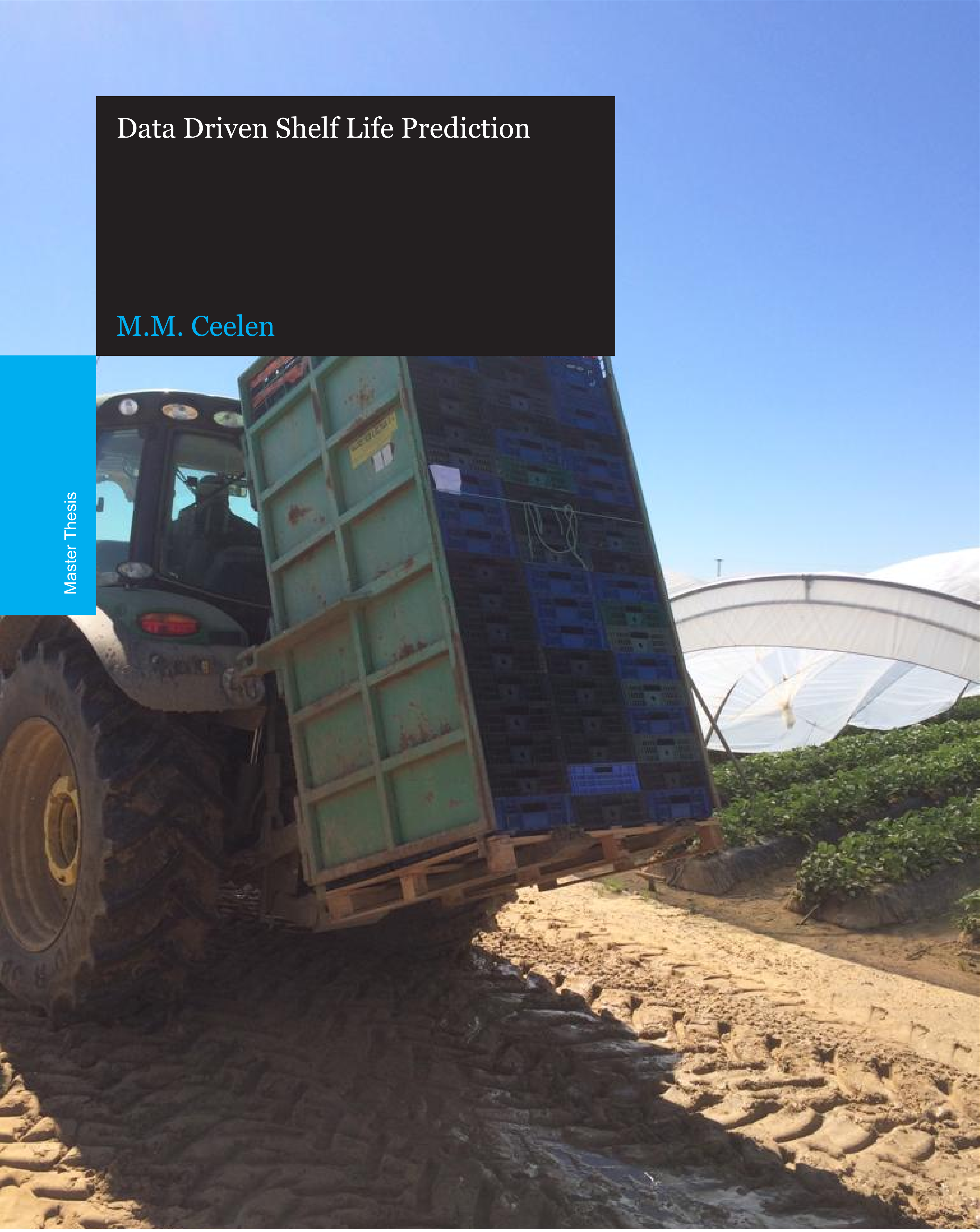
Data Driven Shelf Life Prediction

M.M. Ceelen

Master Thesis



Delft University of Technology

Challenge the future

Data Driven Shelf Life Prediction

by

M.M. Ceelen

in partial fulfillment of the requirements for the degree of

Master of Science
in Mechanical Engineering

Specialisation: Transport Engineering and Logistics

at the Delft University of Technology,
to be defended publicly on Thursday July 18, 2019 at 11:00 AM.

Supervisor:	Dr. ir. H. Polinder	
Daily supervisor:	Ing. V. Garofano	
Thesis committee:	Prof. dr. R. R. Negenborn,	TU Delft
	Dr. A. J. Laguna,	TU Delft
	Ing. D. Deckers,	The Greenery

Report Number: 2019.TEL.8352

TU Delft Delft University of Technology

Preface

This thesis is the final academic project before obtaining a MSc degree in Mechanical Engineering with the specialization in Transport Engineering Logistics. This graduation project is performed at The Greenery in the period from December 2018 till June 2019. The focus of this research is to improve the distribution of fruit and vegetables based .

The Greenery is at the beginning of a transformation from a formal auction to a digital chain manager. This project is a step in the direction of digitization of the supply chain for fruits and vegetables. The results of this research show potential for further research the application of artificial intelligence in the supply chain.

Besides The Greenery, the distribution center and local supermarkets of PLUS where involved in this research. The ability to measure through the whole supply chain was a great challenge but necessary for a successful progression of this research. I'm truly grateful for the support of all participants in the supply chain.

I have to thank Daan Deckers and Vitorrio Garofano for your supervision, enthusiasm and patience. Both of you help me by giving the right amount of freedom and structure when needed.

At last I would like to thank my friends and family for all the support they gave me. It was great to have such a support during my thesis and the years spend at the University.

Marcel Ceelen
Delft, July 2019

Contents

Abstract

In supply chain of perishable food products, like fruit and vegetables, large losses are incurred between the grower and the consumer. Given the ever growing population, and the limited land resources the food losses need to be reduced. Throughout the supply chain, from initial agricultural production down to final household consumption, food is lost or wasted. The mean reason for food waste in the supply chain is deterioration of products during distribution. In the worst case the expiring date is pasted before it arrives at the customer. The waste of product includes a waste of water, energy and soil which is used to produce and transport these products.

One of the reasons for food waste in the fruit and vegetable supply chain are traceability, food date labeling and human quality grading. Those three subjects can be optimised by an accurate and objective prediction of the shelf life of a particular product. In this research the shelf life of a strawberry is predicted by artificial intelligence. The choice for artificial intelligence is based on the fact that shelf life of a products is depends on a large variety of variables. A combination of these variables influences the shelf life period. This combination is always unique. A way to process these unique variables is artificial intelligence. A part of artificial intelligence is machine learning. Machine Learning is a recent, modern technique for data analysis, with promising results and large potential. Different machine learning algorithms are used to predict self life based on temperature measurements through the supply chain.

The prediction of shelf life is based on temperature measurements from the moment the package of strawberries is harvested till the moment this same package is bought by a customer in a local PLUS supermarket. The strawberries are harvested in the south of Spain, near Huelva and distributed to local PLUS Supermarkets near Rotterdam. After the packages with strawberries, including temperature loggers, vised the supermarket shelf. The packages with temperature logger are moved to a shelf life room for visual inspection. During this daily inspection, the actual shelf life of the strawberries is determined by a classified inspector.

The combination of the actual shelf life and the temperature profile through the supply chain is used to train, validate and test different machine learning algorithms. These machine learning algorithms are compared to each other and the best performing algorithms are tested on new data. Based on the prediction accuracy of these algorithms the most reliable algorithms is selected.

The most reliable shelf life prediction algorithm is the Exponential Gaussian Process Regression Algorithm, with the smallest confidence interval and an average deviation of 14.1 %. Finally, the possible improvements in the supply chain based on shelf life prediction, like traceability, food date labeling and quality grading are evaluated.

List of Figures

List of Tables

1 Introduction

1.1. Research Motivation

A large portion of what is produced is never consumed. Throughout the supply chain, from initial agricultural production down to final household consumption, food is lost or wasted. In the case of fresh fruits and vegetables, the food loss can reach as high as one-third of the food produced for human consumption [7]. This means that huge amounts of the resource used in food production are used for nothing. Food losses represent a waste of resources used in production such as land, water, energy and other inputs.

The challenge to deal with increasing food shortages for an ever expanding world population. With a predicted increase of 1.7 billion in world population between now and 2050 [1], the limited resources need to be used efficiently and product waste needs to be reduced.

The different stages in which food waste is considered are primary production, manufacturing, consumer facing business and households, see figure 1.1.

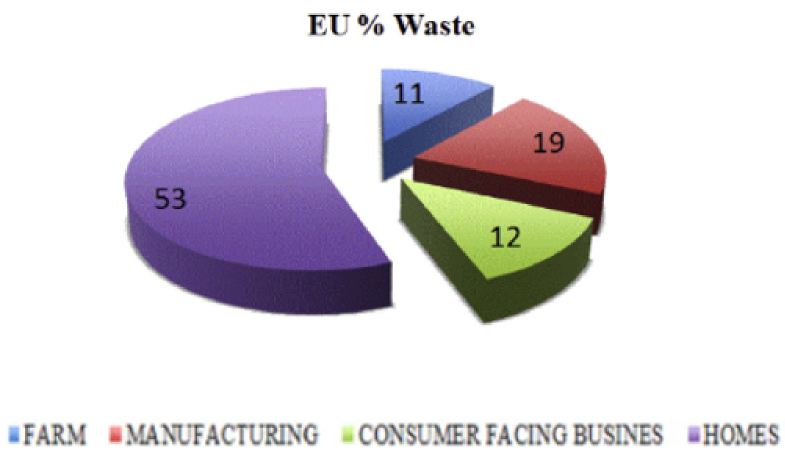


Figure 1.1: Main sources of food waste in the EU [1]

Improvement in global supply chain management have potential to reduce food waste as well as the cost of the consumer given that the cost of food waste is increasing to wards the consumer end of the supply chain. A company participating in global supply chain management of fruit and vegetables is The Greenery. The Greenery distributes fruit and vegetables around the world from growers to supermarkets. A major costs item are the product returns of dissatisfied customers. A significant reason for these returns is the quality of the product. The shelf life period of the products are not as expected. The cost made by the Greenery for each product based on incorrect quality determination or deterioration in the supply chain are given in figure 1.2.

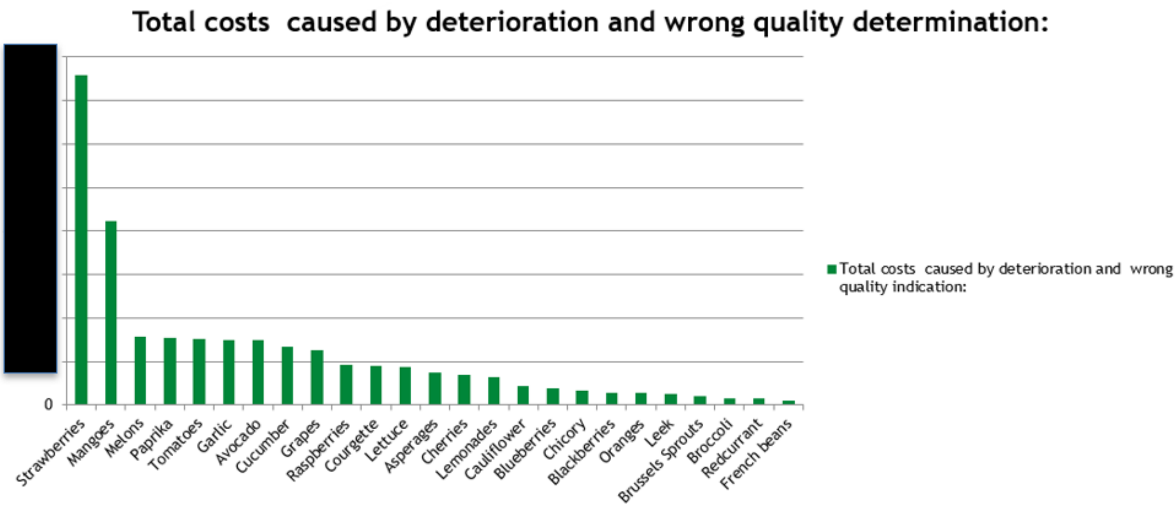


Figure 1.2: Total return costs in euros for each product based on incorrect quality determination or deterioration in the supply chain.

The highest costs for incorrect quality determination or deterioration are due to strawberries, see figure 1.2. Strawberries are a valuable product which quality depends on the conditions in which it is grown, harvested, transported and packaged. By evaluating the strawberry supply chain, improvements can be made in the traceability, food date labelling and the quality grading.

1.1.1. Traceability

The first possible improvement of the strawberry supply chain is traceability. The quality and shelf life of strawberries depend on the growing and distribution conditions. At the moment a strawberry is harvested, the quality of the product is only degrading. To enlarge the shelf life of the strawberry good transport and handling conditions are critical [8].

Improvements can be achieved by implementation of chain wide monitoring systems [9] that facilitate the exchange of information about the product conditions. This development will lead to a higher transparency and security of the food supply chain resulting in a reduction of the global food waste [1].

1.1.2. Food Date Labelling

The second possible improvement is food labelling. The current way of food labelling is one of the reasons for food waste in the final stage of the supply chain. Food chain operators revealed that food producers set best-before dates very conservatively and retailers will not sell product which have passed that date to limit their risk in terms of product liability and potential damage to reputation [10].

To reduce food waste based on the best-before date, a solution could be to enlarge the best-before dates or totally remove them for stable foods like salt, sugar, tea, coffee or rice [11]. But there are concerns [12] that relaxing the rules could weaken the quality and safety of products. The same research concludes that the absence of any shelf life information can possible lead to more food waste because consumers can no longer recall the date of purchase.

A possible solution could be to generate a best-before date based on true shelf life. An innovative suggestion is to integrate intelligent labels. A time-temperature label has the capability to provide information on temperature fluctuations that temperature sensitive food products have experienced. Commercially available labels are based on color change [13] or use data loggers which save the temperature distribution through the supply chain [14]. The consumers perception of time-temperature indication in food packaging is rather open and positive [15]. The availability of product condition data can be used to get more insight into the expiring date of the product.

The knowledge about the expiring date can be used for dynamic pricing. Products which are almost expired should get a price reduction. An idea is to implement the expiring dates into barcodes for dynamic pricing [16], allowing automatic price reduction at the checkout. Another application [17] is launched which connects consumers with food at restaurants and groceries to share fruit and vegetables which are close to the expiring date and would otherwise end up as waste. These initiatives are started to make sure the food is consumed and not wasted.

1.1.3. Human Quality Grading

The third improvement in the strawberry supply chain is based on human quality grading. The currently processed strawberries are non-uniform and go through a supply chain which has variation in the environmental circumstances. The current way of grading the quality of a strawberry is based on experience of inspectors. This grading of quality is based on human knowledge. Manual methods for food quality assessment are challenging even for people who are trained to perform these tasks. The limiting time and large amounts of products lead to quick decision making and large scale evaluation [18].

So far, it is not succeed to express this knowledge in well-established rules. The bottleneck is that frequently, human expert knowledge cannot be summarized systematically. The main advantages of expert systems have been summarized by Linko, McKinion and Lemmon [19, 20]:

- AI may help when expert advice is needed but an expert may not be available
- The expert systems is independent of human errors or moods, it has the ability to explain and justify their line of reasoning.
- Helps to verify a human expert's opinion
- The system is available 24 hours a day
- An expert system can operate in risky situations and conditions unfit for human experts.
- A system can act quickly on the basis of huge databases.
- An expert system use natural language, and requires no complex mathematical expression.

A possible way to summarize the human expert knowledge is Artificial Intelligence (AI). AI can be seen as the part of Computer Sciences that tries to simulate processes that would be described as intelligent behaviour in humans. The advantages of using AI techniques for new tools in food processes are given by Goyache [21]:

- AI techniques are adapted to working in a non-linear behaviour, which generally occurs in food processes
- AI techniques can explicitly explain what is learned.
- AI techniques can be used to indicate variables which are influencing process performance.

Artificial intelligence makes it possible to analyze massive volumes of data in a short time window and accurately. Comparable systems are already used for fraud detection [22] or fault prediction models [23]. The implementation artificial intelligence in the food industry is frequently used for image recognition for product quality analysis [24–26].

Currently done research [27] integrated a continuous monitoring system through a food supply chain. The research compares different measurements systems with each other and recommends the active RFID tags. This active RFID tags send measured data via a gateway to an online database.

A temperature monitoring system has been developed by the project FRISBEE [28]. This web based platform can predict the remaining shelf life based on the conditional data in the previous supply chain by using various scenarios in Monte Carlo simulation.

It is a big challenge to reduce the losses during transportation of perishable products. The combination of reasons for deterioration of perishable products are always unique. To measure the unique combination of reasons and to process these data researchers recommend self learning algorithms like machine learning models [29–31].

1.2. Research Problem

Throughout the supply chain, from initial agricultural production down to final household consumption, food is lost or wasted. The mean reason for food waste in the supply chain is deterioration of products during distribution. In the worst case the expiring date is pasted before it arrives at the customer. The waste of product includes a waste of water, energy and soil which is used to produce and transport these products.

There is a lot of interest from The Greenery in improving the supply chain. This is mainly for two reasons: one is the deterioration of products in the supply chain and the second is incorrect quality determination. The cost made by the Greenery for each product based on incorrect quality determination or deterioration in the supply chain are given in figure 1.2.

The highest costs for incorrect quality determination and deterioration are made in the strawberry supply chain. The strawberry is a valuable product which degrading process highly depends on the surrounding conditions. Problems in the supply chain do result in quality changes and deterioration of the product.

Another problem is the human quality grading. In the current way the inspector is responsible to determine the quality and shelf life of the strawberries. This decision is made once the product is delivered at the distribution centre of The Greenery.

These two aspects are combined into the initial problem definition stated below:

> ”The problem of The Greenery is to keep track on the quality of products through the supply chain and to determine the quality to assign products to particular sales channels or individual customers.”

1.3. Research Question

This research is focused on a reduction of food waste due to incorrect quality control and deterioration in the supply chain. Food waste in the strawberry supply chain can be reduced by improving traceability, food date labelling and quality grading. The major challenge of this research is to find out how reliable artificial intelligence can predict the shelf life of as strawberry and how it can be implemented to optimize traceability, food date labelling and quality grading.

Based on this challenge the following research question is formulated:

> ”How can artificial intelligence contribute to a better shelf life prediction of a strawberry based on temperature measurements from grower till customer?”

To evaluate the supply chain, find the influencing factors on the strawberry shelf life and to learn about the different algorithms able to predict shelf life the following sub-questions are formulated:

a) How to define the supply chain from a transport and logistics perspective?

b) Which parameters need to be measured to predict shelf life of a strawberry?

c) Which machine learning methods are able to predict the shelf life of strawberries based on historical temperature profile?

d) how it can be implemented to optimize traceability, food date labelling and quality grading?

1.4. Research Scope

Out of the range of products the Greenery is offering, the strawberry supply chain is selected. The strawberry is harvested at the moment it reached its optimal sweetness. In comparison to a banana which can be harvested green and will turn into yellow during transportation, a strawberry will become more red but will never become more sweet than the moment it is harvested. This is why the product needs to be shipped to the customer as fast as possible in conditions which enlarge the shelf life of the strawberry.

1.4.1. Measurement Scope

During the time period of this research strawberries were imported from Spain. The grower which participated in this research is growing his strawberries in the south of Spain, Huelva. At this point temperature loggers are added to the strawberries. The strawberries are transported to The Netherlands by truck and distributed at The Greenery located in Breda. The strawberries are packed and shipped to the distribution centre of PLUS, Barendrecht. From this point the strawberries are order picked and shipped to local supermarkets. At the local supermarket the trackers are collected and returned to Breda where the measured data is analysed. An overview of the logger movements through the supply chain is given in figure 1.3.

1.4.2. Initial Condition of Strawberries

The scope of this research is from the moment the strawberry is harvested till local supermarket shelf. The growing process of the strawberry is not inside the measuring scope. The initial condition of the strawberries is assumed to be equal. To make sure the used strawberries are as equal as possible, one grower and one variety of strawberries is selected. In this way the used strawberries have had the same treatment during the growing process.

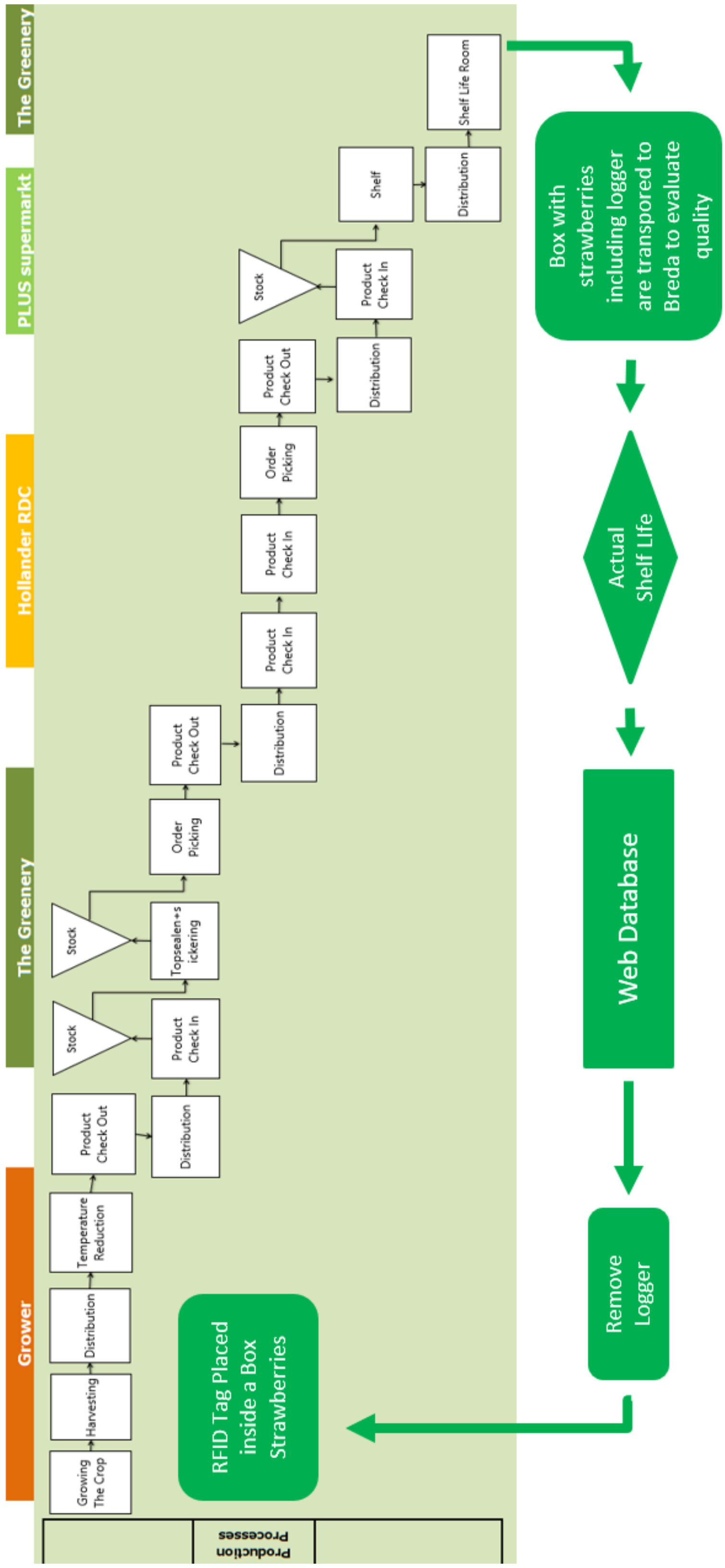


Figure 1.3: Tracker flow overview through the strawberry supply chain.

1.5. Research Objective

This research is executed in collaboration with The Greenery and aims to improve the supply chain of fruit and vegetables. More specifically this research aims to use artificial intelligence to predict shelf life of a strawberries based on temperature measurements in the strawberry supply chain from grower till retail customer. These shelf life predictions can help to improve traceablity, quality grading and product labeling and reduce the costs for in correct quality assignment and deterioration in the supply chain.

For Delft University of Technology the research goal is to get more insight in the distribution of perishable products. Different artificial algorithms are tested and on the measurement data samples to get the most accurate prediction. This results will contribute to knowledge about data research in supply chains.

1.6. Report Structure

The outline of the report is as follows. Chapter 2, gives an overview of the strawberry supply chain. The supply chain from grower till retail customer is evaluated. Chapter 3, 'Shelf Life Prediction' explains the variables who have influence on the shelf life of a strawberry. These variables are input for the artificial intelligence algorithms. The artificial intelligence algorithms used to predict the shelf life of strawberries and the selection criteria for these algorithms are evaluated in chapter 4, ”Predicting Algorithms”. The research mythology is explained in chapter 5, ”Research Methodology”. The results of the measurements through the supply chain and the shelf life predictions of the artificial intelligence algorithms are given in chapter 6, ”Results”. The answer to the research question is finally given in chapter 7, in combination with the recommendations given for further research.

A total of 4 appendices are completing this research. The first appendix, Appendix A, includes the features used to train, validate and test machine learning algorithms. Appendix B contains plots to show the correlation between features and shelf life based on measurement results. Appendix C contains a table of the deviation values of the best performing machine learning algorithms between the predicted shelf life and the actual shelf life. Appendix D, specifies an example of the shelf life observations of strawberries made by the inspector. The report is completed with Appendix E, describing the neural network selection process by using the Neural Network fitting application of Matlab.

2
Supply Chain Overview

2.1. Company Profile

The Greenery is an international sales organization which specializes in fresh fruit and vegetables. The company supplies a full range of fresh fruit and vegetables to supermarkets, wholesalers, caterers and the food processing industry all year round.

The currently known company was established in 1996 after a fusion of 9 fruit auctioneers. By that time it was used as an auction for fruits. From that moment the company transformed into the distribution centre, the organization has acquired a number of robust trading companies and the old auctioning method (the clock) was gradually replaced by a marketing and sales organization.

2.2. Products

The Greenery offers a comprehensive assortment of fresh fruit, vegetables and mushrooms. Products which are not available in the Netherlands are imported such as citrus fruits, bananas and other exotic fruits. In this way The Greenery makes sure that every product can be supplied twelve months a year.

In the distribution of all these products the aim of the company is to make sure the customer receives the products with the quality it is expecting. By incorrect quality determination or deterioration of products during distribution, products do not meet the expectations of the customer. When products do not meet the expectations of the customer, the customer will return the products or ask for a price reduction.

To get an indication of the costs of the problem a costs analysis is made and presented in figure 2.1. The total costs generated in 2018 for each reason are given in the diagram. The return shipments represent the highest expenses.

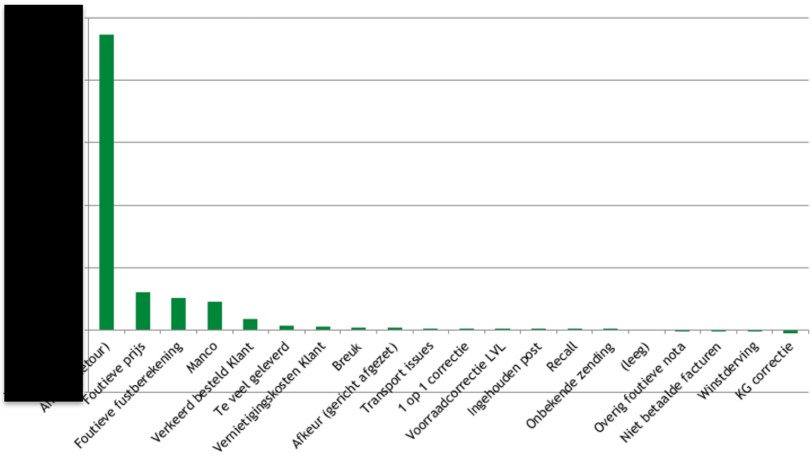


Figure 2.1: Complain Analysis Diagram.

Figure 2.2 is an overview of the reasons and costs of return shipments over the year 2018. The figure points out that the return shipments are mainly caused by depraved fruits or vegetables. On the second place, the costs for misjudgement of the quality related to the product, . On the third place the reason of return is the inconsistency of dimensions of the product with a total expenses of –.

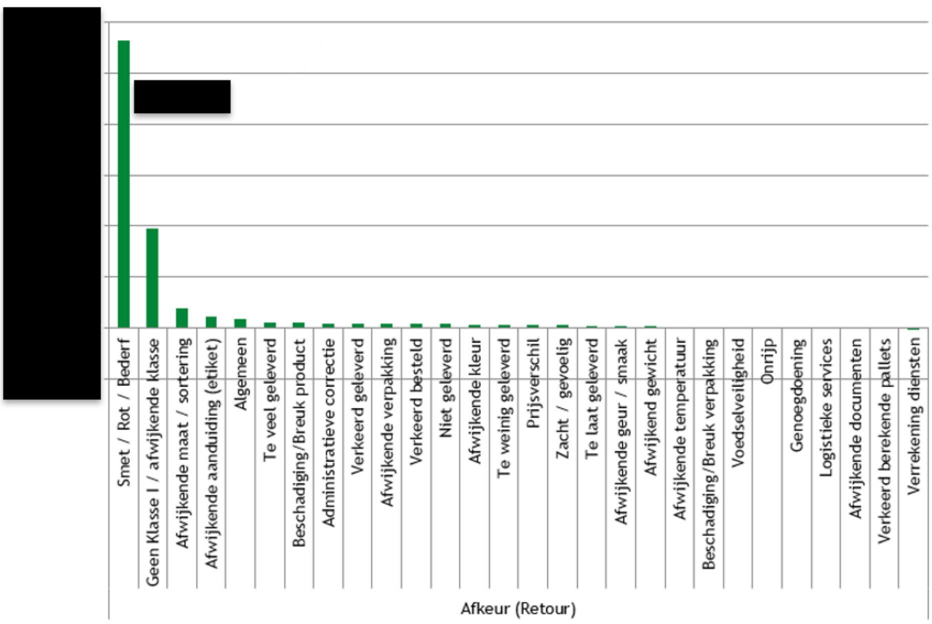


Figure 2.2: Reason and costs of return shipments.

The next step is find out which products are responsible for the return shipment costs. Figure 2.2 shows that the costs of return shipments caused by product rotting and incorrect quality determination.

Supply Chain Selection

To find out which products has problems with deterioration and incorrect quality determination. The costs for deterioration and incorrect quality determination are added for all products and presented in figure 2.3.

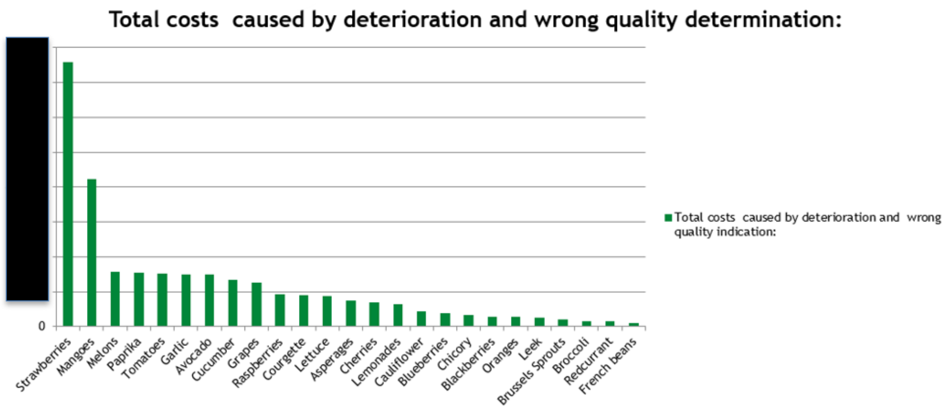


Figure 2.3: Total costs for each product for deterioration or wrong quality determination.

The highest costs are made for strawberries. This is the reason why this product is selected for this research. It is a highly perishable and valuable product, which needs to be distributed in optimal conditions.

2.3. Strawberry Supply Chain

In this section the strawberry supply chain (fig.2.4) is defined. The current supply chain consist of growers, warehouses and retailers. The grower is the manufacturer which will harvest the strawberries. From this point the strawberries are loaded into refrigerated trucks and transported to a distribution centre. The strawberries will be unloaded and stored into cold store. The next step is to sort and wrap the strawberries wit a customer specific package and label. The packed strawberries wait in the cold storage to be transported to the retailers.

2.3.1. Grower

The supply chain of a strawberry starts at the grower. The strawberry is grown and harvested at the grower. From the moment that the strawberry is harvested the decay process starts. To slow down the deteriorate process, the temperature needs to be reduced to 3 degrees Celsius. This is done at a cold store. The cold store is at the same time the storage place where the harvested strawberries wait for transport to the distribution centre.

Harvesting

The harvest conditions do influence the yield, quality, and storage life. For example, harvest is better carried out on sunny or cloudy days, but not when it is raining. To avoid bruising, mechanical damage strawberries are harvested directly in boxes preferred by the customer. In case of this research the packages have a size of 5,5 x 18 x 11 cm and the mass is 500 grams strawberries for each package.

Transportation

To keep the temperatures throughout the whole process of transport at appropriate low temperatures, the strawberries are transported through a cool-chain. The cool chain is the definition of a supply chain in which the temperature is controlled from grower till the customer. To maximize shelf life, a reduced temperature is necessary during the whole process of transport, storage, selling and consumption.

2.3.2. Distribution Centre

By cold truck the strawberries are transported to the distribution Centre. The truck load will be checked by a inspector before it enters the cold storage of the distribution centre. The Strawberries leave the cold storage to be packed for retailers. Packed for retailers the strawberries wait in the cold store for the cold truck which will transport them to the retailers.

Grading

Grading of the fruit is essential for marketing [32]. The current way of grading is based on manual checks. The grades are based on certain criteria, for example color, to achieve commercial standardization. The strawberries are affected by a variety of factors during growth and development, for this reason it is inevitable that they are of different sizes and qualities. By grading the strawberries the price of the products for the growers is specified. Beside the price for the growers, the planning for storage, packing, and marketing is determined [32]. In the current system this information is used to assign the product to a customer with particular specifications.

Packaging

The strawberries are harvested directly in package material which is used through the whole supply chain, from grower to retailer. The package is designed to prevent the fruit from damage during transport and distribution and it should be commercial attractive to customers.

2.3.3. Retailer

The cold truck with strawberries will arrive at the distribution centre of a retailer. At this point the strawberries will be divided over supermarkets. Daily deliveries supply supermarkets with the ordered products. The products which need to be cooled are stored in a cold store before they are presented to the customers.

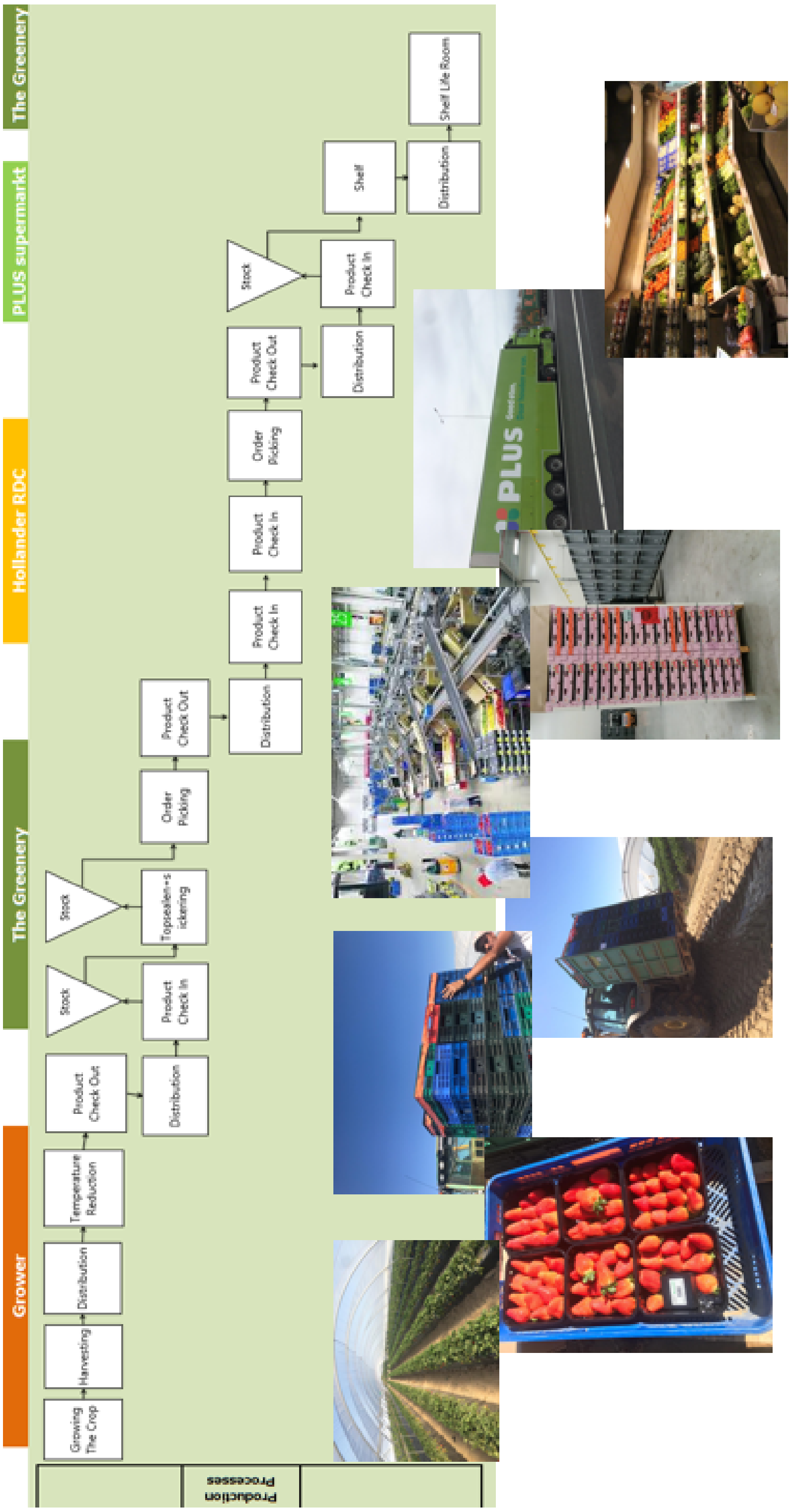


Figure 2.4: Schematic Product flow chart, with pictures of each stage in the strawberry supply chain.

3
Input

To predict the shelf life of a strawberry using artificial intelligence, the algorithm needs input data to be trained on. This input data consist of features. The features have a relation with the output of the algorithm, in this case the predicted shelf life. The stronger this relation, the higher the predictive power. Based on literature study and interviews, relations between shelf life and input features are determined. The potential features need to be measured, so finally the feasibility of measuring the potential features through the supply chain is examined.

The shelf-life of products is not a static parameter but is a highly dynamic variable. The product shelf life is affected by conditions in the environment surrounding the product. A schematic overview of data flow for shelf life prediction is given in figure 3.1. The environmental conditions are the inputs for the shelf life prediction process. The prediction of shelf life are the output of the system. The Artificial Intelligence, visualized as the box, transfers the incoming data into a shelf life prediction.

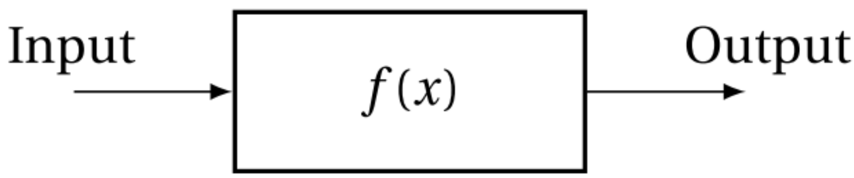


Figure 3.1: Data flow of Strawberry Shelf Life Prediction.

3.1. Shelf Life Influencing Parameters

The first stage of this chapter is focused on the shelf life influencing parameters. This stage will start with a literature survey based on this subject. Besides a literature survey, interviews with experts about input parameters are executed. The information out of these interviews is evaluated.

3.1.1. Literature Survey Shelf Life Prediction Of Strawberries

In the book "Discovering the Future: Modelling Quality Matters", Pol Tijskens describes the most important environmental factors which affect agricultural products [33]. The most important environmental factors described by Pol Tijskens are temperature, relative humidity, and the concentrations of oxygen, carbon dioxide and ethylene. These factors are evaluated in the next section.

3.1.2. Temperature

A cold chain or temperature controlled supply chain provides the essential facilities and methods required to maintain the quality and quantity of foods. Since foods can be time and temperature sensitive in nature, they need to be properly taken care of in terms of harvesting, preparation, packaging, transportation and handling in other words, throughout the entire chain. Temperature is the most important factor in prolonging or maintaining the self life of perishables [34].

Temperature control throughout the logistic chain can optimize the shelf life of a product. Temperature is the main factor affecting all (bio)chemical processes through its effects on activation enthalpy and entropy of the under-laying reactions. This counts for enzymatic and non-enzymatic reactions and therefor applies to a wide range of fresh and processed food products. A normal perception is that low temperatures will extend the shelf life of a particular product, but not all product will extend their shelf-life at low temperatures, some product (e.g. most tropical fruits) are sensitive to low-temperature decay, and will reduce shelf-life [35]. In case of strawberries, the shelf life depends on the ambient temperature [2], figure 3.2. A strawberry is stored at low temperatures around 3 degrees to extend the shelf life. Beside the ambient temperature also fluctuating ambient temperatures have an influence on the shelf life of a strawberry.

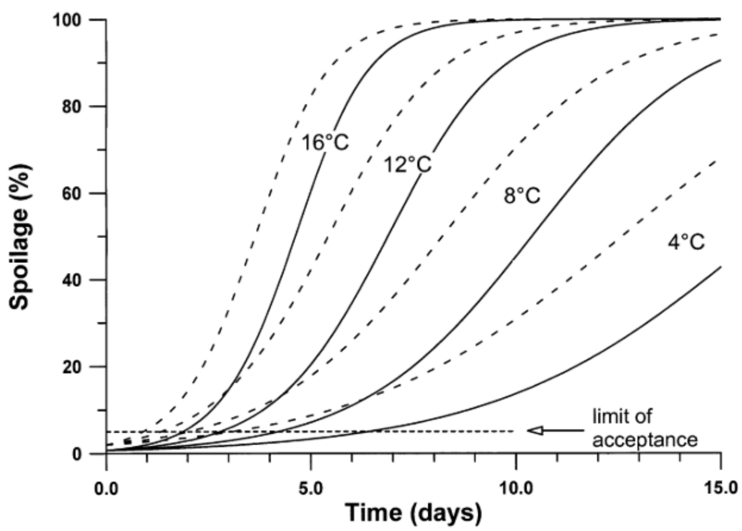


Figure 3.2: Spoilage of 'Elsanta' strawberries as a function of time and temperature as predicted for two different batches of 'Elsanta' strawberries [2].

3.1.3. Fluctuating Temperature Profile

Strawberries with a fluctuating temperature profile will have a reduced shelf life, while those form the semi-constant temperature profile were still have a better shelf life [36]. To estimate the shelf life of a strawberry, the conditions of each transport should be measured to predict the shelf life of the specific batch of strawberries concerned [2].

3.1.4. Humidity

A side effect of mechanical cooling is drying of the air, inducing weight loss and often also affecting the product's appearance through shrivelling. On the other hand, too high humidity can induce moistening and causes rot of many fresh and processed food products. For this reason, relative humidity is the second most important factor affecting quality [37]. No literature is found on the effects of relative humidity on the shelf life of strawberries [38]. But other literature show the effect of Relative Humidity (RH) on the quality and shelf life of tomatoes, bell peppers and bananas:

In case of tomatoes, the relative humidity had a significant effect on the shelf life and over all quality. Tomatoes stored at relative humidity lower than 88 percent were softer, with stems that appeared to be less fresh, maximum shelf life and quality was obtained when tomatoes were stored at 92 percent [39].

The maximum self life of bell peppers was obtained with an relative humidity around 90 - 95 percent. The bell pepper had a better texture and higher nutritional value compared to those exposed to low relative humidity [40].

During the ripening of bananas, it is more important to maintain the temperature at or above 15 degrees than to control the level of relative humidity. The only effect of relative humidity was on weight loss, most likely from the peel [41].

Temperature and relative humidity are critical in minimizing the difference in water vapour pressure between the product and the environment. To minimize the water vapour pressure deficit, the humidity of the surrounding environment should be maintained. The rate of fruit or vegetable transpiration, and thus loss of moisture, can be reduced by raising the relative humidity, by lowering the air temperature, by minimizing the difference between the air temperature and the fruit temperature, by reducing air movement and by protective packaging.

3.1.5. Atmosphere

One of the most effective methods for preserving the quality of fresh fruit and vegetables is to store them in controlled atmosphere or modified atmosphere. The optimum gas mixture depends on the kind of product. In case of the strawberry, a concentration of CO_2 > 20 percent can cause development of off-flavors or changes the color from red to dark red or purple. Since 20 percent CO_2 can cause undesirable changes in strawberry quality, an atmosphere of 5 percent O_2 + 15 percent CO_2 was selected as controlled atmosphere. Results indicate that cold air storage is more effective at 4 degrees than at 10 degrees in maintaining strawberry quality [42].

Another research after controlled atmosphere storage technology to extend shelf life indicate that wild strawberry shelf life can be effectively increased by exposing the fruits to a cold environment and an adequate atmosphere composition. The favourable atmosphere composition is 10 percent CO_2 and 11 percent O_2. This atmosphere appears to prevent fungal growth, reduce the change in soluble solids content and acidity, and maintain an acceptable aroma composition [43]

3.1.6. Interviews about Shelf Life Prediction Of Strawberries

During this research several interviews are conducted. Interviews with strawberry growers, strawberry traders, distributors and inspectors gave the following information about input parameters.

The interviews with growers and experts concluded the following input parameters:

Grower

- **Grower**, every grower follows different ways to grow their products and this will influence the product characteristics.
- **Grow technique** The product can be produced in open air, in tunnels or a greenhouse.
- **Grow technique** The soil which is used to grow the product. A soil, like sand or clay will influence the product characteristics.
- **Temperature** The temperature at the moment the product is harvested.
- CO_2 **rate** in case of strawberry production in greenhouses present, influences the quality of the strawberry.
- **Species**, every species has different characteristics. A strawberry with a high solid rate has better conservation properties.
- **Light intensity** The light availability during the growing process of the strawberry influences the quality of the product.
- **Quality Number**, the quality number is defined at the grower and is checked at the moment the product arrives at the distribution centre. The visual check is done by a experienced inspector.

The interviews with strawberry traders, distributors and inspectors gave the following information about the transport conditions which influence the shelf life of a strawberry:

Transport

- **Temperature conditions** form the moment a strawberry is harvested. The temperature of the strawberry is reduced till 3 degrees Celsius. The higher the ambient temperature the shorter the shelf life of an strawberry. This measurements in this research can be used to see the effects of temperature on the shelf life during the storage and transport to the customer.
- **Transport time** of the product to the customer has influence on the remaining shelf life of the product when it arrives at the customer. The transport time should be kept as short as possible.
- **Humidity conditions** influence the deteriorate process of the product. A high humidity rate will reduce the shelf life of a strawberry.

3.2. Data Logger Selection

The previous section described the input parameters which influence the shelf life of a strawberry based on literature survey and interviews. The scope of this research from starting at the harvest of the strawberry and ending at the moment a consumer buys the strawberries reduces the measureable input parameters to temperature, humidity, CO_2/O_2 and weather conditions during harvesting. In the current supply chain these parameters are not measured. Which means, to measure these parameters on product level a measurement system needs to be selected. This section evaluates five different measurement systems.

Different measurement systems are evaluated to be integrated in the supply chain to measure conditional data. The systems compared in figure 3.3 are compared to each other on the ability to measure the most important conditional data, temperature, humidity and CO_2/O_2. Finally the costs of each logger and the monthly expenses are compared.

	Croptracer	Emerson (1 per pallet)	Emerson (1 per vracht)	Ahrma (1 per pallet)	EPS (1 per krat)	KPN / The Greenery
Traceability	Barcode	Real-Time (Go-Wireless)	Real-Time (Flex Tracker)	Near Field Communication	Not available	Real-Time
Temperature	-	+	+	+	-	+
Humidity	-	-	+	+	-	+
CO_2 / O_2	-	-	+	-	-	-
Logger Costs (€)	-	13-23	80	7,50/0,02 (per dag)	-	70
Gateway Costs(€)	-	-	-	540	-	-
Monthly Costs(€)	-	-	35	(200 x 0,02x30) = 120	-	20
Total Costs(€) (2 maanden)	-	(200 x 23 = 4600	1600 + 700 = 2300	240 + (540) = 760	-	3500 + 40 = 3540

Figure 3.3: Overview of conditional measurements systems to measure data on product level.

3.2.1. CropTracer

The traceability system of CropTracer makes use of barcodes printed on the label of products. The bar codes are scanned when the products are entering a stage in the supply chain en the conditional data in this stage is linked to the product. The barcode printing is a cheap traceability solution but the measurements are not on product level. Besides, the measurements are done in storage or packaging areas, this systems is not able to measure the product on a dock or the moment from harvest till the storage facility.

3.2.2. Emerson

Emerson has the most complete conditional measurement loggers available. The loggers are able to measure temperature, humidity and CO_2/O_2 concentrations in the air. The system is able to measure this data real time. Which means the logger sends an update of the measurements every minute to the data base. In this way the conditions of the product are updated every minute on a user friendly online dashboard. A disadvantage is that the loggers are expensive in comparison with the other data loggers and only made for single use. Which makes as test with more than 50 loggers expensive.

3.2.3. AHRMA

The AHRMA system is able to measure the temperature, and motion. There is no real-time connection with a data base but the loggers are able to store the measured data for 5 days. The logger will transmit the data to a gateway when it is in a range of 80 meters. The loggers are relative to the other systems less expensive and can be reused.

3.2.4. Euro Pool Systems

Fruit and vegetables are distributed in folding trays. These folding trays are used through the whole supply chain. The grower harvest the product in these folding trays and at the end of the supply chain the customer in the supermarket buys the product out of these trays. The trays return to a distribution centre where they will be cleaned to be reused.

The major supplier of these folding trays is Euro Pool Systems. An idea would be to implement sensors into this folding trays which are used from the moment the product is harvested till the product is presented in the supermarket to the customer. The company is testing with integrated sensors but is not able to deliver 50 folding trays with integrated sensors.

3.2.5. KPN/The Greenery

A cooperation between KPN and The Greenery is evaluated to investigate the opportunity to build a measurements system focused on the needs of this research. A reuseable logger which is able to measure real-time the temperature and humidity for about 100 hours costs about 70 euros. To use this concept for a test with the need of 50 loggers, it became to expensive.

3.2.6. Selected Data Logger

To generate a lot of data samples on product level with a small budget, the selected data logger for this research is supplied by AHRMA. The loggers fit in the 500 gram strawberry package and transmit their data wire-less to the gateway or online database. The data is easily accessible by a online dashboard.

AHRMA developed a temperature and motion logger which can store the measured data for 5 days and is able to send this data wireless to a gateway when the logger is near the gateway (+/- 80 meter).

The basic technology is Bluetooth Low Energy which has a range of 30 meter. Ahrma added a RF layout and antenna design which increases the range to 300 meter (open field) and around 80 meter (inside). The temperature logger will store all recorded data with a time stamp of the measurement if the assessed is not in range of a gateway.

The temperature logger infrastructure is transmitting data to the nearest gateway. The gateway will pas on the data to a Microsoft Azure based cloud service. The dashboard inside this cloud helps to manage the measured data.

Table 3.1: Sensor specifications

AHRMA Temperature Sensor Specifications	
Accuracy	+/- 0.5 degrees Celsius
Mearsurement Interval	10 minutes
Total Measurement Period	+/- 5 days
Signal Range	80 meter

3.2.7. Selected Parameters

The scope of this research is limited from the moment the strawberries are harvested till the customer buys the strawberries in local PLUS supermarket. Which means some parameters will not be measured, like the growing process of the strawberries, species, soil and fertilizers. The parameters inside the scope are given in figure 3.4 on the right side ”Shelf Life Influencing Parameters”.

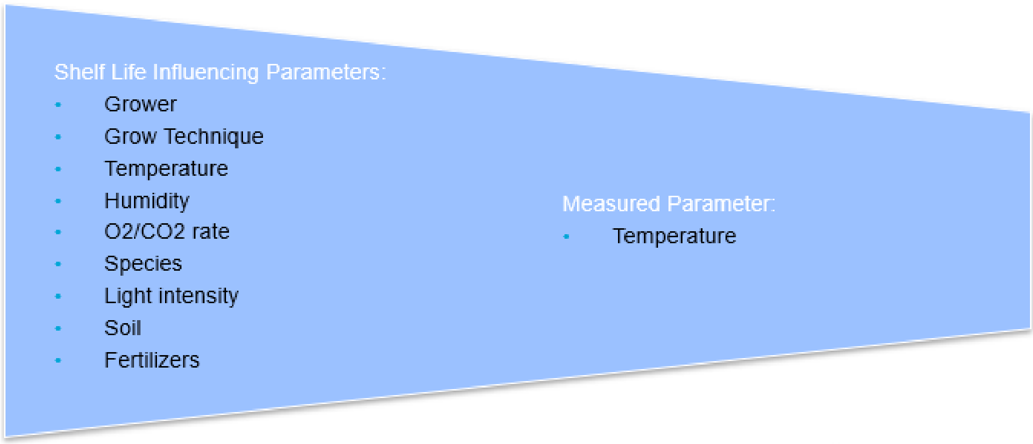


Figure 3.4: Overview of shelf life influencing parameters (left) and measureable parameters (right).

The remaining parameters, temperature, humidity, CO_2/O_2 are input parameters in the scope of this research. The budget limitation forced to choose the AHRA system. The disadvantage of this system is the limitation in measuring humidity or CO_2/O_2 concentrations. Conclusion, out of the Shelf Life Influencing Parameters, the temperature will be measured on product level through the supply chain. The reduction of measured parameters is illustrated in figure 3.4.

3.3. Feature Selection

The selection of features for machine learning algorithms is based on the measured data. Good features are informative for the machine learning algorithm. In this case the output of the machine learning algorithm is the predicted shelf life. In case of this research, an informative feature has an relation to the shelf life of the strawberry. The stronger this relation, the higher the predictive power of the model. To find informative features, the input data is analysed. Based on literature study and interviews, the following relations between the shelf life and the temperature measurements are selected.

3.3.1. Fluctuating Temperature

Fluctuating temperatures are measured when the strawberries move through the supply chain. For example when the strawberries are moved from the field to the storage facility, the temperature changes from 20 degrees to 3 degrees. If the strawberries are loaded into a truck the temperature the product is moved out of the storage facility to a loading dock. From the loading dock it is moved into a truck. The storage facility, the loading dock and the truck have different temperatures. These conditional changes are visual in the temperature profile. To amount of conditional changes can be measured by determining the amount of peaks. The indication of peaks is done in the time period from grower till the strawberries are returned to the shelf life room. The peaks are indicated by a Matlab script after the peak correction, see figure 3.5. The number of peaks is counted and saved to be used as input parameter for the final machine learning algorithm.

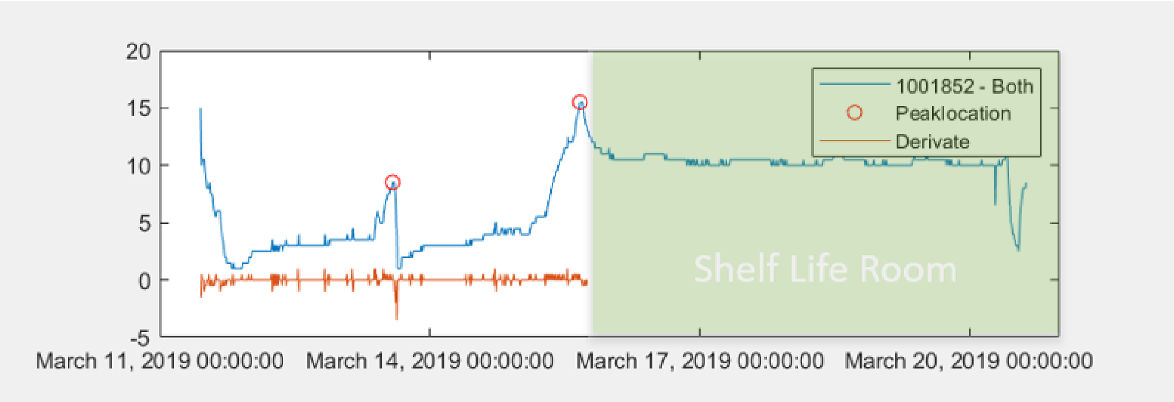


Figure 3.5: Time-Temperature profile strawberry supply chain, peaks are removed and temperature variation of 1,5 degrees or higher are indicated by a red circle.The temperature variation speed for each 10 minutes is visualised by the red line.

3.3.2. Average Temperature

Since the temperature is the main factor affecting all (bio)chemical processes through its effects on activation enthalpy and entropy of the under-laying reactions, a relation between the average temperature and the shelf life is assumed. A strawberry is stored at low temperatures around 3 degrees to extend the shelf life. Beside the ambient temperature also fluctuating ambient temperatures have an influence on the shelf life of a strawberry.

3.3.3. Reduction to Storage Temperature

The moment at which the biggest decrease of temperature is measured is the moment a strawberry is harvested. The strawberry has at that moment the same temperature as the surrounding air. Which is the outside temperature. The strawberry is harvested and transported to a cool storage. In this cool storage the temperature is decreased to the preferred 3 degrees.

Based on interviews with experts the relation between a fast temperature reduction to storage temperature and the shelf life period of a strawberry is made. Which means a faster temperature reduction to storage temperature should have a positive effect on the shelf life of a strawberry.

3.3.4. Time Spend In A Particular Temperature Zone

The temperature zone around 3 degrees Celsius is preferred for strawberry distribution, according to the literature. A constant low temperature will improve the shelf life of the strawberries. To indicate how much time between the harvest moment and the shelf life room is spend in the favourable temperature zone, the temperature profile measured by the temperature logger is divided in temperature zones.

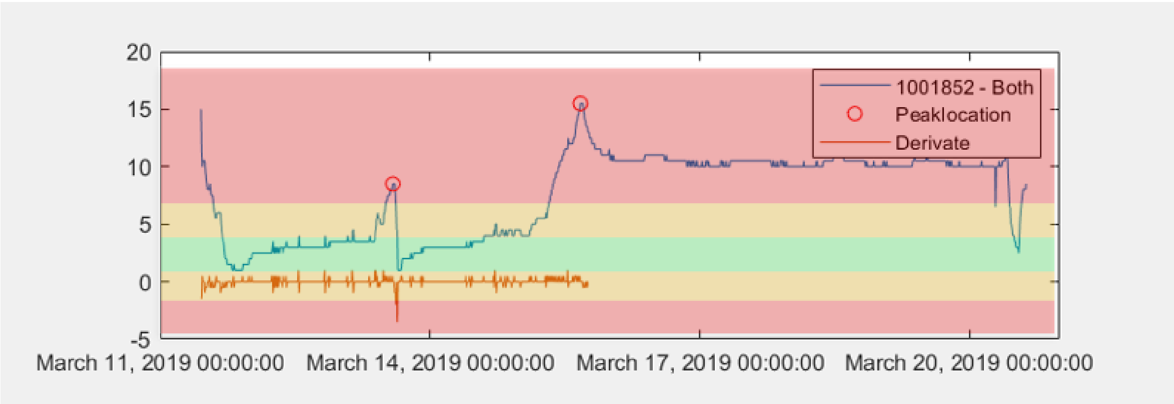


Figure 3.6: Time-Temperature profile strawberry supply chain, temperature zones are indicated by color. The green zones represents temperatures between 2 and 4 degrees Celsius, the orange zones above is between 4 and 7 degrees Celsius and the orange zone under is between 2 and -1 degrees Celsius. The red zones are above 7 degrees Celsius and under -1 degrees Celsius.

Based on literature study, the most preferable storage temperature for strawberries is around 3 degrees Celsius. In figure 3.6 the green zones is between 2 and 4 degrees Celsius. The orange zones above is between 4 and 7 degrees Celsius and the orange zone under is between 2 and -1 degrees Celsius. The red zones are above 7 degrees Celsius and under -1 degrees Celsius.

3.3.5. Selection Of Features

The selection of features for machine learning algorithms is based on literature and interviews with experts. Informative features have high predictive power. The informative features: fluctuating temperatures, average temperatures, temperature reduction to storage temperature and storage temperature will be selected as input features for machine learning algorithms to predict the shelf life of strawberries.

4 Predicting Algorithms

The features selected in the previous chapter are the inputs parameters for the shelf life prediction process. Coming back to the figure 4.1, a schematic overview of the shelf life prediction is given. The prediction of shelf life (output) depends on the measured parameters (input). The prediction algorithm (the box) transfers the incoming data into a shelf life prediction. In this chapter potential prediction algorithms are evaluated.

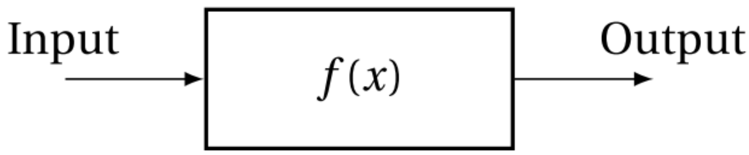


Figure 4.1: Schematic overview of data flow for strawberry shelf life prediction.

Based on this literature survey a selection is made of algorithm which are able to predict the shelf life of a strawberry based on the temperature measurements from the moment the strawberries are harvested till the moment they are bought by customers.

As described in the previous chapter, the relations between input parameters and the shelf life of a strawberry is used to predict the shelf life of a strawberry. The prediction is made by using artificial intelligence. Artificial intelligence is a name for a wide range of algorithms which all have different purposes. In this chapter the selection procedure is explained for shelf life predicting algorithms.

4.1. Artificial Intelligence

Artificial intelligence can be defined as a system which is able to correctly interpret external data, to learn from such data, and to use those learning to achieve specific goals and tasks through flexible adaption [44]. Machine learning is a sub-area of Artificial Intelligence (AI), beside reasoning, Natural Language Processing (NLP) and Planning, see figure 4.2. The different parts of artificial intelligence will be explained in the next section [3].

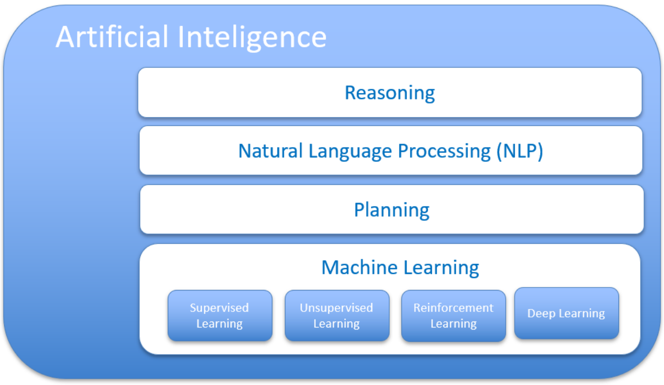


Figure 4.2: Artificial Intelligence is the overall category that includes machine learning and natural language processing [3]

- **Reasoning** allows a system to make inferences based on data. A system is never explicitly trained on a particular question, but it uses the already existing knowledge to answer the question.
- **Natural Language Processing (NLP)** is used to capture the meaning of unstructured text from documents or communication from the user. The system is able to interpret text and spoken language. Users are able to ask as system questions about complex data sets, without a need to code.
- **Planning** is the ability of a system to construct a sequence of actions to reach a final goal. Rather than a pre-programmed decision-making process with a defined order of actions.
- **Machine Learning** is one of the central topics of artificial intelligence, since it is able to learn form environmental data. Machine learning is differentiating from other sub-areas of Artificial Intelligence by the ability to learn and adapt a model based on the data rather than explicit programming.

Because the surrounding conditions of products are always different it is not possible to give a shelf life prediction based on previously enquired knowledge. Reasoning is based on previously acquired knowledge, to answer a new question [45] and makes it less suitable for shelf life prediction.

Natural Language Processing provides implementations for a range of applications. In fact, any application that utilizes text is a candidate for NLP [46]. The application of shelf life prediction is not based on text so Natural Language Processing can not by applied.

Planning is used for mobile robots, planning a travel plan, and producing a process routing for a manufacturing facility. Planning is based on a policy which will be refined based on the resulting trajectory [47]. This refinement policy is different from the machine learning algorithm which learning is based on local measurements.

Based on the fact that the shelf life prediction is based on the environmental characteristics of the product, machine learning is preferred above Reasoning, Natural Language Processing (NLP) or Planning.

4.2. Machine Learning

The definition of machine learning is the ability to learn on their own without being strictly programmed. The methodology of machine learning involves a learning process with the objective to learn from ”experience” (training data) to perform a task. An individual example is described by a set of variables. After the learning process, the trained model can be used to predict new examples (testing data) using the experience obtained during the training process.

4.2.1. Types of learning

This section explains different kinds of learning, supervised, unsupervised, reinforcement learning and neural networks. In the next section the different learning techniques are explained based on the book of Andriy Burkov [4].

- A **supervised learning** algorithm contains a dataset which is labeled. The dataset represents all conditional data. A label is connect to this conditional data and learns the algorithm.

  Classification and regression algorithms are types of supervised learning. Classification algorithms are used when outputs are restricted to a limited set of values, the output is, for example, true or false. Regression algorithms have continuous outputs, the output values are within a range. Most algorithms include linear regression, logistic regression and step wise regression. More complex regression algorithms are ordinary least squares regression, multivariate adaptive regression splines, multiple linear regression, cubist and locally estimated scatter plot smoothing [48].

- In case of **unsupervised learning**, the dataset is a collection of unlabeled examples. The output of an unsupervised learning algorithm is a real number that indicates how the unlabeled examples differ from a typical example in the dataset.

  By clustering analysis elements are categorized. Neither their structure of those clusters nor their number is known in advance. The output is completely derived by the algorithm itself, it is not possible to determine the accuracy or correctness of the resulting output. Speech recognition is an example of unsupervised learning [44].

- In **semi-supervised learning**, the data set contains both labeled and unlabeled examples. By adding more unlabeled datasets it seems like the uncertainty will increase. But adding unlabeled datasets will lead to a better probability distribution.

- **Reinforcement learning** is used for problems where decision making is sequential, and the goal is long-term, such as game playing, robotics, resource management, or logistics.

- **Neural networks** and **deep learning** are often used in image recognition, speech and computer vision applications. A neural network consist of three or more layers, an input layer, one or more hidden layers, and an output layer. A typical neural network consist of thousands or even millions of simple processing nodes that are interconnected. When a neural network consist of multiple layers, it is called deep learning. Deep learning is a machine learning technique that used hierarchical neural networks to learn from a combination of unsupervised and supervised algorithms.

4.3. Supervised Learning

As described in the previous section, supervised learning is based on a data-set which represents the conditional data. The labels connected to particular data will learn the algorithm. Supervised learning is divided in classification, regression and Neural Network algorithms. Classification, Regression and Neural Network represent groups of algorithm which can be used to predict the shelf life of a strawberry. The three groups are tested in Matlab applications which help to generate the algorithms based on the data samples measured in the supply chain. The Matlab application will be explained in the next section.

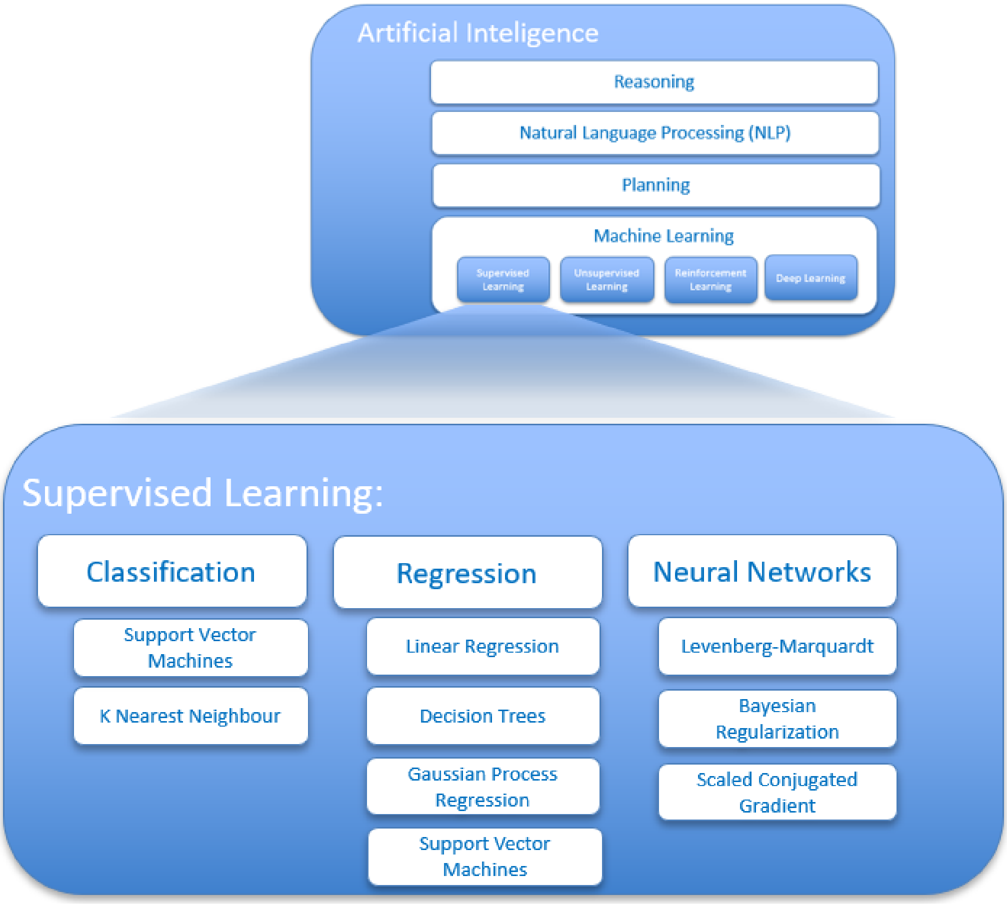


Figure 4.3: Overview of learning algorithm for supervised learning.

4.4. Matlab Regression Learner Application

The matlab regression learner Application helps to choose between various algorithms. The algorithms can be trained and validated. The application displays results of validated models. The models are evaluated on model accuracy, and with plots, such as response plot or residual plot.

The Regression learner application is able to train linear regression models, Support Vector Machines (SVM), regression trees or ensembles of regression trees and Gaussian process regression models. These 4 different models are evaluated in the next section.

4.4.1. Linear Regression

A machine learning algorithm based on linear regression is based on the correlation between input data and output data. This correlation is visualised by a regression line which should be close to the training examples, for example fig. 4.4. If the training examples are far from the linear regression line, the prediction based on the input data would have fewer change to be correct.

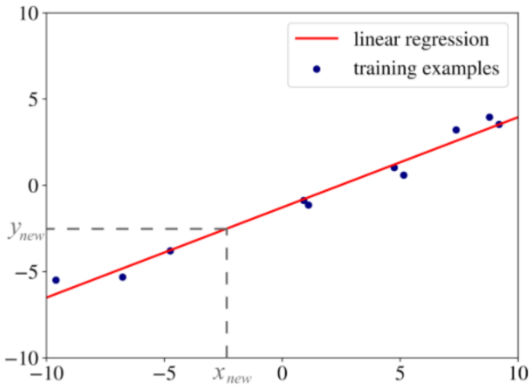


Figure 4.4: Linear regression plot including training examples.

A consideration to use linear regression is that it is simple and will rarely overfit. Overfitting occurs when the model predicts very well labels of the examples used during training but frequently makes errors when applied to examples that weren't seen by the learning algorithm during training [4].

4.4.2. Support Vector Machine

Data is separated by a decision boundary or hyperplane. The equation of the hyperplane is given by two parameters, a real valued vector W, feature vector x and a real number b, given in the next equation:

$$Wx - b = 0 \tag{4.1}$$

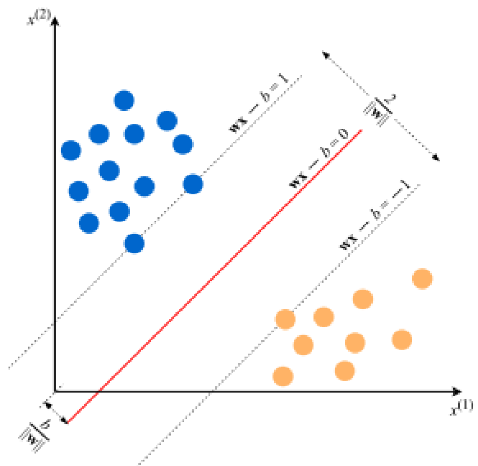


Figure 4.5: Linear regression plot including training examples [4].

To find out how well the algorithm is able to classify new examples, the distance between the closest examples of two classes is important. In case of figure 4.5, the margin between the blue and orange data points. A large margin contributes to a better generalization. The form of the decision boundary determines the accuracy of the model, the form is mathematically or algorithmic ally computed based on the training data. The larger the set of training examples, the more unlikely that the new examples will be dissimilar to the examples used for training.

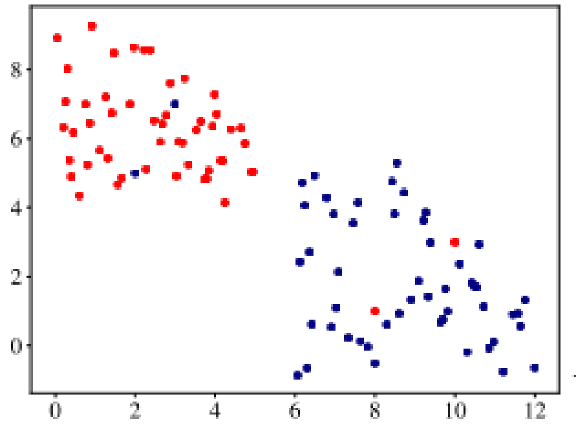


Figure 4.6: Linearly non-seperable cases, the presence of noise [4]

In some cases, it is impossible to separate the two groups of points because of noise in the data, errors of labeling or outliers. For data on the wrong side of the decision boundary, see figure 4.8, the function's value is proportional to the distance from the decision boundary. The following cost function will be minimized:

$$C\|w\|^2 + \frac{1}{N}\sum_{i=1}^{N} max(0.1 - y_i(wx_i - b)) \tag{4.2}$$

The hyperparameter C determines the trade off between increasing the size of the decision boundary and ensuring that each x_i lies on the correct side of the decision boundary. The SVM algorithm will try to find the highest margin by completely ignoring misclassification, because a high value for C will negligible the second term in the cost function.

A larger margin is better for generalization. The value of C regulates the trade off between classifying the training data well (minimizing empirical risk) and classifying future examples well (generalization).

4.4.3. Decision Tree Learning

A decision tree learning consist of branching nodes which each have a particular value. The path followed through the tree is based on the value of the feature vector. If the value of the feature is below a specific threshold, then the left branch is followed. If the feature is above the specific threshold, then the right branch is followed. The final node is called the ”leaf node”, where the decision is made about the class to which the example belongs.

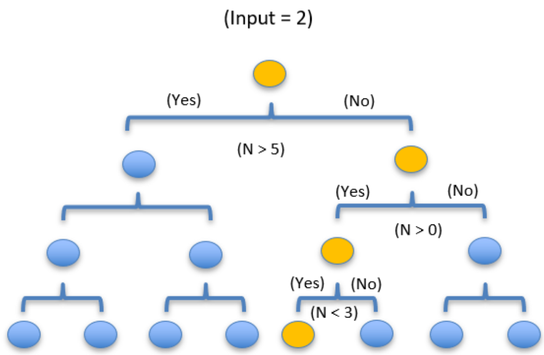


Figure 4.7: Decision Tree learning algorithm overview, example is given with input number 2.

4.4.4. Gaussian Process Regression

Contrary to other learning algorithms that forget training data after the model is built, Gaussian Process Regression keeps all training examples in memory. Once a new example comes in, the algorithm finds training examples closest to the data sample and returns a prediction.

This principle is based on the K-Nearest Neighbors. In which the K represents the number of training data point lying in proximity to the test data. The nearest training data points are used to predict a new data sample. In case of shelf life prediction, the system is trained by data samples with a known shelf life, the ’neighbors’. The shelf life of a new data sample with an unknown shelf life will be predicted based on the nearest ’neighbors’ [5].

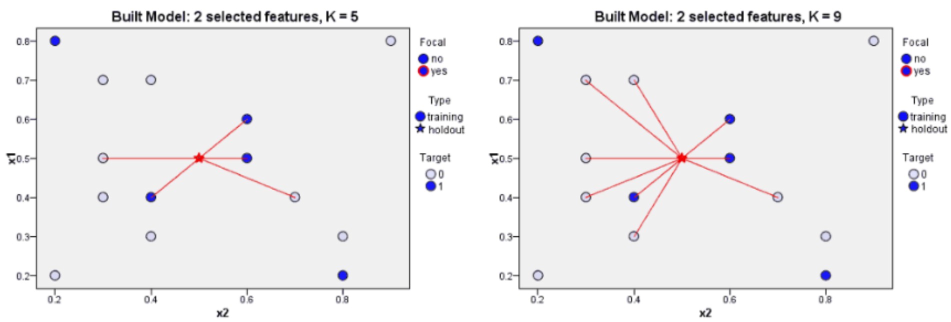


Figure 4.8: K Nearest Neighbours, left K=5, five ’neighbors’ are selected, right K = 9, nine ’neighbors’ are selected [5].

A Gaussian process uses lazy learning and a measure of the similarity between points (kernel function) to predict the value for an unseen point from training data for a really wide range of patterns. Lazy learning is a method which isn’t trained in advance, regarding to Linear Regression the work is done when an input item is presented.

4.5. Neural Network Fitting Application

The integration of neural networks in our common lives is growing. Presently, automated guidance in cars, vehicle scheduling and routing systems make are all applications based on neural networks. The question rises, is it possible to integrate a neural network to predict the shelf life of a strawberry?

To test the prediction accuracy of a Neural Network, the Neural Network Fitting Application is used. This applications is a GUI which automatically generates command-line scripts. The biggest part (95 percent) of the measured data sets are used for training the algorithm and a small part (5 percent) is used to validate the algorithm.

The neural network can optimised in three different ways, with Levenberg-Marquardt, Bayesian Regularization and the Scaled Conjugate Gradient optimisation algorithm. The optimization of the neural network is visualised in figure 4.9. The training of the neural network is based on the measured data samples and the actual shelf life determined by the inspector. The neural network is learning from the data samples and evaluates the predictions with the actual shelf life. This is an iterative process in which the predicts are optimised. The way the neural network is optimising the predictions is based on the optimisation algorithm. Three different optimisation algorithms are evaluated in the next section.

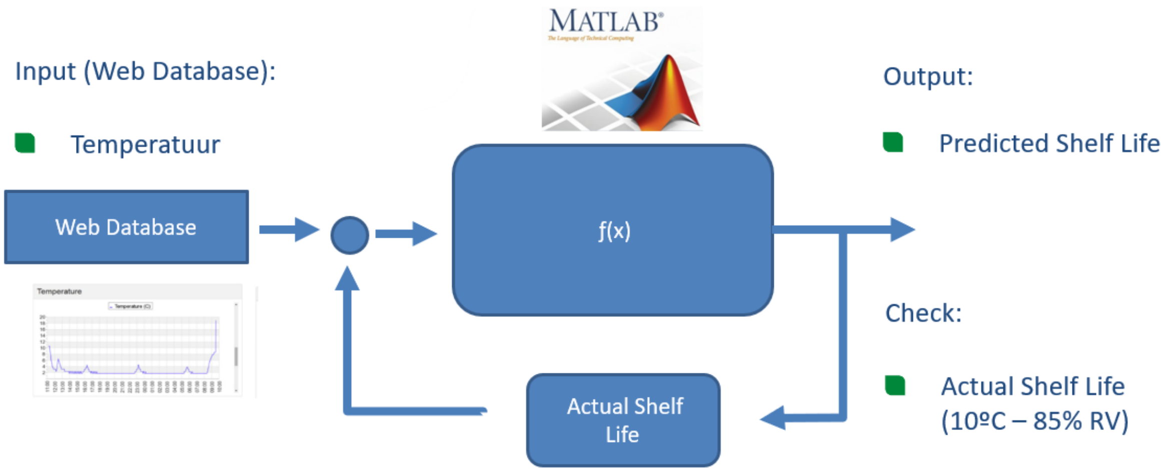


Figure 4.9: Neural Network architecture generated by Matlab Neural Network fitting application.

Optimisation Algorithms

The first evaluated neural network learning algorithm is **Levenberg-Marquardt** This algorithm is a regularly used optimization algorithm [49]. It is efficient when the network contains low number of data samples [50]. So is recommended for most problems. This algorithm typically requires more memory but less time compared to Bayesian Regularization or the Scaled Conjugate Gradient Trainer.

The **Bayesian Regularization** algorithm requires more process time to generate the optimum model but normally results in good generalization for difficult, small or noisy data sets, according to information given by the Matlab Neural Network application.

The **Scaled Conjugate Gradient training algorithm** requires less memory. The algorithm is based up on a class of optimization techniques well known in numerical analysis as the Conjugate Gradient Methods. For large problems Scaled Conjugate gradient is recommend as it used gradient calculations which are more memory efficient that the Jacobien calculations the other two algorithms use.

4.6. Learning Algorithm Selection Criteria

A selection is made in the machine learning algorithms which are able to predict the shelf life of a strawberry. But what kind of regression learning algorithm or neural network suits best these data examples. In the next section the selection criteria are evaluated to find the most accurate shelf life prediction model. The algorithms are trained, validated and tested in Matlab. Matlab helps with the selection of the most accurate algorithm based on the following criteria:

- **Root Mean Square Error (RMSE)**
  One of the validation indicators is the Root Mean Square Error (RMSE). In this research the RMSE value gives the difference between the measured shelf life and the actual shelf life. A learning algorithm with the lowest RMSE is favourable. The RMSE value is also an indication for overfitting. If the RMSE value of the test set is higher than the training set, it is likely that the model overfit the data.

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4.3}$$

- **Mean Squared Error (MSE)**
  The MSE measures average squared error of the shelf life predictions. For each prediction, it calculates square difference between the actual shelf life and the predicted shelf life and then average those values. Like the RMSE value the MSE values is also an indication for overfitting.

$$MSE = \frac{1}{n}\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4.4}$$

The selecting the algorithm based on the RMSE and MSE is based on a research of Chai and Draxler [51] how conclude that RMSE is more appropriate to use when errors follow a normal distribution. The expectation is that the predicted shelf life will have a normal distribution around the actual shelf life.

4.7. Learning Algorithm Evaluation

To evaluate a learning algorithm this section presents a criteria to test a particular learning algorithm. The criteria under- or overfitting is a way to find out if a learning algorithm is able to predict the labels of the data it is trained on.

4.7.1. Divide Data Examples In Three Sets

To build a model it is advised [4] to divide the available data into three distinct sets of labeled examples, the default settings are:

- Training set (70 percent)
- Validation set (15 percent)
- Test set (15 percent)

A model which is able to memorize all training examples and uses this memory to predict their labels will work perfectly. A validation of this algorithm with the same data which is used for training will be useless. The validation is only value-added if it is able to predict based on data which it has never seen.

The training set will be used to develop a model, the validation set will be used to choose the right learning algorithm and the best values of hyperparameters. And last but not least, the test set will be used to asses the final model.

A way to compare the performance of the model on the training and test data is to use a type of average loss function. For example, the mean squared error (MSE) of the model. If the MSE of the model on the test data is higher than the MSE obtained on the training data, this is a sign of overfitting [4].

4.7.2. Underfitting

When an algorithm makes mistakes on the training data, it has a high bias or it underfits. Underfitting is an indication to validate a model to predict well the labels of the data it is trained on. The reasons for underfitting, given by Burkov [4]:

- The model is too simple for the data.
- The features you engineered are not informative enough.

A model is to simple when a dataset looks like a curved line, but our model is a straight line. The features are not informative enough when the features are not good predictors for a particular effect. The model doesn't have a meaning full relationship with the features used.

To solve these kind of problem of underfitting, it is advised by Burkov [4] to generate a more complex model or to engineer features with higher predictive power.

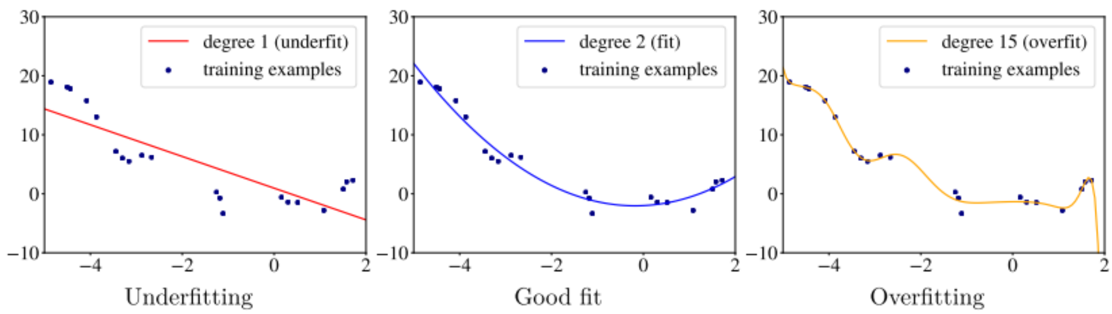


Figure 4.10: Examples of underfitting (left), good fit, (right), and overfitting (right) of a regression model [4].

4.7.3. Overfitting

Another name for overfitting is the problem of high variance. A model that overfits or has a high variance predicts the training data very well, but poorly on the validation data set. Reasons for overfitting are given by Burkov [4]:

- The model is to complex for the data.
- You have too many features but a small number of training examples.

When the model is to complex it means the models is adapted to much on the features of the training data (fig. 4.10), the model would be significant differ when other samples of data are used as training data. This is the reason why a model which overfits performs poorly on new data.

The simplest models can also overfit test data. This happends when data is high-dimensional, but the number of training examples is relatively low. For example a learning algorithm can build a model, which finds very complex relationships between all available features to predict labels of training exampels perfectly. But this model will predict the labels of the validation data set poorly because the model learned also the noise in the values of the features of the training examples.

Solutions to solve overfitting given by Burkov [4]:

- Start with a simple model, linear in stead of polynomial regression, or SVM with a linear kernel instead of RBF, a neural network with fewer layers/units.
- Reduce the dimensionality of examples in the dataset (use dimensionality reduction techniques).
- Add more training data.
- Regularization (most widely used approach to prevent overfitting), regularization leads to a slightly higher bias but will reduce the variance.

4.7.4. Cross-Validation

When training examples are limited, the available data can be split into a training and a test set. The validation set can be simulated by cross-validation. Cross validation works as follows [4]:

The first step is to fix the hyperparameters to evaluate, a hyperparameter is a property of a learning algorithm, usually having a numerical value. The value influences the way the algorithm works and are not learned by the algorithm itself from data. The value of a hyperparameter needs to be set manually.

The hyperparameters need to be found experimentally. A descend way to do that is to use a grid search, random search or Bayesian Hyperparameter optimization. Grid search makes use of a discrete set of values to explore for each hyperparameter. Random search provides a statistical distribution for each hyperparameter from which values are randomly sampled. Bayensian differs from random grid search by using evaluation results to choose the next values to evaluate. The next value is chosen based on those that have done well in the past.

The next step is to divide the training set into subsets of the same size. For example into 5 subsets F_1,F_2,F_3,F_4,F_5. The first model uses the first 4 subsets F_1,F_2,F_3,F_4 for training, the fifth model F_5 is used to validate the trained model. The second model uses F_4 as validation subset and the other subsets are used for training F_1,F_2,F_3,F_5. After interactively building these models, the average of the five values will generate the final value.

4.8. Selected Algorithms

This chapter starts with the introduction of artificial intelligence. Besides machine learning the natural language processing, planning and reasoning models are evaluated for predicting shelf life of a strawberry. The ability to learn from and adapt a model based on the data rather than explicit programming is the reason for selecting machine learning for the prediction of shelf life of a strawberry. The learning process of machine learning is based on features and labels, the labels are the actual shelf life determined by an inspector in the shelf life room. Because of the use of labels, this research is based on supervised learning. Supervised learning can be divided classification, regression and neural network models. The inputs are continuous and the outputs are continuous. This means classification is not suitable, but the regression and neural network models are applicable for predicting shelf life of a strawberry.

5
Research Methodology

5.1. Introduction

The research methodology will explain how the best possible approach is chosen to answer the problem statement and the research question. Mean while, the obstacles in collecting and analysing data will be described. This research methodology will reference to existing research in the field and how this research took a novel approach to address a gap in the literature.

The methodological approach of this research is similar to mathematical modelling. It is similar but not the same, because this research makes used of a black box model instead of building a mathematical model. The black box is the machine learning algorithm. This research is focused on the input and output of this black box. What happens inside the black box, or machine learning algorithm is outside the scope of this research. This research methodology is divided in four categories, system analysis, machine learning selection, testing and use. The steps define the process to build a black box model which can predict self life of a strawberry.

The process starts with the method of data collection. Next subject is the way the data is processed and the computer programs used for this purpose. The method for machine learning feature selection will be discussed before the procedure to select the best suited learning algorithm will be explained.

5.2. System Analysis

In the section System Analysis collection the tools, techniques and procedures used for data collection for this research are evaluated. The characteristics of the temperature logger and the integration of the temperature loggers in the supply chain is described. Before this data is used for this research, it is processed. The process starts with data import from the temperature sensors. This time-temperature profile is analysed and noise is removed. Finally, a description is given of the way features are generated out of the processed data.

5.2.1. Strawberry Supply Chain

The supply chain of strawberries from grower till customer is the scope of this research. In Chapter 2, "Supply Chain Overview", a detailed description of the actors and the processes in the supply chain is given. At the moment a strawberry is harvested, a data logger is implemented into the system. The temperature logger is distributed through the supply chain to the supermarket. In the next section the tools, techniques and procedures used for data collection for this research are evaluated. The characteristics of the temperature logger and the integration of the temperature loggers in the supply chain is described in the next section.

Data Collection

The temperature loggers are supplied by AHRMA. AHRMA, situated in Deventer, has developed a temperature and motion logger which can store the measured data for 5 days and is able to send this data wireless to a gateway when the logger is near the gateway (+/- 80 meter).

During the research period strawberries where imported from the south of Spain. The scope of the supply chain starts at a grower located near Huelva, Spain. At the moment the strawberry is harvested the temperature logger is placed in a 500 grams package filled with strawberries, see figure 5.1.

The package with logger is transported through the supply chain to be removed from the supply chain at the supermarket. To make sure all participants in the supply chain where informed about this measurements, a clear instruction document was generated for each stage. The next section will explain the method used to measure the temperature through the supply chain.

Grower

The supplied temperature loggers where transported to the grower. The grower was instructed to use six loggers a day and put them in separate packages of 500 gram strawberries. These six packages where distributed in separate carton boxes of 12 packages each (fig. 5.2).

Figure 5.1: Sensor in package of 500 gram strawberries.

Figure 5.2: Strawberries distributed in carton box.

Distribution Center The Greenery

The six temperature loggers, are distributed to the distribution centre of The Greenery located at Breda. The loggers connect with the gateway installed at the loading docks to transfer the measured temperature data. The boxes which include a temperature logger are marked with orange tape, see figure 5.3.

Figure 5.3: Pallet loaded with boxes filled with 12 packages of 500 gram strawberries.

An inspector removed the six packages with sensor from the pallet and prepared five empty boxes in which five of the six packages with sensor were transported to a PLUS distribution centre. One of the six packages stayed at the distribution centre of The Greenery in Breda. This package went straight into the shelf life room. The package was stored in the shelf life room under constant conditions (10 degrees Celsius, 85 percent humidity).

The other five packages where packed in prepared boxes (fig. 5.4) and labeled with a code which was unique for each supermarket. These five boxes including 500 gram strawberries and temperature logger are transported to the distribution center of PLUS along with the daily shipment of strawberries. The distribution centre of PLUS is located at Barendrecht.

Figure 5.4: Pallet loaded with boxes filled with 12 packages of 500 gram strawberries.

Distribution Center PLUS

The five packages including temperature loggers arrived at the distribution center of PLUS along with the daily delivery. One of the packages was stored at the distribution centre of PLUS (Hollander), the other four boxes were distributed to four selected local supermarkets. Just like the rest of the strawberries the packages were stored and order picked for particular PLUS supermarkets. Together with other products the strawberries were delivered at local PLUS supermarkets. At the supermarkets the products were unloaded and transported to the cool storage before it is presented for customers or in our case removed from the supply chain to return to the shelf life of The Greenery.

Shelf Life Room

The packages with strawberries including a temperature logger were collected at 4 different supermarket locations and the distribution centre of PLUS. The locations of the supermarkets are selected in the area of the distribution centre of The Greenery. After collecting the 5 packages, they were transported to the distribution centre of The Greenery in Breda.

An inspector will check the quality of the strawberries at the moment they arrive at the distribution center. The inspector is reviewing the quality for 6 days once a day. In this way the shelf life of the strawberries is evaluated.

After 7 days, the logger is removed from the strawberries and the temperature data is saved. All findings of the inspector all registered and the maximum shelf life for each package is documented. The used temperature loggers are collected and shipped to the grower in Spain to be re-used for a second time.

Overview

An overview of the way temperature loggers moved through the supply chain is given in figure 5.5. The overview shows the 6 temperature loggers through the supply chain. The packages of strawberries including a temperature logger are collected and transported to the shelf life room located at the distribution centre of the The Greenery at Breda. In this shelf life room the quality of the strawberries are inspected every day.

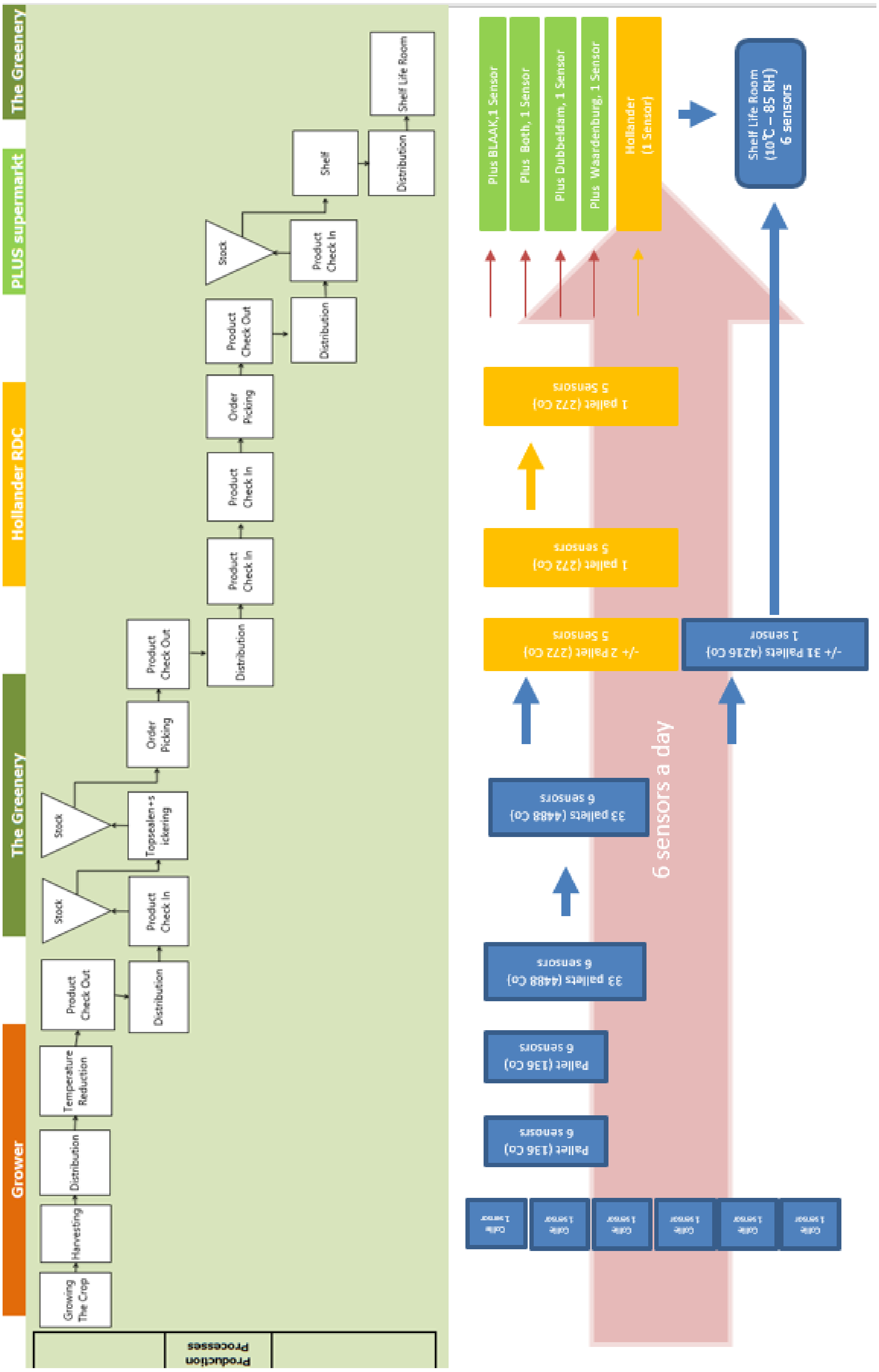


Figure 5.5: Temperature logger integration into strawberry supply chain.

5.2.2. Data Analysis

All collected data is inspected on measurements failures. Data from temperature loggers which did not function properly are rejected. The random peaks are filtered out of the temperature data. Out of this data the features and labels are selected. These features and labels are used to train regression learning and neural network algorithms.

Import Data

At the moment a temperature logger is in the range of a gateway, it will start to transfer the temperature measurements of the last 5 days to the gateway. This data is saved on a server and is available in several minutes. This data is downloaded and opened in excel. This excel sheet shows the temperature and the moment of the measurement measurement with a interval of 10 minutes.

Data Processing

The measured data is checked on missing data points or periods where the sensor didn't measure the temperature, see figure 5.6. Occasionally the temperature data was not available because of the limited storage capacity. The sensor storage is limited to 5 days, after a time period of 5 days the oldest data is overwritten. So it is important that the temperature loggers where located in the rage of the gateway to transfer the measured data every 5 days. Temperature profiles with missing measurement periods are not used in this research.

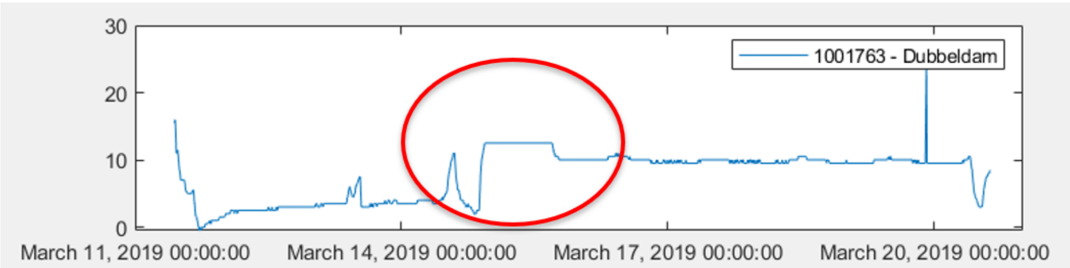


Figure 5.6: Time-Temperature profile strawberry supply chain, period without measurements is indicated.

Another correction on the data used for this research is the removal of peaks which did occur in the time-temperature profile. The peaks did occur randomly and have a temperature difference over 10 degrees with their previous and next measurement. All variations higher than 10 degrees are filtered out. The removed peaks are given in a plot in figure 5.7.

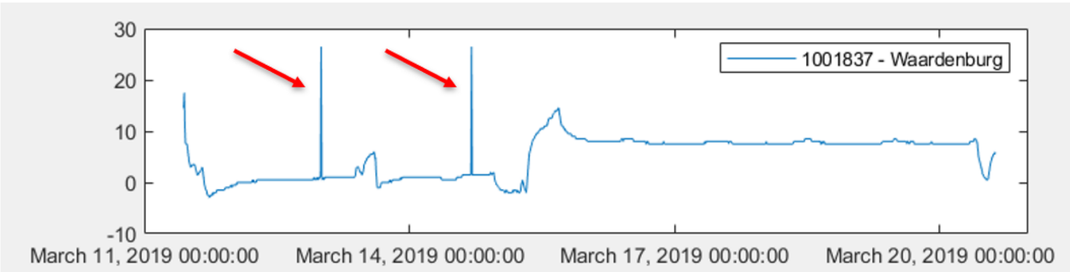


Figure 5.7: Time-Temperature profile strawberry supply chain, peaks are indicated by arrows.

5.2.3. Selection Of Features

The selection of features for machine learning algorithms is based on the measured data. Good features have high predictive power, this means that there is a clear correlation between the feature and the label of the data. The evaluated figures show the correlations between shelf life and fluctuating temperatures, average temperatures, temperature variation, storage temperature and throughput time. The correlations are based on temperature measurements. The measurement period starts from the moment the strawberries are harvested till they arrive at the shelf life room located at the distribution centre of The Greenery at Breda.

5.3. Machine Learning Selection

The measured data is analysed an will be used as input for a machine learning algorithm. Regular machine learning algorithms are discussed in Chapter 4, ”Predicting Algorithm”. To get the most accurate prediction these algorithms are compared to each other based on the measured data samples. The best performing algorithms are selected to be tested on a new data sample which is not used for training or validation.

5.3.1. Regression Algorithm Selection

After selecting features and labels out of the collected data, these features and labels are used to train regression learning algorithms. The regression learner application of Matlab will is used to find out how many folds will result in the lowest RMSE value and to find the most accurate regression algorithm based on the selection criteria.

To find the most accurate algorithm for predicting the shelf life, 19 different regression algorithms are trained by the Matlab regression learner application. These algorithms are evaluated and finally tested on predicting the shelf life based on the available data samples. To use these data samples in an optimal way, cross validation is applied. The next section will evaluated the method of cross-validation and the selection of learning algorithm.

Cross-Validation

This section describes the cross-validation process. To use the available data examples as efficient as possible cross validation is applied [4]. The available data is divided into a training and a test set. The validation set can be simulated by cross-validation. In this way more data is available to train the model which will result in higher accuracy of the trained model. The method of cross validation is explained in the next section with an example of 5 folds, this means the available data samples are divided in 5 groups with a similar amount of samples.

To give an example of cross-validation, cross validation with 5 folds is described. The first step is to dived the test set into 5 subsets F_1,F_2,F_3,F_4,F_5. The first model uses the first 4 subsets F_1,F_2,F_3,F_4 for training, the fifth model F_5 is used to validate the trained model. The second model uses F_4 as validation subset and the other subsets are used for training F_1,F_2,F_3,F_5, see figure 5.8. After interactively building these models, the average of the five values will generate an average value. This average value is representative for the regression model and will be used to select the most accurate regression model.

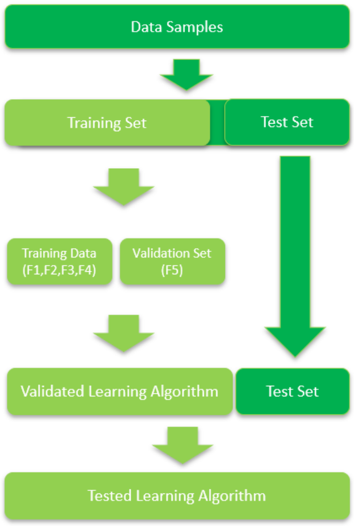


Figure 5.8: Flow chart of cross validation, data samples divided in 5 subsets.

Selection Criteria

Regression algorithms are selected based on the Root Means Square Error value. In the literature survey (chapter 4.6 ”Learning Algorithm Selection Criteria”) the selection criteria for regression learning algorithms are introduced.

When the regression algorithm is validated a Root Mean Square Error value is calculated. This value is the first indication of the final accuracy of the predictions generated by the regression algorithm.

5.3.2. Neural Network Selection

The data samples are used to train a neural network with different training algorithms. The optimisation algorithms used are Levenberg-Marquardt, Bayesian Regularization and the Scaled Conjugate Gradient. The neural networks are tested with different number of layers which showed different deviation values. The number of layers with the least deviation between the predicted shelf life and the the determined shelf life is selected for each different training algorithm. Three different optimisation algorithms will be compared to each other and to the regression learning algorithms in the predicting shelf life test.

5.4. Testing Machine Learning Algorithms

The selected machine learning algorithms are tested on new data. This data is not used for training or validation of the algorithm. By testing the selected algorithms on new data, the accuracy of the system will be evaluated. The following section will explain how the selected algorithms are evaluated.

To validate the accuracy of the system, the selected machine learning algorithms are compared to each other. The difference between the predicted shelf life and the actual shelf life of a strawberry is given for each algorithm. To evaluate the differences and the accuracy of the system the following statistical tools are used:

- Standard Deviation
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)

5.5. Using Models

The test results compare the shelf life predictions of selected machine learning algorithms with each other. These results will be evaluated with a confidence interval and a probability distribution. Based on this result evaluation an algorithm will be chosen which is most suitable to implement in the current strawberry supply chain.

Going back to the research problem, the question was ”How can artificial intelligence contribute to a better shelf life prediction of a strawberry based on temperature measurements from grower till customer?”. The implementation of shelf life prediction is not tested. But the following potential implementations are evaluated:

- Objective Quality Control
- Customer Assignment
- Dynamic Pricing

6
Results

This chapter is divided in three parts based on the schematic overview of the shelf life prediction process, see figure 6.1. This chapter will show the measurements of the temperature loggers. Based on these measurements the machine learning input parameters are determined. These input parameters are used to train, validate and test the regression and neural network models. The results of the regression and neural network models will be evaluated and finally a selection of these models will be tested on new data samples which are not used for training or validation. To compare the results to each other and find how accurate the machine learning algorithms can be. Finally the potential improvements for the strawberry supply chain will be discussed.

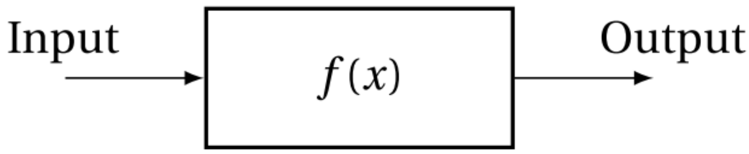


Figure 6.1: Viewing the Strawberry as a system with input and output [6].

6.1. Input Parameters

The data measured by temperature loggers in the supply chain is downloaded from a online database. This data is cleared from measurement faults before it is used to determine the relations between, for example, average temperature and other relations described in chapter 3.1, ”Input Parameters”.

6.1.1. Measured Data

During the research everyday 6 loggers where implemented into the strawberry supply chain. Each logger ended up at a different location. 55 data samples are used in total in this research. For all these data samples features were determined. A plot of the derivative of the temperature profile and the location of peaks is shown for these 6 loggers on the next page, see figure 6.2.

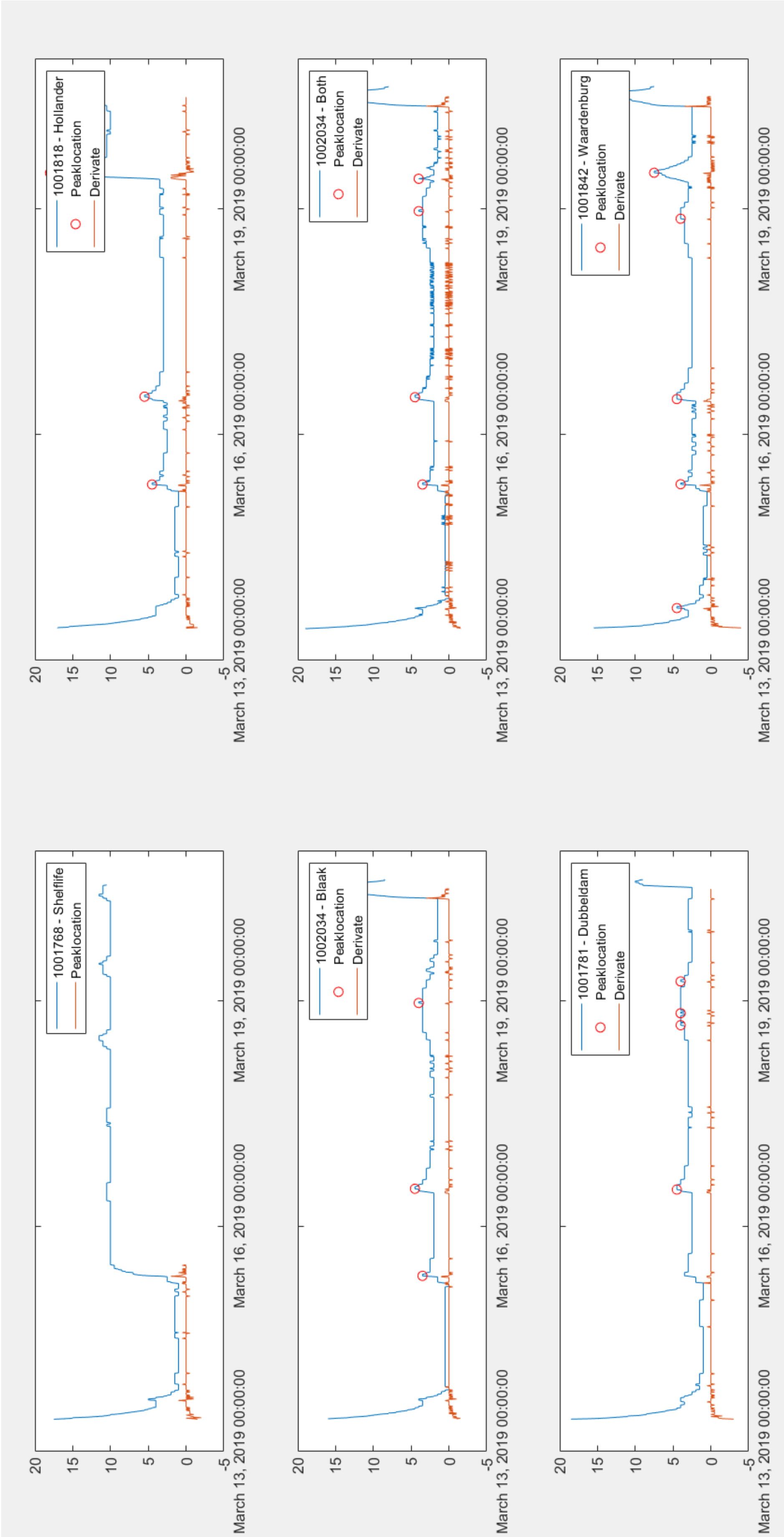


Figure 6.2: Overview of 6 loggers with are implemented in the strawberry supply chain at the same moment, all loggers moved to different locations and show different temperature profiles.

6.1.2. Features

Features are input parameters which are generated based on the measured temperature profile. These features used in this research have a relation with the shelf life. This relations are based on literature and interviews with experts, see chapter 3.1 ”Shelf Life Influencing Parameters”.

The features are determined for each data sample and measured by the temperature logger. The feature values for each data sample are shown in Appendix A. The feature data is used as input for machine learning algorithms.

Feature Correlation

The correlation between features is evaluated by plotting the measured results. The plots are attached in Appendix B. The plots show a wide spread of measurements. Which make the correlation between features and shelf life weak. The least square line is used to give an impression for correlation but is not sufficient to conclude a correlation between the features and shelf life. The amount of data samples is low, when out of bound date points are left out the line changes its direction. Which indicates a weak correlation.

6.2. Prediction Model

The collected temperature data was analysed and used to find the most accurate prediction model. To find the most accurate shelf life prediction given by a regression model the regression learner application of Matlab was used. The regression learning application is able to compare 19 different regression learning algorithms with each other based on the measured data samples.

6.2.1. Selection Regression models

The results of cross-validation of the 19 different regression models for different amounts of folds are compared in the table 6.1. This table is used to select the lowest RMSE values independent of the amount of Cross Validation Folds (CVF).

RMSE Value

One of the validation indicators is the Root Mean Square Error (RMSE). The learning algorithm with the lowest RMSE is favourable, see Chapter 4.6 ”Learning Algorithm Selection Criteria” for more detailed information about the RMSE value.

Folds

The values in bold are the lowest values for five different regression learning models. The values are generated with the data samples divided in 5 and 10 folds. The folds are used for cross validation, explained in chapter 5.3.1, ”Regression Algorithm Selection”. In case 5 folds are used by the Regression Learner Application, the total amount of samples is divided by 5. An example is given in Appendix A. The samples are divided in F_1,F_2,F_3,F_4,F_5.

Table 6.1: Cross validation results for different amounts of folds, RMSE overview of regression learning algorithms generated by Matlab.

	(CVF: 2) RMSE	(CVF: 5) RMSE	(CVF: 10) RMSE	(CVF: 15) RMSE	(CVF: 20) RMSE	(CVF: 25) RMSE
Linear Regression Models						
Linear	1.56	1.31	1.26	1.33	1.25	1.21
Interactions	13.56	25.94	40.02	57.33	56.25	27.13
Robust	1.57	1.40	1.26	1.32	1.26	1.21
Stepwise Linear	1.62	1.30	1.64	1.55	1.52	1.61
Regression Trees						
Fine Tree	1.48	1.43	1.39	1.41	1.40	1.60
Medium Tree	1.33	1.21	1.28	1.30	1.31	1.30
Coarse Tree	1.21	1.23	1.22	1.23	1.23	1.23
Support Vector Machine						
Linear	1.26	**1.10**	1.22	1.21	1.25	1.26
Quadratic	1.39	1.42	1.41	1.31	1.48	1.46
Cubic	1.43	1.66	1.85	1.63	2.12	1.73
Fine Gaussian	1.20	1.21	**1.19**	1.20	1.20	1.19
Medium Gaussian	1.23	**1.13**	1.14	1.17	1.19	1.16
Coarse Gaussian	1.20	**1.15**	1.17	1.18	1.29	1.18
Ensemble						
Boosted Trees	1.36	1.25	1.29	1.36	1.29	1.30
Bagged Trees	1.30	1.19	1.21	1.24	1.21	1.25
Gaussian Process Regression						
Squared Exponential	1.30	1.22	1.21	1.23	1.23	1.24
Matern 5/2	1.29	1.21	1.20	1.22	1.22	1.23
Exponential	1.24	1.18	**1.18**	1.21	1.20	1.22
Rational Quadratic	1.30	1.22	1.22	1.23	1.24	1.24

Selected Algorithms

The regression algorithms with the lowest RMSE values are presented in table 6.2. This table shows the amount of folds used and the RMSE values out of the previous table, table 6.1. The minimum, maximum and average deviations are calculated based on the data samples which are selected for training and validation. The regression algorithms given in table 6.2 will be tested on data samples which is not used for training or validation of the algorithm.

Table 6.2: Overview of prediction accuracy of the selected learning algorithms

		Hidden Layers	RMSE
Support Vector Machine			
	Linear	5	1.10
	Fine Gaussian	10	1,19
	Medium Gaussian	5	1.33
	Coarse Gaussian	5	1.15
Gaussian Process Regression			
	Exponential	10	1.18

Depending on the optimal accuracy every linear regression algorithm has a particular amount of folds. Table 6.2 shows the amount of folds used for each regression learner algorithm in the first column. The accuracy is described by the Root Means Square Error which is shown in the second column.

6.2.2. Selection Neural Network Models

Besides the regression algorithms, the accuracy of neural networks with different kind of optimisation algorithms is evaluated on accuracy. Like the regression models, the same data samples are used to train Neural Network.

6.2.3. Number of Layers

To compare the results of the hidden layer iterations for each learning algorithm, the results are given in a plot. Each learning algorithm was tested 5 times on each layer. The average value of the 5 times is shown as thick line in the plot. The average RMSE for the number of hidden layers between 2 and 16 are provided in the following plots for the three selected optimisation algorithms, Levenberg-Marquardt, Bayesian Regularization and the Scaled Conjugate Gradient Trainer.

Levenberg-Marquardt

The characteristics of this training algorithm is that it uses more memory but less time compared to the other Neural Network learning algorithms used in this research. This optimisation procedure will stop when the mean square error of the validated samples is increasing.

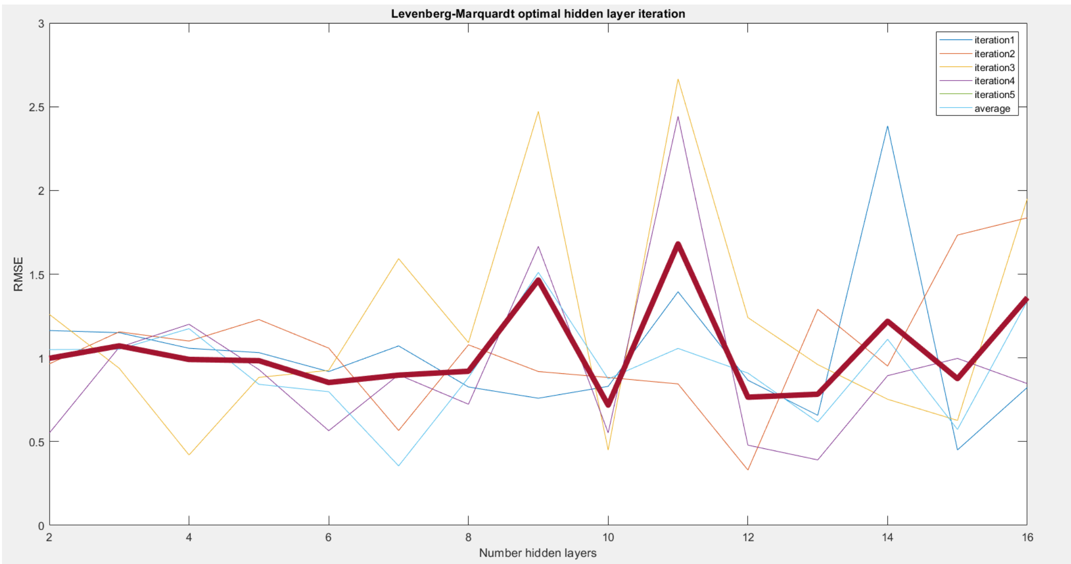


Figure 6.3: Hidden layer selection plot for Levenberg-Marquardt optimisation algorithm.

Figure 6.3 shows the average RMSE using the Levenberg-Marquardt optimisation algorithm for the amount of hidden layers between 2 and 16. The average value of the 5 repetitions for each number of hiddenlayer is given as the thick blue line. The lowest average deviation percentage is generated by 10 hidden layers.

Bayesian Regularization

This algorithm needs more time to train the input data but the results are better for difficult, small or noisy datasets compared to the other Neural Network optimisation algorithms used in this research. The training stops according to adaptive weight minimization (regularization).

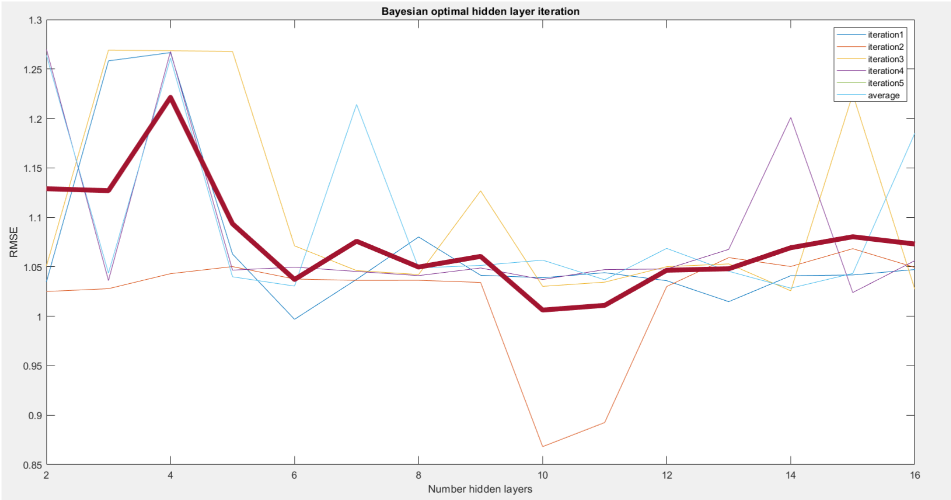


Figure 6.4: Hidden layer selection plot for Bayesian Regularization trainer.

Figure 6.4 shows the average RMSE using the Bayesian Regularization Trainer algorithm for the amount of hidden layers between 2 and 16. The average value of the 5 repetitions for each number of hidden layer is given as the thick blue line. The lowest average deviation percentage is generated by 10 hidden layers.

Scaled Conjugate Gradient

This algorithm is able to train the input data with less computational force. This optimisation algorithm stops the training procedure when the RMSE is increasing.

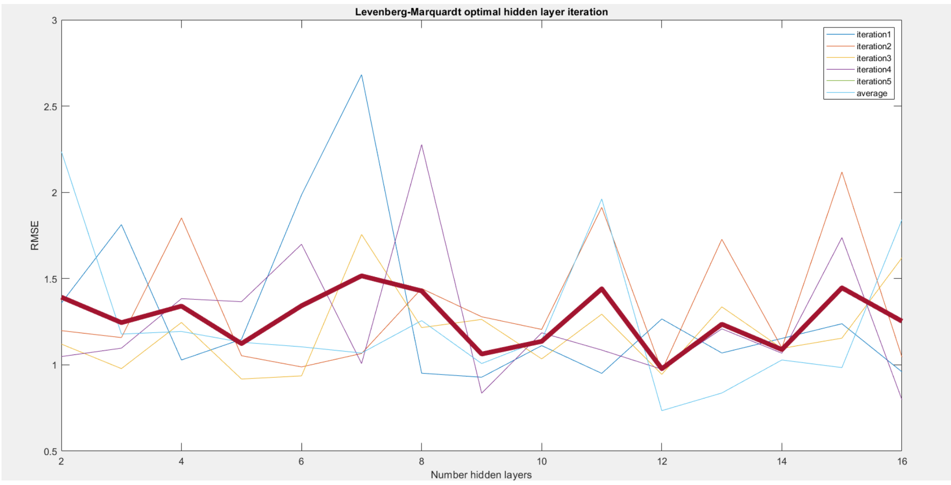


Figure 6.5: Hidden layer selection plot for Scaled Conjugate Gradient trainer.

Figure 6.5 shows the average RMSE using the Scaled Conjugate Gradient Trainer algorithm for the amount of hidden layers between 2 and 16. The average value of the 5 repetitions for each number of hidden layers is given as the thick blue line. The lowest average deviation percentage is generated by 12 hidden layers.

The previous plots show the RMSE value for each neural network layer size. The RMSE values of the three different optimi algorithms are compared to each other in table 6.3. The amount of layers is given in the first column and the lowest average RMSE value based on 5 iterations that can be found in the second column.

Table 6.3: Overview of prediction accuracy of the selected neural networks.

		Hidden Layers	Root Mean Square Error (RMSE)
Neural Network Fitting			
	Levenberg-Marquardt	10	0.72
	Bayesian Regularization	10	1,01
	Scaled Conjugate Gradient	12	1,00

6.3. Shelf Life Prediction Results

The next step is to find out how accurate the regression learning algorithms can be. In the previous section, regression and neural network models were selected. These selected algorithms were trained on the same data samples which are used to select the learning algorithms. To see how accurate machine learning can be, the selected algorithms are tested on data which is not used for training or validation. Each machine learning algorithms will predict the shelf life based on 10 data samples. The outcome of this test will be evaluated.

6.3.1. Data Samples

The data samples used to compare the different machine learning algorithms is shown in the first column of table 6.4. The numbers represent the temperature logger which is used to measure the temperature through the supply chain. This code can be traced back in the data base to give more information about the measured period.

6.3.2. Actual Shelf Life

After running through the supply chain, each package of strawberries including temperature sensor is returned to the shelf life room. The strawberries stay in the self life room for 7 days, during these days an inspector determines the deterioration of the strawberries and decides the moment of expiring. The moment of expiring is the end of the shelf life. This is all documented in an excel sheet, an examples is given in Appendix D.

6.3.3. Predicted Shelf Life

The output of each machine learning algorithm is given in table 6.4. The predicted shelf life for each data sample is generated by selected regression and neural network algorithms. The predicted shelf life is provided in days.

6.3.4. Deviation

The deviation is the difference between the predicted shelf life and the actual shelf life determined by the inspector. The difference is used to determine the most accurate shelf life prediction algorithm. is given in the lowest rows of the table. The average deviation is given in days and in percentage of the actual shelf life.

Table 6.4: Test Results for selected learning algorithm on new data samples.

	Support Vector Machine				Gaussian Process Regression	Neural Network			
	Linear (days)	Fine Gaussian (days)	Medium Gaussian (days)	Coarse Gaussian (days)	Exponential (days)	Levenberg-Marquardt (days)	Bayesian Regularization (days)	Scaled Conjugate Gradient (days)	Actual Shelf Life (days)
1001781	4.8	4.5	5.0	4.4	4.9	4.8	4.3	4.7	4
1001848	4.9	4.3	4.7	4.5	4.3	5.1	4.3	4.7	5
1001768	4.9	4.3	5.0	4.5	4.5	4.8	4.3	4.7	4
1002034	4.3	3.7	4.1	4.2	3.5	3.8	4.3	4.2	4
1002003	5.1	4.3	4.6	4.5	4.3	4.4	4.4	3.8	5
1001811	3.4	4.3	3.8	3.9	4.3	4.1	4.2	4.0	5
1002002	4.1	3.8	4.1	4.2	3.6	3.7	4.2	3.9	4
1002033	3.6	3.8	3.1	4.0	3.6	3.5	4.1	3.5	3
1002038	3.0	4.3	2.8	3.8	3.9	2.9	4.0	2.9	4
1001770	3.2	4.1	2.9	3.9	3.6	3.3	4.1	3.3	3
MAX Absolute Deviation (days)	1.6	1.1	1.1	1.1	0.9	1.1	1.1	1.2	
MIN Absolute Deviation (days)	0.1	0.2	0.2	0.2	0.1	0.1	0	0.1	
Average Absolute Deviation (days)	0.6	0.6	0.5	0.5	0.6	0.5	0.5	0.5	
MAX Deviation (%)	31.8	36.3	31.9	31.9	21.9	27.2	37.8	27.4	
MIN Deviation (%)	2.5	5.8	4.2	4.2	1.9	1.6	0.9	1.4	
Average Deviation (%)	**14.1**	**14.5**	**12.7**	**14.0**	**14.1**	**12.4**	**13.2**	**12.8**	

6.4. Shelf Life Prediction Results Evaluation

The shelf life prediction results are presented in table 6.4. The deviation between the actual shelf life and the predicted shelf life is given in days and in a percentage. In this section the results will be analysed and validated. The reliability of the shelf life prediction generated by the selected algorithms is given in a confidence interval.

6.4.1. Absolute Prediction Accuracy

Based on the test results in table 6.4 the accuracy of learning algorithms is around 0.5 - 0.6 days. The deviation expressed in percentage shows an accuracy between 12.4 and 14.5 percent. The algorithm how has the lowest deviation is the Neural Network trained with a Levenberg-Marquard algorithm,

6.4.2. Underfitting / Overfitting

A way to compare the performance of the model on the training and test data is to use a type of average loss function. For example, the mean squared error (MSE) of the model. If the MSE of the model on the test data is higher than the MSE obtained on the training data, this is a sign of overfitting [4].

Table 6.5: MSE values for training and testing of the selected machine learning algorithms.

	Folds/Layers	Training MSE	Test MSE	Difference Training - Test
Supper Vector Machine				
Linear	5	1.6	0.55	1.05
Fine Gaussian	10	1.57	0.39	1.18
Medium Gaussian	5	1.35	0.52	0.83
Coarse Gaussian	5	1.49	0.41	1.08
Gaussian Process Regression				
Exponential	10	1.57	0.37	1.20
Neural Network				
Levenberg-Marquardt	10	1.85	0.41	1.44
Bayesian Regularization	10	1.15	0.42	0.73
Scaled Conjugate Gradient	12	0.88	0.51	0.37

The training MSE values are given by the Matlab regression learner application, and the neural network fitting application based on the training/validation of the machine learning algorithm. The test MSE values are based on the difference between the predicted shelf life and the actual shelf life.

The MSE of the model on the test data is lower than the MSE obtained on the training data. This means that there is no overfitting in the prediction of shelf life of the strawberries. If there is on overffitting, it could be that there is underfitting. Signs of underfitting is a large bias, the deviation between the predicted shelf life and the actual shelf life in table 6.4 is about half a day. This deviation is good but the fact that there is overfitting means that it can be improved by a more complex model or the features are not informative enough.

6.4.3. Confidence Interval

In this case the only output is the prediction of shelf life of strawberries. To assess the accuracy of the prediction the prediction errors given in table 6.4 are used. Based on the deviations between predicted shelf life and the actual shelf life, given in Appendix C, the 95% confidence interval is determined. The interval values are given in table 6.6.

Table 6.6: Overview of the 95 % confidence interval for each machine learning algorithm.

		Interval Left (days)	Mean (days)	Interval Right (days)	Distance Interval (days)
Supper Vector Machine					
	Linear	-0.53	0.03	0.59	1.12
	Fine Gaussian	-0.43	0.04	0.51	0.94
	Medium Gaussian	-0.63	-0.09	0.45	1.08
	Coarse Gaussian	-0.39	0.09	0.57	0.95
Gaussian Process Regression					
	Exponential	-0.51	-0.05	0.41	0.91
Neural Network					
	Levenberg-Marquardt	-0.54	-0.05	0.42	0.97
	Bayesian Regularization	-0.36	-0.06	0.60	0.96
	Scaled Conjugate Gradient	-0.66	-0.13	0.40	1.06

The exact values for the 95% confidence interval is provided for the selected algorithms are compared in table 6.6. The interval boundaries are given on each side of the mean. The distance between the interval boundaries is provided in the last column. The machine learning algorithm with the smallest boundary is the Exponential Guassian Processs Regression.

The 95% confidence interval is plotted with red lines in the figures on the next page (fig. 6.6 - 6.13). This figures show the probability density for measured deviations in prediciting shelf life of a strawberry. The next section will continue on this topic.

6.4.4. Probability Density Function

In the next section the probability distribution is given based on the deviation between the actual shelf life and shelf life prediction by the selected machine learning algorithms, see figure in Appendix C.

A probability distribution is a mathematical function that provides the probabilities of occurrence of different possible outcomes in a experiment. The are two kind of Probability distributions. A discrete probability distribution, applicable to discrete outcomes, and continuous probability distribution applicable for situations where the outcomes are continuous, like temperature or the shelf life of a strawberry. In this case the continuous probability distribution is used. The probability density of the normal distribution is given in equation 6.1, where μ is the mean of the deviation between predicted shelf life and the actual shelf life. The variable σ is the standard deviation and σ^2 is the variance. The normal distribution equation is plotted in the figures 6.6 - 6.13.

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{6.1}$$

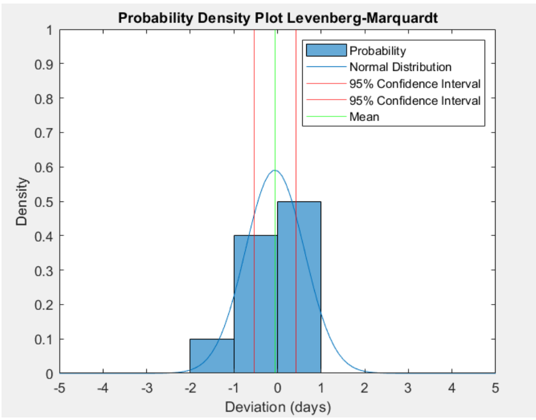


Figure 6.6: Predicted Density Plot Linear (SVM).

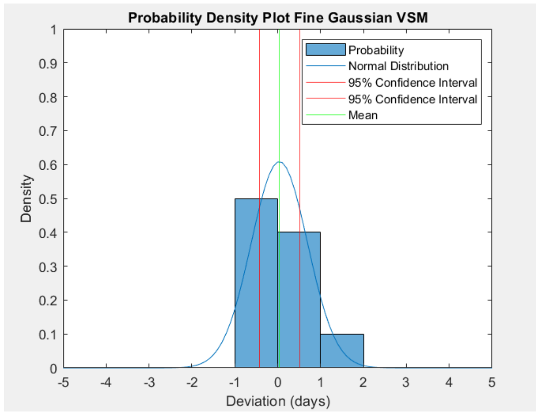


Figure 6.7: Predicted Density Plot Fine Gaussian (SVM).

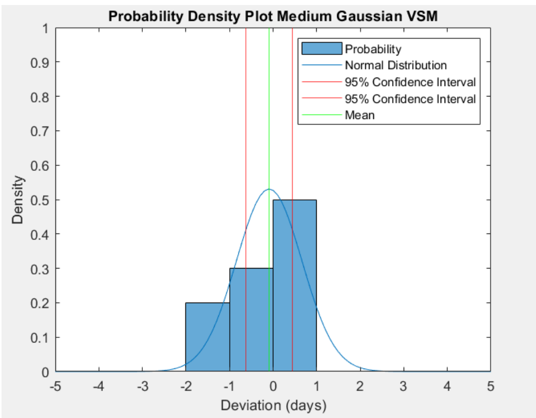


Figure 6.8: Predicted Density Plot Medium Gaussian (SVM).

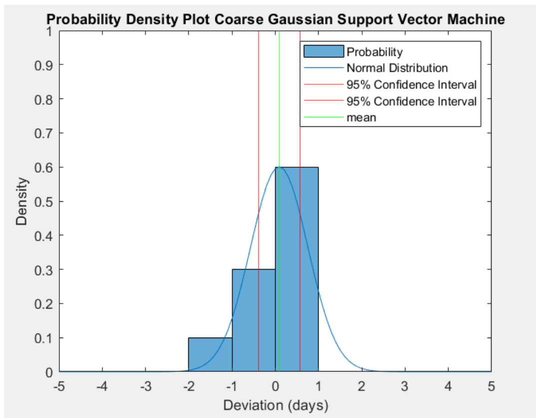


Figure 6.9: Predicted Density Plot Coarse Gaussian (SVM).

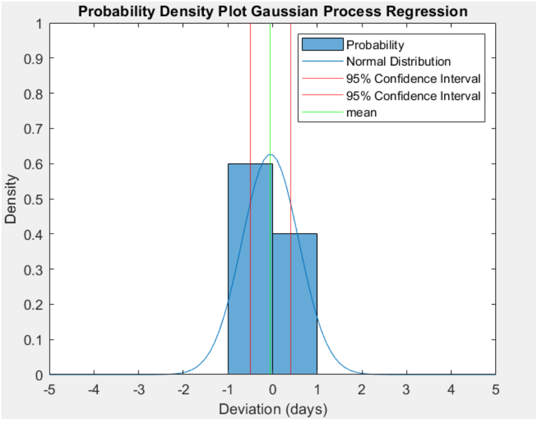


Figure 6.10: Predicted Density Plot Gaussian Process Regression Exponential.

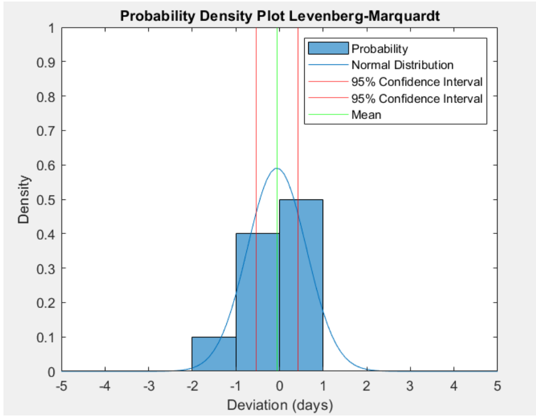


Figure 6.11: Predicted Density Plot Levenberg-Marquardt Neural Network

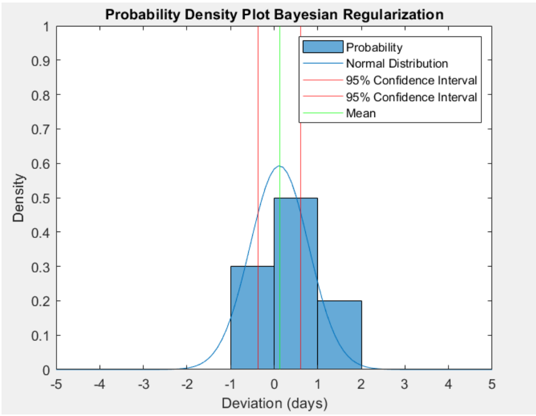


Figure 6.12: Predicted Density Plot Bayesian Neural Network.

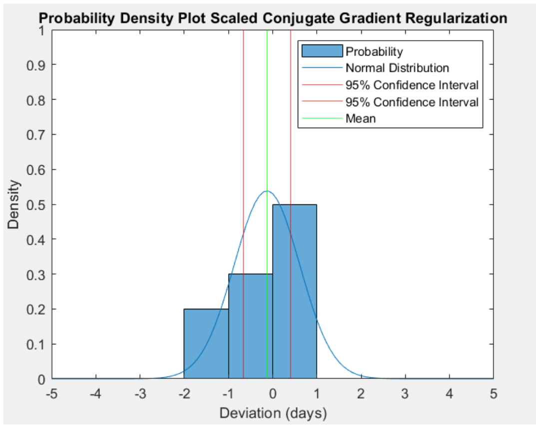


Figure 6.13: Predicted Density Plot Scaled Conjugate Gradient Neural Network.

The probability density plots for each machine learning algorithm show the normal distribution, the mean and a density histogram of the deviations between predicted shelf life and actual shelf life based on the ten test data samples given in Appendix C. The plots give an overview of the deviation spreading of each machine learning algorithm. Compared to the other algorithm, the Gaussian Process Regression Exponential algorithm has the most narrow deviation spreading.

6.4.5. Sample Size

The number of data samples used in this research was limited. To find out if the shelf life predictions depend on the number of data samples, the regression algorithms are trained with a different amount of data samples. The result of this test is given in figure 6.14.

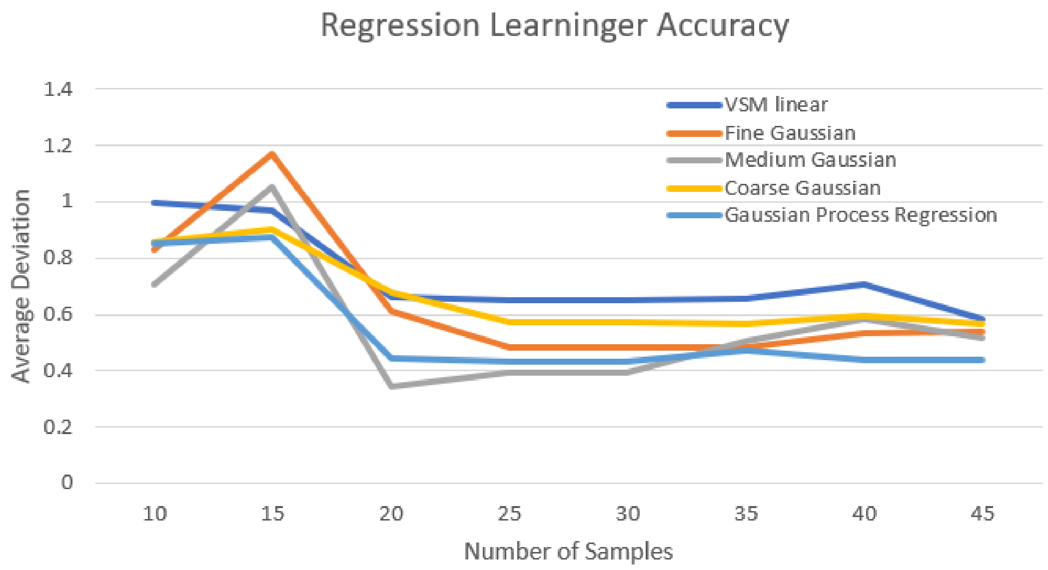


Figure 6.14: Plot of the average deviation of regression learner algorithms based on different numbers of data samples.

The average deviation between the predicted shelf life and the actual shelf life is decreasing in between a sample size of 15-20 samples and is continuing on a constant level. This means the samples size is sufficient to make accurate predictions.

This is different for the neural network predictions. Based on the figures in chapter 6.2.3, the spreading of RMSE values for different numbers of layers is to wide and to horizontal. A regular RMSE - hiddenlayer curve based on enough data samples shows underfitting for a low number of hiddenlayers and overfitting for a high number of hiddenlayers, like the curve in figure 6.15. The curves like figure 6.3 conclude that the sample size is to small to make reliable predictions. More data samples will reduce the spreading and improve the curve shape with underfitting and overfitting.

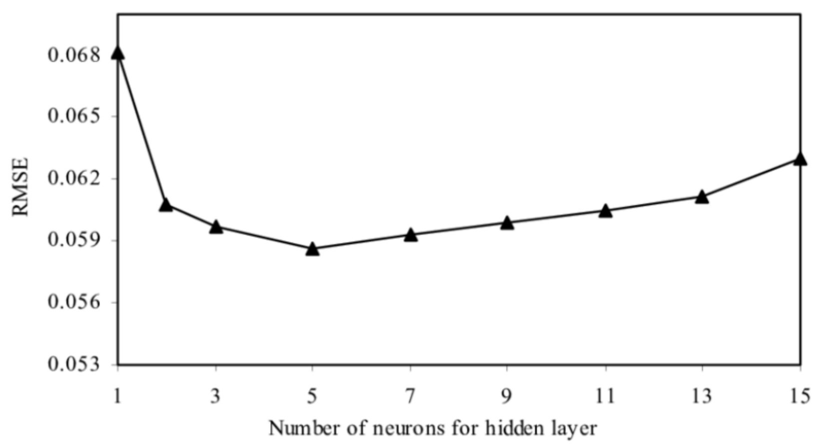


Figure 6.15: Example of a regular RMSE curve, it shows underfitting for a low number of hiddenlayers and shows overfitting for a high number of hiddenlayers.

6.4.6. Machine Learning Algorithm Selection

The test results given in table 6.4 show the average deviation between predicted shelf life and the actual shelf life. The Neural Network optimized by a Levenberg-Marquardt algorithm has the lowest average deviation. The algorithm with the smallest confidence interval and deviation spreading is the Gaussian Process Regression Exponential algorithm, based on the values in table 6.6 and figure 6.10.

When the algorithm is used in practise, the reliability of the system is most important. This means that a narrow 95% confidence interval is important and the data spreading should be small. This is the reason why the Gaussian Process Regression Exponential algorithm is favourable to be used in practise.

6.5. Shelf Life Prediction In Strawberry Supply Chain

In the previous chapter the most reliable machine learning algorithm is selected to predict the shelf life of a strawberry. The introduction of this research described three major problems in the fruit and vegetable supply chain. In this section the potential improvements in traceability, food date labelling and human quality grading will be discussed, figure 6.17 shows an overview of the moments in which shelf life prediction can be implemented.

6.5.1. Traceability

The yearly costs of The Greenery for deterioration of fruit and vegetables during distribution is – euro. The measurement of temperature through the supply chain helps to monitor the conditions in the supply chain and the conditions of the strawberries to reduce deterioration in the supply chain.

Besides the reduction in costs for deterioration in the supply chain, the costs for insurance of products will be reduced as well. The traceability of all products makes it easier to find the cause or the person responsible for the loss of product.

6.5.2. Human Quality Grading

The total yearly cost for incorrect quality determination is about – euros. These costs are based on product returns from customers who expected a different kind of quality of their ordered product.

The shelf life, is one of the quality indications assigned by the inspector. Based on the results of this research the selected algorithm is able to predict the shelf life of a strawberry (in the scope of this research) by half a day, with an accuracy of 95 percent. Based on this accuracy, strawberries could be delivered to customers with a margin of 1 day. A customer how is ordering products with a shelf life of 4 days, gets products with a predicted shelf life of 5 days. To make sure the shelf life of the product is sufficient.

In this way the inspector can be aided by grading quality of the product with an accurate prediction of the shelf life. This will reduce the costs for incorrect quality assignment, but also has potential to assign products to specific customers who order products with specific shelf life periods.

6.5.3. Food Date Labelling

The ability to assign accurate shelf life predictions to products can help to find the right retailer or trade company for the right product and gives the retailer more insight into the resulting shelf life of the product they have to sell to their customers. Possible implementations of shelf life predictions can help to improve customer assignment and dynamic pricing.

Customer Assignment

The transport conditions will be analysed and based on the initial condition of the strawberries, the quality of the arriving strawberries can be monitored through the supply chain. Every customer, online retailer, normal retailer, trade partner or industry has their own requests for product shelf life, this request can be fulfilled by scheduling the product with the predicted shelf life period to the right customer.

The shelf life period of products sold by online retailers are based on trust between the online shopper and the online retailer. The customer does not have the ability to check the product before it is delivered and the online retailer is not able to make sure every sold item has the expected shelf life. To make sure the customer gets the expected product shelf life, an accurate prediction based on data measured on product level is valuable.

The same goes for a retailer with a physical shop, in which case the customer sees what he or she is buying, which makes it easier to select the good products. But if the customer expects a shelf life of 3 days and after 2 days the strawberries are deteriorated, their will be complains. Besides the complains the products with less quality are left in the shelves and will not be sold. This costs money and increases food waste.

The products that a short shelf life could be sold for a lower price to trade companies, who sell their products on local markets. The shelf life expectations are lower and so will be the price. This way the strawberries with less shelf life will be sold and consumed.

If the shelf life of strawberries is not even worth selling to the trade companies, it can be used for industrial purposes. The strawberries can be turned into used for jam. That way the strawberries are sold and consumed in the best way.

Depended on the predicted shelf life of the monitored strawberries, the strawberries can be assigned to the right customer. By doing this customers do not receive products which do not satisfy the customers expectations. Figure 6.16 shows schematic way of product assignment based on the predicted shelf life.

Figure 6.16: Flow chart of product which is assigned to a customer depending on the predicted shelf life.

Dynamic Pricing

The ability to assign accurate shelf life predictions to products gives the retailer more insight into the resulting shelf life of the product they have to sell to their customers. To sell a product which is at the end of its shelf life, the retailer normally reduces the price. In this way the product will be sold and the expectations of the shelf life will be lower, this is called dynamic pricing.

The current way of dynamic pricing is based on the order date. The date the product is ordered is the first date of the shelf life which is requested by the retailer. When the product shelf life is reduced till one day the product will be reduced in price. This price reduction is not based on the actual shelf life of the product, it is assumed that the product will be out shelf life after a certain period. This period is determined in advance and not based on product level.

To make sure the price is determined based on the shelf life of the product, the product is measured throughout the whole supply chain and based on the shelf life prediction based on the measured data the price is determined. A customer will pay a fair price for each product which is also in proportion seen the remaining shelf life of the package of strawberries.

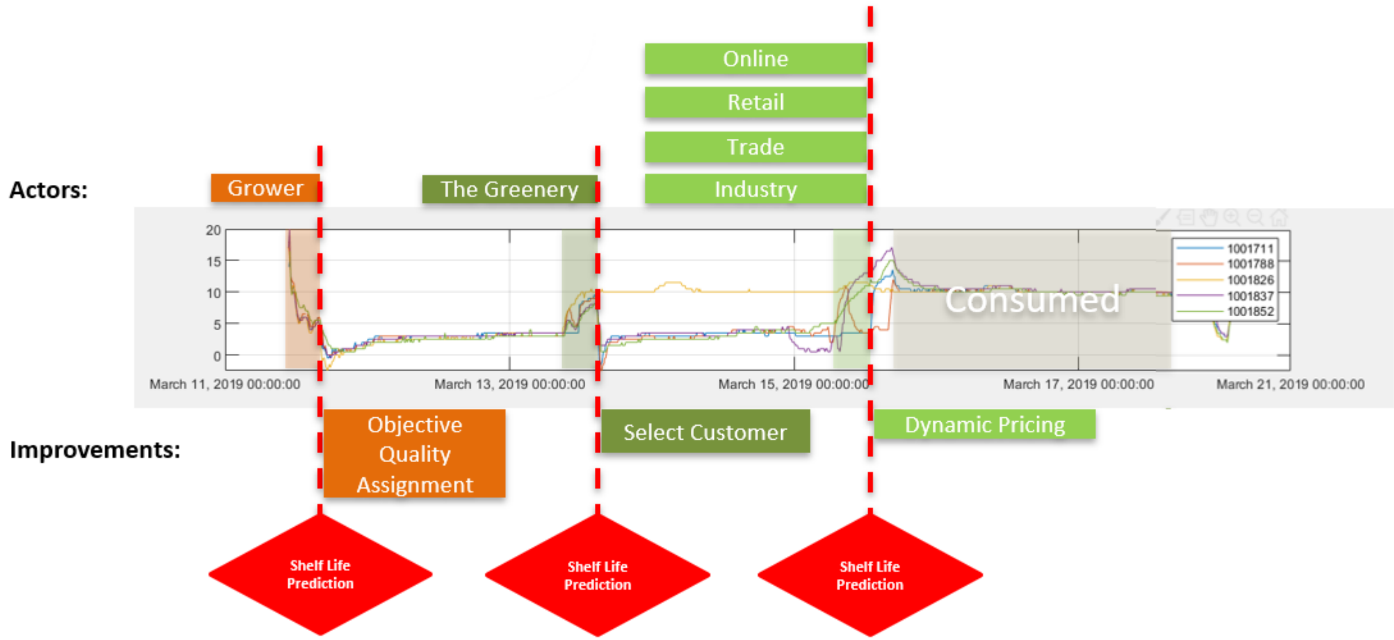


Figure 6.17: Overview of temperature profile, with three stages of shelf life prediction with a potential improvement for each actor in the strawberry supply chain.

7 Conclusion and Recommendations

7.1. Conclusion

It is a global interest to find a way to reduce food waste and to improve sustainability in food distribution. A possible way to achieve this is to find a way to optimize traceability, food date labelling and human quality grading. New developments can initiate optimization of the previous subjects. A new development is artificial intelligence (AI). It has the potential to reduce the human interaction in the supply chain and to make product inspections and/or shelf life predictions more objective. To find out what the potential is of artificial intelligence to predict shelf life, a highly perishable product is selected to be traced through the supply chain from grower till customer. A strawberry is a highly perishable and valuable product. Based on the measurements in the strawberry supply chain the following research question will be answered:

”How can artificial intelligence contribute to a better shelf life prediction of a strawberry based on temperature measurements from grower till customer?”

In this research the surrounding temperature in the strawberry supply chain is measured. The strawberry supply chain is a fast moving supply chain with through put times between 80 and 170 hours from harvesting till customer. The journey from grower till customer includes 4 moments in which the product moves through different surrounding conditions, for example from trucks, on docks and in storage facilities.

The temperatures in those locations is the most important parameter which has influence on the shelf life of a strawberry. Besides the temperature, the growing conditions, humidity and the atmosphere in which the product is produced and transported are relevant.

The temperature data is analysed and parameters, like speed of temperature reduction after harvest, number of peaks in the temperature profile or average temperature and the hours of temperature in particular zones show weak correlations with the shelf life of strawberries. Nevertheless, the accuracy of the prediction of strawberry shelf life is good. Which means the prediction based on individual features is weak, but the prediction based on the combination of the used features is good. This results shows why machine learning has potential to predict shelf life of perishable products.

The determined characteristics are used as input to learn learning algorithms. Regularly used learning algorithms, Regression and Neural Network algorithms, are selected to be tested based on the characteristics of the data samples. The machine learning algorithms are tested on their prediction accuracy. The best performing 8 algorithms are tested on new data samples. The most reliable shelf life prediction algorithm (in the scope of this research) has a 95% confidence interval between -0,5 and 0,4 days deviation of the actual shelf life which satisfies the customers requirements.

The measurement of temperature through the supply chain to predict shelf life will improve the traceability of products, or in this case the strawberries. The traceability makes it possible to find and remove problem sources in the supply chain.

The costs for incorrect quality determination can be reduced by implementing shelf life prediction into the supply chain. One of the inspection actions is to determine the resulting shelf life of the product. The inspector can be advised about the resulting shelf life by the predicted shelf life based on the measured data in the previous supply chain.

The last potential improvement is food date labelling based on product level measurements. The ability to assign accurate shelf life predictions to products can help to find the right retailer or trade company. The shelf life can be predicted with an accuracy of ±0.5 days. The product can be delivered based on a predicted shelf life with a safety margin of 1 day to a customer which expects a particular shelf life.

The package of strawberries has a predicted shelf life based on measurements on product level. This can be used to reduce the price when the strawberry shelf life is close to expiring. this way of dynamic pricing is related to the product and not based on average shelf life of strawberries.

7.2. Recommendations

The recommendations based on this research are given in this section. The recommendations are divided in recommendation for further research, chapter 7.1.1 and recommendations for The Greenery in chapter 7.1.2.

7.2.1. Recommendations For Further Research

The recommendations for the TU Delft are evaluated in this section. Most of the recommendations follow from applying this research to practise.

- **Start condition of strawberries**, the harvest condition of a strawberry at the moment the logger is implemented into the supply chain is considered mutual for all strawberries. Unless the circumstances are made as mutual as possible, there is not a clear correlation found between the features and the shelf life. The initial conditions can be the reason for this error, for example weather conditions or soil characteristics can make big differences for the final the shelf life of a strawberry, these conditions are not part of the scope of this research.

- **Measure more parameters**, parameters like humidity and $CO2/O2$ measurements are not included in this research. More input parameters with a correlation with the shelf life however increases the accuracy of the final prediction. This research focused on the temperature profiles, but a recommendation is to include and measure more parameters and to find (new) features. That way a better indication of the conditions is given to the learning algorithm with a potential of more accurate predictions.

- **Feature Selection**, the current feature selection is based on literature and expert knowledge. A next step would be to find more feature, but also find the optimal combination of features to train, validate and test the machine learning algorithms. An optimization of the features has potential to improve the accuracy of the shelf life prediction.

- **Actual Shelf Life Determination**, during this research the shelf life of strawberries which participated in the experiment was determined once a day by an qualified inspector. Which makes the judgement subjective. An improvement would be to measure the degradation of the strawberry by image recognition. To get an objective indication of degradation.

- **Increase accuracy with more samples**, by executing measurements on a bigger scale, the correlation plots will be more clear. Besides the correlation plots, the influence of data samples on the accuracy of shelf lifer prediction is analysed. For the Regression learning algorithms, more samples did not result in a more accurate prediction. In case of the neural networks, it was clear that the accuracy is very diverse based on the data samples available in this research. The relative small number of data can be the reason for this.

7.2.2. Recommendation For The Greenery

The recommendations for the The Greenery are more practically oriented and follow observations made during the research and analysis of data samples.

- **Implement this system as soon as possible**, all products that go through the systems without any measurements of conditional data coupled to shelf life data is lost valuable data. This research is based on one grower at one place with one supply chain which is measured and tested more than 50 times. To implement this system on a bigger scale, data samples need to be generated in the current process for different kind of products, growers and distribution routes. After gathering enough data, the system is able to predict the shelf life. These predictions can reduce the costs for food waste and customer complains.

- **Commercially interesting**, besides the reduction in costs the prediction of shelf life can be used for commercial purposes like dynamic pricing. By lowering the price the product will be sold and not wasted. The (online) retailer is able to sell the products based on the resulting shelf life.

- **Make the predictions accessible for actors in the supply chain** the predictions of shelf life need to be accessible for all actors in the supply chain. The shelf life of product can change by bad weather conditions during harvest or by condition variations in the supply chain. A changed shelf life period can force the actors in the supply chain to change the destination of the product on a early basis.

A
Appendix

Fold	Logger Number	Grower Number	Variety	Peaks	Average Temp (Celsius)	Cooling slope (Celsius/10 min)	Temp above 7 (hours)	Temp between 7-4 (Celsius)	Temp between 2-4 (hours)	Temp between 2- -2 (hours)	Temp under -1 (hours)	Shelf Life (days)
F1	1001775 - Shelflife	608952	Calinda	1.00	3.41	4.00	5.83	4.00	29.00	11.33	1.67	6.00
F1	1001814 - Hollander	608952	Calinda	2.00	3.47	1.50	6.17	3.33	80.50	9.17	0.67	3.00
F1	1001803 - Blaak	608952	Calinda	4.00	3.47	2.50	6.50	10.67	71.17	11.67	0.67	3.00
F1	1001803 - Both	608952	Calinda	1.00	4.14	3.50	13.00	16.17	60.50	11.17	0.00	3.00
F1	1001759 - Waardenburg	608952	Calinda	3.00	3.88	3.50	10.00	7.50	65.33	17.50	0.00	5.00
F1	1001775 - Shelflife	608952	Calinda	4.00	1.44	10.50	1.83	3.50	2.00	42.50	3.33	7.00
F1	1001814 - Hollander	608952	Calinda	5.00	2.45	9.50	3.83	3.33	117.00	42.17	3.67	7.00
F1	1001803 - Both	608952	Calinda	5.00	2.52	10.00	4.33	7.33	97.83	57.33	3.33	5.00
F1	1002037 - Dubbeldam	608952	Calinda	8.00	2.91	5.00	4.00	9.00	112.00	45.00	0.00	5.00
F1	1001759 - Waardenburg	608952	Calinda	6.00	2.50	6.50	6.67	5.67	97.33	54.50	5.67	5.00
F1	1001768 - Shelflife	608952	Calinda	0.00	2.79	2.00	5.67	3.17	5.33	36.17	0.00	6.00
F2	1001818 - Hollander	608952	Calinda	3.00	4.20	1.50	28.50	5.50	99.67	36.00	0.00	5.00
F2	1002034 - Both	608952	Calinda	4.00	2.32	1.50	5.50	3.00	58.00	103.83	0.00	5.00
F2	1001781 - Dubbeldam	608952	Calinda	4.00	2.82	3.00	2.33	3.17	127.50	36.67	0.00	7.00
F2	1001842 - Waardenburg	608952	Calinda	5.00	2.73	4.00	5.33	8.67	109.83	46.17	0.00	5.00
F2	1001789 - Shelflife	608952	Calinda	1.00	2.73	1.50	3.00	3.67	10.83	34.17	0.00	5.00
F2	1001808 - Both	608952	Calinda	4.00	2.97	3.50	9.33	12.83	23.83	56.67	0.00	3.00
F2	1001849 - Dubbeldam	608952	Calinda	2.00	2.76	1.50	3.67	7.67	50.67	40.33	0.00	2.00
F2	1001844 - Waardenburg	608952	Calinda	4.00	3.37	1.50	11.17	11.00	36.83	43.33	0.00	2.00
F2	1001811 - Shelflife	608952	Calinda	2.00	2.53	1.50	6.17	5.83	8.17	36.33	0.00	5.00
F2	1002002 - Hollander	608952	Calinda	1.00	2.69	1.50	4.83	9.17	51.17	38.17	0.00	4.00
F2	1002033 - Both	608952	Calinda	6.00	2.90	2.50	5.50	18.17	40.33	38.50	0.00	3.00
F3	1002038 - Dubbeldam	608952	Calinda	3.00	3.11	1.50	5.50	21.83	39.33	36.50	0.00	4.00
F3	1001772 - Waardenburg	608952	Calinda	5.00	2.98	2.00	5.00	17.83	44.17	35.17	0.00	3.00
F3	1001784 - Shelflife	608952	Calinda	0.00	1.67	7.00	3.00	2.17	4.00	42.17	0.00	4.00
F3	1001825 - Hollander	608952	Calinda	1.00	2.51	3.00	7.00	3.83	25.17	42.00	0.00	4.00

F3	1001770 - Blaak	608952	Calinda	3.00	2.51	2.50	6.83	12.17	19.50	42.00	0.00	3.00
F3	1001795 - Both	608952	Calinda	2.00	1.89	5.00	3.50	4.83	23.67	46.50	0.00	4.00
F3	1001804 - Dubbeldam	608952	Calinda	3.00	1.87	2.50	1.50	10.17	22.83	41.50	2.00	4.00
F3	1001830 - Waardenburg	608952	Calinda	3.00	2.37	2.50	7.33	8.83	20.00	42.17	1.00	3.00
F3	1001799 - Shelflife	608952	Calinda	1.00	3.56	1.00	10.67	12.17	30.67	42.00	0.00	2.00
F3	1001817 - Hollander	608952	Calinda	4.00	4.00	1.50	17.50	14.33	104.17	11.17	0.00	4.00
F3	1001832 - Blaak	608952	Calinda	6.00	4.23	1.50	14.17	33.83	101.00	0.00	0.00	4.00
F4	1001792 - Both	608952	Calinda	10.00	4.06	3.00	16.00	26.33	100.17	4.83	0.00	4.00
F4	1001766 - Dubbeldam	608952	Calinda	6.00	4.50	2.00	15.50	41.83	91.17	0.00	0.00	4.00
F4	1002030 - Waardenburg	608952	Calinda	6.00	4.44	1.00	23.67	24.83	93.17	6.50	0.00	4.00
F4	1001825 - Shelflife	608952	Calinda	1.00	1.95	2.00	4.17	4.33	2.50	38.33	3.67	7.00
F4	1001844 - Hollander	608952	Calinda	2.00	2.75	2.00	5.67	8.67	70.17	41.67	0.00	6.00
F4	1001759 - Blaak	608952	Calinda	4.00	3.17	3.00	6.33	16.00	61.50	42.83	0.00	4.00
F4	1001775 - Both	608952	Calinda	3.00	2.96	3.00	6.00	7.50	61.67	51.50	0.00	5.00
F4	1001788 - Dubbeldam	608952	Calinda	3.00	2.87	2.00	4.83	5.33	64.33	51.67	0.00	5.00
F4	1001849 - Waardenburg	608952	Calinda	5.00	3.08	4.50	5.33	8.17	68.83	44.00	0.00	4.00
F4	1002003 - Shelflife	608952	Calinda	1.00	3.26	4.50	3.00	2.83	31.33	7.83	0.50	5.00
F4	1001852 - Hollander	608952	Calinda	2.00	3.57	8.50	4.33	7.83	72.67	10.67	0.00	5.00
F5	1001837 - Blaak	608952	Calinda	2.00	3.56	10.00	5.50	8.67	70.17	11.83	0.00	6.00
F5	1001826 - Both	608952	Calinda	3.00	3.21	7.50	4.17	8.33	69.17	14.67	0.00	5.00
F5	1001818 - Dubbeldam	608952	Calinda	3.00	3.70	4.50	3.67	11.00	70.17	11.17	0.00	6.00
F5	1001771 - Waardenburg	608952	Calinda	2.00	3.63	5.50	5.17	7.50	74.67	8.83	0.00	2.00
F5	1001854 - Shelflife	608952	Calinda	1.00	3.14	1.00	7.17	4.67	23.33	37.50	0.00	4.00
F5	1002034 - Hollander	608952	Calinda	3.00	2.66	1.00	4.33	9.83	70.00	63.50	0.00	4.00
F5	1001855 - Blaak	608952	Calinda	3.00	3.24	1.00	5.50	12.00	94.50	36.33	0.00	4.00
F5	1001781 - Both	608952	Calinda	3.00	3.15	3.00	5.17	11.83	94.50	36.67	0.00	4.00
F5	1001848 - Dubbeldam	608952	Calinda	3.00	3.77	4.00	5.17	17.17	120.00	4.83	0.00	5.00
F5	1001768 - Waardenburg	608952	Calinda	3.00	3.45	4.00	6.17	11.17	102.50	27.67	0.00	4.00

B Appendix

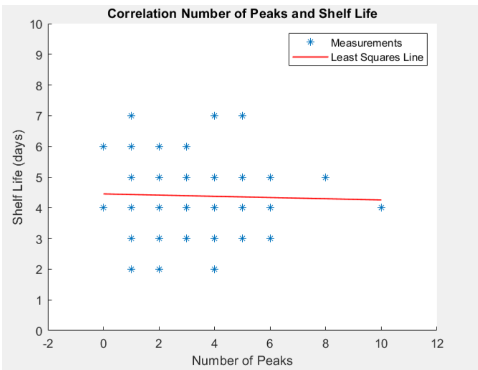


Figure B.1: Correlation plot, between peaks vs shelf life.

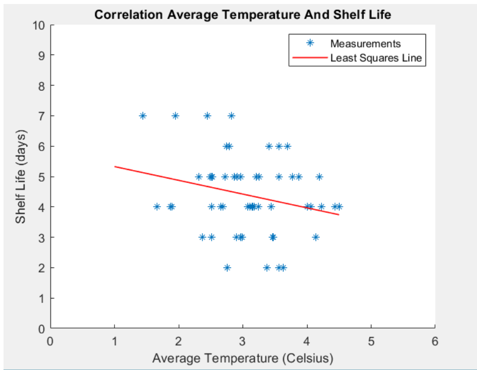


Figure B.2: Average temperature vs shelf life

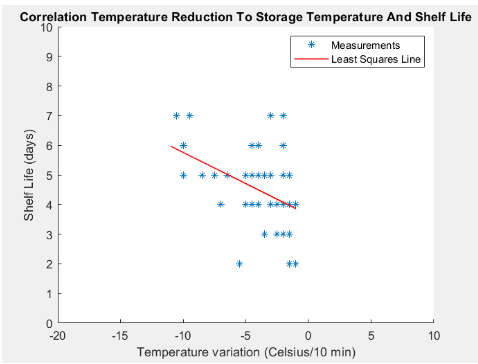


Figure B.3: Temperature reduction vs shelf life.

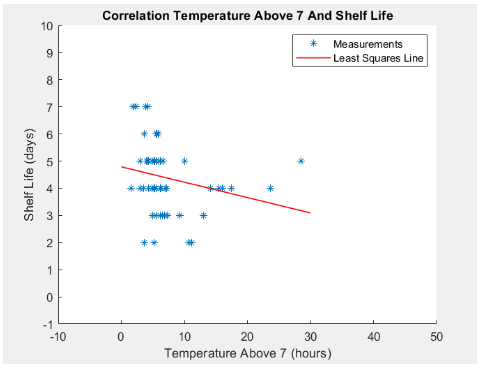


Figure B.4: Temperature 7+ degrees Celsius vs shelf life

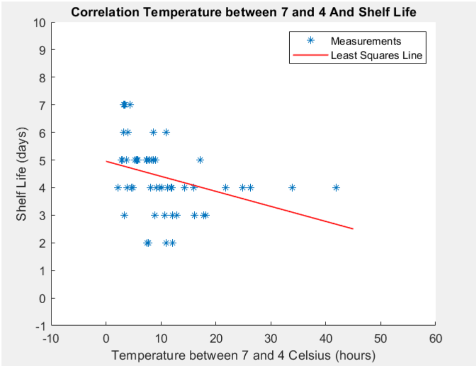


Figure B.5: Temperature 7 - 4 degrees Celsius vs shelf life.

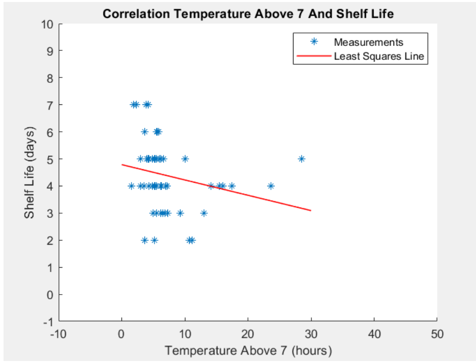


Figure B.6: Temperature 4 - 2 degrees Celsius vs shelf life

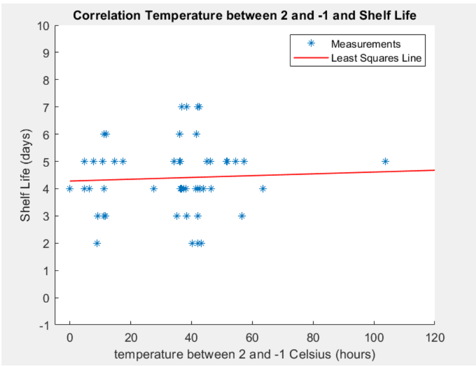


Figure B.7: Temperature 2 - (-1) degrees Celsius vs shelf life.

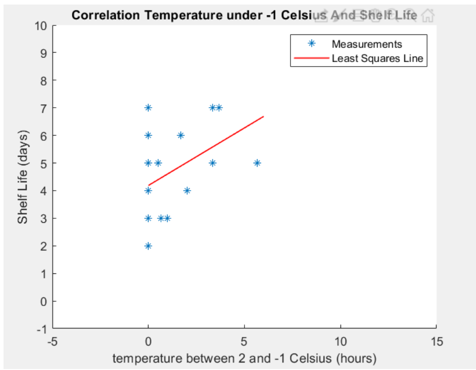


Figure B.8: Temperature under -1 degrees Celsius vs shelf life

C
Appendix

	Supper Vector Machine				Gaussian Process Regression	Neural Network		
	Linear	Fine Gaussian	Medium Gaussian	Coarse Gaussian	Exponential	Levenberg-Marquardt	Bayesian Regularization	Scaled Conjugate Gradient
1001781	0.8	0.5	1	0.4	0.9	0.8	0.3	0.7
1001848	-0.1	-0.7	-0.3	-0.5	-0.7	0.1	-0.7	-0.3
1001768	0.9	0.3	1	0.5	0.5	0.8	0.3	0.7
1002034	0.3	-0.3	0.1	0.2	-0.5	-0.2	0.3	0.2
1002003	0.1	-0.7	-0.4	-0.5	-0.7	-0.6	-0.6	-1.2
1001811	-1.6	-0.7	-1.2	-1.1	-0.7	-0.9	-0.8	-1
1002002	0.1	-0.2	0.1	0.2	-0.4	-0.3	0.2	-0.1
1002033	0.6	0.8	0.1	1	0.6	0.5	1.1	0.5
1002038	-1	0.3	-1.2	-0.2	-0.1	-1.1	0	-1.1
1001770	0.2	1.1	-0.1	0.9	0.6	0.3	1.1	0.3

D
Appendix

Defect Day 0 (day of Final Packing)/ gebreken op dag van inzet	Defects/gebreken Day 1	Defects/gebreken Day 2	Defects/gebreken Day 3	Defects/gebreken Day 4	Defects/gebreken Day 5	Defects/gebreken Day 6	Defects/gebreken Day 7	Defect Comments/opmerkingen gebreken	Day of Shelflife Fail (day of 1st red)/ dag waarop product in shelflife niet meer voldoet
0	onderweg	1	1	1	2	3	4	beschadigingen	6
0	onderweg	3	4	4	5	6	6	bederf/beschadigingen	2
0	onderweg	2	3	3	3	4	4	bederf/schimmel	3
0	onderweg	2	4	4	5	6	6	bederf/schimmel	3
0		3	4	5	6	6	7	schimmel/drukplekken	2
0		3	3	3	4	5	5	drukplekken	2
0	0	1	1	1	2	2	2	bederf	8
0	0	0	0	1	1	2	2	drukplekken	8
0	1	1	1	1	1	2	3	bederf	7
0	0	1	1	1	1	2	3	bederf	7
0	1	1	1	2	3	3	3	bederf	5
0	0	1	1	2	3	3	3	bederf	5
0	0	0	0	1	1	1	2	drukplekken	8
0	0	1	1	2	2	2	3	bederf	7
0	0	1	1	1	2	2	3	bederf	7
0	0	1	2	3	3	3	4	bederf	4
0	0	1	2	2	2	2	4	bederf	7
0	0	1	2	4	4	4	5	bederf/schimmel	4

Figure D.1: Actual shelf life overview of three packages of strawberries with a temperature logger and the actual shelf life determined, red boxes mean product is expired.

E
Appendix

E.1. Neural Network Selection

Besides the regression learning algorithms, neural network fitting is also investigated as potential shelf life prediction algorithm. To find out how accurate neural network fitting can be, a neural network fitting application of matlab is used. The following section explains the steps taken to validate the accuracy of neural network fitting.

E.1.1. Process Description

Steps by step the process to validate the accuracy of Neural Network fitting is described. From selecting data till validation of the final result the next steps describe the process. Figure E.1 gives a flow chart of the algorithm accuracy evaluation procedure.

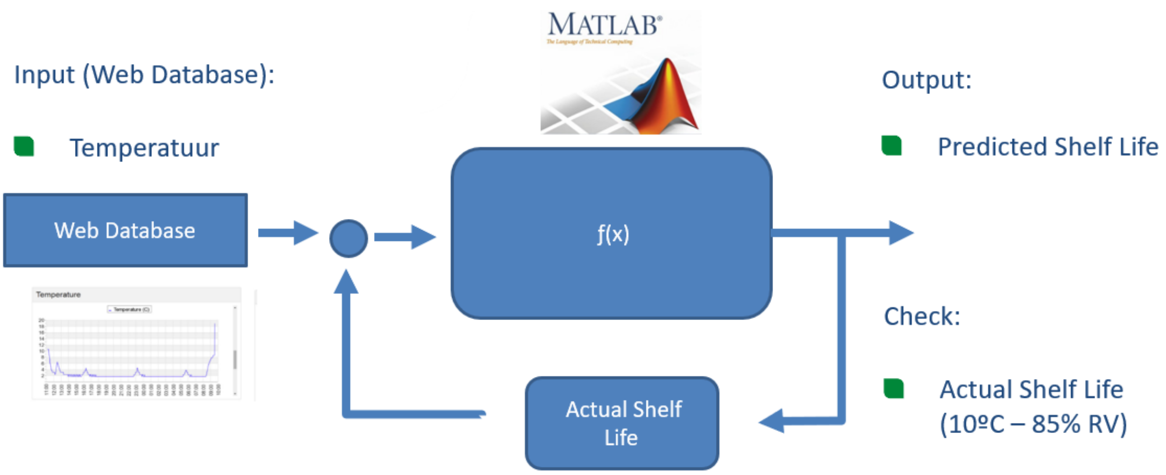


Figure E.1: Neural Network architecture generated by Matlab Neural Network fitting application.

Step 1: Select Input and Target Data

The first step of the Neural Network application integrated in Matlab is to select the input data, used to train the model, and the target data, used to train and validate the algorithm. Figure E.2 shows the data selection dashboard.

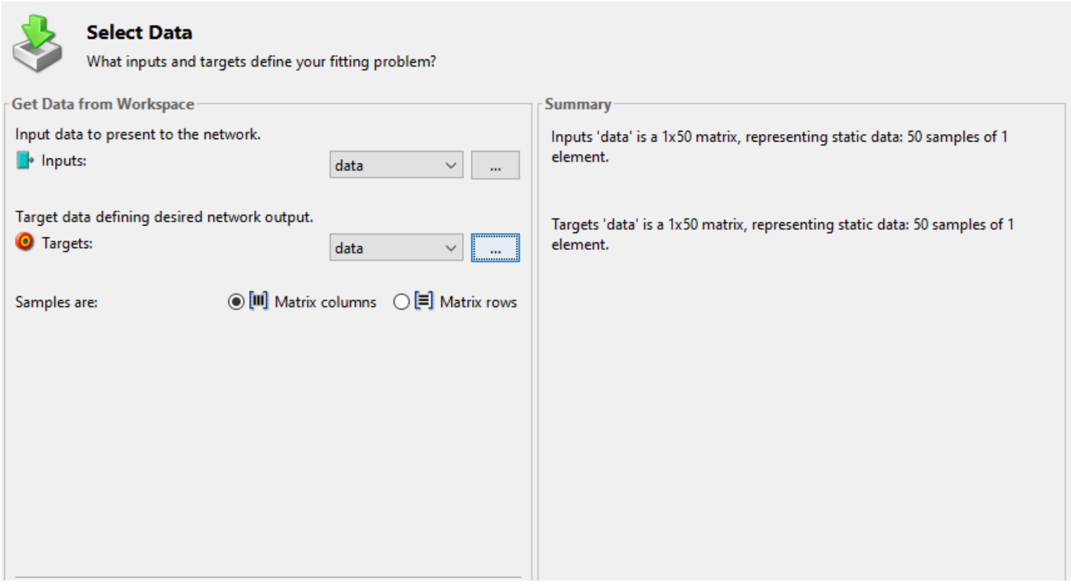


Figure E.2: Neural Network fitting dashboard, import data.

Step 2: Determine The Amount Of Training, Validation And Testing Data

The second step is to select how much percentage of the data samples is used to train, validate and test the generated model, fig. E.3. A higher amount of training data results in a more accurate prediction. In this research, 90 percent of the data samples is used for training, 5 percent is used to validate the trained algorithm and 5 percent is used to test the neural network fitting algorithm.

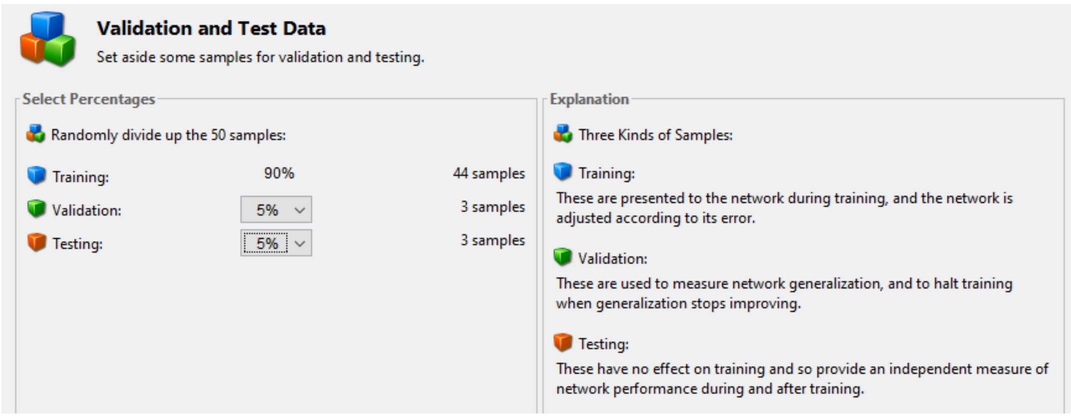


Figure E.3: Select percentage of validation data or test data.

Step 3: Define Network Architecture

A neural network consists of hidden layers which influence the characteristics of the network. The number of hidden layers is variable. This step describes the process to find the optimal number of layers to generate the most accurate prediction. First of all, the selection dashboard is shown in figure E.4. In the right corner the preferred number of hidden layers can be selected.

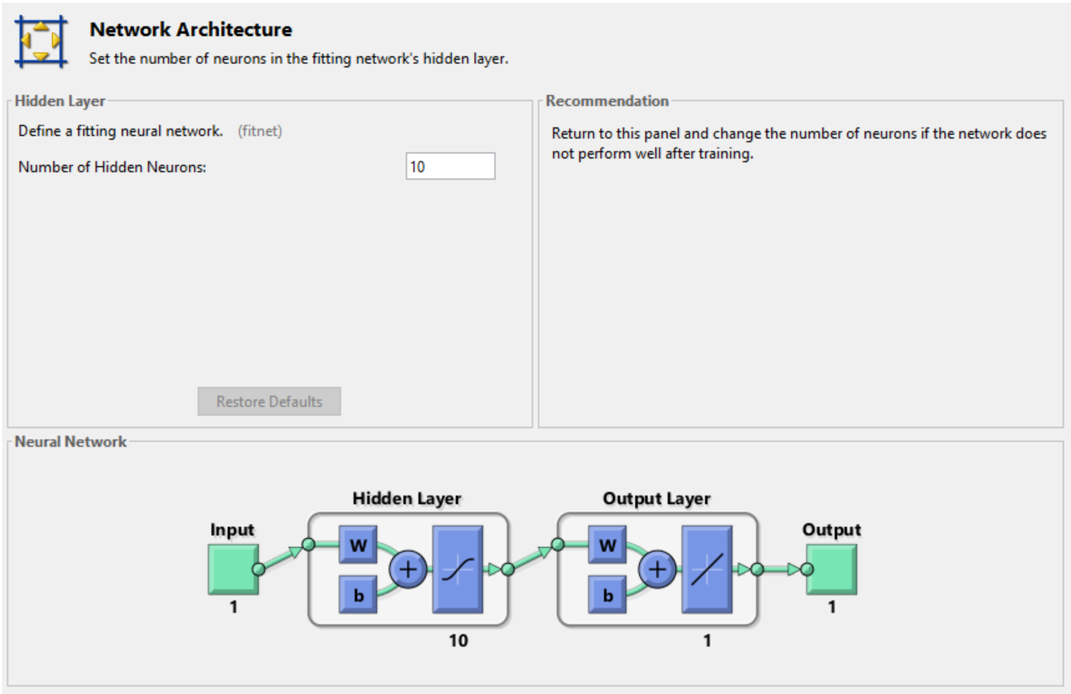


Figure E.4: The hidden layer selection dashboard.

Bibliography

[1] Ultan Mc Carthy, Ismail Uysal, Ricardo Badia-melis, Samuel Mercier, Colm O Donnell, and Anastasia Ktenioudaki. Trends in Food Science & Technology Global food security – Issues , challenges and technological solutions. *Trends in Food Science & Technology*, 77(August 2017):11–20, 2018.

[2] M. L A T M Hertog, H. A M Boerrigter, G. J P M Van Den Boogaard, L. M M Tijskens, and A. C R Van Schaik. Predicting keeping quality of strawberries (cv. 'Elsanta') packed under modified atmospheres: An integrated model approach. *Postharvest Biology and Technology*, 15(1):1–12, 1999.

[3] Daniel Kirsch. *Machine Learning by Judith Hurwitz and.*

[4] Andriy Burkov. THE HUNDRED - PAGE MACHINE LEARNING.

[5] Rohith Gandhi. K Nearest Neighbours - Introduction to Machine Learning Algorithms. *Towards Data Science*, 2018.

[6] Transport Technology. Specialization : Transport Engineering and Logistics Report number : 2016 . TEL . 8087 Title : Redesign the cool chain for air transport of perishable goods by KLM Cargo F . C . T . van der Voort Author :. 2016.

[7] International Congress. *FAO Food waste report.* 2011.

[8] Dana Gunders. *Wasted : How America Is Losing Up to 40 Percent of Its Food from Farm to Fork to Landfill.* Number august. 2012.

[9] Petter Olsen and Melania Borit. How to define traceability. *Trends in Food Science and Technology*, 29(2):142–150, 2013.

[10] Carmen Priefer, Juliane Jörissen, and Klaus Rainer Bräutigam. Food waste prevention in Europe - A cause-driven approach to identify the most relevant leverage points for action. *Resources, Conservation and Recycling*, 109:155–165, 2016.

[11] Yuca Waarts, Mieke Eppink, Elsje Oosterkamp, Sabine Hiller, A van der Sluis, and Toine Timmermans. *Reducing food waste; obstacles experienced in legislation and regulations.* Number 2011-043. 2011.

[12] Jana Valant. 'Best before' date labels: Protecting consumers and limiting food waste. (February), 2015.

[13] J. P. Kerry, M. N. O'Grady, and S. A. Hogan. Past, current and potential utilisation of active and intelligent packaging systems for meat and muscle-based products: A review. *Meat Science*, 74(1):113–130, 2006.

[14] Theofania Tsironi, Eleni Gogou, Eirini Velliou, and Petros S. Taoukis. Application and validation of the TTI based chill chain management system SMAS (Safety Monitoring and Assurance System) on shelf life optimization of vacuum packed chilled tuna. *International Journal of Food Microbiology*, 128(1):108–115, 2008.

[15] By V Rouillard. PAPER PRESENTED AT IAPRI SYMPOSIUM 2013 Quantifying the Non-stationarity of Vehicle Vibrations with the Run Test. (July 2013):203–219, 2014.

[16] Combinating food waste in retail: Pioneering technology cuts price as products expires. 2018 author = Gaynor Selby, number = Jul, url = https://www.foodingredientsfirst.com/news/combating-food-waste-in-retail-pioneering-technology-cuts-price-as-product-expires.html.

[17] David Silverberg. This app lets people buy food cheaply right before it expires. (Jun), 2016.

[18] Shveta Mahajan, Amitava Das, and Harish Kumar Sardana. Image acquisition techniques for assessment of legume quality. *Trends in Food Science and Technology*, 42(2):116–133, 2015.

[19] Susan Linko. Expert systemsÐ what can they do for the food industry ? 9, 1998.

[20] J. M. McKinion and H. E. Lemmon. Expert systems for agriculture. *Computers and Electronics in Agriculture*, 1(1):31–40, 1985.

[21] F. Goyache, A. Bahamonde, J. Alonso, S. Lopez, J. J. Del Coz, J. R. Quevedo, J. Ranilla, O. Luaces, I. Alvarez, L. J. Royo, and J. Diez. The usefulness of artificial intelligence techniques to assess subjective quality of products in the food industry. *Trends in Food Science and Technology*, 12(10):370–381, 2001.

[22] Shiguo Wang. A comprehensive survey of data mining-based accounting-fraud detection research. *2010 International Conference on Intelligent Computation Technology and Automation, ICICTA 2010*, 1:50–53, 2010.

[23] Victoria J. Hodge and Jim Austin. A of Outlier Detection MSurveyethodoligies. *Artificial Intelligence Review*, 22(1969):85–126, 2004.

[24] P. Vithu and J. A. Moses. Machine vision system for food grain quality evaluation: A review. *Trends in Food Science and Technology*, 56:13–20, 2016.

[25] A.; Aly Al-Marakeby A.A.; Salem, F.A. Fast quality inspection of Foof products using computer vision. *International Journal of Advanced Research in Computer and Communication Engineering*, 2(11):4, 2013.

[26] Nandan Thor. Applying Machine Learning Clustering and Classification to Predict Banana Ripeness States and Shelf Life. *Cloud Publications International Journal of Advanced Food Science and Technology*, 2(1):20–25, 2017.

[27] Ganjar Alfian, Jongtae Rhee, Hyejung Ahn, Jaeho Lee, Umar Farooq, Muhammad Fazal Ijaz, and M. Alex Syaekhoni. Integration of RFID, wireless sensor networks, and data mining in an e-pedigree food traceability system. *Journal of Food Engineering*, 212:65–75, 2017.

[28] S. G. Gwanpua, P. Verboven, D. Leducq, T. Brown, B. E. Verlinden, E. Bekele, W. Aregawi, J. Evans, A. Foster, S. Duret, H. M. Hoang, S. Van Der Sluis, E. Wissink, L. J.A.M. Hendriksen, P. Taoukis, E. Gogou, V. Stahl, M. El Jabri, J. F. Le Page, I. Claussen, E. Indergård, B. M. Nicolai, G. Alvarez, and A. H. Geeraerd. The FRISBEE tool, a software for optimising the trade-off between food quality, energy use, and global warming impact of cold chains. *Journal of Food Engineering*, 148(October 2009):2–12, 2015.

[29] Lixing Wang, S. K. Kwok, and W. H. Ip. A radio frequency identification and sensor-based system for the transportation of food. *Journal of Food Engineering*, 101(1):120–129, 2010.

[30] Nodali Ndraha, Hsin I. Hsiao, Jelena Vlajic, Ming Feng Yang, and Hong Ting Victor Lin. Time-temperature abuse in the food cold chain: Review of issues, challenges, and recommendations. *Food Control*, 89:12–21, 2018.

[31] Ricardo Badia-Melis, Ultan Mc Carthy, and Ismail Uysal. Data estimation methods for predicting temperatures of fruit in refrigerated containers. *Biosystems Engineering*, 151:261–272, 2016.

[32] Hongwen HUANG. Chapter 8 - Harvest and Storage. *Kiwifruit*, pages 297–307, 2016.

[33] Pol Tijskens. *Discovering the Future : Modelling Quality Matters*. 2004.

[34] Myo Min Aung and Yoon Seok Chang. Temperature management for the quality assurance of a perishable food supply chain. *Food Control*, 40(1):198–207, 2014.

[35] James M. Lyons. Chilling Injury in Plants. *Annual Review of Plant Physiology*, 24(1):445–466, 1973.

[36] M. C.N. Nunes, J. P. Emond, and J. K. Brecht. Quality of strawberries as affected by temperature abuse during ground, in-flight and retail handling operations. *Acta Horticulturae*, 604:239–246, 2003.

[37] Maarten L A T M Hertog, Ismail Uysal, Bert M Verlinden, and Bart M Nicolaï. Shelf life modelling for warehouse management. *Philosophical transactions of THE ROYAL SOCIETY*, page 15, 2014.

[38] M Cecilia, Nascimento Nunes, Mike Nicometo, Jean Pierre Emond, and Ricardo Badia Melis. Improvement in fresh fruit and vegetable logistics quality. 2014.

[39] Debra Chilson, Astrid Delgado, and Maria Cecilia N. Nunes. Shelf Life of Cluster Tomatoes (Lycopersicum esculentum) Stored at a Non-chilling Temperature and Different Relative Humidity Levels. *Proc. Fla. State Hort. Soc.*, 124(863):246–255, 2011.

[40] M. C.N. Nunes, A. Delgado, and J. P. Emond. Quality curves for green bell pepper (capsicum annuum L.) stored at low and recommended relative humidity levels. *Acta Horticulturae*, 945(April 2012):71–78, 2012.

[41] Cecilia N. Nunes, Yavuz Yagiz, and Jean Pierre Emond. Influence of environmental conditions on the quality attributes and shelf life of 'Goldfinger' bananas. *Postharvest Biology and Technology*, 86:309–320, 2013.

[42] M C N Nunes, A.M.M.B. Morais, J K Brecht, and S A Sargent. Fruit maturity and storage temperature influence response of strawberries to controlled atmospheres. *Journal of the American Society of Horticultural Science*, 127(5):836–842, 2002.

[43] Eva Almenar, Pilar Hernández-Muñoz, José M. Lagarón, Ramón Catalá, and Rafael Gavara. Controlled atmosphere storage of wild strawberry fruit (Fragaria vesca L.). *Journal of Agricultural and Food Chemistry*, 54(1):86–91, 2006.

[44] Andreas Kaplan and Michael Haenlein. Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, 62(1):15–25, 2019.

[45] Léon Bottou. From Machine Learning to Machine Reasoning. 2011.

[46] E D Liddy. Natural Language Processing. In Encyclopedia of Library and Information Science. *Encyclopedia of Library and Information Science*, 2001.

[47] David E. Wilkins. *Practical Planning*.

[48] Konstantinos G. Liakos, Patrizia Busato, Dimitrios Moshou, Simon Pearson, and Dionysis Bochtis. Machine learning in agriculture: A review. *Sensors (Switzerland)*, 18(8):1–29, 2018.

[49] Martin T Hagan and Mohammad B Menhaj. Brief Papers. *Brain and Cognition*, 30(3):275–326, 1996.

[50] Ananth Ranganathan. The Levenberg-Marquardt Algorithm. *Tutoral on LM algorithm*, 11(June):101–110, 2004.

[51] T. Chai and R. R. Draxler. Root mean square error (RMSE) or mean absolute error (MAE)? *Geoscientific Model Development Discussions*, 7(1):1525–1534, 2014.