NONDESTRUCTIVE INTERNAL QUALITY EVALUATION OF PEARS USING X-RAY IMAGING AND MACHINE LEARNING

Tim VAN DE LOOVERBOSCH

Supervisors:
Prof. Bart Nicolaï
Dr. Pieter Verboven
Prof. Jan Sijbers (Universiteit Antwerpen)

Members of the examination committee:
Prof. Marc Hendrickx - chair
Prof. Jan De Beenhouwer (Universiteit Antwerpen)
Prof Wouter Saeys
Prof. Tinne Tuytelaars
Prof. Martine Wevers

Dissertation presented in partial fulfilment of the requirements for the degree of Doctor of Bioscience Engineering (PhD)

Februari 2022

Doctoraatsproefschrift nr. 1751 aan de faculteit Bio-ingenieurswetenschappen van de KU Leuven

Uitgegeven in eigen beheer, Tim Van De Looverbosch, Willem De Croylaan 42, B-3001 Leuven (Belgium)

Dankwoord

In de eerste plaats wil ik mijn promotoren prof. Bart Nicolaï, prof. Jan Sijbers en dr. Pieter Verboven bedanken om mij deze kans te hebben geboden. Bedankt voor jullie begeleiding, jullie inspiratie en inzichten, voor het delen van jullie expertise, en voor de sturing bij het schrijven van papers en dit manuscript. In het bijzonder bedankt ik dr. Pieter Verboven voor zijn dagdagelijkse begeleiding, om in mijn kunnen te geloven, en om mij zoveel bij te brengen. Mede dankzij jullie kijk ik tevreden terug op dit belangrijke hoofdstuk waarin ik zowel op persoon als professioneel vlak ben gegroeid.

Daarnaast wil ik ook graag de leden van de examencommissie, prof. Wouter Saeys, prof. Martine Wevers, prof. Tinne Tuytelaars en prof. Jan De Beenhouwer, bedanken voor de tijd die zij hebben vrijgemaakt om dit manuscript kritisch te evalueren en constructieve feedback te geven.

Ook wil ik de leden de hele MeBioS Postharvest onderzoeksgroep en het VCBT bedanken voor alle hulp. Bedankt aan Jeroen en Ann voor het in goede banen leiden van het verkrijgen en het bewaren van de vruchten. Bedankt aan Wouter en Willem om de veteranen te zijn waarop ik kon rekenen bij de start van mijn doctoraat en de vele uren die we al sportend, lachend en drinkend hebben gedeeld. Agnese, Jakub, Hans, Celine, Maxime, Hui, Bayu, Solomon, Leroi, Valérie, Zi, Astrid, Leen, Michiel, Ujjwal en Bart, bedankt voor de leuke momenten binnen en buiten de werk(m)uren. De afgelopen 2 jaar hebben we elkaar jammer genoeg minder gezien dan gepland, waarvoor geen dank aan de coronapandemie… Dat halen we nog in!

Een bijzondere dank aan mijn gezin; mama, papa, Hanne en Cato, en mijn grootouders voor alle steun en jullie grote vertrouwen en interesse. Wie had ooit gedacht dat jullie 4 jaar lang, vrijwillig, naar een verhaaltje over peren zou luisteren? Ook heel veel dank aan mijn schoonfamilie om naast mijzelf, ook mijn uitleg over mijn onderzoek te omarmen.

Last but not least, bedank ik mijn vriendin Maria Paulina voor haar steun en haar onuitputtelijk geduld en vertrouwen. Dankzij jou liet ik de moed nooit zakken en kon ik mijn gedachten verzetten wanneer ik onnodig zat te piekeren. Met jou aan mijn zij ben ik klaar om weer aan het volgende te beginnen.

Bedankt!

Tim

Merksplas, februari 2022

Abstract

Pear fruit are prone to developing internal disorders that leave consumers dissatisfied and unwilling to repeat their purchase. It must thus be prevented that defect fruit reach the consumer. Internal disorders, however, are invisible from the surface of the fruit. In practice, batches of fruit are, therefore, often discarded based on the result of a manual destructive inspection of a small number of randomly sampled fruit. This leads to unacceptable financial losses and waste due to the disposal of the healthy fruit still present in the batch. Moreover, the sampled fruit may not be representative of the whole batch. Prior research has shown that X-ray imaging is a promising instrumental technique for detecting internal disorders. The aim of this PhD was, therefore, to provide novel, performant, and nondestructive methods to analyze X-ray images automatically in an objective way. This was done by implementing deep learning, which is a new paradigm in machine learning, that allows algorithms to learn directly from data.

First, a method was developed to detect pears with internal disorders with X-ray Computed Tomography (X-ray CT) data using a conventional machine learning strategy. Herein, an image processing algorithm was developed to extract features from the 3D data. Thereafter, a classifier was trained to distinguish healthy and defect pears based on the extracted features. The classifier achieved classification accuracies ranging between 90.2 and 95.1 % depending on the cultivar and number of features that were used. However, the proposed method had several disadvantages. It required a handcrafted feature extraction algorithm. Potentially better features, which were not thought of during development, remained unexplored. In addition, the feature extraction algorithm is possibly application specific. Furthermore, while the method allowed for classifying defect and sound fruit, it could not quantify the severity of the disorders which may be of high importance for consumers and thus for making decisions on discarding fruit or not.

To circumvent these disadvantages, a deep neural network was used to segment different internal structures in CT images, including

healthy tissue, the core, cavities and tissue affected by internal browning. The model was trained on manually annotated CT scans of healthy and defect fruit. On an independent test set, a very good agreement was found between the predicted and ground truth "healthy tissue", "core" and "cavity" labels (average intersection over union (IoU) ≥ 0.95). Interestingly, low IoU scores were found for the "internal browning" label, even though visually most predictions seemed sufficiently accurate. It turned out this was mainly caused by negligible errors on small volumes and volume edges. From the predicted labels of the model, the severity of the internal disorders could be quantified by calculating the affected volumes. The resulting quantitative data was used to classify "consumable" *vs "*non-consumable" fruit at high accuracy (99.4 %) on the one hand and "healthy" *vs* "defect but consumable" *vs* "non-consumable" classification on the other hand (92.2 %). Herein, the identification of "defect but consumable" fruit showed to be the most difficult.

A concern with X-ray CT, however, is that it is currently not applicable inline at the speed of commercial sorting lines (10 fruit/s). X-ray radiography, on the other hand, can easily be implemented inline using an X-ray source and detector on either side of a conveyor belt. A downside of X-ray radiography, however, is that it only produces a 2D projection of the absorption by a 3D object. Compared to X-ray CT, it is thus more challenging to distinguish contrast in the image caused by internal disorders and contrast caused by the shape and internal structure of the fruit. An anomaly detection approach using deep learning was proposed to detect internal disorders in X-ray radiographs of pears, recognizing recent advantages in deep learning, while overcoming the need for annotated data normally required for supervising the model during training.

The anomaly detection model was trained exclusively on X-ray radiographs of healthy pears, after which they were evaluated on a test set with images of healthy and defect pears. Defect pears could be identified based on the anomaly score produced by the model. It was shown that performance could be significantly improved by using synthetic anomalies. Herein fake defects were added to X-ray

images of healthy pears. ROC analysis showed that the proposed method was on par (mean area under the curve (AUC) up to 0.962) with a state-of-the-art benchmark method which was given several advantages (mean AUC = 0.963). The best anomaly detection model achieved an overall accuracy of 95 %, with true positive and false positive rates equal to 91.8 and 0.8 %, respectively. By investigating the performance as a function of internal disorder severity, it was shown that using the proposed method, defect fruit with a cavity percentage > 1.0 % could be detected 100 % accurate, while for lower cavity percentages the accuracy depended on the internal browning severity. The black-box nature of neural networks was addressed by producing heatmaps of the anomalous regions found by the models.

In this research, a large step forward was made towards internal disorder detection in pears using X-ray imaging. It is expected that deep learning and X-ray imaging will increasingly be adopted by various industries for quality inspection. Hereto, the presented methods might be used for the inspection of other fruits or vegetables, or be implemented in other applications, such as foreign object detection in foods. Future work should focus on further investigating deep learning-based approaches, such as unsupervised learning, and to further discover the possibilities, but also limitations, of X-ray based inspection of foods. To bring quality inspection to an even higher level, further research is required in fast inline X-ray CT systems, and other X-ray based methods, such as X-ray phase contrast imaging.

Beknopte samenvatting

Peren zijn vatbaar voor het ontwikkelen van interne gebreken tijdens bewaring. Om ontevreden consumenten, die niet bereid zijn om hun aankoop te herhalen, te vermijden, moet ten zeerste worden voorkomen dat defecte vruchten de consument bereiken. Interne gebreken zijn echter onzichtbaar vanaf het oppervlak van de vrucht. In de praktijk worden partijen fruit daarom vaak weggegooid op basis van een handmatige destructieve inspectie van een steekproef van een klein aantal willekeurige vruchten. Dit leidt echter tot onaanvaardbare financiële verliezen en verspilling door het onnodig afvoeren van het gezond fruit dat nog wel in de partij aanwezig kan zijn. Bovendien kunnen de onderzochte vruchten niet representatief zijn voor de hele partij. Voorafgaand onderzoek heeft aangetoond dat X-stralenbeeldvorming een veelbelovende instrumentele techniek is om interne gebreken op te sporen. Het doel van dit doctoraat was daarom om nieuwe, performante en niet-destructieve methodes te ontwikkelen om X-stralenbeelden automatisch en objectief te analyseren, en interne gebreken te detecteren. Dit werd bereikt door het toepassen van deep learning, een nieuw paradigma in machine learning, waarbij algoritmen rechtstreeks kunnen leren van data.

Eerst werd een methode ontwikkeld om peren met interne gebreken te detecteren met behulp van X-stralen computertomografie (CT) en een conventionele machine learning-strategie. Hierbij werd een beeldverwerkingsalgoritme ontwikkeld om kenmerken uit de 3D-data te extraheren. Daarna werd een classificatiealgoritme getraind op basis van de geëxtraheerde kenmerken om gezonde en defecte peren te onderscheiden. Het classificatiealgoritme bereikte nauwkeurigheden tussen 90,2 en 95,1 %, afhankelijk van de cultivar en het aantal gebruikte kenmerken. De voorgestelde methode had echter enkele nadelen. Het vereiste een handgemaakt algoritme voor het extraheren van kenmerken. Mogelijk bleven nuttigere kenmerken, waaraan tijdens de ontwikkeling niet werd gedacht, onontdekt. Bovendien is het beeldverwerkingsalgoritme algoritme mogelijks specifiek voor de toepassing. Hoewel de methode het

mogelijk maakte om defect en gezond fruit te classificeren, kon het ook de graad van de gebreken niet kwantificeren. Dit is echter van groot belang voor consumenten en dus ook voor het nemen van beslissingen over het al dan niet verwijderen van fruit.

Om deze nadelen te omzeilen, werd een diep neuraal netwerk gebruikt om verschillende interne structuren in CT-beelden te segmenteren, waaronder gezond weefsel, het klokhuis, holtes en weefsel dat is aangetast door intern bruin. Het model werd getraind op handmatig geannoteerde CT-scans van gezond en defect fruit. Op een onafhankelijke test-dataset werd een zeer goede overeenkomst gevonden tussen de voorspelde en de manuele annotaties voor de labels "gezond weefsel", "klokhuis" en "holtes" (gemiddelde intersectie over unie (IoU) $\geq 0{,}95$). Het was opvallend dat lage IoU-scores gevonden werden voor het label "intern bruin", hoewel visueel de meeste voorspellingen voldoende nauwkeurig leken. Het bleek dat dit voornamelijk werd veroorzaakt door verwaarloosbare fouten op kleine volumes en randen. Uit de voorspelde labels van het model kon de graad van de interne gebreken worden gekwantificeerd door de aangetaste volumes te berekenen. De resulterende kwantitatieve gegevens werden gebruikt om enerzijds “consumeerbaar” vs. “niet-consumeerbaar” fruit te classificeren met een hoge nauwkeurigheid (99,4 %) en anderzijds voor de classificatie van “gezond” vs. “defect maar consumeerbaar” vs. “niet-consumeerbaar” fruit (92,2 %). Hierin bleek de identificatie van "defect maar consumeerbaar" fruit het moeilijkst te zijn.

Een bezorgdheid bij het gebruiken van CT is echter dat het momenteel niet aan een hoge snelheid toepasbaar is op commerciële sorteerlijnen (10 vruchten/s). X-stralenradiografie kan daarentegen eenvoudig op sorteerlijnen worden toegepast met behulp van een X-stralenbron en -detector aan weerszijden van een transportband. Een nadeel van X-stralenradiografie is echter dat het enkel een 2D-projectie produceert van de absorptie van een 3D-object. Ten opzichte van CT is het dus een grotere uitdaging om het onderscheid te maken tussen enerzijds contrast in het beeld veroorzaakt door interne gebreken en anderzijds contrast veroorzaakt door de vorm en interne structuur van de vrucht. Een anomaliedetectiemethode werd voorgesteld met behulp van deep learning om interne

aandoeningen in radiografieën van peren te detecteren. Hierbij werden de voordelen van de recente ontwikkelingen in deep learning meegenomen, terwijl de behoefte aan geannoteerde gegevens, die normaal nodig zijn voor het trainen van een model, wordt weggenomen.

Het anomaliedetectiemodel werd uitsluitend getraind op radiografieën van gezonde peren, waarna deze werden beoordeeld op een test-dataset met radiografieën van gezonde en defecte peren. Defecte peren konden worden geïdentificeerd op basis van de anomaliescore die door het model werd voorspeld. Er werd aangetoond dat de prestaties aanzienlijk konden worden verbeterd door gebruik te maken van synthetische anomalieën. Hierin werden gebreken artificieel nagebootst en toegevoegd aan radiografieën van gezonde peren. ROC-analyse toonde aan dat de voorgestelde methode hetzelfde niveau behaalde (gemiddelde oppervlakte onder de curve (AUC) tot 0,962) als een state-of-the-art benchmarkmethode die verschillende voordelen kreeg (gemiddelde AUC = 0,963). Het beste anomaliedetectiemodel behaalde een classificatie nauwkeurigheid van 95 %, met percentages voor echt-positieven en vals-positieven gelijk aan respectievelijk 91,8 en 0,8 %. Door de prestaties te onderzoeken in functie van de graad van de interne gebreken werd aangetoond dat met de voorgestelde methode defecte vruchten met een holtepercentage > 1,0 % met 100 % nauwkeurigheid konden worden gedetecteerd. Voor lagere holtepercentages was de nauwkeurigheid afhankelijk van de graad van intern bruin. De black-box benadering van neurale netwerken werd aangepakt door heatmaps te maken van de afwijkende regio's die door de modellen werden gevonden in de radiografieën.

Een grote vooruitgang werd geboekt in de detectie van interne gebreken in peer met behulp van X-stralenbeeldvorming. Deep learning en X-stralenbeeldvorming zullen in toenemende mate door verschillende industrieën kunnen worden toegepast voor kwaliteitsinspectie. Hiertoe zouden de voorgestelde methodes eenvoudig kunnen worden gebruikt voor de inspectie van ander fruit of groenten, of worden uitgebreid naar andere toepassingen, zoals de detectie van vreemde voorwerpen in voedingsmiddelen. Toekomstig onderzoek kan zich toeleggen op het verder

onderzoeken van op deep learning gebaseerde methodes, zoals *unsupervised learning*, alsook op het verder ontdekken van de mogelijkheden, maar ook beperkingen, van op X-stralen gebaseerde inspectie van voedingsmiddelen. Om kwaliteitsinspectie naar een nog hoger niveau te tillen, is verder onderzoek nodig naar snelle, op productielijnen integreerbare CT-systemen en andere op X-stralen gebaseerde methoden, zoals fasecontrast beeldvorming.

List of abbreviations

1D	1-dimensional
2D	2-dimensional
3D	3-dimensional
AD	anomaly detection
AE	autoencoder
AI	artificial intelligence
ANN	artificial neural network
AUC	area under the curve
BCE	binary cross-entropy
CA	controlled atmosphere
CE	cross-entropy
CNN	convolutional neural network
DAE	denoising autoencoder
DDM	density distribution model
DL	deep learning
FBP	filtered back-projection
FCDD	fully convolutional data description
FNR	false negative rate
FPR	false positive rate
GPU	graphics processing unit
Guided Grad-CAM	Guided Gradient-weighted Class Activation Mapping
HEDDM	heterogeneous density distribution model
HODDM	homogeneous density distribution model
IoU	Intersection over Union
kNN	k-nearest neighbor
KS-test	Kolmogorov-Smirnov test
ML	machine learning

MLP	multi-layer perceptron
MRI	magnetic resonance imaging
MSE	mean squared error
Multisensor HEDDM	Multisensor inspection method using heterogeneous density distribution model
Multisensor HODDM	Multisensor inspection method using homogeneous density distribution model
OE	outlier exposure
R	linear correlation coefficient
ReLU	rectified linear units
RFE method	recursive feature elimination method
RGB	red green blue
ROC curve/analysis	receiver operating characteristic curve/analyses
ROS	reactive oxygen species
SSIM	structural similarity index
SVM	support vector machine
TNR	true negative rate
TPR	true positive rate
Vis-NIR	visual - near infrared
X-ray CT	X-ray computed tomography

Content

Chapter 1

Introduction

1.1 Problem statement

In 2019, Belgium was placed third in the list of top pear producing countries among the European member states (European Commission, 2020). Globally, Belgium ranked 8th for pear production in both value and quantity produced in 2019. Pear production thus represents an important part of the Belgian economy. It had a total value of 161 million euros in 2016, which was around 40 % of the Belgian fruit production (Danckaert et al., 2018). In terms of consumption, pears hold the 5th place in the top fruit consumption in Belgium (VLAM, 2021a). While Belgian pears are highly consumed in the domestic market, it is Belgium's most valuable exported fruit with a value of 243 million euro in 2020. Up to 80 % of the annual pear production is exported (Avermaete et al., 2018; VLAM, 2021b). Delivering quality fruit year-round is thus of high importance to maintain this position on the international market.

To preserve the quality of fresh fruit after harvest, controlled atmosphere storage is applied (Mercier et al., 2017; Thompson et al., 2018). Suboptimal storage conditions, however, can cause severe quality loss by chilling injury, accelerated ripening and senescence, fermentation, stimulated pathogen growth or other physiological decay. Examples are internal browning, watercore, bitter pit or cavities in fruit tissue. Detecting and removing defect fruit from the supply chain is of high importance. It must be prevented that defect fruit can reach consumers and reduce their trust and willingness to pay or to repeat their purchase. Additionally, it enables market players to differentiate themselves from others by providing top quality fruit.

The disorders that develop during growth or storage may not cause externally visible symptoms (Franck et al., 2007; Thompson et al., 2018). As a result, they may be impossible to be detected by current commercial quality grading systems that are mainly focused on the external quality. In practice, whole batches are, therefore, discarded based on the manual destructive inspection of a small sample of fruit rather than of only filtering out the defect fruit. Batch-wise decision-making leads to unacceptable collateral damage, i.e., financial loss and food waste due to the disposal of the healthy fruit still present in the batch. Moreover, the small subset of the batch that is selected for inspection may not be representative of the whole batch. Nondestructive methods for detecting internal disorders are thus required to enable a reliable and inline (i.e., on sorting lines) inspection of each individual fruit.

Research has shown that X-ray imaging is a promising instrumental technique for detecting internal disorders (Arendse et al., 2018; Kotwaliwale et al., 2014; Nicolaï et al., 2014). Thanks to their high energy, X-rays have a good penetration depth through biological material that allows for the visualization of their internal structure. Different materials or material densities inside a sample cause differences in X-ray attenuation. This results in patterns on X-ray images that can be used to analyze the sample under investigation. X-ray imaging methods are mainly applied in two modes. In X-ray radiography, X-rays transmitted through an object are captured by a detector to create a 2D X-ray image, i.e., a radiograph or projection. In X-ray computed tomography (CT), on the other hand, many of such projections are taken from different angles and are combined to produce an information-rich 3D reconstruction of the sample. While X-ray radiography can be easily implemented inline using a source and line detector on either side of a conveyer belt, inline X-ray CT requires more complex and expensive hardware.

Due to the costly hardware requirements and insufficient speed, X-ray CT based methods have mainly been used to characterize internal disorders using manual and semi-automatic image processing workflows rather than to detect internal disorders automatically. Additionally, the described characterization methods are often slow and involve manual annotation that does not scale to

large datasets. Automated X-ray CT based methods would allow for internal disorder detection and quantification. Since continuous developments are made towards inline X-ray CT, inline applications of such automated methods may become feasible in the future. Available X-ray radiography-based methods often use internal disorder and application specific algorithms. This complicates their transferability to other disorders and other biological products with significant differences in shape, size, and composition. A more general purpose multisensor algorithm has been developed, but it still requires application-specific 3D fruit models and complex sensor integration (van Dael et al., 2019, 2017). There is, thus, a need for a generally applicable method for nondestructive internal disorder detection that can be transferred to other applications with less effort, and which can be more easily implemented inline.

1.2 Objectives and outline

To this day, nondestructive detection of internal disorders in horticultural products remains a challenging problem. While X-ray imaging-based methods have been shown to be effective in research, their implementation is hard due to limited transferability to other contexts, or complex sensor integration. Particularly, all proposed methods so far relied on conventional machine learning methods. Herein, experts decide on how underlying algorithms should operate (e.g., image processing pipelines) or on what an algorithm needs (e.g., feature extraction) to solve a task. This results in suboptimal solutions that are biased and limited by the human capability of interpreting the data. Therefore, the aim of this PhD was to develop novel nondestructive methods for internal disorder inspection for X-ray CT and inline radiography using a new paradigm in machine learning, namely deep learning, in which a model can learn to solve a task end-to-end from “raw” data.

In Chapter 2 a background is provided on internal disorders in pears, X-ray imaging and deep learning. Additionally, the current state of the art is reviewed, and the open challenges and opportunities related to the technologies and applications of this work are identified.

In the first research chapter (Chapter 3), a conventional machine learning-based method was developed for X-ray CT to nondestructively detect internal disorders in pears. Herein, a classifier was trained on features extracted by an image processing algorithm to discriminate between defect and sound fruit. Although the classification performed well, there was still room for improvement and the method did not provide output that could be used to further characterize internal disorders, e.g., disorder severity or location.

In response to these shortcomings a deep neural network for image segmentation was developed for X-ray CT images to indicate internal disorders in pears on pixel level and quantify the disorder severity (see Chapter 4). The model was able to segment healthy tissue, the core, tissue affected by internal browning and cavities. Based on the volumes of internal disorders, healthy and defect fruit could be classified downstream at a high accuracy.

In the final research chapter (Chapter 5), a deep learning method is developed for inline X-ray radiography to detect and localize internal disorders in pears. Herein, one of the problems of deep learning, the heavy reliance on labeled data, is tackled by approaching the problem as an anomaly detection challenge in which a model is trained in an unsupervised way.

Finally, Chapter 6 delivers the general conclusions from this research and provides suggestions for future work.

Chapter 2

State-of-the-art

2.1 Introduction

This chapter reviews internal disorders in pears and discusses methods proposed in the literature, with emphasis on X-ray imaging, to detect them. First, common internal disorders in pome fruit are discussed with a brief explanation on how most storage-related disorders develop during controlled atmosphere (CA) storage. Second, the state-of-the-art of internal disorder detection in horticultural products is presented. Third, the basic principles of X-ray imaging are explained, including the interaction of X-rays with matter, X-ray radiography and X-ray CT. Next, relevant image-based machine learning and deep learning methods are discussed. Finally, as a conclusion of this review, the open challenges and opportunities related to internal disorder detection in horticultural products using X-ray imaging are formulated.

2.2 Internal disorders in pome fruit

Pre- and postharvest conditions may disturb the metabolic activity of fruit tissue and lead to various types of physiological disorders, including internal disorders. Due to their economic importance, internal disorders in pome fruit have been studied to better understand and prevent them. While apple and pear are two different species, it is assumed that in both species similar mechanisms are responsible for the development of internal disorders. Due to specific characteristics of different species and cultivars, however, the susceptibility towards different types of internal disorders is species and cultivar dependent. Examples of common internal disorders in pome fruit are internal browning and cavity formation, watercore and watercore breakdown, bitter pit and chilling injury. This section mainly discusses internal browning and cavity formation (see Figure 2.1) since these are the most

common internal disorders in 'Conference' pear, the most widely cultivated pear cultivar in Belgium (Franck et al., 2007).

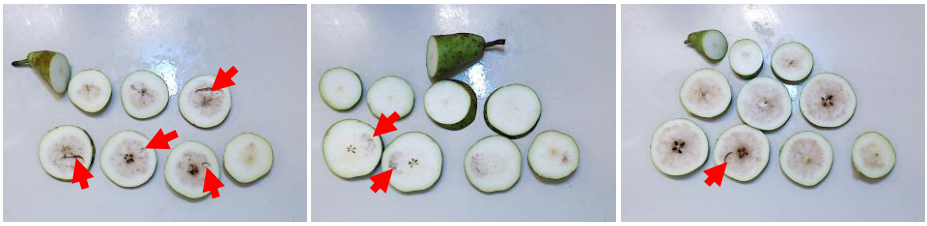

Figure 2.1: Conference pears affected by internal browning and cavity formation. Cavities are indicated by red arrows.

For the development of internal browning during CA storage, the mechanism is presumably the following (Franck et al., 2007; Herremans, Verboven, et al., 2014; Ho et al., 2013; Pedreschi et al., 2009; Veltman et al., 2003). As oxygen concentrations are lowered in CA storage to minimize aerobic respiration, the O_2 concentration inside the fruit may become so low that a local hypoxic or even anoxic state is created due to the limited rate of oxygen supply through the fruit tissue. The rate of oxygen supply is then insufficient to maintain the respiration process and together with an increased CO_2 concentration a shift can take place from respiration to fermentation. Compared to aerobic respiration, however, fermentation is an inefficient process that provides little energy. In addition, fermentation also produces ethanol which causes an off flavor. Due to the lowered energy supply, cells cannot keep up their maintenance processes, such as the repair of cell membranes. Moreover, the shift to fermentation also leads to an imbalance between the oxidative and reductive processes. In oxidative stress conditions, the antioxidant system works insufficiently to eliminate the reactive oxygen species (ROS) that will, among others, further degrade the cell membrane. Eventually, this leads to cell leakage and cell death. Due to cell decompartmentalization, phenols (located in the vacuole) can react with polyphenol oxidase (located in the plastids) to form the polymers that are responsible for the brown color. Cavities form once enough neighboring cells have died and the free liquid resulting from the leakage has diffused towards the boundary of the fruit and evaporated.

Several aspects influence the susceptibility of pome fruit towards internal browning. For instance, larger fruit have a higher risk of entering a hypoxic state due to the larger resistance against oxygen diffusion, resulting in a larger gradient from outer surface to the core of the fruit. Fruit porosity, which is species and cultivar dependent (Z. Wang et al., 2020), is positively correlated with gas diffusivity and thus negatively correlated with internal disorder development. Also, preharvest factors, such as seasonal characteristics and picking date influence internal browning susceptibility (Franck et al., 2007; Lammertyn et al., 2000).

The spatial and temporal distribution of the developmental stages of internal browning has been investigated in apple and pear using X-ray Computed Tomagraphy (CT) and Magnetic Resonance Imaging (MRI) (Gonzalez et al., 2001; Herremans et al., 2013; Lammertyn et al., 2003b, 2003a; Suchanek et al., 2017). It was found that in pear internal browning develops in two distinct patterns, i.e., radial browning and local browning (also called asymmetrical browning). In radial browning, the disorder follows the shape of the fruit concentrically, while in asymmetrical browning the disorder appears in specific local spots (Lammertyn et al., 2003b, 2003a). While the radial browning can be explained from the internal gas gradients, it is not yet fully understood how local defects may develop. Presumably, local microstructural characteristics, e.g., low pore connectivity, are responsible for making specific regions more susceptible (Herremans et al., 2013; Herremans, Verboven, et al., 2014). Moreover, it was found that from the moment the disorder was detected after several weeks of storage, the disorder intensity increased, but, counterintuitively, the affected region did not grow in size (Lammertyn et al., 2003a). On the microstructure level, Herremans at al. (2013) observed the flooding of pores due to cell collapse in brown tissue and the subsequent collapse of tissue and cavity formation during internal browning development in apples using X-ray CT (see Figure 2.2). They also observed that the middle ($R - 5\ mm > X > 0.65 \times R$; R: fruit radius) and inner cortex ($0.65 \times R > X$) were more susceptible to internal browning than the outer cortex ($R - 5\ mm > X$).

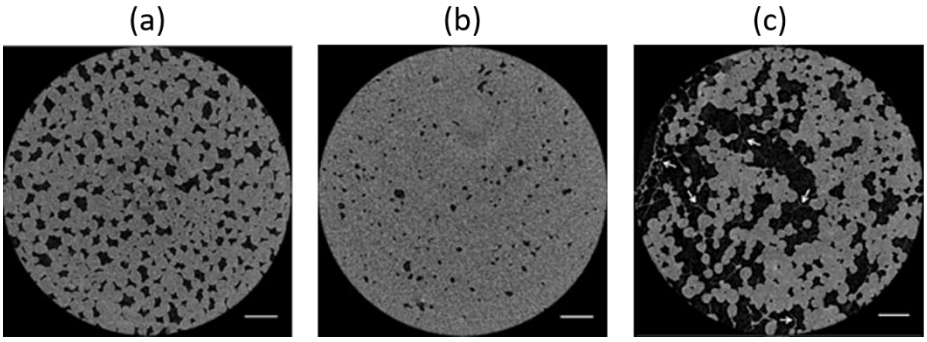

Figure 2.2: Cross-sectional slice of an X-ray CT scan of healthy tissue (a), flooded tissue (b) and tissue with cavities formed (c) of the middle cortex (located at radial distance X; 5 mm > X > 0.65 × R; R: fruit radius) of 'Braeburn' apples. Arrows indicate remains of collapsed cells. Scale bars indicate 500 μm. Adapted from Herremans et al. (2013).

Based on the above discussion on the mechanism of internal disorder development, several aspects about detecting internal disorders non-destructively must be considered. First, instrumental techniques used to detect internal disorders must provide sufficient depth into the fruit, since the sensitivity to internal disorders is a function of the radial distance from the core. Second, sufficient spatial information is required since internal disorders may not only develop in a radial pattern that is spread out over the whole fruit volume, but also locally, which might be more challenging to detect. Third, the success of the instrumental techniques will depend on the physical principles it is based on and on the developmental stage of the disorder. Fourth, severity can be interpreted in two ways: the size of the affected volume, or the degree of tissue degradation inside the affected volume. For clarity, the former will be further referred to as disorder severity, while the later will be denoted as disorder intensity. It is expected that both factors will affect the detectability of the disorder, which can differ for methods using different instrumental techniques. Nonetheless, consumers might tolerate low severity with high intensity (e.g., a cavity) more than high severity with low intensity (e.g., radial browning).

2.3 X-ray imaging

X-ray imaging is a promising technique for internal disorder detection in pears. In the following sections, the interactions of X-rays with matter and its suitability for internal quality inspection are discussed. Next, the process of transmission-based X-ray image generation is described. Thereafter, X-ray radiography and X-ray Computed Tomography (CT), the two most common and mature X-ray imaging techniques, are explained. Finally, X-ray phase contrast imaging, a relatively new X-ray imaging technique that is still far away from practical application, is shortly discussed

2.3.1 X-rays and their interaction with matter

X-rays are, just like visible light, a type of electromagnetic radiation. Electromagnetic radiation are waves that propagate through space and are characterized by their frequency (or wavelength). The higher the frequency, the higher the electromagnetic radiation energy, and vice versa. The wavelength is inversely proportional to the frequency, i.e., a low frequency corresponds to a long wavelength. The electromagnetic spectrum (see Figure 2.3) is divided in different bands of wavelength ranges which are given specific names, e.g., X-rays (0.01 to 10 nm), ultraviolet light (10 to 400 nm), visible light (400 to 700 nm), and infrared light (700 nm to 1 mm). Each band has specific characteristics based on how the radiation is produced, the manner the waves interact with matter, and how they are used in applications. Together with ultraviolet and gamma radiation, X-rays are classified as ionizing radiation because of the high photon energy that can cause the ionization of atoms and subsequent chemical reactions (Kak & Slaney, 2001; Maier et al., 2018).

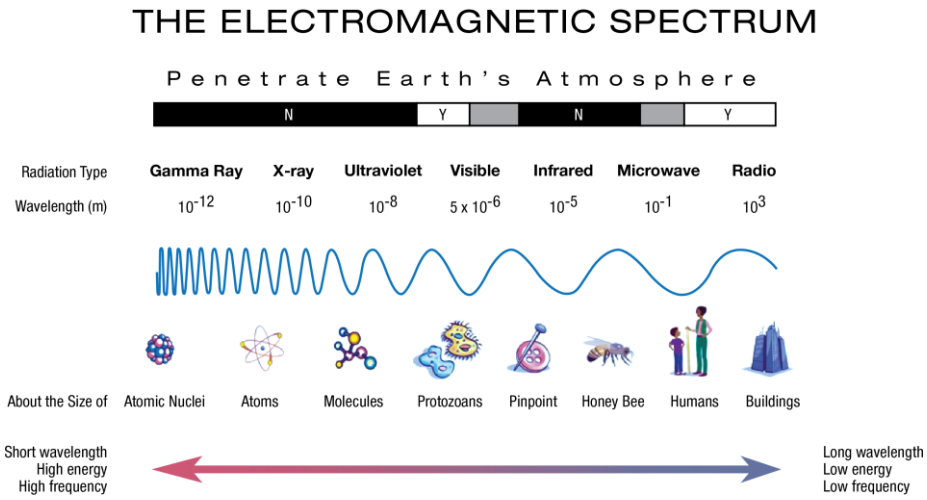

Figure 2.3: The electromagnetic spectrum. Adapted from (My NASA Data, 2021).

Due to their high energy, X-rays can penetrate through material in which they are partly absorbed and scattered via physical interactions. The X-ray beam is thus attenuated while passing through the material. For the energy range of photons most used in diagnostic imaging (20 to 150 keV), the main physical principles responsible for the X-ray attenuation via absorption and scattering are the photoelectric effect and Compton scattering, respectively (see Figure 2.4).

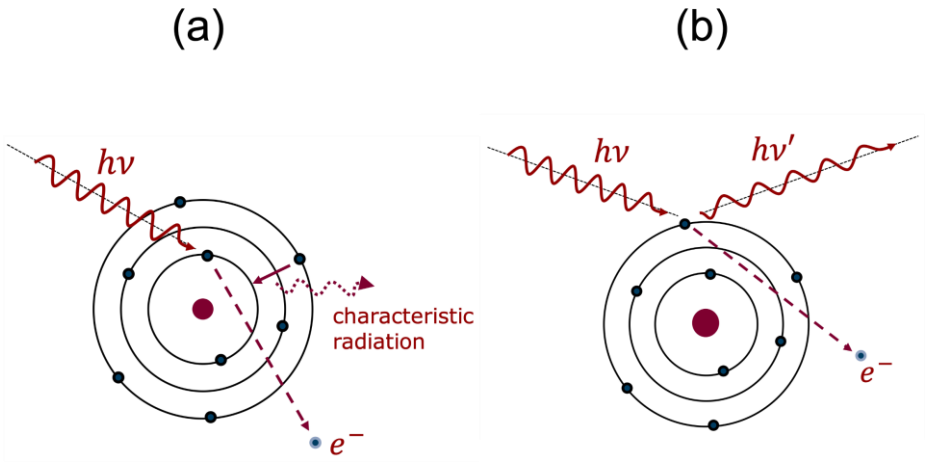

Figure 2.4: Photoelectric absorption (a) and Compton scattering (b) (Sijbers & Jørgensen, 2021).

In the photoelectric effect, an X-ray photon hits an inner electron of an atom. All the energy of the photon is then carried over to the electron, which causes the electron to overcome the binding energy within its shell. The excited electron escapes from its previously tightly bound state, while another electron from a higher energy outer shell takes its place. The latter results in the emission of characteristic radiation that is mostly reabsorbed by the surrounding tissue. While the X-ray photon seizes to exist, the photoelectric effect thus produces a free electron and an ion. Photoelectric absorption depends on the atomic number of the irradiated material and the energy of the incident photon. Compton scattering, on the other hand, is the phenomenon in which a photon hits either a free or a loosely bound outer shell electron. The collision causes the photon to be deflected from its original course into another direction. Herein, part of the photon's energy is carried over to the electron, therefore ionizing the atom. Compton scattering depends on the electron density of the irradiated material (Kak & Slaney, 2001; Maier et al., 2018).

For monochromatic X-rays, the number of X-rays attenuated by a material is described by the Lambert-Beer law, given by

$$I(d) = I_0 e^{-\mu \cdot d}$$

It states that the intensity of incoming radiation (I_0) decays exponentially when it propagates by a distance d through a material with a linear attenuation coefficient μ. The linear attenuation coefficient incorporates the probability of all interactions between the X-ray photons and the material according to the principles described above. It is dependent on the electron density of the material and the energy of the X-ray photons. The electron density of a material, in turn, depends on the atomic number of the material elements and its mass density. While the linear attenuation coefficient is practical to use, it must usually be derived from the mass attenuation coefficient, which is normalized for the mass density and thus includes only atom-dependent effects in function of the radiation energy. For mixed materials, the total mass attenuation coefficient can be calculated as the weighted sum of the individual mass attenuation coefficients proportional to the weights of the

elements. The linear attenuation coefficient can then be calculated by multiplying the mass attenuation coefficient with the density of the material. In practice however, materials are often inhomogeneous in terms of composition. Obtaining linear attenuation coefficients of a material is thus not straightforward. In addition, X-ray radiation is often multispectral, requiring knowledge of the full spectrum of the X-ray source. Since biological tissue is mainly composed of hydrogen and carbon atoms, contrast in X-ray images caused by differences in X-ray attenuation originates mainly from differences in tissue density due to the presence of air pores (Kak & Slaney, 2001; Maier et al., 2018).

2.3.2 X-ray image generation

In X-ray imaging, the X-rays emitted by an X-ray source are produced using a high voltage vacuum tube. Electrons released by the cathode collide with a high velocity onto the anode, i.e., a metal target (e.g., tungsten), resulting in the emission of X-ray photons. The photon energy is limited by the energy of the colliding electrons, which in turn is proportional to the voltage of the tube. The emitted X-rays are generated by two different phenomena, i.e., characteristic X-ray emission and bremsstrahlung (see Figure 2.5). In characteristic X-ray emission, an incident electron can liberate an orbital electron from the inner electron shell of a target atom, provided that the incident electron has sufficient energy to overcome the binding energy of the orbital electron. Thereafter, another orbital electron from a higher energy level can fill the vacancy at the lower energy level. This releases energy that is emitted as an X-ray photon. Since this phenomenon is dependent on the material of the target, it results in typical, or characteristic, discrete frequencies of emitted radiation. In contrast, bremsstrahlung produces a continuous spectrum. In bremsstrahlung, an X-ray photon is released via the deceleration of the incident electron when it is deflected by the electric field of other charged particles, e.g., orbital electrons. The lost kinetic energy is thus released as an X-ray photon (Kak & Slaney, 2001; Maier et al., 2018).

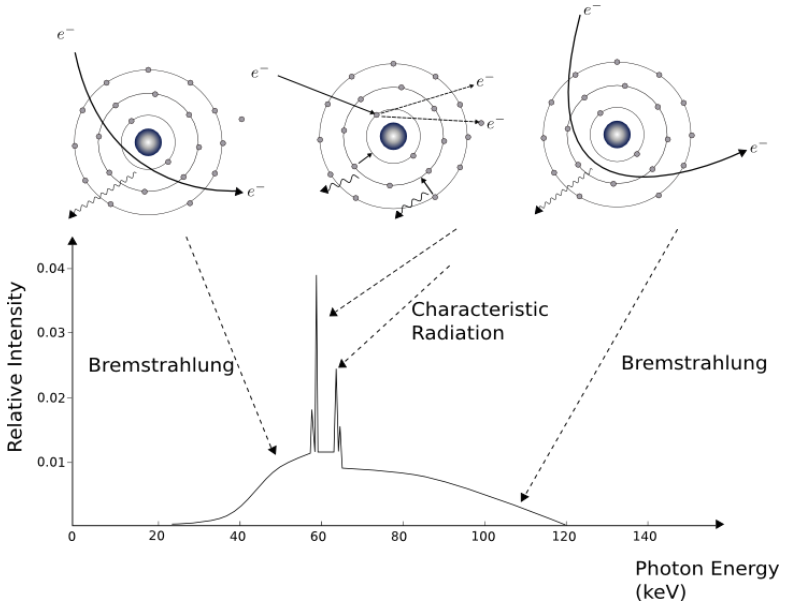

Figure 2.5: X-ray spectrum of a tungsten tube with discrete peaks and the continuous spectrum corresponding to the characteristic radiation and bremsstrahlung, respectively. Adapted from (Maier et al., 2018).

X-ray images are produced from the signal picked up by X-ray sensitive detectors, which in modern systems are flat panel detectors based on semiconductors that convert X-ray photons into an electric signal. Transistors are arranged into a grid, each attached to a light-absorbing photodiode responsible for an individual pixel. An electric signal is created that depends on the intensity of the striking photons that can then be combined to produce a digital image. Typically, a layer of scintillator material is placed before the photodiodes, which first converts X-rays into visible light that is then captured by the photodiodes (Kak & Slaney, 2001; Maier et al., 2018).

2.3.3 X-ray radiography

In X-ray radiography, X-rays transmitted through an object are captured by a detector to create a 2D transmission X-ray image, i.e., a single radiograph or projection. In the radiograph, the intensity scales inversely with the accumulated amount of X-ray attenuation by the sample along the trajectory of the X-rays. Thus, X-ray

radiography does not retain all spatial information and instead produces a cumulative 2D projection of a 3D object (Kak & Slaney, 2001; Maier et al., 2018). The detector can be a 2D array to image the whole sample at once, or a 1D line detector. The latter is often used for inline radiography in which a source and line detector are placed on either side of a conveyor belt. Using line detectors, an arbitrary number of line scans can be stitched together to produce a 2D image. In current inline imaging hardware, detectors with a pixel size of 0.4 mm are readily available for speeds of 55 m/min.

2.3.4 X-ray Computed Tomography

In X-ray CT, many X-ray radiographs, i.e., projections, are taken from different angles and are combined to reconstruct 3D images of the X-ray attenuation by the sample. In micro-CT, i.e., industrial CT, the platform on which the sample is mounted rotates between a source and a detector that are stationary during image acquisition (see Figure 2.6). The position of the detector and sample relative to the source can often be configured to optimize the image quality and field of view. In medical CT, on the other hand, the source and detector rotate around the patient.

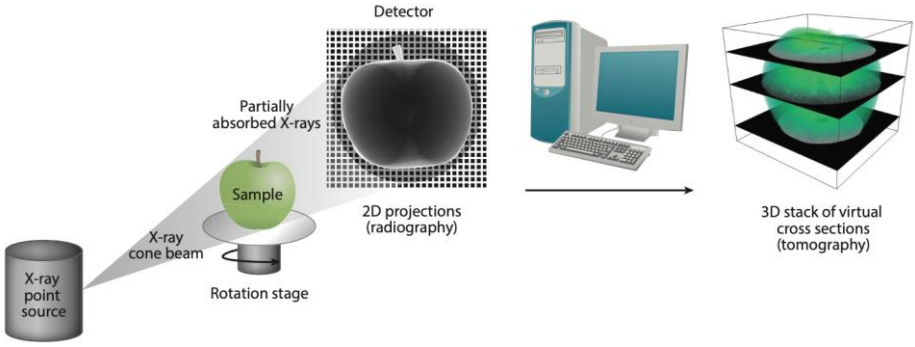

Figure 2.6: The X-ray CT workflow. Multiple 2D projections are taken from different angles after which a 3D reconstruction is computed (Z. Wang, Herremans, et al., 2018).

Here, X-ray CT reconstruction will be described for the 2D case assuming parallel X-ray beams, which can be extended to 3D reconstruction using a cone beam (Kak & Slaney, 2001; Maier et al., 2018). In X-ray CT, the goal is to reconstruct the object function $f(x, y)$, which describes the linear attenuation coefficients at positions (x, y). The image formation is based on two mathematical principles, i.e., the Radon transform and the Fourier slice theorem (Kak & Slaney, 2001; Maier et al., 2018). The Radon transform principle states that any integrable function $f(x, y)$ can be uniquely described as all straight-line integrals over its domain (Maier et al., 2018), i.e.,

$$p(l) = \int_{-\infty}^{+\infty} f(x(l), y(l))dl, \quad \forall l: \left(x(l), y(l)\right)^T \in line\ l \quad \text{(Eq. 1)}$$

In the context of X-ray CT, this means that a slice through the sample volume, described by $f(x, y)$, can be formulated in function of projections following straight lines through the sample. Eq. 1 can be formulated in polar coordinates to prevent duplicate representations of the lines, i.e.,

$$p(\theta, s) = \int \int_{-\infty}^{+\infty} f(x, y)\delta(x\cos(\theta) + y\sin(\theta) - s)dxdy \quad \text{(Eq. 2)}$$

With θ the angle between the x-axis and the normal vector of the line, and s the orthogonal distance between the line and the origin. Herein, only points on the line, i.e., satisfying $x\cos(\theta) + y\sin(\theta) = s$, are selected using the Dirac function δ. The complete set of line integrals required to describe $f(x, y)$ can then be obtained by covering the angles $\theta \in [0, \pi]$ and orthogonal distances $s \in [-\infty, +\infty]$. A single projection is obtained for every fixed angle θ and variable distance to the origin s, i.e., $p_\theta(s) = p(\theta, s)$. All projections can then be arranged side-by-side to produce a 2-D image, i.e., a sinogram, which describes the projected values as a function of θ and s. In the sinogram, every point except for the origin is found at different distances along the s-axis depending on the angle θ (Kak & Slaney, 2001; Maier et al., 2018).

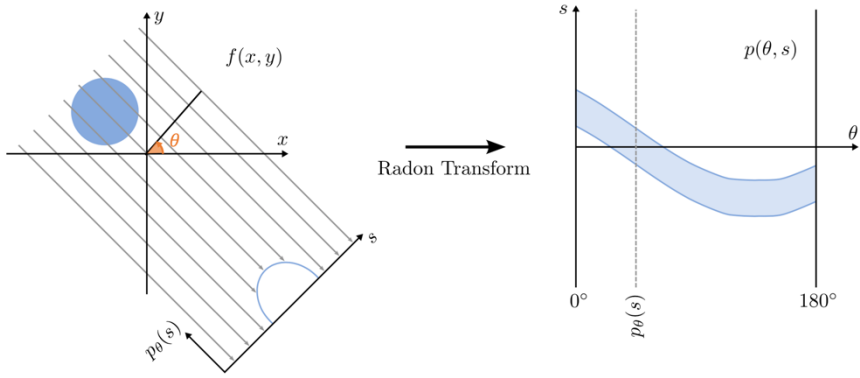

Figure 2.7: The Radon transform. Projections as a function of the angle θ and distance s to the origin (left), and the resulting sinogram (right). Adapted from (Maier et al., 2018).

In X-ray CT, the goal is to compute the inverse Radon transform, i.e., reconstructing $f(x, y)$ from a set of measured projections. Herein, the values of $f(x, y)$ are the linear attenuation coefficients of the imaged sample. To compute the inverse Radon transform, the Fourier slice theorem is used. The theorem states that there is an equivalence between the Fourier transform $P(\xi, \theta)$ of the projection $p_\theta(s)$ and the line through the 2D Fourier transform $F(u, v)$ of $f(x, y)$ following an angle θ through the origin in the 2D Fourier domain. Therefore, a good estimate of $F(u, v)$ can be obtained from many of such lines by using a complete set of projections. The function $f(x, y)$ is then obtained by calculating the inverse 2D Fourier transform of $F(u, v)$. In practice, the projections are back-projected along their corresponding lines to compute $f(x, y)$. To overcome the oversampling of the center of the Fourier domain, a filter is used to enhance and dampen the high and low frequencies, respectively. This reconstruction technique is called the Filtered Back-Projection (FBP) algorithm, which is the most used analytical reconstruction algorithm in X-ray CT. The whole 3D volume can be reconstructed by stacking the computed slices (Kak & Slaney, 2001; Maier et al., 2018).

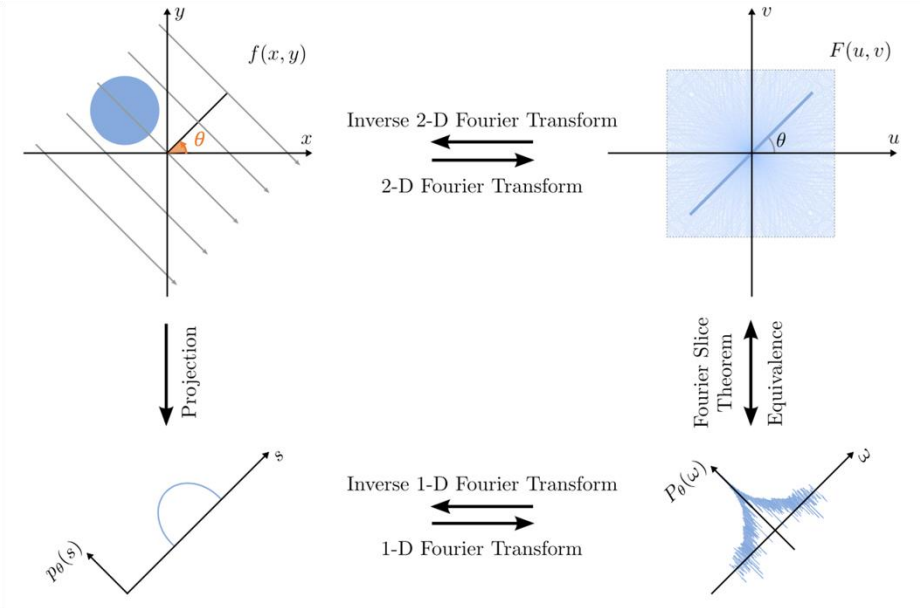

Figure 2.8: Equivalence between the Fourier transform $P(\xi, \theta)$ of the projection $p_\theta(s)$ and the line through the 2D Fourier transform $F(u, v)$ of $f(x, y)$ following an angle θ through the origin in the 2D Fourier domain. Adapted from (Maier et al., 2018).

Alternatively, algebraic reconstruction algorithms and statistical methods can be used that compute the reconstruction iteratively. These techniques can leverage prior knowledge of the acquisition geometry and composition of the sample to produce images with improved resolution, signal-to-noise ratio, and reduction of artefacts. Even though these methods have a high computational cost, contemporary hardware has sufficient performance to implement them (e.g., using parallelization on GPU's) (Beister et al., 2012; Kak & Slaney, 2001; Maier et al., 2018; Palenstijn et al., 2011; Willemink et al., 2013).

Using X-ray CT, 3D images of intact fruit can be produced at a high resolution (Piovesan et al., 2021). However, there is a trade-off between image quality and resolution on the one hand, and speed and equipment cost on the other hand. Due to the costly hardware requirements and insufficient speed, the usage of X-ray CT in industrial applications is currently limited to the inspection of high value products at a relatively slow rate, or to its usage in research and development stages (Buratti et al., 2018; Wevers et al., 2018).

However, recent developments in reconstruction algorithms and hardware have enabled X-ray CT measurements inline. For instance in translational X-ray CT, a sample is rotated and translated simultaneously (De Schryver et al., 2016; Janssens et al., 2016, 2018, 2019; L. F. A. Pereira et al., 2016, 2017). Alternatively, gantry systems, similar to medical CT systems, can be developed for inline industrial CT. The first inline 3D X-ray CT systems are already offered commercially, e.g., "Mito" by BIOMETIC targeted at the food industry (www.biometic.com). In addition, reconstruction algorithms are being developed that incorporate deep learning to reduce the number of projections required to produce a high-quality image and thus to reduce the required acquisition time. However, these systems are not yet suited for sorting of high volumes of low value products.

2.4 Detection of internal disorders in horticultural products

Several instrumental techniques for nondestructive internal disorder detection have been proposed in literature, which are mainly based on visible and near-infrared (Vis-NIR) spectroscopy, MRI, X-ray radiography or X-ray CT, (Arendse et al., 2018; Lu & Lu, 2017; Nicolaï et al., 2014; Piovesan et al., 2021; Srivastava et al., 2018; Z. Wang, Herremans, et al., 2018). In this section, the most studied techniques for internal disorder detection are discussed, with a focus on the application to pome fruit.

2.4.1 Visible and near-infrared (Vis-NIR) spectroscopy

In Vis-NIR spectroscopy, a sample is irradiated with visible (400–750 nm) or near-infrared (750–2500 nm) light which is absorbed and scattered by the material. Because of the limited absorption of Vis-NIR radiation by water, radiation can penetrate up to a few centimeters into biological tissue, depending on the wavelength. Light scattering causes the radiation to diffuse in the sample volume and to be reemitted at the tissue boundaries. The spectrum of the interacted light is measured by a wavelength-sensitive detector, which can then be analyzed (Nicolaï et al., 2007, 2014). It is mainly applied in three modes, i.e., reflectance, interactance and

transmission. In reflectance mode, the spectrum of the reflected light from the surface is measured. However, due to the limited penetration depth of the used wavelengths, it can only be used to spot (sub)surface defects, e.g., bruises. To detect deeper internal defects, transmission is therefore required. However, for large fruit, such as apple and pear, the full transmission of Vis-NIR light is limited. Therefore, as a middle ground, interactance mode is often used to analyze intact fruit. Herein, the light that was partly transmitted through the sample and that was then reemitted via scattering is measured (Lu & Lu, 2017; Walsh et al., 2020). In general, the main downsides of Vis-NIR spectroscopy for internal disorder detection are the lack of spatial information and the difficultly of interpreting the measured spectrum, which depends on i.a., the fruit size, fruit shape, the cultivar, seasonal variation, and temperature. Therefore, large amounts of data are required to perform calibration. Moreover, since internal disorders are not always uniformly distributed in the fruit, long exposure times, a fixed orientation, or multiple measurements from different positions are required (Bobelyn et al., 2010; Lu & Lu, 2017; Nicolaï et al., 2007, 2014; Pasquini, 2003).

Vis-NIR spectroscopy has been used successfully to detect internal disorders in research. Han et al. (2006) reported false positive and negative rates of respectively 4.3 and 5.3 % for brown core detection in 'Yali' pears using transmission Vis-NIR spectroscopy at three different locations per sample with precisely aligned pears. Khatiwada et al. (2016) used transmission Vis-NIR spectroscopy to detect internal browning in 'Pink LadyTM' apples with four spectra acquired of each fruit and reported an accuracy of more than 95 % for classifying acceptable and unacceptable fruit. Additionally, a coefficient of determination (R^2) of 0.83 was found when predicting internal browning severity on a five-point scale. Similarly, Guo et al. (2020) reported an R^2 of up to 0.91 for predicting watercore severity as the percentage of affected area on an equatorial transverse cut in 'Fiji' apples using the average of three transmission spectra by rotating the fruit 120° between measurements. Huang et al. (2020) investigated transmission Vis-NIR spectroscopy to detect internal defects in 'Honeycrisp' apples by acquiring spectra from six different

locations per fruit. Using the mean spectrum of the six measurements, a classification accuracy of up to 93.1 % was reported. However, for fruit with a defective tissue area below 40 % of the total fruit tissue area being rated, the accuracy dropped to 77.3 %.

2.4.2 Magnetic Resonance Imaging (MRI)

MRI is based on the interaction of the nuclear spin with an external magnetic field and radio waves and mainly provides information on the density and mobility of hydrogen nuclei. Different compositions and structures of sample tissue affect the extent to which the protons can interact with the magnetic field and the radio waves. By applying a spatially changing magnetic field across the sample, a signal with spatially varying frequency components is produced. In addition, radio waves are used to cause changes in the directions of the nuclear spin of the protons. In-between pulses of radio waves, the relaxations of the nuclear spins to an equilibrium state along the main magnetic field are recorded. The spatially and temporally varying signals can then be used to reconstruct a multimodal 3D image of the sample. Commonly, three measured tissue properties are visualized, i.e., proton density, the longitudinal relaxation time (T_1) and the transverse relaxation time (T_2) (Brown et al., 2014).

MRI has been used for quality evaluation of fruit, including internal disorder detection (Srivastava et al., 2018). Characterization and detection of bruises was investigated using MRI in apple (McCarthy et al., 1995; Zion et al., 1995) and pear (Razavi et al., 2018). MRI was also used to investigate watercore in apple (Clark et al., 1998; Clark & Richardson, 1999; Herremans, Melado-Herreros, et al., 2014; Melado-Herreros et al., 2013; S. Y. Wang et al., 1988). Internal browning was studied using MRI in apple (Clark & Burmeister, 1999; Defraeye et al., 2013; Gonzalez et al., 2001) and pear (Hernández-Sánchez et al., 2007; Lammertyn et al., 2003b, 2003a; Suchanek et al., 2017; C. Y. Wang & Wang, 1989). MRI has also been used for the detection of cavities in watermelons (Saito et al., 1996) and heat treatment injury in mango (Joyce et al., 1993).

For MRI, the remaining concerns are low image acquisition speed due to physical constraints, the need for a sufficiently powerful and

homogenous magnetic field, high equipment costs, electromagnetic inference and motion artefacts (Brown et al., 2014; Colnago et al., 2014; Nicolaï et al., 2014; Srivastava et al., 2018).

2.4.3 X-ray radiography

X-ray radiography was explored for detecting hollow heart in potato (Finney & Norris, 1973), and internal disorders and insect infestation in mango (Thomas et al., 1993, 1995). Casasent et al. (1998) used X-ray radiography to detect damaged pistachio and reported a classification accuracy of 88 %. Similarly, Kim & Schatzki et al. (2001) tested X-ray radiography to detect damaged almond nuts. A classification accuracy of up to 81 % was achieved. A method for detecting insect damage in wheat kernels was proposed by Karunakaran et al. (2004), who reported a classification accuracy of up to 86 %. Narvankar et al. (2009) developed a method for fungal infection detection achieving true negative and true positive rates of respectively 83 and 89-93 % (depending on the fungus). Kotwaliwale et al. (2007) investigated the use of X-ray radiography for the detection of damaged pecan nuts and reported a true positive and true negative rate of 76 and 100 %, respectively.

The usage of X-ray radiography to detect defects in onions was reported by Tollner et al. (2005) with classification accuracies above 90 %. Haff et al. (2006) showed that the use of X-ray radiography has a high potential for detecting translucency in pineapples. Jiang et al. (2008) proposed the use of X-ray radiography to detect and segment insect damage in guava and peach fruit, reporting detection accuracies for infestation sites of 93 and 96 %, respectively. X-ray radiography was used to detect granulation and endoxerosis in oranges and lemons, respectively, with corresponding classification accuracies of 96 and 94 % (van Dael et al., 2016).

For apple, Shahin et al. (1999, 2001) proposed a sorting system for detecting watercore using X-ray radiography and reported accuracies up to 88 % for classifying healthy, mildly affected and severely affected fruit, which were homogeneous in size. Kim & Schatzki et al. (2000) proposed a method for watercore detection in apple and achieved an overall accuracy of 60 % for classifying healthy, mildly affected and severely affected fruit on a larger

dataset with fruit of various sizes. The system correctly classified apples into the healthy and severe categories with false positive and negative rates in the range of 5 to 8 %. X-ray radiography was used to detect bruises in apple, achieving classification accuracies of up to 93 and 60 % for old and new bruises, respectively (M. Shahin et al., 2002). The potential of X-ray radiography for insect damage detection in apple was explored by Hansen et al. (2005). For the detection of mould core in apple using X-ray radiography, an accuracy of 95 % was reported by Yang et al. (2011).

The X-ray radiography-based methods above mainly use disorder and application specific algorithms, e.g., dedicated feature extraction algorithms. This complicates their robustness and transferability to other biological products with considerable differences in shape, size, and composition. In addition, the contrast in the radiograph may suffer from effects of fruit shape, volume, and internal structure such that internal defects can be less prominent in the image when they are, e.g., shadowed by the core of the fruit. Therefore, a more general purpose multisensor algorithm was developed that combines prior knowledge in the form of shape and density distribution models with X-ray radiography (van Dael et al., 2019, 2017). For instance, for internal disorder detection in pear, the method obtained true positive and true negative rates of 97 and 90 %, respectively, compared to a dedicated reference method that obtained true positive and negative rates of 98 and 84 %, respectively (van Dael et al., 2017). The downside of the method is, however, that it still requires product specific 3D models of the shape and density distribution, and the complex integration of multiple sensors.

2.4.4 X-ray Computed Tomography (CT)

In terms of internal disorders, X-ray CT based methods have mainly been used to characterize them rather than to detect them automatically (Cantre et al., 2017; Diels et al., 2017; Herremans et al., 2013; Herremans, Melado-Herreros, et al., 2014; Herremans, Verboven, et al., 2014; Lammertyn et al., 2003b, 2003a; Mazhar et al., 2015; Muziri et al., 2016; Orina et al., 2017; Si & Sankaran, 2016). Most characterization methods were rather slow and involved semi-

automated steps or manual annotation that does not scale to large datasets. Research was done towards automated internal disorder detection using X-ray CT. A method for the detection of internal decay in chestnuts was proposed by Donis-González et al. (2014), reporting accuracies of 85.9 %, 91.2 % and 96.1 % for classifying in five, three and two classes, respectively. Herremans, Melado-Herreros, et al. (2014) proposed a method for watercore detection in apple, achieving a classification accuracy of up to 89 %. Jarolmasjed et al. (2016) developed a method for bitter pit detection in 'Honeycrisp' apples, reporting accuracies of 70 and 96 % for two different populations.

A few studies compared X-ray CT to MRI for internal disorder detection. Lammertyn et al. (2003a) compared X-ray CT to MRI for the detection of internal browning during its development in 'Conference' pears. They found that both techniques were reliable. However, incipient browning was harder to detect using X-ray CT due to the physical principle according to which the technique operates. In incipient browning, the cellular liquid of death cells has not yet diffused away, resulting in no change in density and X-ray attenuation. On the other hand, changes in proton mobility can already be observed earlier using MRI. For watercore detection in apple, it was found that with their method, better classification of heathy and affected fruit was achieved using X-ray CT, even though MRI provided better contrast between healthy and affected tissue (Herremans, Melado-Herreros, et al., 2014).

2.5 Automated image interpretation

X-ray images provide rich spatial and structural information on the imaged sample. To use X-ray imaging as a tool for internal quality inspection of foods at a high-throughput, automated image processing and interpretation is required. Computer algorithms must, therefore, be developed that "understand" the content of the images to solve various tasks, e.g., classification of healthy and defect fruit, or the segmentation of regions affected by internal browning. In this section, the field of image interpretation is, therefore, discussed in the context of quality inspection. First, the necessary background is provided on how digital images are presented to a

computer (section 2.5.1) and the general challenge in computer vision is described (section 2.5.2). Next, conventional machine learning techniques for interpreting images are discussed (section 2.5.3). Finally, the current paradigm in machine learning, i.e., deep learning, is discussed in the context of images (section 2.5.4). This section is by no means a comprehensive review on deep learning, but introduces the main techniques used in this thesis.

2.5.1 Digital images

Digital images are indispensable tools for data analysis in many fields, e.g., field monitoring, plant phenotyping, medical diagnosis, and nondestructive inspection in manufacturing. They enable the acquisition of rich spatial and spectral data about the imaged object in a standardized way. Digital images comprise a regular grid of picture elements, i.e., pixels, that are the smallest addressable elements in an image. Each pixel is characterized by its position on the image, e.g., x- and y-coordinates, and its intensity. The intensity is typically represented in gray scale, going from black (minimal intensity) to white (maximal intensity). X-ray radiographs, for instance, are gray scale images in which the intensity corresponds to the attenuation of X-rays (see Figure 2.9). Images can comprise multiple channels, i.e., multiple layers of pixels that are superimposed over each other. For instance, a color image typically has a red, a green and a blue (RGB) channel that together contribute to the color of every pixel. Images can also be extended to 3D, e.g., CT volumes, in which the smallest addressable element is called a voxel. Volumetric images can be interpreted as a stack of 2D images, or slices, in which the distance between the slices is defined by the 3rd dimension of the voxel size (i.e., the resolution of the CT image). Volumetric images can also have different channels to represent different image modalities of the same sample, e.g., proton density and T1 and T2-relaxation times in MRI.

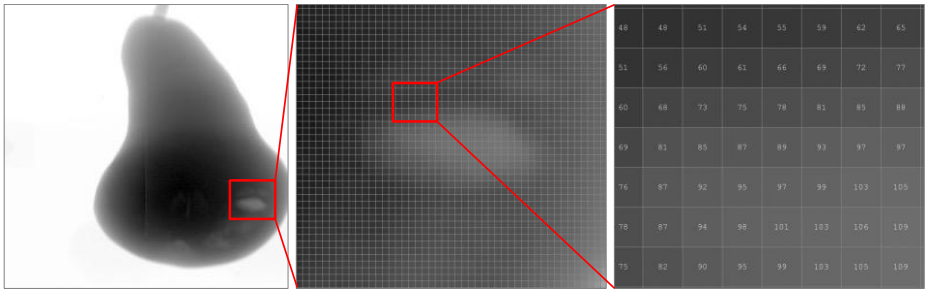

Figure 2.9: Pixels in a digital image illustrated on an X-ray radiograph of a pear with internal disorders. The region indicated in a red box is zoomed in on when going to the left. A digital image is a matrix in which the values represent the color intensities.

2.5.2 Computer vision

Visual perception is the most important mechanism by which humans understand their surroundings. Human visual perception is thus also excellent at analyzing images. However, human interpretation does not scale to a high-throughput or large datasets of images. In addition, it can be subjective and inconsistent over time. Therefore, automated image analysis is strived for in high-throughput, time sensitive or critical applications. While humans can interpret a rendered image easily, to a computer, digital images are nothing more than a grid of numbers, i.e., matrices. It has no knowledge of what the image represents, nor does it understand its content.

With the emergence of computers and digital images, the field of computer vision emerged that deals with exactly how computers can obtain high-level understanding of images and videos (Szeliski, 2010). In the context of quality inspection, typical computer vision tasks are image classification (e.g., sound *vs* defect samples) and segmentation (e.g., identifying the defect region).

To make a computer understand or interpret an image, one could write an algorithm with explicit instructions that must be followed step-by-step by the computer to reach a final decision. However, designing a robust algorithm with fixed rules is extremely challenging for a large dataset of images considering the often large variability between samples, e.g., biological variability. Images can be interpreted as high-dimensional samples, i.e., each pixel or voxel

in the image is a single variable. A low-resolution image of 32 × 32 pixels already contains 1024 variables. Just moving the object in an image slightly in any direction changes the values of many pixels. Interpreting images directly is, therefore, a challenging task. Therefore, huge effort has been made in developing machine learning (ML) techniques.

2.5.3 Machine learning

In ML, an algorithm explores the statistics of a training dataset to identify patterns that are needed to solve a certain task (Bishop, 2006). It enables machines to perform a task by learning from data ("experience"), without being explicitly programmed how to do so. In general, learning can be implemented in three ways, i.e., supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, models are trained using labeled data. Herein, learning occurs by providing direct feedback to the model by optimizing its predictions to match the ground truth labels on a training set. The idea is that a trained model can then be used to provide accurate predictions on unseen data. An example of supervised learning is classification in which the correct answers are available for training.

Unsupervised learning, on the other hand, requires the model to discover the underlying structure in unlabeled data. Finding the underlying structure can be a goal on its own, e.g., clustering, in which samples must be assigned to several groups (clusters). Alternatively, unsupervised learning can be used as an intermediate step to provide useful representations of the data for other tasks. Finally, in reinforcement learning, a model, which in this case is referred to as an agent, learns from trial and error by receiving feedback from operating in an environment without relying on other instructions. In practice, supervised learning is by far the most used ML approach (Bishop, 2006; Goodfellow et al., 2016).

Semi-supervised and self-supervised learning are two special cases of unsupervised learning. Semi-supervised learning assumes that most data is unlabeled, while a small subset of labeled data is available. An example of a semi-supervised learning strategy is to first train in an unsupervised way. Thereafter, the discovered

underlying structure can be used to more efficiently train the model to perform a specific task using the small subset of labeled data. In self-supervised learning, random transformations are applied on unlabeled data. This transformation, which is unknown to the model, must then be estimated or reversed by the model. The idea is that if a model can do this, it must understand certain underlying structures in the data. Again, the learned underlying structure can then be used for clustering or other downstream tasks.

Typically, a dataset is divided in a training, validation, and test dataset. The training dataset is used to train the ML algorithm, while the validation dataset is used for evaluating the sensitivity towards hyperparameters, i.e., parameters that are not learned but that must be set by the operator. Finally, the trained model is tested on the left-out test dataset to gauge its performance on unseen data. The features are often normalized, i.e., centered around the origin and scaled to variance equal to one, to make the features independent of their scale (Bishop, 2006; Goodfellow et al., 2016).

In ML, the data is typically represented as a matrix with samples and variables (or features) as rows and columns, respectively. Each sample is thus represented as a vector, i.e., a feature vector. Therefore, to apply machine learning on images, features must be extracted using image analysis and processing techniques (Bishop, 2006; Russ, 2006; Szeliski, 2010). Image processing algorithms can contain operations at various levels, e.g., low-level features (edge, corner, or blob detection), shape-based features (thresholding, morphological operations, template matching), or whole image analysis (pixel counting, histograms, clustering). Every sample has its own vector of features, which corresponds to a point in feature space. Thereafter, the features are fed to a ML algorithm for learning the task at hand, e.g., classification. For instance, van Dael et al. (2016) developed an algorithm to extract features from X-ray radiographs of oranges to differentiate heathy and defect fruit (see Figure 2.10). The features, i.e., the area, perimeter and solidity of the unaffected endocarp, were used to train a *k*-nearest neighbor (kNN) algorithm for classification. kNN uses the majority vote on the classes of the *k* closest training samples in feature space to predict the class of a new sample. Various other ML algorithms exist for

classification, e.g., Support Vector Machines (SVM), Bayesian classifiers, decision trees and random forests, and for other tasks, e.g., regression (Bishop, 2006).

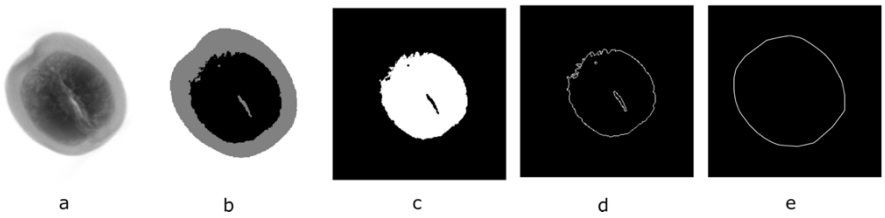

Figure 2.10: Feature extraction on an X-ray radiograph of an orange. Original (a) and segmented (b) image with the extracted features on the unaffected endocarp: area (c), perimeter (d) and solidity (e), calculated as the ratio of the white area in (c) to the area encapsulated by its convex hull in (e). The image in (b) was obtained by applying multiple Otsu-thresholds on image (a) (Otsu, 1979). Adapted from (van Dael et al., 2016)

In the unsupervised case, classification comes down to clustering, e.g., the *k*-means clustering algorithm. Using some similarity measure in feature space, *k*-means clustering iteratively assigns samples to *k* clusters to maximize the similarity inside and minimize the similarity between clusters. Clustering can be applied inside a single image to segment different regions based on their pixel values and coordinates (Pal & Pal, 1993). For instance, the Otsu-threshold separates all pixels in two clusters based on the histogram of pixel values and is often applied to separate the foreground and background (Otsu, 1979). *k*-Means clustering can be used to cluster pixels in *k* clusters corresponding to different regions of interest (Bishop, 2006). Region growing is a semi-supervised region-based image segmentation algorithm. Starting from a given initial seed point, neighboring pixels are added iteratively to the region based on a similarity measure between all pixels already added to the region. Multiple seed points can be used for the same or different clusters (Pal & Pal, 1993).

ML algorithms have proven to be effective in solving various tasks automatically. However, the necessary step of feature extraction has several downsides. First, features are often not transferable to other applications. Relevant features must thus be engineered for every

new application. Second, feature engineering requires expert knowledge and thus comes at a high cost. Third, even for experts finding useful features is labor-intensive and it requires a lot of trial and error. Finally, feature engineering potentially results suboptimal solutions that are biased and limited by the human capability of interpreting the data. Strategies have been developed to partly overcome this issue. For instance, instead of deciding on what features should be extracted from the image, random pattern, or texture, occurrences can be used to create feature vectors of images, e.g., histograms of oriented gradients or local binary patterns (Dalal & Triggs, 2005; Pietikäinen et al., 2011). These methods have a good performance due to their high discriminative power, computational simplicity, and invariance to grayscale changes (e.g., by illumination variations). However, since these methods only provide an unordered statistical distribution of local low-level features, the spatial and conceptual information that can be captured is limited.

Therefore, in deep learning, the current paradigm of machine learning, the intermediate step of feature extraction is removed altogether. Instead, in deep learning models learn to solve a task end-to-end from "raw" data. In the learning process, the model itself decides which features should be extracted from to the data to map the inputs to the target outputs.

2.5.4 Deep learning

In deep learning, artificial neural networks (ANN) are used to learn the mapping from the input to the target output "end-to-end". In classical neural networks, i.e., feedforward neural networks, information flows in one direction from the input to the output through a network of neurons that are connected in layers (see Figure 2.11). The neurons, also called nodes, are connected by weights that must be learned. In this context, the term "deep" refers to usage of multiple layers in the model architecture. A model's depth is thus the number of layers it contains.

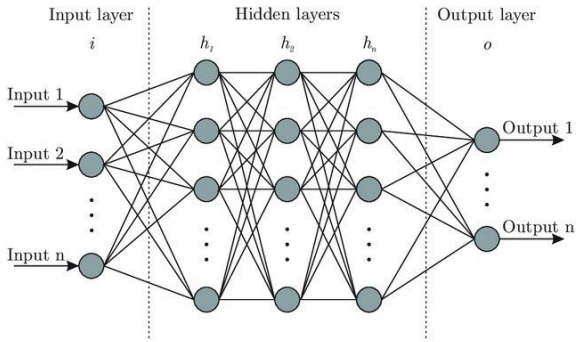

Figure 2.11: An artificial neural network.

2.5.4.1 Artificial Neural Networks

ANNs, originally inspired by the functioning of the brain, were already introduced in the 1960s, but only became state-of-the-art in many ML fields during the last decade. Mainly due to an increase in computation power and the availability of large datasets, deep ANN could be trained at the scale required to do better than other methods. Their dominance in performance on many benchmarks, e.g., AlexNet winning the ImageNet image classification challenge in 2012 (Krizhevsky et al., 2012), was the spark to reignite research on ANN-based methods in many fields, including agriculture, medicine, natural language and transportation.

The single layer perceptron is the simplest neural network (Rosenblatt, 1958). It only contains a single layer of output node(s) and each output is calculated directly as a weighted sum of the inputs. By applying a threshold or sigmoid function on the outputs, it can be used as a classifier. Not long afterwards, the idea of the multi-layer perceptron (MLP) was introduced that has at least one hidden layer, i.e., a layer between the input and output nodes (Rosenblatt, 1961). In contrast to the perceptron, which uses a threshold activation on the outputs, MLPs use in addition nonlinear activation functions (see Figure 2.12) on all nodes except for the input nodes. Additionally, each node can have a bias term. The use of such activations results in a model that can solve nonlinear problems. Since each neuron in a layer is connected to every neuron in the neighboring layers, MLPs are also called fully connected neural networks.

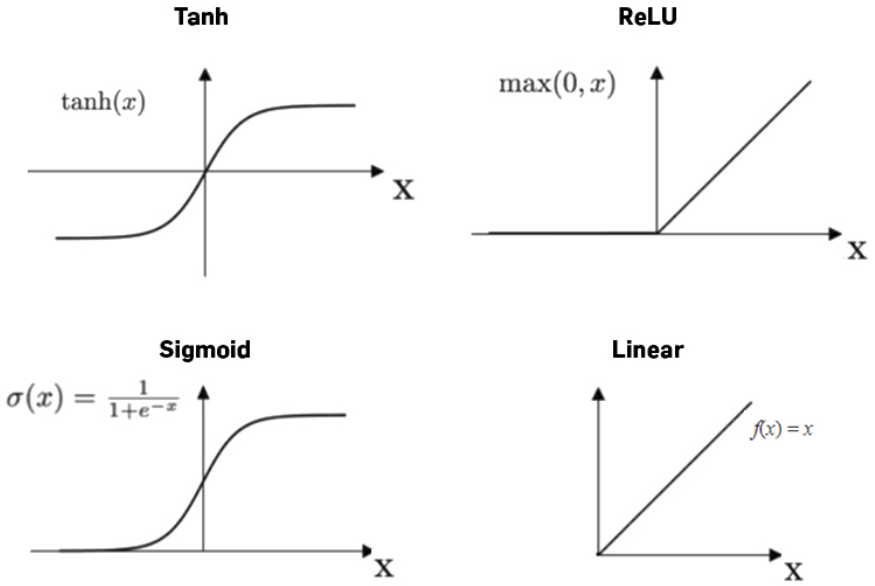

Figure 2.12: Activation functions. Adapted from (Activation Function, 2019).

Zoomed in on a single neuron, its activation is calculated using two operations, i.e., a weighted sum of the activations in the previous layer and a nonlinear function. On the level of whole layers, the activations of $layer_i$ results from a matrix multiplication between a weights matrix and the vector of activations of $layer_{i-1}$ followed by an element-wise nonlinear function. In the context of classification, Figure 2.13 shows the learned representations of a single-layer perceptron and an MLP for linearly inseparable curves in the input space. Using only linear operations, the perceptron is unable to separate the classes. The MLP, on the other hand, can map the input space to a representation in which the curves are linearly separable.

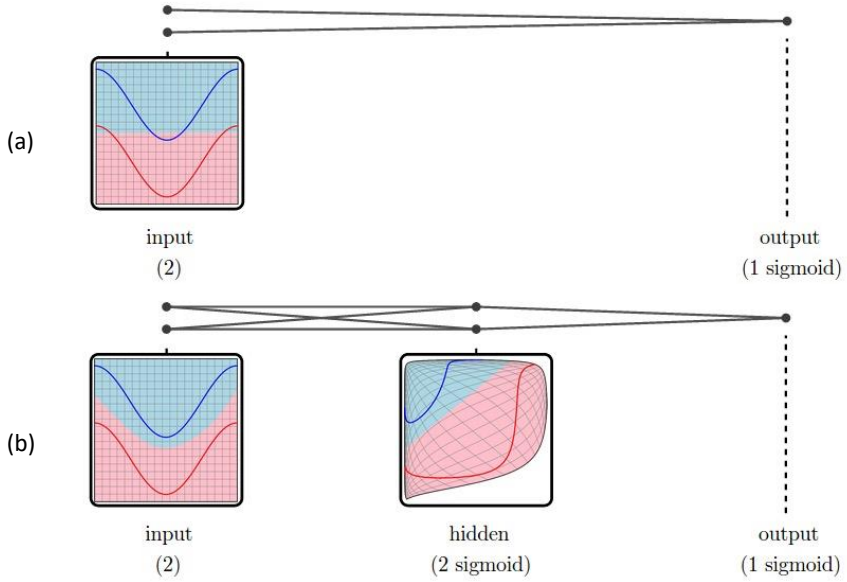

Figure 2.13: Visualization of learned representations in a single-layer perceptron (a) and a MLP (b) for linearly inseparable curves in the input space. The blue and red curves represent data points of two different classes. The blue and red regions in the images indicate regions in the input and feature space that are separated by the classifier. In (a), the perceptron can only construct a line in the input space. Since the two classes are not linearly separable in the input space, some data points are assigned to the wrong class. In (b), a nonlinear transformation is first applied on the input variables. Thereafter, both classes become linearly separable in the transformed feature space. Adapted from (Olah, 2015).

2.5.4.2 Learning by back-propagating errors through the network

To perform a task, ANNs must find the optimal weights that define the inner connections such that their prediction corresponds to the target output. The weights in an ANN are optimized via learning using the backpropagation mechanism (Rumelhart et al., 1986). After randomly initializing an ANN, an error is computed on its output using some objective function. In supervised learning, the objective function measures the similarity between the predictions and the target output. The error, also called loss, is then used to update the network's weights. The optimization happens iteratively using gradient descent, i.e., the weights are updated in the direction that causes the loss to decrease. The partial derivative of the loss

with respect to each weight must thus be computed, which is done using the chain rule. The network's weights θ are then iteratively updated according to

$$\theta \leftarrow \theta \pm \eta \nabla L(\theta) \qquad \text{(Eq. 3)}$$

in which $\nabla L(\theta)$ is the gradient of the loss to each weight and η is the learning rate. η is a hyperparameter that determines the size of the updates. A too small learning rate causes slow convergence, while a too large learning rate can lead to only finding local minima. Gradient estimates are generally done on batches, i.e., random samples of the dataset. This is called stochastic gradient descent and results in a lower chance of getting stuck in local minima. Training is mostly done on several loops, i.e., epochs, over the whole training dataset.

2.5.4.3 Deep learning on 2D and 3D images

To apply fully connected networks on images, the images are flattened into a long vector. However, this vector is very high-dimensional. A large number of weights are thus required to implement these networks directly, resulting in a complex model that requires more time and data to train, and which might easily overfit. Therefore, various weight sharing ideas have been proposed of which convolutional neural networks (CNN) are by far the most effective. In a CNN, the convolutional layers are used in which images are convoluted with kernels, i.e., filters. For a 2D input image with C channels, and H and W pixels in height and width, respectively, each filter has a size of $C \times K \times K$, in which K, a hyperparameter, is the kernel size. Therefore, instead of having each node in the layer being connected to all nodes in the previous one, each node has a small receptive field. Convolutions thus require less weights by limiting the receptive of each node and sharing the weights between nodes of the same layers. An additional benefit of convolutions is that they are invariant to translations in the input image. Multiple filters can be used in the same convolutional layer, which determines the number of channels in the output of the layer. In CNN, the outputs of the layers are often called feature maps.

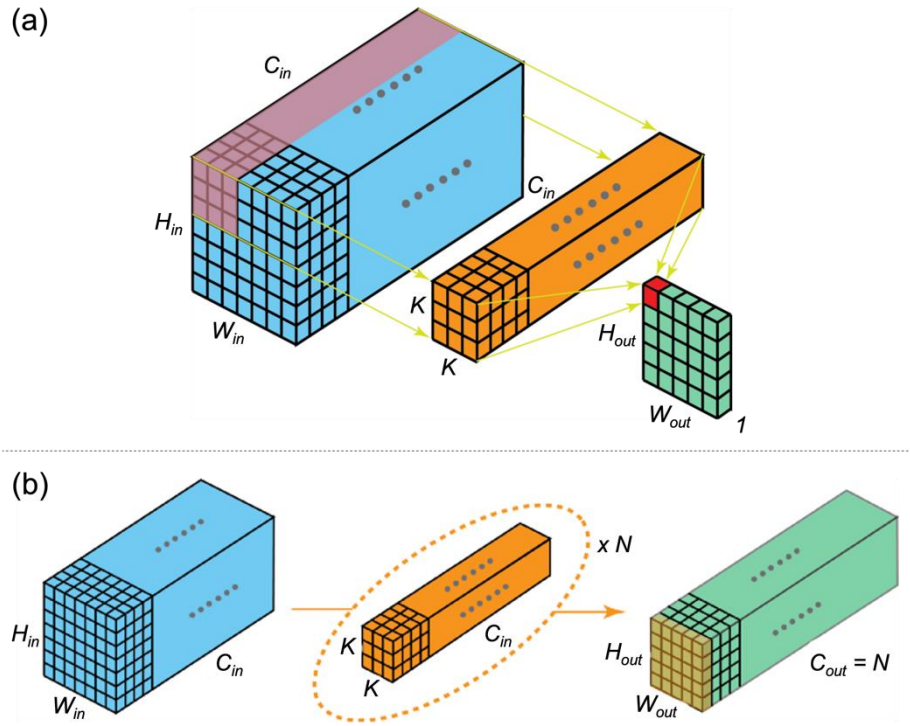

Figure 2.14: 2D convolution of a $H_{\text{in}} \times W_{\text{in}}$ input image that has C_{in} channels. (a) Convolution of the $C_{\text{in}} \times H_{\text{in}} \times W_{\text{in}}$ input image with a single $C_{\text{in}} \times K \times K$ filter resulting in a $1 \times H_{\text{out}} \times W_{\text{out}}$ feature map. The filter slides over the input image along the height and width dimension. At each position, the convolution operates on a $C_{\text{in}} \times K \times K$ region of the input (indicated in red on the left) and produces a single value in the output (indicated in red on the right). (b) Convolution of the $C_{\text{in}} \times H_{\text{in}} \times W_{\text{in}}$ input image with N filters of $C_{\text{in}} \times K \times K$, resulting in a $C_{\text{out}} \times H_{\text{out}} \times W_{\text{out}}$ output. Herein, $C_{\text{out}} = \text{N}$ and H_{out} and W_{out} depend on how the convolution was implemented in terms of kernel size, stride and padding of the input image. The feature map produced by the orange filter is indicated on the right in a transparent orange overlay. Adapted from (Bai, 2019).

A small receptive field, however, prevents the nodes from detecting larger and more abstract features in the image. Therefore, strategies are implemented to increase the receptive field of nodes deeper in the network. For instance, strided convolutions are used (i.e., the filter skips certain positions), or convolutions can be implemented without padding the borders of the image to reduce the height and width of the output. In addition, pooling layers are used to summarize the content of feature maps in a down-sampled form. For instance, a 2 × 2 max-pooling layer with two-pixel stride only keeps

the highest value in every 2×2 region, dividing the spatial dimension of the feature maps in half. Pooling layers are implemented after the non-linear activation function.

After each pooling layer, the receptive field of the nodes in CNNs increases. The first convolution layers generally pick up low-level features, e.g., edges and corners, while the more high-level features are picked up in deeper layers. Layer-by-layer, the spatial resolution of the feature maps decreases, while the level of abstractness increases hierarchically. Often, a CNN consists of convolutional layers followed by fully connected layers that incorporate all activations of the last feature maps. For volumetric images, the convolution and pooling operations in CNNs can be extended to 3D, i.e., using filters with a channel, height, width and depth dimension.

2.5.4.4 Image classification using CNNs

For classification, the feature maps of the last convolution layer are flattened to a long vector, after which fully connected layers are used for final classification. For binary classification, a single output node is used combined with a sigmoid function to convert the output into a probability in the range $[0, 1]$ (Figure 2.15).

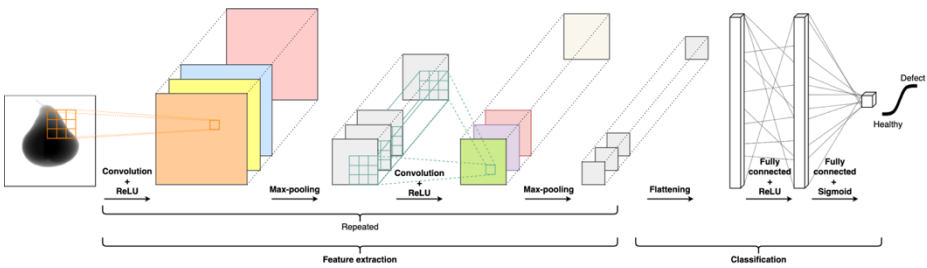

Figure 2.15: Simple CNN for binary classification. Several convolutional blocks, containing convolutional and pooling layers are coupled with fully connected layers and a single output node. The convolutional blocks serve as a learnable feature extractor, while the fully connected layers and output node perform the classification. In this model, ReLU is used as the nonlinear activation function (see Figure 2.12).

As a loss function, the binary cross-entropy (BCE) metric is most often used, i.e.,

$$Loss = \frac{1}{N}\sum_{i=1}^{N}(y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i) \qquad \text{(Eq. 4)}$$

in which N, y_i and $\hat{y}_i$ are the batch size, the ground truth label, and the model output, respectively. For multi-class classification in C classes, the output layer has C nodes and the SoftMax function is implemented to convert the output in probabilities, i.e., the values of output nodes range between 0 and 1, and sum up to 1. The loss on the output vector x is the sum of the cross-entropy (CE) loss calculated for every class c (see Eq. 5). The total loss is then the mean of the loss calculated over the whole batch. For unbalanced datasets, a class-weighted version can be used.

$$Loss(x, c) = -\log\left(\frac{exp(x_c)}{\sum_{i=1}^{C} exp(x_i)}\right) \qquad \text{(Eq. 5)}$$

2.5.4.5 Unsupervised image representation learning

Instead of classifying images in discrete classes, it is often useful to have low dimensional representations of images. These representations can be used for multiple downstream tasks. A way of getting low dimensional representations is by using an autoencoder (AE). An AE is an ANN that learns to map the input back to itself while going through some bottle-neck representation, i.e., a vector of lower dimensions. AEs thus comprise two parts, i.e., an encoder that encodes the input into a low dimensional representation, and a decoder that tries to reconstruct the input from the encoded information. The low dimensional representation, also called code, motivates the model to retain only the most important information. For images, the encoder and decoder contain (multiple) convolutional and fully connected layers. AE are trained by minimizing the reconstruction error, e.g., the mean-squared error (MSE), between the input and output.

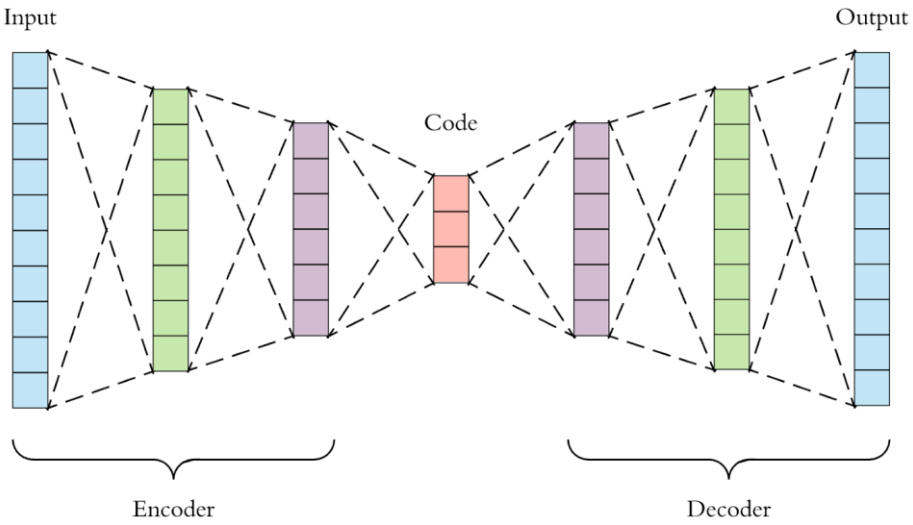

Figure 2.16: An autoencoder comprising an encoder that encodes the input in a low dimensional representation, i.e., code, and a decoder that reconstructs the input from the code.

The encoder is thus pretrained in an unsupervised way and can be used for downstream tasks, such as classification or clustering. Since the dimensions are drastically reduced, not only DL but also conventional ML algorithms can be coupled to the encoder for the downstream tasks. Additional to the encoded representation, also the reconstruction error between the input and output can be useful. For instance, it can be used for anomaly detection. By training an AE on normal data only, the reconstruction error is expected to be higher for anomalies because anomalies were not included in the training set. The reconstruction error represents thus an anomaly score.

To prevent AEs from learning the identity function and to improve the quality of the encoded representation, several regularization techniques can be used. In a denoising AE (DAE), the model is given a corrupted version of the input image and the objective is to restore the original undistorted image. The assumption is that the DAE learns higher-level features that generalize better to unseen data. More broadly, contractive autoencoders learn to map slight variations in the input to the same output. In sparse autoencoders, an extra term is added to the loss function to enforce sparsity by

penalizing activations (nonzero nodes). Therefore, the model learns to activate only a small subset of its hidden nodes when detecting statistical features in the input and leaves the activations of most other nodes close to zero. Hereby, the model is forced to learn more descriptive representations.

A stacked AE refers to an AE that is trained layer by layer instead of jointly. This means that the AE is trained with an increasing number of layers, in which the previously added layers are frozen and only the newly added layers are trainable. This reduces the required memory for training and might speed up convergence compared to jointly training. A downside of a pure stacked AE approach is that learning is greedy. Only a single layer can be optimized at a time, while the previously trained layers cannot be jointly optimized to improve overall performance. Therefore, greedy learning is mostly used as a pretraining step, after which all layers of the AE are fine-tuned together.

2.5.4.6 Semantic image segmentation

Semantic image segmentation is the task of classifying each individual pixel in one of C classes. It is mostly approached as a supervised learning problem with labeled data. The labeled data must thus contain labels on the pixel level, which can be hard to obtain and often requires manual labeling. This is especially challenging for 3D data. A lot of work can be required for labeling a single sample.

The most commonly used and effective neural network for semantic segmentation is the U-Net model (Çiçek et al., 2016; Ronneberger et al., 2015). The U-Net model is a CNN and similar to an autoencoder, it contains an encoder and decoder, which are mirrored versions of each other (see Figure 2.17). The task of the encoder is to capture different semantic concepts, while the decoder maps the semantic concepts to their spatial location in the image. In the encoder, the resolution of the feature maps decreases, while the number of feature maps (channels) increases. In the decoder, the resolution is again increased to the same resolution as the input image while the number of channels decreases. This structure is often depicted in a U-shape; hence the name U-Net.

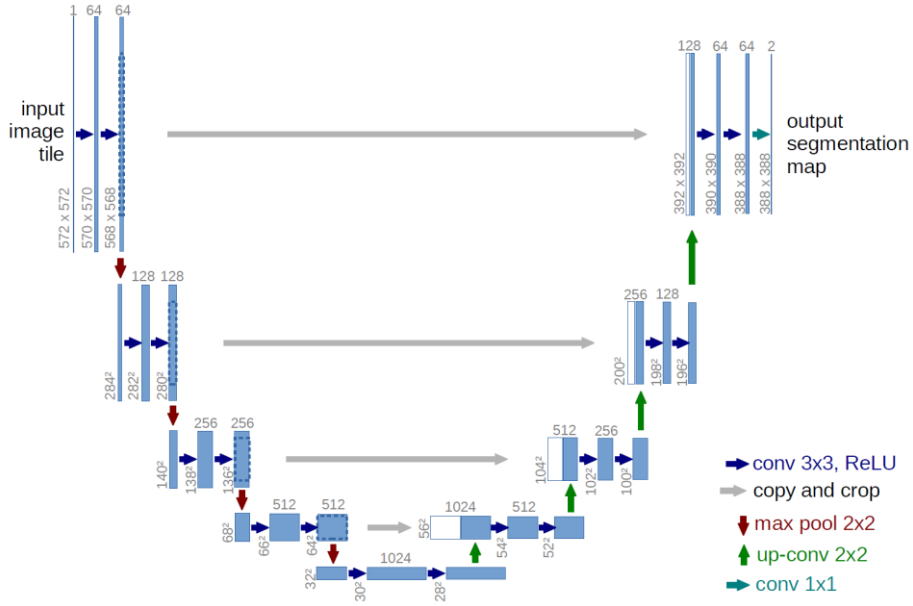

Figure 2.17: U-Net model (Ronneberger et al., 2015)

In the final step of the decoder, an output must be produced with C channels: a single channel for each class. Hereto, a final 1×1 convolution is applied with C filters. Each 1x1 filter combines all info in the previous channels into a single channel, while preserving the spatial resolution. It thus pools the feature maps in the channels dimension. Each pixel can then be assigned to one of the C classes, by finding the channel for which it had the highest value.

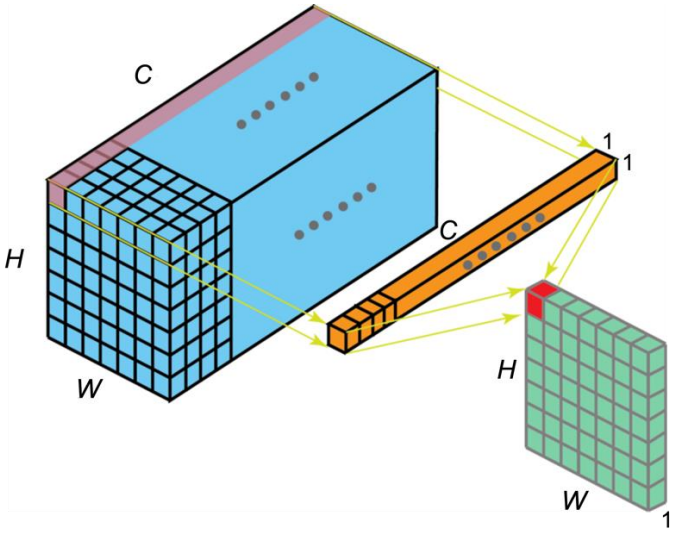

Figure 2.18: Illustration of a 1x1 convolution with a single filter. The filter slides over the input image along the height and width dimension. At each position, the convolution operates on a $C_{in} \times 1 \times 1$ region of the input (indicated in red on the left) and produces a single value in the output (indicated in red on the right). Adapted from (Bai, 2019).

Since its hard for the decoder to produce a detailed prediction from a representation with low spatial resolution, the U-Net model uses skip-connections. The skip-connections concatenate the feature maps of the encoder to the ones of the decoder at the same depth, i.e., with the same spatial resolution. Hereby, the decoder can incorporate more detail in its predictions from low-level features picked up by the first layers of the encoder.

Common loss functions for semantic segmentation are the pixel-wise CE loss, Dice-loss and the intersection over union (IoU) loss. Class imbalances are common for semantic segmentation tasks. While the Dice-score and IoU naturally incorporate class weighing, a class-weighted version of the pixel-wise CE loss is often used. The dice-loss is given by Eq. 6.

$$Loss_{Dice} = 1 - D = 1 - \frac{2}{C}\sum_{c=1}^{C}\frac{\sum_{i=1}^{N} y_{ci}\hat{y}_{ci}}{\sum_{i=1}^{N} y_{ci}^2 + \sum_{i=1}^{N} \hat{y}_{ci}^2} \qquad \text{(Eq. 6)}$$

Herein, D, C, N, y_{ci} and $\hat{y}_{ci}$ are the Dice-score, number classes, the number of pixels, the ground truth, and the prediction, respectively. The ground truth labels are one-hot encoded, i.e., a binary image with C channels in which each class is thus provided as a binary image. The IoU-loss is described by Eq. 7 and illustrated in Figure 2.19. In Eq. 7, A and B are the predicted and ground truth labels, in which the ground truth is one-hot encoded. The IoU can be computed directly from the Dice-score, i.e., $IoU = D/(2 - D)$.

$$Loss_{IoU} = 1 - IoU = 1 - \frac{|A \cap B|}{|A \cup B|} \qquad \text{(Eq. 7)}$$

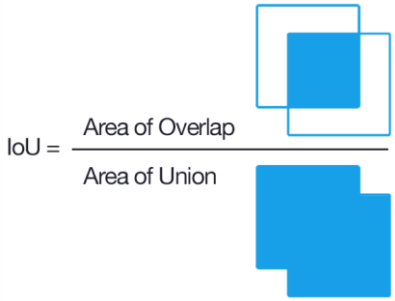

Figure 2.19: Intersection over union.

2.5.4.7 *Transfer learning and data augmentation*

Supervised learning is the most straightforward approach when it comes to applying deep learning. In practice, however, the absence of sufficient training data often limits its performance. Data annotation is a time-consuming job that requires expert knowledge. Various strategies have been proposed to overcome this problem. In transfer learning, a model pretrained on some other dataset is re-used for another task. Most layers of the pretrained model are kept constant, while the final layers of the network are retrained on a smaller dataset. The idea is that since the model is pretrained on a large dataset, it can already capture descriptive features, which can also be used as a feature extractor for other applications. A concern, however, is that the success of the pretrained feature extractor

depends on the similarity between the pretraining dataset (often ImageNet) and the dataset of the application.

Another approach is the use of data augmentation. Herein, the variation in the available dataset is artificially increased by applying transformations on the dataset during training, e.g., rotation, translation, cropping, scaling, or adding noise. This allows to make the model robust against these types of variations.

2.5.4.8 Regularization

A problem in machine learning is the phenomenon of overfitting on the training data. A model that is overparameterized relative to the size of the training dataset can perfectly fit the training data, without having the ability to generalize to unseen data. Deep neural networks have much more parameters than conventional ML algorithms and in theory should, therefore, be much more prone to overfitting. Note, however, that conventional ML algorithms rely on predefined features, while neural networks are trained end-to-end. Part of the model's capacity is thus used for feature extraction.

Several regularization techniques have been proposed to reduce the risk of overfitting in deep learning. The goal of regularization is to make it easier for the model to generalize than it is to overfit. It makes remembering the training data more expensive than finding more general solutions. The most common form of regularization is weight decay, which adds a term to the loss function that penalizes large weights. Another form of regularization is dropout. Herein, a changing set of nodes are randomly turned off during training which makes the training process noisier. The remaining nodes are forced to probabilistically co-adapt, resulting in a more robust model (Goodfellow et al., 2016).

Batch normalization is often implemented between a convolution and the nonlinear activation function. It normalizes the input of the activation function, so that the next layer has a more consistent distribution of inputs. It is said that batch normalization mitigates internal covariate shift. As a layer's weights change during training, it might change the distribution of its activations that will go to the next layer. This effect compounds layer-by-layer and could therefore

make it harder for models to converge. It has also been found that batch normalization improves generalization. It thus has a regularizing effect (Ioffe & Szegedy, 2015).

Additionally, early stopping can be implemented in which training is stopped once the model loses performance on the validation set. Finally, variability introduced by stochastic gradient descent and data augmentation also makes overfitting less likely (Goodfellow et al., 2016).

2.5.4.9 Deep learning in postharvest quality inspection

Twenty years ago, ANN-based methods were already proposed for bruise and watercore detection in apple using X-ray radiography (Kim & Schatzki, 2000; M. Shahin et al., 2002; M. A. Shahin et al., 2001). However, these methods still required feature engineering to provide inputs for a small MLP with a single hidden layer. These methods are, therefore, not considered "deep" learning. In recent years, deep learning has also been applied in agriculture and food science. Examples are disease or weed detection in the field, yield prediction, plant or fruit detection for automated picking, detection of food contamination, and food recognition and calorie estimation (Kamilaris & Prenafeta-Boldú, 2018; Zhou et al., 2019).

Postharvest applications related to quality inspection have mainly been limited to the external quality. A CNN was used to detect mangosteen fruit with surface defects from RGB images, for which a classification accuracy of 97 % was reported (Azizah et al., 2017). Tan et al. (2016) trained a five-layered CNN for detecting apples with skin lesions in infrared images and reported an accuracy of 97.5 %. The detection of apples with external defects in RGB images using a CNN was presented by Fan et al. (2020). A classification accuracy of 92 % was achieved compared to a SVM that scored 87 %. The detection of bruised blue berries based on hyperspectral images was investigated by Wang et al. (2018). They reported classification accuracies of 88 % using ResNet based architectures, which in their work outperformed conventional ML methods including SVM (81 %), linear regression (76 %), random forest (73 %), bagging of multiple decision trees (71 %), and MLP (78 %).

Pesticide residue detection on apples using hyperspectral imaging was tested using an AlexNet-based model for which a classification accuracy of 95 % was reported (B. Jiang et al., 2019). The online detection of surface defects of carrots based on RGB images using various CNNs, including AlexNet, ResNet, ShuffleNet with transfer learning was proposed by Deng et al. (2021). All CNNs had similar binary and multiclass classification accuracies of around 99.8 and 93.0 %, respectively, which all performed better than a Bayesian classifier on extracted features (93.1 and 91.0 % on the same tasks). Damaged root-trimmed garlic detection with RGB images using a CNN based on a pretrained VGG16 model was tested by Thuyet et al. (2020). Their system achieved an overall classification accuracy of 89 %.

Zhang et al. (2020) proposed a method to segment the normal surface, calyx region and bruises in hyperspectral images of blueberries using a fully convolutional network with a VGG-16 backbone. 1 x 1 convolutions with the number of output channels equal to the number of classes were applied on the final and intermediate feature maps, which were up-sampled and merged into a final prediction map. An overall IoU of 0.81 was achieved and it was found that transfer learning did not improve performance. The authors assigned the latter to large differences between their images and the ones in the large dataset used for pretraining, i.e., ImageNet. Additionally, they hypothesized that their integration of the models pretrained on 3-channel RGB images could be improved to better accommodate the application to hyperspectral images.

A method for online detection of citrus fruit with external defects was proposed by Chen et al. (2021). Their system used a pretrained MobileNetV2 model as backbone coupled with a PANet. The backbone CNN captured multiscale features in each of its layers, after which these feature maps were integrated by PANet for classification and bounding box prediction. The predicted bounding boxes were used to track the rolling fruit and integrate the predicted classes from multiple frames into a final prediction to overcome scenarios in which a defect would be invisible on a single frame. An overall multiclass classification accuracy of 87 % was achieved for

single frame classification, which was improved to 93 % using multiple frames via object tracking.

Fazari et al. (2021) deployed a ResNet101 model on hyperspectral images to detect external fungi infection of olives. The model was initialized with the weights pretrained on ImageNet, after which the whole model was fine-tuned on their dataset using data augmentation. To couple the multi-channel hyperspectral input images to the pretrained model that only accepts three input channels, a trainable convolutional layer with 1 x 1 kernels was added to the beginning of the network to bring the number of channels down to three. An overall classification accuracy of 91.8 % was achieved.

Several works have also focused on internal quality inspection. Yu et al. (2018) used a stacked autoencoder and a fully connected neural network to predict firmness and soluble solids content of pears based on hyperspectral imaging and achieved coefficients of determination (R^2) of 0.89 and 0.92, respectively. The deep learning-based method outperformed partial least squares regression and least-squares SVM models that achieved R^2 values up to 0.84 and 0.83 for respectively firmness and soluble solids content.

Liu et al. (2018) used deep learning for inline detection and classification of cucumber surface and internal (watery regions or cavities) defects from hyperspectral images. First, a CNN was coupled with a SVM and trained in a supervised way to classify image patches of normal and defect tissue. Herein, the CNN functioned as a trainable features extractor, while the SVM did the classification. In addition, a stacked sparse autoencoder was trained in an unsupervised way on the spectral signals of defect surfaces and its final representations were used to train a classifier to distinguish between different types of defects. The inspection system was then created by coupling all models and a final classification accuracy of 91.1 % was achieved. This significantly outperformed other tested methods for spectral-spatial classification, i.e., extended morphological profile SVM (68.3 %) and bag-of-visual-words (73.0 %).

A custom CNN and transfer learning with MobileNet and VGG19 models (pretrained on ImageNet) were used to classify crambe seeds from X-ray radiographies. Accuracies of 91, 95 and 82 % were reported for discriminating seeds based on internal tissue integrity, germination capacity and vigor, respectively. It was noted that the custom CNN and pretrained VGG19-based model performed better than the pretrained MobileNet-based model. Additionally, the custom CNN converged faster than the other models. Potentially, the features learned on ImageNet transferred poorly to their dataset (Medeiros et al., 2021).

2.6 Conclusions

Many instrumental techniques for nondestructive internal disorder detection have been proposed in literature. From these techniques, X-ray imaging has been identified as especially interesting for internal disorder detection due to the good penetration depth of X-rays through biological material, and the spatial and structural information that it provides. It can, therefore, be used to detect internal disorders related to density changes in the fruit, which has already been illustrated for some cases.

X-ray radiography is best suited for inline applications. However, the fact that it produces a cumulative 2D projection of a 3D object complicates the detection if internal disorders. Deviating patterns such as internal defects can be less prominent in the image due to the cumulation of information and the variability between samples. Current methods, therefore, rely on application specific algorithms or require product specific 3D models of the shape and density distribution.

By providing 3D information, X-ray CT can overcome these issues. It has shown to be effective for characterizing internal disorders, but methods for automated detection of internal defects are still lacking. Automated methods could potentially be used for more reliable nondestructive detection of internal disorders and could also be a valuable tool for researchers investigating the phenomenon. While X-ray CT is still too expensive and too slow for high volume inline inspection of low value products, it might become feasible in the

future with ongoing innovations in acquisition geometries and reconstruction software.

Current methods for internal disorder detection using X-ray imaging mostly rely on image processing to extract features, followed by a ML algorithm. However, the step of feature extraction has several downsides. First, features are often not transferable to other applications. Second, feature engineering requires expert knowledge which comes at a high cost. Third, even for experts finding useful features is labor-intensive and requires a lot of trial and error. Finally, potentially suboptimal solutions are created that are biased and limited by the human capability of interpreting the data. Therefore, in DL the intermediate step of feature extraction is removed altogether. Instead, models learn to solve a task end-to-end. In the learning process, the model discovers which features should be extracted to map the inputs to the target outputs.

DL already showed to be impactful in several fields, e.g., medical imaging using X-ray or MRI (Lee et al., 2017a; Litjens et al., 2017; Shen et al., 2017; Suzuki, 2017). Postharvest applications of DL to quality inspection have mainly been limited to classification problems related to external attributes. DL remains largely unexplored for the internal quality inspection of fruit. Therefore, there is an opportunity to apply deep learning to internal disorder detection in pears with X-ray imaging.

From the conclusions of the state-of-the-art above, the following targeted contributions are formulated:

- Developing a fully automated method for detecting healthy and defect pears based on 3D X-ray CT volumes
- Using deep learning to quantify disorder severity directly by segmenting internal disorders in 3D X-ray CT volumes of pears to improve classification
- Developing an inline method to detect and localize internal disorders in pears using deep learning in an unsupervised way on X-ray radiographs, which does not require a large labeled dataset

Chapter 3

Nondestructive Internal Quality Inspection of Pear Fruit by X-ray CT using Machine Learning[1]

3.1 Introduction

X-ray CT has shown to be an effective tool to characterize internal disorders in horticultural products. For instance, it has been used to investigate the spatial distribution of internal browning in pear and apple (Herremans et al., 2013; Lammertyn et al., 2003b). However, an automated approach for detecting pears with internal disorders is lacking. Currently, the analysis of 3D data often requires semi-automatic or long-lasting workflows that do not scale to large datasets. Additionally, research in image acquisition geometries and reconstruction algorithms is progressing towards inline implementations of X-ray CT. For instance, in translational CT, the sample is rotated and translated simultaneously. Alternatively, a gantry system can be used in which source and detector pairs rotate around a translating sample. Automated methods for detecting fruit with internal disorders using X-ray CT could therefore become feasible for quality inspection in the future.

The aim of this chapter is to present a nondestructive method for automated internal quality grading of pear fruit using X-ray CT and machine learning. The internal quality grading is presented as a

[1] This chapter is based on: Van De Looverbosch, Tim, Md. Hafizur Rahman Bhuiyan, Pieter Verboven, Manuel Dierick, Denis Van Loo, Jan De Beenhouwer, Jan Sijbers, and Bart Nicolaï. "Nondestructive Internal Quality Inspection of Pear Fruit by X-Ray CT Using Machine Learning." *Food Control* 113 (2020).

classification problem between healthy and defect samples. Hereto, a classical machine learning based approach is preferred over a deep learning based approach due to the limited number of samples (CT volumes). A 3D image processing algorithm is proposed to extract relevant quantitative features to discriminate between healthy and defect pears. Thereafter, a binary linear Support Vector Machine (SVM) is trained with these features and tested on labeled X-ray CT reconstructions of the cultivars *Pyrus communis* L. cv. 'Conference' and *Pyrus communis* L. cv. 'Cepuna'. In addition, a feature selection procedure is implemented to reduce model complexity and duration of the image processing algorithm. 'Conference' is one of the most important commercial cultivars in Europe, represents almost 90 % of the acreage of Belgian pears (Statbel, 2018), and is known to be susceptible to internal browning (Franck et al., 2007). 'Cepuna' is a cross between 'Conference' and 'Doyenné d'Hiver' and used for testing the transferability of the method to other cultivars.

3.2 Materials and methods

3.2.1 Pear fruit and long-term storage

'Conference' and 'Cepuna' pears were respectively harvested on 14 and 25 September 2017 and delivered by a grower member of the Flemish fruit cooperatives BFV and Belorta (Belgium), respectively. Starting from the harvest date, the fruit was stored for six months following two treatments, with approximately 50 kg fruit per treatment. In the first treatment, the storage conditions were set according to the recommendations of the Flanders Centre of Postharvest Technology (VCBT, Leuven, Belgium) for commercial sale (Ultra Low Oxygen treatment, ULO) to deliver control fruit without internal disorders (VCBT, 2017). Herein, the temperature, O_2 and CO_2 partial pressures were set to -1.0 °C, 3.0 kPa and 0.7 kPa, respectively. Prior to the ULO storage, fruit following this treatment underwent an acclimatization period of 21 d at -1.0 °C. In the second treatment, suboptimal storage conditions based on the findings of (Lammertyn et al., 2000), were applied to deliver fruit with internal disorders. Herein, the temperature, O_2 and CO_2 partial pressures were respectively -1.0 °C, 1.0 kPa and 5.0 kPa. A low O_2 partial pressure causes hypoxia in the fruit. In combination with increased

CO_2 partial pressure, this promotes the shift from respiration to fermentation, resulting in a limited availability in energy and an imbalance between oxidative and reductive processes. As such, cell membranes are degraded by reactive oxygen species, leading to cell leakage and cell death, which result in internal browning and cavity formation (Franck et al., 2007; Pedreschi et al., 2009; Veltman et al., 2003).

3.2.2 X-ray CT scans and data labeling

After approximately 6 months, the fruit was removed from storage on 2018-02-27 at the end of the day. Fruit were acclimatized to room conditions before X-ray CT scanning the next day. Minimally 50 fruit per treatment were randomly selected and scanned individually. The fruit were scanned with their stalk-calyx axis approximately aligned with the rotation axis of the scanner. To stabilize the samples during scanning, the fruit was placed on a sample holder consisting of three styrofoam cones glued on a stainless-steel plate which was mounted on top of the rotation table. The system comprised a micro-focus L9181 X-ray source (Hamamatsu Photonics, Hamamatsu, Japan) and a 1512 Dexela CMOS Flat Panel X-ray Detector (PerkinElmer, Waltham, Massachusetts, USA). The rotation table and detector were placed at respectively 674.8 mm and 784.2 mm from the source. The X-ray projections did not fit entirely in the X-ray detector frame. Therefore, two scans per fruit were performed at different heights and stacked together to reconstruct the whole fruit in the CT volume. The scans were performed with a source voltage of 130 kV at 300 mA and pixel size of 598.4 µm. The exposure time was 80 ms. An aluminum filter of 1 mm thickness was used to improve the contrast in the radiographic projections and to reduce beam hardening effects. The projections were obtained with an angular step of 0.9° and were 242 × 192 pixels in size. The samples were rotated over 360° around the central rotation axis of the scanner, resulting in 400 projections. For the acquisition, ACQUILA software was used (Tescan XRE nv, Ghent, Belgium). A 3D image of each fruit was reconstructed with the filtered back-projection algorithm using the ACQUILA-RECON reconstruction software (Tescan XRE nv, Ghent, Belgium). A combination of a polynomial and a Gaussian filter was applied to reduce ring

artefacts. The resulting tomographs had a size of 241 × 241 × 309 voxels, with each isotropic voxel measuring 514.9 × 514.9 × 514.9 μm^3. In total, scanning (32 s/scan), moving the sample stage down and starting the second scan (2 s), stacking and reconstruction (23 s) amounted on average to 1 min and 30 s per sample. The samples were assigned a ground truth label (*'healthy'* or *'defective'*) by visual inspection of the CT reconstruction of each fruit. However, to prevent missing incipient browning and to consider consumer acceptance and preferences in future research, it is suggested to perform an expert panel survey for labeling the fruit based on images of cut fruit in addition to a visual inspection of the CT data.

Figure 3.1 shows the experimental X-ray CT setup and a cut-open image and orthogonal slices through the CT volume of a *'Cepuna'* pear severely affected by internal browning. Internal browning can be observed in the lower intensity regions on the CT slices.

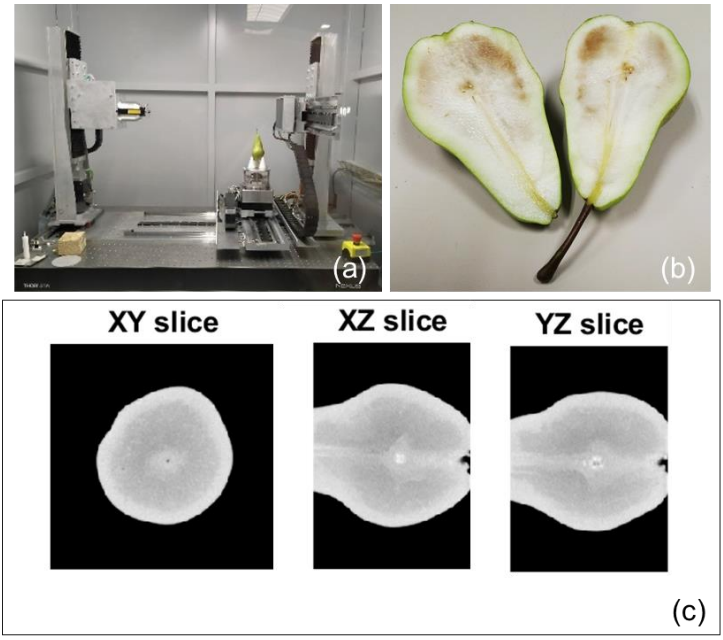

Figure 3.1: Experimental setup and a 'Cepuna' pear's cut-open image and CT scan. (a) Experimental setup with X-ray source (left), and mobile X-ray detector (right) and rotation stage (middle); (b) image of a cut-open 'Cepuna' pear affected by internal browning; (c) orthogonal slices through the CT volume of the same fruit. The XZ and YZ slices are zoomed in on the region affected by internal browning.

3.2.3 Internal disorder detection method for CT images

An algorithm was developed to perform internal disorder detection for pear fruit using the CT images. Recent developments in the medical imaging field reported interesting results for disease detection in CT and MRI data using deep learning based segmentation methods (Lee et al., 2017b; Shen et al., 2017). However, these approaches typically require many manually labelled samples for training. Due to the limited number of samples and the high cost of manually labelling them, a more classical machine learning approach was chosen. First, a feature extraction algorithm was developed to calculate valuable quantities, or features, from the 3D image datasets (see section 3.2.3.1). Subsequently, the features were statistically compared between the cultivars and classes (see section 3.2.3.2). Thereafter, support vector machines (SVM) were trained separately on the 'Conference' feature dataset to classify the fruit. Then, it was investigated whether features could be eliminated while minimizing the reduction in the classification performance. Finally, to test the generalizability of the method, the classifier trained on the 'Conference' data was validated on fruit of the 'Cepuna' cultivar and compared with classifiers trained on the combined dataset (see section 3.2.3.3). All code was written in MATLAB using the Image Processing and Statistics and Machine Learning Toolboxes (MATLAB, 2019b, 2019a).

3.2.3.1 Feature extraction algorithm

A feature extraction algorithm was developed to extract 10 features from the CT volume of each pear fruit, which produced 2 feature datasets, one for each cultivar. To extract the features, five 3D binary masks were generated indicating different regions of the fruit (see Figure 3.2).

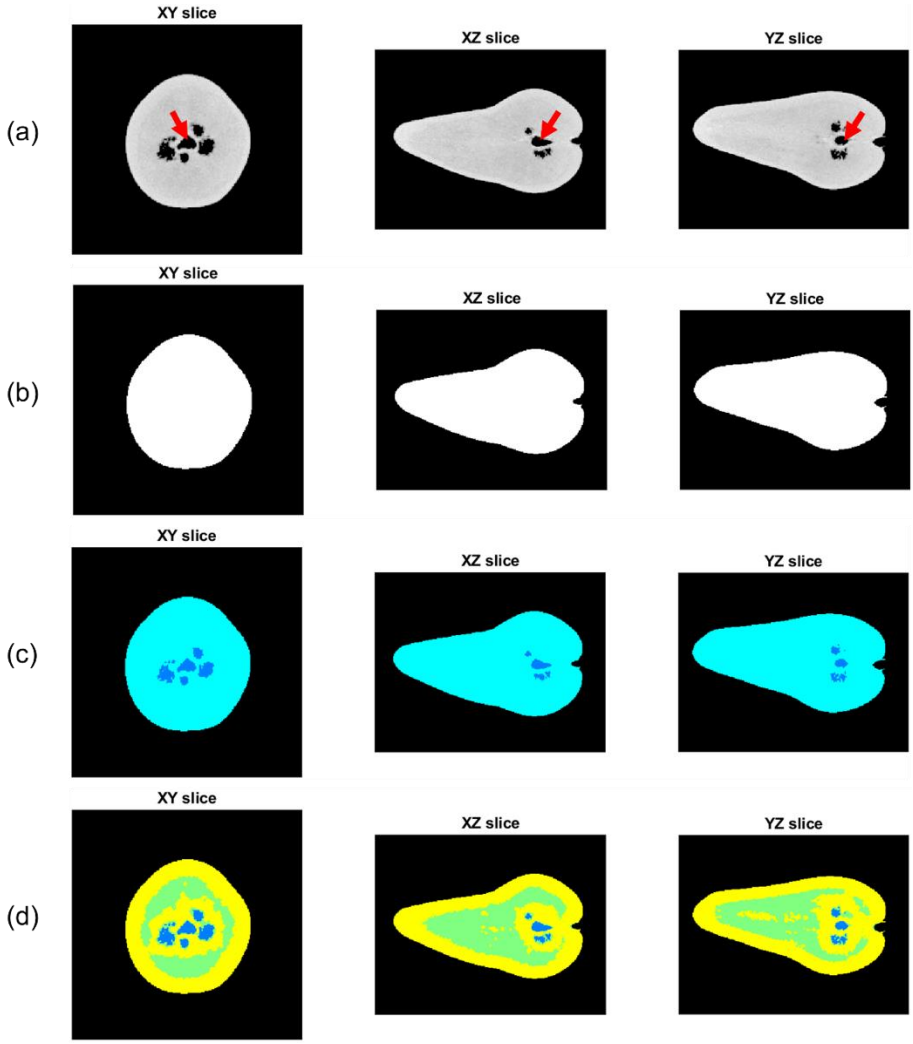

Figure 3.2: Orthogonal slices through the original grayscale CT reconstruction and generated 3D binary masks of a 'Conference' sample with disorders. (a) Orthogonal slices through the original grayscale CT reconstruction of the sample; cavities can be observed around the core (core indicated by red arrow); (b) orthogonal slices through the fruit mask (white); (c) orthogonal slices through the tissue (cyan) and internal air masks (dark blue); (d) orthogonal slices through the internal air (dark blue), low-density tissue (green) and high-density tissue (yellow) masks. Each 3D binary mask indicates whether a certain voxel belongs to a certain segment of the volume (value = 1) or not (value = 0).

Each 3D binary mask specified whether a certain voxel belonged to a certain segment of the volume (value = 1) or not (value = 0). First, a binary mask that indicated which voxels were part of the fruit tissue (tissue mask) was generated using a 3D global Otsu-threshold (Otsu, 1979). Second, a fruit mask was generated by filling up all internal holes of the tissue mask so that voxels outside and inside the fruit had the values 0 and 1, respectively. Third, an internal air mask, only including voxels part of holes, was obtained by subtracting the tissue mask from the fruit mask. Finally, a low-density and a high-density tissue mask were generated by applying a 3D adaptive threshold on all tissue voxels based on the local mean intensity in a 31 x 31 x 31 voxel neighborhood. In pome fruit, tissue of higher density can typically be observed around the core and in the surface region. In-between those regions, typically a higher porosity can be found (Nugraha et al., 2019). Low-density and high-density tissue regions are thus generally always present, but a large difference between those regions can indicate the occurrence of water loss due to internal tissue breakdown. In the reconstructed CT volume, voxels with a relative low intensity value had a lower X-ray attenuation, and thus lower density (higher porosity), than voxels with a higher intensity value.

Features were extracted using the generated masks. By subtracting the tissue mask from the fruit mask and counting the number of remaining voxels, the internal air volume could be calculated. As a first feature, the internal air volume normalized for the total fruit volume was used. For the second to ninth feature, the mean and standard deviation of the intensities of fruit voxels, tissue voxels, low density tissue voxels and high-density tissue voxels were calculated by using the fruit, tissue, low density tissue and high-density tissue mask, respectively. As a final feature, the Kolmogorov-Smirnov test statistic (KS-value) of the Two-Sample Kolmogorov-Smirnov Test between the cumulative intensity distributions of the low- and high-density tissue voxels was used (Massey, 1951; MATLAB, 2019c). Here, the KS-value was interpreted as a measure of homogeneity of the fruit tissue by comparing the intensity distributions of both regions. A lower KS-value indicates that the low-density and high-density tissue regions are of similar density, suggesting that internal

tissue breakdown such as browning is less probable (Franck et al., 2007). This is illustrated for a *'defective'* and *'healthy'* 'Conference' fruit in Figure 3.3. The feature datasets were centered and scaled using the corresponding column mean and standard deviation.

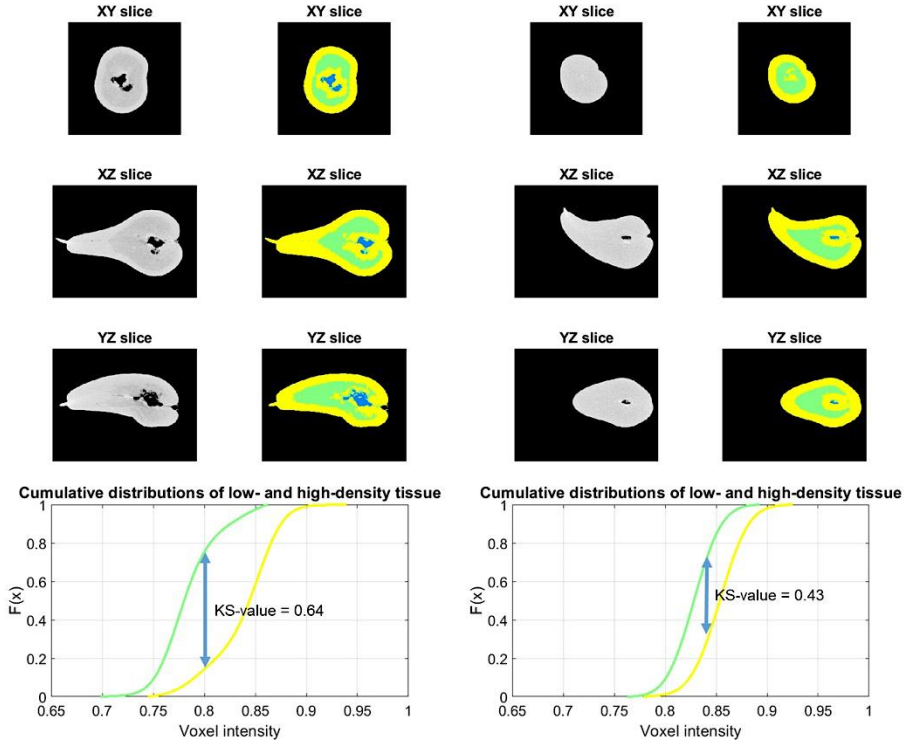

Figure 3.3: Orthogonal slices through CT volume, cumulative intensity distributions and KS-value of the low- (green) and high-density tissue (yellow) in a 'defective' (left) and 'healthy' (right) 'Conference' pear. The 'healthy' sample has more similar intensity distributions and lower KS-value compared to the 'defective' sample

3.2.3.2 Statistical feature comparison

A quantitative feature comparison was performed to explore the data, investigate differences between the cultivars or classes and infer relevant features for classification. Hereto, it was tested if the normal distributions of the features were significantly different between the 'Conference' and 'Cepuna' cultivars on the one hand and the *'defective'* and *'healthy'* classes on the other hand using a Two-Sample t-test at the 5% significance level. Moreover, the linear

correlation coefficients between all features of the 'Conference' feature dataset were calculated.

3.2.3.3 Binary linear Support Vector Machine classifiers

Using the 'Conference' feature dataset, a binary support vector machine (SVM) with a linear kernel was trained and evaluated to classify the fruit. The box constraint, i.e. maximum penalty imposed on margin-violating observations, was set to the default value of 1.0. All variables were standardized using their corresponding mean and standard deviation. A linear kernel was chosen over non-linear approaches because of its simplicity in terms of the number of parameters, because it allows to interpret the importance of each feature for classification and because it is less prone to overfitting compared to non-linear methods.

Confusion matrices were used to present the classification results with true positives (the correctly classified fruit with internal disorders) and true negatives (the correctly classified fruit without internal disorders) shown as a percentage on the matrix diagonal. The false positives and false negatives are shown as a percentage on the bottom left and the top right, respectively.

Thereafter, it was investigated whether the number of features used by the classifier could be reduced without losing classification performance. Hereto the SVM recursive feature elimination method (SVM RFE) as described by (Guyon et al., 2002) was used in which the importance of each feature relative to the other features was evaluated based on the weights that define the decision boundary of the SVM in feature space. The higher the squared weight value, the more important the corresponding feature is for classification. Note that doing it this way is only possible when using a linear kernel in the SVM, as for nonlinear kernels, a generalized version of SVM RFE must be used (Guyon et al., 2002). In practice, a series of classifiers was trained and evaluated on the 'Conference' dataset using 5-fold cross-validation (further referred to as the *'Conference' based SVMs*). In each iteration, the feature with the lowest squared weight value was eliminated. By tracking the average cross-validation classification accuracy and false positive and negative rates, a

decision was made on which features were the most critical and which classifier should be used. In the 5-fold cross-validation, the data was randomly partitioned into 5 sets. Every set was reserved as a validation set after the model was trained using the other four sets.

Next, the generalizability of the trained classifier to other cultivars was evaluated and it was investigated whether the generalizability would increase with a reduction in the number of features. Hereto, the series of trained classifiers was validated on the feature dataset of the 'Cepuna' cultivar. Finally, the 'Conference' and 'Cepuna' datasets were combined and the performance of the series of *'Conference' based SVMs* was compared with two series of SVMs retrained on this combined dataset. The first series was forced to use the same features as the *'Conference' based SVMs* in each iteration, while in the second series the feature elimination algorithm decided which features were retained.

3.3 Results

3.3.1 X-ray micro-CT reconstructions and labeled datasets

For 'Conference', 102 samples were scanned of which 42 and 60 fruit were assigned a *'healthy'* and *'defective'* label, respectively, from expert inspection of the CT images. For 'Cepuna', 15 *'healthy'* and 87 *'defective'* fruit were observed in the 102 scanned samples.

Examples of orthogonal slices and grayscale intensity profiles through CT reconstructed volumes of *'healthy'* and *'defective'* 'Conference' and 'Cepuna' fruit are shown in Figure 3.4. In the *'healthy'* fruit (Figure 3.4, rows b and d) a gradient in voxel intensity can be observed from the center to the fruit surface. Higher intensities due to higher tissue density were observed around the core. When moving from the core towards the fruit surface, the intensities first decreased and thereafter increased again closer to the surface, confirming the observed density distributions from other research (Nugraha et al., 2019). The *'defective'* 'Conference' fruit (Figure 3.4, row a) showed regions of lower voxel intensities that were affected by internal browning (Franck et al., 2007;

Lammertyn et al., 2003b; van Dael et al., 2017). Severe internal browning resulted in cavity formation, which was observed around the core and stalk-calyx axis. The *'defective'* 'Cepuna' fruit (Figure 3.4, row c) were also affected by internal browning, but cavity formation was far less severe. In the grayscale intensity profiles of the *'defective'* fruit, the regions affected by internal browning caused a stronger slope compared to those of the *'healthy'* fruit.

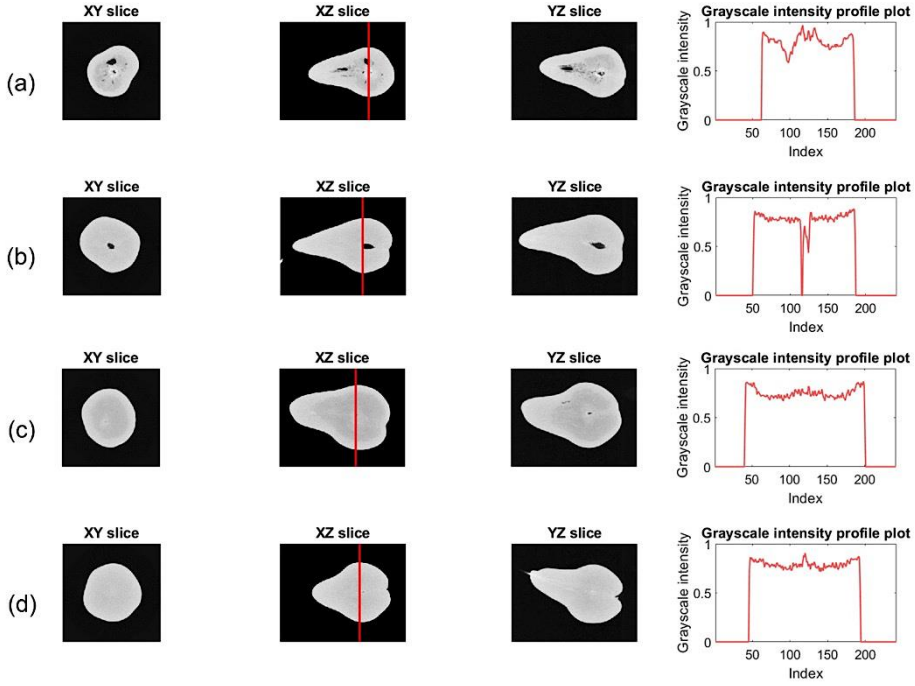

Figure 3.4: Column 1-3: Orthogonal slices through the CT reconstructions of 'defective' and 'healthy' 'Conference' and 'Cepuna' fruit. Column 4: Grayscale intensity profile through the widest position of the fruit in the XZ slices. (a) 'defective' 'Conference' pear; (b) 'healthy' 'Conference' pear; (c) 'defective' 'Cepuna' pear; (d) 'healthy' 'Cepuna' pear.

3.3.2 Quantitative feature comparison

The average of the extracted features and their corresponding standard deviations for 'Conference' and 'Cepuna' *'defective'* and 'healthy' fruit are shown in Table 3.1. Using a Two-Sample t-test at the 5 % significance level, it was tested if the distributions of feature values were significantly different between the 'Conference' and 'Cepuna' cultivars on one hand and the *'defective'* and *'healthy'*

classes on the other hand. Figure 3.5 presents the linear correlation coefficients (R) between all features of the 'Conference' feature dataset.

Table 3.1: The average of feature values and their corresponding standard deviations for 'Conference' and 'Cepuna' 'defective' and 'healthy' fruit. Different letters in superscript indicate significantly different normal distributions at the 5% significance level using the Two-Sample t-test.

Feature	Label	'Conference'	'Cepuna'
Normalized cavity volume [%]	Defective	0.752 ± 0.745^{c}	1.205 ± 1.297^{d}
	Healthy	0.195 ± 0.156^{b}	0.002 ± 0.002^{a}
Mean fruit intensity	Defective	0.829 ± 0.015^{c}	0.802 ± 0.017^{a}
	Healthy	0.835 ± 0.012^{d}	0.824 ± 0.008^{b}
Std fruit intensity	Defective	0.091 ± 0.025^{c}	0.102 ± 0.038^{c}
	Healthy	0.066 ± 0.008^{b}	0.053 ± 0.001^{a}
Mean tissue intensity	Defective	0.835 ± 0.013^{c}	0.812 ± 0.013^{a}
	Healthy	0.836 ± 0.012^{c}	0.824 ± 0.008^{b}
Std tissue intensity	Defective	0.062 ± 0.004^{c}	0.064 ± 0.006^{d}
	Healthy	0.056 ± 0.002^{b}	0.053 ± 0.001^{a}
Mean low-density tissue intensity	Defective	0.814 ± 0.019^{c}	0.782 ± 0.023^{a}
	Healthy	0.825 ± 0.013^{d}	0.806 ± 0.010^{b}
Std low-density tissue intensity	Defective	0.027 ± 0.060^{c}	0.039 ± 0.020^{d}
	Healthy	0.020 ± 0.002^{a}	0.023 ± 0.002^{b}
Mean high-density tissue intensity	Defective	0.856 ± 0.012^{c}	0.838 ± 0.013^{a}
	Healthy	0.854 ± 0.011^{c}	0.846 ± 0.006^{b}
Std high-density tissue intensity	Defective	0.025 ± 0.003^{c}	0.031 ± 0.005^{d}
	Healthy	0.023 ± 0.002^{a}	0.023 ± 0.001^{a}
KS-value	Defective	0.578 ± 0.116^{b}	0.586 ± 0.129^{b}
	Healthy	0.506 ± 0.048^{a}	0.610 ± 0.038^{b}

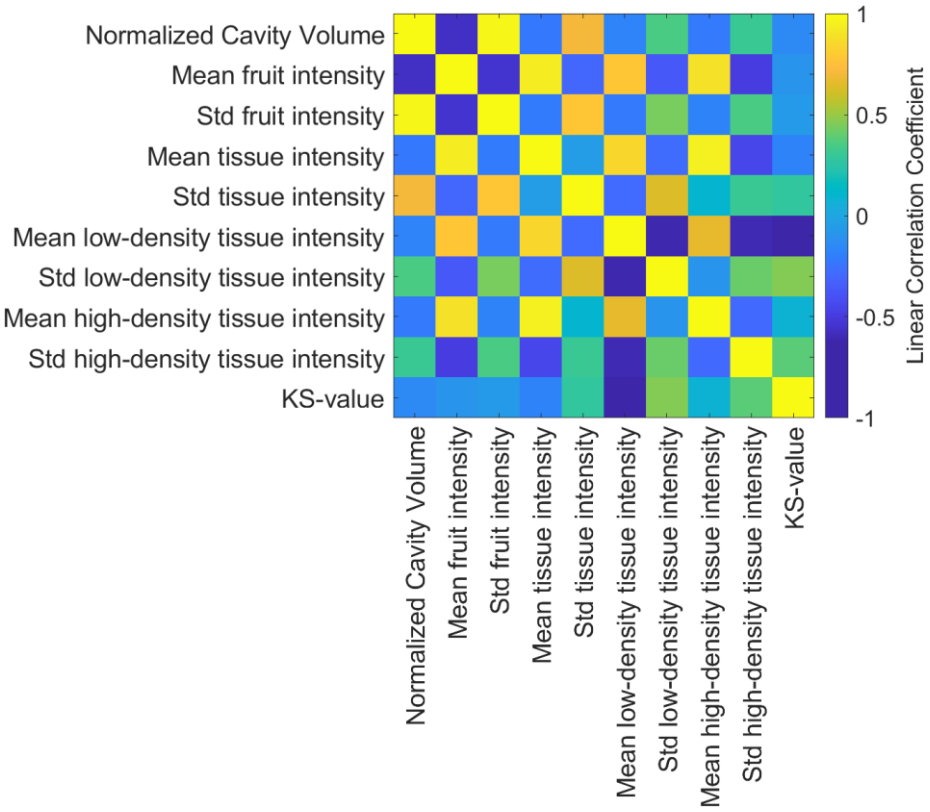

Figure 3.5: Linear Correlation Coefficients between all features of the 'Conference' dataset.

The features *'Normalized cavity volume'* and *'Std fruit intensity'* were highly correlated (R = 0.98). This was expected as fruit with more or large cavities would also have a high variability in voxel intensity. A lower linear correlation was found between the features *'Normalized cavity volume'* and *'Std tissue intensity'* (R = 0.71). The *'Std tissue intensity'* only considers non-cavity voxels, but fruit with a relatively high number of cavities could have more partial volume artefacts (the loss of contrast in voxels that are occupied by multiple types of tissue due to insufficient resolution), or internal browning and thus a higher variability in fruit tissue intensity.

We also observed that the features *'Mean fruit intensity'*, *'Mean tissue intensity'*, *'Mean low-density tissue intensity'* and *'Mean high-density tissue intensity'* were highly correlated, with linear correlation coefficients ranging between 0.66 and 0.95. Not surprisingly, the

correlations were higher between the features of which the regions indicated by the 3D binary masks were more similar, *e.g., 'Mean fruit intensity'* and *'Mean tissue intensity'* had a higher correlation than *'Mean low-density tissue intensity'* and *'Mean high-density tissue intensity'*.

Obviously, as the fruit and tissue masks only differ in the cavity voxels, a rather high correlation (R = 0.78) was found between *'Std fruit intensity'* and *'Std tissue intensity'*. *'Std tissue intensity'* and *'Std low-density tissue intensity'* had a linear correlation of R = 0.64.*'Std high-density tissue intensity'* and *'KS-value'* were not highly correlated with other features, except for moderate negative correlations with *'Mean low-density tissue intensity'*.

3.3.3 Classification results

3.3.3.1 'Conference' based SVM

The classifier trained on the whole 'Conference' feature dataset, comprising 60 'defective' and 42 'healthy' samples, reached an average classification accuracy of 92.2 % for 'Conference' in a 5-fold cross validation with an 88.3 % true positive and a 97.6 % true negative rate, respectively. The confusion matrix with classification results for 'Conference' is shown in Table 3.2. The runtime for feature extraction and classification was on average 2.3 s per sample on a quad-core 3.8 GHz processor with 32 GB of RAM memory.

Table 3.2: Confusion matrix with average classification results of the 'Conference' specific classifier on 'Conference' in 5-fold cross validation.

		Predicted	
		Defective	Healthy
Ground truth	Defective	88.3 %	11.7 %
	Healthy	2.4 %	97.6 %
	Overall classification accuracy:	92.2 %	

The weights that determine the separating plane are shown in Figure 3.6. The features with a high absolute value of the weight are the most important for determining the class of a fruit. The top three features were *'Std tissue intensity'*, *'Std high-density tissue intensity'*

and *'Std low-density tissue intensity'*. Features that measure variability rather than absolute values had higher absolute weights and, thus, were more important for classifying pear fruit. Moreover, fruit with higher values for these features were more likely to be classified as *'defective'* fruit, *i.e.,* the positive class, due to the positive corresponding weights. Both the weights of *'Mean tissue intensity'* and *'KS-value'* features were rather insignificant.

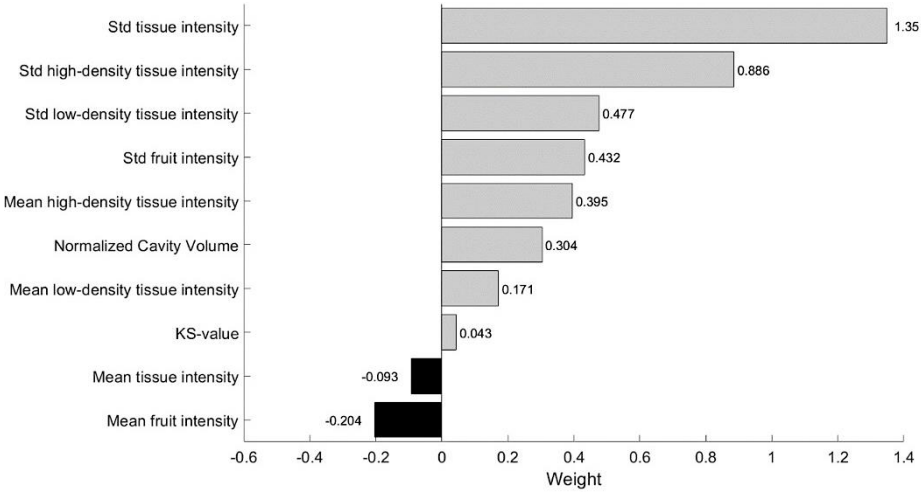

Figure 3.6: Weights of the 'Conference' based SVM sorted by descending weight value.

3.3.3.2 Feature selection

From the previous results, it was observed that not all features were equally important for classification. Some features, *e.g.,* '*KS-Value'*, *'Mean Tissue Intensity'* and *'Mean low-density tissue intensity'*, have relatively low weights compared to others (see Figure 3.6). As explained in section 3.2.3.3, the SVM RFE method was applied to select the most relevant features. A series of classifiers was trained and evaluated on the 'Conference' dataset and in each iteration the feature with the lowest squared weight value was removed for the next iteration. The resulting features used by each classifier and its obtained classification accuracy are shown in Figure 3.7.

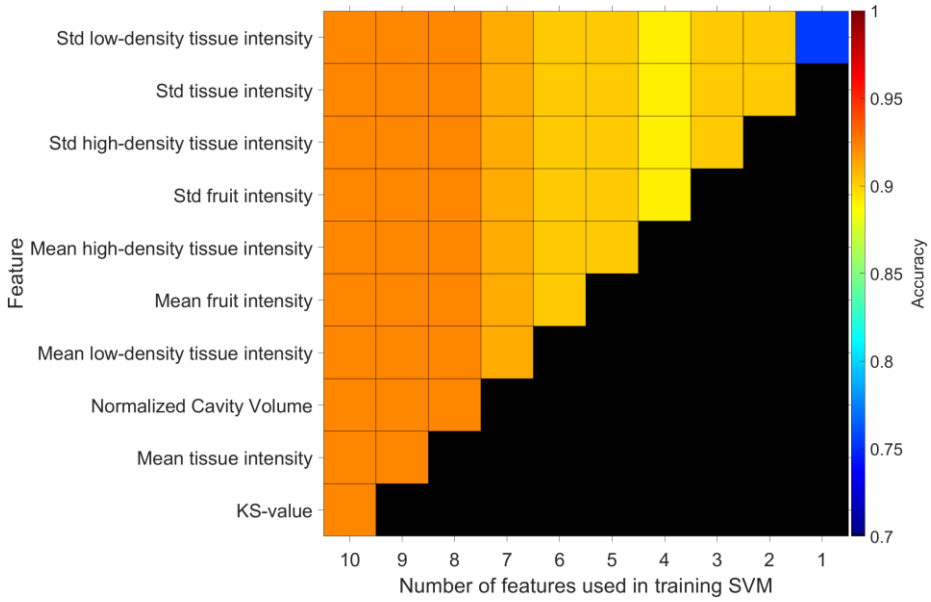

Figure 3.7: Plot of features used for training each SVM on the 'Conference' dataset. Each column represents a classifier in a series of classifiers, in which the number of features allowed to be used decreased from the left (10 features) to the right (1 feature). Every column thus shows the features used by a classifier in the series, while each row shows in which classifiers a certain feature was used. A colored tile indicates a feature was used, while a black tile indicates a feature was eliminated for the specific SVM in the feature elimination procedure. The color of each column represents its obtained classification accuracy.

Figure 3.8 shows a plot of the classification accuracy, true positive, true negative, false positive and false negative rate of the SVM series trained with 1 up to 10 features of the 'Conference' dataset in a 5-fold cross-validation.

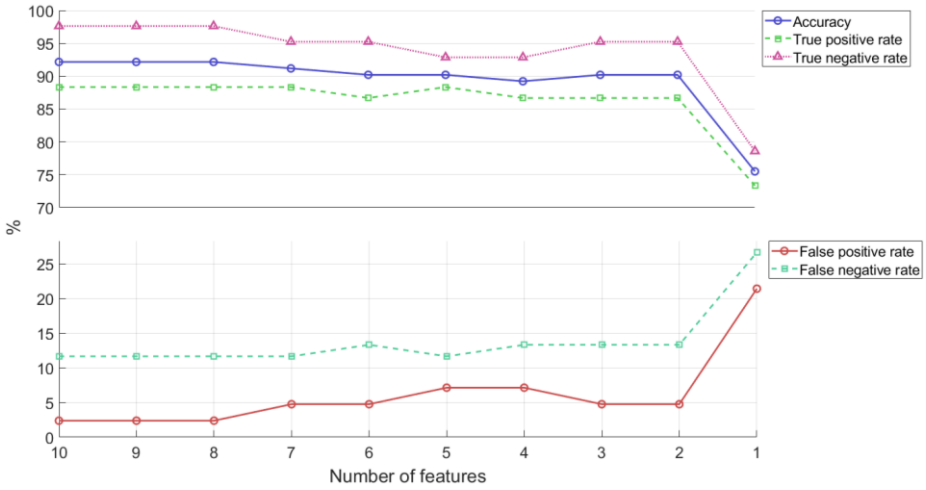

Figure 3.8: Plot of classification accuracy, true positive rate, true negative rate, false positive rate and false negative rate of SVMs trained with 1 and up to 10 features of the 'Conference' dataset.

From Figure 3.8 it can be observed that the performance metrics only slightly decreased up until the last SVM trained with only 1 feature, for which the accuracy dropped to 75.5 %. The SVM trained with two features, (*'Std tissue intensity'* and *'Std low-density tissue intensity'*, see Figure 3.7) still achieved an accuracy of 91.2 % with a false positive rate of 4.8 % and a false negative rate of 13.3 %, which is slightly higher than the 11.7 % for the best classifier.

3.3.3.3 Validating the 'Conference' based SVMs on the 'Cepuna' cultivar

To test the generalizability to other cultivars, the series of 'Conference' based SVMs was validated on the 'Cepuna' cultivar without retraining the classifiers on the 'Cepuna' data. The 'Cepuna' dataset comprised 87 *'defective'* and 15 *'healthy'* fruit. The ten-features classifier reached a good classification accuracy of 95.1 % for 'Cepuna' with 94.3 % true positive and 100.0 % true negative rate, respectively. The confusion matrix with classification results for 'Cepuna' is shown in Table 3.3.

Table 3.3: Confusion matrix with classification results of the ten-features 'Conference' based SVM on 'Cepuna' without retraining the SVM.

		Predicted	
		Defective	Healthy
Ground truth	Defective	94.3 %	5.7 %
	Healthy	0.0 %	100.0 %
	Overall classification accuracy:	95.1 %	

When testing the *'Conference' based SVM* series trained with 1 to 10 features on the 'Cepuna' dataset, the classification performance remained the same for the classifiers using between ten and five features. For the classifiers using between 4 and 2 features, the false positive rate increased from 0.0 % to 6.7 %, while the accuracy remained 95.1 % and the false negative rate reduced from 5.7 % to 4.7 %. With only one feature, the last classifier had a false positive rate of 80.0 %. However, due to the low number of *'healthy'* samples in the 'Cepuna' dataset (fifteen), the accuracy only dropped to 86.7 %.

3.3.3.4 Testing the 'Conference' based SVMs and retrained classifiers on the combined dataset

The series of *'Conference' based SVMs* was tested on the combined dataset. The confusion matrix with classification results of the ten-features classifier is shown in Table 3.4.

Table 3.4: Confusion matrix with average classification results of the ten-features 'Conference' based SVM on the combined dataset.

		Predicted	
		Defective	Healthy
Ground truth	Defective	91.8 %	8.2 %
	Healthy	1.7 %	98.3 %
	Overall classification accuracy:	93.6 %	

Like the previous results, the classification accuracy, true positive rate, true negative rate, false positive rate and false negative rate of the *'Conference' based SVMs* classifiers using between ten and two

features tested on the combined dataset was similar. The two-feature *'Conference' based SVM* achieved an accuracy of 93.1 % and a false positive and false negative rate of 3.5 % and 8.2 %, respectively.

Next, a first series of SVMs was retrained on the combined dataset but was forced to use the same features as their corresponding *'Conference' based SVM* (see Figure 3.7). The SVMs were thus only allowed to change the weight associated to a certain feature. The classifiers using ten and two features achieved the same classification metric scores, with an accuracy of 92.7 % and false positive and false negative rate of 5.3 and 8.2 %, respectively. However, the accuracy and false positive rate were slightly worse compared to the two-feature *'Conference' based SVM*, that achieved the same false negative rate with an accuracy of 93.1 % and false positive rate of 3.5 %.

Finally, a second series of SVMs was retrained on the combined dataset which was now allowed to change the selected features at each iteration. The used features are shown in Figure 3.9. The ten-feature classifier reached an accuracy of 92.7 % and false positive and false negative rates of 5.3 and 8.2 %, respectively. The two-feature classifier scored an accuracy of 91.2 % with a false positive and false negative rate of 12.3 and 7.5 %, respectively. The latter classifier used the features *'Mean fruit intensity'* and *'Mean high-density tissue intensity'* in contrast to the two-feature *'Conference' based SVM* that used the features *'Std tissue intensity'* and *'Std low-density tissue intensity'*.

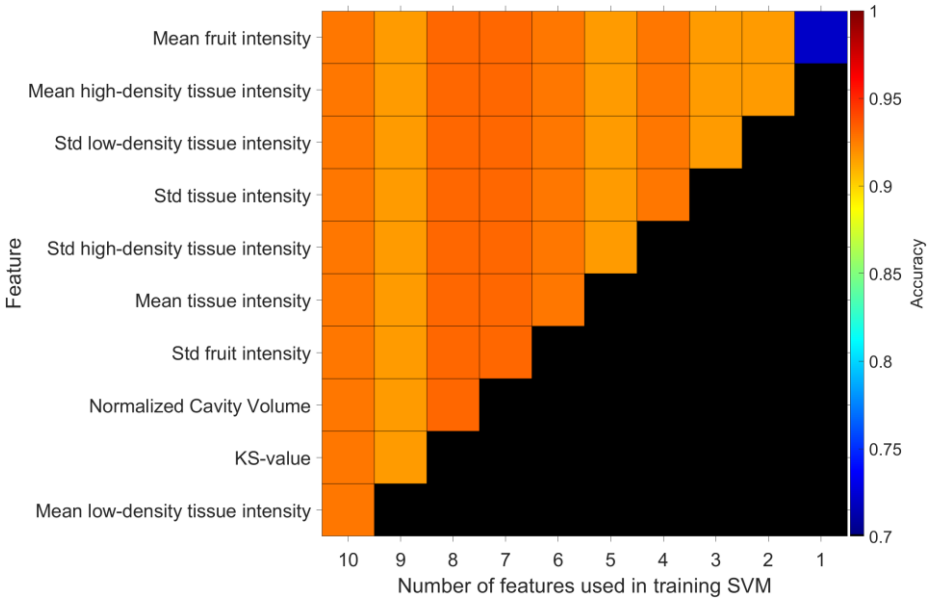

Figure 3.9: Plot of features used for retraining each SVM on the combined dataset. Each column represents a classifier in a series of classifiers, in which the number of features allowed to be used decreased from the left (10 features) to the right (1 feature). Every column thus shows the features used by a classifier in the series, while each row shows in which classifiers a certain feature was used. A colored tile indicates that the feature was used, while a black tile indicates that a feature was eliminated for the specific SVM in the feature elimination procedure. The color of each column represents its obtained classification accuracy.

3.4 Discussion

3.4.1 Internal variability must be measured to separate '*defective*' from '*healthy*' pear fruit

Internal browning and cavity formation were introduced in 'Conference' and 'Cepuna' pears by exposing them to suboptimal storage treatments for six months. The internal disorders differed in severity, location and appearance. Internal browning was characterized by reduced voxel intensity in the CT reconstructions of the fruit due to reduction in tissue density associated with water loss in the affected regions. In regions with severe internal disorder development, cells broke down completely and cavities were observed. This is in line with observations for pear made by (Lammertyn et al., 2003b; Muziri et al., 2016; van Dael et al., 2017).

To further characterize *'defective'* and *'healthy'* pear fruit, features were extracted from the CT volumes and compared between 'Conference' and 'Cepuna' *'defective'* and *'healthy'* fruit (see Table 3.1). First, the features *'Std fruit intensity'*, *'Std tissue intensity'*, *'Std low-density tissue intensity*' and *'Std high-density tissue intensity*' seemed to be the most relevant ones for separating the classes *'defective'* and *'healthy'*, regardless of the fruit cultivar. The *'defective'* class had significantly higher values for these four features due to a wider range in voxel intensity and thus in tissue density. This suggests that when looking for features that separate the *'defective'* from *'healthy'* pear fruit, features that measure variability rather than absolute values will be more performant for classifying pear fruit regardless of their cultivar.

Second, for the 'Cepuna' cultivar a significantly lower *'Mean tissue intensity*' for both *'healthy'* and *'defective'* fruit was observed compared to 'Conference' fruit. This might indicate that on average, the 'Cepuna' fruit have a lower density, i.e. higher porosity, than 'Conference' fruit (Nugraha et al., 2019).

Third, other features like *'Mean fruit intensity*' and *'Mean low-density tissue intensity'* were significantly different between the classes for each cultivar, but no clear threshold can be indicated that works for both cultivars.

Fourth, the KS-value showed to be a good feature to separate *'healthy'* and *'defective'* 'Conference' fruit with higher KS-values for *'defective'* fruit. However, the KS-value was not significantly different for both classes of 'Cepuna' fruit. Even an opposite, although not significant, trend was observed with slightly higher values for *'healthy'* fruit.

Finally, compared to the observed *'healthy'* 'Cepuna' fruit, the *'healthy'* 'Conference' fruit had significantly higher normalized cavity volumes. As such, relative to the total fruit volume, 'Conference' fruit might have larger cores than 'Cepuna' fruit.

One must be careful when generalizing these results because environmental factors potentially influencing the fruit characteristics were not investigated. Fruit characteristics and

susceptibility for internal disorders can be seasonal and location specific. Moreover, only fifteen *'healthy'* 'Cepuna' fruit where included in the dataset. Unfortunately, a large part of the 'Cepuna' fruit subjected tot the control treatment also developed internal disorders. Due to the small sample size of *'healthy'* 'Cepuna' fruit, the observed differences between classes and cultivars must thus be interpreted with caution.

3.4.2 X-ray CT and machine learning can be implemented inline to classify fruit reliably

The large variability in severity, location and appearance of the internal disorders makes it challenging to develop algorithms that detect *'defective'* fruit reliably. However, for the internal disorder detection in 'Conference' pears, a SVM achieved a classification accuracy of 92.2 % with false positive and false negative rates of respectively 2.4 and 11.7 % (see Table 3.2). Moreover, the number of features was reduced from ten to two while keeping the classification performance high by using the SVM RFE method (see Figure 3.7 and Figure 3.8). The classifier trained with the features *'Std tissue intensity'* and *'Std low-density tissue intensity'* still achieved an accuracy, false positive rate and false negative rate of respectively 91.2, 4.8 and 13.3 %.

Furthermore, without retraining or other adaptions to the method the *'Conference' based SVMs* performed excellent on the 'Cepuna' cultivar as well. An overall classification accuracy of 95.1 % with a false positive and a false negative rate of respectively 0.0 and 5.7 %, was achieved by the ten-feature *'Conference' based SVM* (see Table 3.3). Compared to the ten-feature classifier, the two-feature classifier scored the same accuracy with a better false negative rate of 4.6 %, but worse false positive rate of 6.7 %. This shows that the classifiers trained on the 'Conference' cultivar generalize well to the 'Cepuna' cultivar and suggests that the method can be used for other pear cultivars too without much effort. However, an increase in generalizability by reducing the number of features used by the classifiers was not observed, as the performance of all classifiers using between ten and two features was very similar for both cultivars.

The *'Conference' based SVMs* were compared to two series retrained on the combined dataset. The first series was forced to use the same features as the *'Conference' based SVMs*, but was allowed to adapt the weights. In the second series, also the selected features were allowed to be altered by re-implementing the SVM RFE method. In both cases, the *'Conference' based SVMs* scored better, even though no 'Cepuna' data was included in the training process. The two-feature SVM of the second series trained on the combined dataset used the features *'Mean fruit intensity'* and *'Mean high-density tissue intensity'* in contrast to the two-feature *'Conference' based SVM* that used the features *'Std tissue intensity'* and *'Std low-density tissue intensity'*. Differences in performance and selected features are probably caused by the imbalance in the combined dataset. Only 28 % of the combined dataset was *'healthy'* as just fifteen out of the 102 'Cepuna' samples were *'healthy'*, compared to 42 out of the 102 'Conference' samples. The classification metrics for both two-features classifiers on the combined dataset, the 'Conference' subset and the 'Cepuna' subset are summarized in Table 3.5. The classifier retrained on the combined dataset scored very poorly on *'healthy'* 'Cepuna' fruit, with a 40.0 % false positive rate. Due to the small number of *'healthy'* 'Cepuna' samples, however, this only had a limited effect on the overall classification accuracy over the 'Cepuna' and combined datasets.

Table 3.5: Confusion matrix with average classification results on the combined dataset, 'Conference' dataset and 'Cepuna' dataset of the two-features 'Conference' based SVM vs the two-feature classifier that was retrained on the combined dataset with the features 'Mean fruit intensity' and 'Mean high-density tissue intensity'.

Combined dataset		Predicted			
		'Conference' based SVM		Retrained SVM	
		Defective	Healthy	Defective	Healthy
Ground truth	Defective	91.8 %	8.2 %	92.5 %	7.5 %
	Healthy	3.5 %	96.5 %	12.3 %	87.7 %
	Overall classification accuracy:	93.1 %		91.2 %	

'Conference' subset		Predicted			
		'Conference' based SVM		Retrained SVM	
		Defective	Healthy	Defective	Healthy
Ground truth	Defective	86.7 %	13.3 %	85.0 %	15.0 %
	Healthy	4.8 %	95.2 %	4.8 %	95.2 %
	Overall classification accuracy:	91.2 %		89.2 %	

'Cepuna' subset		Predicted			
		'Conference' based SVM		Retrained SVM	
		Defective	Healthy	Defective	Healthy
Ground truth	Defective	95.3 %	4.7 %	97.7 %	2.3 %
	Healthy	6.7 %	93.7 %	40.0 %	60.0 %
	Overall classification accuracy:	95.1 %		92.2 %	

Overall low false positive rates by the *'Conference' based SVMs* were achieved, ranging between 0.0 and 6.7 % (see Table 3.2, Table 3.3, Table 3.4 and Table 3.5). Low false positive rates ensure that the

number of *'healthy'* fruit that are falsely rejected, are minimized. Furthermore, there were false positives that did not have a pronounced internal disorder but showed small deviating characteristics. For instance, one rejected 'Conference' fruit had a relatively big open core which might be indeed disliked by some consumers (see Figure 3.10). Economically, it makes sense to minimize the false positive rate for this application, since during inspection the occurrence of internal disorders in a certain batch might be relatively low and the false negatives might not have severe defects. In contrast, with a high false positive rate to ensure a low false negative rate, the added benefit of detecting *'defective'* fruit might be offset by the falsely rejected *'healthy'* fruit. Of course, this depends on the severity of the internal disorders. To balance the compromise between the false positive and false negative rate in a desired way, one could set a different threshold for the decision boundary of the classifier instead of placing it at f(x) = 0. For 'Conference', the classifier had a true positive rate of 86.7 % with a false positive rate of 4.8 %. To have 0 % false positives, the true positive rate had to drop to around 82 %. To achieve a true positive rate of 100 %, the false positive rate had to be increased to 60 %.

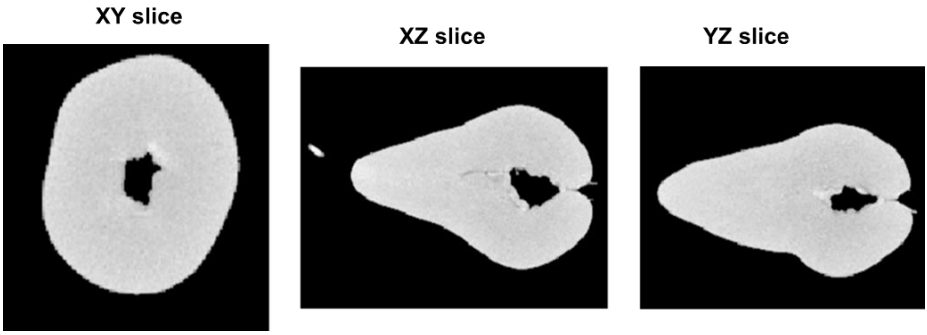

Figure 3.10: Orthogonal slices through the CT reconstructions of a false positive 'Conference' pear example with relatively big open core.

The *'Conference' based SVMs* had false negative rates ranging between 11.7 and 13.3 % for 'Conference' and between 4.7 and 5.7 % for 'Cepuna'. Compared to the false positive rates, the false negative rates are thus higher. As explained above, however, in this application it might be more important to reduce the false positive rate. Moreover, most of the false negative samples had only a very

small internal defects that may not even be noted by the end consumer. As an example, Figure 3.11 shows orthogonal slices through the CT reconstructions of a 'Conference' and 'Cepuna' pear. These examples make clear that for interpreting classification results, it is important to investigate how the data was labeled.

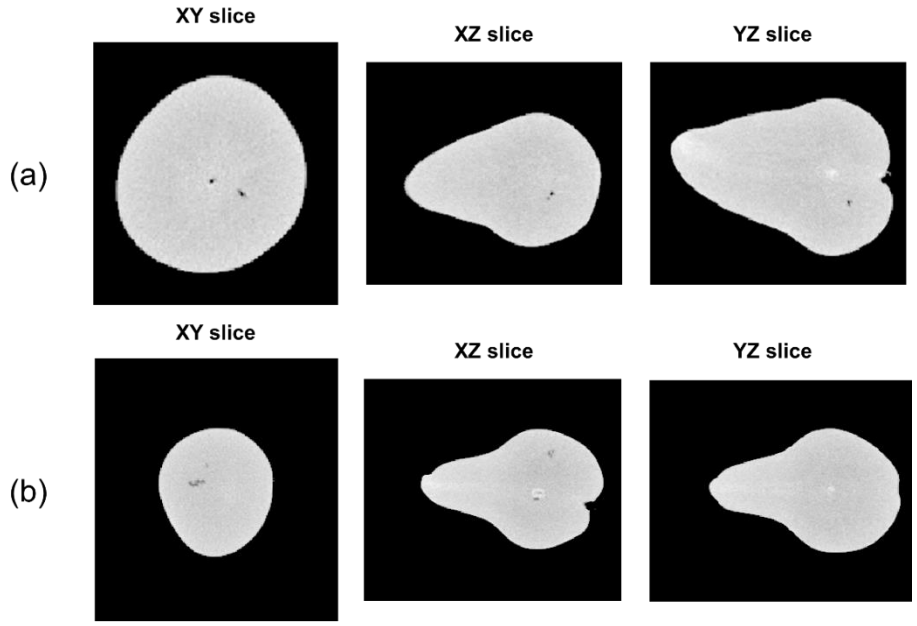

Figure 3.11: Orthogonal slices through the CT reconstructions of a false negative 'Conference' (a) and 'Cepuna' (b) pear examples.

This analysis showed that the inspection of the internal quality of pear fruit can be done in 3D using X-ray CT at high classification accuracies and low false positive rates. In the current experiment, the achieved inspection speed (1.5 minute for scanning, stacking and reconstruction, plus 2.3 s for feature extraction and classification) was not yet compatible with the commercial speed of existing sorting lines, requiring at least 10 fruit per second per lane. Nonetheless, it should be noted that both the experimental CT setup and the feature extraction algorithm were not optimized for reducing the runtime. With advanced reconstruction algorithms available for translational X-ray CT (De Schryver et al., 2016; Janssens et al., 2018), prototypes of inline CT systems can be developed. Speed improvements can be made using a dedicated

system combined with a trained reconstruction algorithm that needs far less projections (Beister et al., 2012; De Schryver et al., 2016; Janssens et al., 2018; Willemink et al., 2013). Moreover, a stacked scan was needed due to the relatively small detector size, requiring two full 360° sample rotations, intermediate height adjustments and a stacking procedure. With a detector of appropriate size, the scanning time can thus already be reduced by 50 %. The method could also be tested with other scanning settings, like exposure time and number of projections to reduce the scanning time, or pixel and voxel sizes to reduce the computational cost during reconstruction and analysis. Additionally, the feature extraction algorithm could be optimized and it was shown that the number of features can be reduced to further reduce the processing time. In terms of hardware, technical challenges must be overcome to transition from prototype systems to fast inline CT systems, *e.g.,* the development sample holders that stabilize and rotate the sample while translating at adequate speed. For industrial application continuing developments in both hardware and software are thus needed to increase the inspection speed and reduce the equipment costs. In addition, the internal quality inspection system can be used on a limited part of the supply in which a high occurrence of internal disorders is expected. In this case, lower inspection speeds might be adequate.

The proposed method uses machine learning in which a classifier is trained on a training set in feature space. However, the features were still hand-crafted and thus possibly suboptimal and application specific. Application on other fruit products, like apples, might require other features to be used making feature determination more difficult. In the feature extraction algorithm, the KS-value was calculated as a measure of similarity between the low intensity (low density) and high intensity (high density) tissue regions. It was expected to be an important feature for separating the classes, but only low corresponding weights were given to this feature by the classifiers. This illustrates that intuitive hand-crafted features that seem smartly designed, may not always be the best choice, and presents a limitation of machine learning. In deep learning, valuable representations of the data are learned and extracted by the model

itself. Hand-crafted features must thus no longer be engineered (Goodfellow et al., 2016). Deep learning on X-ray imaging is increasingly adopted, e.g., in medical applications (Lee et al., 2017b; Shen et al., 2017), and might thus be considered as an alternative method in future research (see Chapter 4).

3.5 Conclusion

A combination of machine learning and X-ray computed tomography was proposed to detect 'Conference' and 'Cepuna' pear fruit with a wide range in internal disorder severity automatically. Trained SVMs achieved good classification accuracies ranging between 90.2 and 95.1 % depending on the cultivar and number of features that were used. Moreover, low false positive rates were obtained, ranging between 0.0 and 6.7 %, while the false negative rates, ranging between 5.7 and 13.3 %, were higher. While the method could detect most defect pears, there is this room for improvement. Classifiers trained on 'Conference' data achieved high validation scores on the 'Cepuna' cultivar suggesting generalizability to other cultivars as well.

With continuing developments in both hardware and software to increase inspection speed and to reduce the equipment costs, the method could be implemented in inline X-ray CT for industrial application. In addition, the methods could be used by researchers investigating internal disorder development to screen for fruit with internal disorders. Further research could test the robustness of the method with faster inline CT image acquisition with reconstructions of potentially lower quality as a compromise.

The proposed method required a hand-crafted feature extraction algorithm. However, potentially better features that were not conceived remained unexplored. In addition, the feature extraction algorithm is possibly application specific. Furthermore, while the method allowed for classifying defect and sound fruit, it could not quantify the severity of the disorders which may be of high importance for consumers and thus for making decisions on discarding fruit or not. The direct quantification of disorder severities, which is tackled in Chapter 4, is thus expected to improve classification performance.

Chapter 4

Nondestructive Quantification of Internal Disorders in Pears using X-ray CT and Deep Learning[2]

4.1 Introduction

A promising technology for nondestructive internal quality inspection is X-ray computed tomography (CT), since spatial information on density differences can be provided in 3D (Chigwaya et al., 2018; Diels et al., 2017; Herremans et al., 2013; Herremans, Verboven, et al., 2014; Lammertyn et al., 2003b, 2003a). Significant progress has been made towards inline X-ray CT applications for food inspection (De Schryver et al., 2016; Janssens et al., 2016, 2018, 2019; L. F. A. Pereira et al., 2016, 2017), with one of the main challenges remaining automatic image analysis. The previous chapter has shown that automated feature extraction on X-ray CT data of pear fruit followed by a classification by means of a support vector machine (SVM), is effective in classifying fruit based on the presence of internal disorders. However, this approach did not allow to quantify the severity of the internal disorders. Moreover, the used features had to be hand-crafted and finding the right features required considerable trial and error. Automatic and accurate segmentation of the internal disorders would enable the

[2] This chapter is based on: Van De Looverbosch, Tim, Ellen Raeymaekers, Pieter Verboven, Jan Sijbers, and Bart Nicolaï. "Non-Destructive Internal Disorder Detection of Conference Pears by Semantic Segmentation of X-Ray CT Scans Using Deep Learning." *Expert Systems with Applications* 176 (2021)

quantification of their severity and potentially improve classification results. However, due to the large biological variability, the development of such algorithms using classical image processing and machine learning approaches is challenging. Fortunately, deep learning has recently become a viable tool for pattern recognition and image interpretation tasks and was found to be successful in tasks using medical X-ray CT data (Goodfellow et al., 2016; LeCun et al., 2015; Litjens et al., 2017; Shen et al., 2017).

The aim of this study is two-fold. First, a nondestructive quality inspection method is developed to quantify internal disorders in pear fruit based on X-ray CT scans. Hereto, a deep neural network for semantic segmentation (U-net) is trained to indicate various structures and tissues, including internal disorders, in the X-ray CT scans. Second, it is investigated if the quantitative data can be used to accurately classify the fruit, on the one hand, in "consumable" or "non-consumable" categories and, on the other hand, into "healthy", "defect but consumable" or "non-consumable" categories. Herein, ground truth classifications were obtained by a survey with images of cut fruit, in which participants indicated if they could spot internal disorders ("defect") or not ("healthy") and whether they would eat the fruit ("consumable") or not ("non-consumable").

4.2 Materials and Methods

The materials and methods are discussed in detail in the following sections. In Figure 4.1, a flowchart is presented to provide a general overview of the materials and methods.

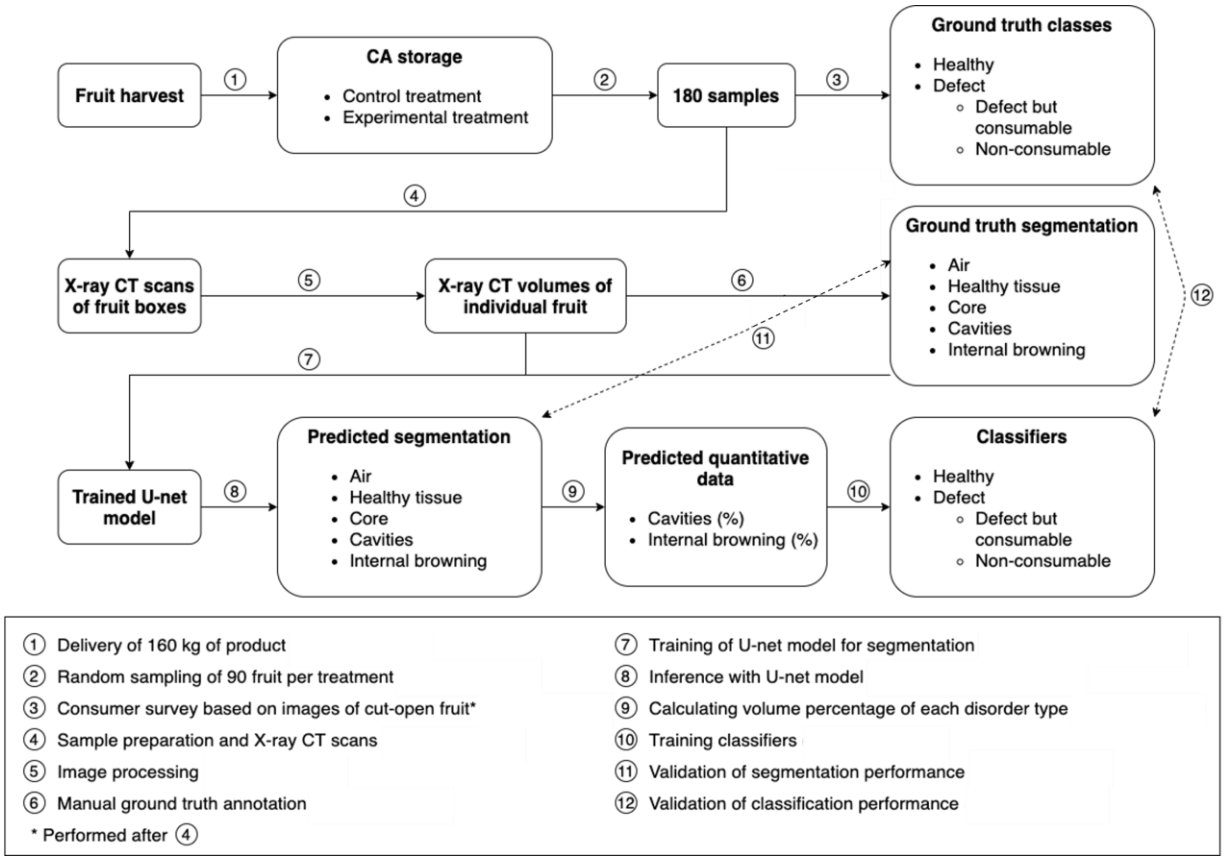

Figure 4.1: Method flowchart as a general overview of the materials and methods.

4.2.1 Pear fruit and storage protocols

A Flemish grower (50°53'06.4" N 5°01'32.2" E, Kortenaken, Belgium) harvested and delivered 160 kg of Conference pears on August 24, 2018. The fruit were randomly divided in two groups and put in long term storage for eight months in controlled atmosphere (CA) storage facilities of the Flanders Centre of Postharvest Technology (VCBT, Leuven, Belgium). For the first 25 days both groups were put at -1 °C in normal atmosphere to acclimatize to the cold temperature, after which different CA conditions for each group were implemented in CA containers with a volume of 1 m^3. The first group received the control optimal storage treatment following the recommendations of the VCBT for commercial sale, with a temperature, O_2 partial pressure and CO_2 partial pressure of -1 °C, 2.5 kPa and 0.7 kPa, respectively (VCBT, 2017). For the second group, experimental conditions were implemented to induce internal disorder development with a temperature, O_2 partial pressure and CO_2 partial pressure of respectively -1 °C, 1.0 kPa and 4.0 kPa. The fruit were removed from CA storage on May 9, 2019. 90 fruit of each treatment were selected randomly and stored at -1 °C in normal atmosphere before further measurements.

4.2.2 X-ray CT scans and data pre-processing

Of each treatment, 90 samples were randomly selected. The fruit were then divided into groups of thirty over six plastic boxes in which they were separated by styrofoam grids. The grids divided each box into two layers of fifteen samples and facilitated image processing afterwards by guaranteeing fruit were not touching each other. Next, the boxes were carefully brought to the UZ Leuven hospital in Leuven (Belgium) for obtaining X-ray CT images with a gantry CT system (SOMATOM Definition Flash, Siemens, Germany). Three boxes could be imaged at once. The system operated at 100 kVp with a voxel size of 0.9766 x 0.9766 x 0.3000 mm^3.

Next, CT data were processed in MATLAB (MATLAB, 2019b, 2019a). Figure 4.2 shows the X-ray CT scan of one box (Figure 4.2a, b and c) and illustrates some of the performed image processing steps. First, the volumes were resampled to a 0.9766 x 0.9766 x 0.9766 mm^3 voxel size. Individual fruit were then cropped (Figure 4.2d)

automatically and identified based on their position to link their CT images to their reference images (see section 2.3). The background of the CT volumes was removed with a global Otsu threshold (Figure 4.2e) (Otsu, 1979) and intensity values of the CT volumes were scaled between [0, 1] by dividing the intensities by the maximum value encountered over all volumes. Since fruit position was not fixed in the boxes, pears in the 3D volumes had to be aligned to the same pose. Therefore, each sample was centralized and rotated so that their principal axis aligned with the Z-axis and the fruit was in upwards orientation. To ensure upwards orientation, the fruit were rotated 180° over the x-axis if the center of mass was in the top half of the volume (Figure 4.2f). Finally, all volumes were padded with background voxels to the maximum size in the X-, Y- and Z-dimension encountered over all 180 volumes, resulting in identical volumes of 184 x 179 x 198 voxels.

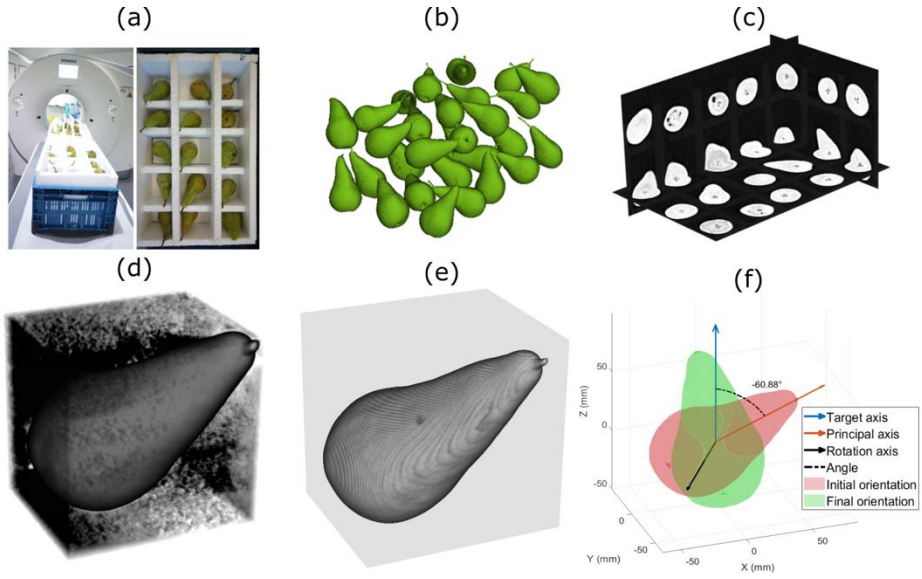

Figure 4.2: X-ray CT scans using a gantry CT system and data pre-processing. (a) Boxes of samples on the CT system on the left and top view of the styrofoam grid on the right; (b) Volumetric representation of one of the boxes; (c) Orthogonal slices through the CT volume of the box in (b); (d) Cropped CT volume containing one sample; (e) Cropped CT volume containing the sample in (d) after removal of the background noise; (f) Illustration of the alignment of the pear volumes with the Z-axis.

4.2.3 Reference images and ground truth classification

Following the CT scans, the fruit were consecutively cut open and reference images of the fruit flesh were taken. Multiple slices were cut perpendicular to the style-calyx axis and spread out below an RGB-camera (see Figure 4.3). Afterwards, an online survey was organized with the reference images in which the participants were asked to indicate on the one hand whether they found the fruit to be healthy or defective and on the other hand if they considered the fruit consumable or not. The survey was sent out to everyone working for the research group and completed by 17 participants. A majority vote was implemented over all participants to decide on the ground truth classification of each fruit ("healthy" *vs.* "defect" and "consumable" *vs.* "non-consumable"). Herein, the strength of the inter-annotator agreement, as measured by Fleiss' kappa statistic, was found to be moderate (0.52) (Fleiss, 1971; Landis & Koch, 1977). This resulted in a dataset of 128, 26 and 26 samples labelled as "healthy", "defect but consumable" and "non-consumable", respectively.

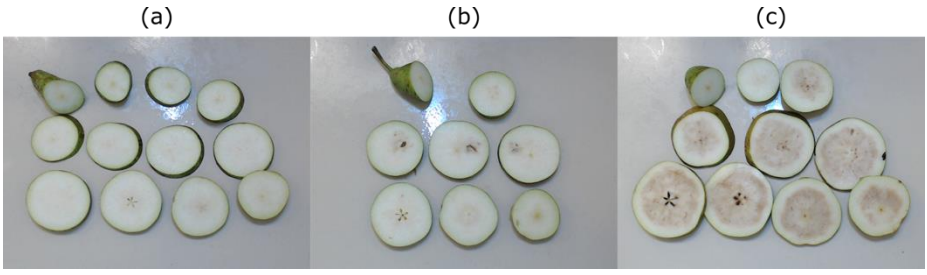

Figure 4.3: Reference images of respectively a "healthy" (a), a "defect but consumable" (b) and a "non-consumable" (c) sample.

4.2.4 Manual ground truth annotation

CT volumes of 51 out of the 180 pears were manually annotated using the free and open-source 3D Slicer software (*3D Slicer*, 2020, p. 3; Fedorov et al., 2012). Prior to manual annotation of each sample, the cut open reference image of each fruit was examined for the presence and severity of internal disorders. In the segmentation software, voxels were given one of five different classes: "external air", "healthy tissue", "core", "cavity" and "internal browning". First,

"external air" was indicated by applying a global Otsu-threshold and preserving the largest island, i.e., the largest connected component. Secondly, the other volumes with a value below the Otsu-threshold inside the fruit were assigned to the "core" or "cavity" class. Hereby, the "core" island was manually indicated. Third, between 5 and 10 seed points were manually placed in regions of "healthy tissue" and regions affected by "internal browning". Then, the region growing algorithm in 3D Slicer ("Grow from seeds") was used to complete the annotation. Herein, the other classes were fixed. Finally, a smoothing operation was done on the "healthy tissue" and "internal browning" regions using the "Joint smoothing" method in 3D slicer with a smoothing factor of 0.50. Figure 4.4 shows the steps in the manual annotation procedure of the CT volumes.

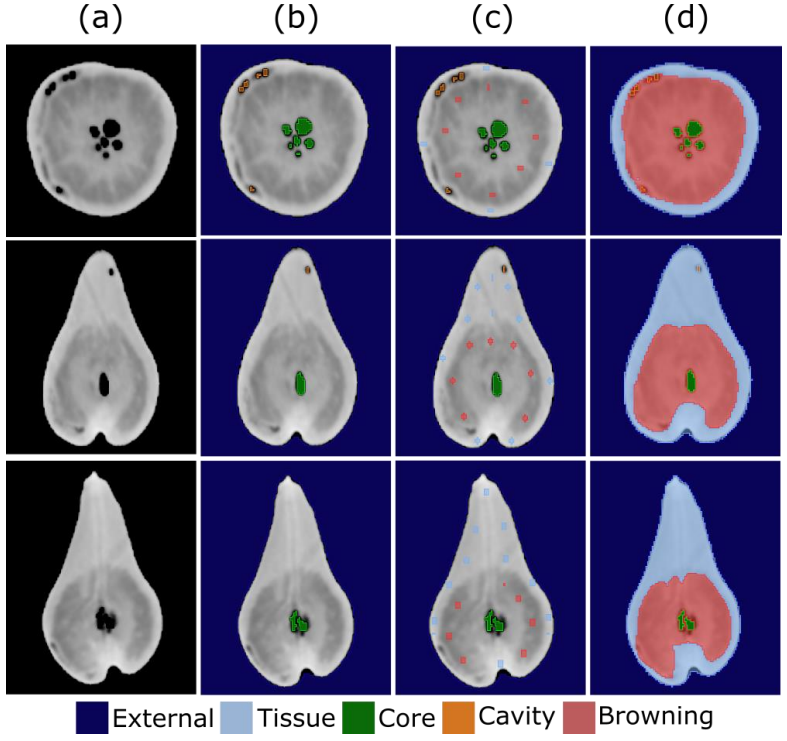

Figure 4.4: Steps in the manual annotation of CT volumes. (a) Initial CT volume after background noise removal with global threshold; (b) CT volume with external (dark blue), core (green) and cavity (orange) voxel indicated; (c) Addition of seed points of healthy tissue (light blue) and browning (red) voxels for subsequent region growing step; (d) Final result of the annotation procedure with healthy tissue (light blue) and browning (red) voxels resulting from region growing using the seeds in (c).

4.2.5 Dataset generation

The YZ slices (179 x 198 pixels) of the pre-processed CT and corresponding manually annotated volumes were used to train the U-net segmentation model. For the input of the model, three consecutive slices were concatenated in the third dimension, resulting in an image with three channels. As target, the manually annotated slice of the middle of the three input slices was used. The slices adjacent to the middle input slice were thus used to provide contextual information. The slices of the 51 annotated volumes were divided in training, validation and test sets. Hereof, slices with only external air were omitted since they are not of interest. The slices of 36 samples were used for training and validation, including 14 "healthy", 10 "defect but consumable" and 12 "non-consumable" samples. 10 % of the slices were used for validation during the training process. The training and validation set included in total 2172 and 241 slices, respectively. Finally, the test set included the slices of 5 "healthy", 5 "defect but consumable" and 5 "non-consumable" samples with 2760 slices in total.

4.2.6 Network architecture and training for segmentation

A U-Net based network of four downsampling and four upsampling blocks was adapted to learn the multi-class semantic segmentation of regions of interest ("external air", "healthy tissue", "core", "cavity" and "internal browning" pixels) in the YZ X-ray CT slices (Milesial, 2019; Ronneberger et al., 2015). The model architecture is presented in Figure 4.5. The input layer had three channels of 179 x 198 pixels. After each 3x3 convolution, batch normalization and the ReLU activation function were applied (Ioffe & Szegedy, 2015; Nair & Hinton, 2010). In the downsampling steps, max pooling with a 2x2 kernel was used. Upsampling was done using a scale factor of 2 and bilinear interpolation. Skip connections were used to concatenate the output of each downsampling block to the output of the corresponding upsampling block. Herein, the output of the upsampling block was padded with zeros to match the height and width of the output of the downsampling block. The output layer had five channels of the same size, one for each class. The segmented

image (only one channel) was then obtained by assigning each pixel to the class with the largest probability in the output layer.

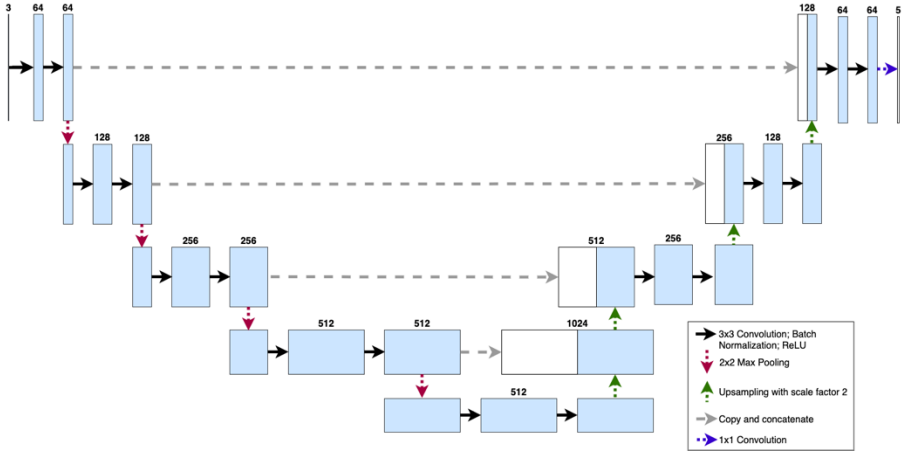

Figure 4.5: U-net model architecture. Blue boxes correspond to multi-channel feature maps, while white boxes correspond to copied feature maps. The number of channels of each feature map is indicated above each box. Adapted from (Ronneberger et al., 2015).

The model was implemented in python using the PyTorch framework version 1.4.0 (Paszke et al., 2019) and trained from scratch for 20 epochs, i.e., for 20 passes over the training set. As loss function, i.e., the function used to calculate the error between the model's prediction and the ground truth, the class-weighted cross-entropy loss function was used to account for class imbalance in the data (Sudre et al., 2017). The used class-weights were determined based on the ground truth data in the training set. Hereto, the differences between the total number of voxels and the number of voxels belonging to each class were first calculated. Then, the ratio of these differences and the total number of voxels was used as class-weight, resulting in the following weights: 0.08 ("external air"), 0.93 ("healthy tissue"), 1.00 ("core"), 1.00 ("cavity") and 0.99 ("internal browning"). The performance on the validation set was tested after every epoch and was used to optimize the batch size and initial learning rate hyper-parameters. The batch size is the number of samples the model handles during one iteration. During training, the weights of the model were updated after each iteration proportional to the loss on each batch. The magnitude of the weight changes depended on the learning rate, which started from the initial

learning rate set by the operator. Here, the Adam optimizer was used to dynamically change the learning rate during training (Kingma & Ba, 2014). The effect of data augmentation in the form of affine image transformations (random mirroring, rotating, shearing and scaling) on the training data was tested.

4.2.7 Segmentation performance validation

After model training was finished, the segmentation performance was validated by comparing the predictions with the ground truth annotation using the intersection over union metric (IoU, also known as the Jaccard index) (Rezatofighi et al., 2019). The IoU is a common metric to evaluate the correspondence between two shapes and has a range between 0 (no correspondence between prediction and ground truth) and 1 (full correspondence between prediction and ground truth). It was calculated for each label on slice level for the test set. The IoU for a certain label was not calculated for slices in which the label was not present in the prediction or ground truth. The mean IoU was calculated per pear and the overall mean and box plot for the IoU of each label were calculated. Finally, the mean and box plots were also calculated for the slices of "healthy", "defect but consumable" and "non-consumable" samples, separately.

4.2.8 Classification

Quantitative data obtained from the output of the trained segmentation model was used to classify the fruit. Therefore, all consecutive slices of all 180 samples were segmented by the U-net model and the percentages of the fruit volumes corresponding to "cavity" and "internal browning" were calculated. The same calculations were performed on the manual ground truth segmentations. First, the accuracy of the predicted percentage of internal browning was investigated. Hereto, bins were made of 2 % in size, resulting in the bins [0.0-2.0 %, 2.0-4.0 %, …, 98-100 %] and it was tested if the prediction fell into the same bin as the ground truth. Second, a binary classifier, a logistic regression model (Kutner et al., 2005), was fitted using a 5-fold cross-validation to separate "consumable" and "non-consumable" samples based on the quantitative data of cavity and brown volume fraction. Finally, a

multiclass classifier, a quadratic discriminant model (Tharwat, 2016), was fitted on the same features using a 5-fold cross-validation to separate "healthy", "defect but consumable" and "non-consumable" samples. The classification results for both classifiers are presented using confusion matrices and receiver operating characteristic (ROC) curves (Fawcett, 2006; Metz, 1978).

4.3 Results

4.3.1 Segmentation model training

The best results were obtained without data augmentation and with a batch size and initial learning rate of 4 and 1.0^{-4}, respectively. Final training and validation losses of, respectively, 0.004 and 0.011 were achieved. The training and validation loss during training is presented in Figure 4.6. Interestingly, data augmentation did not improve the results (final validation loss of 0.016). The remainder of this section therefore discusses the results obtained without data augmentation.

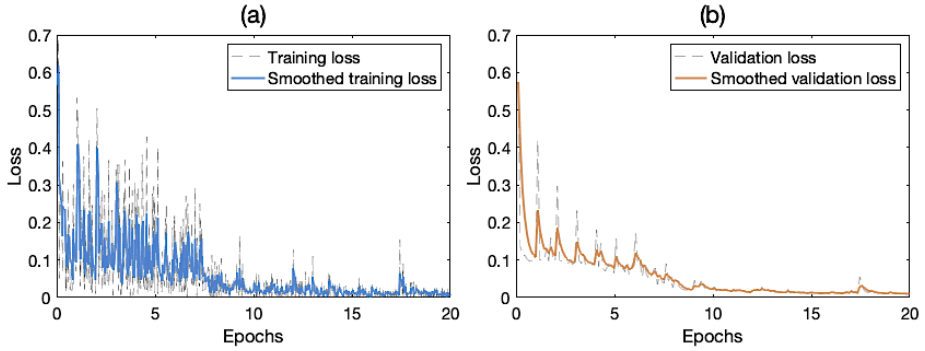

Figure 4.6: Training and validation loss during training. (a) Training loss (gray dashed line) and smoothed training loss (blue solid line) during training with one data point per batch; (b) Validation loss (gray dashed line) and smoothed validation loss (orange solid line) during training with one data point per 60 batches of the training data. Smoothing was done using the exponentially weighted moving average filter with a weight of 0.6.

Figure 4.7 shows a CT slice, its ground truth segmentation, its predicted segmentation and reference image of five samples of the test set. In the first row, a slice of a "defective" pear is shown in which the core and a large cavity were connected, i.e., the core and cavity were indistinguishable. Therefore, no core was labelled during the ground truth annotation, while only cavity is indicated in the

prediction. The second row shows a slice of a "healthy" pear that was correctly segmented. On the third row, the results on a slice of a "defect" pear with cavities and browning is shown. The ground truth and prediction correspond quite well. Finally, in the fourth row a slice of another "defect" pear is shown with good similarity between the ground truth and prediction.

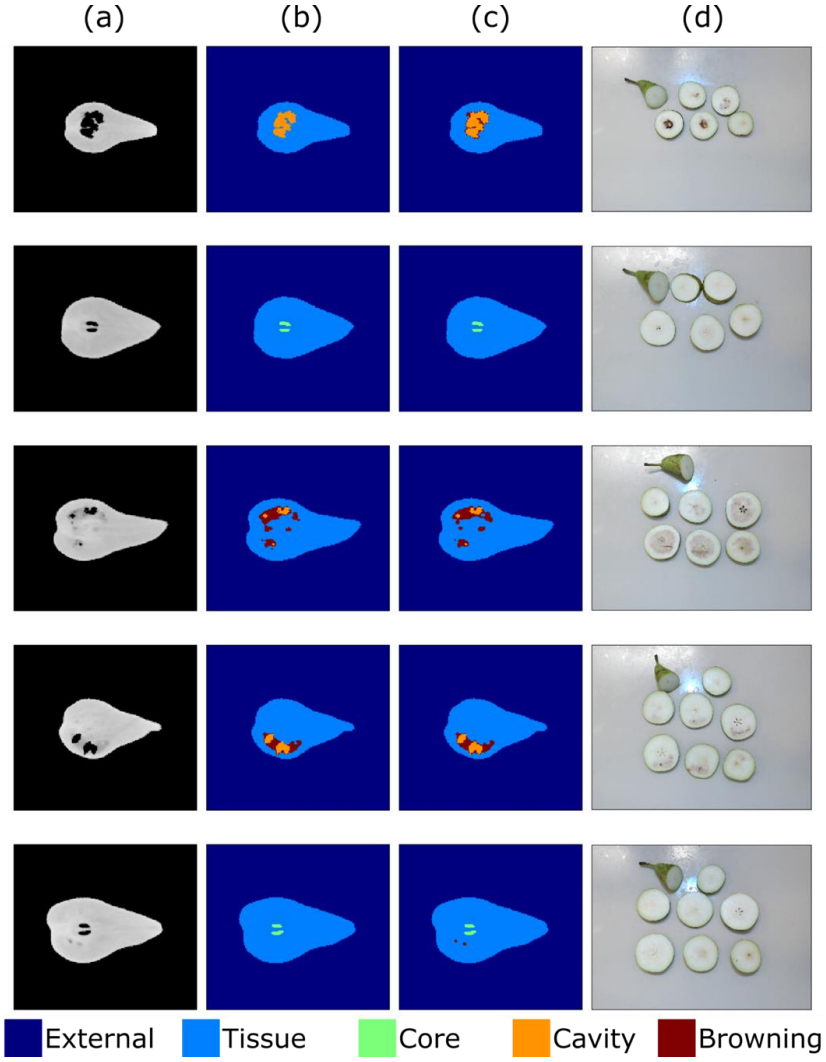

Figure 4.7: Examples of visual segmentation results on the test set. Each row represents another pear. (a) XY slice of the original CT scan; (b) the corresponding ground truth segmentation; (c) the corresponding predicted segmentation; (d) the reference image of the corresponding pear.

4.3.2 Segmentation performance validation

In Figure 4.8, the boxplots of the mean IoU scores, calculated per pear in the test set, are shown for the different label categories. For "external", "healthy tissue" and "core" high mean IoU scores were obtained, meaning there is a good agreement between the ground truth and predicted segmentation. For "core", also outliers with a lower score are present. The median IoU for "cavity" is still quite high (0.95), but the interquartile range is greater, meaning there is a larger spread in the scores. For "internal browning", a median and mean IoU of 0.0 and 0.2 were found. The low median and mean IoU originate from the fact that 8 out of 15 pears of the test set got a mean IoU score of 0.0. In case of IoU score of 0.0, no overlap was found between the ground truth and the prediction. It can also indicate the presence of a certain label in the prediction, which is not present in the ground truth (or *vice versa*).

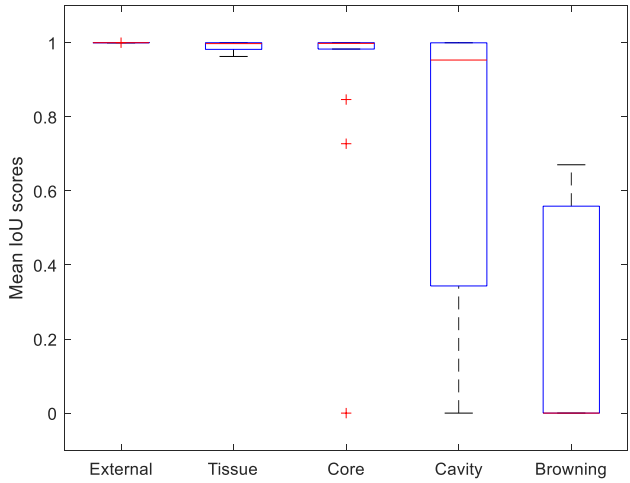

Figure 4.8: Boxplots of the mean IoU scores per pear and per category of the test set. The red line represents the median and the bottom and top line of the blue boxes represent the 25th and 75th percentiles, respectively. The black dashed lines represent the whiskers which extend until the minimum and maximum data point, that are not considered as outliers. Individual outliers are represented as the red plus symbols.

Figure 4.9 shows IoU scores calculated on individual slices of the three different categories of pear ("healthy", "defect but consumable" and "non-consumable") for the test set. For "internal browning", the IoU scores are equal to 0.0 for both the categories "healthy" and "defect but consumable": there is actually no or very little browning present in these pears. For the "non-consumable" category, however, the median is equal to 0.61 and the IoU scores are overall higher. This shows that due to the low IoU scores of the first two categories the overall IoU score of "internal browning" is low. In the plot the first category, high IoU scores are present for "external", "healthy tissue" and "core", which is desirable since this category contains healthy fruit. There is, however, also an IoU score of 0.0 for "cavity" present, meaning an incorrect prediction of cavity. For the other categories, also the first three labels exhibit high scores. For "cavity", the median is high, but outliers are present with lower scores.

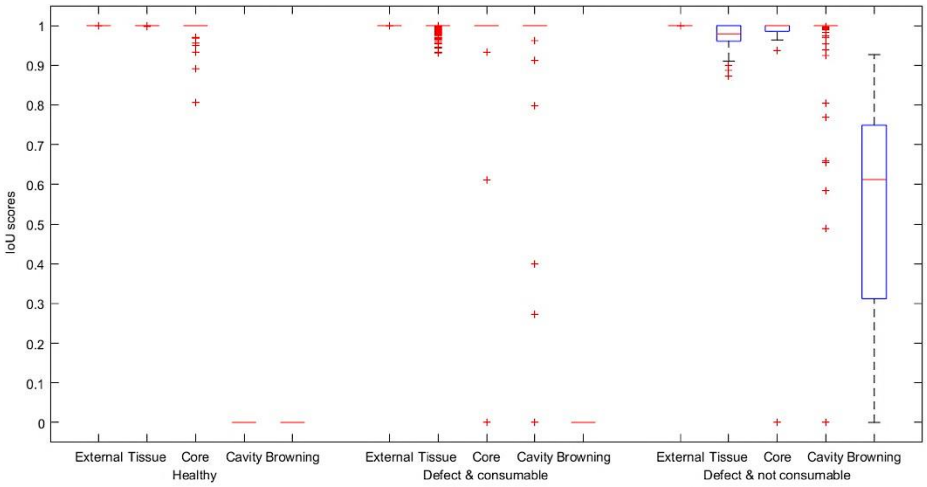

Figure 4.9: Boxplots of the IoU scores per category of the test set. The red line represents the median and the bottom and top line of the blue boxes represent the 25th and 75th percentiles, respectively. The black dashed lines represent the whiskers which extend until the minimum and maximum data point, that are not considered as outliers. Individual outliers are represented as the red plus symbols.

4.3.3 Classification

Based on the majority vote on survey results, all 180 samples were assigned to a ground truth category. 128, 26 and 26 samples were found to be "healthy", "defect but consumable" and "non-consumable", respectively. Thereafter, the trained U-net model was used to segment all slices of each fruit and the percentage of "cavity" and "internal browning" was calculated. Subsequently, the predicted percentages of "internal browning" were compared to the percentages obtained from the ground truth segmentation of the test set by testing if the predicted percentages fell into the same bins as the ground truth ([0.0-2.0 %, 2.0-4.0 %, …, 98-100 %]) (see section 2.8). Figure 4.10 shows a histogram of the predicted and ground truth percentages of "internal browning" for the test set. The prediction of only two samples of the test set was put in the wrong, but adjacent, bin. Note that the test set included samples with internal browning up to around 15 %. However, it was found that for the training set, with internal browning percentages going up to around 45 %, the predicted and ground truth bin were in disagreement for only three samples (results not shown).

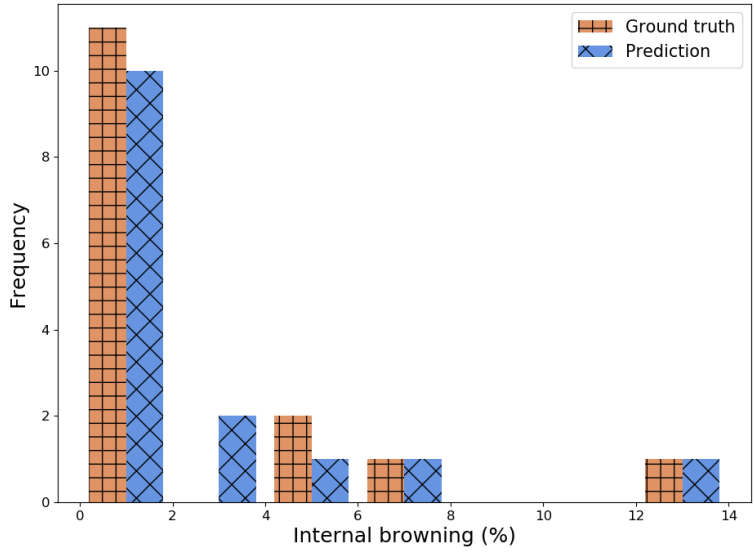

Figure 4.10: Histogram of internal browning percentage in bins of 2 % for the ground truth and as predicted by the model for the test set.

Thereafter, a binary ("healthy" and "defective") and a multiclass ("healthy", "defect but consumable" and "non-consumable") classifier were fitted to the predicted percentages of "internal browning" and "cavity" data of all 180 fruit in a 5-fold cross-validation. Figure 4.11 shows the scatterplots of the dataset with ground truth categories for the binary and multiclass classification.

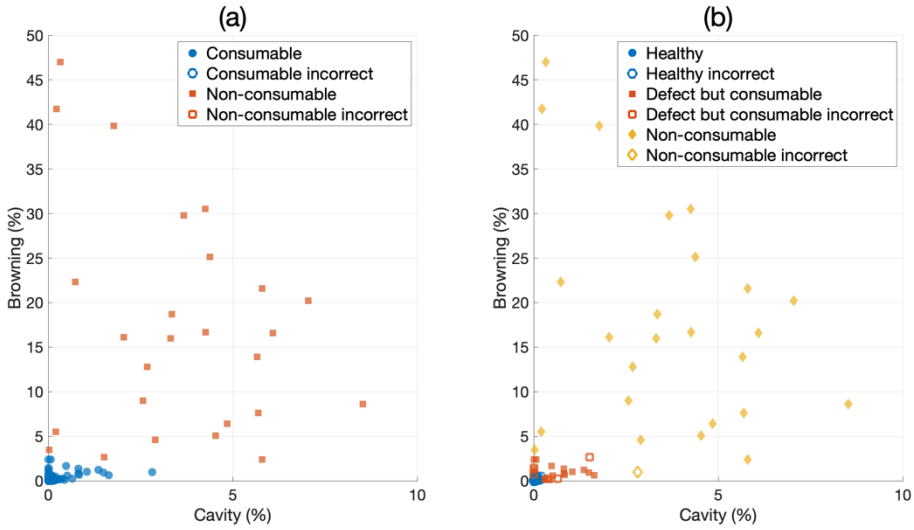

Figure 4.11: Scatterplots of the classification dataset divided in two (a) and three (b) classes. Colors and shape of the marker indicate the category of the samples. Filled markers indicate correct predictions, while open markers indicate incorrect predictions.

The confusion matrices of the binary (logistic regression model) and multiclass classifiers (quadratic discriminant model) are shown in Table 4.1 and

Table 4.2. The binary classifier achieved an overall accuracy, true positive and true negative rate of 99.4, 96.0 and 100.0 % in classifying "consumable" and "non-consumable" samples. The multiclass classifier achieved an overall accuracy of 92.2 % and true positive rates of 97.0, 65.0 and 96.0 % for "healthy", "defect & consumable" and "defect & non-consumable", respectively. While high accuracies were thus obtained for the "healthy" and "defect & non-consumable" categories, a substantial part of the "defect & consumable" category was misclassified as "healthy".

Table 4.1: Confusion matrix of the binary classifier.

		Predicted	
		Consumable	Non-consumable
Ground truth	Consumable	100.0 %	0.0 %
	Non-consumable	4.0 %	96.0 %
	Overall accuracy	99.4 %	

Table 4.2: Confusion matrix of the multiclass classifier.

		Predicted		
		Healthy	Defect but consumable	Non-consumable
Ground truth	Healthy	97.0 %	3.0 %	0.0 %
	Defect but consumable	31.0 %	65.0 %	4.0 %
	Non-consumable	0.0 %	4.0 %	96.0 %
	Overall accuracy	92.2 %		

Since the binary classifier obtained a true positive and false positive rate of respectively 96.0 and 0.0 %, no ROC-curve is shown for this classifier (area under the curve (AUC) = 1). Figure 4.12 shows the ROC curves of the multiclass classifier. Herein, the ROC-curves are presented per category using a one-vs-all approach, *i.e.,* the ROC-curves are shown for each category with the respective category as the positive class and the others combined as the negative class. The

AUC were equal to 0.93, 0.87 and 1.00 for respectively, “healthy”, “defect & consumable” and “defect & non-consumable” as positive class.

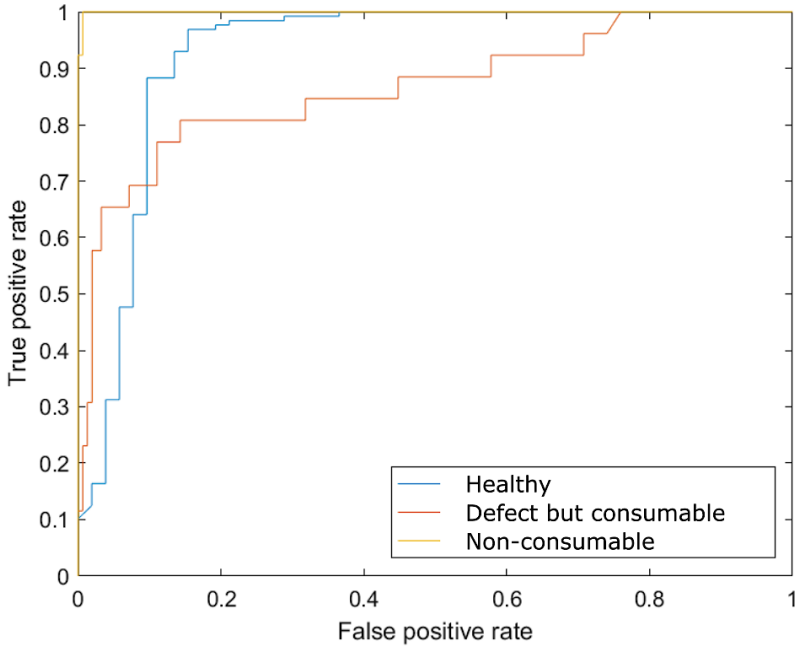

Figure 4.12: ROC-curve of the multiclass classifier.

4.4 Discussion

4.4.1 An efficient and reproducable 3D manual annotation procedure

The manual annotation of 3D data is relatively difficult, especially when it can only be performed on 2D slices. Therefore, we developed a reproducible procedure that uses 3D operations minimally dependent on subjective factors. The threshold to separate voxels comprising air and tissue was determined objectively using Otsu’s method (Otsu, 1979).

The external air, cavities and core could also be objectively determined for most samples. However, some pears were found “difficult” to annotate when cavities were connected with the core of the fruit (see Figure 4.13). During the ground truth annotation, the

connected core and cavity were respectively labelled as “core” and “cavity” when they could easily be distinguished, e.g., only a small hole connecting both areas. However, in case the core and cavity could not be distinguished anymore, the whole opening was labelled as “cavity”. It was found that in these cases, the deep neural network labelled some voxels of the internal space as “core”, even though no clear boundary could be identified between the “core” and “cavity”. Of course, the way of annotating these samples had an impact on the segmentation metrics for the “cavity” and “core” label.

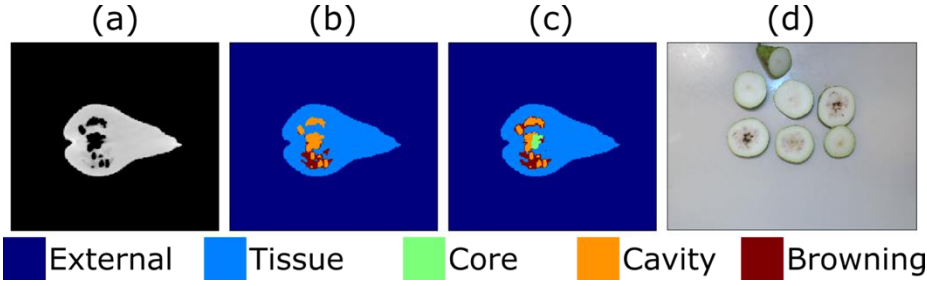

Figure 4.13: Example of a sample with a connected core and cavity. Since no clear boundary could be identified between the two, the whole internal space was labelled as “cavity” during the ground truth annotation. In the predicted image, however, the trained model indicated some pixels as core.

The annotation of “tissue” and “internal browning” was relatively fast using the semi-automatic seeded region growing algorithm (see Figure 4.4b and c). The placements and number of the initial seeds, however, was done subjectively and different seed points could possibly cause different results. Alternatively, the use of a predetermined threshold to segment these regions would make the process easier. However, this was impossible due to the variability in intensity values of the healthy tissue and internal browning between different samples. Lammertyn et al. (2003b) used minimally different thresholds on each pear to separate the tissue and internal browning segments, which is, however, a subjective and equally time-consuming approach. Our annotation process was done by one person. Preferably, this should be done by multiple operators. Although reference images of all fruit were present to guide the operator, this could prevent an operator-specific annotation and improve the performance of the segmentation model.

4.4.2 A segmentation model to quantify the internal disorder severity

A deep neural network was trained to automatically segment different tissue types and structures in slices of X-ray CT scans of pear fruit. Data augmentation in the form of affine transformations on the training data did not improve on the results obtained without data augmentation. In general, data augmentation is useful for preventing neural networks from overfitting on the training data. However, overfitting on the training data was not observed for the model trained without data augmentation (see Figure 4.6). The effect of data augmentation might thus be insignificant. Additionally, the nature of our dataset possibly makes data augmentation less effective. In the preprocessing of the data, all CT reconstructions were put in the same pose (see section 2.2). The preprocessing, which can also be done on new data, makes sure all samples are presented to the model in the same manner. Images resulting from data augmentation might therefore not be encountered in the actual dataset. The model would then generalize to irrelevant cases. It was therefore decided to use the model trained without data augmentation.

Visually, the predicted segmentations of many samples corresponded well with the ground truth annotations (see section 3.1). A validation of the segmentation performance using the IoU metric on an independent test set (see section 3.2) showed high scores for the labels “external”, “healthy tissue” and “core”. A high median IoU was also obtained for the “cavity” label, but a larger range was found including lower values. Interestingly, low IoU scores were found for the “internal browning” label, even though visually most predictions seemed sufficiently accurate. It turned out this was mainly caused by errors on small volumes and volume edges. Since the IoU metric is relative to the ground truth, the absolute size of the volumes did not matter, resulting in low IoU scores even though the error by the model was rather negligible.

Particularly, it was observed that many low IoU scores for “internal browning” were caused by the incorrect labelling of small pixel regions, or regions forming a boundary around the core or cavities

in the predicted image (see the first row in Figure 4.7). These misclassified regions resulted in an IoU of 0.0 if no "internal browning" was present in the ground truth. Typically, these misclassified regions had lower grey scale intensity values similar to the intensity of regions affected by internal browning. The latter had lower intensity values due to a lower density and higher porosity (Nugraha et al., 2019). However, for the experienced eye, supported by the reference images, it was clear that internal browning had often not occurred. Rather, the partial volume effect of the CT imaging was probably responsible for the local decrease in pixel intensity (Barrett & Keat, 2004). Due to the limited resolution of the CT scans (0.9766 x 0.9766 x 0.9766 mm^3 voxel size), voxels in the transition between two regions with different X-ray attenuations (e.g., different densities) have an intensity somewhere in-between. The partial volume effect is thus present at the boundary of the core or cavities, resulting in the misclassification of the affected pixels by the network. Moreover, some cavities were too small to be thresholded in the CT scans and were therefore not annotated as "cavity" during ground truth annotation. Nonetheless, these small cavities caused small spots with lower intensity due to the partial volume effect and were often indicated as "internal browning" in the predicted image (see the last row in Figure 4.7). Even though low IoU scores were obtained for "internal browning", it was shown that the severity of the internal disorders could be accurately predicted, and the misclassified voxels had a relatively small impact on the subsequent classification.

An interesting case was found in which a cavity was predicted to be the core of the fruit (see Figure 4.14). Indeed, when looking at the CT slice in Figure 4.14a, the shape and location of the cavity in the slice resembles the typical shape and location of the core. From examining the CT volume, however, it clearly was a cavity and was labeled as such during ground truth annotation (see Figure 4.14b). Since the model does its predictions slice-by-slice, it lacks the 3D information required to correctly recognize the whole as a cavity. A 3D implementation of the U-Net model could be beneficial to avoid such mistakes (Çiçek et al., 2016). However, to train such model more labelled CT-scans would be required.

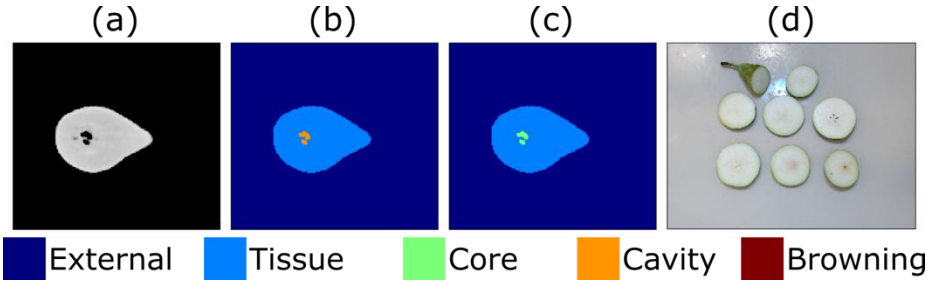

Figure 4.14: Example of a sample where a cavity was predicted as a core due to a similar location and shape of the internal space.

Several strategies can be followed to improve the segmentation results and robustness. First, as stated before, a 3D implementation of the segmentation model could be applied so that the model can use 3D information to prevent mistakes as presented in Figure 4.14. Second, more samples could be manually annotated, including more "difficult" samples. Here, 51 out of the 180 available samples were annotated with a preference for "easier" samples, i.e., with cavities not connected to the core. To segment "difficult" samples, one must attempt to design an objective method. Third, data augmentation, using other techniques than affine transformations, could be applied on the CT scans and/or CT slices to artificially increase the size and variability of the dataset. Finally, additional datasets could be included of fruit from different years, origins and maturity stages.

In addition, the evaluation of the segmentation performance could be improved. For evaluating the performance of segmenting fine structures, Multiscale IoU could be used (Ahmadzadeh et al., 2021). Metrics such as the region-wise over- and under- segmentation measures could be used in conjunction with IoU to evaluate the agreement between the prediction and ground truth in terms of overlap and number of connected regions (Y. Zhang et al., 2021). These metrics could be incorporated in the loss function for training the model.

4.4.3 Classification based on quantitative data

The trained deep neural network was used to obtain accurate predictions on the severity of internal disorders, expressed in percentages of the fruit volume affected by a disorder (see Figure

4.10). Based on the quantitative data, classifiers were trained to separate “consumable” and “non-consumable” fruit on the one hand, and “healthy”, “defect but consumable” and “non-consumable” fruit on the other hand. Excellent classification performance was achieved for the former classifier, with an overall accuracy, false positive and false negative rate of 99.4, 0.0 and 4.0 % (see Table 4.1). “Non-consumable” fruit can thus be reliably separated from sound fruit using our method. For the multiclass classifier, great performance was achieved for “healthy” and “non-consumable” samples, but a true positive rate of only 65.0 % was achieved for the “defect but consumable” category (see

Table 4.2). It was found that 89.0 % of the misclassified samples of this category were predicted to be “healthy”. This could be caused by similar percentages of “cavity” and “internal browning” as “healthy” samples, e.g., only very small internal disorders. This reasoning was confirmed, as all misclassified samples were located in the bottom left corner of Figure 4.11. Additionally, possibly an insufficient segmentation of the defective regions might have occurred, causing an underestimation of the actual disorder severity. Moreover, it should be noted that the alignment step of the pears during pre-processing included interpolation and resampling procedures that might have altered the appearance of the CT scan. However, the majority of this category could be identified correctly, and misclassifications might not be a big concern since the fruit was still considered consumable.

It was found that some internal defects were visible in the reference images of the cut-open fruit, but not in the CT volumes. Recent bruises, for instance, might manifest themselves in the reference image as regions with free water in the intercellular space. Since the water has not evaporated yet, there might not yet be a difference in density and thus X-ray attenuation. An alternative explanation might be that the bruises only occurred after the CT scans were acquired during transport in the period before the reference images were taken. These defective regions were, therefore, visible in the reference images but could not be segmented in the prediction or the ground truth volumes and were, hence, not quantified. Figure 4.15 shows the reference image and orthogonal slices of a bruised

fruit categorized as “defect but consumable”. Contrary, some disorders might have been invisible in the reference images due to the positions where the cuts were made, resulting in an incorrect categorization.

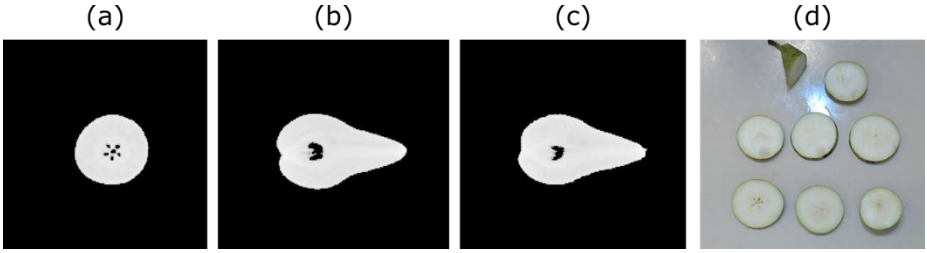

Figure 4.15: Orthogonal slices through the CT volume (a-c) and reference image (d) of a bruised sample classified as “defect but consumable”. While the bruise is visible in the reference image, it is invisible in the CT scan.

Deep learning-based classifiers that are trained end-to-end (from input to output) are black-box systems by nature. Here, we proposed a two-step approach, in which a semantic segmentation model is used to deliver quantitative data based on which the fruit are subsequently classified. The classification is thus not trained end-to-end, which brings a few benefits. First, this approach is transparent, because the output of the segmentation model that precedes the classification can be visualized and expressed quantitatively. Second, the method can be easily adapted for different markets, as it would only require retraining the classifier on categories specific to the different markets. The segmentation can be assumed independent of the market and retraining the segmentation model would thus not be required. End-to-end trained classifiers, on the other hand, would require retraining for each market. In future work, an end-to-end system could be investigated in which a trained 3D segmentation model is used as an initialization of a model with the same architecture, followed by a small convolutional neural network for final classification. To adapt the model to a certain market, the segmentation layers could be frozen, while the classification layers are retrained with market specific categories. Alternatively, the segmentation and classification could be learned simultaneously, using multi-task learning (Ruder, 2017).

4.4.4 Potential applications

Automatic segmentation of CT-scans can significantly speed up the analysis of internal defects in foods. Various promising applications are conceivable. First, the method could be applied to non-destructively inspect the internal quality foods in a quantitative way and to accurately predict the severity of internal defects. This would not only allow separating non-consumable from consumable products, but also to separate consumable products of different internal quality. Since the absence of internal defects can be guaranteed, higher margins are justifiable. However, significant improvements are needed in hardware and software to implement automated X-ray CT analysis at the throughput of commercial sorting lines (e.g. at least ten products per second for fruit sorting). In this regard, research is done on obtaining high quality reconstructions with a minimal number of projections and inline X-ray CT approaches (De Schryver et al., 2016; Janssens et al., 2016, 2018, 2019; L. F. A. Pereira et al., 2016, 2017). Second, this method might facilitate more fundamental research on internal defect development on a larger scale, since cumbersome manual analysis of large datasets can be automated. A larger quantity of samples could be included, or the same samples could also be analysed at different time-points during storage with short intervals for CT scanning. Finally, the method could be adapted for foreign object detection in foods (Edwards, 2004; Graves et al., 1998).

4.5 Conclusion

A non-destructive inspection method was developed to quantify internal disorders in pear fruit in X-ray CT scans using a deep neural network for semantic segmentation. Herein, a model was trained to indicate healthy tissue, core and regions affected by internal disorders, i.e., cavity formation and internal browning, in slices of the scans. The severity of internal disorders could be predicted successfully from the segmentations. Moreover, it was shown that the resulting quantitative data can be used to classify “consumable” *vs* “non-consumable” fruit at high accuracy (99.4 %) on the one hand and “healthy” *vs* “defect but consumable” *vs* “non-consumable” classification on the other hand (92.2 %). For the latter, the

classification performance on the “defect but consumable” fruit was poor (true positive rate of 65.0 %), with most misclassified fruit assigned to the “healthy” class.

Being able to reliably separate fruit from the three classes would facilitate different market strategies and could justify larger margins on the “healthy” category since high internal quality is guaranteed. While the presented method could reliably prevent non-consumable fruit from reaching consumers, additional research is required to improve the separation of top quality from acceptable fruit. In addition, a large scale consumer survey is recommended to better understand consumer tolerances with respect to internal defects.

The X-ray CT-based method presented in this chapter can quantify the internal disorder severity of pears, which can then be used to obtain a high accuracy in internal disorder detection. The presented method might be of great interest to researchers and industry working on quality assurance, product analysis and foreign object detection in foods and other industries. However, an inline implementation at speeds of commercial sorting lines is not yet possible with current hardware.

Chapter 5

Inline Nondestructive Internal Disorder Detection in Pear Fruit using Explainable Deep Anomaly Detection on X-ray images

5.1 Introduction

For inline applications using X-ray imaging, X-ray radiography is currently the most straightforward to implement inline with an X-ray source and line detector on either side of a conveyor belt. X-ray radiography has been studied for internal disorder inspection of fruit using application specific algorithms (Casasent et al., 1998; Kim & Schatzki, 2000; M. A. Shahin et al., 2001; van Dael et al., 2016). Van Dael et al. (2019, 2017) developed a more general purpose method for internal quality inspection. However, it still required application specific shape and density distribution models (DDM). Moreover, the method is relatively hard to implement due to its complexity. Hence, there is a need for a generally applicable method for nondestructive internal disorder detection that can be easily implemented inline.

Over the past decade, deep learning has shown to be very useful in pattern recognition tasks and does not require hand-crafted features or sophisticated image processing pipelines. Instead, each task can be learned end-to-end from data (LeCun et al., 2015). It has been successfully used in many fields, including food and agriculture (Kamilaris & Prenafeta-Boldú, 2018; Zhou et al., 2019). Most

commonly, deep learning is used in a supervised way, i.e., using labeled data. However, acquiring a labeled and balanced dataset of sufficient size is often hard in practice due to high labor or data acquisition costs, or other practical limitations (Goodfellow et al., 2016; Kamilaris & Prenafeta-Boldú, 2018; LeCun et al., 2015; Zhou et al., 2019). In the case of internal disorder detection in pome fruit, it is challenging to acquire sufficient defect samples for several reasons. First, storage conditions are optimized to preserve high quality fruit. Defect fruit are, therefore, exceptions, and their availability is unpredictable. Second, while treatments are available to induce internal disorder development during storage, their success is variable, and the required duration of the storage is a disadvantage. Moreover, it is unclear whether the induced defects are representative of all defect fruit encountered in practice. Third, regardless of the origin of the defect fruit, it is hard to obtain a dataset with a wide range in disorder severity. Fourth, since internal disorders are invisible externally, destructive methods are required to perform ground truth annotations which complicates the tracking of the disorder development over time and the prediction of the utility of the sample in the first place. Finally, it is uncertain whether a model that is trained on a certain dataset is transferable to new data. In contrast, healthy fruit are abundant, immediately available after harvest, and can be easily obtained from various locations and consecutive harvest years.

Therefore, it seems promising to approach the detection of defect fruit as an anomaly detection problem (Chandola et al., 2009). In anomaly detection (AD), a model is constructed using normal data and a certain metric is used as anomaly score to measure the extent to which a new sample deviates from normality. Anomaly detection has been extensively investigated using conventional machine learning techniques for a wide range of applications, but had limited success in high dimensional spaces, e.g., image data. Recently, deep learning-based anomaly detection (deep AD) methods have been developed showing promising results on image datasets (Chalapathy & Chawla, 2019; Ruff, Kauffmann, et al., 2021). Deep AD is mostly seen as an unsupervised learning task in which unlabeled data is used that is assumed to be normal. Alternatively, it can be

done in a semi-supervised setting in which few labeled anomalies are used during training. Recently, a method called outlier exposure (OE) was introduced for semi-supervised AD in the absence of real labeled anomalies for training (Hendrycks et al., 2019). Herein, random natural images from an auxiliary dataset were used as labeled anomalies. It has been shown that OE can improve the performance of AD even though these labeled anomalies are unrelated to the first dataset. In this case, the OE images are referred to as "out-of-class" anomalies, i.e., the anomaly images may contain a totally different content than the normal class of images (Hendrycks et al., 2019). However, in practice one typically also wants to detect anomalies with subtle deviations instead of totally different images. In the context of AD detection in manufactured goods, Liznerski et al. (2021) introduced the usage of synthetic anomalies in which nominal images were subtly distorted using "confetti noise" after which they were used during training as labeled anomalies. Inspired by their method, we developed a pipeline to create synthetic anomalies of X-ray radiographs of pears. We compare this to using the ImageNet dataset as a general auxiliary dataset for OE, and to using "confetti noise".

In this chapter, the application of deep AD is targeted to detect pear fruit with internal disorders using X-ray imaging. Unsupervised (Bergmann, Fauser, et al., 2019; Bergmann, Löwe, et al., 2019) and (semi-)supervised deep AD methods (Liznerski et al., 2021; Ruff, Vandermeulen, et al., 2021) are compared to the multisensor internal disorder detection method (van Dael et al., 2019, 2017) on a dataset of simulated X-ray radiographs. Herein, also the explainability of each method was evaluated using saliency maps, or heatmaps, that highlight the anomalous regions.

5.2 Materials and methods

Three deep AD methods are benchmarked against the multisensor inspection method (van Dael et al., 2019, 2017) for detecting internal disorders in inline X-ray radiographs of pear fruit. Hereto, a simulated dataset of radiography images was generated from X-ray CT scans of pears to allow for a controlled environment and for making abstraction of some technical implementation challenges. In

addition, by simulating the data, a much larger dataset could be generated compared to acquiring it physically, and we prevent biases that could otherwise have been introduced unintentionally during image acquisition. Section 5.2.1 describes the simulated dataset. The different deep AD methods are discussed in Section 5.2.2. The multisensor inspection AD method is explained in Section 5.2.3. Training details are discussed for the deep AD methods in Section 5.2.5. Finally, Section 5.2.6 discusses how the AD methods were evaluated.

5.2.1 Simulated Radiography Datasets

A labeled dataset of inline X-ray projection images of pear fruit was simulated from the CT dataset of 180 samples that was acquired in Chapter 4 (see sections 4.2.2 and 4.2.4). Herein, all samples were labeled "Healthy", "Defect but consumable" or "Defect and non-consumable" (see Figure 5.1). In addition, predictions of the internal disorder severity of each sample, i.e., volume percentages of cavities and internal browning, were available from the trained segmentation model developed in Chapter 4. Here, we defined anomalies as "Defect and non-consumable" fruit (#26), while "Healthy" fruit (#128) were considered nominal. The "Defect but consumable" fruit (#26) were not included in the anomaly dataset, since these samples had only minor defects which were not considered off-putting for consumption. For completeness, the performances of all methods were also evaluated on the "Defect but consumable" group. The simulation of X-ray radiographs is described in Appendix A1.1. The dataset contained 9000 simulated projections in total. Herein, 6400 were nominal, while 1300 were anomalies. The remaining 1300 images were simulated from the "Defect but consumable" class. Every image had a fixed size of 300×300 pixels. Note that for visualization purposes, the grayscale values in the projection were inverted compared to normal transmission images.

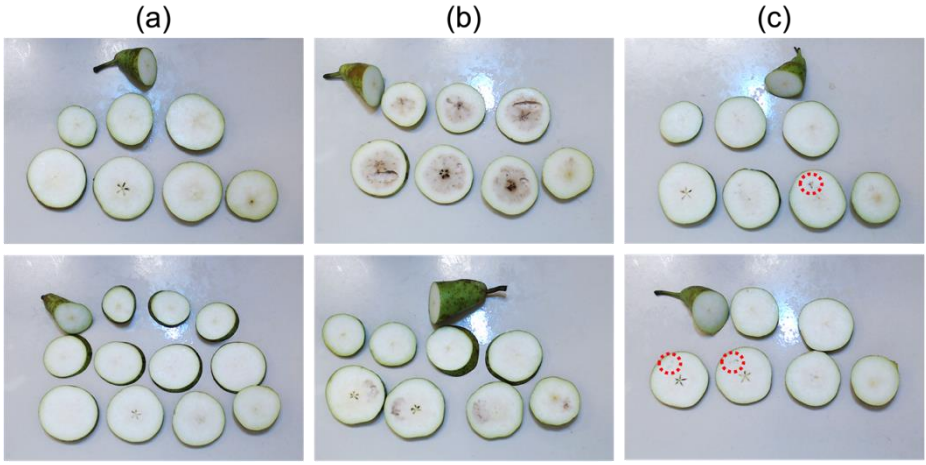

Figure 5.1: Classes in the CT dataset. Column (a) "Healthy", column (b) "Defect and non-consumable" and column (c) "Defect but consumable". Red circles in column (c) indicate minor defects.

5.2.2 Deep AD methods

5.2.2.1 Unsupervised deep AD using an autoencoder

AD is generally approached as an unsupervised learning problem where the goal is to model nominal data and detect samples that deviate significantly from normality. A common approach for deep AD is to train an autoencoder (AE) on nominal data with the task of reconstructing the input (Bergmann, Fauser, et al., 2019; Ruff, Kauffmann, et al., 2021). The architecture of the AE contains an encoder and a decoder connected by a shared latent representation of limited size. This representation acts as a bottleneck, forcing the encoding of only the most informative features in it. The output is then reconstructed from this latent representation by the decoder, after which the reconstruction error between the input and output is calculated. This builds on the assumption that the AE is worse at reconstructing anomalous compared to nominal samples, because it is only trained on nominal data. The reconstruction error can thus be used as an anomaly score. Moreover, the reconstruction error can be visualized as anomaly heatmap, e.g., the pixel-wise squared difference between the input and output, since anomalous regions are assumed to be reconstructed worse.

Here, we train an AE on a training dataset comprising only nominal data, and evaluate it on a test set. The test set comprises an equal number of nominal and anomalous samples. The full architecture for the AE is described in Appendix A4.1. As loss function, both the Mean Square Error (MSE) and Structural Similarity Index (SSIM) between the AE's input and output were tested (Z. Wang et al., 2004). Details on the model training are discussed in Section 5.2.5.

5.2.2.2 Fully Convolutional Data Description (FCDD)

Deep one-class classification methods learn to map nominal data close together in feature space, while mapping anomalies further away (Ruff, Kauffmann, et al., 2021). They can often be trained in a semi-supervised way, where a few example anomalies are available during training. For instance, the semi-supervised AD method Deep SAD (Ruff et al., 2020) maps nominal data close to the center of a hypersphere in feature space and maps anomalies further away from this center, i.e., hypersphere classification. However, in terms of explainability, Deep SAD does not naturally produce an anomaly heatmap. Therefore, a method called Fully Convolutional Data Description (FCDD) (Liznerski et al., 2021) was used to perform explainable deep AD. In FCDD, the output matrix of the model functions as a down-sampled anomaly heatmap. This is accomplished by using only convolutional and pooling layers in the model, limiting the receptive field of each output pixel and preserving spatial information. By up-sampling the anomaly heatmap, anomalous regions can be indicated in the input image. The same FCDD model architecture as the one described by Liznerksi et al. (2021) was used (see Figure 5.2). A detailed overview of all layers is provided in the Appendix A4.2.

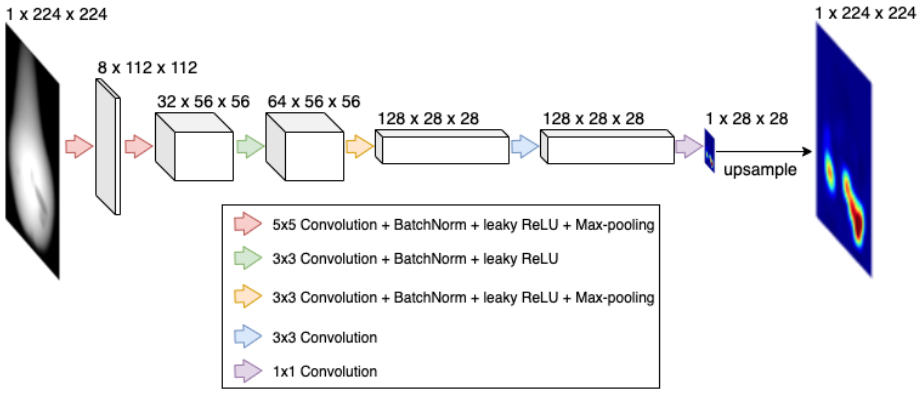

Figure 5.2: Schematic illustration of the FCDD model architecture.

FCDD is trained using nominal and anomaly samples $(X_1, \dots, X_n)$ that are labeled with $(y_1, \dots, y_n)$, where $y_i = 1$ and $y_i = 0$ represent the label of an anomaly and a nominal sample, respectively. The fully convolutional architecture performs the mapping with weights W of the input image X to a feature matrix, i.e., $\phi(X, W)$: $\mathbb{R}^{c \times h \times w} \to \mathbb{R}^{u \times v}$. Herein, c, h and w are respectively the number of channels, the height, and the width of the input image, and u and v are the height and width of the output matrix, respectively. It then utilizes the following objective function (Liznerski et al., 2021):

$$\min_W \quad \frac{1}{n}\sum_{i=1}^{n}(1-y_i)\frac{1}{u \cdot v}\|A(X_i)\|_1 - y_i \log\left(1-\exp\left(-\frac{1}{u \cdot v}\|A(X_i)\|_1\right)\right) \qquad \text{(Eq. 8)}$$

Here, $A(X) = \left(\sqrt{\phi(X,W)^2 + 1} - 1\right)$ is the pseudo-Huber loss function on the output matrix $\phi(X, W)$, with all operations applied element wise, and $\|A(X)\|_1$ is the sum of all elements of $A(X)$, which are all positive. The FCDD loss function minimizes $\|A(X)\|_1$ for nominal data, while maximizing it for anomaly data, which allows the use of $\|A(X)\|_1$ as anomaly score. In other words, output pixels that contribute to $\|A(X)\|_1$ are indicative for anomalous regions in the input image. Training was performed on the nominal training data combined with our OE pipeline to introduce synthetic anomalies (see Section 5.2.2.4).

5.2.2.3 Supervised deep AD

Supervised learning is often infeasible for AD due to limited availability of labeled anomalies, i.e., the dataset insufficiently captures what it means to be anomalous. Therefore, in classical AD methods (e.g., one-class SVM), a classifier is trained to discriminate between concentrated nominal data and unconcentrated anomalies which are assumed to be uniformly distributed (Steinwart et al., 2005). Yet, this approach is assumed to be ineffective for high dimensions because it would require sampling massive amounts of anomalies in feature space. However, it was recently shown that a pure classification-based AD method can work surprisingly well on images when trained using OE (Ruff, Vandermeulen, et al., 2021). The authors hypothesized that this is due to the multiscale structure of images that makes OE samples highly informative for AD. Therefore, a supervised classification-based AD method was evaluated on the simulated dataset. Herein, a ResNet18 model (He et al., 2015) was trained from scratch to classify nominal and anomalous samples using the binary cross-entropy loss function (BCE). This method is further denoted as the BCE classifier. Training was performed on the nominal training data combined with the same OE pipeline as used for FCDD to introduce synthetic anomalies (see section 5.2.2.4). Details on the model training are discussed in Section 5.2.5.

5.2.2.4 Outlier exposure

OE was applied during training for methods that allow for (semi-) supervised training, i.e., FCDD and the BCE classifier. During training, each nominal sample had a 50 % chance of being replaced by an OE image and being labeled as anomalous, resulting in balanced batches for large batch sizes. Hereto, a synthetic defect OE pipeline was developed in which images are being manipulated by adding synthetic defects. This resulted in a projection with blobs resembling real defects. The pipeline is illustrated in Figure 5.3 and discussed in detail in the Supplementary materials. The synthetic defect OE pipeline was compared to using the ImageNet dataset as a general auxiliary dataset for OE, and to using “confetti noise” as proposed by Liznerski et al. (2021) (see Appendix A3). Since the best

performance was achieved when using the synthetic defect OE pipeline, this approach was used in the rest of the paper. To evaluate the benefit of OE, the performance of models trained with 50, 40, 30, 20, 10 and 0 % chance for OE were compared (see section 5.3.2).

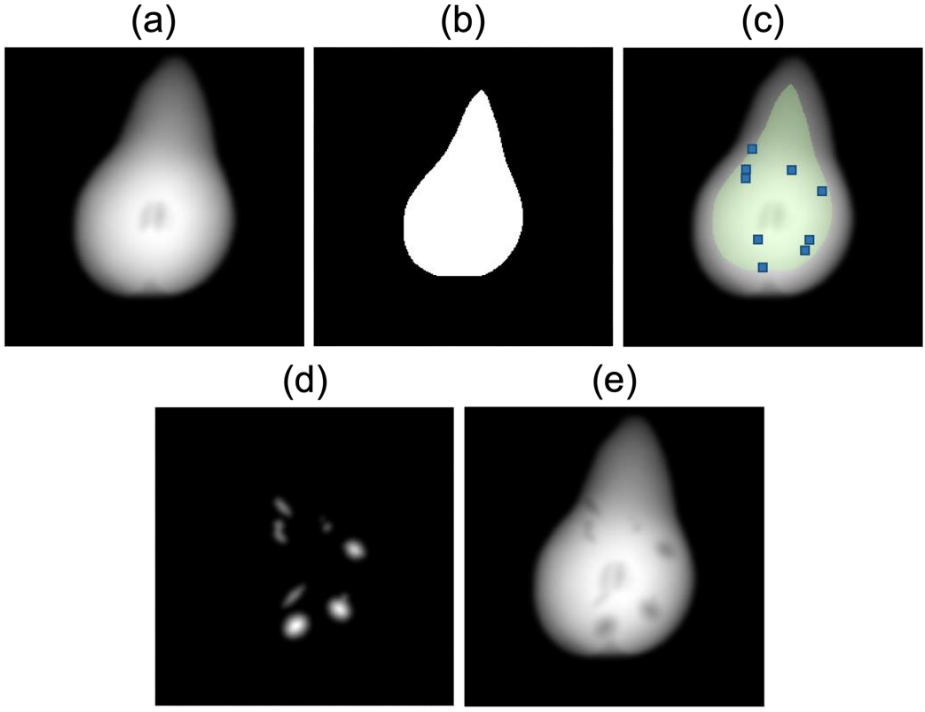

Figure 5.3: Outlier exposure pipeline with synthetic defects. (a) Nominal input image; (b) Area in which seed points for synthetic defects are sampled resulting from threshold and erosion on (a); (c) Sampled seed points (blue squares) in the area in (b), which is overlayed in green; (d) Ellipsoidal blobs resulting from a distance transform followed by a gaussian filter on deformed disks centered around the seed points in (c). Grayscale values are normalized between [0.0, 0.1]; (e) Input image with synthetic defects resulting from subtracting (d) from (e).

5.2.3 Multisensor AD Benchmark

As a benchmark for the deep AD methods, the multisensor AD method (van Dael et al., 2019, 2017) was used in a simulated setup. The method is based on the combination of 3D shape measurement and X-ray imaging and incorporates prior knowledge of the product in the analysis. The principle is that a statistical shape (SSM) and density distribution model (DDM) are first developed offline. In the

inline setup, the fruit is imaged with a 3D-imaging camera and an X-ray radiograph is acquired. The SSM and DDM are then fitted to the 3D point cloud after which a reference X-ray image can be computed. This reference image can be thought of as being the predicted X-ray projection which is expected in case no internal defects are present. Subtracting the reference from the measured radiograph results in a residual image on which deviant internal features light up. This enables the detection of any internal disorders associated with a structural change leading to density differences. A similarity measure between both images can therefore be used as anomaly score. In this work, the Mean Squared Error (MSE) between the reference image and the measured radiograph was used as anomaly score. Although the multisensor method does not require training like the deep learning methods, it does also need prior work and data to determine the SSM and DDM.

For the multisensor AD method, two corresponding reference images were simulated for each simulated inline scan (see Appendix A1.2), i.e., one using a homogeneous DDM (Multisensor HODDM) and another using a heterogeneous DDM (Multisensor HEDDM). The homogeneous DDM assumes no internal density gradients, while the heterogeneous DDM accounts for the overall density gradients in the fruit. The DDMs themselves were generated using the CT data of all healthy samples as described by van Dael et al. (2019, 2017). In Figure 5.4, orthogonal slices through the CT volume and fitted homogeneous and heterogenous DDMs are shown for a healthy fruit. For visualization purposes, grayscale values are plotted in the [0.5, 1] range using a colormap.

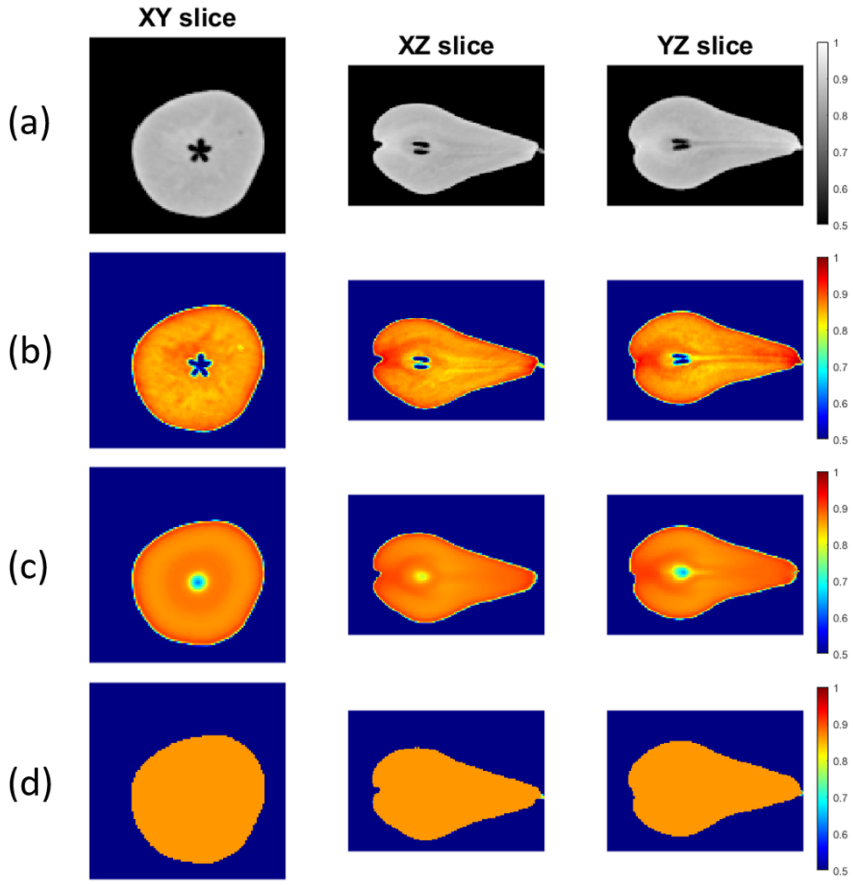

Figure 5.4: Orthogonal slices through the CT volume and fitted DDMs of a healthy fruit; (a-b) Orthogonal slices through the CT volume in grayscale (a) and with colormap visualization (b); Orthogonal slices through the fitted heterogeneous DDM (c) Orthogonal slices through the fitted homogeneous DDM (d). To clearly visualize the internal gradient, the grayscale values in these plots are scaled between [0.5, 1] and a colormap is used.

Figure 5.5 shows these simulated projections for one nominal and two anomalous samples. For the naked eye, using a homogeneous or heterogeneous DDM resulted in approximately the same reference image. However, in the grayscale line profiles (Figure 5.5 (e)) a slight drop in the grayscale values can be observed at the position of the core in the HEDDM reference image. Note that for the first sample, the DDMs can quite accurately predict the grayscale values along the line profile except for the area around the core. The same can be observed for the second sample, even though it has severe internal

browning. For the third sample, the grayscale values along the line profile deviate substantially from the predictions by the DDMs, i.e., lower grayscale values are found than expected.

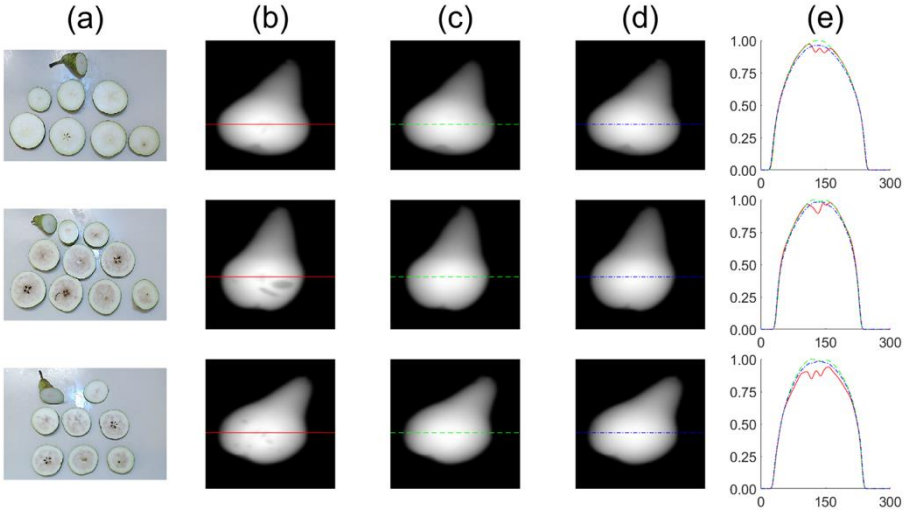

Figure 5.5: Image of cut open fruit (a), simulated projections (b-d) and grayscale line profiles through the core (e) from a nominal sample (1st row) and two anomalous samples (2nd and 3rd row). (b) Simulated projection with the CT volume; (c) Simulated projection with the heterogeneous DDM; (d) Simulated projection with the homogeneous DDM. (e) grayscale line profiles through the core in the images in red for (b), green for (c) and blue for (d).

5.2.4 Dataset splits

Training and test sets were created from the simulated dataset of 9000 samples, with for each sample the following simulated projections: a projection simulated with the CT volume, a projection simulated with the homogeneous DDM and a projection simulated with the heterogeneous DDM. All samples of the fruit labeled as anomaly, i.e., 1300 samples, were included in the test set together with an equal number of nominal samples. The remainder of nominal samples were used for training. Images of every fruit were included in only one of the two sets. To evaluate the effect of how the data was split, a 5-fold cross-validation was implemented. This resulted in five different splits in which the nominal samples were assigned to the training or test set at random while the anomalous samples in the test set were always the same. Thereafter, all

methods discussed above were evaluated on these five dataset splits. Note that only nominal samples were used for training and all available anomalous samples were used for testing. No validation set was used to maximize the number of anomalies available for testing. In addition, all hyperparameters were left to the defaults of the used libraries, except for the learning rates, which were optimized for each method based on the performance on the training set (see section 5.2.5).

5.2.5 Training details

All models were trained on the 5-fold cross-validation data splits for 50 epochs and a batch size of 256. An epoch was defined as a full iteration over all batches in the training set. For each of these data splits, a different seed was used for reproducibility, i.e., determining the random initialization of the models, sample order and composition of the batches, and data augmentation and OE outcomes. On every data split, all methods thus used the same random seed. The initial learning rate that resulted in the lowest final loss on the training set was used. For all models, this resulted in an initial learning rate of 1.0×10^{-4}. During training, the Adam optimization method ($\beta = (0.0, 0.999)$) was used in combination with a learning rate scheduler that reduced the learning rate by a factor 0.1 once a plateau was reached on the training loss (Kingma & Ba, 2014). Herein, a patience of five epochs was used and the learning rate was allowed to drop to a minimum of 1.0×10^{-8}. The data preprocessing consisted of a min-max normalization between $[0, 1]$. For training the AE, random cropping to an image size of 224×224 was used for data augmentation. For the FCDD and the BCE classifier models, the data augmentation during training consisted of random cropping to an image size of 224×224 and OE with the ImageNet OE pipeline or OE with synthetic anomalies using an OE rate of 50 %. In addition, the effect of OE was investigated on one of the dataset splits using a probability of 50, 40, 30, 20, 10 and 0 %.

5.2.6 Test and evaluation details

After training the models, all methods (including the multisensor AD method) were evaluated on the test sets of the 5-fold cross-

validation data splits. Every test image was min-max normalized between $[0, 1]$ and center cropped to a size of 224×224 pixels. A receiver operating characteristic (ROC) analysis (Fawcett, 2006) was performed on the anomaly scores of each method, which is a commonly used method for quantitative evaluation in AD (Ruff, Kauffmann, et al., 2021). The performances of all methods were compared based on the ROC curves and Area Under the (ROC) Curve (AUC) scores. The optimal threshold on the anomaly score of each method anomalies was identified based on the Youden index (Youden, 1950), where the threshold with the highest index was considered optimal. This coincides with the highest point on the ROC curve above the chance line, i.e., the point of highest performance compared to a random classifier. Note, however, that the optimal threshold may also depend on the acceptable trade-off between the true positive and false negative rates which is application specific.

In addition to the quantitative evaluation, the methods were compared qualitatively based on their explainability in terms of anomaly heatmaps. Herein, the goal is to indicate anomalous regions in the input image in case internal defects are present. The AE, FCDD and the multisensor AD methods naturally produce anomaly heatmaps. For AE and the multisensor AD methods, the squared pixelwise difference between the input and output was used as anomaly heatmap. For FCDD, the output was up-sampled to the input image size using nearest neighbor interpolation to produce a full resolution anomaly heatmap. Since the BCE classifier does not naturally produce anomaly heatmaps, a gradient-based method, i.e., Guided Gradient-weighted Class Activation Mapping (Guided Grad-CAM) (Selvaraju et al., 2017), was used to visualize anomalous regions.

5.3 Results

5.3.1 Quantitative evaluation of deep AD methods

The AUC scores, mean AUC score and standard errors of means for all methods on the five data splits using a different random seed are provided in Table 5.1. The highest mean AUC score was obtained by the Multisensor HEDDM (mean AUC = 0.966). This method was closely followed by the BCE classifier (mean AUC = 0.962) and FCDD

(mean AUC = 0.961). Multisensor HODDM obtained the lowest mean AUC score of all methods (0.911). For the multisensor method, using a heterogenous instead of a homogeneous DDM thus significantly improved the results. Relatively high standard errors on the AUC score were found for the AE MSE (0.019) and BCE classifier (0.010), while similar and lower standard errors on the AUC score were found for all other methods (in the range [0.004, 0.006]).

Table 5.1: AUC scores, mean AUC scores and standard errors of means for all methods over the test sets of the 5-fold cross-validation data splits. For each method, the highest AUC score over all random seeds is indicated in bold. Multisensor HEDDM: Multisensor AD using heterogenous DDM; BCE classifier; FCDD: Fully Convolutional Data Description; AE SSIM: Autoencoder using SSIM loss; AE MSE: Autoencoder using MSE loss; Multisensor HODDM: Multisensor AD using homogeneous DDM.

Method	**Random seed**					
	1	**2**	**3**	**4**	**5**	**Mean AUC**
Multisensor HEDDM	0.964	0.973	0.957	0.967	0.968	0.966 ± 0.005
BCE classifier	0.943	0.972	0.968	0.966	0.961	0.962 ± 0.010
FCDD	0.963	0.965	0.953	0.962	0.962	0.961 ± 0.004
AE SSIM	0.922	0.934	0.925	0.928	0.932	0.928 ± 0.004
AE MSE	0.926	0.878	0.913	0.924	0.928	0.914 ± 0.019
Multisensor HODDM	0.918	0.905	0.919	0.908	0.904	0.911 ± 0.006

Figure 5.6 shows the mean ROC curves and AUC scores for all methods over the 5-fold cross-validation data splits. For each method, the local standard deviations are shown calculated over the 5-fold cross-validation data splits. Herein, Multisensor HEDDM, the BCE classifier and FCDD had similar ROC curves. At the most crucial part of the ROC curve, i.e., in the region where the FPR is < 5 %, FCDD performance was the most robust to changes in the training and test datasets, i.e., it had overall the smallest deviations from the mean ROC curve. For the AE-based methods, AE SSIM outperformed AE MSE, which is in line with previous findings that AEs perform better for AD when trained using SSIM instead of MSE (Bergmann, Fauser, et al., 2019). In addition, AE SSIM also had smaller standard deviations along the mean ROC curve, i.e., it had a more consistent performance on different dataset splits compared to AE MSE.

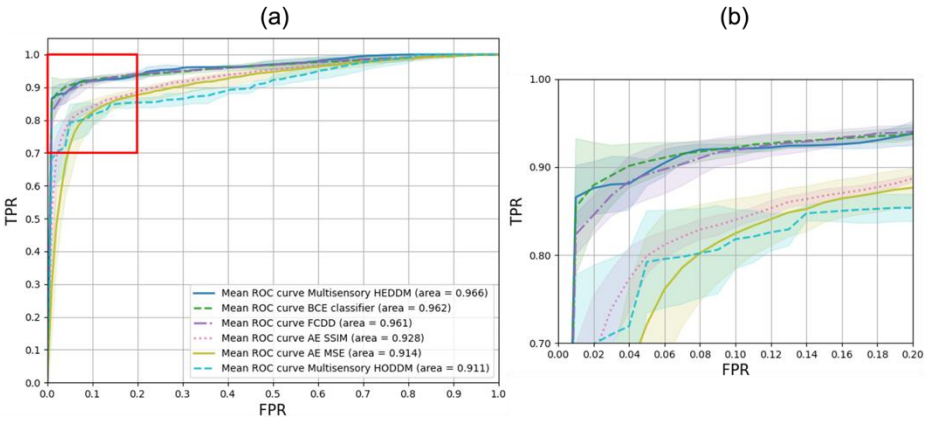

Figure 5.6: (a) ROC curves of all methods with AUC scores; (b) ROC curves in (a) zoomed in on the [0.0, 0.2] and [0.7, 0.1] range for the FPR and TNR, respectively (see red box in (a)). Bold lines indicate the mean ROC curve for each method calculated from the individual ROC curves of each data split. The transparent areas indicate the local standard deviation calculated from the individual ROC curves of each data split. Multisensor HODDM: Multisensor AD using homogeneous DDM; Multisensor HEDDM: Multisensor AD using heterogenous DDM; BCE classifier; FCDD: Fully Convolutional Data Description; AE MSE: Autoencoder using MSE loss; AE SSIM: Autoencoder using SSIM loss.

To investigate the failure cases of the AD methods, the classification performance was tested after setting a threshold on the anomaly scores based on the Youden index on the ROC curve. Herein, the best performing model and corresponding test set for each method was used (see Table 5.1). In Table 5.2, the confusion matrices for each method are shown using the previously described threshold on the anomaly scores. The BCE classifier and FCDD both achieved a TPR and TNR of > 90 %. The multisensor HEDDM and HODDM, respectively, had a TPR of 88.4 and 86.2 %, while the AE-based methods achieved TPRs of around 74 %. All methods had a TNR close to 100 %, apart from Multisensor HODDM, which had a TNR of 86.3 %.

Table 5.2: Classification performance of each method on the anomaly test dataset using the threshold on the anomaly score that maximizes the Youden index of the ROC curve. Columns: TPR: true positive rate (%); FNR: false negative rate (%); FPR: false positive rate (%); TNR: true negative rate (%); ACC: overall accuracy (%). Multisensor HEDDM: Multisensor AD using heterogenous DDM; BCE classifier; FCDD: Fully Convolutional Data Description; AE MSE: Autoencoder using MSE loss; AE SSIM: Autoencoder using SSIM loss; Multisensor HODDM: Multisensor AD using homogeneous DDM.

Method	**TPR**	**FNR**	**FPR**	**TNR**	**ACC**
Multisensor HEDDM	88.4	11.6	0	100.0	94.2
BCE classifier	91.8	8.2	0.8	99.2	95.0
FCDD	90.2	9.8	3.2	96.8	93.5
AE SSIM	74.0	26.0	1.5	98.5	86.2
AE MSE	74.4	25.6	0.5	99.5	87.0
Multisensor HODDM	86.2	13.8	13.6	86.4	86.3

Figure 5.7 shows the accuracy of each method as a function of the disorder severity of the anomaly samples in the test dataset. Every marker represents one fruit with its disorder severity expressed as percentage of the fruit volume affected by cavity formation and internal browning. The color of each marker represents the accuracy of the AD method on the 50 images simulated from the fruit. The accuracy on all the images of each sample is also plotted inside the sample's marker. From the plots, the sample distribution over the cavity and browning percentage ranges can be observed. Anomalies with both a high cavity and browning percentage and samples with very low cavity percentages were underrepresented. All methods, apart from the AE-based methods, had an accuracy of 100 % on all samples with a cavity percentage of > 1.0 %. For samples with a lower cavity percentage, the accuracy depended on the internal browning percentage. Samples with low cavity and browning percentages were especially hard to detect. Remarkably, all methods had a poor accuracy (64 % for the BCE classifier and FCDD, 0 % for the others) on the anomaly with a cavity and browning percentage of 0.3 % and 47.0 %, respectively. On a similar anomaly, with a cavity and browning percentage of 0.2 % and 41.7 %, respectively, Multisensor HEDDM, the BCE classifier and FCDD achieved an accuracy of 98, 94 and 84 %, respectively.

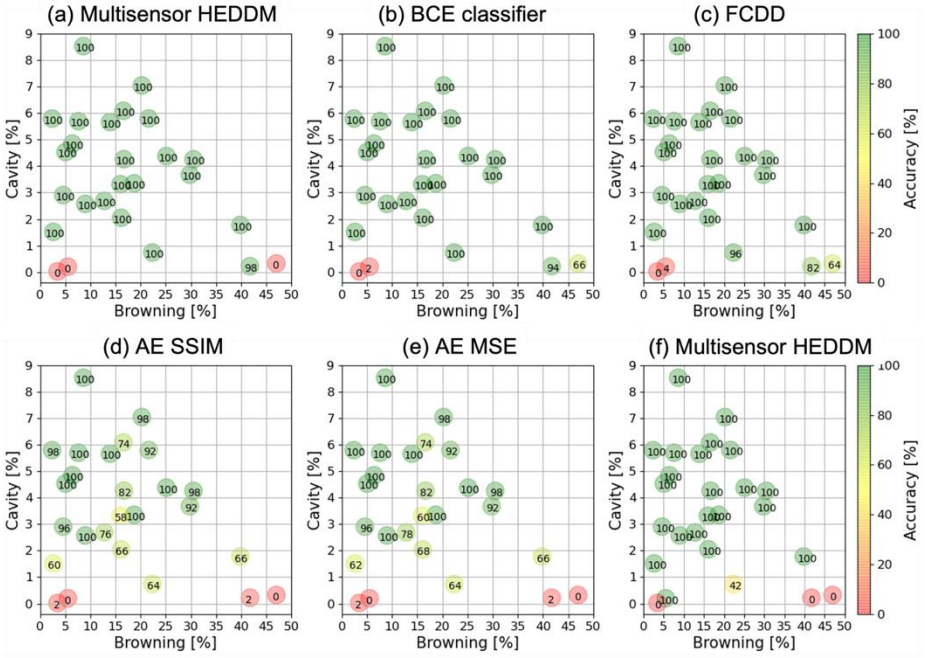

Figure 5.7: AD accuracy of all methods in function of the disorder severity on the anomaly samples in the test dataset using the threshold on the anomaly score that maximizes the Youden index of the ROC curve.

The same analysis was performed on the images of the 26 pears labeled as "Defect but consumable", which were not considered true anomalies. In this analysis, the same nominal samples were used as in the previous test set, while all anomaly images were replaced by "Defect but consumable" samples. The same thresholds on the anomaly scores were used as before. On this second test set, the performance of all methods dropped significantly (see Table 5.3). Note that TNR and FPR remained the same for all methods since the same nominal images and thresholds on the anomaly scores were used as before. While all methods had a poor TPR (≤ 43.8), FCDD and the BCE classifier outperformed both multisensor AD methods on this dataset with small defects.

Table 5.3: Classification performance of each method on the "Defect but consumable" dataset using the threshold on the anomaly score that maximizes the Youden index of the ROC curve. Columns: TPR: true positive rate (%); FNR: false negative rate (%); FPR: false positive rate (%); TNR: true negative rate (%); ACC: overall accuracy (%). Multisensor HEDDM: Multisensor AD using heterogenous DDM; BCE classifier; FCDD: Fully Convolutional Data Description; AE MSE: Autoencoder using MSE loss; AE SSIM: Autoencoder using SSIM loss; Multisensor HODDM: Multisensor AD using homogeneous DDM.

Method	**TPR**	**FNR**	**FPR**	**TNR**	**ACC**
Multisensor HEDDM	7.7	92.3	0.0	100.0	53.8
BCE Classifier	43.8	56.2	0.8	99.2	71.5
FCDD	35.3	64.7	3.2	96.8	66.1
AE SSIM	8.6	91.4	1.5	98.5	53.5
AE MSE	9.6	90.4	0.5	99.5	54.6
Multisensor HODDM	17.3	82.7	13.6	86.4	51.8

Figure 5.8 shows the accuracies of each method in function of the disorder severity of the ''Defect, but consumable'' samples in the second test dataset. It can be observed that apart from three pears, all ''Defect but consumable'' fruit had a cavity and browning percentage below 2.0 %. Samples with a cavity percentage below 1.0 % showed to be especially difficult to detect, regardless of the browning percentage. Most samples with a cavity percentage above 1.0 % could be detected reliably by FCDD and the BCE classifier, while this was not the case for both multisensor methods.

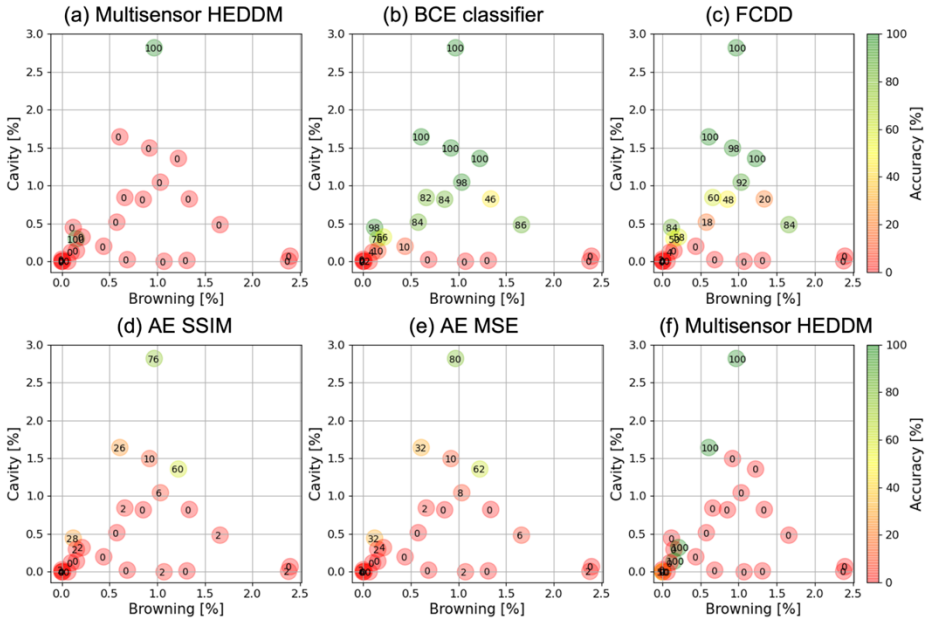

Figure 5.8: AD accuracies of all methods in function of the disorder severity on the "Defect but consumable" samples using the threshold on the anomaly score that maximizes the Youden index of the ROC curve.

5.3.2 The effect of outlier exposure

To evaluate the benefit of OE, the performance of the BCE classifier and FCDD models trained with 50, 40, 30, 20 10 and 0 % chance for OE were compared. Figure 5.9 (a) shows the ROC curves of these BCE classifiers trained and tested with the same random seed (random seed 2), i.e., only differing in the OE probability that was used during training. Even only 10 % of OE (AUC = 0.954) was sufficient to significantly improve the AD performance from the unsupervised case (AUC = 0.917). Figure 5.9 (b) shows the same evaluation for FCDD. Similar to the BCE classifier, all models trained with OE performed equally well. In the unsupervised case, i.e., without OE, FCDD (AUC = 0.934) performed better than the unsupervised AE-based AD method (AUC = 0.928, see Table 5.1). The main observation is that OE, even only applied at a low rate, significantly improved the AD performance for the FCDD and BCE classifier.

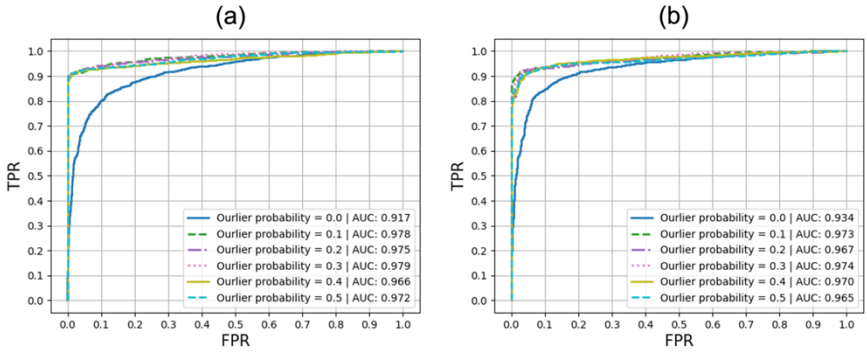

Figure 5.9: ROC curves of the BCE classifier (a) and FCDD (b) and models trained with an OE probability of 50.0, 40.0, 30.0, 20.0 10.0 and 0.0 %.

5.3.3 Qualitative evaluation of the explainability of deep AD methods

All methods were compared qualitatively based on their explainability, i.e., the quality of the anomaly heatmaps. Figure 5.10 shows the heatmaps produced by each method for two nominal samples, two synthetic anomalies produced by our OE pipeline and two anomaly samples. To visualize and compare the anomaly heatmaps for all methods, the heatmaps produced by FCDD were first rescaled between $[0, 1]$. Hereto, all plotted heatmaps produced by FCDD were min-max normalized using the minimal and maximal value found in all the plotted FCDD heatmaps. The heatmaps for AE and both multisensor AD, which were naturally in the $[0, 1]$ range, were for visualization purposes min-max normalized between $[0.0, 0.01]$.

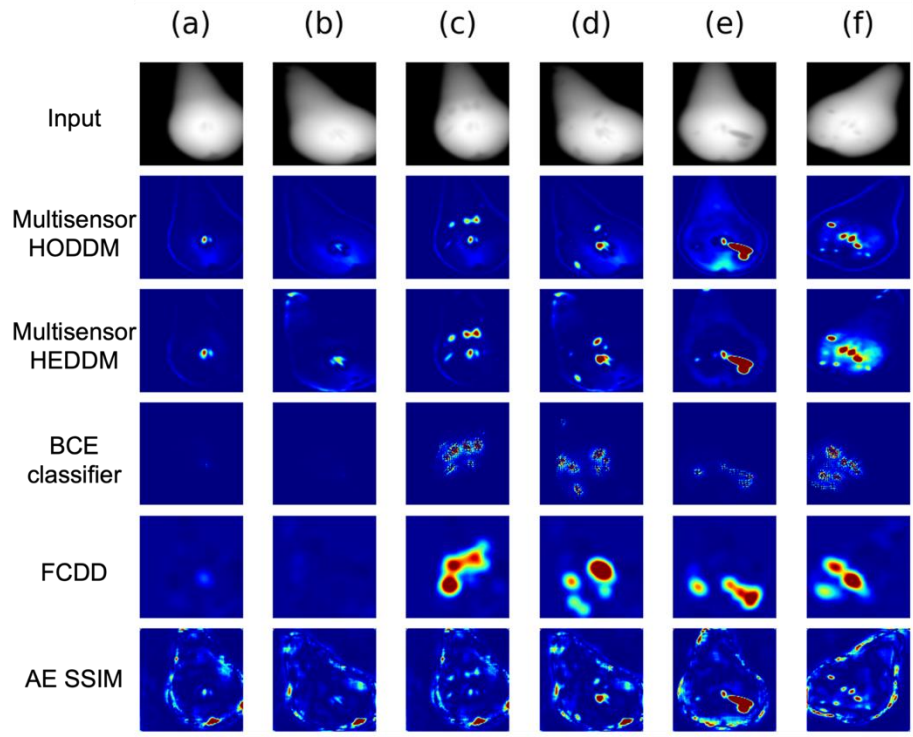

Figure 5.10: Anomaly heatmaps produced by each method for two nominal samples ((a) and (b)), two synthetic anomalies by adding artificial defects to the nominal images in (a) and (b) ((c) and (d)) and two anomalies ((e) and (f)).

The heatmaps of AE were found to be the least informative (last row). While anomalous regions were indicated, many other regions at the edge of the fruit body, around the core and at the fruit calyx were indicated as well. These regions often had higher anomaly scores than the present defects, especially for subtle defects. The multisensor AD methods (2nd and 3rd row) produced more accurate heatmaps than the AE. However, they also often indicated the core as anomalous (see columns (a-b)). Moreover, small defects were often insufficiently highlighted compared to the core area (see column (c)). In the heatmaps of Multisensor HODDM, the outlines of the fruit were highlighted. This was caused by the fact that the homogeneous DDM assumes equal density for all fruit voxels, while in the CT scans voxels at the fruit surface have lower grayscale values due to the partial volume effect. Since the heterogeneous DDM incorporates the partial volume effect, it does not suffer from

this. However, sometimes small errors occurred in the fitting the heterogeneous DDM, resulting for example in heatmaps in which regions close to the pedicel were falsely indicated as anomalous (see 3rd row of columns (b), (d), (e) and (f)). The BCE classifier's heatmaps produced using Guided Grad-CAM was found to be effective in indicating anomalous regions in the image (see 4th row). Overall, these indicated regions were found to be very noisy but still facilitated interpretation. Guided Grad-CAM indicated the areas with defects, while areas without any defects were not indicated. FCDD produced the most informative and clean heatmaps (see 3rd row), with highlighted blobs covering most individual defects (see columns (c-f)) and low values for normal tissue. Compared to the other methods, the BCE classifier and FCDD were overall found to be the least indicating the core as anomalous.

5.4 Discussion

5.4.1 Deep AD methods are on par with the state-of-the-art multisensor method

In a simulated setup, deep AD methods were compared to the recently proposed multisensor AD method (van Dael et al., 2019, 2017), which is a state-of-the-art conventional method for detecting internal defects in X-ray images of agricultural products. For the multisensor AD method, both a homogeneous (Multisensor HODDM) and heterogeneous DDM (Multisensor HEDDM) were used. For the deep AD method, we tested an unsupervised AE-based AD method, FCDD and a BCE classifier-based AD method. Herein, FCDD and the BCE classifier were trained in a semi-supervised way using OE, i.e., the generation of labeled synthetic anomalies by adding fake defects into the images of nominal samples. To evaluate the robustness of the methods towards the randomness introduced in the splitting of the dataset, the order and composition of batches during training and the initialization of the model parameters, all methods were tested using five different random seeds. It was found that the BCE classifier-based (mean AUC = 0.962) and FCDD (mean AUC = 0.961) deep AD methods trained with OE were on par with Multisensor HEDDM (mean AUC = 0.963) and significantly

outperformed AE SSIM (mean AUC = 0.928), AE MSE (mean AUC = 0.914) and Multisensor HODDM (mean AUC = 0.911) (see Table 5.1).

Similar to what van Dael et al. (2019) concluded for apple, it was found that using a heterogenous instead of a homogeneous DDM is beneficial for pear as well, although in previous work it was ignored (van Dael et al., 2017). In practice, however, the method has several downsides. First, to generate accurate reference images, the X-ray geometry, the spectrum of the X-ray source, the detector signal, the sample pose and the sample movement must be precisely known. Second, errors in fitting the SSM result in errors in the reference radiographs. Third, developing the homogeneous and heterogeneous DDM respectively required 3D shapes and CT scans of a representable nominal dataset to be available. Fourth, the DDM, as implemented by van Dael et al. (2019, 2017), only represents the mean density distribution of the cultivar and does not consider potential variations related to the product shape and size, and the shape of the core. Finally, real-time implementation is challenging due to the sequence of steps in which models must be fitted. Given that the multisensor methods were tested in their best-case scenario, it is reasonable to assume that their performance would significantly decrease in practice due to the accumulation of small errors over all consecutive steps. Moreover, provided that nominal data is abundant, the proposed deep AD methods have the benefit of being able to work on the inline X-ray radiographs directly without requiring additional sensors and tedious calibration of a multisensor system. This significantly simplifies their usage on new applications.

Deep AD is often approached as a semi-supervised learning task, since supervised learning is often infeasible due to limited availability of labeled anomalies, i.e., the dataset insufficiently captures what it means to be anomalous. However, Ruff, Vandermeulen et al. (2021) recently showed that for images a pure classification-based AD method worked surprisingly well to detect "out-of-class" anomalies, i.e., anomaly images which contained a totally different content than the normal class of images. Here, we demonstrated that this approach is also effective for detecting subtle anomalies: The BCE classifier performed slightly better than the

semi-supervised FCDD method. In addition, using Guided Grad-CAM, anomaly heatmaps could be produced which were quite accurate in terms of indicating defects. However, the heatmaps were much noisier compared to the ones produced by FCDD.

5.4.2 Outlier exposure with synthetic anomalies can significantly improve AD performance

Deep AD models were trained on a dataset of nominal samples which was augmented using OE with synthetic anomalies by adding fake defects to some of the nominal images during training. Herein, the effect of the OE rate was investigated, i.e., comparing the performance after converting 50, 40, 30, 20, 10 or 0 % of the images into synthetic anomalies during training (see Figure 5.9). OE allowed the FCDD and the BCE classifier to be trained in a semi-supervised way, i.e., on nominal samples and an equal, or smaller, number of labeled synthetic anomalies. It was found that for both methods, OE significantly improved AD performance over the unsupervised case, even when only applied at a low rate (i.e., OE probability of 10 %). For the BCE classifier and FCDD, using a 10 %OE rate improved the AUC from respectively 0.917 and 0.934 to respectively 0.978 (6.7 % improvement) and 0.973 (4.2 % improvement) on the same test dataset. Our results confirm previous findings that OE is effective for improving AD performance even when applied at a low rate (Liznerski et al., 2021; Ruff et al., 2020; Ruff, Vandermeulen, et al., 2021).

We hypothesize that the AD performance is mainly determined by how well the neural network can model nominal data. A low OE rate might improve AD over the unsupervised case because it allows the model to better estimate the boundaries of the nominal distribution. Increasing the OE rate might not improve AD, because the model might try to estimate the distribution of the OE samples. The latter might not be fully comparable to the distribution of the real anomalies. In fact, it was observed that the models trained with a 50 % OE rate were slightly outperformed by the models trained with a lower OE rate, although these differences might not be significant (see Figure 5.9).

At first sight, it might seem surprising that the BCE classifier and FCDD trained in an unsupervised way still achieved AUC scores of 0.917 and 0.934, respectively. Note, however, that the ROC curve is determined by a ranking on the anomaly scores. So, a model could learn the distribution of the nominal training data and consistently return anomaly scores close to zero. However, still slightly higher anomaly scores would be assigned to anomalies that fall outside of this distribution.

Yet, one could expect that the models would suffer from mode collapse, i.e., returning a constant anomaly score of zero and predict all samples as nominal. In that case, the output would be independent from the input and the models would perform much worse on the test set. We hypothesize that mode collapse is prevented by the use of stochastic gradient descent and batch normalization. While the stochasticity during training causes the weights to converge over time, it still causes them to slightly change between batches. Weights are therefore unlikely to become zero. In addition, if weights would converge to very small values and thus result in small activations, batch normalization would still normalize the activations to the same scale. Therefore, the models would not return a constant value.

Given that enough data would be available, training in a supervised way using real anomalies would be preferred over using the OE pipeline. However, a benefit of using the OE pipeline is that it has as a regularizing effect. During training, fake defects are added to nominal samples at random. Over all training epochs, the model thus sees the same image multiple times, each time with or without fake defects that are different in location, size and quantity. It is therefore incentivized to focus on internal defects, because other features, e.g., the fruit shape or size, become uninformative. A model trained in a purely supervised way might be biased towards large fruit because internal disorders are more likely to develop in larger fruit in which hypoxia conditions might occur more frequently or faster. Moreover, the model could also memorize the fruit shape of the anomalies in the training set instead of focusing on the defects. In contrast, a model trained with OE would be penalized if it memorized the fruit shape of an anomaly, since the same image

could later be presented as a nominal sample without defects. Note, however, that these biases could also be partially overcome by using data augmentation techniques, such as resizing and distorting the images during training.

Liznerski et al. (2021) hypothesized that using "out-of-class" anomalies would be unsuited for detecting subtle anomalies. Therefore, they proposed using "confetti noise" to create synthetic anomalies. However, they did not compare their OE method to using a general auxiliary dataset for OE. Here, it was found that models trained using OE with synthetic defects or confetti noise outperformed models that were trained using ImageNet as a general auxiliary dataset for OE (see Appendix A3). Our findings support the hypothesis that using a general auxiliary dataset for OE is uninformative for detecting subtle anomalies at test time. It was also found that in our application, the synthetic defect OE method outperformed the confetti noise OE method, which is presumably due to the closer resemblance of the synthetic defects to real defects, e.g., to the rounded edges, the occurrence of more elongated shapes and an internal gradient (see Appendix A3). To have a generally applicable OE method for image-based anomaly detection problems, e.g., defect or foreign object detection, it is advised to adapt the confetti noise OE method to include more randomness in the shapes of the produced blobs. In addition, blobs with internal gray scale gradients could be explored.

In future work, alternative OE pipelines could be explored for internal disorder detection. For instance, fake defects could be added into CT volumes instead of into radiographs which would allow for defining the disorder severity in terms of volume and changes in X-ray attenuation or density. However, that would in turn lead to other challenges, such as defining the 3D shape and location of the disorder. Alternatively, it would be interesting to investigate if these kinds of OE data augmentation techniques could become learnable as well (Antoniou et al., 2018; Cubuk et al., 2019; DeVries & Taylor, 2017; Lim et al., 2019; Tran et al., 2017), i.e., a trainable data augmentation model which introduces synthetic defects and which is trained either prior to or during the training of the AD model.

5.4.3 Anomaly heatmaps allow for interpreting anomaly detections and localizing disorders in X-ray images

There is a growing interest in explainable artificial intelligence (AI) with the goal of overcoming the black-box nature of trained neural networks. Deep learning-based methods that can provide interpretable decisions are favorable for many reasons, including their validation by human observers, the potential to discover previously unknown factors that might have been found by the model, and the security against adversarial attacks (Adadi & Berrada, 2018; Barredo Arrieta et al., 2020; Gunning et al., 2019; Montavon et al., 2018; Samek et al., 2017). While work has been done to more deeply investigate what neural networks learn internally (Bau et al., 2017), for images it is often sufficiently informative to provide visual explanations, i.e., saliency maps or heatmaps, that indicate the areas in the images which contributed the most to the output of the model. In this work, different AD methods were tested, each requiring a different method for obtaining anomaly heatmaps.

For the AE-based method, anomaly heatmaps were obtained by comparing the model's input and output. The assumption was that since the AE is only trained on nominal data, it is bad at reconstructing the internal defects in anomaly images. The internal defects are thus poorly reconstructed in the output, so that they are highlighted when compared to the input image. While internal defects were indeed indicated, the anomaly heatmaps of the AE were found to be quite noisy. Regions at the edge of the fruit body, around the core and at the fruit calyx had often also high anomaly scores. This may indicate that the AE also had difficulties in reconstructing nominal images due to the limited capacity of the encoder to encode enough information in a latent vector of limited size. Increasing the size of the latent vector could increase the reconstruction capability of the AE for nominal samples. However, this could also increase the risk of overfitting.

The anomaly heatmaps of the multisensor based methods conceptually mostly resembled the heatmaps produced by the AE. In

essence, both the AE-based and the multisensor AD method try to produce a defect-free reference image to which the input image is compared. The AE achieved this by being trained end-to-end on nominal data. In contrast, the multisensor approach predicted the reference image by using prior knowledge in the form of parametrized models developed from a representable nominal dataset and knowledge of the fruit pose and the X-ray system. The multisensor methods produced heatmaps of higher quality than those of the AE. With perfect knowledge, the multisensor AD method can produce high quality heatmaps, however, in the absence of this knowledge it is probable that errors will be introduced, e.g., due to errors in the estimation of the fruit shape or the X-ray geometry.

Due to its fully convolutional architecture, FCDD naturally produced low resolution heatmaps that could be up-sampled to the original resolution of the input size. It produced the most informative and clean heatmaps of all methods with highlighted blobs covering most individual defects and low values for normal tissue. Compared to the other methods, FCDD was overall found to be the least susceptible to indicating the core as anomalous. Presumably, the reason for these high-quality heatmaps of FCDD is that the quality of the heatmaps directly dictates the loss during training (see equation (1)).

For the BCE classifier, the gradient-based method Guided Grad-CAM was used to produce anomaly heatmaps. Guided Grad-CAM could indicate the defect regions to a comparable degree as FCDD, but its heatmaps were much noisier. Similar to FCDD, Guided Grad-CAM was not susceptible to indicating the core as anomalous.

5.4.4 Internal disorder detection depends on the disorder type and severity

To investigate the AD performance of all methods in function of the disorder severity, the methods were tested in detecting the anomalies based on the anomaly score (see section 5.3.1). Herein, the threshold on the anomaly scores was set using the Youden Index, which identifies the point on the ROC curve with the highest performance compared to a random classifier. Note, however, that in practice the optimal threshold depends on the acceptable trade-

off between the true positive and false negative rates which is application specific. All methods, apart from the AE-based methods, had an accuracy of 100 % on all anomalies with a cavity percentage above 1.0 % (see Figure 5.7). This indicates that for detecting defect fruit with cavities of sufficient size, X-ray imaging is a very powerful technique. Cavities cause a large change in X-ray attenuation, which, in relatively homogeneous fruit like pear, can quite easily be observed in the X-ray image even for low cavity percentages. The AE based method was found to be less effective for detecting small cavities, presumably due to the low signal to noise ratio in its anomaly heatmaps. For samples with a cavity percentage < 1.0 %, the accuracy depended on the internal browning percentage. Samples with low cavity and browning percentages were especially hard to detect. Note that the results may be dependent on the used X-ray geometry in the simulated environment, e.g., detector pixel size. However, detecting even smaller cavity proportions may not be commercially relevant anymore.

Remarkably, all methods struggled with a sample which had a cavity and browning percentage of 0.3 and 47.0 %, respectively (see Figure 5.7). On this anomaly, the BCE classifier and FCDD had a poor accuracy (66 and 64 %, respectively), while all other methods achieved an accuracy of 0 %. However, for a similar anomaly with cavity and browning percentage of 0.2 and 41.7 %, respectively, Multisensor HEDDM, the BCE classifier and FCDD achieved an accuracy of respectively 98, 94 and 82 %, while the other methods had an accuracy < 2 %. Figure 5.11 shows an image of the fruit and two simulated X-ray images of a nominal sample (a-b) and both anomalous samples with a browning percentage > 40 % (c-d and e-f). In addition, the corresponding anomaly heatmaps for both multisensor methods, the BCE classifier and FCDD are shown. In the images of the fruit flesh, severe internal browning can be observed for the two anomalous samples (1st row of c-f). In the X-ray images (2nd row of c-f), however, the internal browning cannot be detected with the naked eye. Depending on the angle from which the projection was taken, small cavities can be observed in (d) and (f), but not in (c) and (e). From the FCDD and BCE classifier anomaly heatmaps (5th and 6th row), it can be seen that the models easily

detect small cavities, while internal browning is not indicated. In contrast, larger parts of the fruit flesh, i.e., the regions affected by internal browning, light up in the heatmaps of the multisensor AD methods (also compare to the heatmaps of the nominal sample (a-b)). Presumably, the cellular liquid released by the damaged cells in the regions affected by internal browning has evaporated, resulting in a lower density and X-ray attenuation which can be detected by the multisensor methods. Note, however, that the multisensor methods were not always able to indicate internal browning (see rows 2-3 in Figure 5.10 (e) and row 3 in Figure 5.11 (c-d)) due to the assumption of average density which might not hold true for all fruit, e.g., porosity can differ from fruit to fruit and density can be affected by general dehydration or depend on the stage of internal disorder development.

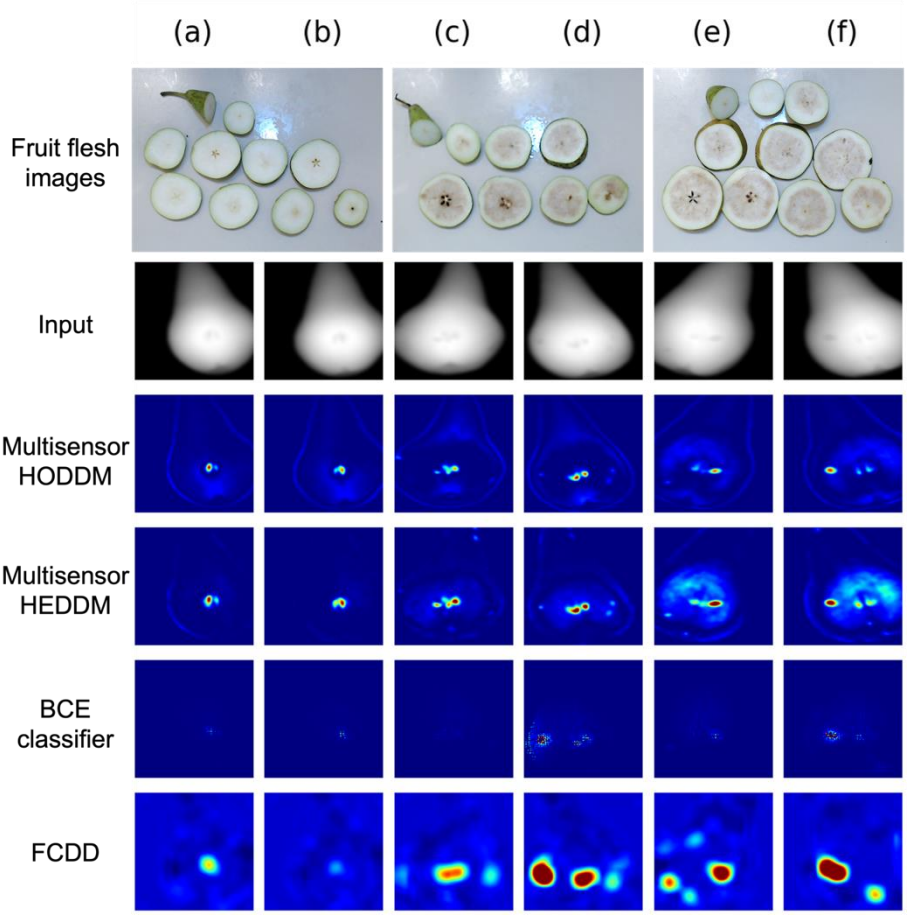

Figure 5.11: Heatmaps of each method for a nominal sample and two samples of the test set with high browning and low cavity percentage. (a-b) Image of the fruit flesh and two X-ray projections of the same nominal fruit; (c-d) Image of the fruit flesh and two X-ray projections of the same anomalous fruit (cavity and browning percentage of 0.3 and 47.0 %, respectively). (e-f) Image of the fruit flesh and two X-ray projections of the same anomalous fruit (cavity and browning percentage of 0.2 and 41.7 %, respectively). The visibility of cavities depends on the projection angle.

The above observations might indicate that for some samples in which internal browning is severe and cavities are absent, knowledge of the fruit shape might be required to detect the disorder in the X-ray radiographs reliably. For instance, brown patches generally have a lower density, and thus lower the X-ray attenuation along the path. The deep AD methods only looked at the X-ray image and had no knowledge of the fruit shape. The observed

X-ray attenuation could therefore be caused by either internal browning, or just by a different fruit shape. For cases in which internal browning has affected a major part of the fruit volume and cavities are absent, the networks were thus less able to detect an anomalous pattern. In contrast, the multisensor AD methods which, in this experiment, had perfect knowledge of the fruit shape, were able to detect this type of anomaly more reliably. Note, however, that it should also be investigated if anomalies with severe browning and cavity percentage < 1.0 % are common, or if cavity development is almost guaranteed when internal browning is severe. An alternative explanation for the poorer performance of the deep AD methods on anomalies with (severe) browning and hardly any cavities is that the models could be biased towards focusing on cavities compared to browning due the implemented OE pipeline. The OE pipeline introduced synthetic defects that might more strongly resemble cavities than browning. As discussed in section 5.4.2, other OE algorithms could be explored in future research.

Future work should extensively test the performance of X-ray based methods on a dataset with large variability in disorder severity, which additionally is uniformly distributed over the disorder severity ranges. In practice, it is hard to obtain such dataset since the outcome of disorder development cannot be guaranteed. Factors such as the fruit origin, environmental conditions during fruit development, the fruit size, and storage conditions and duration are expected to have an important role. In fact, this was one of the main motivations of testing an anomaly detection approach in this work. Our test dataset had a relatively large variability in disorder severity (see Figure 5.7), but contained images simulated from only 26 pears. Moreover, it lacked fruit in the range of high cavity and browning percentages. Additionally, fruit affected by browning, but with low cavity percentages (< 1.5 %) were underrepresented. While it is expected that samples of the former will be easily detected, more samples of the latter would allow for deeper investigation of the limitations of the evaluated AD methods and the capabilities of X-ray based inspection in general. Hereto, a more elaborate study could be done using a simulated dataset with artificial defects (see section 5.4.2) if acquiring such dataset in practice shows to be infeasible.

5.5 Conclusion

Deep anomaly detection (AD) methods were found to be effective for the inline nondestructive detection of internal disorders in X-ray images of pear fruit. The semi-supervised method FCDD and a supervised learning-based classifier were trained using outlier exposure, i.e., with synthetic anomalies, and achieved mean Area Under the ROC Curve scores (AUC) of 0.961 and 0.962, respectively, over all test sets. Both methods were only slightly outperformed by our benchmark, i.e., the multisensor AD method which had perfect knowledge of the fruit shape, pose and position, and the X-ray geometry and characteristics.

While the multisensor method was developed as a generally applicable method for internal quality inspection, it still requires the development of application specific shape and density distribution models. In contrast, it was shown that the deep AD methods can simply work directly on raw X-ray radiographs which allows for an easier transfer of the approach to other applications. By investigating AD performance in function of internal disorder severity, it was shown that using inline X-ray imaging, defect fruit with a cavity percentage > 1.0 % could be detected 100 % accurate, while for lower cavity percentages the accuracy depended on the internal browning severity. In addition, it was shown that anomaly heatmaps can be used for interpreting anomaly detections and localizing internal disorders in X-ray images of pear fruit.

Future research should further investigate AD methods for quality inspection. In addition, capabilities of X-ray radiography-based inspection in general in function of disorder severity should be explored. In particular, the possible uncertainty about detecting fruit with internal browning in the absence of cavities should be tackled.

Chapter 6

General conclusions and future perspectives

6.1 General conclusions

The overall goal of this research was to develop nondestructive inspection methods to detect internal disorders in pear fruit using X-ray imaging and machine learning (ML). The use of X-ray imaging was motivated by the fact that X-rays can easily penetrate through biological material, which allows for the visualization of internal disorders causing density changes. Automated analyses of X-ray data are required for objective and high-throughput inspection. Hereto, machine learning was applied.

In literature, X-ray CT, which provides a 3D image of the fruit, was found to be a very effective tool to study and characterize internal disorders. However, the methods described were mainly semi-automatic and time-consuming. Therefore, in Chapter 3, a method was proposed to automatically extract features from CT scans, after which a classifier was trained to classify defect and sound fruit based on these features. The trained classifiers achieved classification accuracies ranging between 90.2 and 95.1 % depending on the cultivar and number of features that were used. Low false positive rates ranging between 0.0 and 6.7 % were obtained. However, the false negative rates, ranging between 5.7 and 13.3 %, were rather high. Moreover, a downside of the method described in Chapter 3 is that the feature extraction algorithm is application specific and that it cannot quantify the severity of internal disorders. In addition, the features that were extracted might have been suboptimal.

These issues were addressed in Chapter 4, in which the use of deep learning was proposed, which has remained largely unexplored for internal quality inspection. In deep learning, a model learns end-to-

end from data, instead of relying on handcrafted features to be extracted from it. A supervised deep neural network was used to segment internal structures, including internal disorders, in CT scans. Manually annotated CT scans of healthy and defect fruit were used as training data. A high agreement was found between the predicted and ground truth "healthy tissue", "core" and "cavity" labels (average IoU $\geq$ 0.95). Interestingly, low IoU scores were found for the "internal browning" label, even though visually most predictions seemed sufficiently accurate. It turned out this was mainly caused by errors on small volumes and volume edges. Since the IoU metric is relative to the ground truth, the absolute size of the volumes did not matter, resulting in low IoU scores even though the error by the model was rather negligible. From the predicted labels of the model, the severity of the internal disorders could be quantified by calculating the affected volumes. The resulting quantitative data was used to classify "consumable" *vs* "non-consumable" fruit at high accuracy (99.4 %) on the one hand and "healthy" *vs* "defect but consumable" *vs* "non-consumable" classification on the other hand (92.2 %). For the latter, the identification of "defect but consumable" fruit showed to be difficult (true positive rate of 65.0 %), with most misclassified fruit assigned to the "healthy" class. While the presented method could with high certainty prevent non-consumable fruit from reaching consumers, it was not successful in separating top quality from acceptable fruit.

A concern with X-ray CT is that it is currently not applicable inline at the speed of commercial sorting lines (10 fruit/s). X-ray radiography, on the other hand, can easily be implemented inline using an X-ray source and detector on either side of a conveyor belt. For detecting defect fruit using 2D X-ray radiography, conventional machine learning-based methods were already reported in literature, i.e., using image processing to extract features followed by a machine learning algorithm for classification based on these features. However, these methods required application specific feature extraction algorithms. Van Dael et al. (2019, 2017) developed a more general purpose multisensor method for internal quality inspection. However, it still required application specific shape and density distribution models (DDM). Moreover, the

method is relatively hard to be implement due to its complex integration of multiple sensors and model fitting procedures.

Deep learning-based methods could work directly on raw X-ray radiographs to overcome the need for application specific feature extraction algorithms, or shape and density distribution models. Additionally, no extra sensors would be required. However, a large, annotated dataset is typically needed to train neural networks, which is labor intensive to acquire. In addition, in the case of internal disorders in pome fruit, it is challenging to acquire sufficient and representative defect samples with a wide range in disorder severity. In contrast, healthy fruit are abundant, immediately available after harvest, and can be easily obtained from various locations and harvest years.

In Chapter 5, an anomaly detection approach using deep learning was therefore proposed, recognizing recent advantages in deep learning, while overcoming the need for annotated data normally required for supervised learning. In anomaly detection, a model is constructed from nominal data and a certain metric is used as anomaly score to measure the extent to which a new sample deviates from normality. Neural networks were trained exclusively on X-ray radiographs of healthy pears, after which they were evaluated on a test set with healthy and anomalous data. Performance could be significantly improved by using synthetic anomalies in which nominal images were subtly distorted and used during training as labeled anomalies. The proposed method reached a mean AUC of up to 0.962 (area under the ROC curve). This was on par with the multisensor method (mean AUC = 0.963) (van Dael et al., 2019, 2017) which was given the advantage of perfect knowledge of the fruit shape, pose and position, and the X-ray geometry and characteristics. By investigating the performance in function of internal disorder severity, it was shown that using the proposed method, defect fruit with a cavity percentage > 1.0 % could be detected 100 % accurate, while for lower cavity percentages the accuracy depended on the internal browning severity. The black-box nature of neural networks was addressed by producing saliency maps of the anomalous regions found by the models.

In conclusion, it was shown that X-ray imaging is an effective tool for internal disorder detection methods in pear fruit. Significant progress was made using deep learning approaches to work on X-ray images directly. In X-ray CT, the severity of internal disorders could be quantified automatically, allowing accurate detection of non-consumable fruit. For inline inspection, a deep anomaly detection approach was proposed using X-ray radiography. Hereby, the need for having application specific shape and density models or labeled datasets was overcome.

6.2 Future perspectives

A large step forward was made towards internal disorder detection in pears using X-ray imaging. Nonetheless, improvements and further research are required.

In Chapter 4, deep learning showed to be effective for semantically segmenting CT volumes. While the presented method could reliably prevent non-consumable fruit from reaching consumers, additional research is required to improve the separation of top quality from acceptable fruit. In addition, a large scale consumer survey is recommended to better understand consumer tolerances with respect to internal defects.

The model presented in Chapter 4 was trained in a supervised way and segmented the CT volumes in a slice-by-slice fashion. In future work the model could be extended to do 3D segmentation, e.g., using 3D U-Net (Çiçek et al., 2016), to improve performance. A downside of the supervised approach was that a manually annotated dataset was required for training the models. It could be investigated whether defect regions could be detected in an unsupervised way, e.g., using the saliency maps of anomaly detection models as presented in Chapter 5. Cavities and browning could then be distinguished easily based on the grayscale values in the images. However, it seems more difficult to distinguish other defects from each other with an anomaly detection model.

To allow for the method presented in Chapter 4 to be applied inline, it could be coupled to an inline X-ray CT system in which a fast reconstruction is made from a limited number of projections. In

current inline X-ray CT research, artificial neural networks are often used to improve the quality of an initial, coarse reconstruction. Since in practice the CT images would serve as an input for another task, e.g., the detection of internal disorders, a neural network could be trained to directly perform the task on the coarse reconstruction, e.g., semantic segmentation. Herein, the model does not have to precisely predict gray scale values, and instead must only classify pixels. It could, however, be difficult for the model to accurately predict region boundaries from the coarse reconstruction.

For inline detection of internal disorders, additional research is recommended. Since the deep anomaly detection method presented in Chapter 5 was only tested on a simulated dataset, in future work it should be validated on a real dataset. Compared to other methods, e.g., multisensor inspection (van Dael et al., 2019, 2017), the presented method is relatively easy to implement on a real inline X-ray radiography system, as the model works on X-ray images directly. Moreover, the model can be trained on only healthy data that is readily available, with expected performance improvements when using outlier exposure or a limited number of labeled anomalies.

In Chapter 5, it was shown that defect pears with 1 % of their volume affected by cavity formation can be detected reliably using a single X-ray radiograph. In the absence of cavities, however, it might be hard to detect a deviant pattern in the X-ray contrast for defect fruit with severe and uniformly distributed browning disorder. Relatively small defects might show as notable patches, but a large brown defect might affect such a large region in the X-ray image that it might look normal, i.e., the fruit could have another shape instead of being affected by browning. Potentially, multiple radiographs from different angles are required for 1) having the possibility to detect the disorder more easily from another viewpoint; or 2) use the additional radiographs to make inferences about the fruit shape such that anomalous X-ray attenuation can be recognized.

Furthermore, attention should be given to early browning. In this work, fruit were stored for several months to simulate long term CA storage. For defect fruit, this allowed the cellular liquid released

from damaged cells to evaporate and result in density changes. For fruit that are sold earlier, however, internal disorders could also have developed over a period of several weeks. In that case the change in density might not yet be sufficient to be detected in X-ray transmission images. It is expected that this would be less of an issue for X-ray CT, since a reduction in porosity due to the leaked cellular liquid might result in more uniform regions in terms of CT intensities, i.e., a change in texture, which could be detected.

In addition to a validation on a real dataset, a thorough *in silico* study is recommended to investigate the limits of X-ray radiography for internal defect detection in horticultural products. Herein, synthetic samples (e.g., from shape and density distribution models) in random poses could be used that include artificial defects varying in shape, size, location, and relative density. Using deep anomaly detection with a model trained on healthy samples, the feasibility of detecting internal disorders could be mapped as a function of the characteristics of the defects. Early browning could be simulated as defects with a density very close to the density of the surrounding tissue. In addition, it could be tested whether the usage of a multiple radiographs from different angles can improve performance. The results of the *in silico* study could be used to make strategic decisions on the usage and design of X-ray based inline inspection systems. Furthermore, the method could be used to test the feasibility of X-ray radiography based foreign object detection in products with variable shape and density.

While no direct comparison was made between the CT based method presented in Chapter 4 and the inline X-ray radiography based method presented in Chapter 5, it would be interesting to investigate what can be gained from having 3D data available. In addition, it could be investigated if disorder severities could be estimated inline from projection data.

Further research is required into the dynamic process of internal disorder development, e.g., the onset and rate of tissue degradation. Hereto, multiple scans should be taken throughout the storage period. Disorder development following a radial pattern can be understood from simulating the overall gas gradients in the fruit

(Herremans, Verboven, et al., 2014; Ho et al., 2013). However, this is much harder for local defects, since that would require high resolution imaging of the microstructure to map porosity and pore connectivity (Janssen et al., 2020). This poses a problem for current hardware, since high resolution imaging of tissue is mostly done destructively, while the microstructure is needed of a healthy fruit which must be imaged again after internal disorder development to validate where the disorders developed. Potentially, deep neural networks could predict regions with highest probability of developing disorders based on porosity maps (Nugraha et al., 2019). With better understanding of dynamic process of internal disorder development, fruit could potentially be sorted based on how long the fruit can be stored without developing internal disorders.

In the near future, significant technological progress in X-ray imaging is expected. Inline X-ray CT systems are expected to become faster and more affordable. Commercial systems for 3D inspection are already offered today, e.g., "Mito" by BIOMETIC targeted at the food industry (www.biometic.com). Furthermore, X-ray phase-contrast imaging, which can visualize phase shifting and scattering information, could be a great tool for internal disorder detection (Einarsdóttir et al., 2016; Endrizzi, 2018). While some disorders might have absorption characteristics similar to healthy tissue, e.g., early browning, they might differ significantly in scattering behavior. X-ray phase-contrast imaging can be applied in CT and projection modes. In addition, dual energy systems, or the usage of multispectral X-ray detectors, might facilitate better detection of internal disorders, although these methods are mostly targeted at identifying materials with different chemical composition (Andriiashen et al., 2021; Einarsson et al., 2017). Materials are distinguished based on differences in X-ray attenuation for specific energy bands. However, it is expected that healthy and affected tissue may have similar attenuation curves.

Appendix A

A1 Simulation of the radiography dataset

A1.1 Simulated inline scans

A dataset of inline X-ray images was simulated from the CT data in a virtual inline X-ray geometry. Hereto the ASTRA Toolbox (version 1.9.9.dev) was used in MATLAB 2020b (imec-Vision Lab & CWI, 2019; MATLAB, 2020; Van Aarle et al., 2015, 2016). For each simulated inline X-ray image, the CT volume of a sample was rotated randomly in 3D after which 300 consecutive line scans were simulated, while considering the movement of the source and detector relative to the sample, the line rate of the detector and the speed of the conveyor belt. To prevent pears in the rotated CT volumes from pointing directly towards the X-ray source (which is unrealistic on a conveyer system), the rotation of the main fruit axis in that direction was limited to a maximum of 36°. For the simulated system, a source-to-detector and a source-to-object distance of 0.47 and 0.32 m were used, respectively. The line detector had 300 pixels with a pixel size of 0.5 mm. The conveyor belt speed and detector line rate were respectively set to 0.27 m/s and 540 Hz. This resulted in images of 300 × 300 pixels. From each of the 180 CT volumes, 50 inline scans were simulated, resulting in a dataset of 9000 images. Herein, 6400 were nominal, while the 1300 were anomalies. The remaining 1300 images were simulated from the "Defect but consumable" class. The random 3D rotation of the CT volume ensured that each X-ray image was unique and to ensured that models trained on the data can handle various fruit poses.

A1.2 Simulated reference images

Reference images were simulated using the following steps. First, for each image, the CT volume of the fruit was placed in the same pose as was used for the generating the simulated inline scans. Second, the DDMs were fitted to the fruit shape. For fitting the homogeneous DDM, the CT volume was simply thresholded and filled to result in a homogeneous volume, i.e., the same value for all voxels inside the fruit surface. For fitting the heterogeneous DDM, a mesh of the outer

surface was created using the marching cubes algorithm (Lorensen & Cline, 1987) and the DDM was fit to this mesh. Finally, reference images with the homogeneous and heterogenous DDMs were simulated with the same protocol as used for the simulated inline radiographs. Note that by directly using the fruit shape from the CT volume, or a mesh which is directly computed from it, it is assumed that the shape of the fruit is perfectly known.

A2 The synthetic defect OE pipeline

In the synthetic defect OE pipeline, the following image processing steps were taken to introduce synthetic defects (see Figure 5.3). First, the fruit body was segmented from the background using Otsu thresholding (Otsu, 1979). Second, the suspectable area, i.e., the area in which the presence of defects is plausible, was indicated by applying an erosion on the segmented fruit body (Figure 5.3 (b)). This operation removed the outer border of the projected fruit body, as internal defects are more plausible in the center of the fruit compared to close to the fruit surface. Third, pixels in the suspectable area were sampled at random for being a seed point for a synthetic defect (Figure 5.3 (c)). The number of seed points, and thus the number of individual defects in the image, was controlled by uniformly sampling the probability of a pixel being selected for a seed point in a range of [0.01 %, 0.1 %]. Fourth, at every seed point a binary circular area was added, with a radius uniformly sampled between $[3, 15]$ pixels. Fifth, each circular area was deformed using a shearing transformation resulting in ellipsoidal blobs. Sixth, a distance transform followed by a gaussian filter was applied to all pixels within the mask, resulting in irregularly shaped grayscale blobs (Figure 5.3 (d)). Finally, the grayscale values were normalized between $[0.0, 0.1]$ and subtracted from the original projection (Figure 5.3 (e)). The values of the parameters in the pipeline were set based on visual comparison between the resulting synthetic anomalies and real anomalies.

A3 Comparison of OE pipelined

The synthetic defect OE pipeline was compared to using the ImageNet dataset as a general auxiliary dataset for OE, and to using "confetti noise" as proposed by Liznerski et al. (2021). In the

ImageNet OE pipeline, the images were simply replaced by random grayscale versions of the images in ImageNet. In the confetti noise OE pipeline, random grayscale rectangular blobs are subtracted from the input image. Herein, we used the same probability range of a pixel being selected for a seed point, i.e., [0.01 %, 0.1 %], and minimal and maximal blob sizes, i.e., [3, 15] pixels, as were used in the synthetic defect OE pipeline. Similar as to the synthetic defect OE pipeline, the grayscale values of all blobs normalized between [0.0, 0.1] prior to subtracting them form the input image (see section 5.2.2.4). Figure A.1 shows four OE samples of each pipeline. The ImageNet images were of course totally different from the images in our dataset. The synthetic defect OE pipeline produces rounded ellipse-shaped blobs with an internal gradient, while the confetti noise OE pipeline produced homogeneous rectangular blobs.

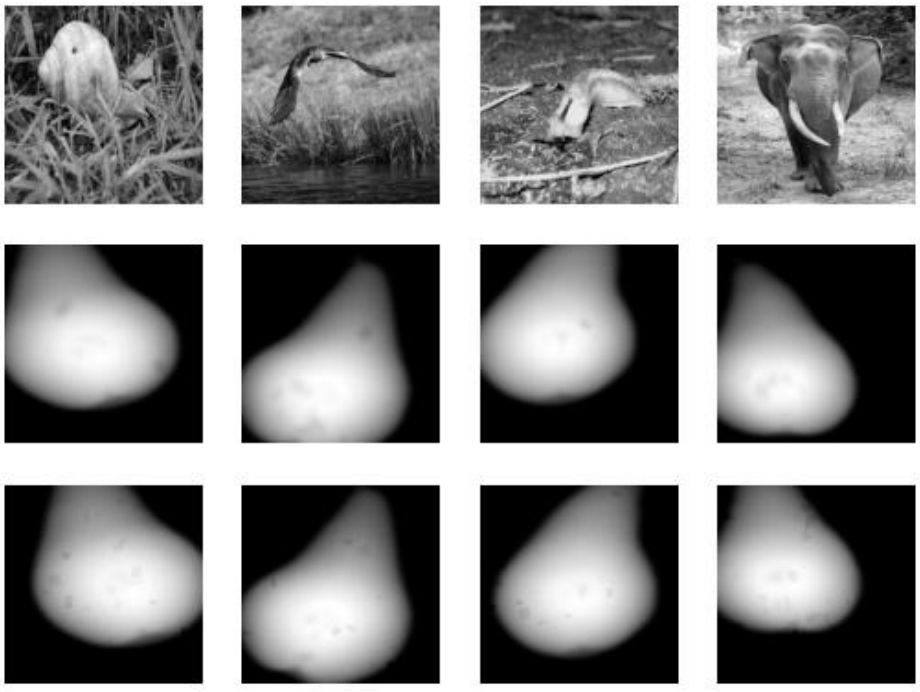

Figure A.1: Four OE samples from each OE pipeline. ImageNet OE pipeline (1st row); Synthetic defect OE pipeline (2nd row); Confetti noise OE pipeline (3rd row).

The performances of the BCE classifier and FCDD trained with each OE pipeline using a 50 % OE rate is shown in Table A.1. Overall, the

best performance was achieved using the synthetic defect OE pipeline, which is presumably due to the closer resemblance of the synthetic defects to real defects, e.g., to the rounded edges, potentially more elongated shapes and an internal gradient.

Table A.1: BCE classifier and FCDD model performance after training with the synthetic defect OE pipeline, general OE using ImageNet, and the confetti noise OE pipeline using a 50 % OE rate. AUC scores, mean AUC scores and standard errors of means for all methods over the test sets of the 5-fold cross-validation data splits. For each method, the highest AUC score over all OE pipelines is indicated in bold for each random seed.

Method	**Random seed**					**Mean AUC**
	1	**2**	**3**	**4**	**5**	
BCE classifier (synthetic defect OE)	0.943	0.972	0.968	0.966	0.961	0.962 ± 0.010
BCE classifier (ImageNet OE)	0.842	0.836	0.781	0.698	0.828	0.797 ± 0.054
BCE classifier (confetti noise OE)	0.917	0.919	0.918	0.944	0.913	0.922 ± 0.011
FCDD (synthetic defect OE)	0.963	0.965	0.953	0.962	0.962	0.961 ± 0.004
FCDD (ImageNet OE)	0.939	0.951	0.906	0.952	0.919	0.933 ± 0.018
FCDD (confetti noise OE)	0.948	0.928	0.941	0.955	0.955	± 0.010

A4 Neural network architectures

A4.1 Autoencoder

Table A.2: Autoencoder model architecture (Liznerski et al., 2021).

Layer	Output shape	Parameters #
Conv2d-1	[-1, 8, 224, 224]	200
BatchNorm2d-2	[-1, 8, 224, 224]	0
MaxPool2d-3	[-1, 8, 112, 112]	0
Conv2d-4	[-1, 32, 112, 112]	6,400
BatchNorm2d-5	[-1, 32, 112, 112]	0
MaxPool2d-6	[-1, 32, 56, 56]	0
Conv2d-7	[-1, 64, 56, 56]	18,432
BatchNorm2d-8	[-1, 64, 56, 56]	0
Conv2d-9	[-1, 128, 56, 56]	73,728
BatchNorm2d-10	[-1, 128, 56, 56]	0
MaxPool2d-11	[-1, 128, 28, 28]	0
Conv2d-12	[-1, 128, 28, 28]	147,456
BatchNorm2d-13	[-1, 128, 28, 28]	0
MaxPool2d-14	[-1, 128, 14, 14]	0
Conv2d-15	[-1, 64, 14, 14]	73,728
BatchNorm2d-16	[-1, 64, 14, 14]	0
MaxPool2d-17	[-1, 64, 7, 7]	0
Linear-18	[-1, 1536]	4,816,896
BatchNorm1d-19	[-1, 1536]	0
Linear-20	[-1, 784]	1,204,224

BatchNorm1d-21	[-1, 784]	0
Reshape-22	[-1, 16, 7, 7]	0
ConvTranspose2d-23	[-1, 64, 7, 7]	9,216
BatchNorm2d-24	[-1, 64, 7, 7]	0
ConvTranspose2d-25	[-1, 128, 14, 14]	73,728
BatchNorm2d-26	[-1, 128, 14, 14]	0
ConvTranspose2d-27	[-1, 128, 28, 28]	147,456
BatchNorm2d-28	[-1, 128, 28, 28]	0
ConvTranspose2d-29	[-1, 64, 56, 56]	73,728
BatchNorm2d-30	[-1, 64, 56, 56]	0
ConvTranspose2d-31	[-1, 32, 56, 56]	18,432
BatchNorm2d-32	[-1, 32, 56, 56]	0
ConvTranspose2d-33	[-1, 8, 112, 112]	6,400
BatchNorm2d-34	[-1, 8, 112, 112]	0
ConvTranspose2d-35	[-1, 1, 224, 224]	200
Trainable parameters	6,670,224	

A4.2 FCDD model

Table A.3: FCDD model architecture (Liznerski et al., 2021).

Layer	Output shape	Parameters #
Conv2d-1	[-1, 8, 224, 224]	200
BatchNorm2d-2	[-1, 8, 224, 224]	0
MaxPool2d-3	[-1, 8, 112, 112]	0
Conv2d-4	[-1, 32, 112, 112]	6,400
BatchNorm2d-5	[-1, 32, 112, 112]	0
MaxPool2d-6	[-1, 32, 56, 56]	0
Conv2d-7	[-1, 64, 56, 56]	18,432
BatchNorm2d-8	[-1, 64, 56, 56]	0
Conv2d-9	[-1, 128, 56, 56]	73,728
BatchNorm2d-10	[-1, 128, 56, 56]	0
MaxPool2d-11	[-1, 128, 28, 28]	0
Conv2d-12	[-1, 128, 28, 28]	147,456
Conv2d-13	[-1, 1, 28, 28]	128
Trainable parameters	246,344	

References

3D Slicer. (2020). https://www.slicer.org/

Activation Function. (2019). AI Wiki. https://docs.paperspace.com/machine-learning/wiki/activation-function

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Ahmadzadeh, A., Kempton, D. J., Chen, Y., & Angryk, R. A. (2021). Multiscale IoU: A Metric for Evaluation of Salient Object Detection with Fine Structures. *ArXiv:2105.14572 [Cs]*. http://arxiv.org/abs/2105.14572

Andriiashen, V., van Liere, R., van Leeuwen, T., & Batenburg, K. J. (2021). Unsupervised foreign object detection based on dual-energy absorptiometry in the food industry. *ArXiv:2104.05326 [Cs, Eess]*. http://arxiv.org/abs/2104.05326

Antoniou, A., Storkey, A., & Edwards, H. (2018). Data Augmentation Generative Adversarial Networks. *ArXiv:1711.04340 [Cs, Stat]*. http://arxiv.org/abs/1711.04340

Arendse, E., Fawole, O. A., Magwaza, L. S., & Opara, U. L. (2018). Non-destructive prediction of internal and external quality attributes of fruit with thick rind: A review. In *Journal of Food Engineering* (Vol. 217). Elsevier Ltd. https://doi.org/10.1016/j.jfoodeng.2017.08.009

Avermaete, T., Bonjean, I., Lievens, E., & Mathijs, E. (2018). *Apple and pear farming in Belgium: An extended summary* (p. 27). SUFISA.

Azizah, L. M., Umayah, S. F., Riyadi, S., Damarjati, C., & Utama, N. A. (2017). Deep learning implementation using convolutional neural network in mangosteen surface defect detection. *2017 7th IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*, 242–246. https://doi.org/10.1109/ICCSCE.2017.8284412

Bai, K. (2019). A Comprehensive Introduction to Different Types of Convolutions in Deep Learning. *Medium*.

https://towardsdatascience.com/a-comprehensive-introduction-to-different-types-of-convolutions-in-deep-learning-669281e58215

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Barrett, J. F., & Keat, N. (2004). Artifacts in CT: recognition and avoidance. *Radiographics*, *24*(6), 1679–1691.

Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). Network Dissection: Quantifying Interpretability of Deep Visual Representations. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3319–3327. https://doi.org/10.1109/CVPR.2017.354

Beister, M., Kolditz, D., & Kalender, W. A. (2012). Iterative reconstruction methods in X-ray CT. *Physica Medica*, *28*(2), 94–108. https://doi.org/10.1016/j.ejmp.2012.01.003

Bergmann, P., Fauser, M., Sattlegger, D., & Steger, C. (2019). MVTec AD — A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9584–9592. https://doi.org/10.1109/CVPR.2019.00982

Bergmann, P., Löwe, S., Fauser, M., Sattlegger, D., & Steger, C. (2019). Improving Unsupervised Defect Segmentation by Applying Structural Similarity to Autoencoders. *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 372–380. https://doi.org/10.5220/0007364503720380

Bishop, C. M. (2006). Pattern recognition and machine learning. *Machine Learning*, *128*(9).

Bobelyn, E., Serban, A.-S., Nicu, M., Lammertyn, J., Nicolai, B. M., & Saeys, W. (2010). Postharvest quality of apple predicted by NIR-spectroscopy: Study of the effect of biological variability on spectra and model performance. *Postharvest Biology and Technology*, *55*(3), 133–143. https://doi.org/10.1016/J.POSTHARVBIO.2009.09.006

Brown, R. W., Cheng, Y.-C. N., Haacke, E. M., Thompson, M. R., & Venkatesan, R. (2014). *Magnetic resonance imaging: Physical principles and sequence design*. John Wiley & Sons.

Buratti, A., Bredemann, J., Pavan, M., Schmitt, R., & Carmignato, S. (2018). Applications of CT for Dimensional Metrology. In S. Carmignato, W. Dewulf, & R. Leach (Eds.), *Industrial X-Ray Computed Tomography* (pp. 333–369). Springer International Publishing. https://doi.org/10.1007/978-3-319-59573-3_9

Cantre, D., Herremans, E., Verboven, P., Ampofo-Asiama, J., Hertog, M. L. A. T. M., & Nicolaï, B. M. (2017). Tissue breakdown of mango (Mangifera indica L. cv. Carabao) due to chilling injury. *Postharvest Biology and Technology*, *125*, 99–111. https://doi.org/10.1016/j.postharvbio.2016.11.009

Casasent, D. A., Sipe, M. A., Schatzki, T. F., Keagy, P. M., & Lee, L. C. (1998). Neural net classification of X-ray pistachio nut data. *LWT - Food Science and Technology*, *31*(2), 122–128. https://doi.org/10.1006/fstl.1997.0320

Chalapathy, R., & Chawla, S. (2019). Deep Learning for Anomaly Detection: A Survey. *ArXiv:1901.03407 [Cs, Stat]*. http://arxiv.org/abs/1901.03407

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, *41*(3), 15:1–15:58. https://doi.org/10.1145/1541880.1541882

Chen, Y., An, X., Gao, S., Li, S., & Kang, H. (2021). A Deep Learning-Based Vision System Combining Detection and Tracking for Fast On-Line Citrus Sorting. *Frontiers in Plant Science*, *12*, 171. https://doi.org/10.3389/fpls.2021.622062

Chigwaya, K., Schoeman, L., Fourie, W. J., Crouch, I., Viljoen, D., & Crouch, E. M. (2018). 'Fuji'' apple internal browning explored via X-ray computed tomography (CT).' *Acta Horticulturae*, *1201*, 309–316. https://doi.org/10.17660/ActaHortic.2018.1201.42

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). *3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation*. http://arxiv.org/abs/1606.06650

Clark, C. J., & Burmeister, D. M. (1999). Magnetic Resonance Imaging of Browning Development in `Braeburn' Apple during

Controlled-atmosphere Storage under High CO2. *HortScience*, *34*(5), 915–919. https://doi.org/10.21273/HORTSCI.34.5.915

Clark, C. J., MacFall, J. S., & Bieleski, R. L. (1998). Loss of watercore from `Fuji' apple observed by magnetic resonance imaging. *Scientia Horticulturae*, *73*(4), 213–227. https://doi.org/10.1016/S0304-4238(98)00076-4

Clark, C. J., & Richardson, C. A. (1999). Observation of watercore dissipation in 'Braebum' apple by magnetic resonance imaging. *New Zealand Journal of Crop and Horticultural Science*, *27*(1), 47–52. https://doi.org/10.1080/01140671.1999.9514079

Colnago, L. A., Andrade, F. D., Souza, A. A., Azeredo, R. B. V., Lima, A. A., Cerioni, L. M., Osán, T. M., & Pusiol, D. J. (2014). Why is Inline NMR Rarely Used as Industrial Sensor? Challenges and Opportunities. *Chemical Engineering and Technology*, *37*(2), 191–203. https://doi.org/10.1002/ceat.201300380

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., & Le, Q. V. (2019). *AutoAugment: Learning Augmentation Strategies From Data*. 113–123. https://openaccess.thecvf.com/content_CVPR_2019/html/Cubuk_AutoAugment_Learning_Augmentation_Strategies_From_Data_CVPR_2019_paper.html

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, *1*, 886–893 vol. 1. https://doi.org/10.1109/CVPR.2005.177

Danckaert, S., Demuynck, E., de Regt, E., De Samber, J., Lambrechts, G., Lenders, S., Vanhee, M., Vervloet, D., Vermeyen, V., & Vrints, G. (2018). *Landbouwrapport 2018: Fruit* (G. Platteau, J Lambrechts, K. Roels, & T. (reds.) Van Bogaert, Eds.; pp. 265–290). Departement Landbouw en Visserij.

De Schryver, T., Dhaene, J., Dierick, M., Boone, M. N., Janssens, E., Sijbers, J., Dael, M. V., Verboven, P., & Nicolai, B. (2016). In-line NDT with X-Ray CT combining sample rotation and translation. *NDT and E International*, *84*, 89–98. https://doi.org/10.1016/j.ndteint.2016.09.001

Defraeye, T., Lehmann, V., Gross, D., Holat, C., Herremans, E., Verboven, P., Verlinden, B. E., & Nicolai, B. M. (2013). Application of MRI for tissue characterisation of 'Braeburn'

apple. *Postharvest Biology and Technology*, *75*, 96–105. https://doi.org/10.1016/j.postharvbio.2012.08.009

Deng, L., Li, J., & Han, Z. (2021). Online defect detection and automatic grading of carrots using computer vision combined with deep learning methods. *LWT*, *149*, 111832. https://doi.org/10.1016/j.lwt.2021.111832

DeVries, T., & Taylor, G. W. (2017). Dataset Augmentation in Feature Space. *ArXiv:1702.05538 [Cs, Stat]*. http://arxiv.org/abs/1702.05538

Diels, E., van Dael, M., Keresztes, J., Vanmaercke, S., Verboven, P., Nicolai, B., Saeys, W., Ramon, H., & Smeets, B. (2017). Assessment of bruise volumes in apples using X-ray computed tomography. *Postharvest Biology and Technology*, *128*, 24–32. https://doi.org/10.1016/j.postharvbio.2017.01.013

Donis-González, I. R., Guyer, D. E., Fulbright, D. W., & Pease, A. (2014). Postharvest noninvasive assessment of fresh chestnut (Castanea spp.) internal decay using computer tomography images. *Postharvest Biology and Technology*, *94*, 14–25. https://doi.org/10.1016/j.postharvbio.2014.02.016

Edwards, M. (2004). *Detecting foreign bodies in food*. Elsevier.

Einarsdóttir, H., Emerson, M. J., Clemmensen, L. H., Scherer, K., Willer, K., Bech, M., Larsen, R., Ersbøll, B. K., & Pfeiffer, F. (2016). Novelty detection of foreign objects in food using multi-modal X-ray imaging. *Food Control*, *67*, 39–47. https://doi.org/10.1016/J.FOODCONT.2016.02.023

Einarsson, G., Jensen, J. N., Paulsen, R. R., Einarsdottir, H., Ersbøll, B. K., Dahl, A. B., & Christensen, L. B. (2017). Foreign Object Detection in Multispectral X-ray Images of Food Items Using Sparse Discriminant Analysis. In P. Sharma & F. M. Bianchi (Eds.), *Image Analysis* (pp. 350–361). Springer International Publishing. https://doi.org/10.1007/978-3-319-59126-1_29

Endrizzi, M. (2018). X-ray phase-contrast imaging. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *878*, 88–98. https://doi.org/10.1016/j.nima.2017.07.036

European Commission. (2020). *The apple market in the EU: production, area and yields*. https://ec.europa.eu/info/sites/info/files/food-farming-fisheries/farming/documents/apples-production_en.pdf

Fan, S., Li, J., Zhang, Y., Tian, X., Wang, Q., He, X., Zhang, C., & Huang, W. (2020). On line detection of defective apples using computer vision system combined with deep learning methods. *Journal of Food Engineering*, *286*, 110102. https://doi.org/10.1016/j.jfoodeng.2020.110102

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Fazari, A., Pellicer-Valero, O. J., Gómez-Sanchıs, J., Bernardi, B., Cubero, S., Benalia, S., Zimbalatti, G., & Blasco, J. (2021). Application of deep convolutional neural networks for the detection of anthracnose in olives using VIS/NIR hyperspectral images. *Computers and Electronics in Agriculture*, *187*, 106252. https://doi.org/10.1016/j.compag.2021.106252

Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.-C., Pujol, S., Bauer, C., Jennings, D., Fennessy, F., Sonka, M., Buatti, J., Aylward, S., Miller, J. V., Pieper, S., & Kikinis, R. (2012). 3D Slicer as an Image Computing Platform for the Quantitative Imaging Network. *Magnetic Resonance Imaging*, *30*(9), 1323–1341. https://doi.org/10.1016/j.mri.2012.05.001

Finney, E. E., & Norris, K. H. (1973). X-ray images of hollow heart potatoes in water. *American Potato Journal*, *50*(1), 1–8. https://doi.org/10.1007/BF02851513

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382. https://doi.org/10.1037/h0031619

Franck, C., Lammertyn, J., Ho, Q. T., Verboven, P., Verlinden, B., & Nicolaï, B. M. (2007). Browning disorders in pear fruit. *Postharvest Biology and Technology*, *43*(1), 1–13. https://doi.org/10.1016/j.postharvbio.2006.08.008

Gonzalez, J. J., Valle, R. C., Bobroff, S., Biasi, W. V., Mitcham, E. J., & McCarthy, M. J. (2001). Detection and monitoring of internal browning development in "Fuji" apples using MRI. *Postharvest Biology and Technology*, *2*(22), 179–188.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press.

Graves, M., Smith, A., & Batchelor, B. (1998). Approaches to foreign body detection in foods. *Trends in Food Science & Technology*, *9*(1), 21–27. https://doi.org/10.1016/S0924-2244(97)00003-4

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics*, *4*(37). https://doi.org/10.1126/scirobotics.aay7120

Guo, Z., Wang, M., Agyekum, A. A., Wu, J., Chen, Q., Zuo, M., El-Seedi, H. R., Tao, F., Shi, J., Ouyang, Q., & Zou, X. (2020). Quantitative detection of apple watercore and soluble solids content by near infrared transmittance spectroscopy. *Journal of Food Engineering*, *279*, 109955. https://doi.org/10.1016/j.jfoodeng.2020.109955

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*(1–3), 389–422. https://doi.org/10.1023/A:1012487302797

Haff, R., Slaughter, D., Sarig, Y., & Kader, A. (2006). *X-ray assessment of translucency in pineapple*. https://doi.org/10.1111/J.1745-4549.2006.00086.X

Han, D., Tu, R., Lu, C., Liu, X., & Wen, Z. (2006). Nondestructive detection of brown core in the Chinese pear 'Yali'' by transmission visible–NIR spectroscopy.' *Food Control*, *17*(8), 604–608. https://doi.org/10.1016/J.FOODCONT.2005.03.006

Hansen, J. D., Schlaman, D. W., Haff, R. P., & Yee, W. L. (2005). Potential Postharvest Use of Radiography to Detect Internal Pests in Deciduous Tree Fruits. *Journal of Entomological Science*, *40*(3), 255–262. https://doi.org/10.18474/0749-8004-40.3.255

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep Residual Learning for Image Recognition*. http://arxiv.org/abs/1512.03385

Hendrycks, D., Mazeika, M., & Dietterich, T. (2019). Deep anomaly detection with outlier exposure. *Proceedings of the International Conference on Learning Representations*.

Hernández-Sánchez, N., Hills, B. P., Barreiro, P., & Marigheto, N. (2007). An NMR study on internal browning in pears. *Postharvest Biology and Technology*, *44*(3), 260–270. https://doi.org/10.1016/j.postharvbio.2007.01.002

Herremans, E., Melado-Herreros, A., Defraeye, T., Verlinden, B., Hertog, M., Verboven, P., Val, J., Fernández-Valle, M. E.,

Bongaers, E., Estrade, P., Wevers, M., Barreiro, P., & Nicolaï, B. M. (2014). Comparison of X-ray CT and MRI of watercore disorder of different apple cultivars. *Postharvest Biology and Technology*. https://doi.org/10.1016/j.postharvbio.2013.08.008

Herremans, E., Verboven, P., Bongaers, E., Estrade, P., Verlinden, B. E., Wevers, M., Hertog, M. L. A. T. M., & Nicolai, B. M. (2013). Characterisation of "Braeburn" browning disorder by means of X-ray micro-CT. *Postharvest Biology and Technology*. https://doi.org/10.1016/j.postharvbio.2012.08.008

Herremans, E., Verboven, P., Defraeye, T., Rogge, S., Ho, Q. T., Hertog, M. L. A. T. M., Verlinden, B. E., Bongaers, E., Wevers, M., & Nicolai, B. M. (2014). X-ray CT for quantitative food microstructure engineering: The apple case. *Nuclear Instruments and Methods in Physics Research, Section B: Beam Interactions with Materials and Atoms*. https://doi.org/10.1016/j.nimb.2013.07.035

Ho, Q. T., Verboven, P., Verlinden, B. E., Schenk, A., & Nicolaï, B. M. (2013). Controlled atmosphere storage may lead to local ATP deficiency in apple. *Postharvest Biology and Technology*, *78*, 103–112. https://doi.org/10.1016/j.postharvbio.2012.12.014

Huang, Y., Lu, R., & Chen, K. (2020). Detection of internal defect of apples by a multichannel Vis/NIR spectroscopic system. *Postharvest Biology and Technology*, *161*, 111065. https://doi.org/10.1016/J.POSTHARVBIO.2019.111065

imec-Vision Lab, & CWI. (2019). *ASTRA Toolbox* (1.8).

Ioffe, S., & Szegedy, C. (2015). *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. https://arxiv.org/abs/1502.03167v3

Janssen, S., Verboven, P., Nugraha, B., Wang, Z., Boone, M., Josipovic, I., & Nicolaï, B. M. (2020). 3D pore structure analysis of intact 'Braeburn'' apples using X-ray micro-CT.' *Postharvest Biology and Technology*, *159*. https://doi.org/10.1016/j.postharvbio.2019.111014

Janssens, E., Alves Pereira, L. F., De Beenhouwer, J., Tsang, I. R., Van Dael, M., Verboven, P., Nicolaï, B., & Sijbers, J. (2016). Fast inline inspection by Neural Network Based Filtered Backprojection: Application to apple inspection. *Case Studies in*

Nondestructive Testing and Evaluation, *6*, 14–20. https://doi.org/10.1016/j.csndt.2016.03.003

Janssens, E., De Beenhouwer, J., Van Dael, M., De Schryver, T., Van Hoorebeke, L., Verboven, P., Nicolai, B., & Sijbers, J. (2018). Neural network Hilbert transform based filtered backprojection for fast inline x-ray inspection. *Measurement Science and Technology*, *29*(3). https://doi.org/10.1088/1361-6501/aa9de3

Janssens, E., Sijbers, J., Dierick, M., & Beenhouwer, J. D. (2019). Fast detection of cracks in ultrasonically welded parts by inline X-ray inspection. *9th Conference on Industrial Computed Tomography*. https://www.ndt.net/article/ctc2019/papers/iCT2019_Full_paper_48.pdf

Jarolmasjed, S., Espinoza, C. Z., Sankaran, S., & Khot, L. R. (2016). Postharvest bitter pit detection and progression evaluation in 'Honeycrisp'' apples using computed tomography images.' *Postharvest Biology and Technology*, *118*, 35–42. https://doi.org/10.1016/J.POSTHARVBIO.2016.03.014

Jiang, B., He, J., Yang, S., Fu, H., Li, T., Song, H., & He, D. (2019). Fusion of machine vision technology and AlexNet-CNNs deep learning network for the detection of postharvest apple pesticide residues. *Artificial Intelligence in Agriculture*, *1*, 1–8. https://doi.org/10.1016/j.aiia.2019.02.001

Jiang, J. A., Chang, H. Y., Wu, K. H., Ouyang, C. S., Yang, M. M., Yang, E. C., Chen, T. W., & Lin, T. T. (2008). An adaptive image segmentation algorithm for X-ray quarantine inspection of selected fruits. *Computers and Electronics in Agriculture*, *60*(2), 190–200. https://doi.org/10.1016/j.compag.2007.08.006

Joyce, D. C., Hockings, P. D., Mazucco, R. A., Shorter, A. J., & Brereton, I. M. (1993). Heat treatment injury of mango fruit revealed by nondestructive magnetic resonance imaging. *Postharvest Biology and Technology*, *3*(4), 305–311. https://doi.org/10.1016/0925-5214(93)90011-Q

Kak, A. C., & Slaney, M. (2001). *Principles of computerized tomographic imaging*. SIAM.

Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, *147*, 70–90. https://doi.org/10.1016/j.compag.2018.02.016

Karunakaran, C., Jayas, D. S., & White, N. D. G. (2004). Identification of wheat kernels damaged by the red flour beetle using X-ray images. *Biosystems Engineering*, *87*(3), 267–274. https://doi.org/10.1016/j.biosystemseng.2003.12.002

Khatiwada, B. P., Subedi, P. P., Hayes, C., Carlos, L. C. C., & Walsh, K. B. (2016). Assessment of internal flesh browning in intact apple using visible-short wave near infrared spectroscopy. *Postharvest Biology and Technology*, *120*, 103–111. https://doi.org/10.1016/J.POSTHARVBIO.2016.06.001

Kim, S., & Schatzki, T. (2000). Apple watercore sorting system using X-ray imagery: I. Algorithm development. *Transactions of the American Society of Agricultural and Biological Engineers.*, *43*(6), 1695–1702.

Kim, S., & Schatzki, T. (2001). Detection of Pinholes in Almonds through X-ray Imaging. *Transactions of the ASAE*, *44*(4), 997–1003.

Kingma, D. P., & Ba, J. (2014). *Adam: A Method for Stochastic Optimization*. https://arxiv.org/abs/1412.6980

Kotwaliwale, N., Singh, K., Kalne, A., Jha, S. N., Seth, N., & Kar, A. (2014). X-ray imaging methods for internal quality evaluation of agricultural produce. *Journal of Food Science and Technology*. https://doi.org/10.1007/s13197-011-0485-y

Kotwaliwale, N., Weckler, P. R., Brusewitz, G. H., Kranzler, G. A., & Maness, N. O. (2007). Non-destructive quality determination of pecans using soft X-rays. *Postharvest Biology and Technology*, *45*(3), 372–380. https://doi.org/10.1016/J.POSTHARVBIO.2007.03.008

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 25). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

Kutner, M. H., Nachtsheim, Christopher J., Neter, John, & Li, William. (2005). *Applied Linear Statistical Models*. McGraw-Hill Irwin. https://books.google.be/books?id=0xqCAAAACAAJ

Lammertyn, J., Aerts, M., Verlinden, B. E., Schotsmans, W., & Nicolaï, B. M. (2000). Logistic regression analysis of factors influencing core breakdown in “Conference” pears. *Postharvest Biology and Technology*, *20*(1), 25–37. https://doi.org/10.1016/S0925-5214(00)00114-9

Lammertyn, J., Dresselaers, T., Van Hecke, P., Jancsók, P., Wevers, M., & Nicolaï, B. M. (2003a). Analysis of the time course of core breakdown in “Conference” pears by means of MRI and X-ray CT. *Postharvest Biology and Technology*, *29*(1), 19–28. https://doi.org/10.1016/S0925-5214(02)00212-0

Lammertyn, J., Dresselaers, T., Van Hecke, P., Jancsók, P., Wevers, M., & Nicolaï, B. M. (2003b). MRI and X-ray CT study of spatial distribution of core breakdown in “Conference” pears. *Magnetic Resonance Imaging*, *21*(7), 805–815. https://doi.org/10.1016/S0730-725X(03)00105-X

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159–174. JSTOR. https://doi.org/10.2307/2529310

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lee, J. G., Jun, S., Cho, Y. W., Lee, H., Kim, G. B., Seo, J. B., & Kim, N. (2017a). Deep learning in medical imaging: General overview. *Korean Journal of Radiology*, *18*(4), 570–584. https://doi.org/10.3348/kjr.2017.18.4.570

Lee, J. G., Jun, S., Cho, Y. W., Lee, H., Kim, G. B., Seo, J. B., & Kim, N. (2017b). Deep learning in medical imaging: General overview. *Korean Journal of Radiology*, *18*(4), 570–584. https://doi.org/10.3348/kjr.2017.18.4.570

Lim, S., Kim, I., Kim, T., Kim, C., & Kim, S. (2019). Fast AutoAugment. *ArXiv:1905.00397 [Cs, Stat]*. http://arxiv.org/abs/1905.00397

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60–88.

Liu, Z., He, Y., Cen, H., & Lu, R. (2018). Deep feature representation with stacked sparse auto-encoder and convolutional neural network for hyperspectral imaging-based detection of cucumber defects. *Transactions of the ASABE*, *61*(2), 425–436.

Liznerski, P., Ruff, L., Vandermeulen, R. A., Franks, B. J., Kloft, M., & Müller, K.-R. (2021). Explainable Deep One-Class Classification. *ArXiv:2007.01760 [Cs, Stat]*. http://arxiv.org/abs/2007.01760

Lorensen, W. E., & Cline, H. E. (1987). Marching cubes: A high resolution 3D surface construction algorithm. *ACM SIGGRAPH Computer Graphics*, *21*(4), 163–169. https://doi.org/10.1145/37402.37422

Lu, Y., & Lu, R. (2017). Non-Destructive Defect Detection of Apples by Spectroscopic and Imaging Technologies: A Review. *Transactions of the ASABE*, *60*(5), 1765. https://doi.org/10.13031/trans.12431

Maier, A., Steidl, S., Christlein, V., & Hornegger, J. (Eds.). (2018). *Medical Imaging Systems: An Introductory Guide* (Vol. 11111). Springer International Publishing. https://doi.org/10.1007/978-3-319-96520-8

Massey, F. J. Jr. (1951). The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, *46*(253), 68–78.

MATLAB. (2019a). *Image Processing Toolbox*. The MathWorks Inc.

MATLAB. (2019b). *MATLAB*. The MathWorks Inc.

MATLAB. (2019c). *Two-sample Kolmogorov-Smirnov test*. https://nl.mathworks.com/help/stats/kstest2.html#btn37ur

MATLAB. (2020). *MATLAB 2020b*. The MathWorks Inc.

Mazhar, M., Joyce, D., Cowin, G., Brereton, I., Hofman, P., Collins, R., & Gupta, M. (2015). Non-destructive 1H-MRI assessment of flesh bruising in avocado (Persea americana M.) cv. Hass. *Postharvest Biology and Technology*, *100*, 33–40. https://doi.org/10.1016/j.postharvbio.2014.09.006

McCarthy, M. J., Zion, B., Chen, P., Ablett, S., Darke, A. H., & Lillford, P. J. (1995). Diamagnetic susceptibility changes in apple tissue after bruising. *Journal of the Science of Food and Agriculture*, *67*(1), 13–20. https://doi.org/10.1002/jsfa.2740670103

Medeiros, A. D. de, Bernardes, R. C., da Silva, L. J., de Freitas, B. A. L., Dias, D. C. F. dos S., & da Silva, C. B. (2021). Deep learning-based approach using X-ray images for classifying Crambe abyssinica seed quality. *Industrial Crops and Products*, *164*, 113378. https://doi.org/10.1016/j.indcrop.2021.113378

Melado-Herreros, A., Muñoz-García, M.-A., Blanco, A., Val, J., Fernández-Valle, M. E., & Barreiro, P. (2013). Assessment of watercore development in apples with MRI: Effect of fruit location in the canopy. *Postharvest Biology and Technology*, *86*, 125–133. https://doi.org/10.1016/j.postharvbio.2013.06.030

Mercier, S., Villeneuve, S., Mondor, M., & Uysal, I. (2017). Time–Temperature Management Along the Food Cold Chain: A Review of Recent Developments. *Comprehensive Reviews in Food Science and Food Safety*, *16*(4), 647–667. https://doi.org/10.1111/1541-4337.12269

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, *8*(4), 283–298. https://doi.org/10.1016/S0001-2998(78)80014-2

Milesial. (2019). Pytorch-UNet. In *GitHub repository*. GitHub. https://github.com/milesial/Pytorch-UNet

Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, *73*, 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

Muziri, T., Theron, K. I., Cantre, D., Wang, Z., Verboven, P., Nicolai, B. M., & Crouch, E. M. (2016). Microstructure analysis and detection of mealiness in 'Forelle'' pear (Pyrus communis L.) by means of X-ray computed tomography.' *Postharvest Biology and Technology*, *120*, 145–156. https://doi.org/10.1016/j.postharvbio.2016.06.006

My NASA Data. (2021). *Electromagnetic Spectrum Diagram*. https://mynasadata.larc.nasa.gov/basic-page/electromagnetic-spectrum-diagram

Nair, V., & Hinton, G. E. (2010). *Rectified Linear Units Improve Restricted Boltzmann Machines*. 8.

Narvankar, D. S., Singh, C. B., Jayas, D. S., & White, N. D. G. (2009). Assessment of soft X-ray imaging for detection of fungal infection in wheat. *Biosystems Engineering*, *103*(1), 49–56. https://doi.org/10.1016/j.biosystemseng.2009.01.016

Nicolaï, B. M., Beullens, K., Bobelyn, E., Peirs, A., Saeys, W., Theron, K. I., & Lammertyn, J. (2007). Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A

review. *Postharvest Biology and Technology*, *46*(2), 99–118. https://doi.org/10.1016/J.POSTHARVBIO.2007.06.024

Nicolaï, B. M., Defraeye, T., De Ketelaere, B., Herremans, E., Hertog, M. L. A. T. M., Saeys, W., Torricelli, A., Vandendriessche, T., & Verboven, P. (2014). Nondestructive Measurement of Fruit and Vegetable Quality. *Annual Review of Food Science and Technology*. https://doi.org/10.1146/annurev-food-030713-092410

Nugraha, B., Verboven, P., Janssen, S., Wang, Z., & Nicolaï, B. M. (2019). Non-destructive porosity mapping of fruit and vegetables using X-ray CT. *Postharvest Biology and Technology*, *150*, 80–88. https://doi.org/10.1016/j.postharvbio.2018.12.016

Olah, C. (2015). Visualizing representations: Deep learning and human beings. *Colah's Blog*. colah.github.io/posts/2015-01-Visualizing-Representations/

Orina, I., Manley, M., & Williams, P. J. (2017). Use of High-Resolution X-Ray Micro-Computed Tomography for the Analysis of Internal Structural Changes in Maize Infected with Fusarium verticillioides. *Food Analytical Methods*, *10*(9), 2919–2933. https://doi.org/10.1007/s12161-017-0831-4

Otsu, N. (1979). A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, *9*(1), 62–66. https://doi.org/10.1109/TSMC.1979.4310076

Pal, N. R., & Pal, S. K. (1993). A review on image segmentation techniques. *Pattern Recognition*, *26*(9), 1277–1294. https://doi.org/10.1016/0031-3203(93)90135-J

Palenstijn, W. J., Batenburg, K. J., & Sijbers, J. (2011). Performance improvements for iterative electron tomography reconstruction using graphics processing units (GPUs). *Journal of Structural Biology*, *176*(2), 250–253. https://doi.org/10.1016/j.jsb.2011.07.017

Pasquini, C. (2003). Near Infrared Spectroscopy: Fundamentals, practical aspects and analytical applications. *Journal of the Brazilian Chemical Society*, *14*. https://doi.org/10.1590/S0103-50532003000200006

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A.,

Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32* (pp. 8024–8035). Curran Associates, Inc. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Pedreschi, R., Franck, C., Lammertyn, J., Erban, A., Kopka, J., Hertog, M., Verlinden, B., & Nicolaï, B. (2009). Metabolic profiling of "Conference" pears under low oxygen stress. *Postharvest Biology and Technology*, *51*(2), 123–130. https://doi.org/10.1016/j.postharvbio.2008.05.019

Pereira, L. F. A., Janssens, E., Cavalcanti, G. D. C., Ren, T. I., Van Dael, M., Verboven, P., Nicolai, B., & Sijbers, J. (2016). Inline X-ray Computed Tomography system based on Discrete Tomography: Application to agricultural product inspection. *Image and Vision Computing*, *submitted*.

Pereira, L. F. A., Janssens, E., Cavalcanti, G. D. C., Tsang, I. R., Van Dael, M., Verboven, P., Nicolai, B., & Sijbers, J. (2017). Inline discrete tomography system: Application to agricultural product inspection. *Computers and Electronics in Agriculture*, *138*, 117–126. https://doi.org/10.1016/J.COMPAG.2017.04.010

Pietikäinen, M., Hadid, A., Zhao, G., & Ahonen, T. (2011). *Computer vision using local binary patterns* (Vol. 40). Springer Science & Business Media.

Piovesan, A., Vancauwenberghe, V., Looverbosch, T. V. D., Verboven, P., & Nicolaï, B. (2021). X-ray computed tomography for 3D plant imaging. *Trends in Plant Science*, *0*(0). https://doi.org/10.1016/j.tplants.2021.07.010

Razavi, M. S., Asghari, A., Azadbakh, M., & Shamsabadi, H.-A. (2018). Analyzing the pear bruised volume after static loading by Magnetic Resonance Imaging (MRI). *Scientia Horticulturae*, *229*, 33–39. https://doi.org/10.1016/j.scienta.2017.10.011

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S. (2019). Generalized Intersection over Union. *The IEEE*

Conference on Computer Vision and Pattern Recognition (CVPR).

Ronneberger, O., Fischer, P., & Brox, T. (2015). *U-Net: Convolutional Networks for Biomedical Image Segmentation*. http://arxiv.org/abs/1505.04597

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408. https://doi.org/10.1037/h0042519

Rosenblatt, F. (1961). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. CORNELL AERONAUTICAL LAB INC BUFFALO NY. https://apps.dtic.mil/sti/citations/AD0256582

Ruder, S. (2017). An Overview of Multi-Task Learning in Deep Neural Networks. *ArXiv:1706.05098 [Cs, Stat]*. http://arxiv.org/abs/1706.05098

Ruff, L., Kauffmann, J. R., Vandermeulen, R. A., Montavon, G., Samek, W., Kloft, M., Dietterich, T. G., & Müller, K.-R. (2021). A Unifying Review of Deep and Shallow Anomaly Detection. *Proceedings of the IEEE*, *109*(5), 756–795. https://doi.org/10.1109/JPROC.2021.3052449

Ruff, L., Vandermeulen, R. A., Franks, B. J., Müller, K.-R., & Kloft, M. (2021). Rethinking assumptions in deep anomaly detection. *ICML 2021 Workshop on Uncertainty & Robustness in Deep Learning*.

Ruff, L., Vandermeulen, R. A., Görnitz, N., Binder, A., Müller, E., Müller, K.-R., & Kloft, M. (2020). Deep semi-supervised anomaly detection. *International Conference on Learning Representations*. https://openreview.net/forum?id=HkgH0TEYwH

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

Russ, J. C. (2006). *The Image Processing Handbook* (5th ed.). CRC Press. https://doi.org/10.1201/9780203881095

Saito, K., Miki, T., Hayashi, S., Kajikawa, H., Shimada, M., Kawate, Y., Nishizawa, T., Ikegaya, D., Kimura, N., Takabatake, K., Sugiura, N., & Suzuki, M. (1996). Application of magnetic resonance imaging to non-destructive void detection in watermelon.

Cryogenics, *36*(12), 1027–1031. https://doi.org/10.1016/S0011-2275(96)00087-2

Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *ArXiv:1708.08296 [Cs, Stat]*. http://arxiv.org/abs/1708.08296

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *2017 IEEE International Conference on Computer Vision (ICCV)*, 618–626. https://doi.org/10.1109/ICCV.2017.74

Shahin, M. A., Tollner, E. W., Evans, M. D., & Arabnia, H. R. (1999). Watercore features for sorting red delicious apples: A statistical approach. *Transactions of the ASAE*, *42*(6), 1889–1896. https://doi.org/10.13031/2013.13354

Shahin, M. A., Tollner, E. W., & McClendon, R. W. (2001). Artificial intelligence classifiers for sorting apples based on watercore. *Journal of Agricultural and Engineering Research*, *79*(3), 265–274. https://doi.org/10.1006/jaer.2001.0705

Shahin, M., Tollner, E., McClendon, R., & Arabnia, H. (2002). *Apple classification based on surface bruises using image processing and neural networks*. https://doi.org/10.13031/2013.11047

Shen, D., Wu, G., & Suk, H.-I. (2017). Deep Learning in Medical Image Analysis. *Annual Review of Biomedical Engineering*, *19*(1), 221–248. https://doi.org/10.1146/annurev-bioeng-071516-044442

Si, Y., & Sankaran, S. (2016). Computed tomography imaging-based bitter pit evaluation in apples. *Biosystems Engineering*, *151*, 9–16. https://doi.org/10.1016/J.BIOSYSTEMSENG.2016.08.008

Sijbers, J., & Jørgensen, J. S. (2021). Just enough physics. In *Computed tomography: Algorithms, insight, and just enough theory* (pp. 35–54). Society for Industrial and Applied Mathematics.

Srivastava, R. K., Talluri, S., Khasim Beebi, S., & Kumar, R. (2018). Magnetic Resonance Imaging for Quality Evaluation of Fruits: A Review. *Food Analytical Methods*, *11*(10), 2943–2960. https://doi.org/10.1007/s12161-018-1262-6

Statbel. (2018). *Tab A landbouwcijfers 2018*. Landbouwgegevens van 2018. https://statbel.fgov.be/sites/default/files/files/documents/landbo

uw/8.1 Land- en tuinbouwbedrijven/DBREF-L05-2018-TAB-A-NL.xlsx

Steinwart, I., Hush, D., & Scovel, C. (2005). A classification framework for anomaly detection. *Journal of Machine Learning Research*, *6*(2).

Suchanek, M., Kordulska, M., Olejniczak, Z., Figiel, H., & Turek, K. (2017). Application of low-field MRI for quality assessment of 'Conference'' pears stored under controlled atmosphere conditions.' *Postharvest Biology and Technology*, *124*, 100–106. https://doi.org/10.1016/j.postharvbio.2016.10.010

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., & Jorge Cardoso, M. (2017). Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. *Lecture Notes in Computer Science*, 240–248. https://doi.org/10.1007/978-3-319-67558-9_28

Suzuki, K. (2017). Overview of deep learning in medical imaging. *Radiological Physics and Technology*, *10*(3), 257–273. https://doi.org/10.1007/s12194-017-0406-5

Szeliski, R. (2010). *Computer vision: Algorithms and applications*. Springer Science & Business Media.

Tan, W., Zhao, C., & Wu, H. (2016). Intelligent alerting for fruit-melon lesion image based on momentum deep learning. *Multimedia Tools and Applications*, *75*(24), 16741–16761. https://doi.org/10.1007/s11042-015-2940-7

Tharwat, A. (2016). Linear vs. Quadratic discriminant analysis classifier: A tutorial. *International Journal of Applied Pattern Recognition*, *3*(2), 145–180.

Thomas, P., Kannan, A., Degwekar, V. H., & Ramamurthy, M. S. (1995). Non-destructive detection of seed weevil-infested mango fruits by X-ray imaging. *Postharvest Biology and Technology*, *5*(1), 161–165. https://doi.org/10.1016/0925-5214(94)00019-O

Thomas, P., Saxena, S. C., Chandra, R., Rao, R., & Bhatia, C. R. (1993). X-ray imaging for detecting spongy tissue, an internal disorder in fruits of 'Alphonso' mango (Mangifera indica L.). *Journal of Horticultural Science*, *68*(5), 803–806. https://doi.org/10.1080/00221589.1993.11516416

Thompson, A. K., Prange, R. K., Bancroft, R., & Puttongsiri, T. (2018). *Controlled atmosphere storage of fruit and vegetables*. CABI.

Thuyet, D. Q., Kobayashi, Y., & Matsuo, M. (2020). A robot system equipped with deep convolutional neural network for autonomous grading and sorting of root-trimmed garlics. *Computers and Electronics in Agriculture*, *178*, 105727. https://doi.org/10.1016/j.compag.2020.105727

Tollner, E., Gitaitis, R., Seebold, K., & Maw, B. (2005). Experiences with a food product X-ray inspection system for classifying onions. *Applied Engineering in Agriculture*, *21*(5), 907–912.

Tran, T., Pham, T., Carneiro, G., Palmer, L., & Reid, I. (2017). A Bayesian Data Augmentation Approach for Learning Deep Models. *ArXiv:1710.10564 [Cs]*. http://arxiv.org/abs/1710.10564

Van Aarle, W., Palenstijn, W. J., Cant, J., Janssens, E., Bleichrodt, F., Dabravolski, A., Beenhouwer, J. De, Batenburg, K. J., & Sijbers, J. (2016). Fast and Flexible X-ray Tomography Using the ASTRA Toolbox. *Optics Express*, *24*(22), 25129–25147. https://doi.org/10.1364/OE.24.025129

Van Aarle, W., Palenstijn, W. J., De Beenhouwer, J., Altantzis, T., Bals, S., Batenburg, K. J., & Sijbers, J. (2015). The ASTRA Toolbox: A platform for advanced algorithm development in electron tomography. *Ultramicroscopy*. https://doi.org/10.1016/j.ultramic.2015.05.002

van Dael, M., Lebotsa, S., Herremans, E., Verboven, P., Sijbers, J., Opara, U. L., Cronje, P. J., & Nicolaï, B. M. (2016). A segmentation and classification algorithm for online detection of internal disorders in citrus using X-ray radiographs. *Postharvest Biology and Technology*. https://doi.org/10.1016/j.postharvbio.2015.09.020

van Dael, M., Verboven, P., Dhaene, J., Van Hoorebeke, L., Sijbers, J., & Nicolai, B. (2017). Multisensor X-ray inspection of internal defects in horticultural products. *Postharvest Biology and Technology*, *128*, 33–43. https://doi.org/10.1016/j.postharvbio.2017.02.002

van Dael, Verboven, P., Zanella, A., Sijbers, J., & Nicolai, B. (2019). Combination of shape and X-ray inspection for apple internal quality control: In silico analysis of the methodology based on X-ray computed tomography. *Postharvest Biology and*

Technology, *May*, 0–1. https://doi.org/10.1016/j.postharvbio.2018.05.020

VCBT. (2017). *Bewaarcondities Appel en Peer*. http://vcbt.be/bewaarcondities_appel_en_peer/

Veltman, R. H., Lenthéric, I., Van Der Plas, L. H. W., & Peppelenbos, H. W. (2003). Internal browning in pear fruit (Pyrus communis L. cv Conference) may be a result of a limited availability of energy and antioxidants. *Postharvest Biology and Technology*, *28*(2), 295–302. https://doi.org/10.1016/S0925-5214(02)00198-9

VLAM. (2021a). *Belgisch thuisverbruik van groenten en fruit (2020)*.

VLAM. (2021b). *Export van vers fruit (2011-2020)*.

Walsh, K. B., Blasco, J., Zude-Sasse, M., & Sun, X. (2020). Visible-NIR 'point' spectroscopy in postharvest fruit and vegetable assessment: The science behind three decades of commercial use. *Postharvest Biology and Technology*, *168*, 111246. https://doi.org/10.1016/j.postharvbio.2020.111246

Wang, C. Y., & Wang, P. C. (1989). Nondestructive detection of core breakdown in "Bartlett" pears with nuclear magnetic resonance imaging. *HortScience (USA)*. https://agris.fao.org/agris-search/search.do?recordID=US8905668

Wang, S. Y., Wang, P. C., & Faust, M. (1988). Non-destructive detection of watercore in apple with nuclear magnetic resonance imaging. *Scientia Horticulturae*, *35*(3–4), 227–234. https://doi.org/10.1016/0304-4238(88)90116-1

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612. https://doi.org/10.1109/TIP.2003.819861

Wang, Z., Herremans, E., Janssen, S., Cantre, D., Verboven, P., & Nicolaï, B. (2018). Visualizing 3D Food Microstructure Using Tomographic Methods: Advantages and Disadvantages. *Annual Review of Food Science and Technology*, *9*(1), 323–343. https://doi.org/10.1146/annurev-food-030117-012639

Wang, Z., Hu, M., & Zhai, G. (2018). Application of Deep Learning Architectures for Accurate and Rapid Detection of Internal Mechanical Damage of Blueberry Using Hyperspectral

Transmittance Data. *Sensors*, *18*(4), 1126. https://doi.org/10.3390/s18041126

Wang, Z., Van Beers, R., Aernouts, B., Watté, R., Verboven, P., Nicolaï, B., & Saeys, W. (2020). Microstructure affects light scattering in apples. *Postharvest Biology and Technology*, *159*, 110996. https://doi.org/10.1016/j.postharvbio.2019.110996

Wevers, M., Nicolaï, B., Verboven, P., Swennen, R., Roels, S., Verstrynge, E., Lomov, S., Kerckhofs, G., Van Meerbeek, B., Mavridou, A. M., Bergmans, L., Lambrechts, P., Soete, J., Claes, S., & Claes, H. (2018). Applications of CT for Non-destructive Testing and Materials Characterization. In S. Carmignato, W. Dewulf, & R. Leach (Eds.), *Industrial X-Ray Computed Tomography* (pp. 267–331). Springer International Publishing. https://doi.org/10.1007/978-3-319-59573-3_8

Willemink, M. J., De Jong, P. A., Leiner, T., De Heer, L. M., Nievelstein, R. A. J., Budde, R. P. J., & Schilham, A. M. R. (2013). Iterative reconstruction techniques for computed tomography Part 1: Technical principles. *European Radiology*, *23*(6), 1623–1631. https://doi.org/10.1007/s00330-012-2765-y

Yang, L., Yang, F., & Noguchi, N. (2011). Apple Internal Quality Classification Using X-ray and SVM. *IFAC Proceedings Volumes*, *44*(1), 14145–14150. https://doi.org/10.3182/20110828-6-IT-1002.01827

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32–35. https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3

Yu, X., Lu, H., & Wu, D. (2018). Development of deep learning method for predicting firmness and soluble solid content of postharvest Korla fragrant pear using Vis/NIR hyperspectral reflectance imaging. *Postharvest Biology and Technology*, *141*, 39–49. https://doi.org/10.1016/j.postharvbio.2018.02.013

Zhang, M., Jiang, Y., Li, C., & Yang, F. (2020). Fully convolutional networks for blueberry bruising and calyx segmentation using hyperspectral transmittance imaging. *Biosystems Engineering*, *192*, 159–175. https://doi.org/10.1016/j.biosystemseng.2020.01.018

Zhang, Y., Mehta, S., & Caspi, A. (2021). Rethinking Semantic Segmentation Evaluation for Explainability and Model

Selection. *ArXiv:2101.08418 [Cs]*. http://arxiv.org/abs/2101.08418

Zhou, L., Zhang, C., Liu, F., Qiu, Z., & He, Y. (2019). Application of Deep Learning in Food: A Review. *Comprehensive Reviews in Food Science and Food Safety*, *18*(6), 1793–1811. https://doi.org/10.1111/1541-4337.12492

Zion, B., Chen, P., & McCarthy, M. J. (1995). Detection of bruises in magnetic resonance images of apples. *Computers and Electronics in Agriculture*, *13*(4), 289–299. https://doi.org/10.1016/0168-1699(95)00027-5

Publications

International journal articles

Published

Van de Looverbosch, T., Raeymaekers, E., Verboven, P., Sijbers, J., Nicolai, B. (2021). Non-destructive internal disorder detection of Conference pears by semantic segmentation of X-ray CT scans using deep learning. *Expert Systems with Applications*, *176*, doi: 10.1016/j.eswa.2021.114925 Open Access

Piovesan, A., Vancauwenberghe, V., **Van De Looverbosch, T.**, Verboven, P., Nicolai, B. with Verboven, P. (2021). X-ray computed tomography for 3D plant imaging. *Trends In Plant Science*, 1-15. doi: 10.1016/j.tplants.2021.07.010

Piovesan, A., **Van De Looverbosch, T.**, Verboven, P., Achille, C., Cabrera, C.P., Boller, E., Cheng, Y., Ameloot, R., Nicolai, B. (2020). 4D synchrotron microtomography and pore-network modelling for directin situcapillary flow visualization in 3D printed microfluidic channels. *Lab on a Chip*, *20* (13), 2403-2411. doi: 10.1039/d0lc00227e Open Access

Van De Looverbosch, T., Bhuiyan, M.H R., Verboven, P., Dierick, M., Van Loo, D., De Beenbouwer, J., Sijbers, J., Nicolai, B. (2020). Nondestructive internal quality inspection of pear fruit by X-ray CT using machine learning. *Food Control*, *113*, doi: 10.1016/j.foodcont.2020.107170 Open Access

Gruyters, W., **Van De Looverbosch, T.**, Wang, Z., Janssen, S., Verboven, P., Defraeye, T., Nicolai, B. (2020). Revealing shape variability and cultivar effects on cooling of packaged fruit by combining CT-imaging with explicit CFD modelling. *Postharvest Biology and Technology*, *162*. doi: 10.1016/j.postharvbio.2019.111098

Under review

Van De Looverbosch, T., Vandenbussche, B., Verboven, P., Nicolai, B. (Submitted April 19, 2021). Nondestructive high-throughput sugar beet fruit analysis using X-ray CT and Deep Learning. *Computers and Electronics in Agriculture*

Van De Looverbosch, T., Jiaqi, H., Tempelaire, A., Kelchtermans, K., Verboven, P., Tuytelaars, T., Sijbers, J., Nicolai, B. (Submitted November 17, 2021). Inline Nondestructive Internal Disorder Detection in Pear Fruit using Explainable Deep Anomaly Detection on X-ray images. *Computers and Electronics in Agriculture*

Conference proceedings

Van De Looverbosch, T., Raeymaekers, E., Verboven, P., Sijbers, J., Nicolai, B. (2020). Non-destructive identification and quantification of internal disorder in of Conference pears using X-ray CT imaging and Deep Learning. Presented at the Postharvest 2020 Webinar Series, New Zealand (held online), 10 Nov 2020-12 Nov 2020.

Piovesan, A., **Van De Looverbosch, T.**, Verboven, P., Achille, C., Parra Cabrera, C., Boller, E., Cheng, Y., Ameloot, R., Nicolai, B. (2020). Dynamic synchrotron microtomography and pore-network modelling for direct in-situ capillary flow observation in 3D printed lab-on-chips. Presented at the Interpore 2020, 12th annual meeting, Qingdao, China. (Held online).

Van De Looverbosch, T., Verboven, P., Nicolai, B. (2019). Non-destructive internal quality inspection of fruit using X-ray imaging and artificial intelligence. Presented at the 33rd EFFoST International Conference 2019, Rotterdam, The Netherlands, 12 Nov 2019-14 Nov 2019.

Van De Looverbosch, T., Verboven, P., Sijbers, J., Nicolai, B. (2019). An efficient X-ray projection simulator of 3D fruit shapes for use in non-destructive internal quality inspection. Presented at the 4th International Conference on Food and Biosystems Engineering, Agia Pelagia, Heraklion, Creta, 30 May 2019-02 Jun 2019.

Piovesan, A., **Van De Looverbosch, T.**, Verboven, P., Achille, C., Boller, E., Cheng, Y., Ameloot, R., Nicolai, B. (2019). Dynamic synchrotron microtomography for direct in-situ capillary flow visualization in functionalized porous material for passive microfluidics. Presented at the Interpore 2019, 11th annual meeting, Valencia (Spain). Open Access

Van De Looverbosch, T., De Beenhouwer, J., Boone, M., Van Loo, D., Wagner, A., Verboven, P., Sijbers, J., Nicolai, B. (2018). Development of a non- destructive inspection algorithm for detection of internal disorders in apple using online x-ray computed tomography. Presented at the 32nd EFFoST International Conference, Nantes, France, 06 Nov 2018-08 Nov 2018.