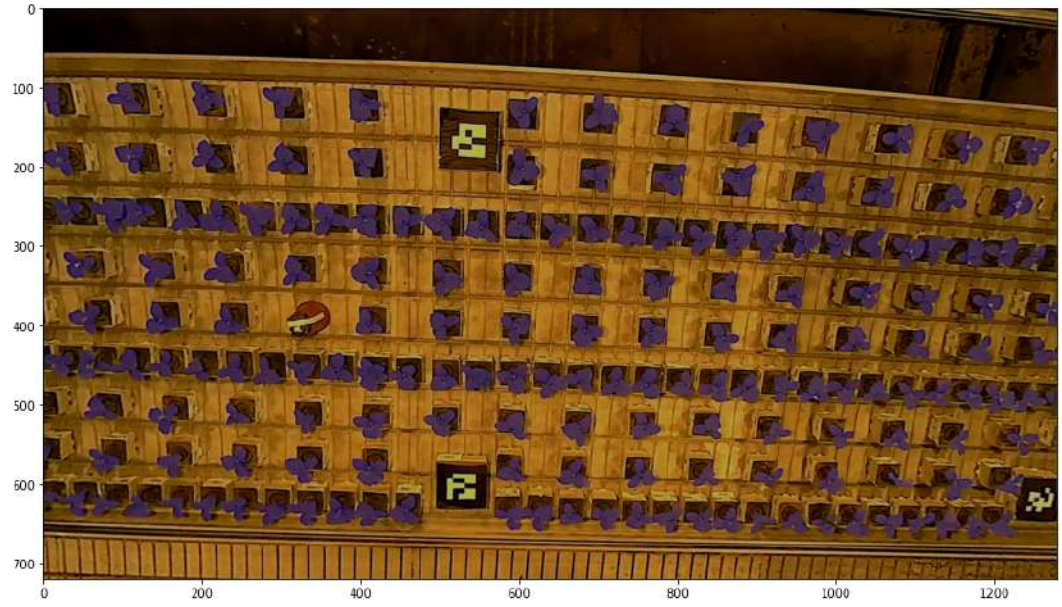
Figure 4-16: (a) Loss dynamics and (b) IoU changing during training and validation procedure.

(a)

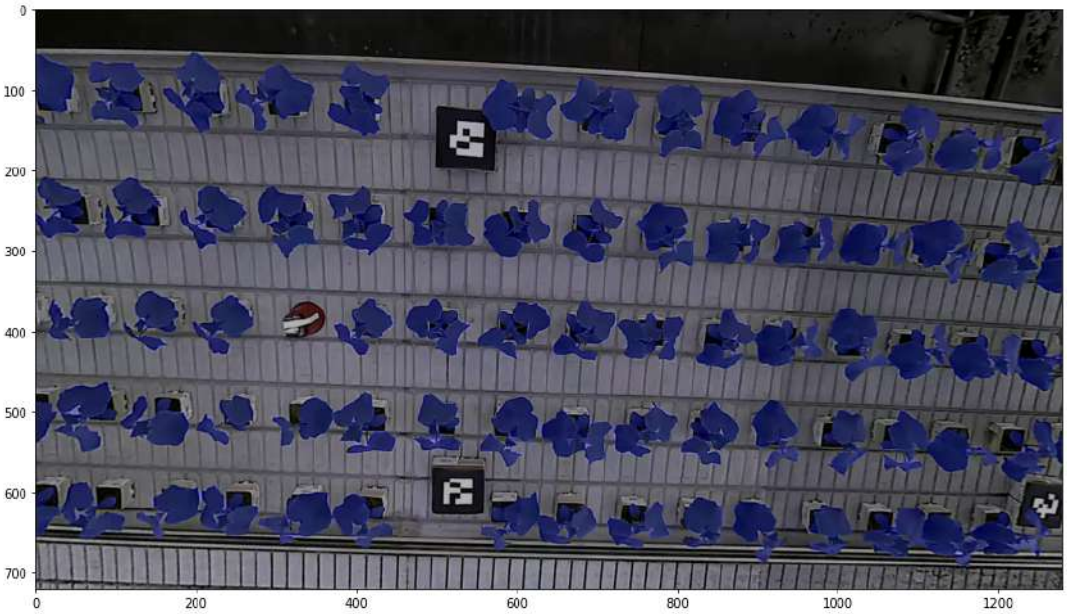


(b)

Figure 4-17: Predicted masks on the validation dataset



(a)



(b)

Figure 4-18: Predicted masks on the test dataset

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}, \tag{4.2}$$

and $\forall\{w_g, w_p\}$ are all the possible pairs of ground truth and predicted masks, while $\#\{w_p\}$ is the number of predicted masks. The average IoU on the validation set using a FCN8s semantic segmentation neural network achieved value of 0.81. The training parameters were selected as follows: batch size = 2, learning rate = 0.008, class weight = 0.5. Images were also resized to 1280x720. Figure 4-16a and Fig. 4-16b present train and validation losses and the corresponding values IoU for 100 epoch. Early stopping criteria was used to retrieve the best model during the process of learning. The examples of predicted masks on the validation dataset are shown in Fig. 4-17a, whereas Fig. 4-17b represents images for different stages of growth. The examples of predicted masks on the test dataset are shown in the Fig. 4-18a and Fig. 4-18b. As it can be noticed from these figures, the predicted masks are very accurate and visually in full correspondence with the actual plants.

Growth dynamics prediction. Using the sequence of selected 975 images from each camera, per-plant projected leaves area was calculated. As there are many plants in the images, also the table can move in horizontal direction, and there were direct biomass measurements of plants, different amount of plants appear in images. This means that the total segmented area should be divided by the actual amount of plants to obtain the averaged area of each plant. An example of the average projected leaves area calculated per plant on the image sequence obtained from one of the cameras is shown in Fig. 4-19. It should be noticed that there was a total power interruption for several days on the 18-th day of the active plant growth. In Fig. 4-19 the first 760 data points, representing the continuous growth are shown. The accuracy of the proposed FCNN used for segmentation is additionally proved by the fact that it was able to capture the diurnal fluctuations in relative leaves location (see Fig. 4-19) of the projection of the leaves area that is caused by biological reasons, specifically respiration. After the power interruption, the system switched on automatically and continued collecting images and data from sensors (215 images). This experience showed the high relevance of the autonomous embedded systems

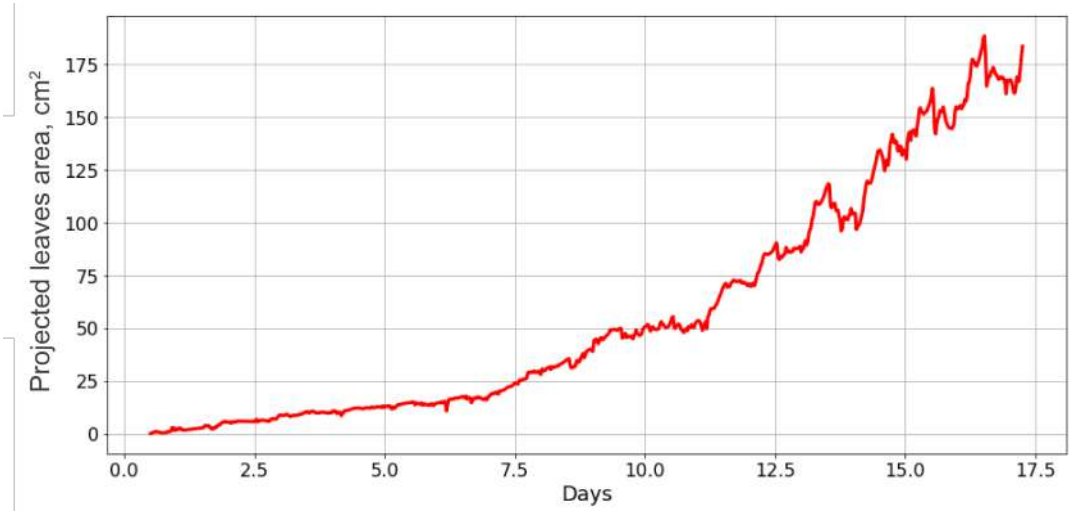


Figure 4-19: Reconstructed dynamics of the specific projected leaves area based on the calculations by using of FCNN

for greenhouses to overcome the problem. Nevertheless, the collected images and biomass measurements were sufficient to find dependency between projected leaves area and biomass (similar that was proposed in the Section 3.5). Figure 4-20 shows the approximated dependency between projected leaves area and biomass using Eq. 3.26.

To construct this dependency, it was used data points representing direct measurements of biomass and corresponding FCNN-calculated projected leaves area during the first 18 days (first 10 points of the biomass measurements form Fig. 4-8).

The derived dependency for cucumbers is the following (Eq. (4.3)):

$$m = 0.00755 * S^{1.57}, \tag{4.3}$$

It should be noticed that the reconstructed coefficient $\gamma = 1.57$ that represents non-linearity for the dependency between biomass and projected leaves area for cucumbers is close to the γ, reconstructed for different individual tomatoes ($\gamma = 1.72; 1.69; 1.67$, see Section 3.5) . This result also proves the high accuracy of the measurements performed in the by 3D scanner (see Section 3.5). Using the obtained dependency it is possible to assess and predict biomass using the projected leaves area predicted by the Verhulst model (see Eq. 3.2).

The non-linear least square method was used for estimation of the parameters

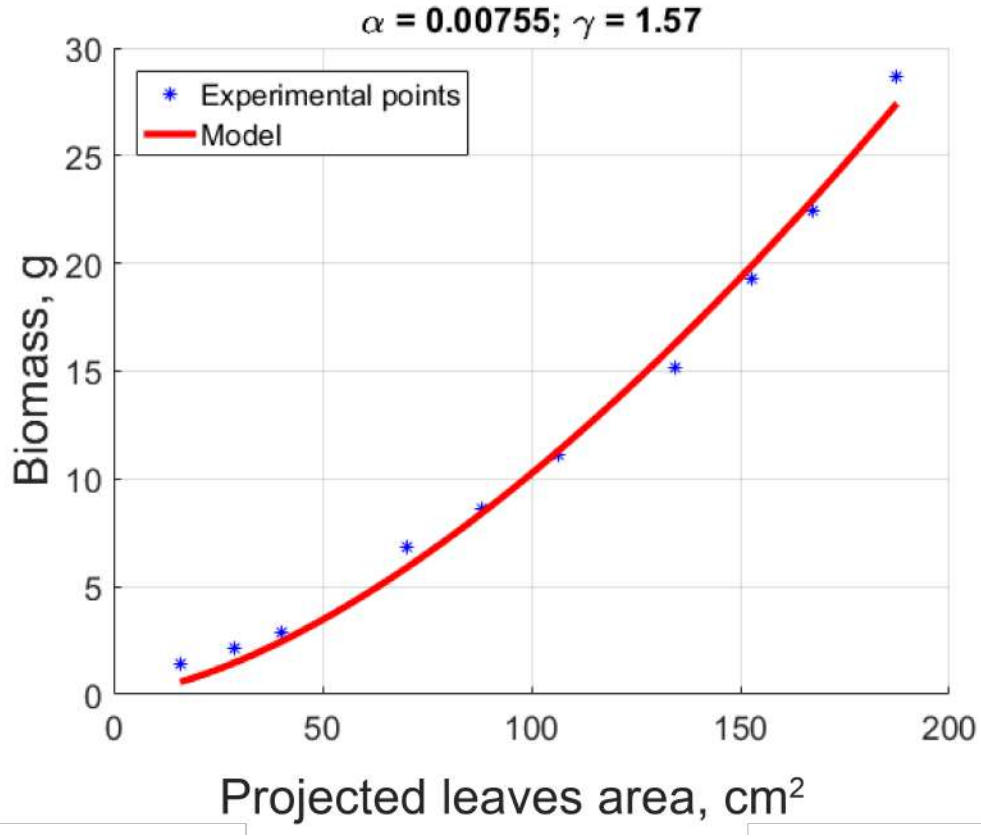


Figure 4-20: Dependency between averaged biomass and specific projected leaves area

in the growth model (Eq. (3.2)) based on the projected leaves area calculations obtained by the FCNN for the first 18 days. The results of the estimation are $\mu = 0.23\ 1/30min$, $S_{max} = 700\ cm^2$ and $S_0 = 5.07\ cm^2$. The relative error of the approximation of the data by the model is 5.5%. The obtained coefficients were used for extrapolation of the projected leaves area growth curve. The result of the fitting to the experimental data and prediction of the projected leaves area for 12 days ahead is shown in Fig. 4-21.

Using these fitted and extrapolated values for the projected leaves area and the derived dependency between projected leaves area and biomass, the biomass for one month including the extrapolation interval (last 12 days) was calculated the predicted. It can be seen from the experiment that after the initial stage of cucumbers growth (3-4 first leaves), they start growing upwards, and the projections obtained from the top-down images don't change significantly (the changing of projected leaves area that can occur due to the overlapping effect starts to prevail). The

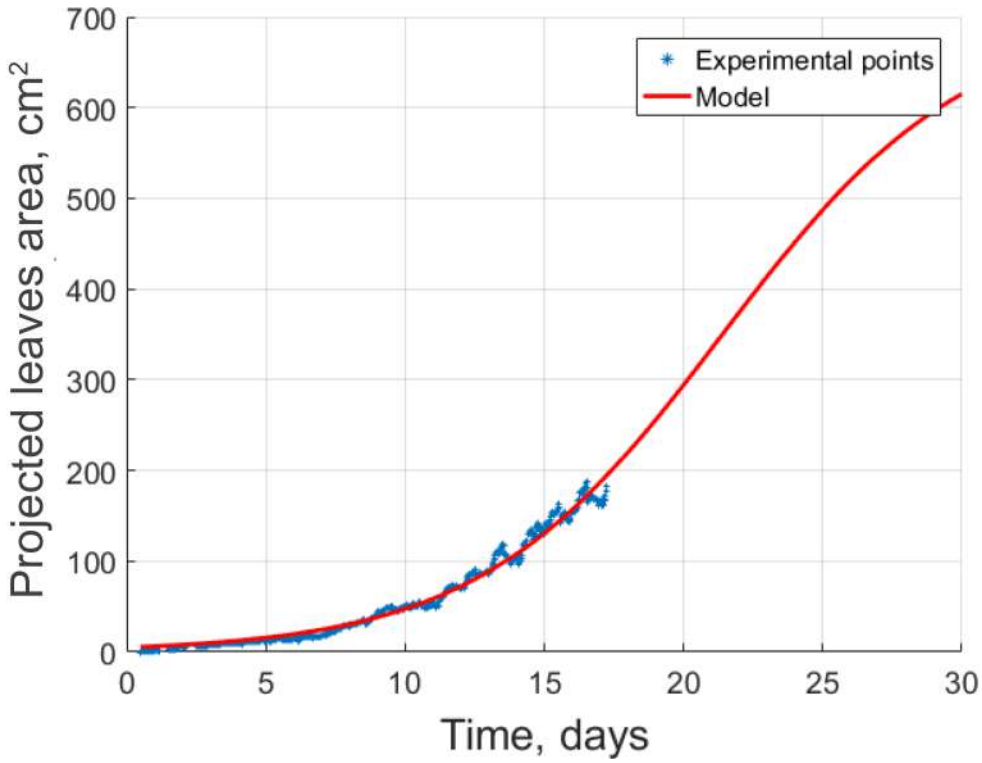


Figure 4-21: Result of the fitting to the experimental data and prediction of the projected leaves area for 12 days ahead based on the Verhulst model and calculations of the projected leaves area obtained by FCNN for the first 18 days of the experiment.

estimation and modeling of the biomass were done at the initial stage of growth. So, the biomass that is accumulated at the initial stages and estimated using the first 3-4 projected leaves area has its limitations. The result of biomass calculation is shown in Fig. 4-22, where the predicted biomass is presented along with the biomass measurements that were used for construction of the dependency. Also, Fig. 4-22 shows the measurements of the biomass that were taken during the last 12 days and that were not included in construction of the dependency. These last 7 points were used to validate of the prediction accuracy. The average relative error of the biomass prediction reached 10.7%.

Conclusions. In this section, the industrial deployment of the AI-enabled sensing system for robust and accurate plant growth dynamics prediction was presented and tested. For the purpose of dataset collection that includes image data, environmental conditions, and biomass measurements, a one-month experiment on cucumbers

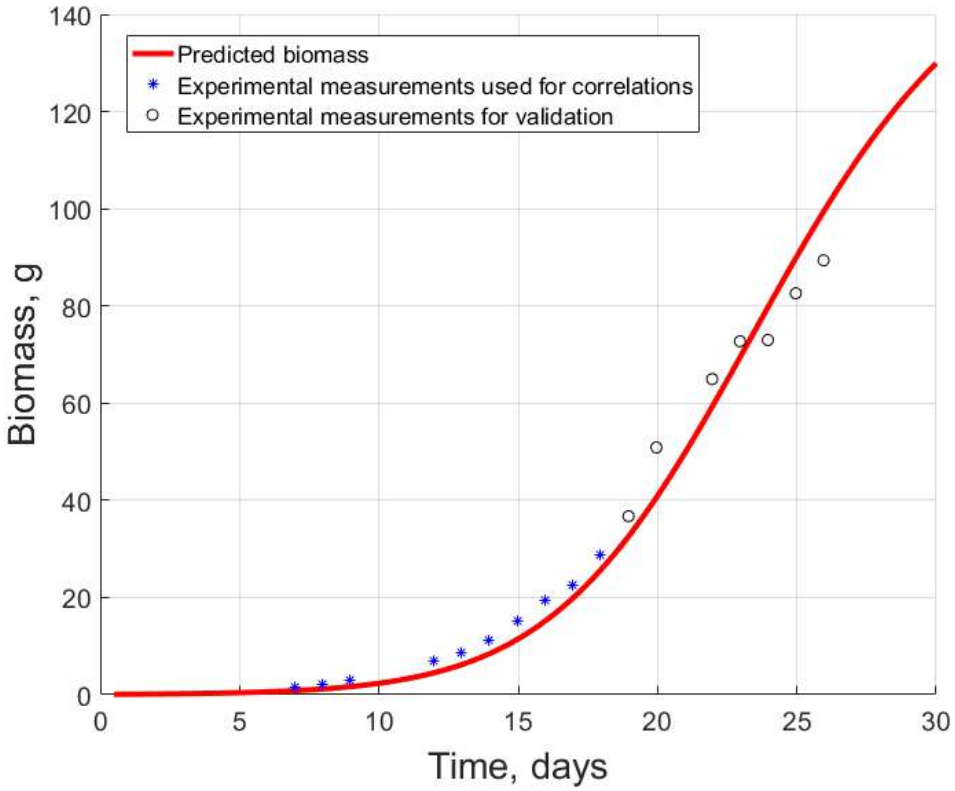


Figure 4-22: Assessment of the biomass based on the correlations and predictions of the projected leaves area vs. experimental measurements.

growth in the greenhouse was conducted. Specifically, a dataset with sequences of 9976 top-down images from 4 cameras, 480 direct measurements of biomass for 17 days period, and environmental data from sensors was created. First, the obtained dataset was labeled and the FCNNs were trained to perform automatic segmentation of cucumbers. The accuracy of the trained FCNN on the validation set was 81% of the IoU. Second, the trained FCNNs were applied to the sequences of images, thus, reconstructing average per-plant projected leaves area and growth dynamics. Then, the correspondence between the area of leaves (projected) and biomass using the direct measurements of the biomass was established. Finally, the dynamics of the biomass was predicted based on the predictions of the projected leaves area within 10% average error for the 12 days prediction horizon. Obviously, the advantage of the proposed methodology is that it allows monitoring huge amount of plants and quantify the actual growth dynamics in real-time using simple cameras. This is impossible to do manually. This, in turn, allows the automatization of the whole

monitoring process and obtain relevant information on the current status of plant growth. One of the limitations of this method is that huge labeled dataset is needed for training of FCNN that will perform segmentation task. Also, FCNNs are computationally complex, however, it was shown that it is possible to implement them using current state-of-the-art technologies into the embedded systems. The other limitation is that in spite that segmentation algorithm works well at all stages of plants growth, the accurate leaves area (projection) assessment by segmentation of plants is only possible on the initial stage of growth when leaves do not overlap. The other benefit of the proposed system is that it provides useful data for the prediction algorithms. So, prediction of plants' growth dynamics becomes possible. The other issue discussed in the section above is biomass prediction. The advantage of the proposed method that it allows to assess and predict biomass using simple 2D cameras. However. the limitation of this method is that a dataset that contains biomass measurements of a particular plant type is needed for calculating the dependency between projected leaves area and biomass and it works only at the initial stage of growth.

Overall, the highly effective and reliable data-driven based pipeline for the plant growth dynamics assessment and prediction was proposed and evaluated. The actual deployment showed the high industrial potential of the implementation of the proposed data-driven approaches for plant growth dynamics assessment and prediction.

4.3 Data-driven computer vision based system for monitoring of seeds germination process

Seeds germination is one of the most important processes in the whole growth cycle defining the future plant development. In this section it will be presented the AI system for monitoring the seeds germination process. The proposed system is a sensor node characterized by sensing, processing and communication capabilities with a special focus on data processing. For this reason it was collected a dataset in an industrial chamber and designed a CNN for germination recognition. First, all

the monitoring system was evaluated on a desktop. Then similarly to the Section 4.1, the developed monitoring system was deployed on the low power embedded system with an external GPU. It was achieved 97% accuracy of seeds recognition and 83% of average IoU score. At the same time, the proposed solution takes advantage of scalability, small size and the ability to be powered by batteries, therefore, ensuring autonomous intelligent operation.

Methodology

The methodology includes a number of important steps needed to monitor the seeds germination process:

- Set up of a continuous experiment for seeds image data collection. Within this step the images of seeds during germination process should be recorded certain timestamps.
- Data annotation (bounding boxes and sprouts) is applied to provide the machine learning (convolutional neural network (CNN)) algorithm with the examples of positive and negative samples for training purposes.
- Training of the CNN algorithm based on the collected data.
- Assembling the embedded platform equipped with external GPU and implementing the designed CNN on it.
- Detection of the seeds germination using CNN followed by the detection of germinated seeds using computer vision techniques.
- Testing of the system.

Data Collection

The experiment was set up on radish seeds which were germinated under controlled conditions. Images of seeds were taken during the continuous experiment. Seeds are initially ordered and located in isolated plastic containers having a size of 18x14x1 *cm* (LxWxH). In the experiment 18 containers were used, with 16 seeds in each.

For germination of seeds two sections with the fiber substrate were wet by 6 *ml* of distillate water. The containers were located vertically into two climate chambers (*Binder* climate chambers) - 9 containers per each climate chamber. Both of the chambers ensure 80% humidity. The temperature was set differently: 21°C for the first chamber and 24°C for the second one. The use of climate chambers for the germination of specific seeds was driven by the only purpose to create the particular environmental conditions.

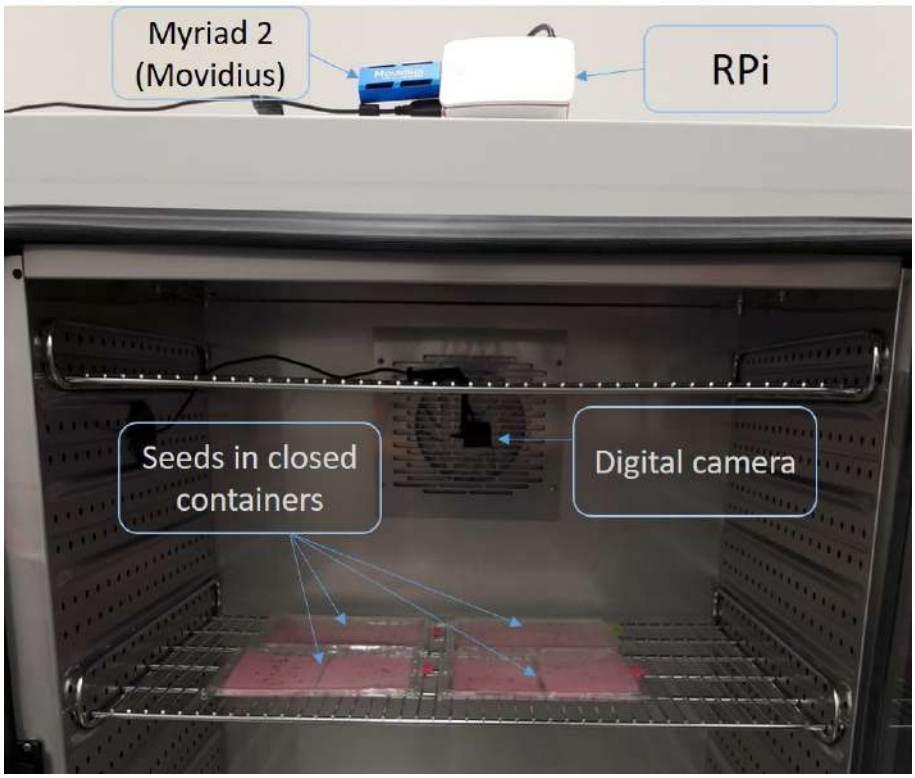
Every 2-4 hours during the experiment and depending on the germination stage all the containers were moved out of the climate chambers to take images of the containers (see Fig. 4-23a). Images of seeds were collected using the 1920×1080 resolution digital camera *Logitech 920* mounted in the stand. The above procedure was repeated 12 times during the experiment. Every iteration 18 images were taken. As an outcome of the experiment an annotated dataset which contains the time ordered sequence of the labelled high-resolution images showing the seeds germination process was obtained [Shadrin, 2018]. Some examples of such images are shown in Fig. 4-24a and Fig. 4-24b. The obtained dataset was used for training the neural network for the identification of germinated seeds and estimation of germination dynamics. In Fig. 4-25a and Fig. 4-25b the dynamics of each of the seed germination processes is shown. It demonstrates that at 24°C germination process in the climate chamber is faster than at 21°C. For testing the developed ML algorithm an embedded system with the external GPU was assembled for conducting high performance calculations, as it is shown in Fig. 4-23b.

Implementation, machine learning and computer vision

The state-of-the-art approach for object recognition based on CNNs. These neural networks (NNs) are effective for extracting local spatial features of an image. For creating, training and validation the CNN, the PyTorch framework was used [Paszke et al., 2017]. The structure of the proposed custom CNN is presented in Table 4.3.

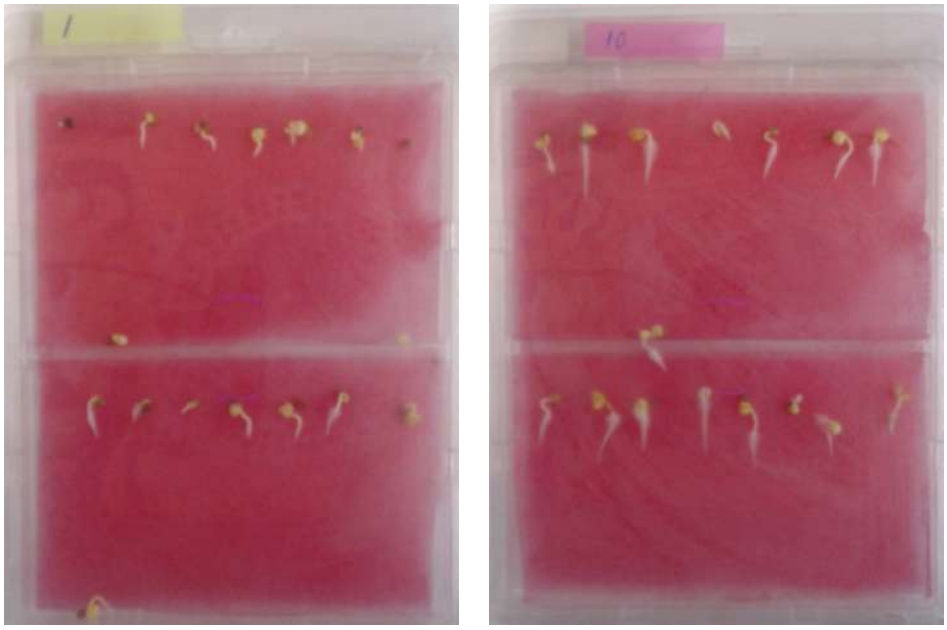
The CNN proposed in Table 4.3 includes 2 convolutional blocks, 2 linear blocks and a sigmoid block. The convolutional block consists of the convolutional, batch normalization, max pooling, ReLU activation and dropout layers. The linear block

(a)



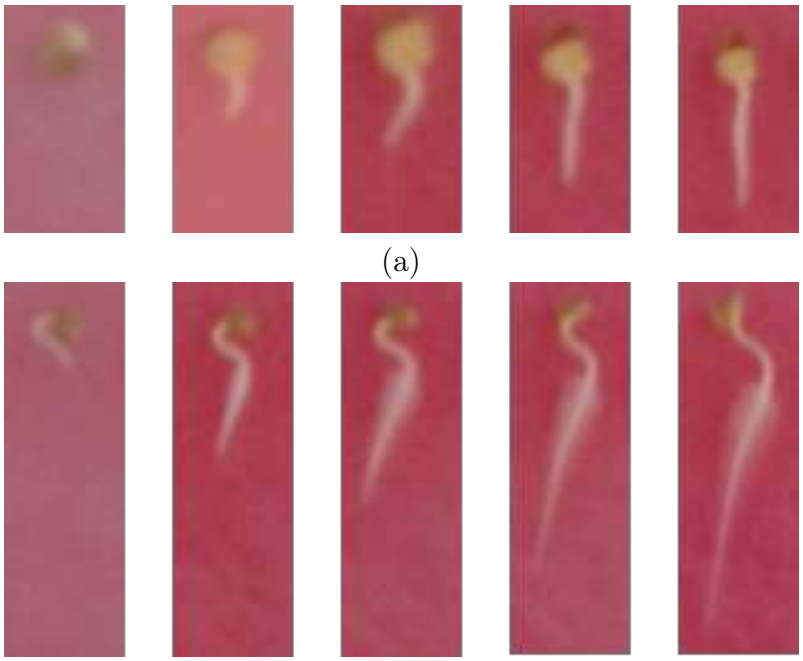
(b)

Figure 4-23: Climate chamber (*Binder*) (a) Process of germination for obtaining dataset; (b) Embedded system assembly for testing of machine learning (CNN) algorithm



(a) (b)

Figure 4-24: Example of images of the containers with seeds taken during the experiment. Seeds germinated in different conditions: (a) 21°C; (b) 24°C.



(a)

(b)

Figure 4-25: Images of seeds germination process taken during the experiment. All the images have time reference and were taken with a 3-hour time period. The top and bottom images were taken at the same time, but illustrate the seeds germinated in different conditions: (a) 21°C; (b) 24°C.

Table 4.3: Convolutional Neural Network Architecture.

#	**Layer**	**Dimension**			**Kernel**	**Stride**
		Width	*Height*	*Depth*		
0	Input	90	90	3	-	-
1	Convolution	40	40	48	11	2
2	Normalization	40	40	48	-	-
3	Pooling	20	20	48	2	2
4	ReLU	20	20	48	-	-
5	Dropout(p=0.1)	20	20	48	-	-
6	Convolution	16	16	96	5	1
7	Normalization	16	16	96	5	1
8	Pooling	8	8	96	2	2
9	ReLU	8	8	96	-	-
10	Dropout(p=0.1)	8	8	96	-	-
11	Fully Connected	1	1	100	-	-
12	Normalization	1	1	100	-	-
13	ReLU	1	1	100	-	-
14	Dropout(p=0.1)	1	1	100	-	-
15	Fully Connected	1	1	100	-	-
16	Normalization	1	1	100	-	-
17	ReLU	1	1	100	-	-
18	Dropout(p=0.1)	1	1	100	-	-
19	Fully Connected	1	1	1	-	-
20	Sigmoid	1	1	1	-	-

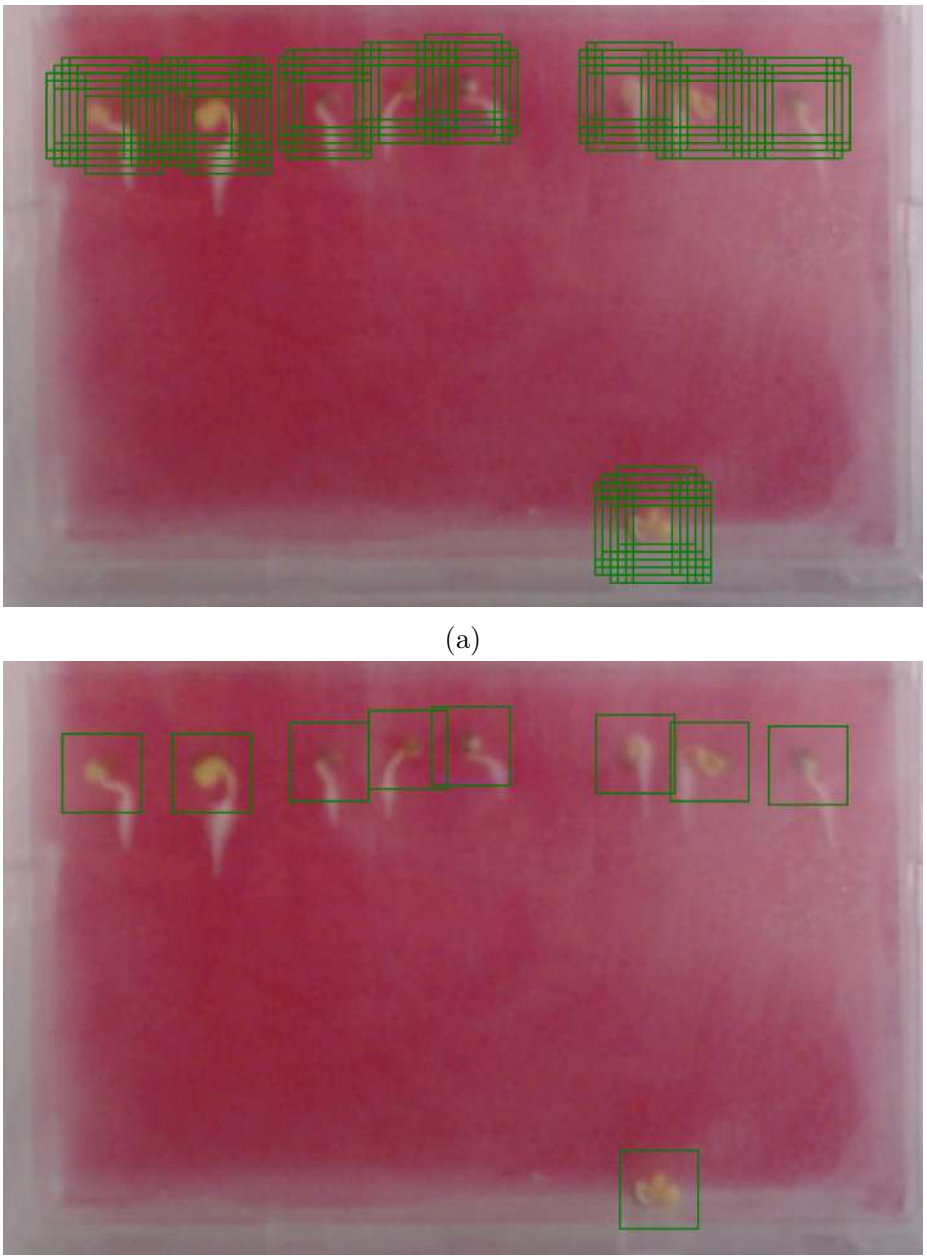
includes the fully connected, batch normalization, ReLU activation and dropout layers. The sigmoid block is composed of the fully connected and sigmoid activation layers. The convolutional layer takes the input frame and sums the frame values with weights. Thus, the local spatial features of the input are extracted. The batch normalization layer centers the input with the mean and scales it with the variance. Commonly, the batch normalization layer is used to make the convergence more stable. The max pooling layer takes the input frame and finds the frame maximum over the kernel with predefined size. It helps to extract the local peaks and compress the input size. The ReLU activation layer turns the negative input values into zeros. This speeds up the convergence process. The dropout layer turns the input values into zeros with the probability p. It makes network robust to overfitting and responsible for neural network generalization. The fully connected layer sums all the input values with weights for every value in it. It helps find the meaningful connections between the input values. The sigmoid activation function of x is $\sigma(x) = \frac{1}{1+e^{-x}}$. It is used for converting the values into probabilities. When, for example, the picture frames 90x90x3 ($pixels \times pixels \times channels$) is used as the input, the probability belonging to the seed class will be the output. It should be noted that, we tried to adapt the advanced deep CNN architectures (e.g. VGG16 [Kucer et al., 2018]) to solve this task. However, the increase in accuracy is insignificant and is about 1%. On the other hand, VGG16 has much more parameters compared to the proposed architecture and this leads to difficulties to running it on an embedded system. Thus, the proposed architecture is a trade-off between accuracy and complexity.

The proposed CNN architecture (as well as VGG16 etc.) was trained to perform the seeds recognition task ('exist or not exist') in the predefined window area (90×90). Using this trained CNN, the recognition and localization of the seeds in the picture was performed with the application of the sliding window technique [Noh et al., 2016]. For this purpose, 90×90 windows that overlap less than 90%: every next window was obtained from the previous one by a small shift in the horizontal or vertical direction. After cropping images from the picture, they were used as the input of the pre-trained CNN to recover the labels (existing or not existing seed). All the 90×90 pixels windows with positive labels ('seed exists') were combined and

non-maximum suppression technique was applied to them [Oro et al., 2016]. This technique allows to merge close images (windows) into one in order to avoid multiple references to the same place of the picture. The result is presented in Fig. 4-26. Finally, image covered by the windows where the seeds should exist was obtained the. Using windows obtained by CNNs then it should be detected in them whether the seeds have or have not germinated. This was performed with computer vision techniques that make possible to calculate the amount of white pixels referring to sprouts and compare pixels to the manually defined threshold.

Dataset labeling. In the experiment, data from 18 containers with seeds of 12 time periods (which leads to 216 raw pictures) were collected. For every picture, it was manually defined the ground truth boxes (windows) where seeds are located. 4 containers, #8-11, were referred to test dataset, while other containers, #1-7 and #12-18, were referred to the train dataset. From pictures of the containers in the train dataset 90×90 pixels images with seeds (ground truth boxes) and background were cropped. As a result, 2400 images of seeds and 3300 images of background were acquired. The train dataset was split into the train (80%) and validation (20%) parts. During the training and validation phases, random horizontal and vertical flips along with color jitter were applied to increase the model robustness. Color jittering grants the benefits of including the simulation of different illumination sources. During the validation phase the model predicted labels of images from the validation part. During the test phase a picture of the containers from the test dataset was taken. Then using this image it was obtained the 90×90 pixels windows with the window scrolling and for them labels were predicted. After that NMS was applied to leave only one predicted window per group and to estimate the average IoU between the ground truth windows and the predicted windows.

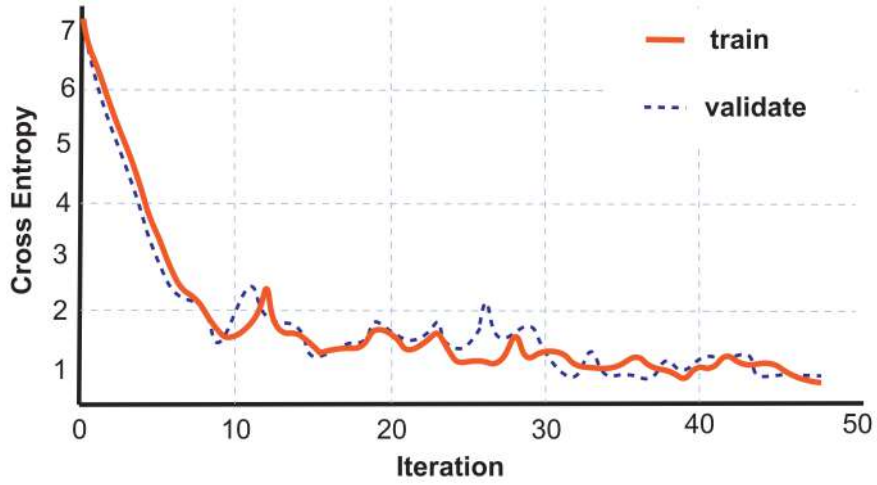
CNN Performance Evaluation. The CNN was trained for 50 epochs (iterations) with the cross-entropy loss for the train and validation datasets shown in Fig. 4-27a. For each epoch the random horizontal and vertical flips and color jittering were applied to ensure the data augmentation. The accuracy of trained CNN on the validation is more than 97%. (See 4-27b). It means that the CNN mismatch



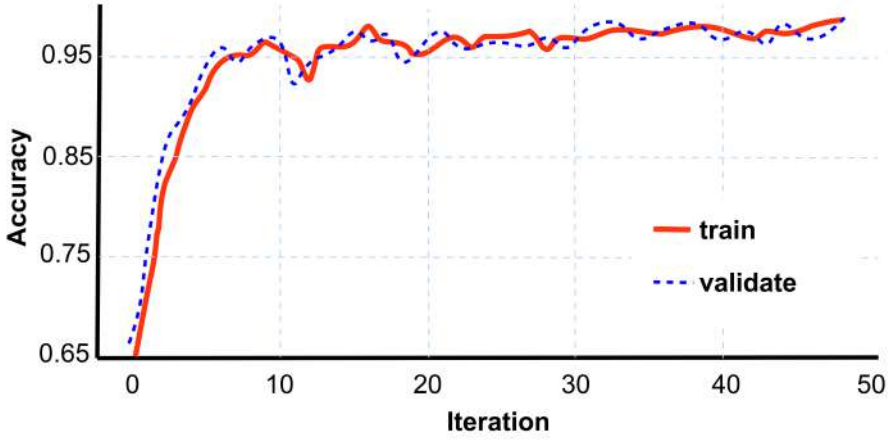
(a)

(b)

Figure 4-26: Non-maximum suppression application. In (a) seeds are covered with multiple windows, while in (b) one window per seed was used. This was obtained by grouping the windows and keeping one window per group.



(a)

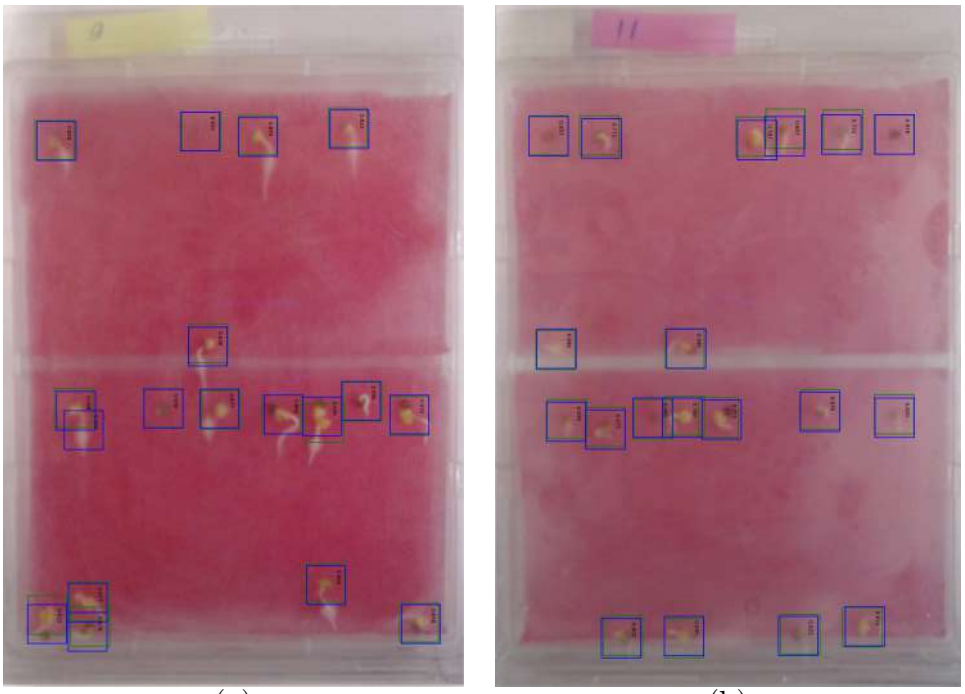


(b)

Figure 4-27: Cross entropy loss(a) and accuracy(b) on CNN training for 50 iterations(epochs).

3 windows with seeds or background out of 100. In the test phase the method was applied for containers #8-11. The results are shown in Fig. 4-28. To assess the quality of the developed model the average IoU between the predicted windows and the ground truth windows were evaluated using the formula (4.1). As a result, obtained of the average IoU was 0.83.

Germination Detection. A typical problem associated with the detection of seeds germination using images and by application of traditional computer vision algorithms is distinguishing the white pixels belonging to sprouts. These pixels may belong to recently germinated seeds or to other objects close to white pixels, e.g.



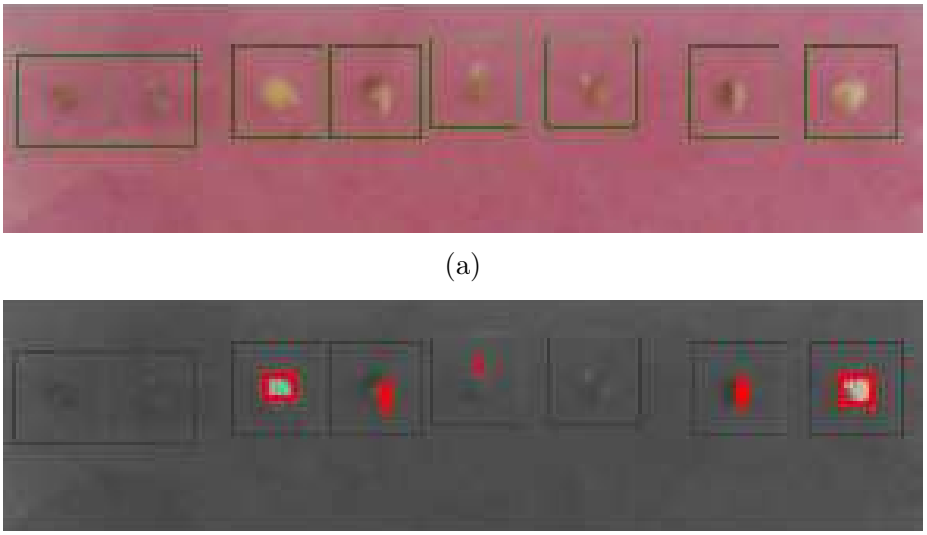
(a) (b)

Figure 4-28: Seed recognition in the 9^{th}(a) and 11^{th}(b) containers.

the drops resulted from humidity and appeared on the container. The key feature of the designed CNN is to propose the regions for detecting further germinations within the regions identified by it. Figure 4-28 shows that all the seeds germinated within a certain period of time are characterized by the reasonable quality of the seeds used. In most cases the germination rate is 80-90%.

Fig. 4-29a demonstrates the example where the seeds have been localized in a container using CNN. Then it was recognized which seeds out of the localized ones are germinated. For this reason it was developed the following algorithm based on the Pythons library *skimage*. The algorithm was then applied to each bounding boxes and works as follows:

- It converting the RGB image into a grey-scale one.
- Using the Otsu algorithm (for each of the proposed regions) the image, is turned into binary.
- Grey-scale morphological closing of the image is obtained.
- The bounding for each instance is obtained. Threshold of 100 pixels was



(a)

(b)

Figure 4-29: Detection of seed germination (b) in the regions proposed by CNN (a).

chosen. Thus, seeds with more than 100 white pixels are assumed to have already been germinated.

- Presenting the instances: the germinated seeds and background.

Figure 4-29b shows the outcome of the application of the proposed methodology: five seeds out of all seeds in the container are recognized as germinated ones on the 26-th hour after starting the experiment.

Implementation and evaluation of ML algorithms on the smart sensing embedded platform

All the developed machine learning algorithms mentioned above were implemented on the smart sensing low-power embedded platform (see Fig. 4-30). Similarly to the platfrom described in the Section 4.1 this platform is based on a Raspberry Pi 3 single board computer and an external GPU Intel Movidius with Myriad processor. The platform can easily run the pre-trained CNNs and deep neural networks (DNN). Although Raspberry Pi has restricted computational capabilities, the external GPU significantly expands the platform performance. The power consumption of GPU is 1 W with 100 GFLOPS performance at this power consumption. The system

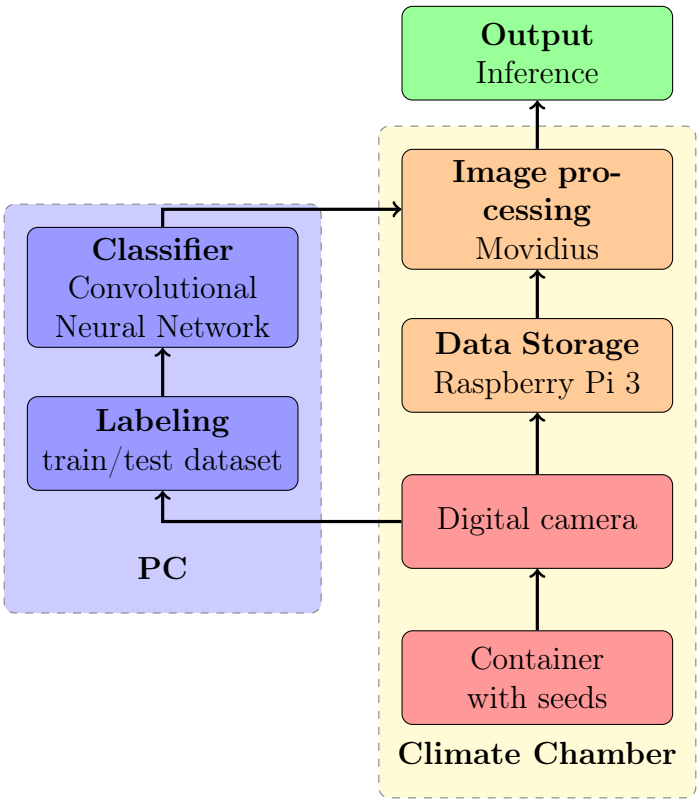


Figure 4-30: System block diagram

is powered by an external battery. The proposed platform successfully works with the pre-trained NN. The CNN used in the investigation was implemented on the embedded system in the following way. First, it was trained on a desktop computer using the *PyTorch* library. Then, the trained model was converted to the caffe model by open source software and finally was compiled into the binary graph to upload it on the GPU using the Intel's Movidius Neural Compute Stick library. NCS with Myriad processor was chosen because it has comparable performance per watt capabilities to Tesla K40 and Tegra K1.

The test dataset and the pre-trained CNN were uploaded into embedded system. Then the system run the pre-trained CNN to perform sequence of 1000 predictions on the test dataset. Such characteristics as CPU and RAM usage, time for prediction and power consumption were measured. Power, current and voltage measurements of the embedded system were collected every 100 ms.

For the developed prototype the mean CPU and RAM usage were about 37.04% and 30.67% respectively. In fact, if comparison is done with a desktop computer,

it will use much more CPU and RAM in absolute values, because there are a lot of background processes on the desktop, that consume additional resources. The mean period for prediction is 1.98 s. The mean power consumption of the prototype is 2.5 W. The proposed prototype dissipates 10 times less power consumption and can operate even if it is powered by an off-the-shelf battery. Assuming that a single prediction is accomplished within a 0.5-3 hours period and the device is in hybernate mode between the predictions, the operation time provided by a single battery (with 2600 mAh power capacity) could be significantly extended.

Conclusions

In this section, the application of CNNs in couple with CV algorithms for germination detection was demonstrated. A custom CNN architecture for seeds recognition achieves 97% accuracy and 83% of IoU. Using the CNN and computer vision, the sensing system is able to, first, localize them in the container and, then, detect the germinated seeds. It is beneficial for the emerging autonomous applications in the scope of IoT. For implementing the proposed approach it was collected a dataset which contains the sequentially time-ordered images of seeds germination process at different stages. Also, this solution was deployed on a low-power embedded system. One of the limitations of the developed system is that a labeled dataset is needed for training of CNNs. Most probably the CNN should be retrained for usage for other types of seeds or in other germination systems. The application of the system can also lead to wrong results if germinated seeds will overlap. One of the advantages of the proposed system is the possibility to provide a real-time quantitative assessment of the germination rate. This could be useful for conducting laboratory experiments aimed at optimization of the germination process and finding optimal environmental parameters for a particular type of seeds. The other advantage is that the proposed approach is scalable and has a strong industrial impact as a powerful tool for assessing the performance of germinating systems and predicting future harvest. At the same time, it provides an opportunity for making optimization at the initial stage of plant growth. This optimization will further result in the optimal management of resources in the context of precision agriculture.

4.4 Sustainability of plant growth: early remote diseases detection

Plant diseases can lead to dramatic losses in yield and quality of food. Data-driven approaches being accurate and fast become more and more useful for real-time, remote, all-encompassing plant diseases detection. Automation and early diseases detection using data-driven technologies in turn will allow to reduce labour aimed at manual disease monitoring and fungicide usage. In the following section it will be proposed and evaluated the approach for finding optimal wavebands in NIR and MIR for further plant diseases detection at early stages. Spectral analysis is widely used for detection and discrimination of plant diseases. Discrimination among yellow rust, powdery mildew and wheat aphid by using of multispectral analysis and Fisher's linear discriminant analysis was proposed in [Yuan et al., 2014]. Development of spectral vegetation indices for detection of sugar beet diseases in the spectral range from 450 nm to 900 nm by using of RELIEF-F algorithm was proposed in [Mahlein et al., 2013]. A variety of spectroscopic and imaging techniques for plant diseases detection was presented in a review [Sankaran et al., 2010]. One of the aim of the proposed in following research approach is provide simple and interpretable tool for finding optimal wavebands for diseases discrimination. This approach will be evaluated on apple tree diseases. However, as physical and biological essences of fungal diseases appearing in similar way, the proposed approach can be easily expanded and adapted to the investigation other plants such as tomatoes, discussed before [Pujari et al., 2015]. As it was shown in Section 4.3 and Section 4.1, it is easier to implement light-weight neural networks into the embedded systems. One of the benefit of proposed approach is that feature extraction will allow to detect diseases with usage of light-weight CNNs or FCNNs. These NNs will use input images in certain optimal spectra for detecting particular disease; thus it is not necessary for NN to extract spectral features from the lower layers, so it is possible to remove them. Also, it is not necessary to collect a huge dataset, as features have already been extracted. The workflow is following: first, spectral data in IR reflective spectra for different apple tree diseases (apple scab, moniliasis and powdery mildew) on different

stages of diseases development were collected. Then, using these spectra, the optimal wavebands based on the proposed discriminating coefficient, was obtained. The collected dataset as well as scripts in *Matlab* for processing data and finding optimal spectral bands are available link: https://yadi.sk/d/ZqfGaNlYVR3TUA

Dataset collection and spectral analysis

Apple scab is a disease of apple trees caused by fungi and it affects leaves and fruits. Seven samples of the apple leaves were selected for obtaining spectra: four leaves were infected with apple scab at different stages, two leaves were cured of apple scab, and one healthy leaf was used as a reference. On each leaf, in a small region of 1000 μm^2, 5-6 sub-regions of 10 μm^2 were allocated. For each of these sub-regions, the spectrum of reflected light was measured in the infrared region with the wavelength of 1.6-18 μm. The achieved spectral data were used to distinguish between infected, diseased, cured, and healthy leaves. The 35 spectra were obtained to get reliable results. The examples of leaf spectra and leaf samples are presented in the Fig. 4-31. Importantly, it was noted, that there is no much difference between the recorded spectra the regions with visible signs of scab, and other regions where there are no visible signs of scab, i.e., regions at the earliest stage of disease (Fig. 4-31). Similar results were obtained for fully damaged leaves (Fig. 4-31b) and for treated scab leaves (Fig. 4-31c). Unsurprisingly, spectra for healthy leaf significantly differentiate from infected samples in all investigated sub-regions (Fig. 4-31d). This observation creates the possibility for remote detection of apple scab. Using the same approach, data on leaves infected by moniliasis and powdery mildew were also collected. In total, it was obtained 20 spectra for moniliasis and 16 spectra for powdery mildew. The full dataset was published, which may provide the chance to further expand this study in the future.

Finding optimal spectral wavebands for apple tree diseases detection

The averaged spectra for healthy and infected by apple scab leaves are shown in the Fig. 4-32a. These averaged spectra were used to reveal the optimal bandwidth which can be used for for suitable infrared cameras selection for in vivo studies.

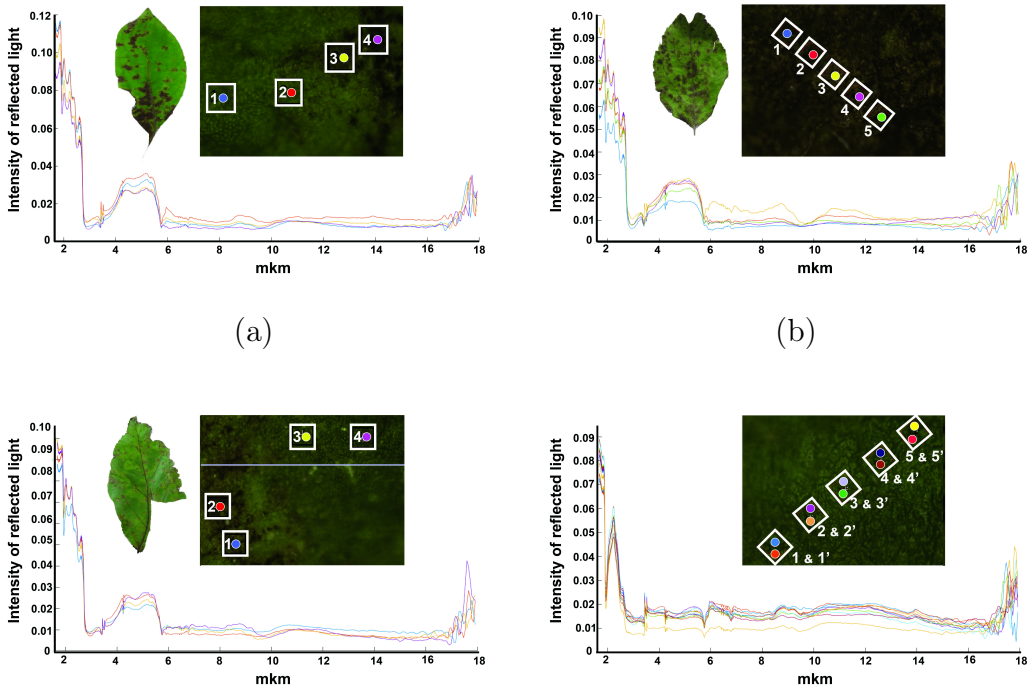
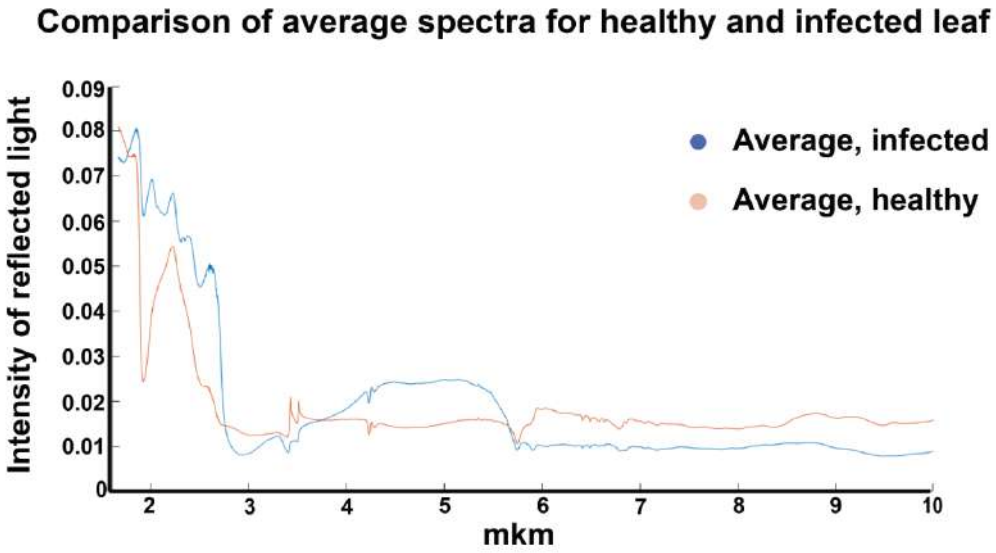


Figure 4-31: Obtained spectra and leaf samples for: (a), (b) four sub-regions of infected leaves respectively, (c) four sub-regions of cured leaf, (d) 8 sub-regions of healthy leaf.

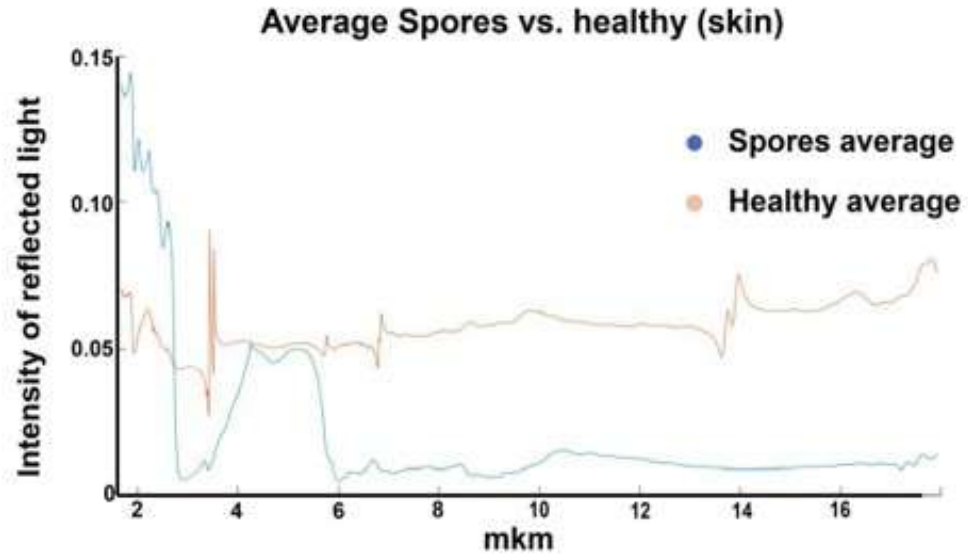
Based on the averaged spectra presented in the Fig. 4-32, discriminating coefficients were simulated in order to solve the classification problem. Using them, optimal bandwidths for disease detection were defined. The simulation was carried out in *MATLAB*. It was proposed to introduce a new discriminating coefficient similar in structure to the normalized difference vegetation index (NDVI). The proposed coefficient is the absolute difference of the reflection of the bands divided by their sum for normalization as given by equation(4.4):

$$DiscriminatingCoef(i,j) = \frac{|AUC_1(i,j) - AUC_2(i,j)|}{AUC_1(i,j) + AUC_2(i,j)}, \tag{4.4}$$

where AUC is the area under an averaged spectrum of wavebands, i is the wavelength from each waveband started, and j is the width of the waveband. The coefficients were calculated for the available spectra for all wavelengths and for all possible bands. The minimum bandwidth step used in the simulation is 50 nm, due to the



(a)



(b)

Figure 4-32: (a) Averaged spectra for infected by apple scab and healthy leaves, (b) averaged spectra for spored and healthy apples.

fact that more narrow-band cameras cannot be used in the field. The results are presented in the form of 2-d graph (see Fig. 4-33), representing the distribution of the values of the discrimination coefficient for different wavelengths and bandwidth for apple scab. It should be noted, that in the Fig. 4-33 simulation presented only up to 3.2 μm, because the value of discriminating coefficient for larger wavelength is not significant for this particular case (see Fig. 4-32). Areas in Fig. 4-33 with relatively high coefficient values represent the most selective wavebands. It can be noticed from Fig. 4-33, that the regions starting from 1.8-2.0 μm wavelength with bandwidth 0.2-0.4 μm, as well as the regions starting from 2.4-2.6 μm with bandwidth 0.1-0.4 μm have a good selective ability, since the coefficient value is relatively high. This result is very well explained by theoretical assumptions: Fungi, while destroying cells, causes decreasing of water content, which can be clearly seen in the infrared waveband of spectral region. The resulting classification bands coincide with the water absorption spectrum.

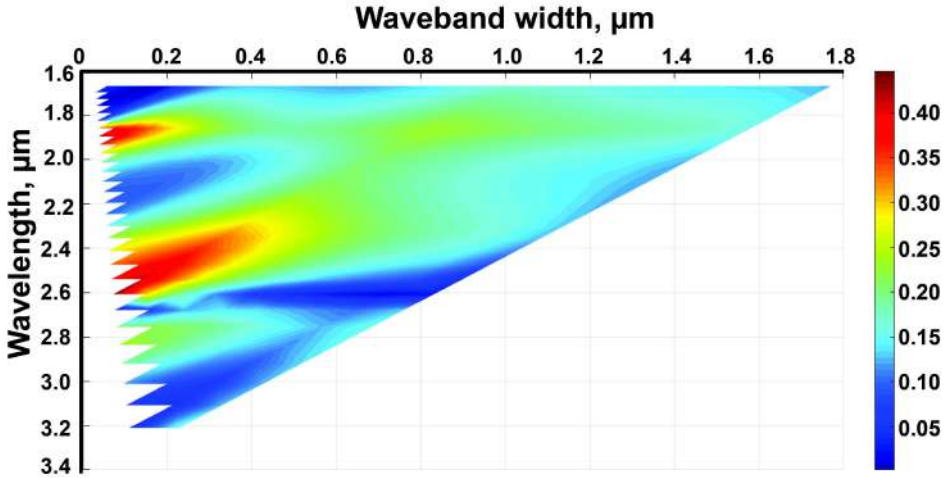


Figure 4-33: Distribution of the discriminating coefficient for different spectral wavebands, y-axis is wavelength from which waveband started, x-axis is the width of waveband.

Using the same approach discussed above and spectral data for other apple tree diseases, simulations of the distribution of discriminating coefficients were performed. Diagrams for Moniliasis and for Powdery mildew are shown in Fig. 4-34 and Fig. 4-35 respectively. Results of simulations are summarized in Table 4.4. From Table 4.4 it can be noted that optimal spectra for discriminating of different diseases

do not overlap. This is very important as it allows to discriminate different diseases using data, recorded from one multispectral camera.

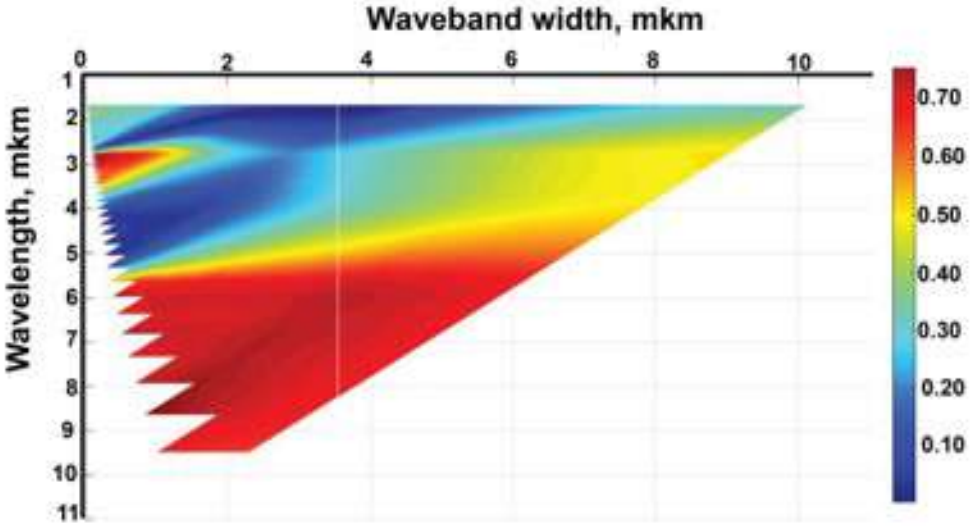


Figure 4-34: Distribution of the discriminating coefficient for Moniliasis for different spectral wavebands, y-axis is wavelength from which waveband started, x-axis is the width of waveband.

Table 4.4: Summary of the highest values of the discriminating coefficient representing the best wavebands in near and short infrared spectra for detecting apple tree diseases.

Disease	**Waveband,** μm	**Coef.value**
Apple scab, healthy/infected	1.8-2; 2.4-2.8	0.5
Moniliasis, spores/infected apple	2.8-3.1	0.7
Moniliasis, healthy/infected skin of apple	1.6-1.8	0.25
Moniliasis, healthy/spores	2.9-3.2; 5.9-9	0.8
Powdery mildew healthy/infected	2.7-2.9	0.41

Conclusions

In this section, the generic approach for finding optimal wavebands in IR reflectance spectra for early-stage disease detection was shown. This approach was evaluated with different apple tree diseases. For performing an evaluation of the proposed approach, a new dataset with spectral data in near-infrared and mid-infrared spectrum

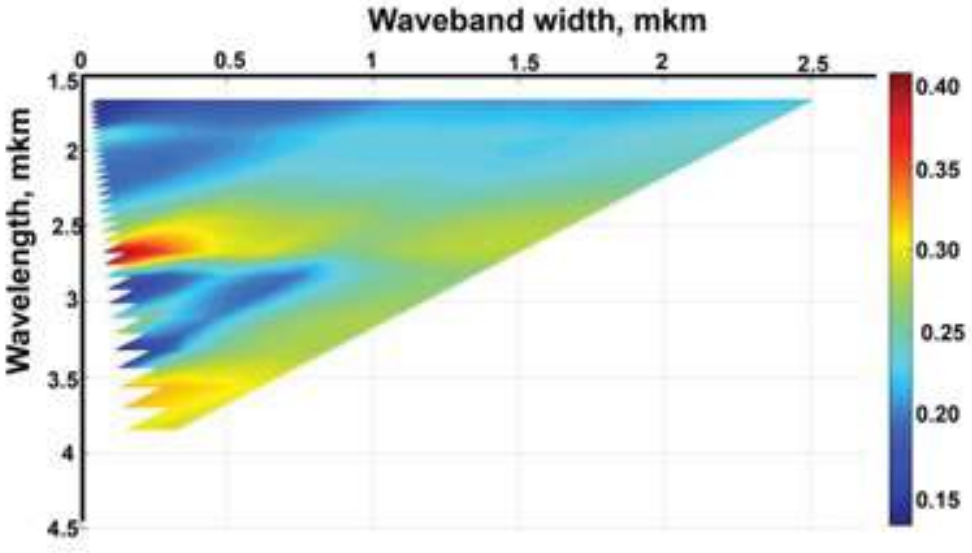


Figure 4-35: Distribution of the discriminating coefficient for Powdery mildew for different spectral wavebands, y-axis is wavelength from which waveband started, x-axis is the width of waveband.

range was built. This dataset consists of 51 spectra obtained in the spectrum range 1.6 - 18 μm for different apple tree diseases. The approach was successfully tested on the obtained dataset for detecting three different diseases. The obtained results coincided with theoretical assumptions and showed its accuracy. The drawback of the proposed method is that hyper-spectral data are needed for obtaining optimal wavebands. This method can be useful for designing a system that will detect diseases in field conditions. Using the obtained optimal wavebands it is possible to choose a proper multispectral camera for further evaluation in field conditions. However, the illumination in field conditions can slightly differ from that was used in laboratory investigations. This can lead to the tuning of the equipment, but anyway, the proposed method can provide the first assumptions of the optimal wavebands that are useful for detecting the particular disease. The advantage of this method is that it is universal. It can be applied to the investigation of the variety of plant diseases. Also, features, designed from spectral data, can be useful for a deep learning approach, in case of implementation on embedded systems for field disease detection. Using these features will potentially allow to decreasing the number of layers, making the deployed networks more shallow and easier to run on embedded systems. Overall

this study shows the potential of spectral analysis coupled with CV algorithms for remote plant diseases detection at early stages of plant growth.

4.5 Discussion

To compare developed data-driven techniques with hybrid model-based approaches, the developed data-driven approaches were evaluated on the same datasets that were used in Chapter 3 for creating hybrid models. First of all, the proposed in Section 4.1 RNN based technique for growth dynamics prediction can be directly compared to the DMD approach, developed the Section 3.4. For this purpose, the LSTM NN, with the same architecture as described above, was trained and evaluated on the same dataset that was used for modeling by DMD method in Section 3.4 (four tomato plants). LSTM was trained on the first half of the data and tested on the second half. The results of evaluation using the data for the 4-th plant showed the average relative error is 9.3% for the prediction horizon up to 5 hours (10 time steps). While testing of DMD for the same data (presented in Fig 3-22) showed the average relative error 16% for the whole interval of the prediction. Using the same training procedure, the LSTM was evaluated on the data obtained for all four tomato plants. The mean relative error among predictions for all four plants is 8.35% for the prediction horizon up to 5 hours. A similar study of errors was conducted for DMD during cross-validation procedure when data for different three out of four tomato plants were used for training DMD and data for the remaining plant were used for testing of the algorithm. The obtained relative error for 4-fold cross-validation is 18.5%. It is important to notice that for training RNNs only previous points in time sequence are used. However, for the DMD algorithm the data for the whole investigated period are needed for training procedure. This means that to train DMD algorithm it is needed to perform preliminary experiments to obtain data for similar plants for the whole period of modeling. This, in turn, allows the DMD algorithm to perform such long-term predictions, but it is needed to obtain preliminary data for training. On the other hand, RNNs predictions are relying on the data that are receiving from the same plant in real-time. RNNs are adaptive to

different plants and are able to take into account the current and previous growth dynamics of a particular plant. They can also be retrained while new measurements are coming, which makes them more accurate. Thus, it is beneficial to use them, as the universal tool for growth dynamics prediction for a variety of plants and under various environmental conditions. However, if the growth dynamics of the particular type of plant under certain environmental conditions was investigated and the data were obtained for the whole period of interest, DMD allow to perform long-term predictions (see Fig. 3-22). One of the limitations of RNNs is that they are essential "black-boxes", so the results can't be interpreted directly. On the other hand, by reconstructing the evolution operator in the DMD approach, it is possible to assess the impact of each parameter on the predicted values. However, as it was stated in the Section 3.4, fine-tuning of a set of features that defines a state vector is needed. The other limitation of RNNs that they are much more computationally complex compared to DMD. Thus, DMD is much easier to deploy into the low-power embedded systems.

For comparison of the data-driven approaches with Kalman filter, RNN was trained and evaluated using the same data as that were used for Kalman filtering (9 lettuce plants, see Section 3.2). The first 400 data points (out of 820) were used for training RNN, the rest data points for testing. Data obtained for 2-nd plant were used as an example for training/evaluation procedure. Using this data, the average relative error 5.7% was achieved for the prediction horizon up to 5 hours (10 time steps). Similar results were obtained for other plants. The accuracy of the Kalman filter was evaluated by reconstruction of projected leaves area, using predicted values of μ and S_{max} for the same tested data. The average relative error for predictions by Kalman filter was 10.2%. Such low accuracy compared to the RNNs appears due to the model (Verhulst) that is used for the Kalman filter. This model doesn't capture diurnal fluctuations in the projected leaves area. This drawback can be a benefit of usage Kalman filter because if a more complex model will be adapted to the Kalman filter, it will be able to provide more accurate results. Also, the Kalman filter, being computationally simple has a high computational efficiency which allows its usage on embedded devices. The other advantage of Kalman filter compared to RNNs is

that no training procedure is needed. Also, no prior information such as for DMD is required. Thus, the Kalman filter has shown itself as the most straightforward method for raw evaluation of growth dynamics among all other investigated.

Table 4.5: Summary of models comparison

Method	Advantages	Limitations	Relative error, %
RNN	+Universal +Uses the current state and past dynamics +Robust to real-time environmental changes +Automatically adaptive	-Difficult to interpret -Computationally complex	Tomato dataset Section 4.1, 5.4% Tomato dataset Section 3.4, 8.35% Lettuce dataset Section 3.2, 5.7%
DMD	+Long-term predictions +Model-based	-Uses prior information for similar plants -Fine-tuning of state vector	Tomato dataset Section 3.4, 16%
Kalman filter	+Straightforward +Model-based +Computationally effective	-Complex models are needed for accurate estimations -Linearization errors -Convergence	Lettuce dataset Section 3.2, 10.2%

The main obtained results as well as advantages and limitations of methods are shortly summarized in Table 4.5.

4.6 Conclusions

In this chapter, a set of approaches for assessment and prediction of plant growth dynamics at different stages using data-driven approach was presented. More specifically the following studies were conducted:

- Implementation and evaluation of RNNs, in particular, LSTM for growth dynamics prediction. This approach was applied to an own dataset obtained on the developed artificial system equipped with CV and sensing systems. The proposed approach, being accurate, showed the high potential in industrial application due to easy end-to-end deployment and universality.

- Implementation of CV and sensing system in an industrial experiment. This experiment allowed to collect comprehensive imaging and environmental data and biomass measurements. These data were used for training FCNNs to perform semantic segmentation and to calculate projected leaves area of each plant automatically. After this, growth dynamics was reconstructed using sequences on images.

- Creation of the system for germination rate detection based on CV approaches. CNNs were trained for seeds localization based on own experimental data. Using the proposed regions, CV techniques were applied for germination quantification and rate detection. The developed autonomous system for germination monitoring is scalable accurate and autonomous, showing high practical usefulness for industrial application.

- Evaluation of an approach for finding optimal wavebands in IR reflectance spectra for early-stage disease detection. The data that describe plant diseases at the different stages in near-infrared and mid-infrared spectrum range were collected. The results showed that it is possible to detect and to discriminate diseases at early stages, when they can not be detected in visible spectrum. This approach can also help to extract features making CV-based approaches for diseases detection less complex and more reliable.

To sum up, in this chapter an end-to-end implementation and evaluation (on own experimental datasets) of data-driven approaches for growth dynamics assessment and prediction were proposed. Such approaches proved to be robust, universal and useful in supporting sustainable plant growth.

Chapter 5

Data-driven modeling of environmental parameters for improvement of plant growth prediction

5.1 Problem statement and proposed solutions

Controlling and predicting plants growth on the field is a complicated task due to a high variety of parameters that influence growth dynamics. There is a huge amount of factors that can a have positive or negative effect on the growth rate, the final yield and the quality of production. These factors can be divided into controlled and uncontrolled. There are some model-based methodologies, implemented in software, that can model the yield taking into account different controlled and uncontrolled parameters [Nendel et al., 2011]. There are two problems that should be solved for improving the accuracy of such model-based techniques:

- Accurate prediction of the spatial distribution of the uncontrolled (or slightly controlled) environmental parameters that are included in the models.
- Accurate assessment of the effects of controlled environmental parameters, that are not included in the models, on growth.

Prediction of spatial distribution of uncontrolled parameters. For solving this issue, a robust and accurate method for prediction of the spatial distribution of environmental parameters should be proposed. The example set of soil and environmental parameters the spatial distribution of which should be modeled before including them into the model for yield prediction [Krishnan and Aggarwal, 2018]:

- Thickness, sand content, silt content, clay content
- Hydraulic conductivity initial moisture
- Ammonium content, nitrate content, pH, electrical conductivity
- Groundwater composition

More comprehensive knowledge of these parameters, their distribution and accuracy of measurements leads to more precise predictions. Currently used solutions for modeling of spatial distribution of the environmental parameters are mainly based on the ordinary Gaussian process regression (GPR). The procedure of obtaining the interpolated maps of each modeled parameter includes a manual selection of the kernels (variograms) and their parameters. This leads to an important drawback: The results of modeling can differ dramatically for the same used data because different kernels and their parameters can be selected. The process of tuning model by selection of different kernels and parameters is also time-consuming. In order to overcome these problems, it was proposed an advanced technique based on Gaussian process regression (GPR) and optimal kernel structure selection using Bayesian information criteria (BIC) (see Section 5.2). The main advantage of the proposed method is that it automatically finds the best possible kernel structure which also has the least complexity for the particular problem (dataset). The proposed technique was validated on a new dataset for freshwater chemical composition, which is the important uncontrolled factor that should be included in the modeling of the plant growth dynamics and final yield. The dataset was obtained in the newly added to Moscow territories in 2011 and contains data samples (chemical composition) form 1194 wells, 222 small rivers, and 153 springs. As an additional enhancement of the developed methodology, it was proposed the PCA-based aggregated index that automatically defines the most influential chemical water properties to construct one

aggregated water quality index (WQI). The proposed GPR with BIC method for spatial interpolation was evaluated on the aggregated WQI. However, it should be noticed that any of the chemical compounds that are included in the WQI can be interpolated directly in the same way as for WQI by using the proposed technique. This technique can also be applied directly to obtain the spatial distribution of any soil and environmental parameters that are listed above. According to basic methods for assessing goodness of fit and cross-validation (average R^2 is about 0.64 by 5 fold cross-fold validation, and average root mean squared error RMSE is about 0.065), it was concluded that the proposed method is more accurate than traditional interpolation techniques such as ordinary and universal kriging. The proposed method is also universal because of automatical kernel structure selection. Thus, the powerful mathematical approach for modeling of the spatial distribution of the most important environmental parameters was developed and evaluated.

The proposed methodology and its evaluation as well as experimental data are discussed in details in the Section 5.2.

Assessment of the effect of controlled parameters on growth. There are still some factors that are not included in the modeling since they have a complex and unstudied effect on plant dynamics. One of the most important factors that was not taken into consideration is phytotoxicity. This factor can affect growth and plants quality dramatically [Nagajyoti et al., 2010]. However, there are no robust models that can predict phytotoxicity effects in a huge amount of different types of soils. Machine learning approaches opens huge perspectives for solving this problem. There are two main sources of toxicity: (1) direct exposure of harmful substances (such as oil contamination) and (2) concomitant insertion of fertilizers into the soil. The advantages of ML approaches over other methods for solving the problem of phytotoxicity prediction were shown based on investigations focused on total petroleum hydrocarbons (TPH) phytotoxicity assessment (which represents as direct exposure by pollutants) and on the assessment of effects of mineral waste-based fertilizer insertion. Two use cases were studied for evaluating the proposed ML techniques that are able to predict the effect of controlled parameters on growth.

The first one is modeling the TPH acute phytotoxicity effects that was performed on eleven samples of soils from Sakhalin island in greenhouse conditions. Different soils were contaminated with crude oil in different doses ranging from the 3.0 to 100.0 g•kg-1. Measuring the *Hordeum vulgarie* root elongation, the crucial ecotoxicity parameter was estimated. Also, the contrast effect in different soils was investigated. To predict TPH phytotoxicity, different machine learning models were used, namely artificial neural network (ANN) and support vector machine (SVM). These models were proved to be valid using the mean absolute error method (MAE), the root mean square error method (RMSE), and the coefficient of determination (R^2). It was shown that ANN and SVR can successfully predict barley response based on soil chemical properties (pH, LOI, N, P, K, clay, TPH). The best achieved accuracy was as following: MAE – 8.44, RMSE –11.05, and R^2 –0.80. The proposed methodology and its evaluation as well as experimental data are discussed in details in the Section 5.3.1.

The second use case is modeling the toxicity effect of mineral waste-based fertilizer insertion in soil. The phytotoxicity was evaluated by the quantification of the effects of different doses input of phospogypsim (PG). The results show a similarity between the 0%, 1% and 3% PG treatments at all collection times based on toxicological and biological properties. Beyond 7.5% PG, some biological test was significantly inhibited in response to trace element stress. Among all tested parameters, soil urease activities, soil respiration activities after glucose addition, *S. alba* root lengths and *E. fetida* survival rates show a sensitivity to PG addition. This means that the prediction and quantification of the effects of different doses input of phospogypsim (PG) is crucial for plant growth modeling. Machine learning algorithms revealed that only several elements (mobile and water-soluble forms of Ca, Ba, Sr, S, and Na, water-soluble F) could be responsible for elevated soil toxicity for those indicators. SVR models able to predict soil biological and ecotoxicity properties, and increasing numbers of randomly selected training examples from 50% to 90% of initial experimental data significantly improved model performance. The benefits of unsupervised and supervised machine learning methods for investigating the toxicity of man-made substances in soil were shown. Tracking and assessment

of phytotoxicity effects on plant growth based on ML methods will give possibility to complement the existing models with missing but important parameters allowing to decrease modeling uncertainties. The proposed methodology and its evaluation as well as experimental data are discussed in details in the Section 5.3.2.

Overall, the proposed in the following sections data-driven modeling of environmental parameters gives an opportunity:

- To include controlled parameters such as toxicity effects by inserting of fertilizer in existing model-based algorithms for yield prediction making them more precise and robust.
- To predict more accurately the spatial distribution of the uncontrolled environmental parameters making the existing model-based algorithms for yield prediction more precise.

5.2 Machine learning approaches for assessment of water quality distribution

Introduction

Accurate and large-scale monitoring of freshwater and groundwater quality is one of the main tools used to assess current ecological situation and to indicate the drivers and trace sources of pollution, which particularly can affect the agriculture industry by yield losses [Han et al., 2016, Berger et al., 2017]. Since water resources can be described by a large number of chemical, physical and biological parameters, normally, studies dedicated to water quality assessment implement one integral parameter to characterise the overall water state: the water quality index (WQI). The WQI aims to reduce the wide range of individual parameters to one joint descriptive characteristic of high practical importance and easy to interpret at the same time. Normally, monitoring parameters for the calculation of the WQI with specific levels of importance are selected in two different ways: subjectively by expert opinion [Ramakrishnaiah et al., 2009] or objectively on the basis of statistical methods

[Sun et al., 2016, Tripathi and Singal, 2019b], such as principal component analysis, factor analysis and cluster analysis for the next stage of data transformation into a mathematical expression.

Big data approaches to the collection, evaluation and prediction of environmental data, including water quality, have become more popular due to recent advances in machine learning applications and the availability of modern, high-accuracy equipment and sensors for water quality measurement [Alilou et al., 2019, Mitrović et al., 2019, Karami et al., 2014]. For example, the Next Generation Weather Radar, and the Global Precipitation Measurement Mission collect tens of terabytes of data every year. In the meanwhile, the measurement of water quality in many countries, including Russia, is still based on spot networks; more importantly, the overall density of sample collection may be very low [Zhulidov et al., 2000]. Concerning this fact, the development of adaptable algorithms to select the best sampling point locations is a matter of topical interest, and using ML methods for data extrapolation have a great potential to become a "workhorse" of scientific community as well as being implemented for management issues [Madrid and Zayas, 2007, Keskin and Grunwald, 2018].

Regarding water quality parameter spatial prediction, different tools (SAGA GIS, QGIS, ArcGIS, special packages in the environments of R and Python) exist and provide frameworks for kriging and mapping the spatial distribution of properties of natural objects, including forestation, soil properties and groundwater properties, on the basis of limited data [Barzegar et al., 2019, Khaki et al., 2018, Sajedi-Hosseini et al., 2018]. The basic limitations of these tools are the non-automatic method of variogram fitting and the manual selection of calculation parameters (e.g. search distance and maximum points in search distance), which should be discussed clearly every time to obtain the same result on the same data, or, in contrast, hidden options in closed-source software. Meanwhile, the state-of-the-art approaches now involve ML algorithms coupled with geostatistical data processing; these approaches allow more precise, high-resolution establishment of the spatial distribution of characteristics.

The main aims of the following study are:

- The development and validation of the improvement of the standard Gaussian process regression (GPR) technique.
- The demonstration of the practical usefulness for environmental parameters interpolation by investigation of the use case of water quality prediction.

Materials and methods

Site description and available dataset. The object of current study is the newly added territories to Moscow, located adjacent to the city of Moscow in the Central European part of Russia (55°N, 37°E) and extends over 1480 km^2 area. This territory accommodates a wide variety of land-use types, including farmlands, croplands, natural grasslands and forests, apart from suburban settlements and some industrial operations. The mean annual temperature of this region is about 3-4°C. The mean temperature in the coldest month of the year (January) ranges between -9.5°C and -11.5°C, while in the warmest month, July, mean temperatures are between +17°C and +18.5°C. The average annual precipitation is approximately 500-520 mm, with approximately with approximately two thirds rainfall and the rest snow. The predominant types of natural vegetation are coniferous and broad-leaved forests, while agricultural lands include pastures and arable land mostly for feed crops and cereals. Main specific of the investigated territory is that it has been rapidly urbanised during the last decade.

The set of samples, using in this study, covers almost all the investigated territory. A total of 1600 water samples were collected during 2017-2018 from wells (1215 samples), rivers (225 samples) and springs (160 samples) covers the whole region (see Fig. 5-1). Water samples were collected from wells, rivers, or springs by using a 2-L stainless-steel container. The samples were bottled and then immediately transported to the laboratory for chemical analysis, eliminating the need for conservation methods.

For each water sample, 25 parameters were measured. The pH was measured by using a HANNA pH-meter 213. Anions (NO3, NO2, PO4, SO4 and Cl) were measured by ion chromatography using a Dionex 1100 instrument. NH4 content was obtained on an HACH DR2800 using colorimetric determination with Nessler's

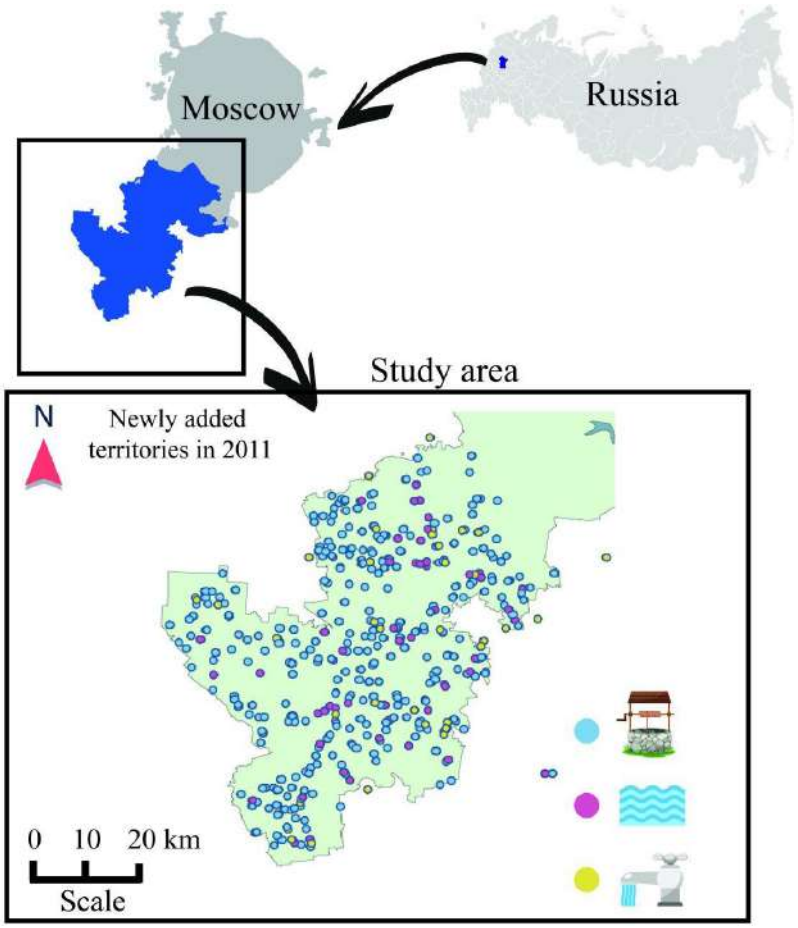


Figure 5-1: Location map of the study area. Different colours mark source of collected water samples - wells coloured in blue; rivers coloured in purple; and springs coloured in yellow.

reagent. Cation (K, Cr, Ni, Ca, Zn, Fe, Mn, Na, Cu, Mg) contents were obtained by inductively coupled plasma atomic emission spectroscopy with an ICP-OES Agilent 5110 spectroscope. Mineralization was measured by gravimetric analysis consisting of evaporation at 105°C in a drying chamber. Alkalinity was obtained by titration with 0.05N HCl. Hardness was measured by titration with Trilon B and eriochrome black. Overall, it was obtained relatively large size of the dataset which contains more than 1600 samples (each with 25 measured chemical parameters). It might be useful for validation and other methodological research in community and the dataset was shared through the following reference [Pukalchik et al., 2020].

Data preparation and methodology.

An end-to-end solution for geospatial water quality assessment using modern machine learning methods such as Gaussian process regression and Bayesian information criteria was proposed and evaluated. Figure 5-2 presents a brief summary of the steps involved in this procedure.

Water quality index calculation based on PCA and weighted factors. A PCA model was used to assess the pollutant loads integral to water quality and to avoid data redundancy. Raw data were filtered to eliminate anomalies: missing coordinates, incorrect record type etc. After this initial pre-processing step, the total number of useful samples decreased from 1600 to 1569. Then it was decided to remove Hg, Cd, Co and Pb from the dataset for further analysis, as their concentrations were insignificant (much lower than toxic levels) and did not exceed the required water quality standards in Russia. Twenty-one water quality parameters were included in the PCA model. Only those components for which the corresponding eigenvalue was higher than or equal to 1 following *Varimax* rotation, and PCs that explained at least 5% of the observed data variation were considered for further examination. Moreover, those parameters that were correlated with other significant parameters (correlation more than 0.6) were eliminated only if they had the smallest loadings among the correlated parameters. The weight scores (w_i) derived from PCA were used as weighted factors for the significant variables (indicators)

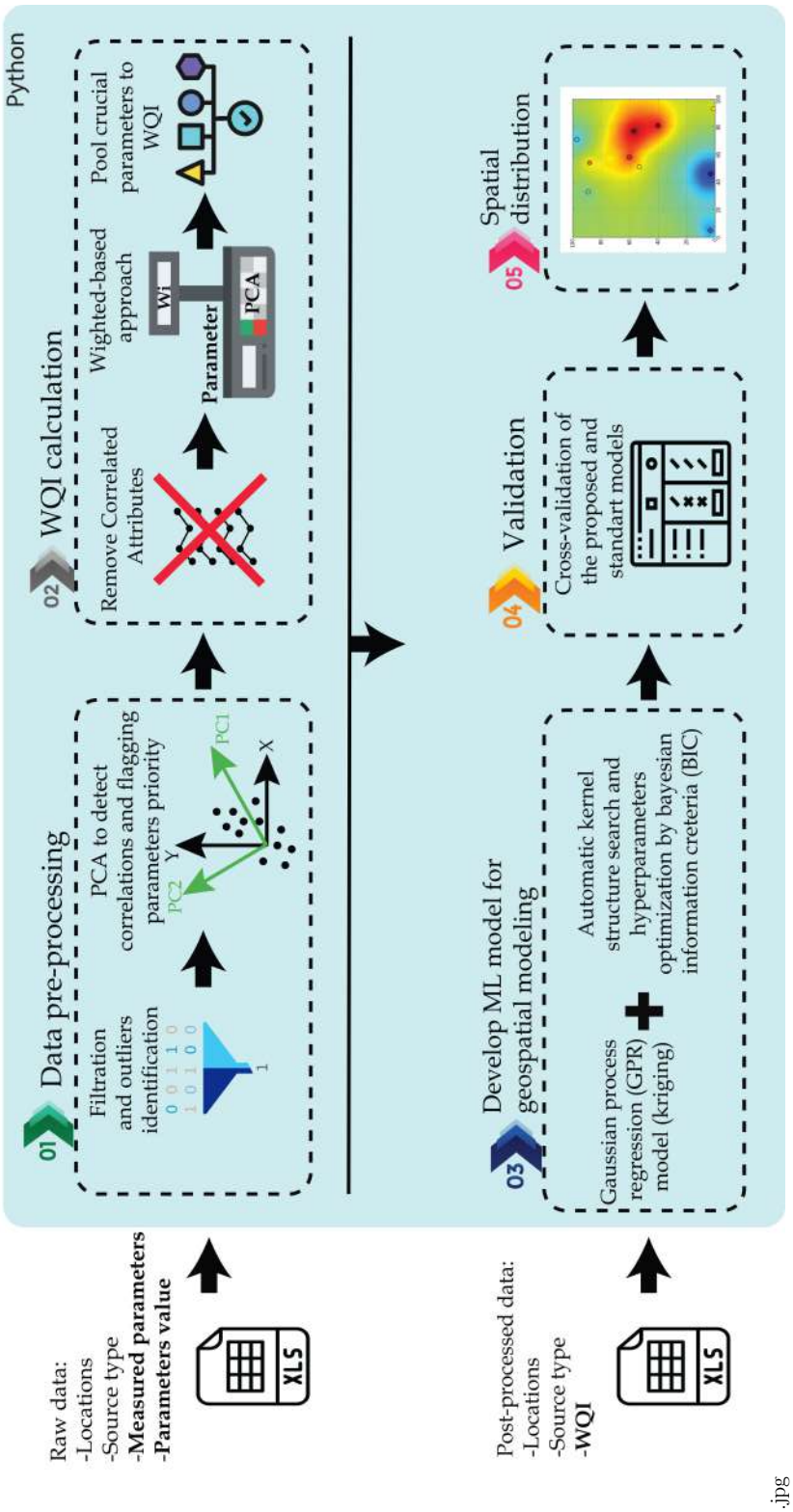


Figure 5-2: Methodology for using machine learning methods for weighted WQI calculation (Steps 1 and 2) and geospatial WQI prediction by using Gaussian process regression with automatic kernel search (Steps 3-5).

from respective PCs, and the WQI was calculated by using eq. 5.1:

$$WQI = \sum_{i=1}^{S} L_i \cdot w_i, \tag{5.1}$$

where S is the number of significant principal components, L_i denotes the loading on values of each selected water property included in the particular principal component and w_i denotes the weight of the corresponding component. In order to scale WQI to the [0,1] range, the weight scores was normalized (w_i) by using eq.5.2:

$$w_i := \frac{w_i}{\sum_{i=1}^{S} w_i} \tag{5.2}$$

Machine learning approach for geospatial modelling of WQI with automatic kernel detection

Gaussian process regression: general overview of the methodology. In order to perform geospatial modeling of multiple water properties from the collected dataset, *Gaussian Process Regression* (GPR) framework was used, more commonly known as *kriging* in geostatistics [Williams and Rasmussen, 2006]. A stationary Gaussian process is completely determined by its *mean* $\mu(\cdot)$ and *covariance* (kernel) $k(\cdot, \cdot)$ functions:

$$\begin{aligned} f(\mathbf{x}) &\sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')), \\ \mu(\mathbf{x}) &= \mathbb{E}\ f(\mathbf{x}), \\ k^x(\mathbf{x}, \mathbf{x}') &= \mathbb{E}\ [(f(\mathbf{x}) - \mu(\mathbf{x}))(f(\mathbf{x}') - \mu(\mathbf{x}'))], \end{aligned}$$

where $\mathbf{x} \in \mathbb{R}^2$ is a vector of d input parameters. In this particular case, $d = 2$ and $\mathbf{x}$ represents a vector of spatial coordinates. Let us Consider a simple GPR model with additive Gaussian noise:

$$y(\mathbf{x}) = f(\mathbf{x}) + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Given the training data $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^\intercal \in \mathbb{R}^{n \times d}$, $\mathbf{y} = (y_1, \ldots, y_n)^\intercal \in \mathbb{R}^n$, where n is the number of samples and $(\cdot)^\intercal$ denotes the transpose operator, the predictive distribution at the unobserved point $\mathbf{x}^*$ is given by

$$\begin{aligned} \hat{f}(\mathbf{x}_*) &\sim \mathcal{N}(\hat{\mu}, \hat{\sigma}^2), \\ \hat{\mu}(\mathbf{x}_*) &= \mu(\mathbf{x}_*) + k_*^x \Sigma(\mathbf{y} - \mu(\mathbf{X})), \\ \hat{\sigma}^2(\mathbf{x}_*) &= k(\mathbf{x}_*, \mathbf{x}_*) - k_*^{xT} \Sigma^{-1} k_*^x, \\ \Sigma &= K^x + \sigma^2 I, \end{aligned} \tag{5.3}$$

where I is an identity matrix, $K^x = k^x(\mathbf{X}, \mathbf{X}) = k(\mathbf{x}_i, \mathbf{x}_j)$, $i, j = 1, \ldots, N$ is a spatial covariance matrix between all of the training points, $k_*^x = k(\mathbf{X}, \mathbf{x}_*)$ is a spatial covariance between training points and the single prediction point and $\mu(\mathbf{X}) = \mu(\mathbf{x}_i)$, $i = 1, \ldots, n$ is the mean function calculated at the training points. The particular choice of the kernel function depends on the assumptions about the model and a particular application, e.g., *Gaussian Kernel* (corresponding to the Gaussian variogram). Kernel hyperparameters are usually optimized using the *Maximum Likelihood Estimation* (MLE) or its variations [James et al., 2013].

Figure 5-3 shows an example of GPR using a Gaussian kernel over the observations sampled from the sine function with random noise. The prediction variance increases at points with missing observations, and increases significantly outside of the interpolation region with the mean failing to capture the true function trend. This emphasizes the need for a better method to select kernel hyper-parameters.

Hyper-parameter Selection using Bayesian Information Criteria. Common approaches to hyper-parameter optimisation are *Maximum Likelihood Estimation* (known model, continuous parameters), and *Cross-Validation* (model is unknown, discrete parameters). Typically, one could select multiple combinations of different kernels, perform MLE for each of them and then compare the models using cross-validation to choose the best overall model. It was decided to follow the approach in [Duvenaud et al., 2013] using Bayesian Information Criteria (BIC) which represents a space of covariance function as a combination of a small number of base

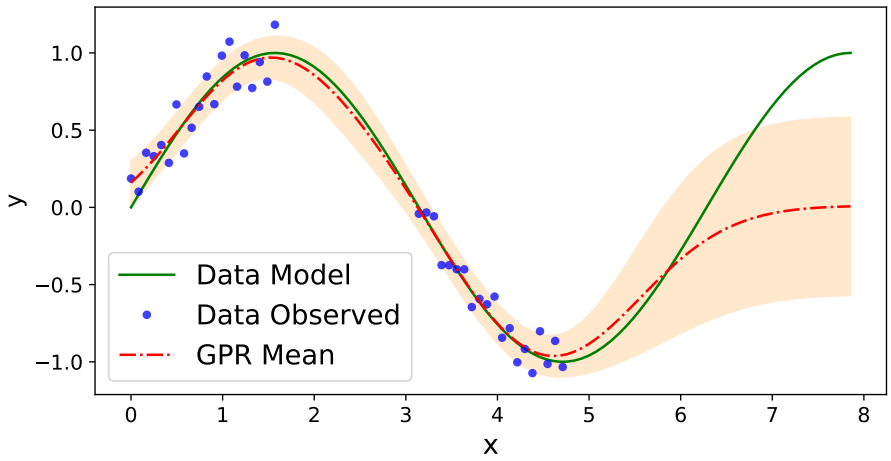


Figure 5-3: Gaussian Process Regression (red dashed line depicts the predictive mean and orange fill depicts the standard deviation intervals) with noisy measurements (blue dots) of the sine function (solid green line) using RBF kernel.

covariance functions using sum and product operations, and may be represented as:

$$\begin{aligned} BIC &= -2 \cdot \text{Log-likelihood} + m \cdot \log n, \\ \text{Log-likelihood} &= -\frac{n}{2} \cdot \log 2\pi - \frac{n}{2} \cdot \log |\Sigma| - \frac{1}{2} \cdot (\mathbf{y} - \mu)^T \Sigma^{-1} (\mathbf{y} - \mu) \end{aligned} \tag{5.4}$$

where n is the number of samples, m is the total number of optimised parameters, and Σ is defined as in the eq. 5.3. To construct the optimal kernel, it was considered a basic set of operations, such as *plus* and *multiplication*, These operations were applied to the following kernel functions: polynomial (Eq. (5.5)), Gaussian (Eq. (5.6)), periodic (Eq. (5.7)) and exponential (Eq. (5.8)). Thus, the final automatically constructed kernel, for example, can be the multiplication of the polynomial kernel on Gaussian, plus periodic, etc. Optimal kernel structures can include the multiplication of the same types of elementary kernels:

$$k_{poly}(\mathbf{x}, \mathbf{x}' | \, \theta_1, \theta_2, \theta_3) = \theta_1 \left(\sum_{i=1}^{d} \theta_2 \mathbf{x}_i \mathbf{x}'_i + \theta_3 \right)^{deg}, \tag{5.5}$$

$$k_{gaussian}(\mathbf{x}, \mathbf{x}' | \, \theta_4, \ell) = \theta_4 \exp \left(-\frac{1}{2} \sum_{i=1}^{d} \frac{(\mathbf{x}_i - \mathbf{x}'_i)^2}{\ell_i^2} \right), \tag{5.6}$$

$$k_{periodic}(\mathbf{x}, \mathbf{x}' | \, \theta_5, s, T) = \theta_5 \exp\left(-\frac{1}{2}\sum_{i=1}^{d}\frac{1}{s_i}\sin^2\left(\frac{\pi}{T_i}(\mathbf{x}_i - \mathbf{x}'_i)\right)\right), \tag{5.7}$$

$$k_{exp}(\mathbf{x}, \mathbf{x}' | \, \theta_6, l) = \theta_6 \exp\left(-\sqrt{\frac{1}{2}\sum_{i=1}^{d}\frac{(\mathbf{x}_i - \mathbf{x}'_i)^2}{l_i^2}}\right), \tag{5.8}$$

where $d = 2$, polynomial was taken of degree 2; θ_1, θ_4, θ_5, θ_6 are the variances; θ_2, ℓ, l and s are length scales; θ_3 is the bias; T is period. In performed calculation experiments all of the kernels were considered isotropic and the following constraints during hyper-parameter optimization were applied: $T_1 = T_2 = T \in [1, 10]$, $s_1 = s_2 = s \in [0.1, 10]$ and $\ell_1 = \ell_1 = \ell \in [0.1, 10]$, other parameters were left unconstrained. The best kernel is a combination (structure) of the elementary kernels with optimized parameters that gives the minimal BIC value. This way, it is possible to model a variety of stationary kernels and control the accuracy by selecting basic kernels and boundary values for their hyper-parameters. The main goal of introducing such boundary values is to avoid over-fitting and ensure the robustness of the performance of the obtained optimal kernel (composition of the basic kernels). Moreover, the aim is to reduce the model complexity by decreasing the number of tuned hyper-parameters in the optimal kernel.

The procedure of fitting the Gaussian process is quite computationally expensive $O(n^3)$, where n is the number of training data points. Hence, instead of brute-force search of the best kernels, a greedy search was implemented in the current study. Greedy search in general means that the extension with the lowest BIC is selected for each extension of the current kernel. The main advantage of this approach is that it does not require any handcrafting of potentially effective kernels, but instead enables an automatic search for the best kernel structure and hyper-parameter optimisation.

Universal and Ordinary kriging. To compare the proposed method with baseline geospatial modelling techniques, Ordinary Kriging (OK) and Universal Kriging (UK) using the *GPy* library were preformed. Since this library allows to perform Gaussian process regression, the connection between Gaussian process regression and kriging methods is following: a) basic kernel functions k_{poly}, $k_{gaussian}$, $k_{periodic}$,

k_{exp} correspond to the respectively variograms, b) GPR with a constant mean function $\mu(\mathbf{x}) = \mu$ corresponds to OK, and c) GPR with a linear mean function $\mu(\mathbf{x}) = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2$ corresponds to UK with linear trend. Hyper-parameters of the kernel and mean functions are optimized using MLE approah during the training phase.

Approach to geospatial modelling. Firstly, spatial coordinates were converted from EPSG:4326 (latitude, longitude) format to EPSG:32637 (UTM zone 37N) format. Then, the converted coordinates were scaled down to the [0,10] range. Some measurements of the water quality measurements were taken spatially far from the main investigated area (Moscow region). Thus, to filter outliers, using the scaled coordinates, clustering of the water sampling locations was performed using the density-based DBSCAN method [Ester et al., 1996]. In particular, the hyper parameter $\epsilon = 1.0$ in the DBSCAN method was used and allowed to allocate a main cluster and outliers. The parameter ϵ allows to tune the size of a cluster and serves to set the permissible distance to the point to be included into the cluster. After clustering and removing the outliers, the coordinates again were re-scaled within [0,10] range. Finally, it was decided to use the data only from the large class (wells, 1215 data points). In total, 391 data points were removed from the dataset (37 data points out of them were removed by DBSCAN) and 1178 data points were kept for further investigation. The WQI was calculated for each data sample and a rectangular 100 x 100 grid was used for geospatial modelling and mapping. The boundaries of the selected grid were defined by the minimum and maximum coordinates of the kept water sampling locations.

Validation procedure. To validate the developed model, it was applied a standard validation approach with 5 random splits on training and testing datasets of relative size 90% and 10%, respectively. For each training/testing split, a) the model to the training data was fitted, then, b) the values of WQI for the test data point locations were predicted and c) the coefficients of determination R^2 and the root mean squared error (RMSE) were calculated:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}, \tag{5.9}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}, \tag{5.10}$$

where $\hat{y}_i$, y_i are the predicted and observed values, respectively, and $\overline{y}$ is the average of the observed value.

The RMSE is a good comparative statistic for assessing model output, as it provides a global indication of how similar the interpolated values are to the observed or measured values [MacCormack et al., 2013]. When analysing the RMSE statistics, a small RMSE value indicates that the interpolated values for the output model are more similar to the observed values, whereas a large RMSE value suggests that the interpolated model values are less similar to the observed data points. Thus, the RMSE values are used here to determine how well the model fits the observed data values, with low RMSE values indicating a high degree of model accuracy [MacCormack et al., 2018, Mueller et al., 2004]. All the calculations were carried out in Python using the following libraries: scikit-learn, [Pedregosa et al., 2011], GPy, [GPy, since 2012] and Folium.

Results

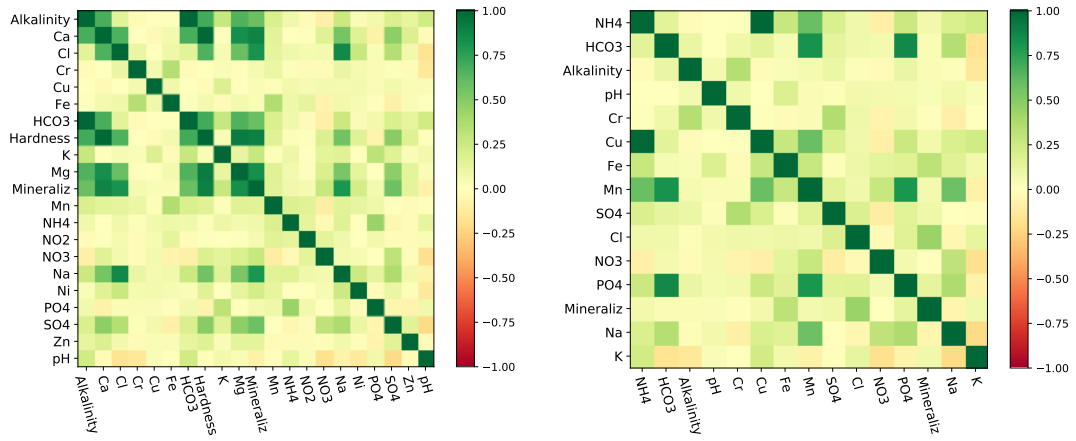
PCA-based weighted water quality index. The PCA method was used to reveal the significant contaminants among samples and calculate the weighted-loads of tested parameters in WQI. In total, five PCs with loads above 1 and a cumulative variance of about 61% (see Table 5.1) were observed. Then, the parameters of each PC that were correlated significantly with others and had the lowest loading's among them were eliminated (see Fig. 5-4). Finally, the WQI includes only non-correlated parameters with loadings greater than 0.3 to the contributed PCs. The Varimax rotation was used for PCA calculation and it helped to reveal the PCs

with the exact chemical properties of water, which were clearly interrelated and showed specific types of pollution. As an example, the chemical indicators usually linked with organic pollution were coupled to PC3, whereas parameters of water mineralization were coupled to PC1 (please see Table 5.1)

Table 5.1: Chemical components loading attributed to each PCs based on the PCA with Varimax rotation

Principal components	**Comp1**	**Comp2**	**Comp3**	**Comp4**	**Comp5**
Eigenvalues	6.116	2.057	1.856	1.543	1.237
Variance (%)	29.12	9.79	8.84	7.35	5.89
Cumulative variance (%)	29.12	38.92	47.76	55.10	61.00
Parameters loadings					
NH_4	0.0794	0.0041	**0.5602**	0.0279	-0.0603
HCO_3	-0.0363	**0.5385**	0.0041	0.0229	0.0137
Alkalinity	-0.0364	**0.5386**	0.0041	0.0228	0.0136
pH	-0.1731	**0.3074**	0.2065	-0.0889	-0.1959
Hardness of water	0.2960	0.2583	-0.1245	-0.0123	0.0035
Cr	0.0076	-0.0764	-0.0718	**0.5049**	0.1270
Cu	-0.1188	0.0103	0.0489	0.2093	**0.4262**
Fe	-0.0179	0.0199	-0.0408	**0.6504**	-0.0269
Mn	0.0557	0.0913	0.1145	**0.4557**	-0.1452
Ni	0.2217	-0.1376	-0.0030	-0.1010	-0.0475
Zn	-0.0368	0.1017	-0.1915	0.0638	0.1721
SO_4	0.1987	-0.0145	-0.1570	-0.0894	**0.3695**
Cl	**0.5033**	-0.1380	0.0726	0.0079	-0.1002
NO_3	0.0666	-0.1398	-0.0800	-0.1048	**0.5048**
NO_2	0.0518	-0.0645	0.1705	0.1495	0.0442
PO_4	0.0223	-0.0059	**0.6047**	-0.0642	0.1163
Mineralization	**0.3729**	0.1255	0.0228	-0.0215	0.1407
Ca	0.2973	0.2457	-0.1414	-0.0098	-0.0152
Mg	0.2552	0.2604	-0.0634	-0.0169	0.0540
Na	**0.4440**	-0.0817	0.1863	0.0101	-0.0330
K	-0.1235	0.1455	0.2777	-0.0010	**0.5150**

In fact, each PC contributed to a series of chemical parameters in the tested dataset. For example, the PC1 was linked to the chloride content, overall mineralization and sodium content of water (with loadings greater than 0.3). However, all three of these parameters were correlated: r(Na & Cl) = 0.856; r(Cl & Mineralization) = 0.819; and r(Na & Mineralization) = 0.800. Thus, the final shortlisted parameters from these PCs were a subset of the co-correlated parameters to prevent overlooked results and include only Cl. A similar case with co-correlated parameters



(a) (b)

Figure 5-4: The correlation heatmap for chemical parameters in tested freshwater samples. Figure [A] present correlation coefficient) between all measured chemical parameters, while figure [B] present correlation coefficient only for parameters with significant PCA loading. Initial number of water quality parameters for WQI constriction was reduced from twenty-one to fifteen after PCA.

was observed in parameters attributed to PC2. The PC2 revealed three main characteristics of water pollution: hydrocarbonates (HCO_3), alkalinity and pH. At the same time, only HCO_3 & Alkalinity was characterized by r as 1.0, while two other parameters revealed low values of co-correlaton r(pH & HCO_3) = 0.227, r(pH & Alkalinity) = 0.228 and were included in the shortlisted parameters. All correlations among significant parameters for PC3, PC4 and PC5 were low (see Fig. 5-4); thus, all parameters with $loads > 0.3$ (Table 5.1) were used for the WQI calculation. In detail, r(NH4 & PO4) = 0.437 in PC3, r(Cr & Fe) = 0.336, r(Cr & Mn) = 0.097 and r(Mn & Fe) = 0.353. The last PC, PC5, consisted of four significant parameters with extra low co-correlations: r(Cu & SO4) = 0.0642, r(K & NO) = 0.1376, r(K & SO4) = 0.1637, r(Cu & NO3) = 0.0571, r(SO4 & NO3) = 0.2970 and r(Cu & K) = 0.1769.

The resulting WQI is a combination of 12 parameters with different normalized weighted factors:

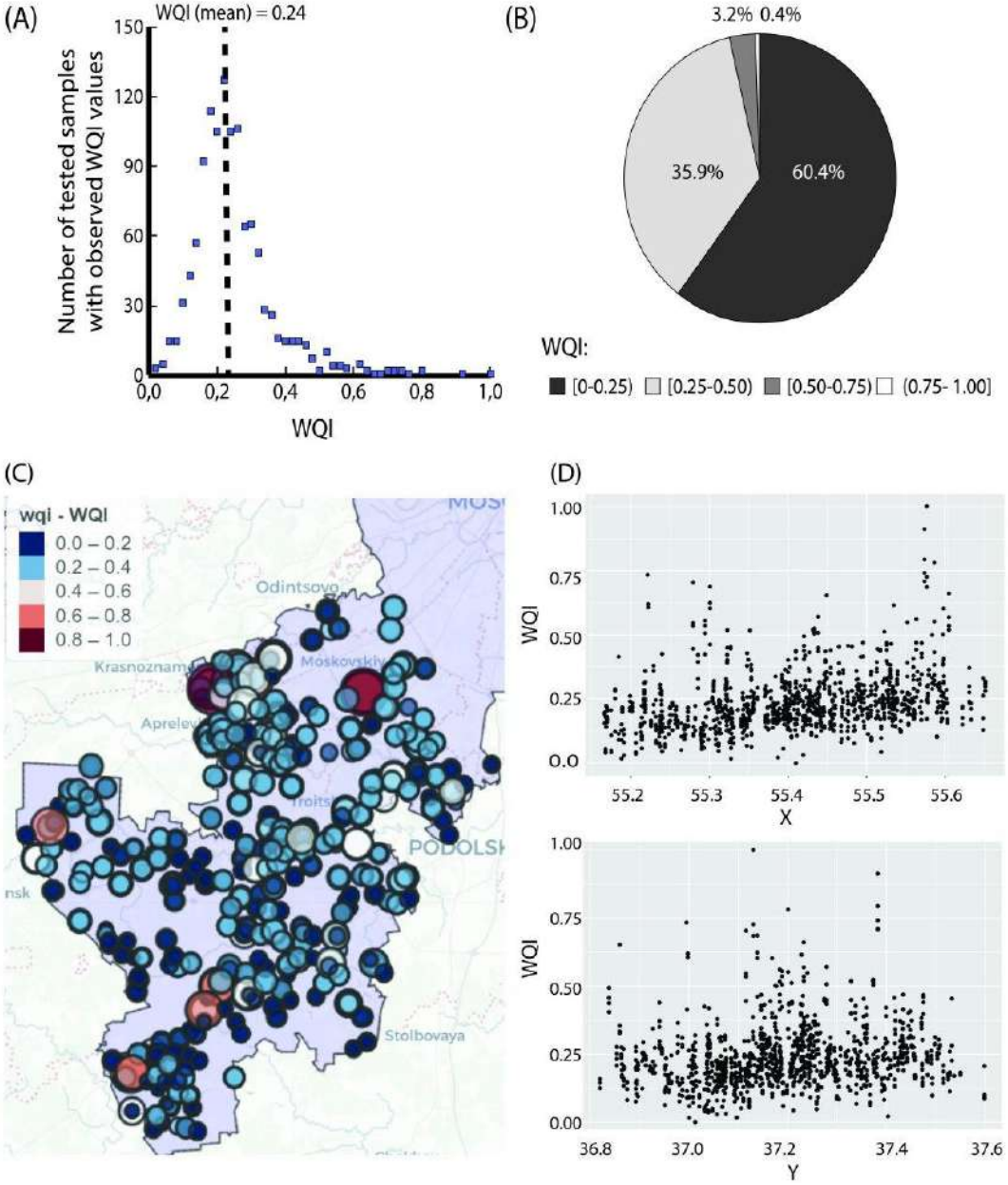


Figure 5-5: The overall distribution of WQI in tested samples. [A] The graph presents the number of tested samples with observed WQI and the mean value of the WQI. [B] Pie chart of statistical distribution of WQI for tested samples. [C] Distribution of points with estimated WQI across the study area, lower WQI values are corresponding to good groundwater quality, and higher – to poor groundwater quality. [D] Ratio of WQI to spatial coordinates: X – Latitude, Y – Longitude.

$$WQI = 0.2912 \cdot (Cl) + 0.0979 \cdot (pH + Alkalinity) + 0.0884 \cdot (NH_4 + PO_4) + \\ + 0.0735 \cdot (Cr + Fe + Mn) + 0.0589 \cdot (Cu + SO_4 + K + NO_3) \tag{5.11}$$

Table 5.2: The optimal kernel parameters for the tested Gaussian kernel with periodical kernels.

Parameter	**Value**
Gaussian kernel variance, θ_4	0.0367
Gaussian kernel length scale, l	4.86
Periodic kernel variance, θ_5	0.0204
Periodic kernel period, T	5.67
Periodic kernel length scale, s	0.1

The distribution of the calculated WQIs among the tested samples is presented in Fig. 5-5. The mean WQI was 0.24 in the tested locations, and the median was 0.22. These values signalled that less than 0.4% of the tested samples were actually characterized as highly polluted, with a $WQI > 0.75$. Distribution of WQIs across the spatial coordinates – latitude and longitude – does not show any significant trends.

BIC method for geospatial modelling vs ordinary and universal kriging. The technique based on GPR and kernel structure selection using BIC was proposed and validated. The optimal kernel structure obtained by the BIC method was found to be a sum of Gaussian and periodic kernels (see Eq. (5.12)). The optimized hyperparameters can be found in Table 5.2 with $\ell_1 = \ell_2 = \ell$, $s_1 = s_2 = s$, $T_1 = T_2 = T$ (isotropic case).

$$k^*(x, x') = \theta_4 \exp\left(-\frac{1}{2}\sum_{i=1}^{2}\frac{(x_i - x_i')^2}{\ell^2}\right) + \theta_5 \exp\left(-\frac{1}{2}\sum_{i=1}^{2}\frac{1}{s}\sin^2\left(\frac{\pi}{T}(x_i - x_i')\right)\right) \tag{5.12}$$

To validate the proposed approach and to compare it to baseline methods (OK and UK with different kernels), the 5 different random splits scheme (90% and 10% train/test split) was applied. Table 5.3 shows the corresponding R^2 and $RMSE$ values obtained for different validation splits. It can be seen that optimal kernel selection gave the best R^2 compared to the standard kriging methods. RMSE for obtained model was comparable to other methods. However, the standard deviation of errors on different validation data subsets was minimal compared to the

other approaches which make the proposed method beneficial. In the case of RMSE assessment, it is also important to compare the obtained RMSE values with the average value for WQI in the tested dataset. As can be seen, the average value of WQI was 0.24 (Figure 5-5A). Therefore, the calculated $RMSE = 0.065$ indicates that the proposed GPR model coupled with BIC is suitable for modelling. Finally, Figure 5-6 shows the results of geospatial modelling of WQI values and the corresponding uncertainty maps, obtained with different approaches. The results clearly demonstrate the advantages of automatic kernel selection using BIC, allowing to recognize the local pollutant areas.

Table 5.3: Results of the obtained models on 5 random validation splits.

		1	2	3	4	5	mean	std
Kriging with BIC	R^2	0.729	0.487	0.609	0.641	0.702	0.637	0.098
approach	RMSE	0.060	0.072	0.071	0.062	0.059	0.065	0.0063
Ordinary Kriging	R^2	0.580	-0.075	0.599	0.625	0.575	0.461	0.300
gaussian kernel	RMSE	0.068	0.076	0.056	0.060	0.059	0.064	0.0085
Universal Kriging	R^2	0.610	0.014	0.604	0.646	0.622	0.499	0.271
exponential kernel	RMSE	0.070	0.077	0.056	0.060	0.058	0.064	0.0088
Universal Kriging	R^2	0.544	-0.052	0.600	0.631	0.590	0.463	0.289
gaussian kernel	RMSE	0.071	0.076	0.055	0.059	0.058	0.064	0.0093
Universal Kriging	R^2	-11.205	-9.316	-11.042	-6.693	-9.860	-9.623	1.820
polynomial kernel	RMSE	0.129	0.113	0.109	0.097	0.103	0.110	0.0122
Universal Kriging	R^2	0.415	-0.038	0.579	0.637	0.593	0.437	0.278
periodic kernel	RMSE	0.080	0.076	0.057	0.059	0.058	0.066	0.0114

Discussion

PCA-weighted approach in WQI construction. The WQI was proposed for the first time in 1965 by [Horton, 1965]. The implementation of weighted factors for quality index construction is currently becoming very popular in environmental science. This procedure was applied earlier in the environmental sustainability index [Esty et al., 2005] and the Langat River water quality index [Mohd Ali et al., 2013]. Nevertheless, the large diversity of approaches of WQI construction shows a list of vulnerabilities of the idea, and many details remain unclear: the high diversity of types of water resources on the global scale that cannot be described by the same measure, the consequent diverse number of parameters used, and, finally, the high

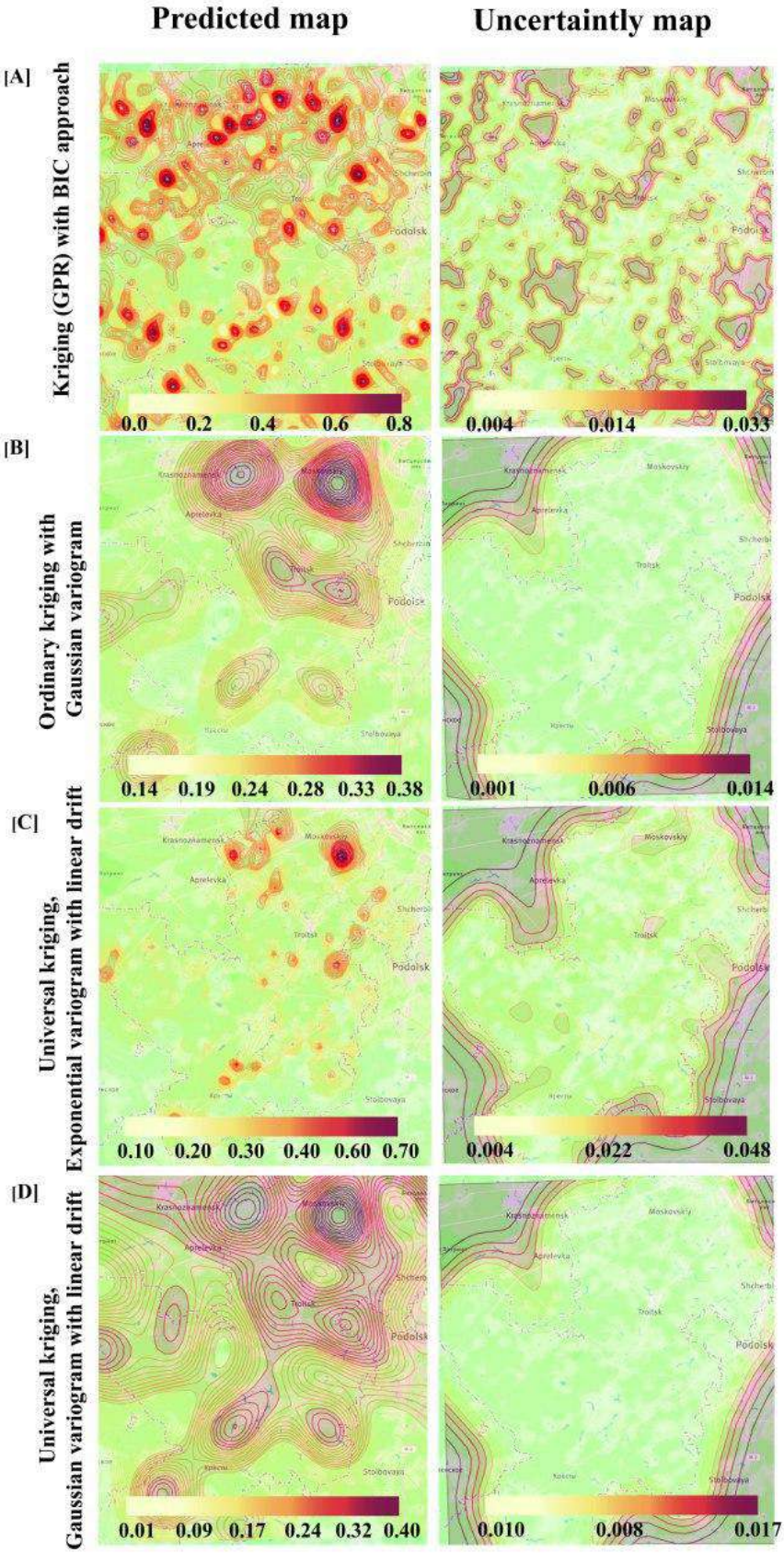


Figure 5-6: Geospatial prediction of Water quality index and uncertainty maps based on different techniques: A - GPR coupled with BIC; B - Ordinary kriging with Gaussian variogram; C - Universal kriging, Exponential variogram+linear drift; D - Universal kriging, Gaussian variogramm+linear drift.

level of subjectivity [Sun et al., 2016]. For these reasons, most existing WQIs are not universal and may be used only in case studies [Tyagi et al., 2013]. As a suggestion, the WQI should be based on an algorithm including parameters with maximum loads into general variability (thus, excluding subjectivity being adaptive).

The proposed WQI, which involves the most influential parameters, allows to model the environmental situation in the investigated area, thus including these features into agricultural modeling. Obviously, this simplification is a logical step toward the description such of complicated object as water resources , and it is convenient for use in both scientific and practical applications. PCA-based approach helps to reveal the 12 crucial parameters of water quality (Cl, pH, Alkalinity, NH_4, PO_4, Cr, Fe, Mn, Cu, SO_4, K, and NO_3) instead of the 25 parameters initially measured, basing on 1569 sampling points. For example, pH is a crucial water-quality parameter that affects water chemistry, including alkalinity, speciation and solubility.

A similar PCA-weighted approach for the water quality index, where authors applied the PCA with Varimax rotation to select the most important features of water quality and reduced the original dataset from 13 parameters to 9, was proposed by [Tripathi and Singal, 2019a]. Authors used all important features of water quality; however, unlike the proposed approach, they included even correlated parameters, which in practice led to an overestimation of the final values.

It can be highlighted at least one possible disadvantage of the proposed approach, which may be connected with the data size required for PCA. For example, in [Hutcheson and Sofroniou, 1999] it was recommended that at least 150 cases are needed to obtain satisfactory results in using this method. At the same time, not every study of environmental parameters assessment includes more than 150 collection points due to high installation, operational, and maintenance costs for each sampling representative of the whole environmental conditions (as an examples, [Ouyang, 2005, Chen and Han, 2018]).

Automatic approach to geospatial mapping. It was proposed the improvement of the Gaussian process regression based on the Bayesian information criteria

for automatic kernel structure search. The proposed approach allows to model the distribution of water quality precisely and to detect multiple local foci of the deviations, compared with commonly used ordinary kriging and universal kriging, which were inaccurate and ineffective for this particular problem (Fig. 5-6).

The proposed approach permits determination of the most realistic spatial distribution of the WQI due to the application of the algorithm for automatical kernel construction, which consists of the basic, non-linear kernels. Actually, when it comes to the end-to-end implementation in operational data processing chains, like geospatial modelling, it is mandatory to invest in models that are both accurate and robust but also require minimal user intervention for fitting parameters. An automatic kernel search helps to solve the problem of manual hyper-parameter and kernel structure selection. According to the performed validation, the developed model showed lack of overfitting and provided an accurate prediction on the test dataset according to the used metrics. Recently, similar approach for automatic kernel selection, was used successfully in several cases, e.g. the estimation of chlorophyll-a concentrations from remote sensing data, delineation referents of city centres by topographic data, soft-sensor modelling for algal bloom monitoring [Gómez-Chova et al., 2011, Lüscher and Weibel, 2013, Wang et al., 2014]. However, to date, it has not been transferred to geospatial modelling.

Conclusions

An end-to-end high accurate framework that allows to automatically model the geospatial distribution of ecological factors that have the effect on crop yields was developed. This approach states the clear methodology from the step of initial data pre-processing and detection of the driving factors from PCA to the automatic kernel search for geospatial mapping. The feasibility and robustness of the proposed methodology in the case of water quality estimation in the newly added territories to Moscow was shown. The novel approach of an automatic kernel structure search was adapted and applied in this framework, and this approach allows to achieve detailed results for geological modelling, compared with ordinary and universal kriging methods. Overall, the developed methodology opens wide possibilities for solving

similar problems for parameters distribution modeling in the most accurate and efficient way.

5.3 Machine learning approaches for phytotoxicity effects assessment

5.3.1 Machine learning methods to predict acute phytotoxicity in petroleum contaminated soils

Introduction

Crude oil contamination, arising from oil production and transporting procedures, has a devastating impact on the surrounding terrestrial ecosystems entailing agricultural production and thus human health [Larive, 2008, Khan et al., 2018]. Crude oil includes various aliphatic and aromatic hydrocarbons, which are rich in petroleum hydrocarbon and non-hydrocarbon compounds, yet are deficient in any nutritional elements. Since total petroleum hydrocarbons (TPH) structurally belong to the group of complex chemical compounds that are either found within crude oil they can be used as a means of measuring soil contamination with crude oil. The level of TPH in polluted soils largely depends on soil properties and can rise almost by 10 times as compared to the background level. What is more important, TPH input to the soil may trigger immediate changes in the environment and thus induce adverse biological and ecological effects [Garcia et al., 2019, Hunt et al., 2019]. Among different soil quality indicators, acute phytotoxicity is a conventional yet very efficient method to assess the extent to which soil is polluted [Gerber et al., 2017]. It is often measured by calculating the seed germination inhibition, root growth inhibition, or any other adverse effects on plants. Several studies demonstrated TPH contamination to be closely related to soil phytotoxicity [Molina-Barahona et al., 2005, Kanarbik et al., 2014, Kaur et al., 2017]. The mechanism by which TPH may induce soil phytotoxicity is rather complex as the effects produced by various environmental factors and soil properties are synergistic. Therefore, it is essential to

perform an accurate simulation of the threefold relationship between plant response, TPH content, and physicochemical factors. Thus, quantifying and predicting TPH phytotoxicity in the soil is a crucial issue for soil planning and remediation, resulting in the improvement of the crop yield. Meanwhile, experimental measurements assessing the correlation between soil properties and phytotoxicity of the contaminants are time-consuming and complicated, and, what is more, it seems impossible to test all the naturally existing variants with utmost accuracy. Hence, the quest to find a universal approach to determine a non-linear correspondence between bio-indicators and the physicochemical data observed in the soil is necessary. These demands are met by machine learning techniques (ML) (that demonstrated a rapid surge of interest to designing models to simulate and predict soil processes Goodarzi et al. [2016], Olawoyin [2016], Cipullo et al. [2019], Sayyad Amin et al. [2019]). Modern supervised machine learning methods, such as support vector machine (SVM) and artificial neural networks (ANN), are considered to be promising and efficient tools aimed at interpreting high-dimensional and high-nonlinear data in environmental science. Previously, SVM was used to predict spatial distribution of soil organic carbon, soil nitrogen stock [Kou et al., 2019], and soil salinity [Wu et al., 2018]. ANNs were successfully applied to model soil physical properties in case of temperature fluctuations [Ozturk et al., 2011] and erosion [Gholami et al., 2018]. It was useful in determining soil chemical properties [Fernandes et al., 2019] as well as soil biological activity [Jha and Ahmad, 2018, Ebrahimi et al., 2019]. The main aim of the following research was to show that a few observations and measurements of the soil properties provide the possibility to predict their phytotoxicity effect on the plants based on contaminants.

Materials and methods

Top soils samples from eleven field sites on the Sakhalin island were brought to the laboratory, where they were thoroughly mixed and quartered (see Table A.3)). The total weight of one wet soil sample was 4 kg. The soils were stored in airtight containers at a room temperature before being analysed. The soil samples were classified according to WRB [Chesworth, 2007].

The experiment was conducted in greenhouse conditions (22±1°C). Crude oil was characterized by the bulk density 0.83 g cm^{-3}, 2.6% of the $C_5 - C_9$ fraction, 21.3% of the $C_9 - C_{15}$ fraction, 2.6% of the $C_{16} - C_{20}$, 72% of the asphaltene and mazut fractions, and low content of sulfur (148 mg kg^{-1}) [Kovaleva et al., 2017]. Different doses of crude oil (from the 3.0 to 100.0 g kg^{-1}) were manually added to the soils based on the expert opinion about their possible sorption capacity to TPH (https://doi.org/10.6084/m9.figshare.9114638.v2). Each toxicity test consisted of 6 treatments — one control treatment and five increasing oil concentrations — each conducted in triplicate.

Barley toxicity tests. Spring barley (*Hordeum vulgare L.*) is one of the most essential grain crops [Arendt and Zannini, 2013]. According to the international standard ISO 11269-1, *H. vulgare L.* is also recommended for bioassay investigations. The seeds were pre-sterilized by orbital agitation with 70% ethanol for 2 min and then with 5% sodium hypochlorite adding several drops of Tween 80 for 30 min. They were rinsed six times and sterilized in distilled water. All seeds were transferred in sterile conditions into 10 mm Petri dishes containing the studied soils (20 g of dry mass for organo-mineral soils and 10 g of dry mass for peat soils) covered with a filter paper (Ø 90 mm Whatman #1) soaked with distilled water to achieve the moisture level equal to 60% of the soil water holding capacity. The Petri dishes were taped and placed in the dark at 21°C for five days under the same conditions to evaluate root elongation. Finally, barley root elongation was calculated either as the average mean, or as the sum of all roots emerged from each seed. The experiments were performed three times.

Soil properties measurements. The contents of soil organic carbon by Walkley-Black protocol, soil total nitrogen by ISO 11261, soil total phosphorus and potassium were determined, soil in water 1:5 extract by pH-meter. Particle-size distribution in soil samples was measured by the aerometric method. TPH in soil was determined by the gas chromatography technique (GC-FID) in GC 6890N Agilent Technologies with a flame ionization detector.

Statistical analysis and machine learning. Before performing statistical analysis, normality of the data were checked by using the Kolmogorov-Smirnov test, and data conversion was applied in cases where they were not normally distributed. Treatments were replicated for three times. The data were processed following the analysis of variance; the means of treatments were compared using Tukey's minimum significant difference test at the 0.05 probability level. Two approaches were evaluated - support vector regression (SVR) and artificial neural networks (ANN) - to predict root lengths using *Python*. In the course of the research the following packages were used: *Scikit-learn Python* library, as well as *Keras* and *Tensorflow* libraries for building and training neural networks. See Fig. 5-7 for more details on the pipeline of the implemented approach to predict soil phytotoxicity:

The dataset was randomly split into training and validation datasets using the *train test split* method from *Scikit-learn Python* library. The train sub-dataset consisted of 172 randomly selected records, which accounted approximately for 80% of the total amount. The test sub-dataset included other 44 records, comprising the remaining 20% of records.

Support vector machine. Support vector regression is a learning regression algorithm developed from the Support vector machine; the mathematical formulation of SVR is provided and discussed in detail by [Drucker et al., 1997]. The strength of SVR is the ability of the model to establish complex nonlinear relationships in the multidimensional or hyper-dimensional feature space. Training dataset was used for initial fit of the SVR with the Gaussian kernel (Radial basis function - RBF) in Python, while the optimal hyper-parameters of the model were obtained by solving an optimization problem (minimization of the RMSE) by varying of the hyperparameters C and γ. The parameter search space was a priori set to $0.01 \leqslant C \leqslant 10000$ and $0.01 \leqslant \gamma \leqslant 50$. A logarithmic grid for parameters search space was used. This approach for SVR modelling is widely presented in the literature as example of a successful application of the method to the problems of multidimensional regression in various research fields [Hjorth, 2017].

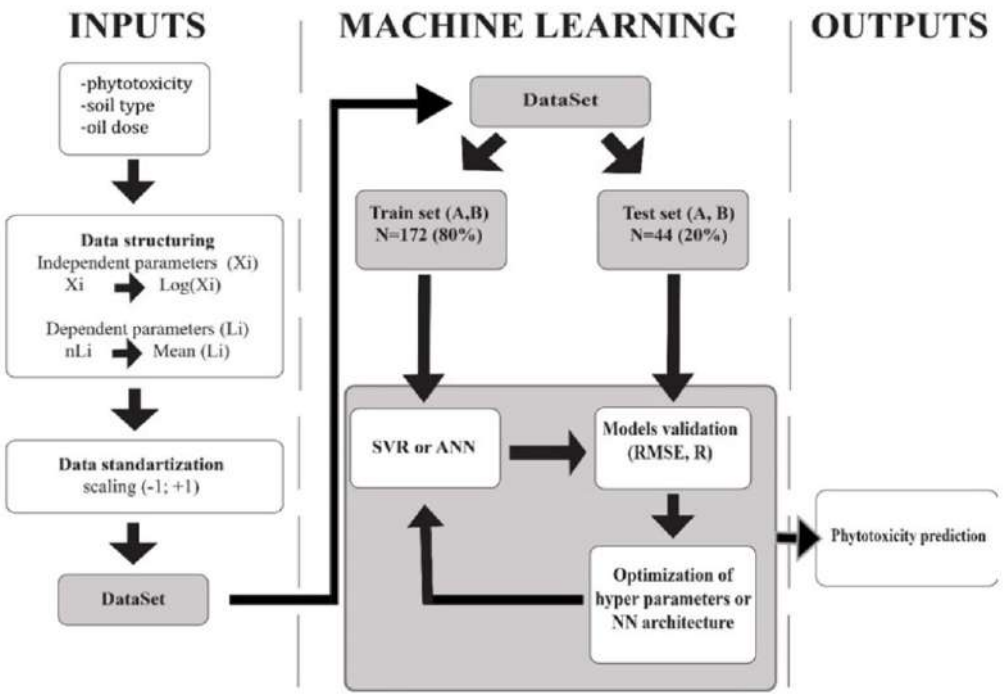


Figure 5-7: Workflow for using machine learning methods for TPH phytotoxicity prediction.

Neural network design and training ANN modelling. For the purposes of modelling, it was used one hidden fully connected layer and several input and hidden layers consisting of different amounts of neurons: from 16 to 256 ones were tested. A rectified linear unit (ReLU) was used as an activation function, exhibiting strong biological and mathematical underpinning [Hahnloser et al., 2000]. The neural network architecture and its training procedure were implemented in Python using the Keras library with the computational core TensorFlow. In the process of learning, the adam algorithm was used instead of the classical stochastic gradient descent to increase the efficiency of calculations [Kingma and Ba, 2014]. Mean squared error was used as a loss function.

Performance evaluation. Traditionally, several metrics are used to assess the accuracy of ML models. The quality of the trained model was evaluated using the root mean square error (RMSE) (5.10) as well as mean absolute error (MAE) (5.13):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{5.13}$$

where n is the number of training compounds, $\hat{y}_i$ and y_i are the estimated and observed responses, respectively. Being independent of the response scale, contrary to RMSE, the coefficient of determination R^2 (5.9) is also considered as a useful metrics for accuracy.

Results of modeling

Changes in physicochemical and biological properties of the soil under crude oil pollution. The soils used for this study exhibited a wide range of soil properties (Table A.3) and belonged to the main types of zonal and intrazonal soils of the Sakhalin island. Soil pH ranged from 4.30 at Soil #3 (Carbic Podzol) to 5.65 at Soil #5 (Livic Stagnosols Dystric), while in seven samples pH was lower 5.5. The LOI varied greatly: the maximum organic carbon was at Soil #1 (Fabric Histosols Dystric) 97.18%, while Soil #9 (Umbric Fluvisols Oxyaquic) had the lowest values <2%. Clay content ranged from 0 to 28.20%. The initial concentrations of TPH in soils ranged between 0.3–3.0 g kg^1; the mean barley root length in the controls was from 39 to 109 mm; the longest roots detected in the soils were with a high content of organic carbon and NPK. Injection of crude oil significantly increased TPH content in the soils (Data represented in https://doi.org/10.6084/m9.figshare.9114638), but phytotoxicity effects produced on barley varied greatly among different soil types and crude oil treatments (see Fig. 5-8). As compared to the controls, the marked phytotoxicity was measured in soils #9 (Umbric Fluvisols Oxyaquic), #10 (Haplic Cambisols Dystric) and #11 (Umbric Fluvisols Oxyaquic) with doses of 20 and 30 g kg^{-1} of crude oil in the soils. A slight hormesis effect (i.e. the stimulation of response at low doses followed by the inhibition at high doses) was observed in several soils (e.g. Rustic Podzols, Carbic Podzols, Histic Podzols, etc), but it was significant ($p < 0.05$) only for soil #5 - Luvic Stagnosols Dystric (see Fig. 5-8). Being applied to various soils, crude oil affected barley in such factors as intensity

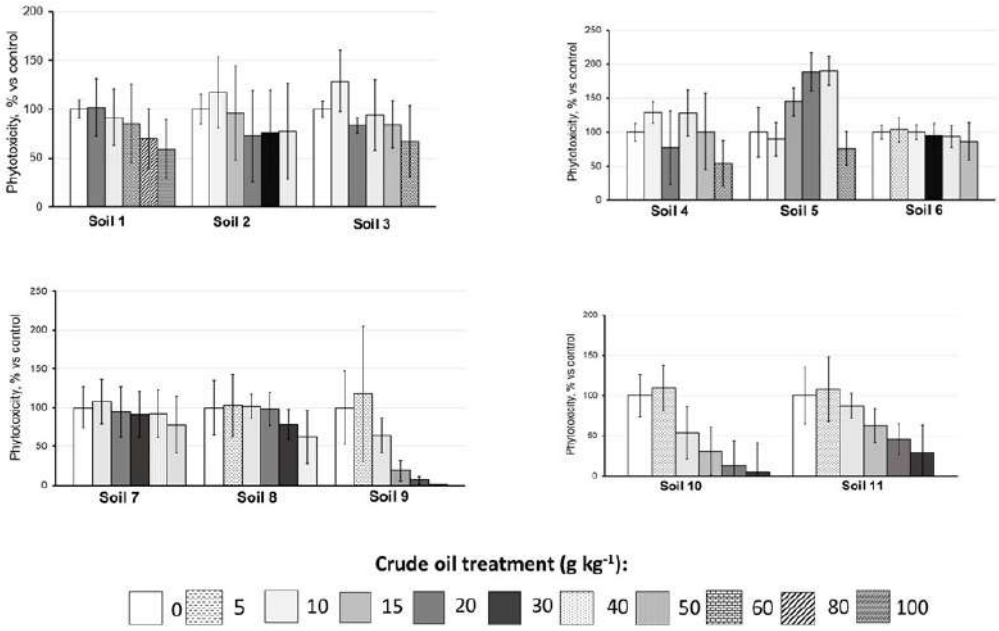


Figure 5-8: The effects different crude oil treatments (g/kg) produce on barley root lengths phytotoxicity in different soils; the error bars represent a standard deviation of the mean (n = 90). Soils description by WRB: 1 - Fibric Histosols Dystric; 2 - Rustic Podzols; 3 - Carbic Podzols; 4 - Histic Podzols; 5 - Luvic Stagnosols Dystric; 6 - Histic Gleysols Dystric; 7 - Fibric Histosols Eutric; 8 - Umbric Fluvisols Oxyaquic; 9 - Umbric Fluvisols Oxyaquic; 10 - Haplic Cambisols Dystric; 11 - Umbric Fluvisols Oxyaquic.

and direction depending on the type of the soil. Key factors influencing the root length were determined by agglomerative hierarchical clustering and principal component analysis; and presented in Fig. 5-9. For example, a high content of organic matter may determine the bioavailability of TPH in the soils, because of their high affinity to organic contaminants. This statement has been proved for soil samples #1, 2, 3, 4 (see Fig. 5-8, 5-9). Since these soil samples were characterized by a high value of LOI, even extremely high doses of crude oil (>30 g kg^{-1}) appeared to be less phytotoxic than other samples with a lower content of LOI. Nevertheless, other complementary factors also tended to affect the root lengths as in the case of other soil samples. For instance, PCA revealed that in soil samples #3 and #4 barley response was determined not only by TPH content, but also by initial NPK values.

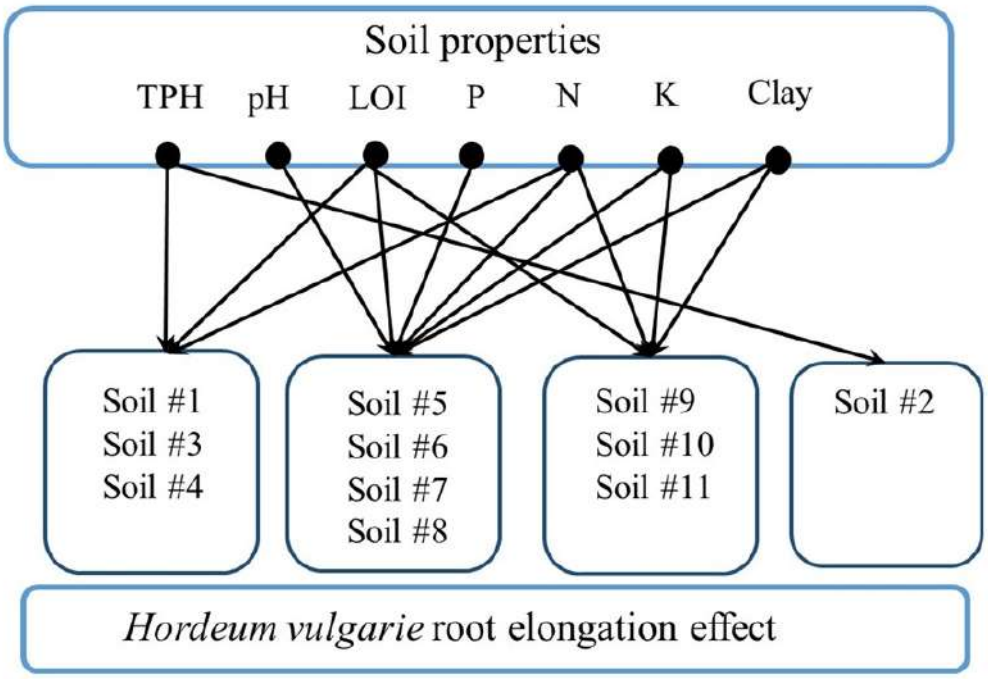


Figure 5-9: Drivers for barley root lengths depend largely on the tested soils according to agglomerative hierarchical clustering and principal component analysis

Development of predictive toxicity models

Support vector regression performance. SVR method was trained to predict barley phytotoxicity (mean root length) depending on the basic soil characteristics and TPH content. RBF kernel was used in the SVR algorithm. C and γ kernel parameters varied to produce a model with optimal performance and were cross-validated using subsets A and B, which included 44 random records from the training dataset. The desired level of accuracy was controlled by measuring the average mean of RMSE and R^2 with different values of C and γ. Figure 5-10 shows the plots with predicted vs actual values for root lengths for two-cross validation with different combinations of hyperparameters C and γ. Figure 5-11 summarizes in a heatmap the calculations of R^2 and RMSE on a grid of C values (varying from 0.01 to 10 000) and γ values (from 0.01 to 50). Judging from the data in Fig. 5-10 and 5-11, it is obvious that the lowest RMSEs were achieved with the hyperparameters C = 900, $\gamma = 1$ (Fig. 5-10A) and C = 900, $\gamma = 0.1$ (Fig. 5-10F). Thus, it was concluded that by setting C = 900 and varying γ from 0.1 to 1, RMSE values close to 11.0 with a

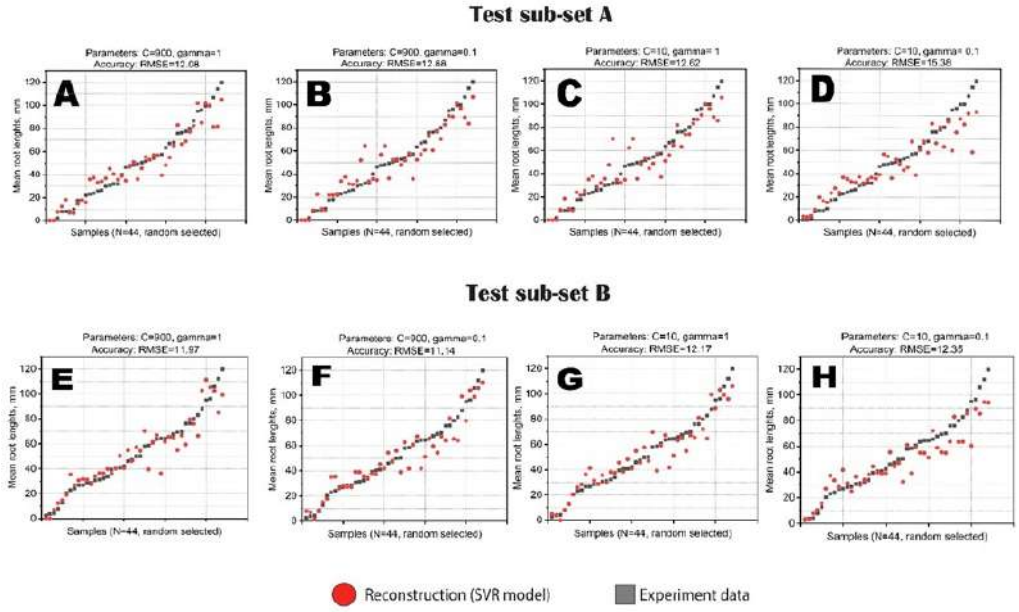


Figure 5-10: Results of the reconstruction of the mean root length obtained by applying the SVR method to two-test sub-sets A and B with different hyperparameters.

determination coefficient R^2 of 0.8 could be achieved.

Artificial neural networks performance. To predict soil phytotoxicity using artificial neural network models under different TPH concentrations in soils, soil physicochemical and phytotoxicity properties were used as input data, as in case of SVR. After determining the complexity of training and testing data, ANN was designed with different numbers of neurons in the hidden and input layers. Then the optimum structure of the networks was determined using the RMSE and MAE criteria. Figure A-1 gives an overview of ANN model performance with a different amount of neurons in the hidden layer and 128 neurons in the input layer (testing sub-sets). It can be seen that the prediction accuracy of ANN is close to that of SVR (RMSE varies in the range from 11.14 to 15.38), and does not change significantly depending on the properties of the sample, though the number of hidden neurons was expected to have an effect on ANN performance. The accuracy of the tuned ANN models using RMSE metrics ranges from 11.05 to 14.79, whereas the MAE metrics ranges between 8.44 and 12.10. It should be highlighted that architectures with a

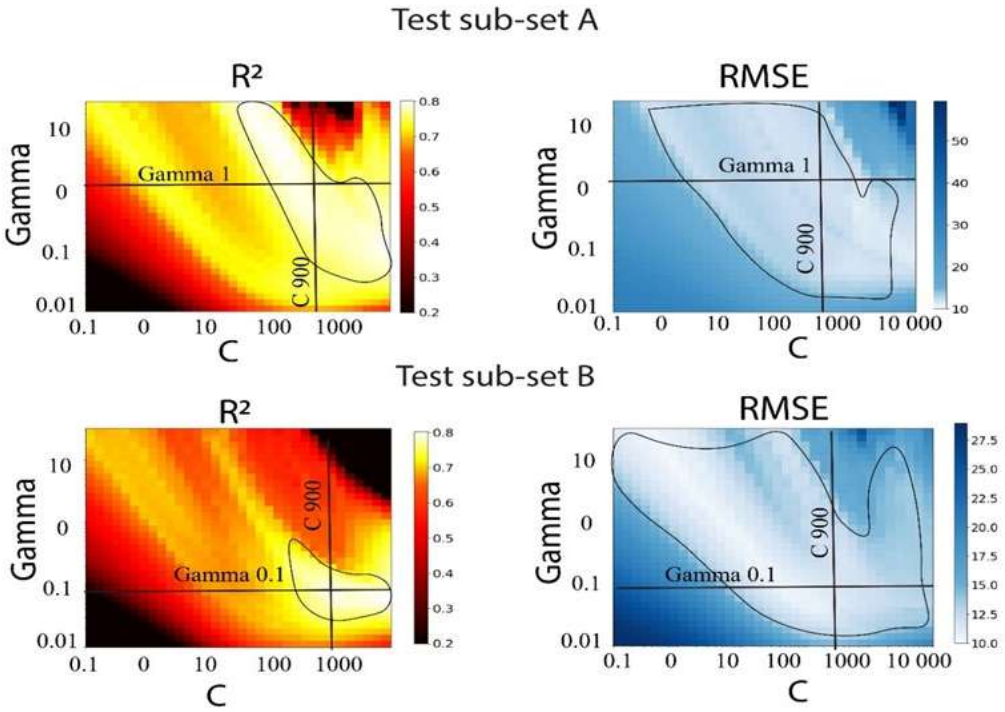


Figure 5-11: Results of the calculations of the determination coefficient R^2 and the root mean squared error RMSE on the grid of SVR model hyper parameters γ and C for two different test sub-sets A and B. The contour features out the best areas with the highest R^2 and the lowest value RMSE. Perpendicular lines correspond to the coefficients of Fig. 4A and 4F. Note: Attention. There are different scales of RMSE values for test sub-set A and B.

similar amount of neurons in the hidden and input layer turned out to have larger RMSE and MAE values than other combinations (see Fig. A-1). The best prediction with the lowest RMSE 11.05 and MAE 8.44 was achieved with the architecture of 256 neurons in the input layer and 64 in the hidden one.

Discussion

Plant bioassay is an effective and popular tool for estimating the soil quality and assessing ecotoxicological risks [Ghosh et al., 2017]. Rapid-cycling plants like barley (*Hordeum vulgarie*) allow to reveal a short-term influence from organic and inorganic contaminants in the soil [Kim et al., 2019, Nikolaeva et al., 2019]. According to the ISO 11269-1:2012, root elongation may be a crucial endpoint for ecotoxicological assessment. Normally, elongation tests imply plant development during a short 5-day timespan. This period corresponds to the imbibition of dry seeds and an intensive water uptake for the radicle protrusion through seed covering layers and roots development, so at this stage plants are usually sensitive to contaminants [Weitbrecht et al., 2011]. In this study the focus was on standardized plant toxicity endpoints in different soils polluted by TPH. While phytotoxicity assessment is a time-consuming method, requiring the development of advanced prediction models and getting accustomed to soil types and pollutants, it was assumed that SVR and ANN may be promising as the candidates for phytotoxicity prediction in different soils [Cipullo et al., 2018].

The main types of zonal and intrazonal soils obtained from different parts of Sakhalin island were studied. For each soil sample it was conducted a greenhouse experiment to test TPH phytotoxicity of barley planted in differently TPH-treated soils. As a result, it was received and analyzed a dataset comprising 11 types of soils and described them according to barley phytotoxicity response, TPH content, and some chemical properties. Many researchers previously reported significant inhibitory effects due to the elevated dose of TPH in the early-stage plants growth. Masakorala et al. [2013] detected a 50% and 97% reduction in *Lactuca sativa L.* root lengths at 1% and 3% TPH, respectively, in freshly-contaminated soils. The TPH contaminated soil usually had significant phytotoxic effect on the root growth even

at low contamination level (0.5% according to [Gaskin et al., 2008]. This problem was previously singled out for different kinds of heavy metals or TPH pollution [Said et al., 2019]. According to the results Fig. 5-8 obtained, the traditional dose-response curve approach seems to be unsuitable for data analysis because of the non-linear response of root elongation under increasing TPH content and exciting hormesis effects in some dose of crude oil in several soils. Notably, the effect of early stimulation following TPH adding was observed in [Kirk et al., 2002, Shahsavari et al., 2013]. Overall, the high variability of this dataset poses a challenge to predict the dose-dependent relationships and barley response using traditional approaches. The ML methods are very advantageous in this case because ANN and SVR can perform nonlinear regression efficiently for high dimensional datasets.

Recently, several models have been developed to assess the quantity relationships for various toxicity responses on TPH of terrestrial plants, microorganisms, and soil fauna [Hentati et al., 2013, Bori et al., 2016, Tran et al., 2018, Cruz et al., 2019, Soroldoni et al., 2019]. These relationships were defined either for soils with different properties or those from different countries. At the same time, it is possible that the developed models and dependencies on these data may not be proved on the samples from other regions because of differences in soil physicochemical properties. The deviation of the modelled soil phytotoxicity from the measured ones across the machine learning models is shown in Fig. 5-12, where ANN and SVR predicted values were plotted against the average measured values for the root length in a randomly split dataset. The SVR model had a higher correlation and a lower RMSE than other models, while ANN models were approximately equal in this regard. Both groups of models were characterized by a suitable accuracy for predicting TPH phytotoxicity effect on barley root length in direct soil bioassay tests. The SVR model achieved a sufficient performance accuracy for both predicted and measured values of phytotoxicity; using the combination of hyper parameters C=900 and $\gamma = 0.1$ it was achieved R^2 of about 0.80 and RMSE values of about 11-12 (see Fig. 5-11). ANN showed the minimal calculated $RMSE = 11.05$ and $MAE = 8.44$ with 256 inputs and 64 hidden neurons, respectively (see Fig. A-1). Previously, a similar approach was used observed for soil heavy metals prediction [Sergeev et al., 2019],

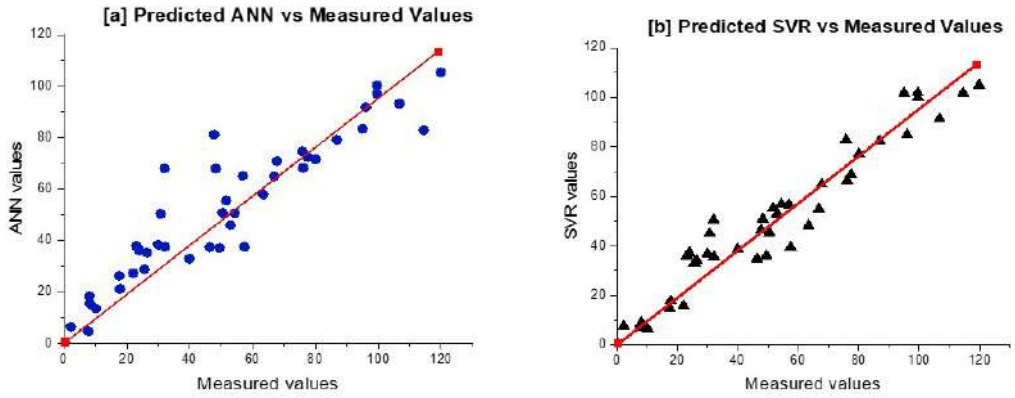


Figure 5-12: Measured and estimated root lengths for test subset-A of ANN (A) and SVR (B) models. Red lines stand for 1:1 line.

and soil salinity prediction [Pouladi et al., 2019]. These results indicate that both models could be effectively adapted in the course of training and to be further used as a general model for predicting TPH phytotoxicity in soils.

Conclusions

This study shows the complexity and significance of the investigation of phytotoxicity effect of contaminated soils. The modeling of the phytotoxicity effects was demonstrated on the example of the TPH soils contamination. As it was revealed, the range of the effect depend largely on the properties of the soil. In most tested soils, adding TPH in the doses over 15 g kg^{-1} tended to increase the barley root lengths and induced marked phytotoxicity. Nevertheless, in some cases the addition of TPH positively affected the growth of the roots as compared to the test with a non-polluted control. It was demonstrated the benefits of applying the machine learning approach to the prediction of TPH phytotoxicity in eleven types of soils. In the course of research, SVR and ANN algorithms were trained based on the data received from the greenhouse experiment. As a result, the best performance was detected for SVR model with R^2 - 0.8 and RMSE - 11. This predictive model can definitely provide additional information about the quality of the soil on a regional scale what it turn can improve the agricultural management and lead to increasing crop yield.

5.3.2 Machine learning methods for assessment and prediction of phosphogypsum influence on soil

Introduction

Usage of wastes for agricultural needs is becoming more and more popular. However, the effect such utilization is not always unambiguous. Waste production is an increasing global concern that is projected to worsen with the accelerating world's population. To be adduced just as an example, in the Russian Federation more than 31.5 billion tons of waste were accumulated and identified by 2016 (including 140 million tons of phosphogypsum (PG)) and 100 million tons were landfilled [Report, 2015]. Given the large quantities that are produced, and keeping in mind that only 14% of PG is used in the construction industry, it is necessary to dispose the surpluses [Tayibi et al., 2009]. For instance, the land application of PG in agricultural fields could be an important recycling alternative aiming to reduce landfilling sites [Saadaoui et al., 2017].

The application of PG as an amendment has generally shown a positive effect on soil chemical properties, including an increase in the available sulphur and phosphorus content, improvement of soil structure and crop yield [Carmeis Filho et al., 2017, Kammoun et al., 2017]. Furthermore, PG amendment is recommended in ameliorating salinity in damaged soils, providing a source of Ca to replace the excess Na in cations' exchange [Hurtado et al., 2011]. However, there are several difficulties in expanding the use of PG for an agronomy purpose, due to its complex structure.

Only a fraction of ecotoxicological studies have been performed to evaluate the ecological impact of PG application on soil. PG information is particularly fragmentary especially regarding their inclusion of trace element pollutants and other compounds as its specific composition and characteristics change considerably depending on the geographical origin. This waste typically comprises mainly gypsum and phosphate, but may also include the potentially hazard elements, such as fluoride, strontium, barium. The presence of the latter at high levels in PG may have hazardous impact on the soil in general and on humans and plants, in particular. Pollutants from PG may adversely affect the soil environment by retarding the plant

growth, and enhancing the soil toxicity [Ayadi et al., 2015, Yakovlev et al., 2013, Hentati et al., 2013].

The use of supervised ML methods trained on empirical data could be advantageous to make predictions on the potential toxicity effects of exogenous substances in soil [Deng et al., 2017, Cipullo et al., 2019], properties of drug-like molecules [Palmer et al., 2015], and biomonitoring the pesticide toxicity [Zhu et al., 2018, Niell et al., 2018]. ML models are able to learn the relationships between input variables (e.g. soil amendment, soil type) and output variables (e.g. changes in soil toxicity, or bioassay response) from a training dataset, these relationships can then be generalized to make informed decisions in new cases. The interest to ML methods definitely rises, especially when we deal with soil systems, because the traditional statistical extrapolation techniques do not fit well in case of complex environment [Jager, 2011, Fox, 2015]. Overall, it can concluded that the application of ML to environmental issues (that are closely related to the precision agricultural domain) is the latest cutting edge research direction.

Materials and methods. Greenhouse experiment includes different PG doses (0, 1%, 3%, 7.5%, 15%, 25% and 40%) and two times-collection points after treatments - 7 and 28 days. For each treatment and each times-collection point it was measured: i) soil pH, bioavailable (H_20 and NH_4COOH-extractable) element content (S, P, K, Na, Mg, Ca, Fe, Zn, Sr, Ba, F); ii) soil enzyme activities – dehydrogenase, urease, acid phosphatase, FDA; iii) soil CO_2 respiration activity with and without glucose addition; iv) *Eisenia fetida*, *Sinapis alba* and *Avena sativa* responses. The dataset that was used for simulations is available via the link: `https://doi.org/10.1007/s11368-019-02253-2`. The ordinary chemical, toxicology and biological measuring of soil properties was combined with state-of-the-art mathematical analysis, namely: i) support vector machines (used for prediction); ii) mutual information test (variable importance tasks).

Mutual information test. The investigation of the factors relevance was carried out by mutual information test [Kraskov et al., 2004]. This method studies probabilistic dependencies between the target vectors and considered factors. These

complementing measures can be useful to analyze the data from different points of view. In contrast to correlation analysis, this test allows identifying nonlinear relations between given factors and target vectors. Moreover, it gives a degree of dependencies for every pair of considered factors or between factors and a target vector. These degrees help to eliminate the most redundant and the least relevant factors. The elimination can be based on some threshold number of required factors or the threshold value of mutual information score. In this study, with a help of mutual information test, the load of individual chemical variables in biological and ecotoxicity responses was practically assessed. The most important chemical features that were identified by mutual information algorithm for the estimation of the PG biological and ecotoxicological influence in soil are represented in Fig. 5-13. The heatmap (5-13) shows that different features from PG were dominated for each biological and ecotoxicity variables.

According to the experimental results, among the 9 measured soil biological and ecotoxicity variables, only URE and SIR soil activities, *S. alba* root lengths and *E.fetida* survival rate were negatively affected by PG treatments [Pukalchik et al., 2019]. The results obtained from mutual information tests suggest, that only a few elements may be dependent on the exacerbating effects of PG on the mentioned variables, in particular: F(w), P(m), Sr(m) have the greatest load on earthworm's toxicity; S(m), Ca(m) and Na(m) influenced *S. alba* root lengths toxicity; Ca(m), Ba(w) and Sr(m) mostly affected on soil URE activities; finally, F(w), P(w) and P(m) affected in SIR values. These results were supported by the mutual information scores (see Fig. 5-13). Previous studies highlighted the key role of the exacerbated F, Sr and P soil content in toxicity to earthworms. In particular, fluorine and strontium may have led cytotoxicity effects, and phosphorus addition with fertilizer may also induced earthworm's mortality [Chae et al., 2018]. The sulfur, calcium and sodium phytotoxicity effects may be connected with their possible accumulation in roots and ion relations effects [Negrão et al., 2017]. Inhibition activities of barium and strontium to soil URE activities were earlier observed by [Tabatabai, 1977]. The lack of effect of PG on soil enzyme activities like AP, FDA and DHA, looks controversial. However, it provides the evidence in favor of the sensitivity of these enzymes to soil

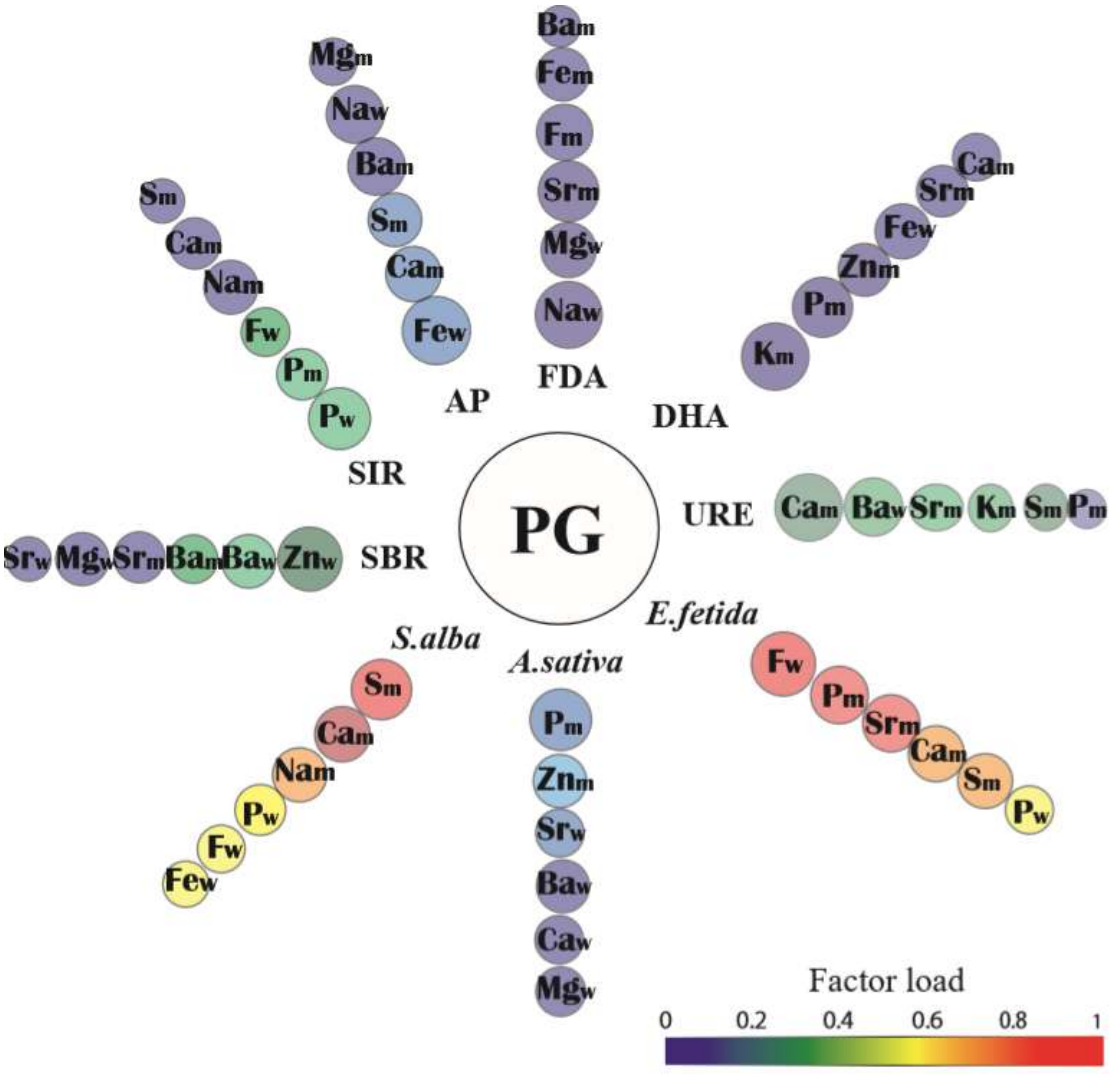


Figure 5-13: Influence of measured chemical elements in soil after PG addition to soil biological and toxicological responses from the mutual information test. Balls are coloured according to calculated load (from 0 to 1): the higher values coloured in read, and the lowest values - in blue.

contamination that could be overvalued for soil monitoring purposes. In general, the enzyme activities are considered to be the first to respond to soil contamination; due to their high sensitivity to react to environmental changes. Moreover, they play a fundamental role in the dynamics of C, N, P, S which in turn have an effect on plant growth dynamics [Caldwell, 2005]. However, obtained results in general make it possible to assume that a high amount of fertilizer elements could interfere the effect of trace elements on hydrolysis enzymes. As it can be seen from the mutual information scores shown in fig. 5-13, the P, K, Na, Mg and S addition has the highest load in AP, DHA and FDA responses among all the other elements. Thus, it was concluded that the chosen machine learning techniques are useful to further studies in the issues in questions and potentially help elucidate quite 'in-obvious' relations.

Support vector machine. In this research, the dataset was randomly split into a training and validation datasets with different ratios (90% or 116 observations, 70% or 88 observations, and 50% or 63 observations were used as a training datasets, 10%, 30% and 50% observations - as a text dataset). The input data (all measured chemical, biological and ecotoxicity data) were log-transformed prior to model development, biological and ecotoxicity data were scaling from 0 to 100 scale in compare with NA control samples.

The training dataset was initially used to fit the SVR model with the Gaussian kernel function and the optimal model's hyper parameters were obtained by solving an optimization problem (minimization of the RMSE) on a grid of hyper-parameters: C and γ. The parameter search space was a priori set to $0.001 \leqslant C \leqslant 1000$ at an incremental ratio of 10 and $0.001 \leqslant \gamma \leqslant 0.3$ at steps of 0.001. After training, the derived SVR models were applied to the validation datasets to produce the apparent soil biological and ecotoxicity properties. The performance of SVR modeling was evaluated by the root mean squared error (RMSE) (5.10).

The trained SVR model was used to predict the soil biological and ecotoxicity properties in the presence of different PG doses. Table 5-14 shows the performance indicators for SVR-1, SVR-2 and SVR-3 models with varying input size on training

	DHA	URE	AP	FDA	SBR	SIR	*E.fetida*	*S.alba*	*A.sativa*
SVR-1. Training dataset 116 observations and validation datasets 12 observations									
RMSE	8.64	7.04	12.15	8.13	4.25	6.18	9.19	6.01	8.57
SVR-2. Training dataset 88 observations and validation datasets 38 observations									
RMSE	9.35	10.37	11.64	11.91	6.13	11.48	12.33	8.10	8.12
SVR-3. Training dataset 63 observations and validation datasets 63 observations									
RMSE	13.55	11.94	12.97	11.93	11.16	12.16	12.50	9.84	12.64

Figure 5-14: Influence of training set size on SVR models performance to predict soil biological and ecotoxicity properties after PG addition

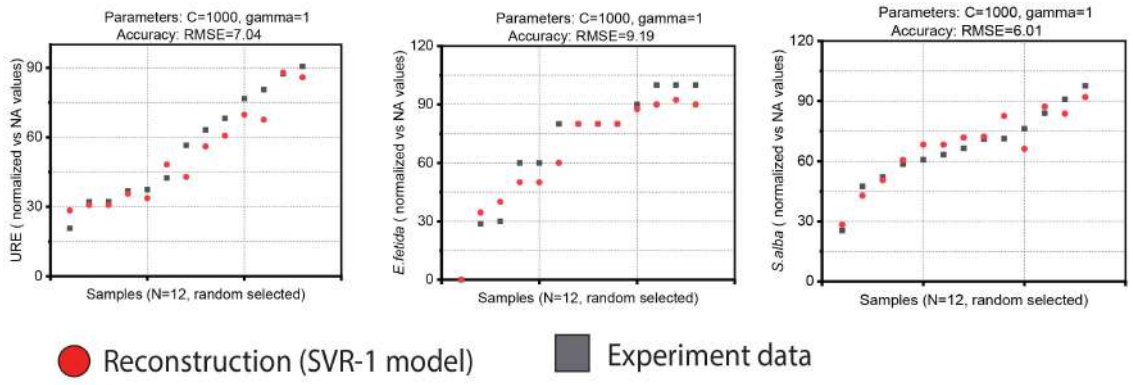


Figure 5-15: Prediction accuracy for the selected soil toxicity data using SVR-1 model

datasets (90%, 70% and 50% of the input data were used). The RMSE values ranged from 4.25 to 12.15 for SVR-1 model provide better accuracy for modeling response parameters than SVR-2 and SVR-3 models. Figure 5-15 shows predicted and experimental values for the selected biological and ecotoxicological parameters (as an example the URE, *S.alba*, and *E.fetida* were chosen) based on the SVR-1. As can be seen, visual correlation between measured and predicted values for the random selected samples were satisfactory.

Models based on biological indicators could become a powerful tool in soil ecotoxicology and could help to reduce the amount of analysis needed to adequately monitoring soil systems quality [Cipullo et al., 2019]. The results of SVR performances revealed that model prediction ability consistently improved with the increasing size of training sets. The SVR model was able to predict the toxicity and biological prop-

erties with adequate accuracy only in case when 90% of received data were used as a training dataset (see Fig. 5-15). When the training dataset was reduced to 70 or 50% of experimental data the accuracy of modeling dramatically decreased (see Table 5-14). A similar influence of training set size on SVM-based prediction have already been investigated the phosphogypsum toxicity using the simple linear probit model [Rodríguez-Perez et al., 2017, Yakovlev et al., 2013, Hentati et al., 2015]. In particular, no observed effect concentration of PG in soil were determined from 1.24% (F. candida) to 24.61% (*E. crypticus*), and no toxic effect was detected for *Zea mays*, and *Lactuca sativa* up to 25% PG was descrribed in [Hentati et al., 2015]. According to [Yakovlev et al., 2013] the most sensitive indicator of an ecosystem stress for PG application was a microbial respiration activity, and the calculated (not observed) effect concentration was 10.8% in artificial soil.

Conclusions. The proposed approach to identify the PG influence in soil with advanced ML models looks beneficial in comparison with previous studies, which can be explained by better applicability of the available knowledge because it relied on both qualitative data (biological and ecotoxicological properties) and quantitative data (chemical properties of soils with different doses of PG) for models training. Keeping in mind that the relationship among pollutants and even the chemical composition of waste is highly nonlinear and very complex, it was mandatory to use more accurate analysis tools based on statistical learning such as the support vector regression. It was noted that size of training datasets significantly influenced the SVR-models performance and even a "small" amount of data could be enough to train SVM models. At the time when the ecological monitoring programs are declining in a cost-effective manner and it is not always possible to receive a "Big Data" in soil environment, the usage of ML methods may be a promising candidate tools to prevent soil degradation and contamination, thus addressing problems for effective agricultural land fertilizers usage.

5.4 Conclusions

In this chapter a set of data-driven methods for more accurate modeling of the environmental parameters was proposed. This gives the opportunity to include modeled parameters into existing model-based algorithms for yield prediction to improve them. It was proposed and evaluated end-to-end methodology for improving accuracy of prediction of spatial distribution on the environmental parameters. The method is based on the Gaussian process regression with an automatical kernel structure search based on Bayesian information criteria. This method was tested on the proposed water quality index which in turn plays a crucial role in the irrigation process defining the crop yield. The developed methodology for creating of water quality index that describes integral characteristics of pollution can be applied to other environmental parameters. The results of modeling showed the improving accuracy of prediction of spatial distribution of parameter compared to standard techniques as well as more automated process of modeling. The other techniques described in this chapter are dedicated to prediction of toxicity effects of soil contamination. There is no precise method up to day that allows to model the toxicity effects of soil contamination. It was demonstrated that toxicity effects of such complex systems as contaminated soils can be assessed precisely using ML techniques. In particular, the advantages of using of ML techniques for prediction of toxicity effects were shown on the problem of TPH phytotoxicity assessment. Also, the possibilities of ML methods to identify quantitative effect of addition of a fertilizer such as phosphogypsum into soil were demonstrated. The results of modeling are more accurate compared to previous studies and opens many possibilities to analyze soil properties and important factors that were unavailable before.

Overall, the proposed and evaluated ML methods for modeling the environmental parameters and their effects on growth showed lots of advantages compared to previous studies allowing to include these parameters into existing methodologies for plant growth prediction and taking into account effects that were not considered before. Application of these approaches can be easily extended to greenhouses giving the opportunity to perform accurate modeling of spatial distribution of envi-

ronmental parameters and effects of different fertilizers on plant growth dynamics.

Chapter 6

Conclusion

The present thesis focuses on data-driven and hybrid methods to propose a set of solutions to solve one of the most crucial problems of modelling plant growth in the controlled environment to achieve a sustainable food production. The research emphasizes universality and robustness of the proposed approaches as the main feature and considers different aspects of plant growth: seed germination process, vegetation stage of growth and disease detection.

To carry out the study most effectively novel automated artificial growth systems equipped with a sensor and non-invasive machine vision systems were developed and constructed. Using these systems unique and relevant datasets were collected to describe plant growth dynamics and environmental conditions. Leaves area (projected), being the main target parameter, was estimated using a self-developed accurate computer vision methods. These datasets were used for the following validation of the proposed methods. The set of the proposed methods was divided into two groups based on the principals of applying the method and limiting its implementation. The former is the hybrid approach which includes data-driven and model-based methods. Such approaches such as Kalman filter, dynamic mode decomposition, merging 2D/3D data, fully convolutional neural networks were adapted and implemented to the obtained dataset. The research has shown that Kalman filter has a high computational efficiency without reducing the accuracy, which cannot be achieved using more straightforward methods. Applying dynamic mode decomposition in couple with a set of differential equations showed that physical principles

could be included to modelling; this method also showed a high accuracy on a small train dataset, though it appeared to require fine-tuning. Instance segmentation method used along with other CV methods allowed us to track and predict growth dynamics of each leaf, which could be used for a detailed plant growth dynamics assessment. Merging 2D/3D promised the prediction of the plant's biomass basing on 2D data.

To enhance the hybrid methods, pure data-driven approaches, such as recurrent neural networks and convolutional neural networks, were also proposed to be used for growth dynamics assessment and prediction. Recurrent neural networks were implemented to projected leaves area prediction. The dataset used for validation was collected on the own small-scale experimental setup equipped with machine vision and sensors systems. The result of the modelling showed accurate, long-term predictions of the projected leaves area. After that, machine vision and sensing systems were deployed into an industrial experiment, in the course of which a huge dataset describing growth dynamics and growing conditions was collected. Using this dataset FCNNs were trained to perform segmentation tasks for an automatic leaves area (projection) calculation. Basing on sequences of images the growth dynamics of plants was reconstructed. In addition, the biomass of the plants was measured, which allowed us to find the dependencies between the projected leaves area and the biomass and to predict the biomass using the obtained 2D images. The thesis also proposes a methodology for seed germination rate assessment; it is based on CNNs to propose the regions presumably containing seeds and on CV techniques to perform a quantitative analysis of germinated seeds in the proposed region. The methodology was evaluated on a constructed experimental setup. Another vital topic discussed in this thesis in sustainability of plant growth. The research proposes the methodology for finding optimal wavebands in the infrared spectrum to detect diseases at the early stages. This method was evaluated on the own obtained dataset and was proved to be of a high practical usefulness.

Finally, a research to model environmental parameters in field conditions that play a crucial role in plant development was conducted. Application of machine learning methods for modeling of the spatial distribution of highly variable en-

vironmental parameters was proposed. The proposed techniques improve current state-of-the-art results on modeling of environmental parameters. Noteworthy, the approaches initially developed for open systems could be relevant for greenhouses as well, as they can be easily transferred to artificial systems.

To sum up, all the proposed and evaluated data-driven and hybrid techniques showed high accuracy, universality, and robustness for solving plant growth assessment and modeling tasks. These methods are highly useful for industrial applications. The developed experimental setups and the collected relevant datasets could be used in further evaluation of newly appeared methods for growth dynamics assessment and prediction.

Among the future objectives of the present research the following steps are planned: to test the proposed methods on a larger variety of plants, to include to the study of other environmental parameters, such as photosynthesis, and to preform more industrial deployments.

Bibliography

NNAPI. https://developer.android.com/ndk/guides/neuralnetworks, 2019. [Online; accessed 07-May-2019].

TensorFlow Mobile. https://www.tensorflow.org/lite, 2019. [Online; accessed 07-May-2019].

Mahyar Aboutalebi, Alfonso F Torres-Rua, and Niel Allen. Multispectral remote sensing for yield estimation using high-resolution imagery from an unmanned aerial vehicle. In *Autonomous Air and Ground Sensing Systems for Agricultural Optimization and Phenotyping III*, volume 10664, page 106640K. International Society for Optics and Photonics, 2018.

Edmundo Acevedo, Theodore C Hsiao, and DW Henderson. Immediate and subsequent growth responses of maize leaves to changes in water status. *Plant Physiology*, 48(5):631–636, 1971.

Amir H. Alavi, Pengcheng Jiao, William G. Buttlar, and Nizar Lajnef. Internet of things-enabled smart cities: State-of-the-art and future trends. *Measurement*, 129:589 – 606, 2018. ISSN 0263-2241. doi:https://doi.org/10.1016/j.measurement.2018.07.067.

Joan Albiol, Jordi Robusté, Carles Casas, and Manel Poch. Biomass estimation in plant cell cultures using an extended kalman filter. *Biotechnology progress*, 9(2): 174–178, 1993.

Bashar Alhnaity, Simon Pearson, Georgios Leontidis, and Stefanos Kollias. Using deep learning to predict plant growth and yield in greenhouse environments. *arXiv preprint arXiv:1907.00624*, 2019.

Hossein Alilou, Alireza Moghaddam Nia, Mohsen Mohseni Saravi, Ali Salajegheh, Dawei Han, and Bahram Bakhtiari Enayat. A novel approach for selecting sampling points locations to river water quality monitoring in data-scarce regions. *Journal of hydrology*, 573:109–122, 2019. doi:10.1016/j.jhydrol.2019.03.068.

Moustafa Alzantot, Yingnan Wang, Zhengshuang Ren, and Mani B Srivastava. Rstensorflow: Gpu enabled tensorflow for deep learning on commodity android devices. In *Proceedings of the 1st International Workshop on Deep Learning for Mobile Systems and Applications*, pages 7–12. ACM, 2017.

Nan An, Christine M Palmer, Robert L Baker, RJ Cody Markelz, James Ta, Michael F Covington, Julin N Maloof, Stephen M Welch, and Cynthia Weinig. Plant high-throughput phenotyping using photogrammetry and imaging techniques to measure leaf length and rosette area. *Computers and Electronics in Agriculture*, 127:376–394, 2016.

Victor H Andaluz, Andrea Y Tovar, Kevin D Bedón, Jessica S Ortiz, and Edwin Pruna. Automatic control of drip irrigation on hydroponic agriculture: Daniela tomato production. In *Automatica (ICA-ACCA), IEEE International Conference on*, pages 1–6. IEEE, 2016.

Elke K Arendt and Emanuele Zannini. *Cereal grains for the food and beverage industries*. Elsevier, 2013.

Danny Awty-Carroll, John Clifton-Brown, and Paul Robson. Using k-nn to analyse images of diverse germination phenotypes and detect single seed germination in miscanthus sinensis. *Plant methods*, 14(1):5, 2018.

Amal Ayadi, Amal Chorriba, Amine Fourati, and Radhia Gargouri-Bouzid. Investigation of the effect of phosphogypsum amendment on two arabidopsis thaliana ecotype growth and development. *Environmental technology*, 36(12):1547–1555, 2015.

Xingzhen Bai, Zidong Wang, Li Sheng, and Zhen Wang. Reliable data fusion of hierarchical wireless sensor networks with asynchronous measurement for greenhouse monitoring. *IEEE Transactions on Control Systems Technology*, 27(3):1036–1046, 2018.

Saraansh Baranwal, Siddhant Khandelwal, and Anuja Arora. Deep learning convolutional neural network for apple leaves disease detection. *Available at SSRN 3351641*, 2019.

Rahim Barzegar, Asghar Asghari Moghaddam, Jan Adamowski, and Amir Hossein Nazemi. Assessing the potential origins and human health risks of trace elements in groundwater: A case study in the khoy plain, iran. *Environmental geochemistry and health*, 41(2):981–1002, 2019. doi:10.1007/s10653-018-0194-9.

Jan Behmann, Kelvin Acebron, Dzhaner Emin, Simon Bennertz, Shizue Matsubara, Stefan Thomas, David Bohnenkamp, Matheus T Kuska, Jouni Jussila, Harri Salo, et al. Specim iq: evaluation of a new, miniaturized handheld hyperspectral camera and its application for plant phenotyping and disease detection. *Sensors*, 18(2): 441, 2018.

Pedro Bello and Kent J Bradford. Single-seed oxygen consumption measurements and population-based threshold models link respiration and germination rates under diverse conditions. *Seed Science Research*, 26(3):199–221, 2016.

Elisabeth Berger, Peter Haase, Mathias Kuemmerlen, Moritz Leps, Ralf Bernhard Schaefer, and Andrea Sundermann. Water quality variables and pollution sources

shaping stream macroinvertebrate communities. *Science of the Total Environment*, 587:1–10, 2017. doi:10.1016/j.scitotenv.2017.02.031.

Nikolai Bessonov and Vitaly Volpert. *Dynamic models of plant growth*. Editions Publibook, 2006.

Scott Billings. Industrial hydroponic control apparatus, March 8 2018. US Patent App. 15/256,585.

Jaume Bori, Bettina Vallès, Lina Ortega, and Maria Carme Riva. Bioassays with terrestrial and aquatic species as monitoring tools of hydrocarbon degradation. *Environmental Science and Pollution Research*, 23(18):18694–18703, 2016.

Jonas Bühler, Louai Rishmawi, Daniel Pflugfelder, Gregor Huber, Hanno Scharr, Martin Hülskamp, Maarten Koornneef, Ulrich Schurr, and Siegfried Jahnke. phenovein—a tool for leaf vein segmentation and analysis. *Plant Physiology*, 169(4): 2359–2370, 2015.

Bruce A Caldwell. Enzyme activities as a component of soil biodiversity: a review. *Pedobiologia*, 49(6):637–644, 2005.

Zachary C Campbell, Lucia M Acosta-Gamboa, Nirman Nepal, and Argelia Lorence. Engineering plants for tomorrow: how high-throughput phenotyping is contributing to the development of better crops. *Phytochemistry reviews*, 17(6):1329–1343, 2018.

Sebastian Candiago, Fabio Remondino, Michaela De Giglio, Marco Dubbini, and Mario Gattelli. Evaluating multispectral images and vegetation indices for precision farming applications from uav images. *Remote sensing*, 7(4):4026–4047, 2015.

Antonio CA Carmeis Filho, Chad J Penn, Carlos AC Crusciol, and Juliano C Calonego. Lime and phosphogypsum impacts on soil organic matter pools in a tropical oxisol under long-term no-till conditions. *Agriculture, Ecosystems & Environment*, 241:11–23, 2017.

Yooeun Chae, Dokyung Kim, and Youn-Joo An. Effects of fluorine on crops, soil exoenzyme activities, and earthworms in terrestrial ecosystems. *Ecotoxicology and environmental safety*, 151:21–27, 2018.

Ayan Chaudhury, Christopher Ward, Ali Talasaz, Alexander G Ivanov, Norman PA Huner, Bernard Grodzinski, Rajni V Patel, and John L Barron. Computer vision based autonomous robotic system for 3d plant growth measurement. In *2015 12th Conference on Computer and Robot Vision*, pages 290–296. IEEE, 2015.

Yiheng Chen and Dawei Han. Water quality monitoring in smart city: A pilot project. *Automation in Construction*, 89:307–316, 2018. doi:10.1016/j.autcon.2018.02.008.

Ward Chesworth. *Encyclopedia of soil science*. Springer Science & Business Media, 2007.

Anna Chlingaryan, Salah Sukkarieh, and Brett Whelan. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture*, 151:61–69, 2018.

Woo Jae Cho, Hak-Jin Kim, Dae Hyun Jung, Dong Wook Kim, Chang Ik Kang, and Gyeong Lee Choi. An embedded system for control of hydroponic nutrients. In *2015 ASABE Annual International Meeting*, page 1. American Society of Agricultural and Biological Engineers, 2015.

Sabrina Cipullo, G Prpich, P Campo, and F Coulon. Assessing bioavailability of complex chemical mixtures in contaminated soils: Progress made and research needs. *Science of the Total Environment*, 615:708–723, 2018.

Sabrina Cipullo, Boris Snapir, George Prpich, P Campo, and Frederic Coulon. Prediction of bioavailability and toxicity of complex chemical mixtures through machine learning models. *Chemosphere*, 215:388–395, 2019.

Lucian Codrescu, Willie Anderson, Suresh Venkumanhanti, Mao Zeng, Erich Plondke, Chris Koob, Ajay Ingle, Charles Tabony, and Rick Maule. Hexagon dsp: An architecture optimized for mobile multimedia and communications. *IEEE Micro*, 34(2):34–43, 2014.

RHR Costa, CT Zanotelli, DM Hoffmann, P Belli Filho, CC Perdomo, and M Rafikov. Optimization of the treatment of piggery wastes in water hyacinth ponds. *Water Science and technology*, 48(2):283–289, 2003.

Jaqueline Matos Cruz, Nádia Aline Corroqué, Renato Nallin Montagnoli, Paulo Renato Matos Lopes, Maria Aparecida Marin Morales, and Ederio Dino Bidoia. Comparative study of phytotoxicity and genotoxicity of soil contaminated with biodiesel, diesel fuel and petroleum. *Ecotoxicology*, 28(4):449–456, 2019.

Jeffrey A Cruz, Xi Yin, Xiaoming Liu, Saif M Imran, Daniel D Morris, David M Kramer, and Jin Chen. Multi-modality imagery database for plant phenotyping. *Machine Vision and Applications*, 27(5):735–749, 2016.

Andre Daccache, Jerry W Knox, EK Weatherhead, Alireza Daneshkhah, and TM Hess. Implementing precision irrigation in a humid climate–recent experiences and on-going challenges. *Agricultural water management*, 147:135–143, 2015.

Jifeng Dai, Kaiming He, and Jian Sun. Convolutional feature masking for joint object and stuff segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3992–4000, 2015.

Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016.

Abhiram Das, Hannah Schneider, James Burridge, Ana Karine Martinez Ascanio, Tobias Wojciechowski, Christopher N Topp, Jonathan P Lynch, Joshua S Weitz,

and Alexander Bucksch. Digital imaging of root traits (dirt): a high-throughput computing and collaboration platform for field-based root phenomics. *Plant methods*, 11(1):51, 2015.

N. Davies and S. Clinch. Pervasive data science. *IEEE Pervasive Computing*, 16(3): 50–58, 2017a. ISSN 1536-1268. doi:10.1109/MPRV.2017.2940956.

Nigel Davies and Sarah Clinch. Pervasive data science. *IEEE Pervasive Computing*, 16(3):50–58, 2017b.

Bert De Brabandere, Davy Neven, and Luc Van Gool. Semantic instance segmentation with a discriminative loss function. *arXiv preprint arXiv:1708.02551*, 2017.

Jianqiang Deng, Xiaomin Chen, Rui Wang, Jiangkuan Nan, and Zhenjie Du. Ls-svm data mining analysis: how does biochar influence soil net nitrogen mineralization in the field? *Journal of Soils and Sediments*, 17(3):827–840, 2017.

Stijn Dhondt, Nathalie Gonzalez, Jonas Blomme, Liesbeth De Milde, Twiggy Van Daele, Dirk Van Akoleyen, Veronique Storme, Frederik Coppens, Gerrit TS Beemster, and Dirk Inzé. High-resolution time-resolved imaging of in vitro arabidopsis rosette growth. *The Plant Journal*, 80(1):172–184, 2014.

Harris Drucker, Christopher JC Burges, Linda Kaufman, Alex J Smola, and Vladimir Vapnik. Support vector regression machines. In *Advances in neural information processing systems*, pages 155–161, 1997.

T Duan, SC Chapman, Y Guo, and B Zheng. Dynamic monitoring of ndvi in wheat agronomy and breeding trials using an unmanned aerial vehicle. *Field Crops Research*, 210:71–80, 2017.

S Ducournau, A Feutry, P Plainchault, P Revollon, Bertrand Vigouroux, and MH Wagner. Using computer vision to monitor germination time course of sunflower (helianthus annuus l.) seeds. *Seed Science and Technology*, 33(2):329–340, 2005.

David Duvenaud, James Robert Lloyd, Roger Grosse, Joshua B Tenenbaum, and Zoubin Ghahramani. Structure discovery in nonparametric regression through compositional kernel search. *arXiv preprint arXiv:1302.4922*, 2013.

Hsien Ming Easlon and Arnold J Bloom. Easy leaf area: Automated digital image analysis for rapid and accurate measurement of leaf area. *Applications in plant sciences*, 2(7), 2014.

Mitra Ebrahimi, Mohammad Reza Sarikhani, Ali Akbar Safari Sinegani, Abbas Ahmadi, and Saskia Keesstra. Estimating the soil respiration under different land uses using artificial neural network and linear regression models. *Catena*, 174: 371–382, 2019.

Ngozi Clara Eli-Chukwu. Applications of artificial intelligence in agriculture: A review. *Engineering, Technology & Applied Science Research*, 9(4):4377–4383, 2019.

Olakunle Elijah, Tharek Abdul Rahman, Igbafe Orikumhi, Chee Yen Leow, and MHD Nour Hindia. An overview of internet of things (iot) and data analytics in agriculture: Benefits and challenges. *IEEE Internet of Things Journal*, 2018.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

Daniel C Esty, Marc Levy, Tanja Srebotnjak, and Alexander De Sherbinin. Environmental sustainability index: Benchmarking national environmental stewardship. *New Haven: Yale Center for Environmental Law & Policy*, pages 47–60, 2005.

Patrick Eugster, Vinaitheerthan Sundaram, and Xiangyu Zhang. Debugging the internet of things: The case of wireless sensor networks. *IEEE Software*, (1):1–1, 2015.

Kevin G Falk, Talukder Z Jubery, Seyed V Mirnezami, Kyle A Parmley, Soumik Sarkar, Arti Singh, Baskar Ganapathysubramanian, and Asheesh K Singh. Computer vision and machine learning enabled soybean root phenotyping pipeline. *Plant methods*, 16(1):5, 2020.

Mingsheng Fan, Jianbo Shen, Lixing Yuan, Rongfeng Jiang, Xinping Chen, William J Davies, and Fusuo Zhang. Improving crop productivity and resource use efficiency to ensure food security and environmental quality in china. *Journal of experimental botany*, 63(1):13–24, 2011.

Konstantinos P Ferentinos. Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture*, 145:311–318, 2018.

Konstantinos P Ferentinos, Nikolaos Katsoulas, Antonis Tzounis, Thomas Bartzanas, and Constantinos Kittas. Wireless sensor networks for greenhouse climate and plant condition assessment. *Biosystems engineering*, 153:70–81, 2017.

Mariele Monique Honorato Fernandes, Anderson Prates Coelho, Carolina Fernandes, Matheus Flavio da Silva, and Claudia Campos Dela Marta. Estimation of soil organic matter content by modeling with artificial neural networks. *Geoderma*, 350:46–51, 2019.

Fabio Fiorani and Ulrich Schurr. Future scenarios for plant phenotyping. *Annual review of plant biology*, 64:267–291, 2013.

Fabio Fiorani, Uwe Rascher, Siegfried Jahnke, and Ulrich Schurr. Imaging plants dynamics in heterogenic environments. *Current opinion in biotechnology*, 23(2): 227–235, 2012.

S Fishman and M Génard. A biophysical model of fruit growth: simulation of seasonal and diurnal dynamics of mass. *Plant, Cell & Environment*, 21(8):739–752, 1998.

Frank Forcella, Roberto L Benech Arnold, Rudolfo Sanchez, and Claudio M Ghersa. Modeling seedling emergence. *Field Crops Research*, 67(2):123–139, 2000.

David R Fox. Selection bias correction for species sensitivity distribution modeling and hazardous concentration estimation. *Environmental toxicology and chemistry*, 34(11):2555–2563, 2015.

Grégoire T Freschet, Elferra M Swart, and Johannes HC Cornelissen. Integrated plant phenotypic responses to contrasting above-and below-ground resources: key roles of specific leaf area and root mass fraction. *New Phytologist*, 206(4):1247–1260, 2015.

Daiane Frighetto Frighetto, Gustavo Maia Souza, and Alexandre Molter. Spatio-temporal population control applied to management of aquatic plants. *Ecological Modelling*, 398:77–84, 2019.

Marina Reback Garcia, André Pereira Cattani, Paulo da Cunha Lana, Rubens César Lopes Figueira, and César C Martins. Petroleum biomarkers as tracers of low-level chronic oil contamination of coastal environments: A systematic approach in a subtropical mangrove. *Environmental pollution*, 249:1060–1070, 2019.

Sharyn Gaskin, Kathleen Soole, and Richard Bentham. Screening of australian native grasses for rhizoremediation of aliphatic hydrocarbon-contaminated soil. *International Journal of Phytoremediation*, 10(5):378–389, 2008.

Michel David Gerber, Thomaz Lucia Jr, Luciara Correa, José Eduardo Pereira Neto, and Érico Kunde Correa. Phytotoxicity of effluents from swine slaughterhouses using lettuce and cucumber seeds as bioindicators. *Science of the Total Environment*, 592:86–90, 2017.

Michel Edmond Ghanem, Hélène Marrou, and Thomas R Sinclair. Physiological phenotyping of plants for crop improvement. *Trends in Plant Science*, 20(3): 139–144, 2015.

V Gholami, MJ Booij, E Nikzad Tehrani, and MA Hadian. Spatial soil erosion estimation using an artificial neural network (ann) and field plot data. *Catena*, 163:210–218, 2018.

Pooja Ghosh, Indu Shekhar Thakur, and Anubha Kaushik. Bioassays for toxicological risk assessment of landfill leachate: a review. *Ecotoxicology and environmental safety*, 141:259–270, 2017.

Jonathon A Gibbs, Michael Pound, Andrew P French, Darren M Wells, Erik Murchie, and Tony Pridmore. Approaches to three-dimensional reconstruction of plant shoot topology and geometry. *Functional Plant Biology*, 44(1):62–75, 2017.

Mario Valerio Giuffrida, Hanno Scharr, and Sotirios A Tsaftaris. Arigan: Synthetic arabidopsis plants using generative adversarial network. In *Proceedings of the 2017 IEEE International Conference on Computer Vision Workshop (ICCVW), Venice, Italy*, pages 22–29, 2017.

Kamlesh Golhani, Siva K Balasundram, Ganesan Vadamalai, and Biswajeet Pradhan. A review of neural networks in plant disease detection using hyperspectral data. *Information Processing in Agriculture*, 5(3):354–371, 2018.

Mahmood R Golzarian, Ross A Frick, Karthika Rajendran, Bettina Berger, Stuart Roy, Mark Tester, and Desmond S Lun. Accurate inference of shoot biomass from high-throughput images of cereal plants. *Plant Methods*, 7(1):2, 2011.

Luis Gómez-Chova, Jordi Muñoz-Marí, Valero Laparra, Jesús Malo-López, and Gustavo Camps-Valls. A review of kernel methods in remote sensing data analysis. In *Optical Remote Sensing*, pages 171–206. Springer, 2011. doi:10.1002/9780470748992.

Mohammad Goodarzi, Leandro dos Santos Coelho, Bahareh Honarparvar, Erlinda V Ortiz, and Pablo R Duchowicz. Application of quantitative structure-property relationship analysis to estimate the vapor pressure of pesticides. *Ecotoxicology and environmental safety*, 128:52–60, 2016.

GPy. GPy: A gaussian process framework in python. `http://github.com/SheffieldML/GPy`, since 2012.

Christine Granier and Denis Vile. Phenotyping and beyond: modelling the relationships between traits. *Current opinion in plant biology*, 18:96–102, 2014.

Peter J Gregory, A Glyn Bengough, Dmitri Grinev, Sonja Schmidt, W Bill TB Thomas, Tobias Wojciechowski, and Iain M Young. Root phenomics of crops: opportunities and challenges. *Functional Plant Biology*, 36(11):922–929, 2009.

Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Li Wang, Gang Wang, et al. Recent advances in convolutional neural networks. *arXiv preprint arXiv:1512.07108*, 2015.

Richard HR Hahnloser, Rahul Sarpeshkar, Misha A Mahowald, Rodney J Douglas, and H Sebastian Seung. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, 405(6789):947–951, 2000.

Esmael Hamuda, Brian Mc Ginley, Martin Glavin, and Edward Jones. Improved image processing-based crop detection using kalman filtering and the hungarian algorithm. *Computers and electronics in agriculture*, 148:37–44, 2018.

Dongmei Han, Matthew J Currell, and Guoliang Cao. Deep challenges for china's war on water pollution. *Environmental Pollution*, 218:1222–1233, 2016. doi:10.1016/j.envpol.2016.08.078.

Joe J Hanan. *Greenhouses: Advanced technology for protected horticulture*. CRC press, 2017.

Ahmad Nizar Harun, Robiah Ahmad, and Norliza Mohamed. Plant growth optimization using variable intensity and far red led treatment in indoor farming. In *Smart Sensors and Application (ICSSA), 2015 International Conference on*, pages 92–97. IEEE, 2015.

Olfa Hentati, Radhia Lachhab, Mariem Ayadi, and Mohamed Ksibi. Toxicity assessment for petroleum-contaminated soil using terrestrial invertebrates and plant bioassays. *Environmental monitoring and assessment*, 185(4):2989–2998, 2013.

Olfa Hentati, Nelson Abrantes, Ana Luísa Caetano, Sirine Bouguerra, Fernando Gonçalves, Jörg Römbke, and Ruth Pereira. Phosphogypsum as a soil fertilizer: ecotoxicity of amended soil and elutriates to bacteria, invertebrates, algae and plants. *Journal of hazardous materials*, 294:80–89, 2015.

JS Urban Hjorth. *Computer intensive statistical methods: Validation, model selection, and bootstrap*. Routledge, 2017.

Dennis Robert Hoagland, Daniel Israel Arnon, et al. The water-culture method for growing plants without soil. *Circular. California agricultural experiment station*, 347(2nd edit), 1950.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Robert K Horton. An index number system for rating water quality. *Journal of Water Pollution Control Federation*, 37(3):300–306, 1965.

Lillian J Hunt, Daiana Duca, Tereza Dan, and Loren D Knopper. Petroleum hydrocarbon (phc) uptake in plants: A literature review. *Environmental pollution*, 245:472–484, 2019.

María Dolores Hurtado, Santiago M Enamorado, Luis Andreu, Antonio Delgado, and José-María Abril. Drain flow and related salt losses as affected by phosphogypsum amendment in reclaimed marsh soils from sw spain. *Geoderma*, 161(1-2): 43–49, 2011.

Graeme D Hutcheson and Nick Sofroniou. *The multivariate social scientist: Introductory statistics using generalized linear models*. Sage, 1999. doi:10.2307/2681277.

Hirofumi Ibayashi, Yukimasa Kaneda, Jungo Imahara, Naoki Oishi, Masahiro Kuroda, and Hiroshi Mineno. A reliable wireless control system for tomato hydroponics. *Sensors*, 16(5):644, 2016.

Andrey Ignatov, Radu Timofte, William Chou, Ke Wang, Max Wu, Tim Hartley, and Luc Van Gool. Ai benchmark: Running deep neural networks on android smartphones. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

Mircea Horea Ionica and David Gregg. The movidius myriad architecture's potential for scientific computing. *IEEE Micro*, 35(1):6–14, 2015.

Tjalling Jager. Some good reasons to ban ec x and related concepts in ecotoxicology, 2011.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

Sunil Kr Jha and Zulfiqar Ahmad. Soil microbial dynamics prediction using machine learning regression methods. *Computers and electronics in agriculture*, 147:158–165, 2018.

J Benton Jones Jr. *Hydroponics: a practical guide for the soilless grower*. CRC press, 2016.

D Jung, JE Son, IB Lee, and MM Oh. Real-time control of hydroponic macronutrients for closed growing system. *Acta horticulturae*, pages 657–662, 2014.

Dae-Hyun Jung, Hyoung Seok Kim, Changho Jhin, Hak-Jin Kim, and Soo Hyun Park. Time-serial analysis of deep neural network models for prediction of climatic conditions inside a greenhouse. *Computers and Electronics in Agriculture*, 173: 105402, 2020.

Murat Kacira, Peter P Ling, and Ted H Short. Machine vision extracted plant movement for early detection of plant water stress. *Transactions of the ASAE*, 45 (4):1147, 2002.

R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Trans. ASME J. Basic Eng.*, 82:35–45, 1960.

Lev V Kalmykov and Vyacheslav L Kalmykov. A white-box model of s-shaped and double s-shaped single-species population growth. *PeerJ*, 3:e948, 2015.

Alexandros Kaloxylos, Robert Eigenmann, Frederick Teye, Zoi Politopoulou, Sjaak Wolfert, Claudia Shrank, Markus Dillinger, Ioanna Lampropoulou, Eleni Antoniou, Liisa Pesonen, et al. Farm management systems and the future internet era. *Computers and electronics in agriculture*, 89:130–144, 2012.

Andreas Kamilaris and Francesc X Prenafeta-Boldú. Deep learning in agriculture: A survey. *Computers and electronics in agriculture*, 147:70–90, 2018.

Mariem Kammoun, Imen Ghorbel, Safa Charfeddine, Lotfi Kamoun, Radhia Gargouri-Bouzid, and Oumèma Nouri-Ellouz. The positive effect of phosphogypsum-supplemented composts on potato plant growth in the field and tuber yield. *Journal of environmental management*, 200:475–483, 2017.

Liina Kanarbik, Irina Blinova, Mariliis Sihtmäe, Kai Künnis-Beres, and Anne Kahru. Environmental effects of soil contamination by shale fuel oils. *Environmental Science and Pollution Research*, 21(19):11320–11330, 2014.

Yukimasa Kaneda, Hirofumi Ibayashi, Naoki Oishi, and Hiroshi Mineno. Greenhouse environmental control system based on sw-svr. *Procedia Computer Science*, 60: 860–869, 2015.

Mengzhen Kang and Fei-Yue Wang. From parallel plants to smart plants: intelligent control and management for plant growth. *IEEE/CAA Journal of Automatica Sinica*, 4(2):161–166, 2017.

W-Y Kao and IN Forseth. Dirunal leaf movement, chlorophyll fluorescence and carbon assimilation in soybean grown under different nitrogen and water availabilities. *Plant, Cell & Environment*, 15(6):703–710, 1992.

Jalal Karami, Abbas Alimohammadi, and Tayebeh Seifouri. Water quality analysis using a variable consistency dominance-based rough set approach. *Computers, Environment and Urban Systems*, 43:25 – 33, 2014. ISSN 0198-9715. doi:https://doi.org/10.1016/j.compenvurbsys.2013.09.005. URL `http://www.sciencedirect.com/science/article/pii/S0198971513000859`.

Navjot Kaur, Todd E Erickson, Andrew S Ball, and Megan H Ryan. A review of germination and early growth as a proxy for plant fitness under petrogenic contamination—knowledge gaps and recommendations. *Science of The Total Environment*, 603:728–744, 2017.

Kent Kernahan. Aeroponic growth system wireless control system and methods of using, January 28 2016. US Patent App. 14/341,781.

H Keskin and S Grunwald. Regression kriging as a workhorse in the digital soil mapper's toolbox. *Geoderma*, 326:22–41, 2018. doi:10.1016/j.geoderma.2018.04.004.

M Khaki, Ehsan Forootan, Michael Kuhn, Joseph Awange, AIJM van Dijk, M Schumacher, and MA Sharifi. Determining water storage depletion within iran by assimilating grace data into the w3ra hydrological model. *Advances in Water Resources*, 114:1–18, 2018. doi:10.1016/j.advwatres.2018.02.008.

Muhammad Atikul Islam Khan, Bhabananda Biswas, Euan Smith, Ravi Naidu, and Mallavarapu Megharaj. Toxicity assessment of fresh and weathered petroleum hydrocarbons in contaminated soil-a review. *Chemosphere*, 212:755–767, 2018.

Muhammad Attique Khan, M Ikram Ullah Lali, Muhammad Sharif, Kashif Javed, Khursheed Aurangzeb, Syed Irtaza Haider, Abdulaziz Saud Altamrah, and Talha Akram. An optimized method for segmentation and classification of apple diseases based on strong correlation and genetic algorithm based feature selection. *IEEE Access*, 7:46261–46277, 2019.

Shin Woong Kim, Seung-Woo Jeong, and Youn-Joo An. Application of a soil quality assessment system using ecotoxicological indicators to evaluate contaminated and remediated soils. *Environmental geochemistry and health*, pages 1–10, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Jennifer L Kirk, John N Klirnomos, Hung Lee, and Jack T Trevors. Phytotoxicity assay to assess plant species for phytoremediation of petroleum-contaminated soil. *Bioremediation Journal*, 6(1):57–63, 2002.

Kiryushin. Tekhnologicheskaya modernizatsiya zemledeliya — neotlozhnaya zadacha. *Ekonomika sel'skogo khozyaystva Rossii.*, 2:17–25, 2009.

Werner Kloas, Roman Groß, Daniela Baganz, Johannes Graupner, Henrik Monsees, Uwe Schmidt, Georg Staaks, Johanna Suhl, Martin Tschirner, Bernd Wittstock, et al. A new concept for aquaponic systems to improve sustainability, increase productivity, and reduce environmental impacts. *Aquaculture Environment Interactions*, 7(2):179–192, 2015.

Dan Kou, Jinzhi Ding, Fei Li, Ning Wei, Kai Fang, Guibiao Yang, Beibei Zhang, Li Liu, Shuqi Qin, Yongliang Chen, et al. Spatially-explicit estimate of soil nitrogen stock and its implication for land model across tibetan alpine permafrost region. *Science of The Total Environment*, 650:1795–1804, 2019.

EI Kovaleva, AS Yakovlev, MG Nikolaenko, AO Makarov, and AA Makarov. Ecological evaluation of oil-contaminated soils (sakhalin) using enchytraeidae. *Eurasian soil science*, 50(3):350–358, 2017.

Toyoki Kozai, Genhua Niu, and Michiko Takagaki. *Plant factory: an indoor vertical farming system for efficient quality food production.* Academic Press, 2015.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

P Krishnan and Pramila Aggarwal. Global sensitivity and uncertainty analyses of a web based crop simulation model (web InfoCrop wheat) for soil parameters. *Plant and soil*, 423(1-2):443–463, 2018.

M. Kucer, A. C. Loui, and D. W. Messinger. Leveraging expert feature knowledge for predicting image aesthetics. *IEEE Transactions on Image Processing*, 27(10): 5100–5112, Oct 2018. ISSN 1057-7149. doi:10.1109/TIP.2018.2845100.

Matjaž Kukar, Petar Vračar, Domen Košir, Darko Pevec, Zoran Bosnić, et al. Agrodss: A decision support system for agriculture and farming. *Computers and Electronics in Agriculture*, 161:260–271, 2019.

Gennady Yu Kulikov and Maria V Kulikova. The accurate continuous-discrete extended kalman filter for radar tracking. *IEEE Transactions on Signal Processing*, 64(4):948–958, 2015.

Victor Kulikov, Victor Yurchenko, and Victor Lempitsky. Instance segmentation by deep coloring. *arXiv preprint*, 2018.

MT Kuska and A-K Mahlein. Aiming at decision making in plant disease protection and phenotyping by the use of optical sensors. *European journal of plant pathology*, 152(4):987–992, 2018.

Sagar Lachure, Amol Bhagat, and Jaykumar Lachure. Review on precision agriculture using wireless sensor network. *International Journal of Applied Engineering Research*, 10(20):16560–16565, 2015.

KKR Lakkireddy, K Kasturi, and Sambasiva Rao KRS. Role of hydroponics and aeroponics in soilless culture in commercial food production. *Research & Reviews: Journal of Agricultural Science and Technology*, 1(3):1–8, 2018.

Nicholas D Lane, Sourav Bhattacharya, Akhil Mathur, Petko Georgiev, Claudio Forlivesi, and Fahim Kawsar. Squeezing deep learning into mobile and embedded devices. *IEEE Pervasive Computing*, 16(3):82–88, 2017.

JF Larive. Performance of european cross-country oil pipelines. statistical summary of reported spillages in 2006 and since 1971. Technical report, CONCAWE Oil Pipelines Management Group's Special Task Force on oil pipeline . . . , 2008.

Seyyed Salar Latifi Oskouei, Hossein Golestani, Matin Hashemi, and Soheil Ghiasi. Cnndroid: Gpu-accelerated execution of trained deep convolutional neural networks on android. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1201–1205. ACM, 2016.

Hoonsoo Lee, Tran Quoc Huy, Eunsoo Park, Hyung-Jin Bae, Insuck Baek, Moon S Kim, Changyeun Mo, and Byoung-Kwan Cho. Machine vision technique for rapid measurement of soybean seed vigor. *Journal of Biosystems Engineering*, 42(3): 227–233, 2017.

Lei Li, Qin Zhang, and Danfeng Huang. A review of imaging techniques for plant phenotyping. *Sensors*, 14(11):20078–20111, 2014.

Jie Liang, Ali Zia, Jun Zhou, and Xavier Sirault. 3d plant modelling via hyperspectral imaging. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 172–177, 2013.

Kaiyan Lin, Jie Chen, Huiping Si, and Junhui Wu. A review on computer vision technologies applied in greenhouse plant stress detection. In *Chinese Conference on Image and Graphics Technologies*, pages 192–200. Springer, 2013.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

Patrick Lüscher and Robert Weibel. Exploiting empirical knowledge for automatic delineation of city centres from large-scale topographic databases. *Computers, Environment and Urban Systems*, 37:18 – 34, 2013. ISSN 0198-9715. doi:https://doi.org/10.1016/j.compenvurbsys.2012.07.001. URL `http://www.sciencedirect.com/science/article/pii/S0198971512000609`.

Kelsey MacCormack, Emmanuelle Arnaud, and Beth L Parker. Using a multiple variogram approach to improve the accuracy of subsurface geological models. *Canadian Journal of Earth Sciences*, 55(7):786–801, 2018. doi:10.1139/cjes-2016-0112.

Kelsey E MacCormack, Jason J Brodeur, and Carolyn H Eyles. Evaluating the impact of data quantity, distribution and algorithm selection on the accuracy of

3d subsurface models using synthetic grid models of varying complexity. *Journal of Geographical Systems*, 15(1):71–88, 2013. doi:10.1007/s10109-011-0160-x.

David Macii, Anton Ageev, and Andrey Somov. Power consumption reduction in wireless sensor networks through optimal synchronization. In *2009 IEEE instrumentation and measurement technology conference*, pages 1346–1351. IEEE, 2009.

Yolanda Madrid and Zoyne Pedrero Zayas. Water sampling: Traditional methods and new approaches in water sampling strategy. *TrAC Trends in Analytical Chemistry*, 26(4):293–299, 2007. doi:10.1016/j.trac.2007.01.002.

A-K Mahlein, T Rumpf, P Welke, H-W Dehne, L Plümer, U Steiner, and E-C Oerke. Development of spectral indices for detecting and identifying plant diseases. *Remote Sensing of Environment*, 128:21–30, 2013.

Anne-Katrin Mahlein, Erich-Christian Oerke, Ulrike Steiner, and Heinz-Wilhelm Dehne. Recent advances in sensing plant diseases for precision crop protection. *European Journal of Plant Pathology*, 133(1):197–209, 2012.

Hanping Mao, Hongyan Gao, Xiaodong Zhang, and Francis Kumi. Nondestructive measurement of total nitrogen in lettuce by integrating spectroscopy and computer vision. *Scientia Horticulturae*, 184:1–7, 2015.

Charalampos Marantos, Nikolaos Karavalakis, Vasileios Leon, Vasileios Tsoutsouras, Kiamal Pekmestzi, and Dimitrios Soudris. Efficient support vector machines implementation on intel/movidius myriad 2. In *2018 7th International Conference on Modern Circuits and Systems Technologies (MOCAST)*, pages 1–4. IEEE, 2018.

Kanaji Masakorala, Jun Yao, Huan Guo, Radhika Chandankere, Jingwei Wang, Minmin Cai, Haijun Liu, and Martin MF Choi. Phytotoxicity of long-term total petroleum hydrocarbon-contaminated soil—a comparative and combined approach. *Water, Air, & Soil Pollution*, 224(5):1553, 2013.

Hipólito Medrano, Magdalena Tomás, Sebastià Martorell, Jaume Flexas, Esther Hernández, Joan Rosselló, Alicia Pou, José-Mariano Escalona, and Josefina Bota. From leaf to whole-plant water use efficiency (wue) in complex canopies: limitations of leaf wue as a selection target. *The Crop Journal*, 3(3):220–228, 2015.

Manav Mehra, Sameer Saxena, Suresh Sankaranarayanan, Rijo Jackson Tom, and M Veeramanikandan. Iot based hydroponics system using deep neural networks. *Computers and electronics in agriculture*, 155:473–486, 2018.

Gerard Rudolph Mendez, Mohd Amri Md Yunus, and Subhas Chandra Mukhopadhyay. A wifi based smart wireless sensor network for monitoring an agricultural environment. In *2012 IEEE International Instrumentation and Measurement Technology Conference Proceedings*, pages 2640–2645. IEEE, 2012.

Ralf Metzner, Anja Eggert, Dagmar Van Dusschoten, Daniel Pflugfelder, Stefan Gerth, Ulrich Schurr, Norman Uhlmann, and Siegfried Jahnke. Direct comparison of mri and x-ray ct technologies for 3d imaging of root systems in soil: potential and challenges for root trait quantification. *Plant methods*, 11(1):1–11, 2015.

Massimo Minervini, Hanno Scharr, and Sotirios A Tsaftaris. Image analysis: the new bottleneck in plant phenotyping [applications corner]. *IEEE signal processing magazine*, 32(4):126–131, 2015.

Massimo Minervini, Andreas Fischbach, Hanno Scharr, and Sotirios A Tsaftaris. Finely-grained annotated datasets for image-based plant phenotyping. *Pattern recognition letters*, 81:80–89, 2016.

Massimo Minervini, Mario V Giuffrida, Pierdomenico Perata, and Sotirios A Tsaftaris. Phenotiki: an open software and hardware platform for affordable and easy image-based phenotyping of rosette-shaped plants. *The Plant Journal*, 90 (1):204–216, 2017.

Daniele Miorandi, Sabrina Sicari, Francesco De Pellegrini, and Imrich Chlamtac. Internet of things: Vision, applications and research challenges. *Ad hoc networks*, 10(7):1497–1516, 2012.

Orazio Mirabella and Michele Brischetto. A hybrid wired/wireless networking infrastructure for greenhouse management. *IEEE Transactions on Instrumentation and Measurement*, 60(2):398–407, 2011.

PD Mitchell. Methods and assumptions for estimating the impact of neonicotinoid insecticides on pest management practices and costs for us corn, soybean, wheat, cotton and sorghum farmers. Technical report, AgInfomatics Research Report. Madison, WI; 2014. 96 p. Available from: http . . . , 2014.

Tatjana Mitrović, Davor Antanasijević, Saša Lazović, Aleksandra Perić-Grujić, and Mirjana Ristić. Virtual water quality monitoring at inactive monitoring sites using monte carlo optimized artificial neural networks: a case study of danube river (serbia). *Science of the Total Environment*, 654:1000–1009, 2019. doi:10.1016/j.scitotenv.2018.11.189.

Naoko Miyamoto, Ernst Steudle, Tadashi Hirasawa, and Renee Lafitte. Hydraulic conductivity of rice roots. *Journal of Experimental Botany*, 52(362):1835–1846, 2001.

Zalina Mohd Ali, Noor Akma Ibrahim, Kerrie Mengersen, Mahendran Shitan, and Hafizan Juahir. The langat river water quality index based on principal component analysis. In *AIP Conference Proceedings*, volume 1522, pages 1322–1336. American Institute of Physics, 2013. doi:10.1063/1.4801283.

George Mois, Silviu Folea, and Teodora Sanislav. Analysis of three iot-based wireless sensors for environmental monitoring. *IEEE Transactions on Instrumentation and Measurement*, 66(8):2056–2064, 2017.

L Molina-Barahona, L Vega-Loyo, M Guerrero, S Ramirez, I Romero, C Vega-Jarquín, and A Albores. Ecotoxicological evaluation of diesel-contaminated soil before and after a bioremediation process. *Environmental Toxicology: An International Journal*, 20(1):100–109, 2005.

JI Montero, E Baeza, E Heuvelink, J Rieradevall, P Muñoz, M Ercilla, and C Stanghellini. Productivity of a building-integrated roof top greenhouse in a mediterranean climate. *Agricultural Systems*, 158:14–22, 2017.

Mohammad Motamedi, Daniel Fong, and Soheil Ghiasi. Cappuccino: efficient cnn inference software synthesis for mobile system-on-chips. *IEEE Embedded Systems Letters*, 11(1):9–12, 2019.

TG Mueller, NB Pusuluri, KK Mathias, PL Cornelius, RI Barnhisel, and SA Shearer. Map quality for ordinary kriging and inverse distance weighted interpolation. *Soil Science Society of America Journal*, 68(6):2042–2047, 2004. doi:doi:10.2136/sssaj2004.2042.

J. Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society of Industrial and Applied Mathematics*, 5(1):32–38, March 1957.

P Muñoz, A Antón, M Nuñez, A Paranjpe, J Ariño, X Castells, JI Montero, and J Rieradevall. Comparing the environmental impacts of greenhouse versus open-field tomato production in the mediterranean region. In *International Symposium on High Technology for Greenhouse System Management: Greensys2007 801*, pages 1591–1596, 2007.

P C Nagajyoti, K D Lee, and TVM Sreekanth. Heavy metals, occurrence and toxicity for plants: a review. *Environmental chemistry letters*, 8(3):199–216, 2010.

Nefedov. Innovatsionnyye tekhnologicheskiye protsessy i mashiny dlya vnutripochvennogo vneseniya mineral'nykh udobreniy v sisteme tochnogo zemledeliya: monografiya. *MESKH*, 2e, 2015.

Sónia Negrão, SM Schmöckel, and M Tester. Evaluating physiological responses of plants to salinity stress. *Annals of botany*, 119(1):1–11, 2017.

C Nendel, M Berg, K Ch Kersebaum, W Mirschel, X Specka, M Wegehenkel, KO Wenkel, and R Wieland. The MONICA model: Testing predictability for crop growth, soil moisture and nitrogen dynamics. *Ecological Modelling*, 222(9): 1614–1625, 2011.

Sergey Nesteruk, Dmitrii Shadrin, Vladislav Kovalenko, Antonio Rodríguez-Sánchez, and Andrey Somov. Plant growth prediction through intelligent embedded sensing. In *IEEE International Symposium on Industrial Electronics 2020, IEEE ISIE 2020 (Accepted)*, 2020.

Chuong V Nguyen, Jurgen Fripp, David R Lovell, Robert Furbank, Peter Kuffner, Helen Daily, and Xavier Sirault. 3d scanning system for automatic high-resolution plant phenotyping. In *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE, 2016.

Silvina Niell, Florencia Jesús, Rosana Díaz, Yamandú Mendoza, Gastón Notte, Estela Santos, Natalia Gérez, Verónica Cesio, Héctor Cancela, and Horacio Heinzen. Beehives biomonitor pesticides in agroecosystems: Simple chemical and biological

indicators evaluation using support vector machines (svm). *Ecological indicators*, 91:149–154, 2018.

Artyom Nikitin, Ilia Fastovets, Dmitrii Shadrin, Mariia Pukalchik, and Ivan Oseledets. Bayesian optimization for seed germination. *Plant methods*, 15(1):43, 2019.

Olga Nikolaeva, Vladimir Tikhonov, Maxim Vecherskii, Natalia Kostina, Elena Fedoseeva, and Angelika Astaikina. Ecotoxicological effects of traffic-related pollutants in roadside soils of moscow. *Ecotoxicology and environmental safety*, 172: 538–546, 2019.

Maria-Elena Nilsback and Andrew Zisserman. Delving deeper into the whorl of flower segmentation. *Image and Vision Computing*, 28(6):1049–1062, 2010.

Tobias Nilsson, Benedikt Soja, Maria Karbon, Robert Heinkelmann, and Harald Schuh. Application of kalman filtering in vlbi data analysis. *Earth, Planets and Space*, 67(1):136, 2015.

S. Noh, D. Shim, and M. Jeon. Adaptive sliding-window strategy for vehicle detection in highway environments. *IEEE Transactions on Intelligent Transportation Systems*, 17(2):323–335, Feb 2016. ISSN 1524-9050. doi:10.1109/TITS.2015.2466652.

Maroua Nouri, Nathalie Gorretta, Pierre Vaysse, Michel Giraud, Christian Germain, Barna Keresztes, and Jean-Michel Roger. Near infrared hyperspectral dataset of healthy and infected apple tree leaves images for the early detection of apple scab disease. *Data in brief*, 16:967–971, 2018.

Kazunari Nozue and Julin N Maloof. Diurnal regulation of plant growth. *Plant, Cell & Environment*, 29(3):396–408, 2006.

Richard Olawoyin. Application of backpropagation artificial neural network prediction model for the pah bioremediation of polluted soil. *Chemosphere*, 161:145–150, 2016.

Orbis Research. Global precision farming market 2018-2023. 2018.

D. Oro, C. Fernández, X. Martorell, and J. Hernando. Work-efficient parallel non-maximum suppression for embedded gpu architectures. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1026–1030, March 2016. doi:10.1109/ICASSP.2016.7471831.

Ying Ouyang. Evaluation of river water quality monitoring stations by principal component analysis. *Water research*, 39(12):2621–2635, 2005. doi:10.1016/j.watres.2005.04.024.

Murat Ozturk, Ozlem Salman, and Murat Koc. Artificial neural network model for estimating the soil temperature. *Canadian journal of soil science*, 91(4):551–562, 2011.

R. Pahuja, H. K. Verma, and M. Uddin. A wireless sensor network for greenhouse climate control. *IEEE Pervasive Computing*, 12(2):49–58, April 2013a. ISSN 1536-1268. doi:10.1109/MPRV.2013.26.

Roop Pahuja, HK Verma, and Moin Uddin. A wireless sensor network for greenhouse climate control. *IEEE Pervasive Computing*, 12(2):49–58, 2013b.

Pedro Palencia, Jordi Giné Bordonaba, Fátima Martínez, and Leon A Terry. Investigating the effect of different soilless substrates on strawberry productivity and fruit composition. *Scientia horticulturae*, 203:12–19, 2016.

David S Palmer, Maksim Misin, Maxim V Fedorov, and Antonio Llinas. Fast and general method to predict the physicochemical properties of druglike molecules using the integral equation theory of molecular liquids. *Molecular pharmaceutics*, 12(9):3420–3432, 2015.

Xanthoula Eirini Pantazi, Dimitrios Moshou, and Alexandra A Tamouridou. Automated leaf disease detection in different crop species through image features analysis and one class classifiers. *Computers and electronics in agriculture*, 156: 96–104, 2019.

Dae-Heon Park, Beom-Jin Kang, Kyung-Ryong Cho, Chang-Sun Shin, Sung-Eon Cho, Jang-Woo Park, and Won-Mo Yang. A study on greenhouse automatic control system based on wireless sensor network. *Wireless Personal Communications*, 56(1):117–130, 2011.

Jorge Parraga-Alava, Kevin Cusme, Angélica Loor, and Esneider Santander. Rocole: A robusta coffee leaf images dataset for evaluation of machine learning based methods in plant diseases recognition. *Data in brief*, 25:104414, 2019.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

Jayamala K Patil and Raj Kumar. Advances in image processing for detection of plant diseases. *Journal of Advanced Bioinformatics Applications and Research*, 2 (2):135–141, 2011.

Stefan Paulus, Jan Dupuis, Anne-Katrin Mahlein, and Heiner Kuhlmann. Surface feature based classification of plant organs from 3d laserscanned point clouds for plant phenotyping. *BMC bioinformatics*, 14(1):238, 2013.

Stefan Paulus, Jan Behmann, Anne-Katrin Mahlein, Lutz Plümer, and Heiner Kuhlmann. Low-cost 3d systems: suitable tools for plant phenotyping. *Sensors*, 14(2):3001–3018, 2014.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

Nils Pletschen and Klaus J Diepold. Nonlinear state estimation for suspension control applications: a takagi-sugeno kalman filtering approach. *Control Engineering Practice*, 61:292–306, 2017.

Johannes A Postma, Ulrich Schurr, and Fabio Fiorani. Dynamic root growth and architecture responses to limiting nutrient availability: linking physiological models and experimentation. *Biotechnology Advances*, 32(1):53–65, 2014.

Nastaran Pouladi, Ali Asghar Jafarzadeh, Farzin Shahbazi, and Mohammad Ali Ghorbani. Design and implementation of a hybrid mlp-ffa model for soil salinity prediction. *Environmental earth sciences*, 78(5):159, 2019.

Parisa Pouladzadeh, Shervin Shirmohammadi, and Rana Al-Maghrabi. Measuring calorie and nutrition from food image. *IEEE Transactions on Instrumentation and Measurement*, 63(8):1947–1956, 2014.

Michael P Pound, Andrew P French, Erik H Murchie, and Tony P Pridmore. Automated recovery of 3d models of plant shoots from multiple colour images. *Plant Physiology*, pages pp–114, 2014.

Joshua L Proctor, Steven L Brunton, and J Nathan Kutz. Dynamic mode decomposition with control. *SIAM Journal on Applied Dynamical Systems*, 15(1):142–161, 2016.

Jagadeesh D Pujari, Rajesh Yakkundimath, and Abdulmunaf S Byadgi. Image processing based detection of fungal diseases in plants. *Procedia Computer Science*, 46:1802–1808, 2015.

Maria Pukalchik, Dmitrii Shadrin, and Maxim Fedorov. Global trends and perspective development directions in precision agriculture. *APK Russia Journal (in Russian)*, 2018.

Maria A Pukalchik, Alexandr M Katrutsa, Dmitry Shadrin, Vera A Terekhova, and Ivan V Oseledets. Machine learning methods for estimation the indicators of phosphogypsum influence in soil. *Journal of Soils and Sediments*, 19(5):2265–2276, 2019.

Mariia Pukalchik, Dmitrii Shadrin, Artyom Nikitin, Raghavendra Jana, Polina Tregubova, and Sergey Matveev. freshwater chemical properties for New Moscow region. 7 2020. doi:10.6084/m9.figshare.10283225.v2. URL https://figshare.com/articles/dataset/freshwater_chemical_properties_for_New_Moscow_region/10283225.

P Agung Putra and Henry Yuliando. Soilless culture system to support water use efficiency and product quality: a review. *Agriculture and Agricultural Science Procedia*, 3:283–288, 2015.

Long Quan, Ping Tan, Gang Zeng, Lu Yuan, Jingdong Wang, and Sing Bing Kang. Image-based plant modeling. In *ACM Transactions on Graphics (TOG)*, volume 25, pages 599–604. ACM, 2006.

KARTHIKA Rajendran, MARK Tester, and STUART J. Roy. Quantifying the three main components of salinity tolerance in cereals. *Plant, Cell & Environment*, 32(3):237–249. doi:10.1111/j.1365-3040.2008.01916.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-3040.2008.01916.x.

CR Ramakrishnaiah, C Sadashivaiah, and G Ranganna. Assessment of water quality index for the groundwater in tumkur taluk, karnataka state, india. *Journal of Chemistry*, 6(2):523–530, 2009. doi:10.1155/2009/757424.

Lino J. Ramírez-Pérez, América B. Morales-Díaz, Adalberto Benavides-Mendoza, Karim De-Alba-Romenus, Susana González-Morales, and Antonio Juárez-Maldonado. Dynamic modeling of cucumber crop growth and uptake of n, p and k under greenhouse conditions. 234:250–260, 2018. ISSN 03044238. doi:10.1016/j.scienta.2018.02.068. URL https://linkinghub.elsevier.com/retrieve/pii/S0304423818301523.

Mengye Ren and Richard S Zemel. End-to-end instance segmentation with recurrent attention. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA*, pages 21–26, 2017.

Russian National Report. Water quality analysis using a variable consistency dominance-based rough set approach. 2015.

Howard M Resh. *Hydroponic food production: a definitive guidebook for the advanced home gardener and the commercial hydroponic grower*. CRC Press, 2016.

F Xavier Rius-Ruiz, Francisco J Andrade, Jordi Riu, and F Xavier Rius. Computer-operated analytical platform for the determination of nutrients in hydroponic systems. *Food chemistry*, 147:92–97, 2014.

Robotics Microfarms. Improved risk prediction for precision agriculture: automated monitoring of pathogen movement. 2018.

Francisco Rodríguez, Manuel Berenguel, José Luis Guzmán, and Armando Ramírez-Arias. *Modeling and control of greenhouse crop growth*. Springer, 2015.

Raquel Rodríguez-Perez, Martin Vogt, and Jurgen Bajorath. Influence of varying training set composition and size on support vector machine-based prediction of active compounds. *Journal of chemical information and modeling*, 57(4):710–716, 2017.

Bernardino Romera-Paredes and Philip Hilaire Sean Torr. Recurrent instance segmentation. In *European Conference on Computer Vision*, pages 312–329. Springer, 2016.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

Jacques Roy. *Response of plants to multiple stresses*. Academic Press, 2012.

Vincent E Rubatzky and Mas Yamaguchi. *World vegetables: principles, production, and nutritive values*. Springer Science & Business Media, 2012.

A Ruíz-García, IL López-Cruz, A Ramírez-Arias, and E Rico-Garcia. Modeling uncertainty of greenhouse crop lettuce growth model using kalman filtering. In *International Symposium on New Technologies for Environment Control, Energy-Saving and Crop Production in Greenhouse and Plant 1037*, pages 361–368, 2013.

T Rumpf, A-K Mahlein, U Steiner, E-C Oerke, H-W Dehne, and L Plümer. Early detection and classification of plant diseases with support vector machines based on hyperspectral reflectance. *Computers and Electronics in Agriculture*, 74(1): 91–99, 2010.

Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1-3):157–173, 2008.

R.P. Rötter, M. Appiah, E. Fichtler, K.C. Kersebaum, M. Trnka, and M.P Hoffmann. Linking modelling and experimentation to better capture crop impacts of agroclimatic extremes—a review. *Field Crops Research*, 221:142 – 156, 2018. ISSN 0378-4290. doi:https://doi.org/10.1016/j.fcr.2018.02.023.

Ezzeddine Saadaoui, Naziha Ghazel, Chokri Ben Romdhane, and Nouman Massoudi. Phosphogypsum: potential uses and problems–a review. *International Journal of Environmental Studies*, 74(4):558–567, 2017.

Ibrahim Said, Salman Abd El-Raof Salman, Yousria Samy, Samir Ahmed Awad, Ahmed Melegy, and Andrew S Hursthouse. Environmental factors controlling potentially toxic element behaviour in urban soils, el tebbin, egypt. *Environmental monitoring and assessment*, 191(5):267, 2019.

Farzaneh Sajedi-Hosseini, Arash Malekian, Bahram Choubin, Omid Rahmati, Sabrina Cipullo, Frederic Coulon, and Biswajeet Pradhan. A novel machine learning-based approach for the risk assessment of nitrate groundwater contamination. *Science of the total environment*, 644:954–962, 2018. doi:10.1016/j.scitotenv.2018.07.054.

Sindhuja Sankaran, Ashish Mishra, Reza Ehsani, and Cristina Davis. A review of advanced techniques for detecting plant diseases. *Computers and Electronics in Agriculture*, 72(1):1–13, 2010.

Esther Sanyé-Mengual, Francesco Orsini, Jordi Oliver-Solà, Joan Rieradevall, Juan Ignacio Montero, and Giorgio Gianquinto. Techniques and crops for efficient rooftop gardens in bologna, italy. *Agronomy for sustainable development*, 35 (4):1477–1488, 2015.

Muhammad Sarfraz. *Computer Vision and Image Processing in Intelligent Systems and Multimedia Technologies*. IGI Global, 2014.

Swaytha Sasidharan, Andrey Somov, Abdur Rahim Biswas, and Raffaele Giaffreda. Cognitive management framework for internet of things:—a prototype implementation. In *2014 IEEE World Forum on Internet of Things (WF-IoT)*, pages 538–543. IEEE, 2014.

Javad Sayyad Amin, Hossein Rajabi Kuyakhi, and Alireza Bahadori. Prediction of formation of polycyclic aromatic hydrocarbon (pahs) on sediment of caspian sea using artificial neural networks. *Petroleum Science and Technology*, 37(18): 1987–2000, 2019.

Hanno Scharr, Massimo Minervini, Andrew P French, Christian Klukas, David M Kramer, Xiaoming Liu, Imanol Luengo, Jean-Michel Pape, Gerrit Polder, Danijela Vukadinovic, et al. Leaf segmentation in plant phenotyping: a collation study. *Machine vision and applications*, 27(4):585–606, 2016.

Peter J Schmid. Dynamic mode decomposition of numerical and experimental data. *Journal of fluid mechanics*, 656:5–28, 2010.

JW Scott, BK Harbaugh, and EA Baldwin. Micro-tina'andmicro-gemma'miniature dwarf tomatoes. *HortScience*, 35(4):774–775, 2000.

AP Sergeev, AG Buevich, EM Baglaeva, and AV Shichkin. Combining spatial autocorrelation with machine learning increases prediction accuracy of soil heavy metals. *Catena*, 174:425–435, 2019.

D. Shadrin, A. Somov, T. Podladchikova, and R. Gerzer. Pervasive agriculture: Measuring and predicting plant growth using statistics and 2d/3d imaging. In *2018 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, pages 1–6, May 2018. doi:10.1109/I2MTC.2018.8409700.

D. Shadrin, A. Menshchikov, D. Ermilov, and A. Somov. Designing future precision agriculture: Detection of seeds germination using artificial intelligence on a low-power embedded system. *IEEE Sensors Journal*, 19(23):11573–11582, Dec 2019. ISSN 2379-9153. doi:10.1109/JSEN.2019.2935812.

Dmitrii Shadrin. SeedGermination. `https://github.com/DmitriiShadrin/SeedsGermination/`, 2018. [Online; accessed 19-May-2019].

Dmitrii Shadrin, Andrey Somov, Tatiana Podladchikova, and Rupert Gerzer. Pervasive agriculture: Measuring and predicting plant growth using statistics and 2d/3d imaging. In *2018 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, pages 1–6. IEEE, 2018a.

Dmitrii Shadrin, Artem Chashchin, George Ovchinnikov, and Andrey Somov. System identification-soilless growth of tomatoes. In *2019 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, pages 1–6. IEEE, 2019a.

Dmitrii Shadrin, Alexander Menshchikov, Dmitry Ermilov, and Andrey Somov. Designing future precision agriculture: Detection of seeds germination using artificial intelligence on a low-power embedded system. *IEEE Sensors Journal*, 19(23): 11573–11582, 2019b.

Dmitrii Shadrin, Alexander Menshchikov, Andrey Somov, Gerhild Bornemann, Jens Hauslage, and Maxim Fedorov. Tomato Growth Dataset. https://github.com/DmitriiShadrin/TomatoesGrowth, 2019c. [Online; accessed 21-January-2019].

Dmitrii Shadrin, Alexander Menshchikov, Andrey Somov, Gerhild Bornemann, Jens Hauslage, and Maxim Fedorov. Enabling precision agriculture through embedded sensing with artificial intelligence. *IEEE Transactions on Instrumentation and Measurement*, 2019d.

Dmitrii Shadrin, Tatiana Podladchikova, George Ovchinnikov, Artem Pavlov, Maria Pukalchik, and Andrey Somov. Kalman filtering for accurate and fast plant growth dynamics assessment. In *2020 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*. IEEE, 2020a.

Dmitrii Shadrin, Mariia Pukalchik, Ekaterina Kovaleva, and Maxim Fedorov. Artificial intelligence models to predict acute phytotoxicity in petroleum contaminated soils. *Ecotoxicology and Environmental Safety*, 194:110410, 2020b.

Dmitrii Shadrin, Mariia Pukalchik, Anastasia Uryasheva, Evgeny Tsykunov, Grigoriy Yashin, Nikita Rodichenko, and Dzmitry Tsetserukou. Hyper-spectral nir and mir data and optimal wavebands for detection of apple tree diseases. In *ICLR (CV4A)*, 2020c.

Dmitrii G Shadrin, Victor Kulikov, and Maxim Fedorov. Instance segmentation for assessment of plant growth dynamics in artificial soilless conditions. In *BMVC*, page 329, 2018b.

Esmaeil Shahsavari, Eric M Adetutu, Peter A Anderson, and Andrew S Ball. Tolerance of selected plant species to petrogenic hydrocarbons and effect of plant rhizosphere on the microbial removal of hydrocarbons in contaminated soil. *Water, Air, & Soil Pollution*, 224(4):1495, 2013.

L. Shatalina. Tochnoye zemledeliye kak odin iz putey k energosberezheniyu resursov sel'skokhozyaystvennom proizvodstve. *APK Rossii*, 24:949–953, 2017.

Weisong Shi, Jie Cao, Quan Zhang, Youhuizi Li, and Lanyu Xu. Edge computing: Vision and challenges. *IEEE Internet of Things Journal*, 3(5):637–646, 2016.

Yakusheva V. P. Shpaara D., Zakharenko A. V. Tochnoye sel'skoye khozyaystvo. *ucheb.-prakt. Posobiye. SPb*, 2009.

De Silva. Ipanera: An industry 4.0 based architecture for distributed soil-less food production systems. In *Manufacturing & Industrial Engineering Symposium (MIES)*, pages 1–5. IEEE, 2016.

Pedro FB Silva, Andre RS Marcal, and Rubim M Almeida da Silva. Evaluation of features for leaf discrimination. In *International Conference Image Analysis and Recognition*, pages 197–204. Springer, 2013.

Arti Singh, Baskar Ganapathysubramanian, Asheesh Kumar Singh, and Soumik Sarkar. Machine learning for high-throughput stress phenotyping in plants. *Trends in plant science*, 21(2):110–124, 2016.

Baihaqi Siregar, Syahril Efendi, Heru Pranoto, Roy Ginting, Ulfi Andayani, and Fahmi Fahmi. Remote monitoring system for hydroponic planting media. In *2017 International Conference on ICT For Smart Society (ICISS)*, pages 1–6. IEEE, 2017.

A. Somov, D. Shadrin, I. Fastovets, A. Nikitin, S. Matveev, I. Oseledets, and O. Hrinchuk. Pervasive agriculture: IoT-enabled greenhouse for plant growth control. *IEEE Pervasive Computing*, 2018. doi:10.1109/MPRV.2018.2873849.

A. Somov, D. Shadrin, I. Fastovets, A. Nikitin, S. Matveev, I. Oseledets, and O. Hrinchuk. Pervasive agriculture: Iot-enabled greenhouse for plant growth control. *IEEE Pervasive Computing*, 17(4):65–75, Oct 2018. ISSN 1558-2590. doi:10.1109/MPRV.2018.2873849.

Andrey Somov, Christine C. Ho, Roberto Passerone, James W. Evans, and Paul K. Wright. Towards extending sensor node lifetime with printed supercapacitors. In Gian Pietro Picco and Wendi Heinzelman, editors, *Wireless Sensor Networks*, pages 212–227, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-28169-3.

Andrey Somov, Dmitry Shadrin, Ilia Fastovets, Artyom Nikitin, Sergey Matveev, Oleksii Hrinchuk, et al. Pervasive agriculture: Iot-enabled greenhouse for plant growth control. *IEEE Pervasive Computing*, 17(4):65–75, 2018.

Anne Sophie et al. Disease decision support systems: their impact on disease management and durability of fungicide effectiveness. In *Fungicides*. InTech, 2010.

Sanye Soroldoni, Graciane Silva, Fabio Veríssimo Correia, and Marcia Marques. Spent lubricant oil-contaminated soil toxicity to eisenia andrei before and after bioremediation. *Ecotoxicology*, 28(2):212–221, 2019.

D. Spirjakin, A. Baranov, A. Karelin, and A. Somov. Wireless multi-sensor gas platform for environmental monitoring. In *2015 IEEE Workshop on Environmental, Energy, and Structural Monitoring Systems (EESMS) Proceedings*, pages 232–237, July 2015. doi:10.1109/EESMS.2015.7175883.

Marijn F Stollenga, Wonmin Byeon, Marcus Liwicki, and Juergen Schmidhuber. Parallel multi-dimensional lstm, with application to fast biomedical volumetric image segmentation. In *Advances in neural information processing systems*, pages 2998–3006, 2015.

Wei Sun, Chunyu Xia, Meiying Xu, Jun Guo, and Guoping Sun. Application of modified water quality indices as indicators to assess the spatial and temporal trends of water quality in the dongjiang river. *Ecological Indicators*, 66:306–312, 2016. doi:10.1016/j.ecolind.2016.01.054.

Yuan Sun, Kelin Hu, Kefeng Zhang, Lihua Jiang, and Yu Xu. Simulation of nitrogen fate for greenhouse cucumber grown under different water and fertilizer management using the EU-rotate_n model. 112:21–32, 2012. ISSN 03783774. doi:10.1016/j.agwat.2012.06.001. URL https://linkinghub.elsevier.com/retrieve/pii/S0378377412001424.

MA Tabatabai. Effects of trace elements on urease activity in soils. *Soil Biology and Biochemistry*, 9(1):9–13, 1977.

Hanan Tayibi, Mohamed Choura, Félix A López, Francisco J Alguacil, and Aurora López-Delgado. Environmental impact and management of phosphogypsum. *Journal of environmental management*, 90(8):2377–2386, 2009.

Kerry Taylor, Colin Griffith, Laurent Lefort, Raj Gaire, Michael Compton, Tim Wark, David Lamb, Greg Falzon, and Mark Trotter. Farming the web of things. *IEEE Intelligent Systems*, 28(6):12–19, 2013.

Thanh Hoai Tran, Einav Mayzlish Gati, Amram Eshel, and Gidon Winters. Germination, physiological and biochemical responses of acacia seedlings (acacia raddiana and acacia tortilis) to petroleum contaminated soils. *Environmental pollution*, 234:642–655, 2018.

Mansi Tripathi and Sunil Kumar Singal. Allocation of weights using factor analysis for development of a novel water quality index. *Ecotoxicology and environmental safety*, 183:109510, 2019a. doi:Allocation of weights using factor analysis for development of a novel water quality index.

Mansi Tripathi and Sunil Kumar Singal. Use of principal component analysis for parameter selection for development of a novel water quality index: A case study of river ganga india. *Ecological Indicators*, 96:430–436, 2019b. doi:10.1016/j.ecolind.2018.09.025.

Jonathan H Tu, Clarence W Rowley, Dirk M Luchtenburg, Steven L Brunton, and J Nathan Kutz. On dynamic mode decomposition: theory and applications. *arXiv preprint arXiv:1312.0041*, 2013.

Muammer Turkoglu, Davut Hanbay, and Abdulkadir Sengur. Multi-model lstm-based convolutional neural networks for detection of apple diseases and pests. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–11, 2019.

Katrine Grace Turner, Sharolyn Anderson, Mauricio Gonzales-Chang, Robert Costanza, Sasha Courville, Tommy Dalgaard, Estelle Dominati, Ida Kubiszewski, Sue Ogilvy, Luciana Porfirio, et al. A review of methods, data, and models to assess changes in the value of ecosystem services from land degradation and restoration. *Ecological Modelling*, 319:190–207, 2016.

Shweta Tyagi, Bhavtosh Sharma, Prashant Singh, and Rajendra Dobhal. Water quality assessment in terms of water quality index. *american Journal of water resources*, 1(3):34–38, 2013.

Jordan Ubbens, Mikolaj Cieslak, Przemyslaw Prusinkiewicz, and Ian Stavness. The use of plant models in deep learning: an application to leaf counting in rosette plants. *Plant methods*, 14(1):6, 2018.

Jordan R Ubbens and Ian Stavness. Deep plant phenomics: a deep learning platform for complex plant phenotyping tasks. *Frontiers in plant science*, 8:1190, 2017.

Hideaki Uchiyama, Shunsuke Sakurai, Masashi Mishima, Daisaku Arita, Takashi Okayasu, Atsushi Shimada, and Rin-ichiro Taniguchi. An easy-to-setup 3d phenotyping platform for komatsuna dataset. In *Computer Vision Workshop (ICCVW), 2017 IEEE International Conference on*, pages 2038–2045. IEEE, 2017.

Dagmar van Dusschoten, Ralf Metzner, Johannes Kochs, Johannes A Postma, Daniel Pflugfelder, Jonas Bühler, Ulrich Schurr, and Siegfried Jahnke. Quantitative 3d analysis of plant roots growing in soil using magnetic resonance imaging. *Plant physiology*, 170(3):1176–1188, 2016.

Fred van Eeuwijk, Daniela Bustos-Korts, Emilie J Millet, Martin Boer, Willem Kruijer, Addie Thompson, Marcos Malosetti, Hiroyoshi Iwata, Roberto Quiroz, Christian Kuppe, et al. Modelling strategies for assessing and increasing the effectiveness of new phenotyping techniques in plant breeding. *Plant Science*, 2018.

James Edward Vanderplank. *Disease resistance in plants*. Elsevier, 2012.

Manuel Vázquez-Arellano, Hans W Griepentrog, David Reiser, and Dimitris S Paraforos. 3-d imaging systems for agricultural applications—a review. *Sensors*, 16(5):618, 2016.

Harry Vereecken, Andrea Schnepf, Jan W Hopmans, Mathieu Javaux, Dani Or, Tiina Roose, Jan Vanderborght, MH Young, Wulf Amelung, Matt Aitkenhead, et al. Modeling soil processes: Review, key challenges, and new perspectives. *Vadose Zone Journal*, 15(5), 2016.

Fernando Vicente-Guijalba, Tomas Martinez-Marin, and Juan M Lopez-Sanchez. Crop phenology estimation using a multitemporal model and a kalman filtering strategy. *IEEE Geoscience and Remote Sensing Letters*, 11(6):1081–1085, 2013.

Zijian Wang, Ze Zhao, Dong Li, and Li Cui. Data-driven soft sensor modeling for algal blooms monitoring. *IEEE Sensors Journal*, 15(1):579–590, 2014. doi:10.1109/JSEN.2014.2350497.

Tim Wark, Peter Corke, Pavan Sikka, Lasse Klingbeil, Ying Guo, Chris Crossman, Phil Valencia, Dave Swain, and Greg Bishop-Hurley. Transforming agriculture through pervasive wireless sensor networks. *IEEE Pervasive Computing*, 6(2): 50–57, 2007.

Karin Weitbrecht, Kerstin Müller, and Gerhard Leubner-Metzger. First off the mark: early seed germination. *Journal of experimental botany*, 62(10):3289–3309, 2011.

Sarathi M Weraduwage, Jin Chen, Fransisca C Anozie, Alejandro Morales, Sean E Weise, and Thomas D Sharkey. The relationship between leaf area growth and biomass accumulation in arabidopsis thaliana. *Frontiers in Plant Science*, 6:167, 2015.

Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

Sjaak Wolfert, Lan Ge, Cor Verdouw, and Marc-Jeroen Bogaardt. Big data in smart farming–a review. *Agricultural Systems*, 153:69–80, 2017.

Weicheng Wu, Claudio Zucca, Ahmad S Muhaimeed, Waleed M Al-Shafie, Ayad M Fadhil Al-Quraishi, Vinay Nangia, Minqiang Zhu, and Guangping Liu. Soil salinity prediction and mapping by machine learning regression in c entral m esopotamia, i raq. *Land degradation & development*, 29(11):4005–4014, 2018.

Kerui Xia, Haibo Gao, Liang Ding, Guangjun Liu, Zongquan Deng, Zhen Liu, and Changyou Ma. Trajectory tracking control of wheeled mobile manipulator based on fuzzy neural network and extended kalman filtering. *Neural Computing and Applications*, 30(2):447–462, 2018.

AS Yakovlev, MA Kaniskin, and VA Terekhova. Ecological evaluation of artificial soils treated with phosphogypsum. *Eurasian soil science*, 46(6):697–703, 2013.

Petrushin A.F. Yakushev V.P., Lekomtsev P.V. Tochnoye zemledeliye: opyt primeneniya i potentsial razvitiya. *Informatsiya i kosmos*, 3:50–56, 2014.

Xiaoyuan Yang, Alan H Strahler, Crystal B Schaaf, David LB Jupp, Tian Yao, Feng Zhao, Zhuosen Wang, Darius S Culvenor, Glenn J Newnham, Jenny L Lovell, et al. Three-dimensional forest reconstruction and structural parameter retrievals using a terrestrial full-waveform lidar instrument (echidna®). *Remote sensing of environment*, 135:36–51, 2013.

L. Yao and Z. Ge. Deep learning of semisupervised process data with hierarchical extreme learning machine and soft sensor application. *IEEE Transactions on Industrial Electronics*, 65(2):1490–1498, Feb 2018. ISSN 0278-0046. doi:10.1109/TIE.2017.2733448.

Lin Yuan, Yanbo Huang, Rebecca W Loraamm, Chenwei Nie, Jihua Wang, and Jingcheng Zhang. Spectral analysis of winter wheat leaves for detection and differentiation of diseases and insects. *Field Crops Research*, 156:199–207, 2014.

D Yumeina, GK Aji, and T Morimoto. Dynamic optimization of water temperature for maximizing leaf water content of tomato in hydroponics using an intelligent control technique. In *V International Symposium on Applications of Modelling as an Innovative Technology in the Horticultural Supply Chain-Model-IT 1154*, pages 55–64, 2015.

Jianming Zhou, Guoxiang Gu, and Xiang Chen. Distributed kalman filtering over wireless sensor networks in the presence of data packet drops. *IEEE Transactions on Automatic Control*, 64(4):1603–1610, 2018.

Lianjie Zhou, Nengcheng Chen, Zeqiang Chen, and Chenjie Xing. Roscc: An efficient remote sensing observation-sharing method based on cloud computing for soil moisture mapping in precision agriculture. *IEEE Journal of selected topics in applied earth observations and remote sensing*, 9(12):5588–5598, 2016.

X Zhou, HB Zheng, XQ Xu, JY He, XK Ge, X Yao, T Cheng, Y Zhu, WX Cao, and YC Tian. Predicting grain yield in rice using multi-temporal vegetation indices from uav-based multispectral and digital imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 130:246–255, 2017.

Jingbo Zhu, Jing Wang, Yan Ding, Baoyue Liu, and Wei Xiao. A systems-level approach for investigating organophosphorus pesticide toxicity. *Ecotoxicology and environmental safety*, 149:26–35, 2018.

Alexander V Zhulidov, Vladimir V Khlobystov, Richard D Robarts, and Dmitry F Pavlov. Critical analysis of water quality monitoring in the russian federation and former soviet union. *Canadian Journal of Fisheries and Aquatic Sciences*, 57(9): 1932–1939, 2000. doi:10.1139/cjfas-57-9-1932.

Appendix A

Additional Resources

RMSE						
		Hidden neurons				
		16	***32***	***64***	***128***	***256***
Inputs neurons	***16***	14.79	14.37	12.81	12.30	12.40
	32	12.39	13.82	11.86	12.42	11.52
	64	13.18	11.57	13.50	12.13	11.42
	128	12.10	11.35	11.48	12.65	12.00
	256	12.14	11.88	11.05	11.83	11.99

MAE						
		Hidden neurons				
		16	***32***	***64***	***128***	***256***
Inputs neurons	***16***	12.10	11.69	10.66	10.00	9.23
	32	9.45	11.18	8.99	9.58	8.45
	64	11.30	9.45	10.11	9.10	8.95
	128	9.82	8.96	9.09	10.09	9.25
	256	9.72	9.46	8.44	9.27	9.20

Figure A-1: Results of prediction errors calculation for test sub-sets depending on the amount of neurons on input and hidden layers.

Table A.1: Watering schedule, (d.w. - distillate water).

#	date	EC, mS/cm	pH
1	17.05	1.6	5.5
2	21.05	1.7	5.5
3	25.05	1.8	5.5
4	27.05	1.8	5.5
5	29.05	1.9	5.5
6	31.05	1.9	5.5
7	02.06	1.9	5.5
8	03.06	1.9	5.5
9	04.06	1.9	5.5
10	05.06	1.9	5.5
11	06.06	d.w.	d.w.
12	07.06	d.w.	d.w.
13	08.06	d.w.	d.w.
14	09.06	d.w.	d.w.
15	10.06	d.w.	d.w.

Table A.2: Measurements of EC and pH changing dynamics during the experiment.

Date	1		2		3		4		5		6	
	EC, mS/cm	pH	EC, mS/cm	pH	EC, mS/cm	pH	EC, mS/cm	pH	EC, mS/cm	pH	EC, mS/cm	pH
18.05	1.62	8.12	1.59	8.03	1.50	7.97	1.58	8.11	1.55	8.0	1.53	7.89
19.05	1.75	8.13	1.70	8.15	1.69	8.12	1.69	8.20	1.66	8.20	1.68	8.04
21.05	1.78	6.64	1.74	6.58	1.72	6.83	1.69	6.75	1.70	6.83	1.62	6.68
22.05	1.82	7.76	1.81	7.68	1.72	7.54	1.68	7.65	1.74	7.75	1.66	7.73
25.05	1.86	7.11	1.82	6.98	1.79	6.97	1.80	7.02	1.82	7.06	1.82	7.01
26.05	2.01	7.55	1.90	7.88	1.92	7.76	2.02	7.60	1.97	7.81	1.83	7.71
28.05	2.24	8.08	2.00	8.07	2.02	7.61	2.17	7.52	2.24	7.82	2.01	7.79
30.05	2.25	8.02	2.13	8.09	2.05	7.82	2.20	8.01	2.18	7.95	2.12	7.87
1.06	2.43	7.77	2.08	7.62	2.08	5.56	2.12	7.28	2.00	7.41	1.85	7.40
3.06	2.74	8.17	2.36	8.15	1.96	8.43	-		-		-	
4.06	3.18	7.73	2.32	8.17	2.13	8.26	-		-		-	

Table A.3: The location and selected surface soil properties (expressed on a dry-weight basis) for sites of the Sakhalin island

#	Soil type according to WRB	pH_{H2O}	LOI (%)	N	P	K	Clay content (%)	$TPH_{initial}$ mg/kg
1	Fibric Histosols Dystric	5.57	97.18	42.40	2.00	23.60	0.00	2908.00
2	Rustic Podzols	5.50	86.11	39.60	2.00	37.40	0.00	5139.00
3	Carbic Podzols	4.30	81.95	72.30	79.00	225.40	0.00	3139.67
4	Histic Podzols	5.21	35.44	27.30	34.20	247.20	0.00	4088.00
5	Luvic Stagnosols Dystric	5.65	3.60	15.30	10.00	26.40	10.90	713.00
6	Histic Gleysols Dystric	5.10	81.09	63.60	113.10	543.00	0.00	1770.33
7	Fibric Histosols Eutric	4.89	87.71	67.30	94.70	383.80	0.00	1063.67
8	Umbric Fluvisols Oxyaquic	5.52	1.28	31.30	56.50	122.70	3.80	1188.33
9	Umbric Fluvisols Oxyaquic	5.50	67.04	290.10	516.30	616.30	22.97	627.33
10	Haplic Cambisols Dystric	4.50	47.36	181.10	501.90	179.80	27.60	370.67
11	Umbric Fluvisols Oxyaquic	5.00	41.57	71.90	1184.50	441.00	38.20	514.00