Strategic relevance of Artificial Intelligence

A case study on image recognition in the medical sector

Thesis by Rasmus Emil Hansen

Cand.merc.it, CBS

Co-authored by Giorgios Kritsotakis

Department of Computer Science, KU

Submission date: 15-09-2017

Supervisor: Daniel Hardt

Number of characters: 172,200 (with illustrations)

Abstract

In order to discover how a medical company can model a Machine Vision application and communicate the performance findings to the CEO, this thesis has based its research on a case study of Højgaard Equine Hospital. The case study allowed the research to identify, how automation of a specific process could provide benefits to a medical company, including labour cost savings, mitigation of false negative diagnoses and increase the concistency of diagnoses across the industry.

The identified process included detection of the disease OCD on the hock of horses. This process could be automated by use of image recognition (and therefore Machine Vision) as it involves a computer-aided detection task, where a pattern is discovered from looking through x-rays. In order to understand how to model this type of process, the research has applied concepts from Deep Learning, Machine Learning and Convolutional Neural Networks.

The research analysis followed the CRISP-DM model in order to integrate data mining into the business of the case company. The specific business goal was to achieve a model performance above 0,50 (Recall) and a positive expected value. To overcome the issue with lack of data, different data augmentation techniques have been used to scale up the data to approximately 2600 unique x-rays. The training and testing of model 1 have been executed on the IBM Watson platform by use of a pre-trained CNN, where only the final layer has been trained.

The performance of model 1 includes a Recall of 0,47 and an expected profit loss of -60,45 DKK for the Discrete Classifier and a Recall of 0,75 and max profit of 170,72 DKK for the Ranking Classifier. The performance of model 2 includes a Recall of 0,66 and an expected profit of 173,28 DKK for the Discrete Classifier and a Recall of 0,85 and max profit of 211,74 DKK for the Ranking Classifier. The performance of model 2 is superior to model 1 and model 2 fulfills the business goal. The performance of the models can be communicated to the CEO by use of profit curves and ROC graphs, which are useful metrics for business decision-making.

The deployment of the model is based on a 4-step approach, which takes into account the associated risk & maturity when integrating the model into the existing business landscape. The recommendation is to use the model for decision support.

Reflections on the research process produced findings that are relevant to implement for a future research, i.e. more frequent meetings with domain experts as well as starting the data collection much earlier in the research process. Reflections on ethical issues suggested that humans should not prevent automation, but accept it, and direct their focus towards more value-adding activities.

The findings of this research contribute to the research field of image recognition applied within the medical sector as they explain how to model and communicate two models that are based on convolutional neural networks.

Preliminary comments & acknowledgements

This thesis is written by the author himself and with contributions from his thesis partner. The thesis partner is a master student at Copenhagen University – Department of Computer Science (DCS). The scope of collaboration is restricted to the data mining part, including data collection, data pre-processing & data augmentation, as well as knowledge sharing and discussions on data science concepts. As this part of the research traditionally is very time-consuming, the collaboration resulted in significant timesavings as well as important knowledge sharing on the findings.

Each student has conducted most of the research themselves, including the modelling phase. However, both thesis complement each other, as their common purpose is to discover useful applications of image recognition on the same dataset, but with different modelling approaches. This thesis includes a benchmark of each model's performance and it therefore includes the technical performance results and architecture of model 2, which has been created by the thesis partner. This will provide a solid baseline for deciding on both models' feasibility with the case study.

We wish to extend our heartfelt gratitude to all the people who helped us make this research possible. First, we would like to thank our supervisors, Daniel Hart (DOD), Stefan Sommer (DCS) and Sune Darkner (DCS). We are also grateful for the collaboration with Jørgen Michael Hansen, whom without access to his x-ray database and insights on the veterinarian industry, would not have allowed us to conduct any research at all. In addition, a great thank to the expert interviewee, Jacob Axelsen from Deloitte, for taking his time to provide useful insights and feedback.

Table of contents

1 Introduction

In 2015, a deep-learning machine named Enlitic competed against expert human diagnostic radiographers in diagnosing lung cancer. Enlitic won the competition. It was built to read X-rays and CT scans and was about 50 percent better at classifying tumors and had a false-negative rate of zero (where the disease is missed), compared with 7 percent for humans. (Straitstimes, 2017)

The above accomplishment is remarkable. Today we are halfway through 2017 and the smart machines (another word for a machine with deep learning capabilities) are expected to replace traditional human jobs. Many consider this wave of smart machines a feature of the fourth industrial revolution (RBR, 2016). The machines leverage artificial intelligence to perform tasks that until recently were perceived to be human domain. This is mainly due to the explosion of data and improvements in hardware and software infrastructure, which have enabled smart machines to apply deep learning models (a subarea of Artificial Intelligence) in advanced analytics. In the example above, Enlitic is able to diagnose tumors since it has been trained with thousands or more images consisting of X-rays and CT scans. This has created a model that can look for patterns in the data including patterns associated with tumors. Not only is the machine more efficient, it is also much faster at providing a diagnosis compared to human radiologists.

Currently, Artificial Intelligence ("AI") is at the top of Gartner's hype curve, which also includes fields like machine learning, cognitive advisors/chatbots and smart robots (Gartner, 2016). At this point in time it can therefore be difficult to cut through the noise and decide where AI is going to have a practical impact. This is also because most AI research has been limited to academic fields, and still lacks exploration of practical applications. Many businesses are therefore experimenting with these techniques in order to unveil if deep learning can be used within their business area. As an example, Enlitic shows how smart machines can influence the medical sector and the jobs of radiologists.

Within the medical sector, especially IBM Watson ("Watson") has proven to be at the "*forefront of AI in the medical sector*" and the machine has successfully been used for diagnosing lung cancer and heart diseases (ITN, 2017). Companies can license Watson and build their own models on top of the powerful machine that provides them with the necessary infrastructure. As opposed to building up a deep learning model from scratch in e.g. Google Tensorflow, business can take advantage of Watson's pretrained model. This can be beneficial for companies that do not have the capabilities to experiment with deep learning models on a very technical level, but still wants to look into the opportunities it has.

Considering the potential of AI applied in the medical sector, it is interesting to further explore the practical impact of AI. Detecting diseases based on patterns derived from thousands of images (e.g. X-rays) is a task that potentially could be managed by smart machines. This application area is called Machine Vision "MV" and the specific task for detecting a disease (called "image recognition") is often solved by a CNN algorithm ("Convolution Neural Network"), like the case of Enlitic (Venturebeat, 2014). CNNs are a type of deep neural network that are especially

useful for image recognition. The practical application of image recognition has proved itself for computer-aided detection of lung cancer (example above) and in general, a lot of research is being conducted within this field. However, there are still diseases that could be detected by use of image recognition and further research ought to be conducted in this field.

Besides a certain level of technical capabilities, conducting research within image recognition requires access to sufficient data to train the models on. This can be difficult in the medical sector, where most data are very confidential and therefore influenced by strict policies and regulations. Thus, getting access to relevant data is a fundamental condition for doing experimentations with image recognition.

1.1 Research question

Based on the above considerations, the purpose with this thesis is to further investigate practical applications of image recognition within the medical sector, including how image recognition can be used to detect diseases on medical data. In specific, this thesis aims at solving the following research question:

How can a medical company develop a machine vision application and communicate its performance to the CEO?

This research question does not only include the modelling of an MV application, but also an explanation of the business relevance. This is important, since business people might not have the necessary technical understanding to really capture the impact of AI on their business. If the CEO does not understand the impact of the MV model on his business, the model will not be adopted, and it will not have any practical relevance.

"Medical company" refer to a case company that fulfill three conditions:

1) The company has a relevant business problem that could be solved with MV
2) The company has access to sufficient and relevant medical data
3) The company accepts a collaboration that allow flexibility for doing experimentations with their data

With "develop" this research refers to the data mining process of creating a model by use of data science techniques and "machine vision application" refer to a model that is capable of doing image recognition. This delimits the scope of analysis from other machine vision tasks like motion analysis and scene reconstruction (Klette, 2014). Finally, "communicate" refers to the process of documenting the model's performance and using visualizations to communicate the impact to the business.

1.2 Contribution to research

This thesis contributes to the research field of artificial intelligence, by being one of the few scholarly attempts to examine practical applications of image recognition in the medical sector. Scholars have paid only limited amount of attention to detection

of *osteochondritis dissecans* (OCD), since most research is focused on diagnosing cancer (Forbes, 2017).

Specifically, to the author's knowledge, this thesis is the first study that applies image recognition on x-rays from pre-existing clinical data warehouses in order to detect OCD diseases on equines. According to a comparative study by A.M. McCoy et al. (2013) there is a broad range of similarities between OCD affecting humans and animals including "*radiographic"* similarities (the study involved horses, pigs and humans) (McCoy, et al., 2013). This suggests that the thesis' research could contribute to comparative research studies on OCD diseases affecting humans as well, which makes the research even more interesting.

Since the medical data is confidential, it is difficult if not impossible for "outsiders" to replicate the results of this research. However, this is the case with most medical data and it is therefore not considered to be a specific problem for this particular research, but a necessary circumstance for most research within the medical domain.

1.3 Assumptions

It is acknowledged that the author does not have full knowledge of the problem domain prior to the research, and he is therefore aware that he continuously learns more throughout the process. However, during this process, he has become increasingly aware of the biases, as-is assumptions, and his prejudice towards creating a successful MV model. These biases and as-is assumptions have gradually been modified by his interaction with theory and empirical data.

One assumption is that we assume deep learning algorithms to be the most useful for solving the specific business problem. However, according to the authors of the *Deep Learning* book, machine learning is "*the only viable approach to building AI systems*" (Goodfellow, Bengio, & Courville, 2016, p. 8), which means that other machine learning algorithms might as well be useful. However, within this thesis time constraints and limited available resources, it has only been possible to focus on the application of a specific kind of machine learning algorithm, which is also included within the scope of deep learning. A benchmark with *other types* of machine learning algorithms, like SVMs, is therefore out of scope for this thesis. However, the research will provide a benchmark with a model developed from the *same type* of algorithm, but on a different platform.

Another assumption includes the amount of data. During the data preparation process the author and his thesis partner believed there would be plenty of data to reach at least 5000 labeled images for both positive and negative classifications. However, as the data collection progressed, it was apparent that not enough data were labelled as positives. This was due to different uncertainties that we discovered along the way, including:

- We did not know the exact number of positives or negatives beforehand. This was something we would find out after documenting all the journals in the spreadsheet.
- The exact number of positives was far lower than what we expected.
- Some positive diagnosed horses that were documented from the physical journals could not be found in the clinical database containing the x-rays.
- Only 2-4 x-rays per positive diagnosed horse (out of 14) could be labeled with OCD on the hock.
- Some positive diagnosed horses had wrong diagnoses.

As deep learning algorithms usually require a lot of data, the data preparation part proved to be challenging. The lack of data has clearly influenced the performance of our models even though we have applied different techniques to scale it up. Our initial assumption has therefore influenced the direction and final results of our research.

1.4 Limitations

The focus of this thesis is not on tuning parameters that may increase model performance – nor is it a focus to design a completely new neural network. This will be the focus area of the thesis created by the author's thesis partner. Instead, the training of data will happen on a pretrained model. This will limit the complexity of developing a model, and leave room for considering the performance of the model and its feasibility with the business.

1.5 Research approach

In order to answer the above research question, this thesis will explain the research design, including the philosophy of science and methodological approach. This chapter also presents the data collection methods and introduce a case study of a Danish veterinarian company that has access to sufficient medical data.

Since AI covers different research fields, the thesis will conduct a literature review to identify the most relevant concepts. This will help understand relevant methods and techniques used to build a MV model that can perform image recognition.

As the process of discovering patterns in datasets is a data mining activity, the thesis will base its model building approach on the CRISP model. This will provide a structured procedure that also considers a business perspective. The CRISP procedure will include business topics like "process automation", "profit models" and "risk management" along the way in order to provide sufficient depths to relevant business issues.

In order to put the performance of the final model into perspective, the thesis will include a benchmark with another MV model that has been developed from the same medical data, but with another platform. This will allow the case company to better evaluate the impact of the model.

The analysis & results sections will be followed up with a discussion of important empirical insights from the data mining process, as well as a reflection on the ethical concerns when deploying and putting AI into practice.

Finally, the thesis will conclude on the research and account for how to proceed with these findings for future research.

2 Methodology

This section will describe the methodology of the thesis. It will be structured in accordance with "the research onion" which gives a structured approach to navigate between underlying assumptions and choices that existed prior to the data collection and analysis in this research (Saunders, Lewis, & Thornhill, Research methods for business students, 2009).

2.1 Research design

This section will describe the research design of the thesis, which has been structure in accordance with the research onion. (Saunders, Lewis, & Thornhill, Research methods for business students, 2009)

2.1.1 Research philosophy

The pragmatic philosophy has been chosen as the focus of research, as thisphilosophy emphasizes the findings' practical consequences (Saunders & Tosey, The layers of research design, 2012). This paradigm accepts both positivism and constructivism as ontological frameworks as long as they support the purpose of discovering practical findings that can be adopted by the case company. This is useful to this research as we are interested in research that both includes interpreting the contextual situation faced by the case company, as well as objectively deducing knowledge from the medical data. (Saunders, Lewis, & Thornhill, Research methods for business students, 2009)

2.1.2 Research approach

With the pragmatic research philosophy, the research approach is focused on what provides practical findings, which often favors a combination of the deductive and inductive approaches (Saunders & Tosey, The layers of research design, 2012). Thus, this thesis will start with an inductive approach to gain insights on the research phenomenon of deep learning and the case company by researching secondary data from the Internet and primary data from interviews. The inductive approach will help us formulate the necessary criteria for the model to be put into practice. When building the model and performance metrics, the thesis will deduct results that justify whether the model is suitable for being put into practice. (Saunders, Lewis, & Thornhill, Research methods for business students, 2009)

2.1.3 Research strategy

The purpose with the research strategy is to conduct an explorative study. Exploratory studies are valuable in finding insights on "*what is happening; to seek*

new insights; to ask questions and to assess phenomena in a new light" (Saunders et al., 2009, p. 139). Since there is, to the author's knowledge, no existing academic research on how to apply image recognition on medical data to detect OCD disease on horses, the primary data found in this thesis will lay the foundation for understanding this subject.

The research strategy is based on both *experimental research* and a *single case study*. This hybrid approach allows the research to assess the feasibility of the generated models with the case company's business. (Saunders, Lewis, & Thornhill, Research methods for business students, 2009)

The experimental research process examines the results of an experiment against the expected results (Saunders, Lewis, & Thornhill, Research methods for business students, 2009). With this strategy, the thesis intends to iteratively produce results from the model experimentations and assess the final model's feasibility with the case company's business.

The case study is a qualitative approach that is used to address a specific challenge or theory (Tellis, 1997). This allow us to conduct an in depth investigation of the research area where real-life conditions can be used (Zainal, 2007). This is necessary to understand how the results from the experiments fit into the business of the case company. More specifically: What business criteria should the model meet in order to be put into practice.

2.1.4 Methodological choice

This thesis applies a multi method approach, which refer to the use of both a qualitative and a quantitative methodology. This approach splits the research into separate segments, with each producing a specific dataset (Saunders, Lewis, & Thornhill, Research methods for business students, 2009).

The qualitative techniques, including the interviews, will provide an in-depth understanding of the motivations behind the case company's situation including *why* image recognition can provide value to the business. The quantitative techniques will be used to build the model and performance metrics from the collected medical data. These techniques will help understand *how* the case company can benefit from applications of image recognition. The techniques come from the deep learning and machine learning academic fields and are mathematically founded. They include consequence chains of input action/data/methods, which, used in the same combinations, will follow consistent paths towards the same output and are therefore based on an exact science.

Time horizon

AI is a rapidly evolving field, which favors a cross sectional study that is focused on a phenomenon at a particular time (Saunders, Lewis, & Thornhill, Research methods for business students, 2009). Saunders et al. (2009) describes a cross sectional

study as a "*snapshot taken at a particular time*". This snapshot will help describe the current situation of image recognition applied on a specific kind of medical data. However, further studies and more research should be conducted following this study, since this field is evolving rapidly.

2.1.5 Techniques and procedures

This section looks closer at the data analysis and data collection approaches.

2.1.5.1 Data analysis

The choice of data analysis tools has a huge impact on the research product. Especially the choice of platform for developing the model will decide the flexibility in terms of parameter tuning. This influences what options are available for optimizing the performance of the model and thereby the model's final feasibility with the business. This will be discussed further below.

For now it is relevant to know that IBM Watson will be used as platform, including the Visual Recognition Service, which allows the research to benefit from Watson's infrastructure and processing power and is necessary for training the model. The command procedures for training and testing are very straightforward and can be found on https://www.ibm.com/watson/developercloud/doc/visual-recognition/tutorial-custom-classifier.html. In order to use Watson one only has to create a Bluemix account, which is free for one month.

The testing of the model is based on reused code from Github: https://github.com/joe4k/wdcutils. This includes installing the Jupyter Notebook package, which contains the iPython command shell. This shell is used for testing the model and outputting results to csv files.

The processing of results, including calculations and computations of graphs, has been conducted in Microsoft Excel and Nodepad++.

In the "Future Outlook" section this research portrays an image, which shows the x-ray with tiles. This analysis is based on reused code from Github: https://github.com/IBM-Bluemix/Visual-Recognition-Tile-Localization?cm_mc_uid=96470940168114929797367&cm_mc_sid_50200000=1501614401&cm_mc_sid_52640000=1501614401. It includes installing node.js as runtime system and npm as package manager. It allows the user to upload medical data to the node.js application in the browser, which then "chops" the data into tiles, which are then analysed by Watson and finally output as "heat-map" visualizations.

2.1.5.2 Data collection

Secondary data

Secondary data is data that has already been collected for a purpose other than for this thesis (Saunders, Lewis, & Thornhill, Research methods for business students, 2009). In order to gain insight on deep learning theories including how to develop a MV model, this thesis has included external secondary data sources from the Internet and CBS library research database. The main academic research literature includes the "Deep Learning" book by Ian Goodfellow, Yoshua Bengio and Aaron Courville (Goodfellow, Bengio, & Courville, 2016) and "Data Science for Business" by Foster Provost and Tom Fawcett (Provost & Fawcett, 2013). Information from Stackoverflow and IBM Watson websites have been collected in order to gain an understanding of the necessary techniques and methods used to develop models.

Since the original purpose with creating the medical data was a purely business purpose and not a research purpose, the medical data collected for this thesis are considered to be internal secondary data sources. However, considering that this thesis is a feasibility study and that the medical data has been automatically produced and processed with almost no human involvement (besides conducting the x-ray examination), we have no issues trusting the integrity of the core dataset. But since the diagnosis of the horse is very much influenced by human judgement, the labeling of the medical data (our data preparation), i.e. positive or negative classifications, has some influence on the integrity of the data. This has resulted in some misclassifications, which will be further explained in the analysis section.

As the medical data are confidential, the author has secured preapproval from the CEO of the case company to include any kind of reference to these data in the thesis.

Primary data

The collected primary data includes a semi-structured interview with the CEO of the case company and an in-depth interview with a domain expert (within deep learning). We were motivated to collect primary data rather than secondary data from a multiple case review, as there does not exist much case literature on this particular research field.

The interviewees were chosen, as they comply with the following criteria:

- Representative from a medical company with in-depth knowledge about:
    - The business in general
    - The use of medical data
- Domain expert within practical applications of:
    - Deep learning
    - Image recognition

<u>Semi-structured and in-depth interviews</u>

Semi-structured interviews provide an opportunity to 'probe' answers, where it is desired that the interviewees explain, or elaborate on, their responses (Saunders, Lewis, & Thornhill, Research methods for business students, 2009). This is important when the focus is on understanding the meanings that participants ascribe to various phenomena. It will allow the research to understand the business of the case company, including elaborations on the specific parts of the business that comprise

medical data. The findings will be implemented in the case study as well as in the analysis section. The semi-structured interview should cover the following *themes*:

- Strategy
- Industry
- Clients
- Activities involving medical data
- Use of medical data

In-depth interviews, also called unstructured interviews, are informal and are used to explore in depth a topic of interest (Saunders, Lewis, & Thornhill, Research methods for business students, 2009). There is no predetermined list of questions to work through. However, there needs to be a clear idea about the aspects that are to be explored.

This interview method will allow the research to gain insights on "*image recognition and how to put it into practice*". Specifically, it will help the author understand how to approach the research field and understand which platforms are most suitable for solving the research question considering the time limits and available resources. The interviewee was presented only to the scope of research and case company, which formed the basis for how the conversation evolved.

Ad. Semi-structured interview with Jørgen Michael Hansen

Jørgen Michael Hansen, CEO at Højgaard Equine Hospital, was chosen as interviewee due to his daily management of the business as CEO and his experience with x-ray examinations and medical data (32 years of experience). The full interview is attached in Appendix 2.

Ad. Indepth interview with Jacob Axelsen

Jacob Axelsen from Deloitte was chosen as interviewee, due to his academic background and experiences from putting deep learning into practice. He holds a Ph.D. that involves biological neural networks and has contributed, through his work as management consultant, to several proofs of concepts for businesses that involve the development of deep learning models and their feasibility with the businesses. The full interview is attached in Appendix 1.

2.1.6 Subconclusion

To conclude on the methodology section, the research onion has structured the methodological approach consistently throughout the thesis. Hence, the research illustrates a coherent and applicable methodology, enabling the conclusions reached to be logical, valid, relevant and of high quality.

Following the pragmatic approach, it is now relevant to look at the case study that can give an understanding of the case company's business, including how they use medical data. The insights from this study will be used in the "business understanding" section, which is the first step in developing a feasible model.

2.2 Case study – Højgaard Equine Hospital

2.2.1 Introduction

Højgaard Equine Hospital ("Højgaard") is the largest animal hospital in Denmark, which is specialized in equines. The hospital has 30 employees, including 13 veterinarians, 6 veterinary nurses, one radiographer, one farrier, office and stable staff (Højgaard, 2017). On an annual basis, they execute approximately 12.200 consultations at the hospital. Their target markets include "patients" from Denmark, Sweden, Norway and Germany and all together they have around 15.000 clients. Their CEO is Jørgen Michael Hansen ("Michael"), who is also the chief surgeon. (Højgaard, 2017)

Their vision is to "*put the horse in focus*" and "*make a difference*". Their mission is to be the leading hospital in terms of professional knowledge as well as being able to offer clients and "patients" the best available diagnostics and treatment. This also includes acquisition of the most advanced devices in order to better diagnose the animals. (Højgaard, 2017)

Currently, Højgaard's turnover amounts to 25 million DKK (2016) per year and aims at a surplus of ten percent annually. This goal is not always achieved, but this is not an issue as Michael puts it: *"Our current focus is to keep a "healthy" business in order to be able to keep developing our business"* (fyens.dk, 2016).

2.2.2 Focus on innovation

Højgaard sees itself as a *knowledge company*, which emphasizes their focus on research and development. They believe innovation is a necessity to stay competitive. That is why Højgaard seeks to attract veterinarians with highly specific specializations in equine knowledge domains. Also, they have invested in e.g. a MRI scanner, and are currently the only ones in the Danish equine industry who can provide MRI scans. Even though this investment is Højgaard's largest device investment, it has not generated any positive ROI for years. However, this is not a problem according to the CEO, since this device is important for attracting new segments, acquiring knowledge and in general developing the business. Clients come from countries abroad just to have their horse MRI scanned (fyens.dk, 2016).

In general – when benchmarking with the human industry, Michael believes the Equine industry is a couple of years behind in terms of technology developments, but the many technologies for diagnosing humans are quickly adopted in the equine industry as well - like the MRI scanner.

To sum up, Højgaard's focus on innovation and their willingness to invest in highly sophisticated technology gives them a competitive edge in the market.

2.2.3 The industry for equine x-ray examinations

Højgaard competes in different industries depending on what kind of consultations they provide. Højgaard has the necessary equipment for producing x-rays of horses and one of their major markets for equine consultations is the industry for x-ray examinations. In this industry, they compete on three segments that complement each other: X-ray production, diagnostic and surgery segments. These will be elaborated on in the following.

The x-ray production segment

The x-ray production segment is targeted by several equine veterinarians and not only by veterinary businesses, but any business that has access to x-ray equipment. All players provide cheap productions of x-rays, which is possible since the equipment for producing x-rays is cheap and accessible. This market competes on price.

The diagnostic segment

The diagnostic segment is targeted by fewer players, including Højgaard. Højgaard offers one service that includes x-ray production and diagnosis as a bundle package. This segment generates a major part of Højgaard's revenue and is characterized by two types of clients (their names are fictive and only made up for reference purposes in this thesis):

1) "Clients": They are requesting routine examinations of horses in which the owners want to know the radiographic status of their horse and if it affects the use of the horse.
2) "Traders": They want to sell or buy a horse and need an examination of the horse in connection with the trade in order to be able to estimate the value of the horse. Typically, both the buyer and the seller want to see the examinations. As Michael puts it: *"You can hardly trade any horse these days without radiographic examination"*. This type of client is very common and therefore important to Højgaard, cf. Appendix 2.

Since it takes 1-3 years of experience with diagnosing x-rays before one has enough experience to be able to commercialize diagnostics as a service, there is a barrier of entrance to this segment. Smaller veterinary businesses will usually direct their clients to the larger hospitals where diagnosis services are offered. Having experience with diagnosis of x-rays is therefore an advantage for Højgaard.

However, in some cases clients buy x-rays elsewhere and e-mails the x-rays to a veterinarian in order to have him diagnose them. In these cases the clients are not charged anything. Højgaard estimates that they are loosing 10% of revenue in this segment, which is due to cases like these. This also suggests that the segment is cost sensitive.

Moreover, mistakes in a diagnosis can severely hurt the reputation of a veterinary business, which leads to loss of earnings as well. Some horses are worth millions of

DKK and an x-ray diagnose of very poor quality can induce huge costs on the client, e.g. if a horse is diagnosed to be healthy and it later turns out it is not (a false-negative decision). This is therefore a major risk. It also suggests that this segment is sensitive to risk.

Additionally, industry players face challenges with consistency in x-ray evaluations according to an e-mail correspondence between equine veterinarians cf. Appendix 9. As clients tend to request second opinions on x-ray evaluations from another veterinarian, the client sometimes receive a different diagnosis based on the same x-rays. This creates distrust to the evaluations of x-rays, thus damaging the reputation of the industry. Lack of consistency in x-ray diagnoses with OCD is therefore a risk factor.

The surgery segment

The surgery segment is only targeted by three equine hospitals in Denmark, including Højgaard. Clinical operations are very complex and therefore much more expensive than producing x-rays and diagnostic services. The three players offer more or less the same prices for surgeries. The surgery service is also complementary to diagnostic services and usually a client will accept a surgery if the veterinarian recommends it. Thus, having the x-rays produced and diagnosed at Højgaard's facilities increases the likelihood of generating additional revenues from clinical operations.

Subconclusion

Højgaard benefits from their experience with diagnosing x-rays, which, together with clinical operation, generates a major part of Højgaard's revenue. But what provide Højgaard with a competitive advantage in the market are economies of scale, research & development activities and knowledge differentiations (e.g. ability to do surgery). Due to these advantages, Højgaard is considered to be a leader in the x-ray examination industry for equines.

Since the diagnostic segment is a key account group to Højgaard, Michael is very concerned about Højgaard's capabilities for detecting diseases from x-rays examinations. Especially one disease is very important to look for, since it often occurs on the x-ray. This disease will be explained in the following.

2.2.4 Diagnosing OCD

Osteochondritis dissecans (OCD) is "*a common developmental disease that affects the cartilage and bone in the joints of horses. It causes clinical signs of disease in 5-25 % of all horses and can occur in all horse breeds"* (ACVS, 2017). According to Michael, OCD is "*a top priority disease to look for"*. There are three typical places on the horse where OCD are most likely to occur, including the hock, the fetlock and the stifle. However, OCD can also occur other places.

Michael points out where OCD is typically located on the hock (author's private foto)

The development of OCD typically occurs within the first year of life. If a horse does not have OCD at an age of 1 year, it will not develop it later on. Clients therefore normally request an x-ray examination of their horse after it has turned one year old and before the horse will be used for training purposes (normally before the horse has turned 2 or 3 years). This allows the veterinarian to examine the horse and make decisions in chronological order:

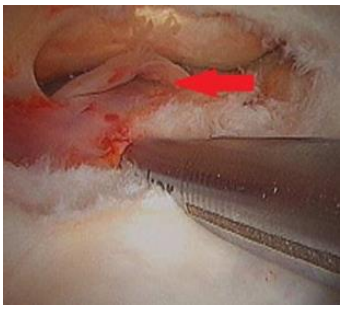
Step 1: Does the horse have OCD?

Step 2: Does OCD affect the use of the horse (for clients) and/or will it affect the possibility of selling the horse at a later time (for traders)?

Step 3: Is a clinical operation needed?

If a horse has OCD and is supposed to be included in a trade, or if OCD affects the use of the horse, the owner usually requests a clinical operation, since it is very difficult to sell a horse with OCD. However, sometimes the barrier to affecting the "use" of the horse is very high, when the horse is not supposed to be used for anything in particular. In this case no clinical operation is recommended.

If the client accepts a clinical operation to remove OCD, Højgaard's surgical department will be in charge of conducting the clinical operation, including orthopedic surgeries by use of arthroscopy. Arthroscopy is s a *"minimally invasive surgical procedure on a joint in which an examination and sometimes treatment of damage is performed using an arthroscope - an endoscope that is inserted into the joint through a small incision"*. (Järvinen, Guyatt, & Gordon, 2016). The image to the right (from Højgaard Hospital) displays Arthroscopy in action.



In order to understand how OCD diseases are identified on the x-ray, the following will describe how x-ray examinations are conducted.

2.2.5 X-ray examination and medical data

X-rays of horses are produced at Højgaard Equine Hospital whenever it is necessary for detecting certain diseases, including OCD. The x-rays are the most essential medical data that are used to diagnose a horse. Usually the client pays for a package of 14 x-rays that covers different body parts of the horse.

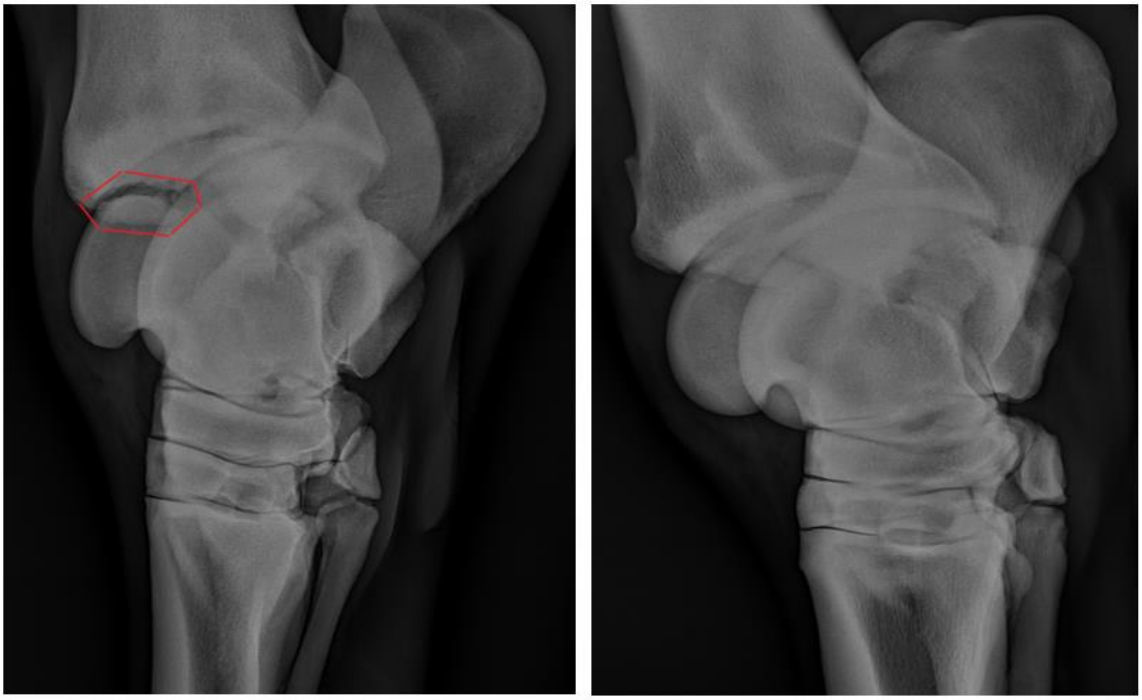
According to Michael, Højgaard has above 15.000 x-rays located on two different virtual databases that can only be accessed through office computers. He is not sure about how many of them that has OCD, but he estimates it to be around 5.000 x-rays.

He also explains that the journals containing information about the diagnoses are physically located in archives at the hospital and therefore separated from the virtual x-rays.

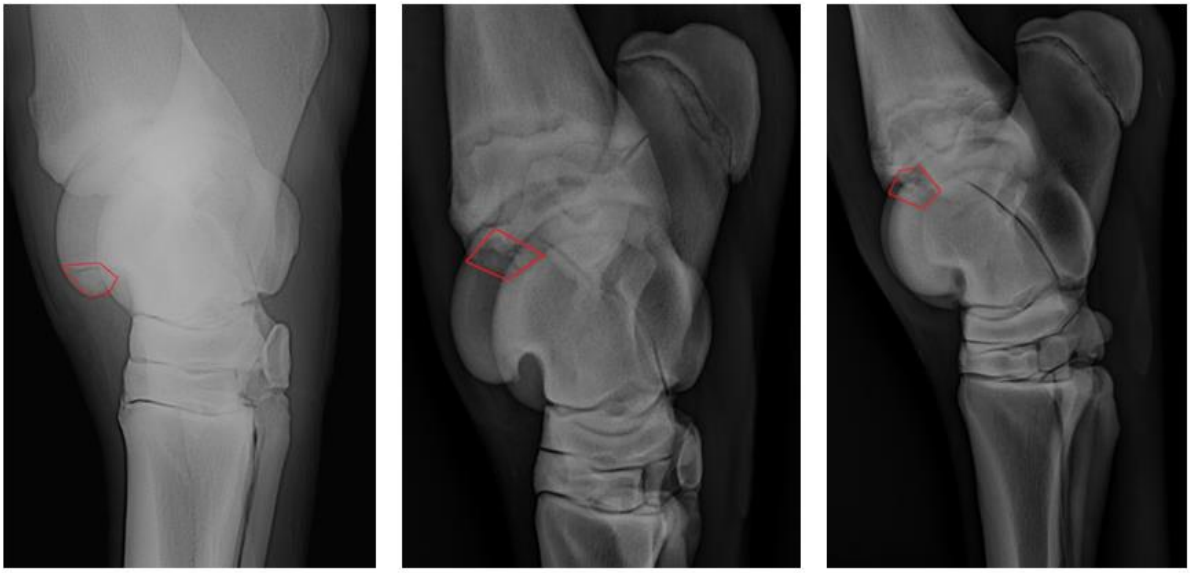
Since the hock is an essential body part of the horse where OCD is typically detected, Appendix 8, describes the three x-ray projections on this part of the body. For now, it is important to notice that three different x-ray projections on the hock (LM, DMPLO and DP) are the most relevant x-rays to look at, when trying to detect if there exists OCD disease in the hock.

The following images display x-rays of the hock containing OCD diseases (red marking shows where OCD is located on the x-ray). In general, Michael explains that OCD follows a certain pattern that he is looking for when examining the x-rays:

"*OCD is identified by a fragment (white texture) that has been separated from its place on the bone*. *An x-ray with OCD therefore both needs to show a fragment and a "hole" in the bone from where the fragment has been separated".*



X-rays DMPLO: OCD Crista Intermedia (left) and no OCD (right) (The x-ray's are owned by Højgaard Equine Hospital)



X-ray DMPLO: OCD Lateral Trochlea

X-ray DMPLO: OCD Crista Intermedia

X-ray LM: OCD Crista Intermedia

The x-ray's are owned by Højgaard Equine Hospital

The first x-ray is a very clear OCD detection that shows how a fragment (white texture) is clearly separated from the bone. The other x-rays with OCD also displays fragments separated from places on bones, but they are more difficult to see for the "untrained eye". They also show that OCD can be located different places on the x-

ray, in this case "Crista Intermedia" and "Lateral Trochlea". Also, there might be several smaller OCD fragments on the same x-ray.

Michael estimates that a veterinarian needs at least 2 years of training with examining x-rays, before he is capable of detecting OCDs.

2.2.6 Validating the case company

Based on the above case study it is estimated that Højgaard comply with the formulated conditions for selection of case company. This is because detection of OCD on x-rays is a process that could be automated by use of image recognition, since x-rays are basically images and OCD disease is a white texture that follows a pattern explained above.

As will be described in the following, they have also accepted a very flexible collaboration.

2.2.7 Challenges & next steps

As described above, detecting OCD is an essential part of Højgaard's x-ray examinations. An improvement of this process will therefore optimize a major revenue driver and therefore add value to the business. It is estimated that an intelligent automation of this process can provide three types of benefits:

1) Lower the cost of producing diagnoses and training the staff due to automation of the detection process. Also, this will allow staff to focus on more value-adding activities (labor cost savings).
2) Mitigate risks of false negative diagnoses (increased accuracy/recall).
3) Mitigate the risks of inconsistent diagnoses in the industry (increased consistency).

Since Højgaard is generally focusing on innovating their business, they are very interested in exploring new ways of solving problems. They have therefore agreed to collaborate with the author and his thesis partner, and they have allowed full access to their databases and journal archives, as well as full guidance by the CEO of the company through the whole research period. The purpose with the collaboration is to develop an innovative solution to Højgaard's challenges that can be communicated to the CEO.

3 Theoretical underpinnings

This section provides an overview of the literature and research in academic fields that are relevant to this thesis. It accounts for academic literature within three research areas: AI concepts, business application concepts and data mining concepts. This is because the focus area of the thesis comprises all three areas.

The following section will break down the concept of AI to understand how deep learning, machine learning and convolutional neural networks relate to each other. The business application of these concepts will be explained in the Computer Vision section. Finally, the process of modelling an AI application will be explained in the data mining section.

3.1 Artificial Intelligence

*Artificial Intelligence (*AI) is a concept that refer to computers/machines that are intelligent, i.e. they can solve problems that previously only humans were able to solve. Historically, AI has been very successful at solving tasks that are intellectually challenging for humans, but can be explained with a set of formal mathematical rules. Some of the most challenging problems to solve with AI include the ones that are intuitive to humans (easy to solve), and feel automatic, but are very hard to explain formally, including what knowledge is required to solve these problems. These include intuitive tasks like recognizing spoken words or detecting a disease on an x-ray. Computers need to capture this knowledge in order to solve the issues in an intelligent way. (Goodfellow, Bengio, & Courville, 2016)

3.2 Deep Learning

Hard-coding this knowledge about the world into the computers by use of logical-inference rules has not been successful (Goodfellow, Bengio, & Courville, 2016, 2-6). This is called the knowledge-base approach. Instead, the solution to the intuitive problems is:

"*to allow computers to learn from experience and understand the world in terms of a hierarchy of concepts, with each concept defined in terms of its relation to simpler concepts*".

This approach to AI is called *Deep Learning* (DL), which is a subfield of representation learning, explained below. DL explains how concepts are built on top of each other, forming a hierarchy of layers of simple concepts. Deep learning has two intrinsic benefits:

- *Learning from experience* avoids the need for human operators to formally specify all of the knowledge that the computer needs
- Hiearchy of *concepts* allows the computer to learn complicated concepts by building them out of simpler ones

(Goodfellow, Bengio, & Courville, 2016)

As machines cannot rely on hard-coded knowledge, they need the ability to acquire their own knowledge, by extracting patterns from raw data. This capability is called *machine learning (M-L)*.M-L includes, among others, algorithms like linear regression, logistic regression, naïve Bayes and SVMs, but also deep learning algorithms like NLP and CNNs.

3.3 Representation learning

The performance of M-L algorithms depends heavily on the representation of data, including choice of features that represent the data (a feature could be the shape, texture or line that is present on an image). The M-L algorithm then learns how different features correlates with different outcomes. However, self-learned representation often yield much better performance, which is why *representation learning*, a sub-field within M-L, is very useful. This allows the AI system to adapt to new tasks with minimal human intervention. (Goodfellow, Bengio, & Courville, 2016)

When designing algorithms to learn features, the goal is usually to separate the *factors of variation* that explain the observed data. These factors can be thought of as abstractions that help o make sense of the rich variability in the data. E.g. When analysing an x-ray image displaying a disease, the factors of variation include the position of the disease, its colour, and the angle and brightness of the light (Goodfellow, Bengio, & Courville, 2016). However, a major source of difficulty for AI applications is that many of the factors of variation influence every single piece of data (e.g. every pixel) that can be observed. It is therefore useful to disentangle the factors of variation and discard the ones that are not important to the application.

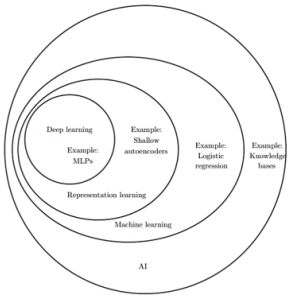


Figure: Explains how the concepts relate to each other (source: (Goodfellow, Bengio, & Courville, 2016, p. 9)

This central challenge in representational learning of extracting high-level abstractions (learning features) can be solved with deep learning by introducing representations that are expressed by other simpler representations. E.g. DL can represent the concept of an image by introducing simpler concepts as corners and contours (features), which then can be defined in terms of edges. (Goodfellow, Bengio, & Courville, 2016)

3.4 Machine learning

As mentioned earlier, AI applications need capability to tackle the hard issues in the real world. This capability comes from the M-L algorithm, which is a learning algorithme that can "learn from data". Mitchell (1997) defines M-L as:

"A computer program that is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured provides the definition by P, improves with experience E."

(Goodfellow, Bengio, & Courville, 2016, p. 99)

The following will adress the different components in the M-L definition.

3.4.1 The task T

This thesis is specifically looking into the task called *classification*. An example of a classification task is object recognition (or image recognition), where the input is an

image (usually described as a set of pixel brightness values), and the output is a class or a probability of classes that identifies the object in the image. Modern object recognition is best accomplished with deep learning according to Krizhevsky et al. (2012) and Ioffe and Szegedy (2015). (Goodfellow, Bengio, & Courville, 2016, p. 100).

An M-L task includes building a model that performs well on new, previously *unseen* inputs. This task is called *generalization*.

The training of a M-L model includes using a training set, in which one can compute some error measure on the training set called the *training error*, and then focus on reducing this training error. However, what seperates M-L from optimization problems, is that M-L does not only focus on minimizing the training error, but also a low *generalization error*/*test error.* The generalization error is the error we get from testing the model on the test set. This set is collected separately from the training set. (Goodfellow, Bengio, & Courville, 2016, p. 110)

Collection of relevant data for training and testing is based on *the data generation process*, which includes two assumptions on the probability distribution of the two sets:

- The data in each dataset are *independent* from each other
- The training set and test set are *identically distributed*, drawn from the same probability distribution as each other.

(Goodfellow, Bengio, & Courville, 2016, p. 111)

3.4.2 The performance measure, P

For classification tasks, the performance of the model is usually measured in terms of *accuracy*. Accuracy is the proportion of examples for which the model produces the correct output. Equivalent information can be obtained by measuring *the error rate*, which is the proportion of examples for which the model produces an incorrect output (1-accuracy). (Goodfellow, Bengio, & Courville, 2016)

Another performance measure is *Precision.* This includes the proportion of all positive predictions that are correct. It is a measure of how many positive predictions were actual positive observations. (Albon, 2017)

Recall (also known as *sensitivity* or *True positive* rate) measures the proportion of all real positive observations that are correct. It is a measure of how many actual positive observations were predicted correctly. (Albon, 2017)

The F1 score is an 'average' of both precision and recall (also called the “harmonic mean”). It is used to average ratios and calculate a single score (Albon, 2017):

$$F1 = 2\frac{Precision * Recall}{Precision + Recall}$$

In general, the performance of a M-L algorithme is based on its ability to:

- Make the training error small (reduce *underfitting*)
- Make the gap between training and test error small (reduce *overfitting*)

One can control whether a model is more likely to overfit or underfit, by varying the models *capacity*, which is defined as the model's ability to fit a wide variety of functions. Models with low capacity may struggle to fit the training set and models with high capacity can overfit because it memorizes properties of the training set that does not serve well on the test set (does not generalize well). (Goodfellow, Bengio, & Courville, 2016)

3.4.3 The experience, E

M-L algorithms can generally be categorized as either *unsupervised* or *supervised* depending on what kind of experience they are allowed to have during the learning process. (Goodfellow, Bengio, & Courville, 2016)

This thesis will focus on supervised learning. This type of learning algorithm uses dataset that contains features, but each data example is also associated with a label or target, e.g. an x-ray dataset can be annotated with different diseases. A supervised learning algorithm can then study the x-ray dataset and learn to classify x-rays with different diseases. (Goodfellow, Bengio, & Courville, 2016)

On a more technical level, the M-L algorithms requires a high amount of numerical computation, which involve solving different mathematical functions by use of iterative methods. An important operation in these functions include *optimization*, where the general purpose is to find a value of an argument x that minimizes or maximizes the function. (Goodfellow, Bengio, & Courville, 2016)

Most DL algorithms are based on an optimization algorithm called *stochastic gradient descent (SGD).* SGD is an extension of the gradient algorihtm, where the specific purpose is to minimize or maximize the objective function. When minimizing the function, the function is a *cost function, loss function or error function*. The goal is to reduce the cost function to a global minimum, i.e. the lowest possible error value. Reducing the cost function will therefore optimize the learning and result in better performance. (Goodfellow, Bengio, & Courville, 2016, p. 98).

One way to reduce the generalization error, but not the training error, is by use of *regularization techniques*. This includes modifying the learning algorithms by adding a penalty called a regularizer to the cost function, e.gmany DL algorithms apply *weight decay* as the regularizer (Goodfellow, Bengio, & Courville, 2016, p. 120).

Another way to influence the behaviour of the algorithme, is by adjusting (or tuning) the parameters, which are often chosen manually. Usually a validation data set is used to test the models parameters and based on the performance results, the parameters will be adjusted accordingly. The validation set is always taken from the same distribution as the training set. Specifically, the training data is split into two disjoint subsets. The training set is used to learn the parameters and the validation set is used to estimate the generalization error during or after training. (Goodfellow, Bengio, & Courville, 2016)

3.5 Neural networks & deep learning

Now that we know the general capability behind AI, we can dig further into the specifics of DL algorithms. Deep learning algorithms were previously (dating back to the 1950s) named *artificial neural networks* or just neural networks, since they were intended to reflect computational models of the biological brain. In a brain, neurons communicate with each other in a network using synapses, which are electrical-chemical signals. This corresponds to the layers in the neural network which consists of units (or neurons) that act in parallel. It is important to note that they do not reflect real models of biological function and most researchers today do not make use of DL in order to simulate the processes in a brain. (Goodfellow, Bengio, & Courville, 2016, pp. 13-19).

The multilayer perceptron (MLP), which is also named *feedforward neural network,* is one type of DL algorithm that consists of a mathematical function that maps input values to output values. It is one function composed of many simpler functions where each function provides a new representation of its input, which is fed forward through the layers of functions to the final output layer (Goodfellow, Bengio, & Courville, 2016, pp. 13-19).

As mentioned in the limitation section above, this thesis will not conduct parameter tuning and neither will it design a model from scratch, but instead apply a pretrained model. It is therefore not in the scope of this work to go into depth with the specific algorithm design However, the following section will include some brief details on what is the most common design for the models used in this thesis.

3.5.1 Important components of a feedforward neural network

F*eedforward neural networks* includes hidden layers and one has to choose an activation function that is able to compute the hidden layer values. It is usually recommended to use the the rectified linear unit (or ReLU) which is a non-linear function (Jarrett et al., 2009; Nair and Hinton, 2010; Glorot et al., 2011a) (Goodfellow, Bengio, & Courville, 2016, pp. 170-173).

The output function for a classification task in a neural network is usually the softmax function. The softmax function is used to represent a probability distribution for a number of different classes. (Goodfellow, Bengio, & Courville, 2016, pp. 183).

The algorithme used for computing the gradient in a neural network (see further below) is typically based on the backpropagation algorithme. Backprop computes the gradient by allowing information from the cost function to flow backwards through the network. (Goodfellow, Bengio, & Courville, 2016, p. 203).

Designing the architectureof a network involves considerations on the *depth* (number of layers) *Depth* allows the algorithme to learn through multiple steps. Each layer of a representation (see further up) can be thought of as the state of the model's memory after executing another set of instructions in parallel (this is also explained further below). This helps the model organize its processing. There is no single correct answer to what is the most appropriate depth (Goodfellow, Bengio, & Courville, 2016, p. 201).

3.6 Convolutional Neural Networks

Convolutional neural networks (CNNs) are a type of feedforward neural network that:

"uses convolution in at least one of their layers."

(Goodfellow, Bengio, & Courville, 2016, p. 330).

Convolution indicates that the network employs a mathematical operation called convolution. The following paragraphs will describe how CNNs can be applied to find patterns in images.

An image is basically a matrix of pixel values. If it is a two dimensional greyscale image (like most x-rays) the value of each pixel in the matrix could range from 0 to 255, where zero is indicating black and 255 indicating white. The primary purpose of CNNs is to extract features from the input image. Convolution preserves the spatial relationship between pixels by learning image features using small squares of input data. This will be explained graphically by use of the following figure 1:

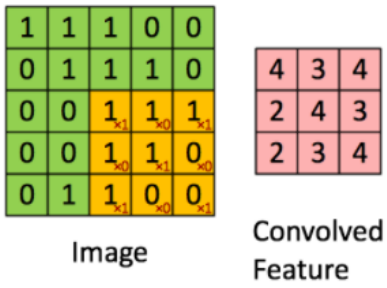


Figure 1: The convolution operation (The_data_science_blog, 2017)

The figure above shows an orange matrix that is "slided" over the original image (green colour). Based on the covered pixel values in the green matrix, the orange matrix calculates one element. After having calculated the element the orange

matrix moves one pixel and calculates a new element. For every position, one element is computed and added to the final output matrix (The pink matrix). In other words the pink matrix "sees" parts of the image every time the orange matrix changes position. (The_data_science_blog, 2017)

In CNN terminology, the orange matrix is called a filter (or "kernel"). Its purpose is to detect features. The product of the convolution (the pink matrix) is called the "convolved feature" or "feature map". A CNN can have many kernels in each layer to filter out new features, e.g. edges in one layer and shapes in another. By changing the parameter valuesn the kernel different features can be detected. (The_data_science_blog, 2017)

During training, CNNs learn the parameter values of the kernels filters by its own. However, still the parameters need to be specified, including number of filters, filter size, architecture of the network etc. before the training is started The more filters, the more image features get extracted and the better the network will recognize patterns on unseen images. The following will explain the most important steps of how CNNs learn to recognize a pattern:

Step 1:

The size of the feature map is based on different parameters including the number of filters used for the convolution operation (same as *Depth* explained above) and the number of pixels in the input matrix that the filter matrix will cover (this is called *Stride*) (The_data_science_blog, 2017).

Step 2:

After the convolution operation, the ReLu activation function is used. Its purpose is to add non-linearity to the network, since convolution is only a linear operation. This is because most real-world data is non-linear (The_data_science_blog, 2017).

Step 3:

Spatial pooling will reduce the dimensionality of each feature map, but retain the most important information. Spatial pooling can be of different types, including Max Norm, Average and Sum. Max Norm pooling includes defining a spatial neighbourhood (e.g. 2*2 window) and take the largest element from the rectified feature map within that window. (The_data_science_blog, 2017)

Ultimately, the function of pooling will progressively reduce the spatial size of the input representation, including:

- Make the feature dimensions smaller
- Reduce the number of parameters and computations, resulting in better control of overfitting
- Make the network invariant to small transformations or distortions in the input image

- Reach an equivariant representation of the image. This is very useful as it allows the network to detect objects no matter where it is located in an image.

(The_data_science_blog, 2017)

In the analysis section the architecture of a CNN model will be described in order to show how the concepts and functions described above relate to each other.

3.7 Business application area: Computer vision and Computer-aided detection

As the case problem described above includes recognizing a disease on images by use of computer technology, this thesis will look into the specific application field of computer vision.

Computer Vision (CV) is the science of looking at visual media from a computer's point of view. In this research, the application area is restricted to cover 2D images. One area of CV is Image learning which refer to M-L applied to images. One type of CV task within Image Learning is *Image Recognition*, of which there are at least three variations (Goodfellow, Bengio, & Courville, 2016):

- Object recognition: Recognizing objects on an image usually together with the objects 2D position.
- Identification: Recognizing an individual instance of an object, e.g. a specific person's face.
- Detection: Recognizing a specific condition, e.g. detection of abnormal cells on a medical image.

This thesis will look into recognition of a specific condition on x-rays including detection of the OCD fragment in the hock. The specific application field is therefore "Detection", which is also called "Computer-aided Detection" ("CAD") (Forsyt & Ponce, 2003). This application scenario excludes detection of how many conditions are present on the image, e.g. if there are more than one OCD fragment present on the x-ray.

Based on the author's literature review on CV and CAD research, only the most recent literature has been chosen, as the field is rapidly evolving and new findings emerge on a monthly basis. One case study on a clinical pharmaceutical company, Novartis, emphasized the need to start the design process with identifying the important business results first and then continue with the image acquisition process afterwards using best practices. (Cast Study - Image Data Management System - Oracle - Novartis, 2008)**.** This case specifically looked into large scale implementation of x-rays to speed up drug testing. This thesis will therefore design a model in accordance with this procedure, starting with the business understanding of the case problem.

Fenn et al. (2015) also explains why it is important to exclude irrelevant features on an image as early as possible and already during the image acquisition phase

(Fenn, Mendes, & Budden, 2015). Russakovsky et al. (2014) agrees on that, based on a study of different CV algorithms and their applications. He also emphasized the importance of labelling the image correctly as part of the preparation phase. (Russakovsky, 2014)

In order to detect OCD diseases it is necessary to exclude bone areas that are not relevant for OCD detection as OCD would never appear in that area (based on previous experience). This is where cropping is a useful technique, which will be explained further below.

Constantiou and Kallinikos (2015) points out that heterogeneous data such as images, should be tagged with homogenous additional data (also known as master data) as much as possible in order to include all descriptive data, e.g. time of x-ray production, the circumstances and any other related data links. This should be acquired as early as possible during the image acquisition phase. (Constantiou & Kallinikos, 2015).

This research has included relevant metadata and documented it in a spreadsheet, cf. Appendix 4. As mentioned below this has helped the researcher to better evaluate and discover x-rays that are relevant or not to this research.

Rosenbeck (2015) explains the importance of visualizing the results to the business in order to help them better understand and capture the value from CV. Also, is important to include the business by use of agile feedback loops during the process (Rosenbeck, 2015). This research produces visualizations results that gives a foundation for better business decision-making. In addition, during the experimentation process, the authors have consulted the business on several occasions to better understand the data and how a model can provide any value to it.

Russakovsky et. al. (2014) points out that major region of deviations in an image can be identified by use of trial batches in order to discover what outliers look like and re-label them accordingly. This process is best carried out by a convolutional neural network that discovers the relevant pattern and labelling all by itself (explained above) (Russakovsky, 2014). Other relevant image recognition techniques (or "image descriptors") include scale-invariant feature transform (SIFT) and histogram of oriented gradients (HOG), which have been widely used for object detection and segmentation in medical image analysis (Shin, et al., 2016).

Based on the above process insights and best practices, the application of CNNs for CAD appears relevant to solve the research question.

3.7.1 Pre-trained modelling

According to Shin et al. (2016), there exist currently three major techniques to employ CNNs for medical image classification:

- Training the CNN from scratch end-to-end
- Fine-tuning the CNN (supervised) by using off-the-shelf pre-trained CNN features from other tasks
- Conducting unsupervised CNN training with supervised fine-tuning.

Recent research (2016) suggests that pretrained CNNs can be effectively utilized on x-rays (Shin, et al., 2016) (Spampinato, Palazzo, Giordano, Aldinucci, & Leonardi, 2016) (Razavian, Azizpour, Sullivan, & Carlsson, 2014) (Ribeiro, Uhl, Wimmer, & Häfner, 2016). The specific cases included x-rays of the human chest, lymph node and lungs. One of them used a pretrained model from GoogleLeNet and produced an accuracy result of 70%, which are good considering the limited amount of data (55 chest X-ray scans - 25 positives and 30 negatives). This suggest that fine-tuning a pretrained CNN requires less data and could outperform (or performs as well) as network learnt from scratch. This is helpful if one has only a limited amount of data. However, a pretrained model cannot change network architecture, which limits the options for parameter tuning (Github - CS231n, 2017).

The fine-tuning of the CNN (training) only takes place in the final classifier layer, where the pretrained model will learn to detect high-level features on x-rays relevant to the case study. The approach is especially useful, if the dataset on which the network was trained, is similar to the target dataset, as the patterns learned by the convolution kernels are likely to be equally discriminative. (Spampinato, Palazzo, Giordano, Aldinucci, & Leonardi, 2016)

This research chose to focus on how to fine-tune an existing pre-trained CNN and benchmark it with a CNN model that has been trained end-to-end. This choice was made, since the researchers expected that it would be hard to gather enough data. In addition, Jacob Axelsen recommends experimenting with a pretrained model initially as a lot of time can be saved compared to building up a model from scratch, cf. Appendix 1.

IBMs Visual Recognition service is a pre-trained M-L model based on the Watson Developer Cloud (IBM1, 2017). Specifically, it applies a convolutional neural network that has been pre-trained from millions of labeled images. On top of it has a custom classifier that can be trained with new input images. (IBM2, 2016) (IBM3, 2915). The service will then output a probability score for how certain it as at a custom class. Moreover, IBM Watson has been cited as being in the forefront of medical AI. In 2015, IBM purchased Merge Healthcare for $1 billion, partly to get an established foothold in the medical IT market. This purchase gave Watson millions of radiology studies to help train the AI in evaluating patient data (ITN, 2017). This suggest that Watson is a strong platform for building models based on medical data.

The following section will continue to explain the process approach behind the model experimentations.

3.8 Data mining process

The baseline process for producing the research is based on the CRISP-DM model (See figure 2 below). It is an iterative process that explains how to go about integrating data mining in a business environment (Shearer, 2002). However, the actual process changed over the course of producing the results and the thesis itself. This is due to the explorative nature of the research in which ideas emerge during the process, requiring the data mining process to sometime change direction. This research therefore assume some flexibility to the CRISP-DM approach.

Conceptually, data mining is an activity where a team seeks a goal on a particular dataset in a specific situation, rather than letting the data talk and see what comes out (which is M-L) (Provost & Fawcett, 2013, p. 27). However, these are interchangeable definitions and this thesis will not dvelve into differences as it adds nothing to the product of the thesis. The CRISP model is therefore used for conducting the data analysis as a starting point.

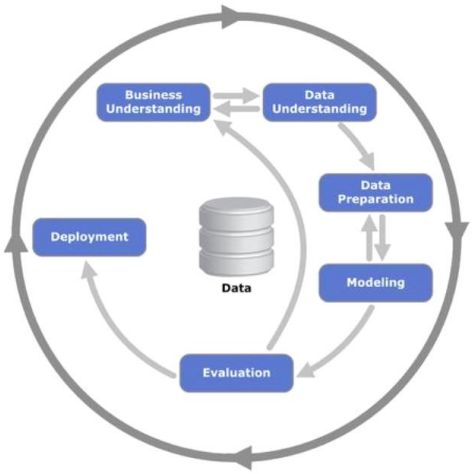


Figure 2: CRISP-DM model (Provost & Fawcett, 2013, p. 27)

The model describes how an initial *understanding of the business situation* is required as a starting point. This include considerations on the use case scenario and how the model is relevant to the business. This is followed up by an *understanding of data*, which is the raw material used for solving the business problem. It includes an understanding of the strengths and limitations of data. Next the *data is prepared*, which often include data transformations like cropping, scaling etc. This allows the data to be compared to each other and more suitable for modelling techniques. (Provost & Fawcett, 2013, pp. 27-30)

During the *modelling phase*, algorithms and data science methods will be used to predict patterns in the data, i.e. create a model. The research will account for the model experimentations done with the pre-trained model, as well as the design choices for the model developed in Tensorflow. However, the primary focus is put on the pre-trained model. (Provost & Fawcett, 2013, p. 31)

Afterwards, the model will be *evaluated* in order to find out if it is valid and reliable before moving on. This include considerations on whether there exist any pattern and if the results are feasible to the business. It usually include both quantitative and qualitative assessments in order to ensure full comprehensibility of the model to stakeholders. This phase is the focus area of this research and it will include the following assessments (Provost & Fawcett, 2013, p. 32):

- Accuracy and related technical measures
- Expected value framework to calculate the value add
- Profit curves
- ROC graphs and ROC curves

Finally, depending on the evaluation, the model is being *deployed* in some information system or business process. Deploying a model into production typically requires compatibility with an existing system (Provost & Fawcett, 2013, p. 33). This may incur substantial expenses and investments, which is why *risk* & *maturity* considerations ought to be taken into account.

The four assessments mentioned above will be explained in the following.

3.8.1.1 Accuracy and related technical measures

As mentioned in the M-L section, accuracy is a very common performance measure of M-L models. Additionally, accuracy will be complemented with other performance measures like the error rate, precision, recall and the F1 score, in order to provide a more comprehensive view of the model's technical performance.

3.8.1.2 Expected value framework

As the performance also needs to be conveyed to business stakeholders that do not necessarily understand data science performance measurements (or how these measurements influences the business), other frameworks need to be provided. One useful framework is the expected value framework. It decomposes data-analytic thinking into (Provost & Fawcett, 2013, p. 194) the structure of the problem (the formular), the elements of the analysis that can be extracted from the data (confusion matrix and expected rates) and the elements of the analysis that need to be acquired from other sources (cost/benefit matrix).

This research will use the expected value framework to evaluate classifiers. Calculating the expected value provides a suitable numeric measure for benchmarking the models, since the framework aggregates all possible cases. The expected value is calculated using three information sources (Provost & Fawcett, 2013, p. 197):

- Confusion matrix
- Expected rates
- Cost/benefit matrix

A classifier produces a *confusion matrix* which will allow the research to identify the number of x-rays that are predicted to be:

- True positives (The x-ray is predicted to show OCD and it is true)
- False positives (The x-ray is predicted to show OCD and it is not true)
- True negatives (The x-ray is predicted not to show OCD and it is true)
- False negatives (The x-ray is predicted not to show OCD and it is not true)

(Provost & Fawcett, 2013, p. 197)

The *expected rates* are calculated from the confusion matrix and they are used as estimates of probabilities of whether an x-ray is a true positive ("tp rate"), false positive ("fp rate"), true negative ("tn rate") and false negative ("fn rate"). E.g. tp and fp rates are calculated using the total distribution of positives divided by the number in the confusion matrix and tn and fp rates are calculated using the total distribution of negatives divided by the number in the confusion matrix. (Provost & Fawcett, 2013, p. 198)

The *cost/benefit matrix* specifies, for each (predicted, actual) pair in the confusion matrix, displays what is the cost or benefit of making such a decision. This matrix generally depend on external information provided via analysis of the consequences of decisions in the context of the business problem. Usually, they cannot be specified exactly, but only as approximate ranges. (Provost & Fawcett, 2013, p. 199)

The expected value/profit in this research will factor out the probabilities of seeing each class by including the *class priors*. The class priors, *p(p)* and *p(n)*, specify the likelihood of seeing positive and negative instances, respectively. This will allow the research to separate the influence of class imbalance from the predictive power of the model. (Provost & Fawcett, 2013, p. 202).

The overall calculation of the expected profit can be expressed with the following formular:

$$\begin{aligned}\text{Expected profit} = \quad & p(\mathbf{p}) \cdot [p(\mathbf{Y} \mid \mathbf{p}) \cdot b(\mathbf{Y}, \mathbf{p}) + p(\mathbf{N} \mid \mathbf{p}) \cdot c(\mathbf{N}, \mathbf{p})] + \\ & p(\mathbf{n}) \cdot [p(\mathbf{N} \mid \mathbf{n}) \cdot b(\mathbf{N}, \mathbf{n}) + p(\mathbf{Y} \mid \mathbf{n}) \cdot c(\mathbf{Y}, \mathbf{n})]\end{aligned}$$

Each mathematical expression in the formular refer to the information sources above and will be put into action in the analysis section. This formular is also useful for comparing models that are tested on different distributions of data as it is just a matter of replacing the priors. (Provost & Fawcett, 2013, p. 201).

3.8.1.3 Profit curves

A more intuitive way of expressing a models performance is by using visualizations instead of just calculating a single number from a formular. While the expected profit formular can be used to compute a decision for each classifier based on the expected

value, another strategy is to make decisions based on ranking a set of classifiers by their scores, and then take actions on the ones ranked in the top of the list. (Provost & Fawcett, 2013, p. 209)

This strategy is useful in situations where the model gives a classifier score that ranks images by their likelihood of belonging to a class, but which is not a true probability. It is also useful when costs and benefits cannot be specified precisely, but that it should still be possible to take actions (e.g. on the highest likelihood scores). As will be explained later, both of these practical issues are apparent in this research. (Provost & Fawcett, 2013, p. 210)

The images can be ranked by putting a threshold on the classifier score. A high threshold correspond to a conservative strategy, in which high certainty is necessary for taking any action. Conversely, a low threshold correspond to a permissive strategy. Appendix 3 displays a classifier with a threshold, and how confusion matrices changes when the threshold changes. (Provost & Fawcett, 2013, p. 211).

Now since different thresholds produces different classifiers and therefore different expected profits, the purpose is to choose the threshold that produces the highest expected profit. Putting these values onto a graph will display the profit curve. This curve will show the expected cumulative profit for that classifier as progressively larger proportions of images are included (by varying the threshold). (Provost & Fawcett, 2013, p. 213). The proportion of images on the x-axis is also called "coverage". It is the "*fraction of examples for which the machine learning system is allowed to produce a response"* (Goodfellow, Bengio, & Courville, 2016, p. 426).

3.8.1.4 ROC graphs and ROC curves

Profit curves relies on the conditions that both class priors and cost-benefit estimates are known and are expected to be stable. However, if these conditions are unstable it is better to use another visualization technique that can accommodate these uncertainties. The *Receiver Operating Characteristics (ROC) graph* is a two-dimensional plot of a discrete classifier with fp rate on the x-axis against tp rate on the y-axis. It depicts relative trade-offs that a discrete classifier makes between benefits (true positives) and costs (false positives). A discrete classifier is a classifier that only outputs a label (e.g. OCD or Not OCD) as opposed to a ranking model, which also outputs a score. A *ROC curve* depicts multiple points (a curve) of a ranking model, which corresponds to the different thresholds as mentioned above. This allows one to limit the model to only take decisions on the ones that are above a certain threshold. (Provost & Fawcett, 2013, p. 215-218)

3.9 Sub conclusion

Based on the above theoretical concepts and approaches this thesis will chose to follow the CRISP-DM process approach starting out with a description of the use case scenario for a DL model. This will be followed up by the data understanding

and data preparation phase, which will account for different data transformation techniques.

IBM Watson's pre-trained Visual Recognition Service will be used for developing the model. This allow the research to build a model on a CNN without spending time and ressources on the end-to-end development, and instead focus on the business aspects of the research. The explanations of the DL, M-L and CNN concepts will help understand how the IBM service works, but as the inner functions are proprietary knowledge, it is not possible to say exactly what algorithms the service has been built upon. The concepts will also help explain the main design choices behind the model build in Tensorflow, but it is out of scope to go into technical details of that model for this thesis.

The evaluation phase will include evaluations of both models (Watson and Tensorflow) as well as a benchmark between them. The evaluation will be based on classic performance measurements (e.g. Accuracy and Recall) as well as the expected value framework, profit curves, ROC graphs and ROC curves.

Finally, the deployment phase will include considerations of risk & maturity, and how to effectively deploy the model(s) into the existing business landscape.

4 Analysis

This section will present the analysis, which is structured in accordance with the CRISP-DM and applying the concepts described in the theoretical section.

4.1 Business understanding

In order to understand the use scenario for Højgaard the following section will elaborate on the x-ray examinations. The following figure 3 displays a high-level end-to-end business process architecture for this type of consultation including how the three segments connects to these processes. This allows identification of which process is suitable for intelligent automation.

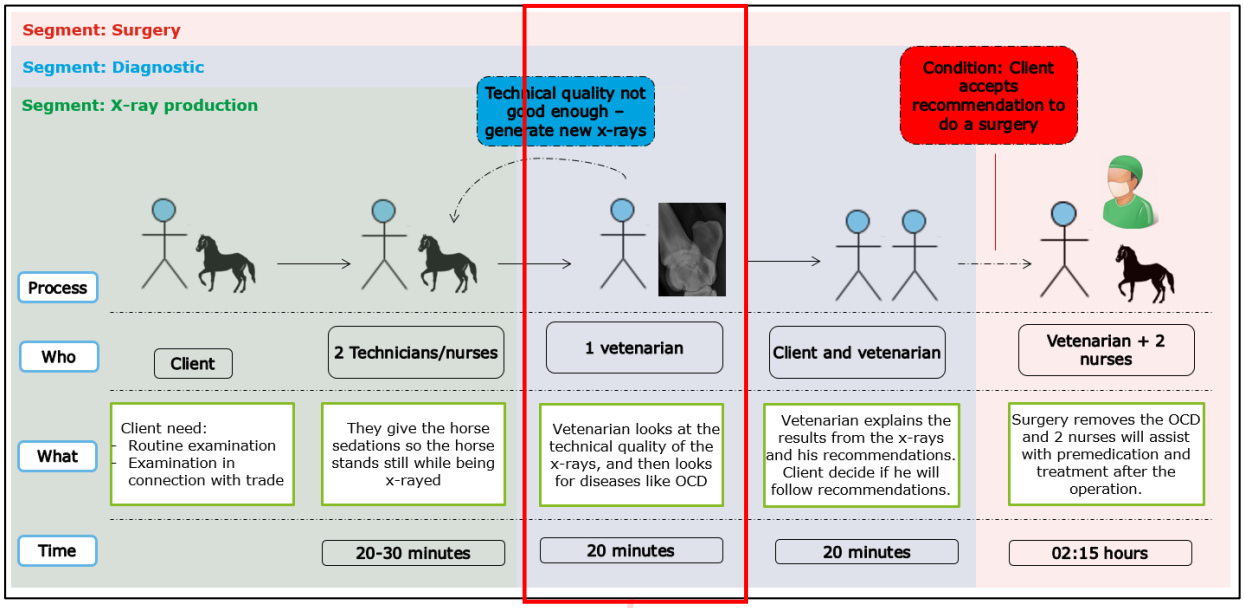
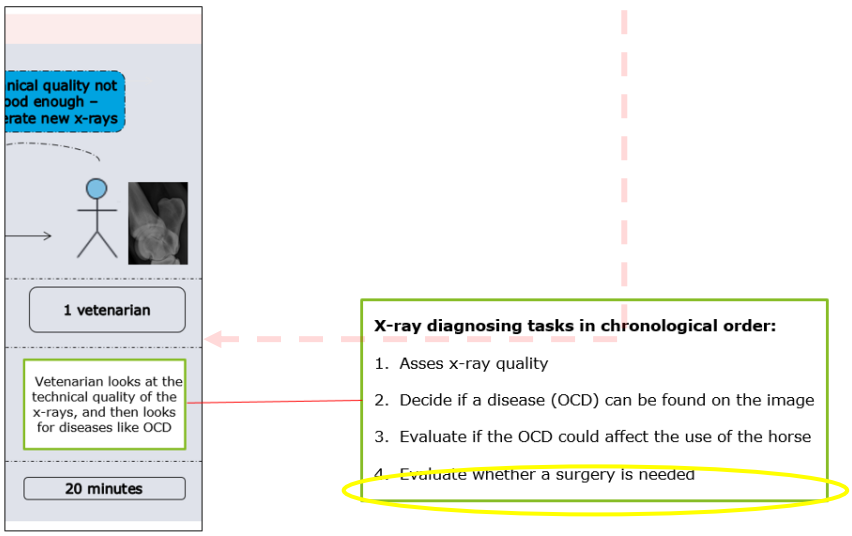


Figure 3: Shows the business architecture of x-ray examinations. At Højgaard, they do not distinguish between the "x-ray production" and the "diagnostic" segments. (Source: Author's own making)

Since this thesis is looking into the application of artificial intelligent technology including machine vision, the focus is placed on processes that involve images. It is



important to note that a complete diagnose is not only based on an x-ray examination, but also takes into account the horse's behavior and past clinical operations.

Figure 4: The diagram elaborates on the concrete tasks a veterinarian has to go through in order to perform an x-ray examination.

However, the examination of the x-ray image is the most influential activity in connection with discovering a disease, cf. Appendix 2. The above figure 4 elaborates on the specific tasks carried out in connection with the examination of the x-rays.

Among the tasks mentioned above, task 2: *"Decide if a disease can be found on the image"*, includes looking at the image, discover a white texture separated from the bone, make an evaluation and then take a decision based on that evaluation. As this task involves recognizing a specific condition on an image, the task can be classified as a CAD task. This task is therefore suitable for intelligent automation by use of image recognition. The research will therefore focus on how to automate this specific task/process by use of image recognition.

However, as will be further elaborated in the deployment section, full automation is not necessarily the best use scenario depending on the performance of the model and deployment scenarios. Besides, Michael expects the system to be at least 99,5 % correct all the time if he is to consider deploying it into production. This is due to the risk of false negatives (to a lesser degree false negatives), which has a major negative influence on the business. However, as explained in the M-L section, Recall explains what amount of the positive classifications were actually correct, and is therefore a better technical measure for deciding on the amount of false negatives. Optimizing Recall is therefore a primary business goal.

As the expectations for this academic research is to test out a DL model on a specific business case, focus is on the methodologies and concepts used and not on the model's results. However, to provide a direction for the research, a concrete business goal is formulated. Moreover, the model is not expected to fully automate the process, but rather be deployed for decision support. This will be elaborated in the deployment section.

Based on the above considerations, the specific business goal is:

"*To automate the above process by use of CAD and achieve a recall above 0,50 and a positive expected value*."

4.2 Data understanding

The available data consist of more than 15.000 x-rays, of which 5.000 are estimated to show OCDs (positives). However, as mentioned above the process of discovering what x-rays have OCD and then label them accordingly, is very time-consumring as the physical medical records containing information about the diagnoses are

separated from the digital x-rays, cf. Appendix 7 (it portrays the physical medical records). A lot of time has been spent on documenting diagnoses in a spreadsheet and acquiring data from the cloud systems. The researchers therefore decided to experiment early on with a few data and continue collecting data on the side in order to start the modelling phase as quick as possible. This would allow early reflections on how to conduct experiments and what the results could look like.

The amount of original data collected in the end was only around 600 x-rays without OCD and 200 with OCD. In addition, for the final experimentation the researchers decided to discard collected x-rays that were taken from the front (also known as the DP projection, cf. Appendix 8. This was because almost none of the x-rays were diagnosed with OCD, and the projected image deviated a bit from the two other projections. This limited the amount of x-rays used for training, but it allowed more similarity between the x-rays, which could benefit generalization. Unfortunately, it also limited the business scope of the model, as it could not be used to discover patterns from this projection of the hock.

Despite the heavy time-burden for collecting the data, the data proved to be useful for supervised learning, as it is easy to match the diagnose with an x-ray and then label the x-ray accordingly. However, some diagnoses had mistakes and the following describes typical veterinarian (human) errors when detecting OCDs on x-rays:

- Differentiating between the diseases OC and OCD. OC is just a "hole" in the bone, whereas OCD includes both a "hole" in the bone and a fragment separated from the "hole".
- Some OCDs are located at a place on the x-ray, which are unusual. They are sometimes not discovered.
- Some OCDs are so small that it is very difficult for the human eye to spot them and differentiate between an x-ray which has no disease and one which shows OCD – see Appendix 5: X-rays with OCD (difficult cases). These are called "borderline x-rays".
- Young horses (below 1 year old) tend to have immature bone developments, which could be seen as OCD disease on the x-ray, but it is actually not. Detecting this error require external knowledge about the age of the horse.

As was discovered during the experimentations the above error types have also affected the labelling of the data, as some of the x-rays ought to be catagorised as negatives even though they were diagnosed as positives. These images have been re-labelled by the researchers and with guidance from Michael Hansen.

The third error type above is the one that most often leads to false negatives. Michael specificly hopes that the model is able to detect these cases. This error type was also discovered during the experimentations, and the model also had difficulty in labelling these correct. As the produced model does not take into account age of the horse, the fourth error type is expected. Taking into account the age of the horse ought to be done in a follow-up research.

As can be seen in Appendix 8, x-ray examinations of the hock are carried out from the same three angles. This create a high similarity between the projected x-rays which benefits generalization. Moreover, since OCD is typically located more or less the same place on the x-ray, this also add on to the OCD pattern. This is also confirmed by the M-L expert, Jacob Axelsen, cf. Appendix 1. However, as Jacob also notices, a biological fragment that has been breaked from the bone is an *irregular object*. This means that it does not follow an evolutionary law, since there is no selection of breaks. Therefore, the object is more or less co-incident in its shape. According to him, one way to go about this problem is to use segmentation (which is a different M-L problem).

Despite the lack of data and the irregular nature of OCD, the collected x-ray data are still considered to be reliable and valid for experimentations with a model based on image recognition.

4.3 Data preparation

The first step in the data preparation was to document the diagnoses of the horses in a spreadsheet by manually looking through the physical journals, cf. Appendix 4.

The documentation of the diagnoses would allow the researchers to discover the x-rays that should be labelled with OCD and which one should be labelled with "Not OCD". Based on these findings, the relevant x-rays could be discovered on the cloud database and downloaded locally.

As mentioned above the data collection was very time-consuming and not enough data was collected due to this. The classes are therefore skewed, since the final collection of data totalled approx. 600 data (75%) that are negatives (labelled "Not OCD") and 200 data (25%) that are positives (labelled with "OCD"). Also, the total data (800) is usually not enough for training aneural network. Therefore, in order to get more data the researchers decided to scale up data. More specificly, the data was scaled up to approximately 1300 positives and 1300 negatives.

In order to scale up the data, the researches have used different data augmentation techniques including rotation, which rotates the image in one direction. Also, image translation has been used to translate the x-ray a few pixels in each direction and zooming in on the x-ray. These techniques have generated the extra data needed, where each data is an unique x-ray.

Since the training time increases with the amount of pixels in the image (as mentioned above, the pixels are the input to the network), the image resolution has been resized down to 256 * 256 for the final experimentation. This has reduced training time, which was very necessary for producing model 2.

The x-rays have also been cropped down to the following:

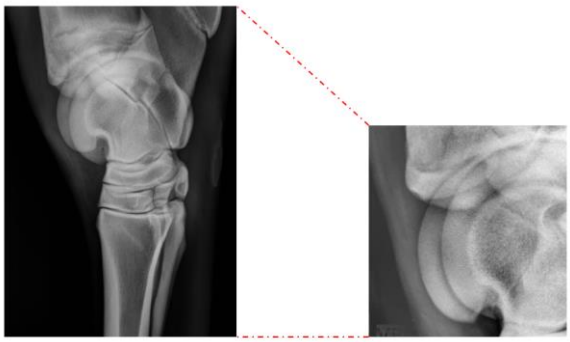


Figure 5: Cropping down the x-ray to the region of interest

The cropping allows the filters to only focus on the region of interest, where OCD is most commonly found. This will reduce training time, but also increase generalization as the x-rays now appear even more similar.

For the training and testing of the models, the total amount of data has been split into a training set and a test set. Since the modelling of model 1 is based on a pre-trained model, no data has been used for validation and parameter tuning, as there is no parameters to tune.

In relation to the data generation process and its assumptions, the training and test data are independent from each other, but for model 1 - the training and test set are not entirely identically distributed as the following table shows:

	Model 1	Model 2
Training set	1344 Not OCD (56%) + 1068 OCD (44%)	1344 Not OCD (56%) + 1068 OCD (44%)
Test set	167 Not OCD (64%) + 93 OCD (36%)	100 Not OCD (56%) + 79 OCD (44%)

Nevertheless, Model 1's test set concist of the same test data as Model 2' test set, but with additional data that has been added. This small variation in test data ought to be levelled in a follow up experiment, to provide a more precise benchmark.

4.4 Modelling

The following section describes how the author performed data modelling in IBM Watson. Also, it will provide a descriptive analysis of model 2, which has been modelled in GooglenTensorflow. For the final experimentation, both models have

been trained on the same data set and approximately the same test set. This provide a useful foundation for benchmarking.

4.4.1 Model 1: Modelled in IBM Watson – visual recognition platform

During the modelling phase different experimentations were conducted. IBM Watson Visual Recognition service were used as the platform to train a custom made classifier that labels the images into negatives (Not OCD) or positives (OCD). The process is very straightforward as the author only needs to upload the data to the network and decide on number of classes and name of labels. Watson then outputs a label on the image as well as a confidence score. The scores are comparable and range from 0.0 to 1.0. The one with the highest score are more likely to appear in the image than the lower one. However, they may both be present. (IBM4, 2017) It is estimated that an intelligent automation of this process can provide three types of benefits:

1) Lower the cost of producing diagnoses and training the staff due to automation of the detection process. Also, this will allow staff to focus on more value-adding activities. (labour cost savings)

2) Mitigate risks of false negative diagnoses (increased accuracy/recall)

3) Mitigate the risks of inconcistent diagnoses in the industry (increased concistency)

There is an option to use the custom-made model in conjunction with Watson's base labelling service, which would add a set of built-in labels. However, based on some initial experimentations the built-in tagging service did not seem to provide any real value, as the tagging included tags like "black" and "ligament". The author therefore decided to focus on the custom-made labels. Also, the training time was only about 2 minutes for the largest training set (2669 images), which is mainly due to the fact that most of the network are pretrained, and that it is only the high level layer that has to be trained. In comparison, the average training time for model 2 was two hours.

The initial experimentations quickly showed an acceptable performance on the training set, which only concisted of 326 images. However, the performance on the test sets were much worse. To proceed from there the author evaluated the cost and feasibility of gathering more data as an approach to increase performance. Compared to large internet companies that has easy access to data, gathering data from a medical application seems to be more costly in terms of time. However, as there were no other real alternative, the author decided to proceed with data collection.

4.4.1.1 Overfitting evaluation

As mentioned in the M-L section, two factors decide how well the M-L algorithme is performing: *The ability to make the training error small and making the gap between training and test error small*.

For the final experimentation with model 1, the training error was 0%. However, the test error was 22,69%, which could suggest overfitting. This means that the model tend to tailor the model to the training data, at the expense of generalizing to previous unseen data points. This could be due to the complexity of the model. One way to better diagnose overfitting is by crating a fitting graph. However, since the "complexity" of Watson is propritory knowledge it is not possible to get that information. For the same reason it is not possible to apply regularization techniques in order to lower the test error.

The only approach is therefore to add more data to the training set.

4.4.2 Model 2: Modelled in Google Tensorflow

This model is build in Google Tensorflow and by use of the Keras library. This model is build from scratch, which gives certain advantages in terms of what is possible to do with the network architecture, e.g. customize number of layers and filters.

The following figure 6 shows the network architecture of model 2:

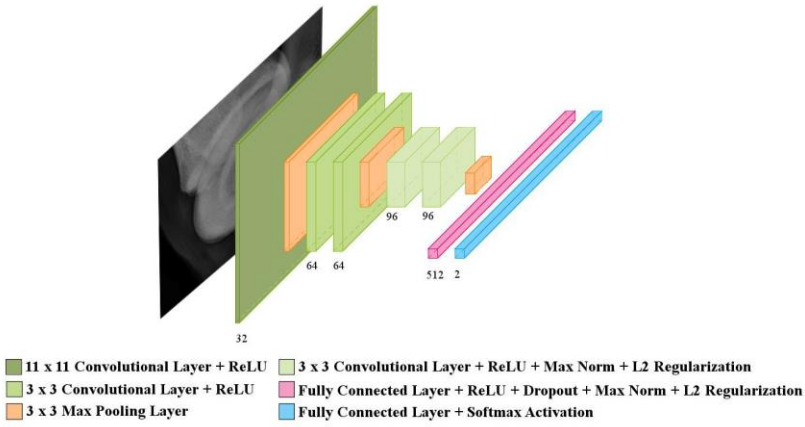


Figure 6: Final network architecture (Created by the thesis partner)

The architecture portrays the different layers in the model. Next to each layer is a number describing the depth of each filter and the number in front of the layer name, e.g. "11 * 11", describes the size of the kernel filters. The layers applies both convolution, pooling, the activation function ReLu, softmax and fully connected layers as explained in section 5.1.1.4.

The above architecture also has fully connected layers. These are layers that implies that every neuron in the previous layers is connected to every neuron in the next layer. The purpose with the fully connected layer is to use the output from the convolutional and pooling layers (high-level features), for classifying the input into

various classes, e.g. disease or no disease. (Goodfellow, Bengio, & Courville, 2016, p. 202)

Also, the network uses regularization techniques including Dropout, Max Norm and L2 regularization. These techniques are used to minimize the generalization error/test error.

The selection of the above parameters can influence both the time or memory cost of running the algorithme or the quality of the model, i.e. its performance. For this model, the parameters have been chosen based on Grid earch. In short Grid search involves picking some values for each parameter and then the grid search algorithme trains the model. The validation set error will then decide which parameter is the best. (Goodfellow, Bengio, & Courville, 2016, p. 428)

4.5 Performance evaluation

The following section will address the performance results from the final experimentation with the models.

4.5.1 Accuracy and other technical measures

Accuracy:	77,31%
Error rate:	22,69%
Precision:	0,81
Recall:	0,47
F-score:	0,60
Coverage:	100%

Table X: Performance stats for model 1

Accuracy:	82,68%
Error rate:	17,32%
Precision:	0,93
Recall:	0,66
F-score:	0,77

Coverage:	100%

Table X: Performance stats for model 2

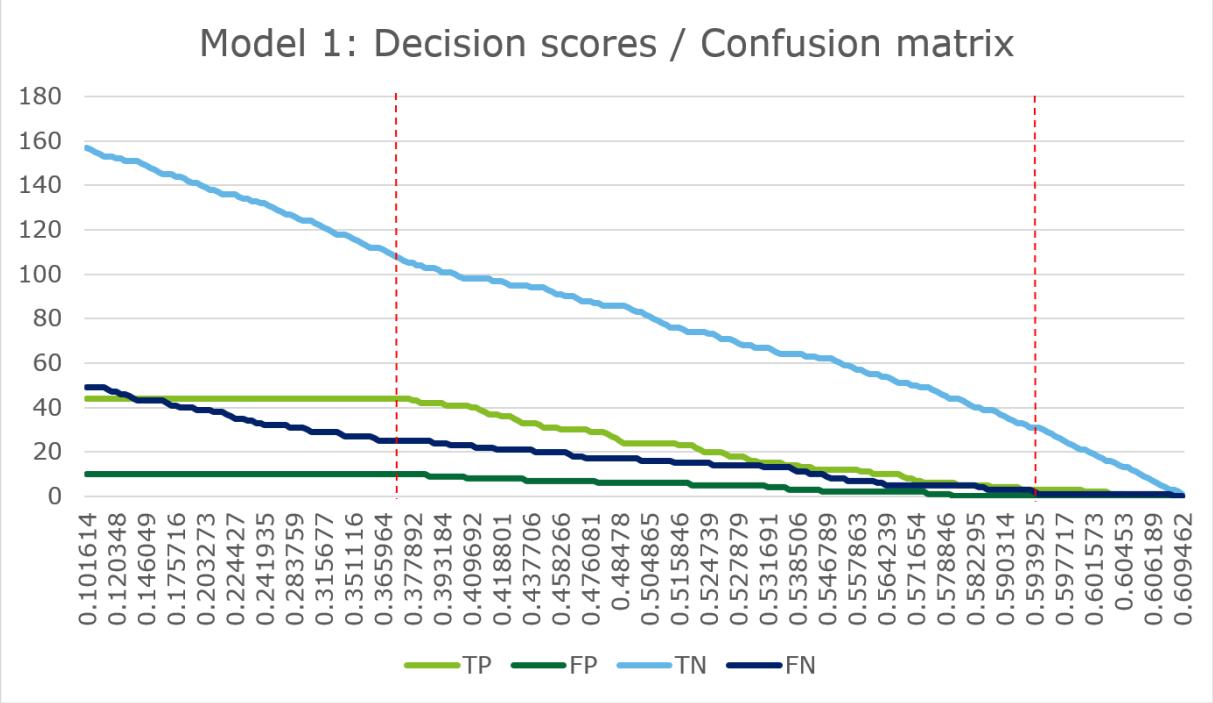
The technical performance measures shows that model 2 outperforms model 1 in all performance stats. The 77% and 83% in accuracy for the test set is neither bad nor a very good generalization capability, but somewhere in the middle. The approach to increase the accuracy of both models is to add more data. Also, for model 2, parameter tuning and regularization techniques also has the potential to increase accuracy.

As will be portrayed below in the confusion matrix, accuracy might be miss leading. Accuracy does not distinguish between types of correct and wrong predictions, which makes it difficult to know the distribution of false negatives. Recall is a better measure for that. However, recall does not provide any information about how that performance affects the business and neither is a single number very informative. In terms of recall, Model 2 has a recall of 0,66 compared to Model 1's 0,47. Based on these results, only model 2 fulfill the business goal of a recall above 0,50. This suggest that for both models not many actual positive observations were predicted correctly, i.e. many false negatives.

The precision of both models are relatively good, which means that most positive predictions were actual positive observations. The F1 score is also relatively good. However, the F1 score weights precesion and recall equally, but this research is more concerned with recall. The F1 score is therefore not perfectly suitable for this case scenario.

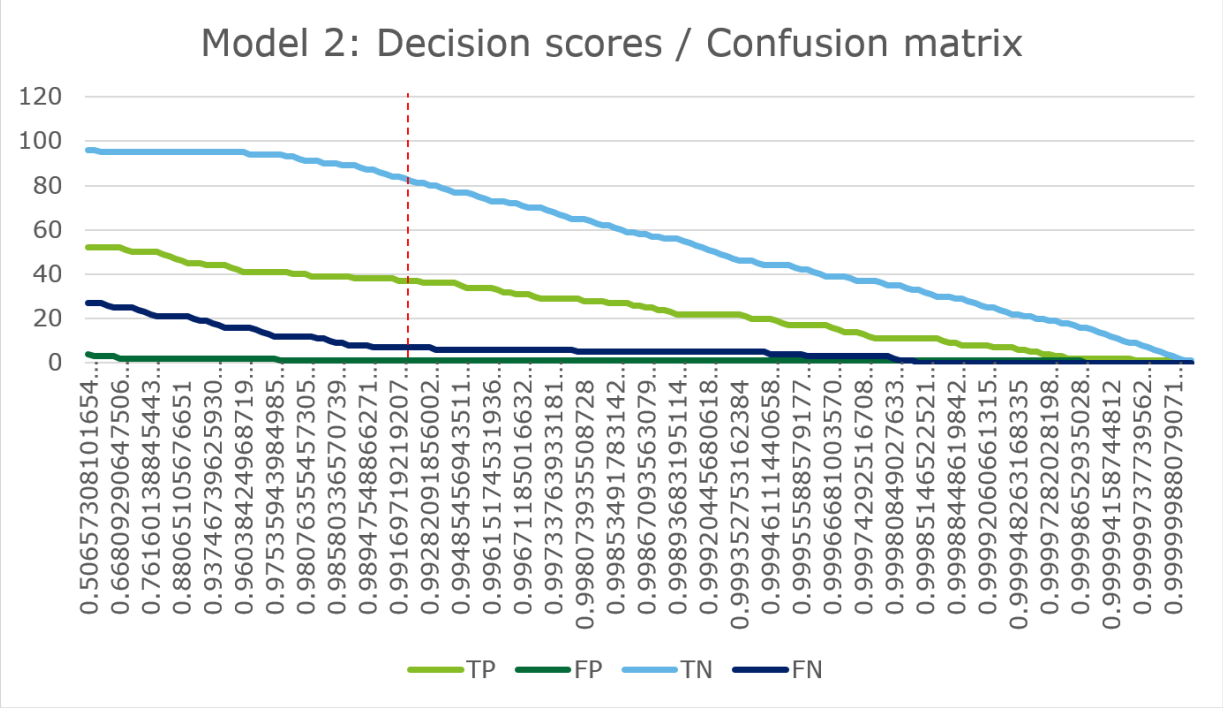
4.5.1.1 Decision threshold based on ranking classifier

Since every prediction by the models create a ranking classifier with a specific confusion matrix categorization (TP, FP, TN, FN) and a confidence score, it is possible to plot the categorization (y-axis) along with its confidence scores (x-axis). This allow the researchers to discover whether the distribution of categorizations have any relationship with the confidence scores. It also allow the researchers to decide on a specific threshold that will maximize recall.



This graph describes the distribution of confusion matrices based on the decision score. Recall can be calculated from the confusion matrix and it is therefore possible to decide from the graph what is the size of Recall (by looking into TP and FN), i.e. a decision threshold at 0,593 corresponds to a recall at 0,75 and a decision threshold at 0,370 corresponds to a recall of 0,64 (marked with red dotted lines). The lower the decision threshold, the more data is being put into the model. There is therefore a trade off between a low decision threshold and recall. The threshold could therefore be increased in order to increase recall. However, that would decrease the relevance of the model, as it would only take a decision on the few data it is most confident about. This is also why both graphs indicate that the models predict incorrect labels with a lower confidence score compared to the correct labels which have a higher score. As seen on the graph TPs and TNs are the ones that increases fastest in the higher end of the decision scores. In the lower end FNs surpasses TPs.

Moreover, according to IBM scores around 0,5 suggest that there is a significant amount of similarity between the classes. The data are therefore not distributed in distinct clusters, and the scores reflect this closeness to the best boundary between positives and negatives that the model can learn. (IBM4, 2017). This also suggest that the error type mentioned in the Data Understanding section (borderline x-rays) is present in this model, which create a lot of false negatives.



Similarly, model 2 has the highest recall at a decision threshold around 0,991, which corresponds to a recall of 0,84. Since this model has very high confidence scores, compared to model 1, it would still include most data even if the decision threshold has increased to 0,991. This suggest that it might be worth putting a threshold on model 2 in order to optimize recall. This model has also more scores distributed in one end of the confidence score ratio, which suggests it is better at distinguishing between negatives and positives.

4.5.2 Expected value framework to frame classifier evaluation

The following section describes the different information sources in order to be able to calculate the expected value for each model.

4.5.2.1 Confusion matrix, model 1

Total number of samples: 260	Predicted Positive	Predicted Negative
Actual Negative	False positive 10	True negative 157
Actual Positive	True pos 44	False negative 49

4.5.2.2 Confusion matrix, model 2

Total number of samples: 179	Predicted Positive	Predicted Negative
Actual Negative	False positive 4	True negative 96
Actual Positive	True pos 52	False negative 27

4.5.2.3 Expected rates

The calculations of the rates are based on the data in the confusion matrices above.

Expected rates, model 1	**Expected rates, model 2**
Total = 260 Tp rate = 44/93 = 0.47 Tn rate = 157/167 = 0.94 Fp rate = 10/167 = 0.06 Fn rate = 49/93 = 0.53 Positives (p) = 93 Negatives (n) = 167 p(p) = 0.36 p(n) = 0.64	Total = 179 Tp rate = 52/79 = 0.66 Tn rate = 96/100 = 0.96 Fp rate = 4/100 = 0.04 Fn rate = 27/79 = 0.34 Positives (p) = 79 Negatives (n) = 100 p(p) = 0.44 p(n) = 0.56

4.5.2.4 Cost-benefit matrix

Factoring the weights

While documenting the journals digitally the researchers also documented the conclusion of the diagnoses. Conclusions of a diagnosis can include the following: "*Nothing has been found*", "*The diagnoses don't affect the use of the horse*", "*It*

cannot be stated that the diagnoses will not have an affect on the use of the horse" and "T*he diagnoses do not have/have a significant influence on the use of the horse*". The third conclusion *"It cannot be stated that the diagnoses will not have an affect on the use of the horse"* is typically followed up by a recommendation from the veterinarian to do a surgery.

Based on the results in the excel file "*overview of journals*", it is calculated that out of a sample of 144 horse journals that are diagnosed with OCD in the hocks, 74 of them were given this conclusion. That is around 50% for every horse diagnosed with OCD. Although it cannot be argued that for every horse diagnosed with OCD other contextual factors (like other diseases) might not have contributed to that conclusion, it is still a reasonable indicator for how often the veterinarian assumes that OCD in the hock affects the use of the horse. However, the veterinarian only recommends a surgery if the use of the horse demands it or if the horse is to be traded. Also, the client has to accept that a surgery is to be executed. The researchers therefore assume that the final probability for a surgery to be executed based on x-rays displaying OCD in the hock, is around 33,75%.

Cost-benefit calculation

The cost-benefit matrix is based on information collected from the interview with Jørgen Michael Hansen. It is therefore external information that seeks to describe the consequences of decisions in the context of our detection problem. The calculations are based on assumptions and approximate ranges. This has been used for simplicity of calculation. (Provost & Fawcett, 2013) (199)

All values will be expressed as benefits and with costs being negative benefits. The calculations and assumtions are displayed in Appendix 6. The following describes the cost-benefits for each decision:

A *false positive* occurs when the model classify a x-ray with OCD, but there is no OCD. This decision includes revenue from an atroscopi investigation (surgery) with a factor of 33,75%. Costs include salary to two technicians and one veterinarian for doing the atroscopi. The benefit in this case is 3437,60 DKK= b(Y, n). Even though it is a monetary benefit, veterinarians suffer from an uethical decision. As Michael puts it: *"On a long-term basis false-positives will damage our reputation, and therefore should be avoided. However, they are preferred from false negatives",* cf. Appendix 2*.*

A *true positive* occurs when the model classify an a x-ray with OCD, and it is correct. The revenue from an atroscopi investigation is included with a factor of 33,75%. Costs include salary to two technicians and one veterinarian for doing the atroscopi. The benefit in this case is 3437,60 DKK = b(Y, p)

A *true negative* occurs when the model classify a x-ray with "Not OCD", and it is correct. In this case no revenues or costs is generated since the horse is "healthy". The benefits are 0 = b(N, n).

A *false negative* occurs when the model classify a x-ray with "Not OCD", but there is actually OCD. This is the worst case scenario, since the business might have to compensate for the client's loss of earnings in relation to a dismissed sale or other associated costs. The main costs will be covered by the business' insurance company, and the self-payment is rather small. However, this scenario also include related costs in terms of loss of future earnings and reputation that could amount to much higher costs. Since these costs are harder to estimate specifically they have not been taken into account for calculation of the cost matrix. This ought to be done in a follow-up study. The benefits are -5000 = c(N, p)

These cost-benefit estimations are summarized in the following 2 × 2 cost-benefit matrix:

Cost-benefit matrix	Predicted Positive	Predicted Negative
Actual Negative	3437,60 DKK	0 DKK
Actual Positive	3437,60 DKK	-5000 DKK

Figure 7: Cost-benefit matrix for OCD detection.

4.5.2.5 Calculate expected profit of models

Given the above matrix of costs and benefits, these are multiplied cell-wise against the matrix

of probabilities, then summed into a final value representing the total expected profit:

Model 1 – expected profit

Expected profit: $= p(p)\cdot [p(Y \mid p)\cdot b(Y,p) + p(N \mid p)\cdot c(N, p)] + p(n)\cdot [p(N \mid n)\cdot b(N , n) + p(Y \mid p)\cdot c(Y , n)$

$= 0.21 \cdot [0.47 \cdot 3437{,}60 \text{ DKK} + 0.53 \cdot (-5000) \text{ DKK}] + 0{,}79 \cdot [0.94 \cdot 0 \text{ DKK} + 0.06 \cdot 3437{,}60 \text{ DKK}]$

$= 0.21 \cdot (1615{,}67 - 2650) + 0.76 \cdot 206{,}26 \text{ DKK}$

$= 156{,}76 - 217{,}21$

$\approx -60{,}45 \text{ DKK}$

This expected value means that if this model is applied to a population of (horse) patients and a surgery is done on the horses due to a positive classification, Højgaard can expect to make an average loss of about -60,45 DKK per x-ray.

Model 2 – expected profit

Expected profit: $= p(p) \cdot [p(Y \mid p) \cdot b(Y,p) + p(N \mid p) \cdot c(N, p)] + p(n) \cdot [p(N \mid n) \cdot b(N , n) + p(Y \mid p) \cdot c(Y , n)$

$= 0.44 \cdot [0.66 \cdot 3437{,}60 \text{ DKK} + 0.34 \cdot (-5000) \text{ DKK}] + 0{,}56 \cdot [0.96 \cdot 0 \text{ DKK} + 0.04 \cdot 3437{,}60 \text{ DKK}]$

$= 0.44 \cdot (2268{,}81 - 1700) + 0.56 \cdot 137{,}504 \text{ DKK}$

$= 250{,}28 - 77$

$\approx 173{,}28 \text{ DKK}$

This expected value means that if this model is applied to a population of (horse) patients and a surgery is done on the horses due to a positive classification, Højgaard can expect to make an average profit of about 173,28 DKK per x-ray.

Based on these calculations the business ought to chose model 2 as it is the only one which is profitable. However, to better visualize the impact on the business the following section will create profit curve for both models.

4.5.3 Profit curves

The following graph displays the profit curves of the two models:

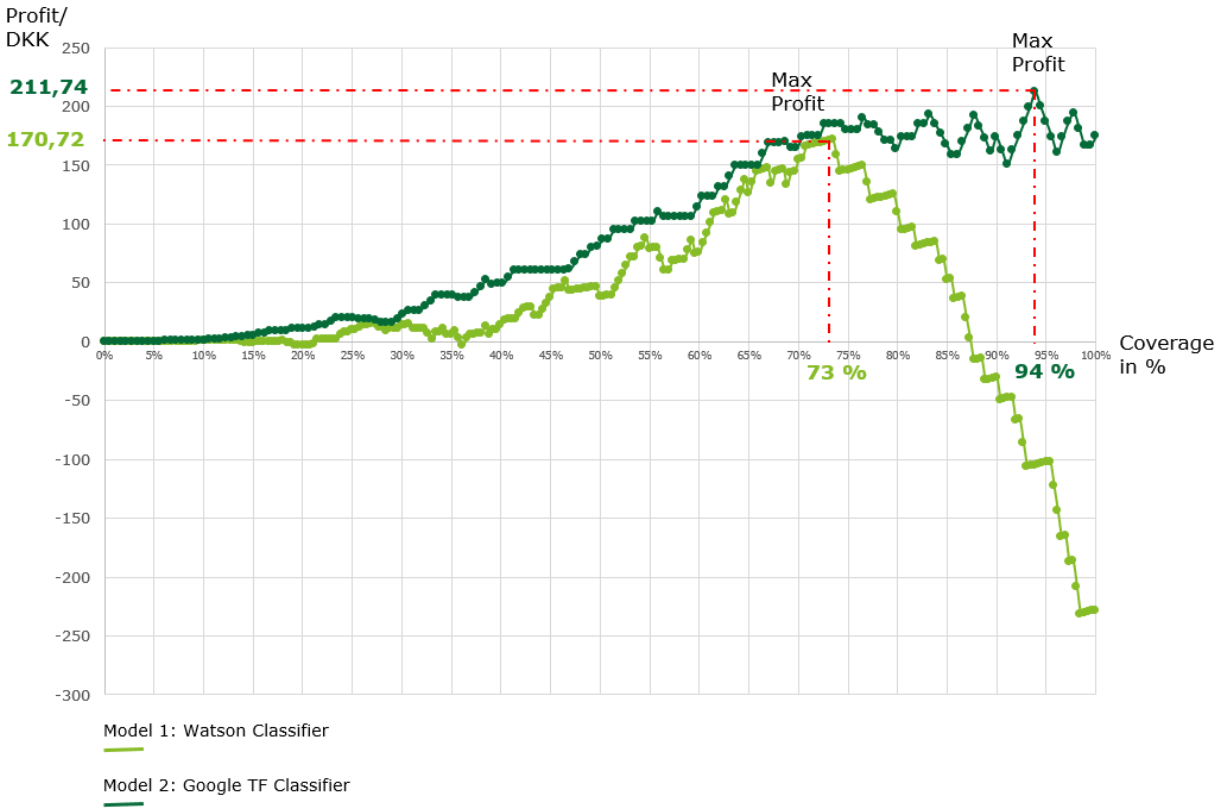


Figure 8: Profit curves of the two models. Each curve shows the expected cumulative profit for that model as progressively larger proportions of x-rays are targeted.

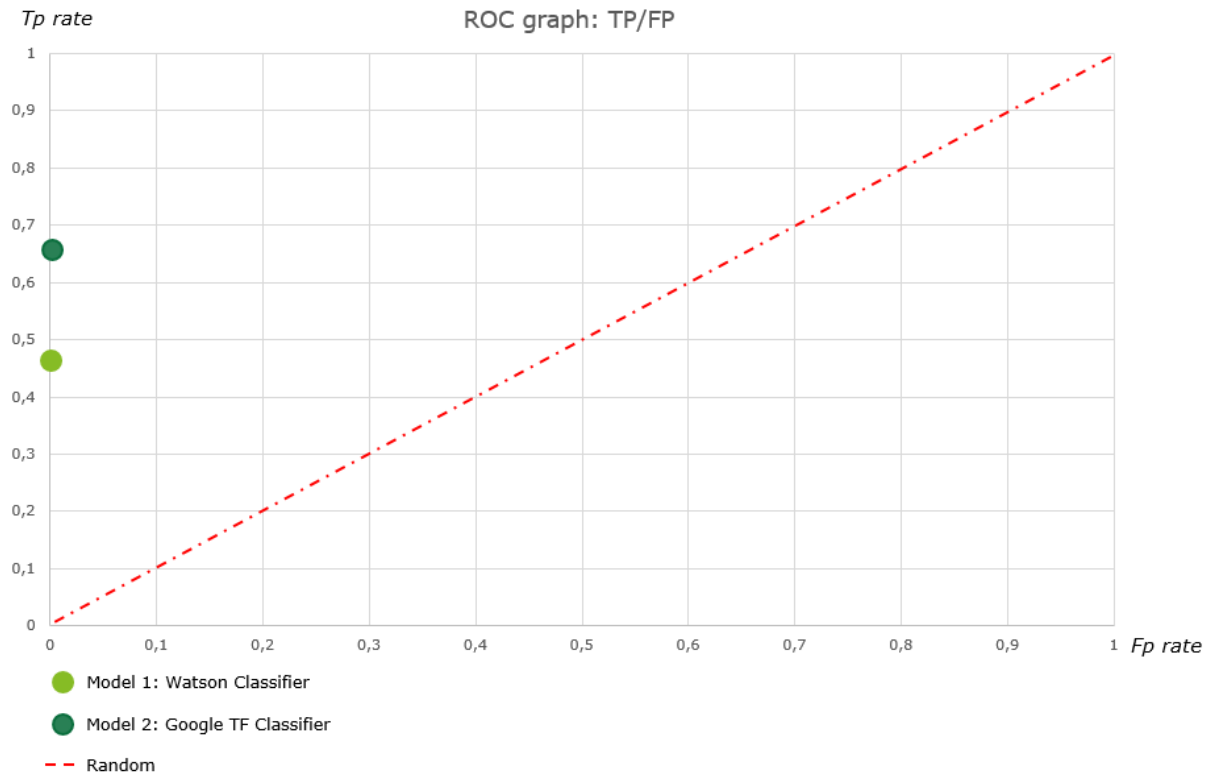
Model 2 produces the maximum profit of 211,74 DKK by targeting the top-ranked 94% of x-rays. If the goal was simply to maximize profit and the researchers had unlimited resources, the top 94% of x-rays should be chosen. The profit curve shows that profit for model 1 goes towards negative after having reached 170,72 DKK. In order to stay positive this curve suggest that the model should only cover 83% of the data.

In general model 2 outperforms model 1 everywhere on the curve. Højgaard is therefore more certain to gain value from model 2.

4.5.4 ROC graph

Profit curves rely on the conditions that the class priors and cost/benefit matrix are clearly identified. Since the true cost of a FN and FP does not include loss of reputation and future profits, it is difficult to get a precise measure of the cost. Also, the class priors may change, including the proportion of negatives and positives in the target population of x-rays. This is because it is very difficult to tell the exact frequency of OCD diseases among all x-rays taken.

The ROC graphs and curves can accommodate these uncertainties, which will be explained in the following.

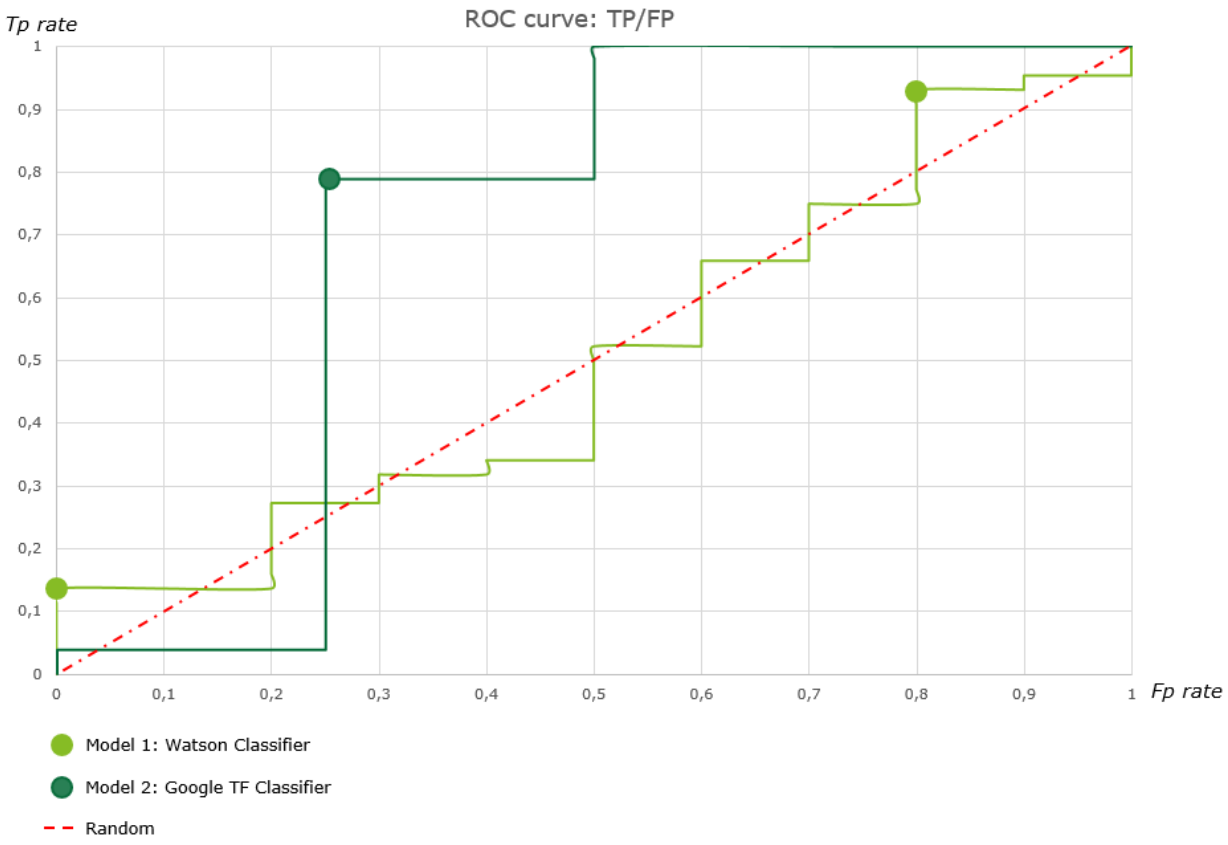


Model 1's classifier rate (0,06; 0,47) appear very conservative as it is placed on the lefthand side of a ROC graph, near the x axis. This type of classifier raise alarms (manke positive classifications) only when there is a strong evidence. They make few false positive erros, but they also have a low true positive rate.

Model 2's classifier rate (0,04; 0,66) is a bit closer to perfect classification (0, 1).

4.5.5 ROC curve

The following ranking model produces a curve in ROC space. It is used with a threshold to produce two discrete (binary) classifiers.



Model 1 shows that the ranking model is more or less random and sometimes worse than random. Model 2 is nearest perfection (the upper left corner) at (0,25; 0,79). This corresponds to a decision threshold of 0,976.

4.6 Deployment

The models have to be maintained and improved continuously and possibly adjusted for new business requirements. It is therefore important to get experience and confidence with using the models before they are used in a fully automated process. Therefore, Højgaard needs to take into account different maturity steps in order to gain a fully automated process. The potential of automation and cost reductions will amongst others depend on what risk the CEO of Højgaard accepts to take. The following figure 9 describes the necessary steps.

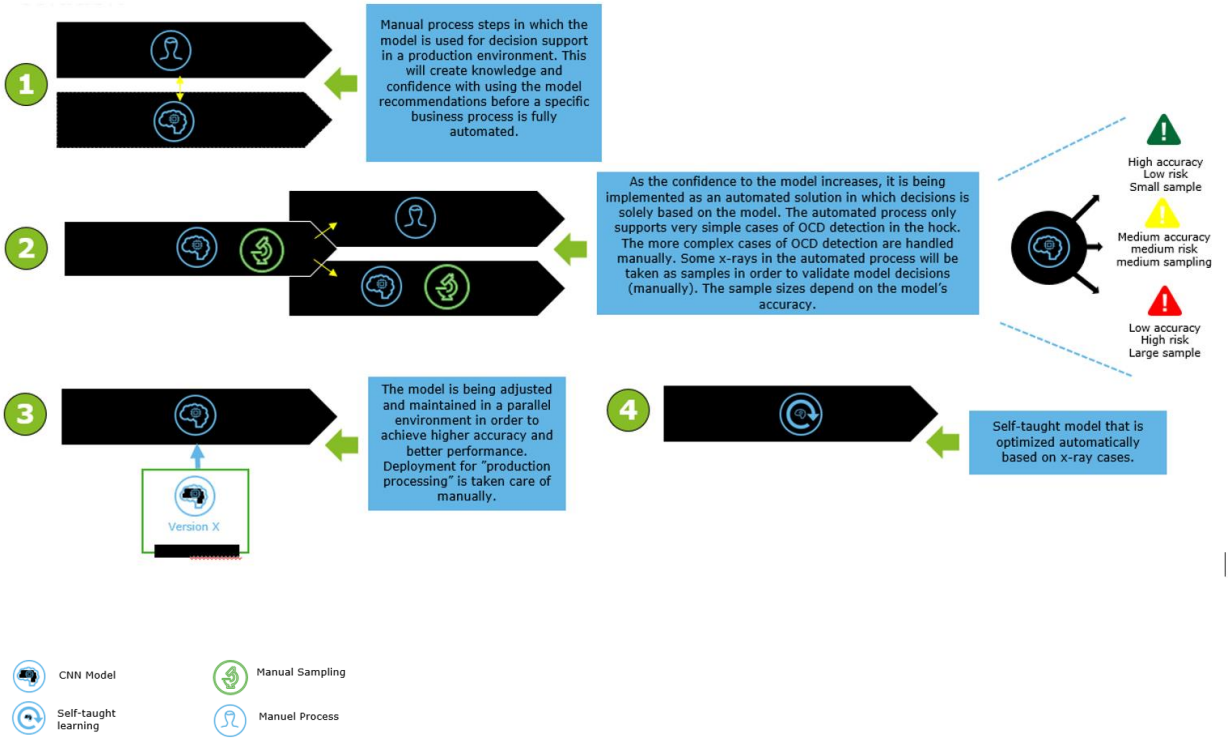


Figure 9: Recommendation for automation approach – 4 maturity steps towards an AI solution (author's own making)

4.6.1 Maturity step 1 – Model used as decision support

This process is manual and decisions are only supported by recommendations from the model. This use case only allows the model to give "second-opinions" on x-rays if needed. It is a safe deployment of the model that contributes to the quality of x-ray diagnoses. It also allows Højgaard to gain confidence in the model. However, this use case will not allow the business to gain any substantial benefits from automation yet.

This is the deployment approach initially recommended for the business.

4.6.2 Maturity step 2 – Model takes some decisions

As a starting point this process is fully automated. Depending on the model's recommendations, this process can be either fully automated or manual with

support from model. Recommendations with high confidence levels are more likely to be fully automated, since the business risk is low and only a small validation sample is needed. Recommendations with lower confidence levels ought to be guided by manual decisions, since the business risk is high and bigger validation samples is required. Overall, this use case allows the business to gain experience and confidence with process automation under "low-risk" circumstances.

4.6.3 Maturity step 3 - Towards a Model that can take more decisions

New x-rays will continuously be collected in order to improve the model. The model's accuracy is depending on the amount of collected data. Therefore, all x-rays of hocks ought to be collected in order to be used for future improvements of the model. Results from manual samples will also be added to the training data in order to optimize the model. The model will be adjusted and analysed continuously in order to discover whether new features should be included in the model. This will ensure that the model complies with current business challenges and issues with detection of OCD diseases. Finally the model should be further testet and evalutated to achive improvements and robustness.

4.6.4 Maturity step 4 – Semi-supervised model will drive full automation

The whole process is fully automated in this model and the model is able to learn by itself, e.g. by use of generative models and unsupervised learning. (European Research Council, 2016)

4.6.5 Other deployment concerns

Other concerns not addressed, but equally important for deployment is to look into the platform & architecture relevant for deployment of this system. The model has to fit right into the the process landscape for x-rays. Data flows from the x-ray proprietory database to the model should happen automatically in order to make the model simple to train and test. Also, the platform should be suitable for future developments. If the system scales and is applied for other automation tasks the business also needs to consider creating governance and organization around it, including a support team to maintain it.

5 Results & model benchmark

Performance metrics	***Model 1: IBM Watson Visual recognition***	***Model 2: Google Tensorflow***	***Comments***
Discrete classifier			
Accuracy	77,31%	82,68%	
Error rate	22,69%	17,32%	

Precision	0,81	0,93	
Recall	0,47	0,66	
F-score	0,60	0,77	
Coverage	100%	100%	
Expected profit	-60,45 DKK	173,28 DKK	
ROC graph (FPrate;TPrate)	(0.06; 0.47)	(0.04; 0.66)	Model 2 is nearest perfection
Ranking classifier			
Decision threshold	0,594 (Recall: 0,75)	0,991 (Recall: 0,84)	Model 1's threshold is very high, which decreases the relevance of the model.
Max profit	170,72 DKK (target top 83%)	211,74 DKK (target top 94%)	
ROC curve (FPrate;TPrate)	Close to random	(0.25; 0.79) – corresponds to decision threshold 0,976	Model 2 is nearest perfection

The table above summarizes the performance results of the two models. Model 2 outperforms model 1 in all performance metrics. Also, it is the only model that has a discrete classifier that achieves a recall above 0,50 and a positive expected profit, which complies with the required business goals.

However, if a decision threshold is put on model 1 (at 0,594), it will also comply with the recall requirement, but it will also limit the relevance of the model. Also, what is not reflected in the performance metrics, is that model 1 is easier to train, as it does not require any parameter tuning and does not require comprehensive developer skills. The Watson platform is build for novice developers, and allows easy access to Watson's infrastructure and access to a customer support team 24/7 hours a day. But it is also based on a subscription and therefore has an additional cost, if it is to be used above a period of one month. Using the free version also has restrictions on training size (20 data instances per batch).

On the other hand, Google Tensorflow is free and an open source library, and much more advanced in terms of technical options. An open source network lowers the barriers for using it, but since it it also demands more technical knowledge, only

people with a strong technical background will benefit from using it. Also, Google Tensorflow does not provide access to any cloud GPU or sufficient processing power. If a business wants to gain any value from modelling large amount of data with DL, it should therefore include additional infrastructure costs, when using this library.

To sum up - if Højgaard was to deploy a model today based on the current performance, it is recommended to choose model 2. However, each platform has its strength and weaknesses for modelling a CNN. The author recommends the business to chose a model based on a holistic decision that includes considerations on business requirements, performance requirements, platform and architecture requirements.

Moreover, it is recommended that the deployment of the model follows a step-wise process and initially only deployed for decision support purposes. This allows the business to better understand how the model can be used and in time, when the business is more mature, it will be applied more widely. This approach mitigates the risk of failure.

As the current accuracy is not high enough for the CEO to accept deployment of the model into his business, there are currently no business opportunity in this business. However, this might change if more data is added and the performance increases. Also, according to Michael, smaller veterinarian businesses, auctions for horses and insurance companies are also interested in a system that can detect diseases on x-rays. These businesses do normally not evaluate x-rays themselves, since they do not have the necessary experience. Nevertheless, that is not needed if they have a system that can do this for them. Future research should also include these stakeholders.

6 Discussion

This section will discuss important reflections on the research approach, as well as the ethical considerations surrounding the topic of artificial intelligence.

6.1 Reflections on the research proces

When reflecting on the whole research process, several things could have been done in order to optimize certain processes and activities. In terms of data collection this could have been started earliere, i.e. already before the thesis research begin. This is because data collection always appears to be more difficult than expected. Beginning earlier with the data collection would provide a better overview of accessible ressources already in the beginning of the process. This is because most research projects tend to lack data – including this research project.

Moreover, as the research field of AI and image recognition is considered to be an advanced topic, even for computer scientists, early guidance by experts within these domains is important. This research received guidance from domain experts within academia (supervisor), industry (Deloitte Consultant) and case company (the CEO).

Still the frequency of meetings could have been increased in order to better handle mistakes along the way, i.e. reduce the amount of time spent on correcting mistakes and focus more on data collection.

In general, simplicity is the way ahead when trying to understand the complex conceps behind AI and deep learning. That is why this research has taken an approach that de-conceptualizes the concept of AI into sub-concepts, methods and procedures - breaking it down concept by concept until the necessary understanding is gained. This has assured faster learning within a limited amount of time.

Benchmarking two models developed on different platforms have assured diversity in the research. It has allowed the research to exploit what platform is better at solving the specific case problem and it increases the robustness of the performance results. Also, adding the expected value framework and profit curves to the performance evaluation have provided a better idea of how the models affect the business model. These are very useful for communicating the results to business stakeholders that can better understand these measures. It also assures easier stakeholder buy-ins on the model. One way to contribute with more qualitative performance measures is to retrieve user-satisfaction surveys from the veterinarians, who have tested the use of the system. Finally, the x-ray existing x-ray examination processes ought to be changed and adapted in order to maximize the automation potential of the model.

6.2 Ethical considerations

Since the model contributes to automation and thereby eliminating a manual task performed by a human, it is relevant to discuss if it is ethically correct to replace a human with a machine.

According to Andrew McAfee, MIT research scientist and author to the book "Race Against the Machines", machines like Enlitic (mentioned in the introduction) points to the fact that many jobs, including radiographers, are in danger of being taken over by robots and computers. "*If a job is a routine, it can be done by a machine*", according to McAfee. Since radiographers' job is to match patterns, they are basicly doing what a machine can do "*very well, much faster and without getting tired*" (Straitstimes, 2017). Moreover, as examining x-rays is an important part of the veterinarian's job at Højgaard, there is a possibility that parts of a veterinarian's job will be automated by artificial intelligent machines. The model developed in this research is a step in that direction.

However, McAfee also mentions that the category of jobs that require social interaction is safe from being automated. This is also the opinion of Michael, who believes that communication with the client is an emotional thing that requires a human discussion, especially because the owners are very emotionally attached to their horses, cf. Appendix 2. McAfee recommends one to accept that technology will *race ahead*, since this will also bring many good things, i.e. higher quality and low-cost medical diagnostics could be extended to people who currently don't have great healthcare. In the case of veterinarians, one could argue that veterinarians could be

retrained to focus on more value creating activities, like client management. This type of task require negotiation, empathy and problem solving skills, which is hard for a computer to learn. These skills will allow humans to compete *with* and *not against* machines. (Straitstimes, 2017)

Based on McAfee's opinion one should therefore not prevent automation, but accept it, and change focus to other value creating activities. Based on his opinion, it is ethically correct to replace a human with a machine.

7 Conclusion

The results of the research clearly states that Machine Vision application can be used to detect OCD patterns in x-ray data and that performance results can be communicated with profit curves and ROC graphs to better target business stakeholders. The customized model outperforms the pre-trained model on all performance metrics, which suggest that the flexibility gained from parameter tuning also benefits the final performance and even for a smaller set of training data. However, the general performance results also indicate that the models need to be improved before they are ready for deployment and used as decision support. The models need to be trained with more data that can be collected from the database, and eventually scaled up by use of data augmentation techniques. Future research should also implement best practices, including prioritizing data collection techniques and frequent support meettings with domain experts.

7.1 Future outlooks

Future research should first and foremost focus on how to improve the performance of the models as this will increase the business relevance. Increasing the amount of data used for training the model, is a very useful approach for increasing model performance (Goodfellow, Bengio, & Courville, 2016, p. 428). Future research should therefore focus on collecting more data.

Also, it could be interesting to explore whether it is possible to train a model to report more meaningful decisions to the veterinarians. This will be explained in the following.

7.1.1 Semantic modelling

One approach to create a more meaningful model is to make it able to detect where OCD is located on the image. This can be achieved by use of *segmentation* techniques. Image segmentation is the process of partitioning the image into multiple segments (sets of pixels). By assigning a label to every pixel in the image, pixels with the same label will share certain characteristics, e.g. same color, intensity or texture. This will result in a set of segments that covers the entire image (Shapiro & Stockman, 2001). Since OCD diseases share textures that are more or less similar, this approach could potentially help detecting where OCD is located.

Another approach is to use a pre-processing technique that *tiles-up the image* into smaller images, which are then analyzed individually against the custom classifier.

The results are then assembled into a bigger image to provide a view into where within that image OCD conditions are recognized by the classifier. (IBM5, 2017)

The below image displays an example of how the x-ray can be tiled up into smaller images. In this case the procedure was to drop the x-ray image into a browser that uploads the image to a Node.js. application. Then the image is "chopped" into smaller tiles, where each tile is analysed with the Watson Visual Recognition service. The results are then visualized in a heatmap visualization. The colourization is based on confidence scores returned from the custom classifier created in this thesis (model 1). The red colour displays the highest confidence score, i.e. where the model is most certain to discover OCD. (Github1, 2017). The x-ray below could be tiled up even further in order to recognize more precisely where OCD is located on the image. However, it still suggest that it is more confident on certain areas than others.

If the model could be trained on more data and modelled with e.g. segmentation or tiling techniques, it could be possible to detect where OCD is located on the x-ray, and thereby add more meaning to the model.

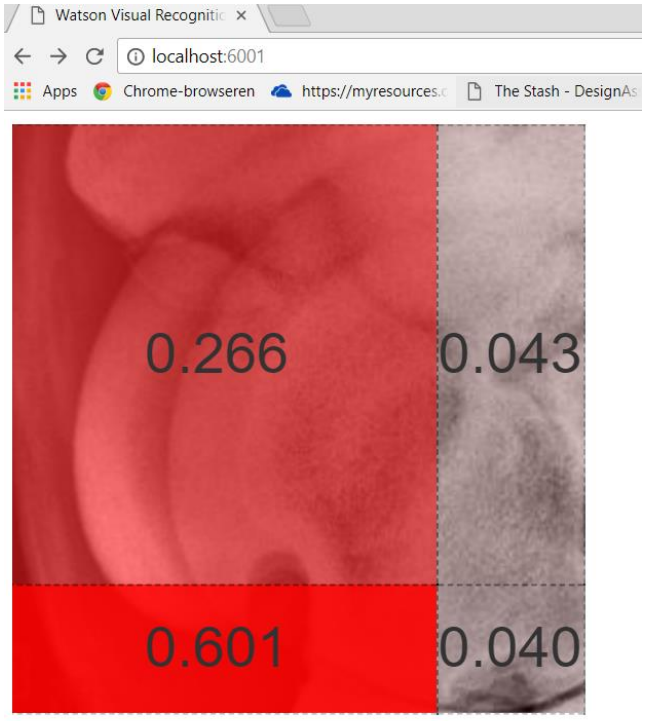


Figure 10: Tiled-up x-ray with heatmap colourization

8 Bibliography

ACVS. (2017, July 25). *Osteochondritis Dissecans (OCD) in Horses*. Retrieved from large animal: https://www.acvs.org/large-animal/osteochondritis-dissecans-horses

Albon, C. (2017, August). *Precision, Recall, and F1 Scores*. Retrieved from Model evaluation: https://chrisalbon.com/machine-learning/precision_recall_and_F1_scores.html

Cast Study - Image Data Management System - Oracle - Novartis (IBIS 2008).

Constantiou, I., & Kallinikos, J. (2015). New games, new rules: big data and the changing context of strategy. *Information Technology*, 44-57.

European Research Council. (2016, July 22). *SELF-LEARNING AI EMULATES THE HUMAN BRAIN*. Retrieved from Stories: https://erc.europa.eu/projects-figures/stories/self-learning-ai-emulates-human-brain

Fenn, S., Mendes, A., & Budden, D. (2015). Addressing the non-functional requirements of computer vision systems: a case study. 77-86.

Forbes. (2017, September 7). *IBM Invests $240 Million Into AI Research Lab With MIT As It Struggles In AI Battle*. Retrieved from NewTech: https://www.forbes.com/sites/aarontilley/2017/09/07/ibm-invests-240-million-for-ai-research-lab-with-mit-as-it-struggles-in-ai-battle/#51c5e207612e

Forsyt, D. A., & Ponce, J. (2003). Computer Vision, A Modern Approach. *Prentice Hall*.

fyens.dk. (2016, march 29). *Fynsk hestehospital populært i hele Norden*. Retrieved from Erhverv: http://www.fyens.dk/modules/fsArticle/index.php?articleid=2968286

Gartner. (2016, August 16). *Gartner's 2016 Hype Cycle for Emerging Technologies Identifies Three Key Trends That Organizations Must Track to Gain Competitive Advantage*. Retrieved from Hype Cycle: http://www.gartner.com/newsroom/id/3412017

Github - CS231n. (2017, August 30). *Transfer Learning*. Retrieved from Github: http://cs231n.github.io/transfer-learning/

Github1. (2017, september). *Visual-Recognition-Tile-Localization*. Retrieved from Github: https://github.com/IBM-Bluemix/Visual-Recognition-Tile-Localization?cm_mc_uid=96470940168114929797367&cm_mc_sid_50200000=1501614401&cm_mc_sid_52640000=1501614401

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning.

Højgaard. (2017, July 1). *din hest i fokus*. Retrieved from Højgaard hestehospital: http://www.din-hest-i-fokus.dk/hojgaard-hestehospital/

IBM1. (2017, May). *The developer blog*. Retrieved from Train and evaluate custom machine learning models of Watson Developer Cloud: https://developer.ibm.com/dwblog/2017/machine-learning-custom-models-watson-developer-cloud/

IBM2. (2016, October). Retrieved from DeveloperWorks: https://developer.ibm.com/answers/questions/316140/visual-recognition-capabilities/

IBM3. (2915, april 8). *Five new services expand IBM Watson capabilities to images, speech, and more*. Retrieved from https://developer.ibm.com/watson/blog/2015/02/04/new-watson-services-available/

IBM4. (2017, August 17). *Watson*. Retrieved from IBM: https://www.ibm.com/watson/developercloud/doc/visual-recognition/customizing.html

IBM5. (2017, August 2). *How to sharpen Watson Visual Recognition results with simple preprocessing*. Retrieved from Bluemix Blog: https://www.ibm.com/blogs/bluemix/2017/03/sharpen-watson-visual-recognition-results/

ITN. (2017, February 24). *How Artificial Intelligence Will Change Medical Imaging*. Retrieved from Artificial Intelligence: https://www.itnonline.com/article/how-artificial-intelligence-will-change-medical-imaging

ITN. (2017, February 24). *How Artificial Intelligence Will Change Medical Imaging*. Retrieved from https://www.itnonline.com/article/how-artificial-intelligence-will-change-medical-imaging

Järvinen, Guyatt, & Gordon. (2016). *Arthroscopic surgery for knee pain.* BMJ.

Klette, R. (2014). Concise Computer Vision. Springer.

McCoy, A., F., T., N. I., D., S., E., J., E., K., O., . . . C.S., C. (2013). Articular osteochondrosis: a comparison of naturally-occurring human and animal disease. *Elsevier*, Volume 21, Issue 11, Pages 1638-1647. Retrieved from http://www.sciencedirect.com/science/article/pii/S106345841300914X#fig6

Provost, & Fawcett. (2013). *Data Science for Business.* United States of America: O'Reilly Media, Inc.

Razavian, A. S., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). CNN Features off-the-shelf: an Astounding Baseline for Recognition. *Royal institute of Technology*.

RBR. (2016, May 11). *Industry 4.0: Robotics Presents a Golden Opportunity*. Retrieved from Manufacturing: https://www.roboticsbusinessreview.com/manufacturing/industry_4-0_robotics_presents_a_golden_opportunity/

Ribeiro, E., Uhl, A., Wimmer, G., & Häfner, M. (2016, October 26). *Exploring Deep Learning and Transfer Learning for Colonic Polyp Classification*. Retrieved from NCBI: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5101370/

Rosenbeck, A. (2015). Practical Python and OpenCV. *2nd edition*.

Russakovsky, O. D.-F. (2014). ImageNet Large Scale Visual Recognition Challenge.

Saunders, M., & Tosey, P. (2012). *The layers of research design.* Retrieved from http://www.academia.edu/4107831/The_Layers_of_Research_Design

Saunders, M., Lewis, P., & Thornhill, A. (2009). Research methods for business students. *Pearson Education.*

Shapiro, L. G., & Stockman, G. C. (2001). *Computer Vision.* New Jersey.

Shearer, C. (2002). *The CRISP-DM Model: The New Blueprint for Data Mining* . Retrieved from https://mineracaodedados.files.wordpress.com/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf

Shin, Roth, Gao, Lu, Xu, Nogues, . . . Summers. (2016). Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE*.

Spampinato, Palazzo, Giordano, Aldinucci, & Leonardi. (2016). Deep learning for automated skeletal bone age assessment in X-ray images . *Elsevier*, 41-51 (2017).

Stackexchange. (2017, April 17). *Questions*. Retrieved from Stackexchange: https://stats.stackexchange.com/questions/273189/what-is-the-weight-decay-loss

Straitstimes. (2017, March 6). *Race with machines, not against them: MIT research scientist Andrew McAfee*. Retrieved from straitstimes: http://www.straitstimes.com/singapore/education/race-with-machines-not-against-them

Tellis, W. (1997). *Application of a case study methodology.* Retrieved from http://www.nova.edu/ssss/QR/QR3-3/tellis2.html

The_data_science_blog. (2017, August). *An Intuitive Explanation of Convolutional Neural Networks*. Retrieved from https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/

Venturebeat. (2014, October). *Enlitic picks up $2M to help diagnose diseases with deep learning*. Retrieved from Entrepreneur: https://venturebeat.com/2014/10/28/enlitic-funding/

Wang, & Summers. (2012). *Machine Learning and Radiology.*

Zainal, Z. (2007, June). *Case study as a research method.* Retrieved from Journal Kemanusiaan.

9 Appendixes

This section contains appendixes that has been referenced in the thesis

9.1 Appendix 1: Interview with Machine Learning expert

Date: 15-03-2017

Location: Deloitte

Time spent: 01:10:03

Attendees: Interviewer and Jacob Axelsen

I: One thing we are looking into is how many layers do we need and the depth of each layer

J: That is hyperparameter tuning..

I: Yes, we have tried to do that…

J: You do not need to worry – it is very trivial

J: If you need to do image recognition you can just do transfer learning… you take a platform…

J: I would recommend to use Tensorflow – if you are good – just stick to that. If you have not triede it before, but e.g. think that the mathematical operations and graphical operations are difficult, then you can use some APIs. The most common used specificationAPI is keras

I: Well, he is already doing that…

J: The problem with keras is however that at some time you run into a wall.. these are the most used… Keras is just a package – it pack things into tensorflow… but the problem with Keras is that it only solves the trivial part of the problem – it looks like it does not, but it actually is the case.. Therefore, no matter what, you should not spends too much time training networks up from the ground, unless you have a lot of data?

I: We do not have enough data – we do not have enough time to gather everthing. We have approx.. 2000 images and of those maybe only 500 images have OCD disease…

J: Yeah, but with 500 images you have to do transfer learning, because you don't have a chance to design a network that is able to have great success on data at that size… but transfer learning can do that… you can do that very quickly in Keras.. you just write: Load and inception version 3 or rest-net or exception.. Google have a new one named MobileNet.. these are pretrained models that contains best practices… take the new ones since rest-net and version 3 are some years old and do not use the new best practices – they use the most heavy procedures. With this you can train a model in a couple of hours with transfer learning – training time is therefore not an issue. The problem is whether you can recognize those weighs you would like to find with that type you are looking for.

I: You have previously said there was a problem with looking at an irregular object as a fragment

J: What you can recognize with these graphs are regular stuff like faces – since they have been chosen by evolution. Bones are also typically regular – however, your fingerprints, hair and so might be irregular. A break on a bone is not a regular thing since it follows a physical law. It does not follow the evolution, since there is no selection on breaks. Therefore, it will have a "breaked" line, which is co-incident. What you therefore need is to do segmentation of images… that is what I am doing. It is quite difficult and requires a lot more data material… When you e.g. make it look for cats it does not just see a cat but it sees the cat body, a cat laying down, cats from different angles, small cats – you therefore have a lot of microclasses that is all directed towards the cat. So the problem is how do you make the network point at the cat? That can be even more problematic if you have an irregular object. So when filters learn a lot of different forms of lines and edges in all angles – you can do that with a

lot of tools from Keras – you can pull out those filters. They are typically good at catching lines from many different angles – these lines are regular and not irregular. It will look after elipses and squares instead – so it will look after some microclasses. You should therefore hope that the fragment you are looking for in most cases have a regular form. You just have to experiment, I cannot gurantee it works…

Another problem is that you cannot make it count the number of fragments on the picture. That you can do if you have segmented it beforehand. Segmentation concists of different activities. One of them is that you cover the images until the p-value of the object you are looking for will decrease – importance card, focus card… you will cover the image and make an unlimited amount of predictions for each small part of the image – what if what you are looking for occurs in different places on the image? So when you are covering one part of the image the p-value decrease a little, but you still see it.. then you should cover both places at the same time and then the combination possibilities increases and then you have to make a lot of predictions.. You need to know what you are doing when looking into focus-cards… Another thing to do, which you do a lot of within bioinformatics, is to make some segments which contain some boundaries and within you have a colour, which is the class that the object represent on the picture. E.g. if you have a bike, then there is a line around the motorbike and within it is a colour of the motorbike, and the biker is on top and there is a line beneatch the biker and a colour within it so that he has been filled out.. since his legs cover parts of the bike, his colour will cover the bike colour. This is a semantic scheme analysis – but it requires very advanced dataset. You need someone to draw lines around the bones and fragments. Then you need a palley, fill it out and do it for all your 500 images. It could be a good investment for you… you can also use Amazon's Mechanical Turk - that would be my biggest recommendation.

I: Have you yourself experiences with using Mechanical Turk?

J: No, but I know that a lot of people are using it. We would like to use Deloitte's own channels. We have Deloitte Pixels, which also uses Mechanical Turk, so we will go that way, and expect to do a crowdsourcing experiment, where they will segment cars for us, since there exists many different images of cars as well taken from many different angles. There can easily be hundreds of different classes… It was especially difficult since we did not have very good data. I had to create data with a 3D program.

I: In relation to lack of data, do you have any recommendation for how to overcome that?

J: You can increase the size of the images by making a lot of operations on it – add noise, translate it – you can also do symmetrically operations. There exists packages in Keras that can do that for you. Then you have to train with a generator, but when you use a generator, you become more and more sensitive to what you can do and what you cannot do… I would recommend that you stick to Tensorflow instead.. a real computer scientist ought to start in Tensorflow. People who have knowledge with computer algorithms don't have to work with specificationAPIs. Only use these APIs for operations on the graph itself, since this graph we get for free if we use regular image recognition. If we do segmentation then there exists some new tools that are able to deal with less data material – I have also worked with those. These are very new. It is a network that generates images and then there is two models: One of them creates the images and the other "one says Yes or No". One of them is just an autocoder – we ask it to create itself or create a goal – the most simple generator (not always an autoencoder) is actually a network that generates the input it gets. When it has gone very effective and the error function on your input and output has been minimized then it has created the best possible autogenerated output that it is able to do. But maybe you want it to create something else like e.g. segment stuff. Then you need a tool that is able to look at the qulity of these segments – there you have a discriminator that can do that. All kinds of machine learning are discriminators, incl. SVM, regression models, classic NNs, decision trees. They discriminate on inputs. Discriminators are very simple… these other advanced method can generate data for you. That is a good thing if you do not have a lot of data. But they are called "GANner" generative adversarial networks. You need to look at these, and if you want to go in that direction, and have the time, then you should look into WasserStein GANner – they have solved the problem of how to train them. With these you can always train them. But this case where you both have a discriminator and a generator, they cannot do anything in the beginning. But there is a dilemma, if one of them becomes very good the other one might become bad… it is very complicated – but they are good if you don't have so much data… This is a very advanced method.

I would recommend you to focus on transfer learning. You can focus on the most already trained models – you can use those and freeze the most ordinary layers and open up the higher layers – and train them to look for those objects you would like to find. Start with those and see how long you can get. But don't be sad if the data material do not support it enough… 500 labelled images are a very few – but it could be enough. It is not that bad.

I: It is also a matter of how much time we have

J: One thing you should be aware of is that one thing is knowing where the data is and another thing is actually getting them out…

I: Yes, the images have also been cropped

J: You should also notice that it does not matter if you use HD images – it is not so important – You can see what you need to see with just 256 * 256 resolution – 512 * 512 should also be more thatn necessary. This is due to the filters that are looking into transfers – if you can see it then the filters ought to be able to see that as well… But sometimes you can also design noise that can ruin what the filters can actually see. You should consider cleaning up noise on the images… But the images you showed me seemed quite clear to me.

I: Yeah well some of them are more clearer than others

J: It is quite easy to eliminate noise. You just take all your images and put on different kinds of noises and then you create an autoencoder on the noisy picture image the more clearer image. You could start out doing that just to get some experience with how you can actually do that. There is an example on that on the Keras blog. But just notice that if you do not find anything then it could be because something is "fooling you". There is 10,000 of things that could go wrong. Maybe the books of deeplearning seems quite straight forward, but they are not. It is much more difficult than how it appears.

I: We have considered bechmarking our project with IBM Watson e.g. – have you had any experience with that…

J: You can provide classes to IBM yes – but I only believe Bluemix looks after classes. But you can give it images of the different images – it will be like your own transfer learning. That you can do if you like, but that will then be pure transfer learning. You cannot do autoencoding or check for noise… But if common transfer learning solves your problem then you can use IBM Watson, so just put images with fragments in one of the classes and then the rest (with no fragments) in the other classes, and then some images without labels and take e.g. 100 of each of them an check how it is doing… The training time might take 1 hour or so – it is a good idea and then you don't need to do anything

J: However, I have not worked a lot with those vendor platforms, I have more used the classic tools…

So transfer learning is where you will begin and then you get some benchmarks and performance metrics – but when you need to tweak it there is a lot of other stuff you can try out and see what you think.

I: We consider using recall instead of precision

J: Yes, you can do that. There is a matrix on Wikipedia that is explained in many ways. It is about having as many observations within the diagonal..

I: Have you previously heard of any other experiments that have been looking into bones?

J: There are plenty – I have a contact who also looked into this, but was very skeptical about the results… but I have seen some image repositories around that uses image transcription to identify different kinds of tissues…

J: Maybe image transcription can be used – currently I am using pix-to-pix, but that is oriented image transcription. You have two images and the object is totally fixed oriented. They will be located on precisely the same place on the picture. That is pix-to-pix, but it made a splash some time ago… You

can e.g. transcribe a line drawing to an object. If you have drawed a cat it can transcribe it to a cat. Maybe you want to throw away TNN'er completely away and then look into segmentation… currently I am looking into holistic embedded edge detection, HED, they have also used pix-to-pix and that one can be used without training – they have been pretrained in Caffe, especially Caffe 2.0 should be useful. But difficult to learn about many platforms… Experts from the different platform will usually advocate for their own. When I started I used Theano – then I realized that the once who made Theano have also made Tensorflow. They just translated Theano. Keras therefore works well on both platforms. Some years ago you could only work on Theano, whish is a very heavy platform. What you can do today is much more easy than working on Theano. It was totally impossible to work well with Theano on a Microsoft machine. Whereas Tensorflow runs directly in Windows. In February I went to a data science meeting for data science in Deloitte and spoke about GPU supported tensorflow – nobody knew what it was. But it is surely the best… Also, even though it is not the best platform, Torch is also good, but not supported by Windows, but Linux. I like to work with a platform that works on all machines. Tensorflow is good, but some of the other platform also have their advantages. One barrier you might meet is training time, influenced by how you choose your parameters. So either you run a lot of models on a lot of small nodes at the same time and then you have put your hyperparameters so you don't feel that you have wasted your time or else you do data parallisation, but that is difficult with Tensorflow. With data parallisation you make replication of the model that all with have a small part of data, and they then forward their predictions to a parameter server, which will then decide what gradient needs to be and then it is send back to all the replications. Settting up that is pretty difficult in Tensorflow. It is more easy in Keras. All the time new tools is added to this library. But if you receive huge datasets then you need data parallisation if you have to just train once or twice. Big data will mean a lot and also memory – big GPUs are very expensive, HPC card is also useful… Sometimes people forget that a really expensive GPU card can do the same as a super server or super cluster that 10 years ago would have costed 10 times as much. These are only useful for a specific type of parallisation, which is the close matrix intensive calculations – you throw a highdimensional matrix object around the different nukes – that is what "kudu" is doing. You cannot parallize everything, e.g. you cannot throw Twitter on a GPU.

I: Okay, so interesting. Do you have any experiences with working with ImageNet?

J: There are a lot of different classes on ImageNet. Transfer learning has been trained on imageNet, so the filters are pre-trained. But all of the objects are regular.

I: We look into x-rays and they are black and white. Do you know if it is easier to work with that kind of image format?

J: It is easier to make a gradient on white-black because you can almost see where it is located. But you can make gradient on everything… But if you do that, then you are also doing what your network is supposed to do. In principal the filters ought to do it. Let us say that you look into histrograms on colourisation. On x-rays you will have a histogram with most being in the lower end, so a lot of black pixels, a bit of very white and some grey. If you want to level it out that it is possible. But I would rather work with super resolution. You can work with network that can change the image or add information… but there are millions of things that you can do. Are your x-rays always taken from the same angle?

I: The images we are looking into are always taken from the same two sides

J: That is a very big advantage to you. You should really appreciate that. Because then you might not have at all to work with machine learning, but can maybe work with a SVM with a kind of Gaus "kerne". If I was you I would do both ML and SVM at the same time. If the angle on the images is always the same, then there might not be a reason for doing ML. I would not even consider ML if I had images taken always from the same sides.

J: Also consider if the crack on the bone or fragment is more or less located at the same coordinates on the image?

I: It is usually located on the same place on the image

J: If I was you then I propose that you do a HED segmentation and a SVM – you would save a lot of time. I have used ML to do text classification, but that was because there were no other way I could

change it or transform it to a SVM. You can do ML, but then you need to create a parser, and only do it, if there is no other easy way.

On the previous project I did we created two models, with one of them being a deep learning model. Look into SVM and the right kind of "kerner". If you have really good edge detection and use K-means clustering and decide on how many clusters you are using – then probably a SVM can solve the problem.

J: How long time do you have?

I: We only have July month left.

J: Then do transfer learning with Keras and immediately thereafter look into SVMs. Have you run any classifiers yet?

I: Yes, my partner has

J: You are in time shortage. My experience from working with clients is that it takes a lot of time.

Start immediately with looking into this docker file from Github: his username is s9xie/hed..

Go into that place and he has it and the docker file is placed under pix-to-pix repository under Tensorflow. The docker file can be run without a GPU. It is run very slowly – but it runs. It is located on afinelayer/pixtopix/tensorflow. The docker file can make it possible for your HED to work. Then you have state of the art edge detection and probably your SVM can use that. If the image always is from the same side then you can much easier work without deep learning.

9.2 Appendix 2: Interview with Michael Hansen

Date: 24-03-2017

Location: Højgaard Hestehospital

Time spent: 01:02:03

Participants: Interviewer (Rasmus) and Jørgen Michael Hansen, CEO of Højgaard Equine hospital

I:

OK - today is the 21st of July 2017 and the clock is 3pm and I'm doing an interview with the CEO of Højgaard hestehospital Jørgen Michael Hansen.

I:

So Michael let's just begin. Michael, the topic we are going to discuss is about how the X-ray diagnosis services are carried out at the vet hospital. So I'll just give you some questions related to that topic. Initially, I would like to ask you about the business model around the X-race diagnosis and if you have a budget for doing X-ray innovation at Højgaard Hestehospital?

M:

We don't have a budget specifically related to that part of our business - X-rays are mainly part of two business areas for us. And the one is routine examinations of horses where owners wants to know the radiographic status of their horse and the other main area is horses being traded and sold from one to another where were both buyer and seller wants to know the radiographic status on that horse.

I:

And if you compare x ray diagnosis service between these two services are they any different in terms of which of those are more expensive than the other or you generate more revenue for one of them.

M:

The x-ray part is comparable from one to another. The only difference is that when the horse is being traded there is a clinical examination correlated with the radiographic examination. The price and the amount of time spent on the radiographic examination is the same for either of these two.

I:

If you compare radiographic examination services to other services at the hospital – does it amounts to a huge part of the revenue of HH?

M:

This is a major business for us and for many of my colleagues. You can hardly traded any horse these days without radiographic examination.

I:

So as you know we are specifically focusing on developing a model for OCD detection in the hock area and we are trying to find out what is the cost associated with this service.

Obviously when you have bought the x ray hardware, which is the heacy cost, the only cost is employee salary, like time spent on the service and maybe training. So how long time does it take to do an x-ray examination?

M:

I think the price of the hardware and software that you need to invest in before amounts to around half a million DKK. And then you come to the examination.

We have technicians to cover the radiological examination, as we are quite a specialized place. Two two nurses and two technicians doing the examination. They start giving the horse the sedation so the horse stand still while the horse while being X-rayed. It is a standard procedure. The standard procedure for taking 14 x-rays takes around 20-30 minutes. So you have two persons occupied with the examinations. After the examination the Veterinarian looks at whether the technical quality is good. And that is his responsibility to check if there are any abnormalities like OCD and afterwards to have a discussion with the owner who has required the examination. So take two technicians 20 minutes all together and and say 15-20 minutes afterwards for the Veterinarian.

I:

What the salary for these involved employees?

M:

A technician will have around 25.000-30.000 DKK a month. An employed Veterinarian will have around 40.000-45.000 DKK a month. A nurse is in the same range as a technician.

I:

If you had to do a surgery - how long time will that take?

M:

If you have to find an OCD leasion and you have to carry out an operation it will take two nurses the pre-induction of inastesia – when they have finished doing that the surgery is then carried out by a

Veterinarian. The two nurses spend around an hour and a half on the surgery and the vet will spend around half an hour to a full hour depending on how complicated the operation is.

I:

Does it require any experience to to diagnose horse? Is it something that requires some sort of certification in order to be able to know how to do the diagnosis - do you need to be trained before you are able to do that?

M:

You don't have to be certified to do the evaluation, but it does take a lot of experience to be able to do it. So you can if you want to start evaluating X-rays the day you leave the vet school, but in my opinion it takes at least 1-3 years looking at x-rays, before you have the necessary experience and qualifications.

I:

Let us speak about the cases that you're looking for on the x rays. OCD is one of them. How would you - is it a primary focus area when you do the x ray diagnosis. Is it a top priority to look for or do you have sort of like a ranking of diseases to look for?

M:

OCD is a top priority lesion or disease to look for in most breeds in Denmark and all over the world. Some breeds don't have the disease or to a very small degree, but other races have a lot of OCD. So so X-raying horses for OCD is is the major reason we are doing x-rays.

I:

OK. And where do you typically find OCD on the horse?

M:

There are like four or five pre-dialectician places where you often see the OCDs and pre-dialectican joints like we have the hock, the fetlock and we got the stifle and you also will see OCD at other joints, but these are the three main joints where you see OCD.

I:

Okay, and you said something about that OCD might depend also on which race you are looking at. Is there a difference between horses who carry OCD?

M:

Some breeds very rarely have OCD – Like like Icelandic horses. Why the warm bloods and the Trotter's and thoroughbreds, which are the major horse breeds – they typically have OCD.

I:

Do you see any trends in the diagnosis of OCD. Has it increased recently or is there no trends?

M:

There is a lot of work being carried out on the different breeds - trying to get rid of the OCD, but none of them has had any success. And the problem is that the OCD is probably linked to some of the genes

that are being selected for in other means, e.g. if you have the trotter you choose to focus on breeding genes that will make the horse faster. You select genes that are able to run fast or have an off-spring that can run fast. But somehow the genes for generating OCDs, they seemed to be linked to some of these factors. So if you select horses with these genes you also get more OCDs. It is the same for other genes, like larger horses that can run fast or jump higher. If you do your selection on that you will not see a decrease in OCD, but if you do a selection for lower range OCD you of course will see fewer of them. But so far those investing in horses, they have no for who is the fastest horse. So you use those

So far the onese investing in horses – their major interest is in creating a horse that can be sold for a lot of money, and not so much thinking about the increase of OCD side effects. They are not focusing on avoiding on OCD. But if you concentrate on that - there is someone creating an index that tells what are the risk of getting OCD depending on your breed. This information is available for breeders, but they choose differently. It is not a major criterium, when they select their stallion.

I:

S let's figure out the processes for doing X-ray examination. If you could put me through the whole process from when the client is thinking that there is some issue here and then contacts the vet to get it fixed.

M:

OCD is a disease or a lesion that is being developed when the horse is growing and the whole process has run to an end when the horse is 1 year old. So, if the horse has OCD it'll be there when the horse is one year old and it will not develop afterwards. So if you have a horse that is free of OCD, when it's one yeay old it is free forever. So what we do is we recommend our clients to have their horses X-rayed after they've been 1 year old and before they start training the horse which for forabyte is two years old and for most races 2-3 years old. The reason we recommend that is that if you have your horse at that time, we can give the client a view if there is any OCD lesion and we can give our opinion on if this OCD lesion is likely to affect 1) the use of the horse and 2) how it will affect the possibility of selling the horse at a later time. It is a very big issue if the horse has OCD or has not OCD or in between there, has been operated for OCD. Many horse owners will realise that selling horses with OCD fractions is very difficult as most buyers has heard about stories of horses with fractions that did not work or only for a limited time. So having the x-rays taking before you start working with the horse or before you want to sell the horse is in my opinion very important for two reasons: 1) to have a clinically sound horse for the rest of its life and 2) it is also a valuable investment if you want to sell it at a later time.

So the process is that the horses are brough to the clinic – it is being sedated and x-rayed – we have a discussion about the horse with the owner. If it is a normal horse, the client is happy and no problem. The client will start training the horse 1 or 2 years later. If it has an OCD lesion we will have a discussion about whether it should be operated on or if we say it does not amount to any risk. Also, the owner has to make a judgement by himself and decide whether he wants the horse to be operated. We could state that it would not have any affect on the use of the horse, but maybe on the selling of the horse. So it could be a good investment for having it operated. Also, we can never be totally sure of our judgement. It happens every year that we misjudge – we give the opinion that this fragment is probably going to cause clinically problems and it turns out that the horse is working the whole life withour problems – and we also say that this fragment should not be removed and then suddenly the horse shows clinical problems. This will happen.

I:

So how often would the owner accept your recommendation to do a surgery? Or are they reluctant to do it?

M:

If I think or recommend the fragment should be removed, because it's in my clinical experience will cause problems at a later time. Almost everyone will have the fragment removed. If I say that this fragment to my opinion will hardly cause any clinical problems, they will have to consider if it is a good idea to have the fragments removed, because if they intend to sell the horse at a late problem, there will be some discussion about that fragment. So then then it's up to the owner themselves to make a judgment of it's worth investing in an operation that probably won't make the problem any better or worse, but it will be a better price it can be sold for.

I:

OK. So if it is a case where he wants to sell the horse probably he also wants to accept the surgery.

M:

Yes but if he didn't came there with the intention to sell the hores, when if he wants to keep the horse for himself - I don't see any reason for not doing this.

I:

Could you just briefly describe the nature of these x-rays. I guess they're confidential so it is only the client who have access to them and you?

M:

Yes.

I:

And do you know of any public sources where you can find x-rays of horses?

M:

Many of these x rays are taken because the owner wants to sell the horse or wants to sell the horse at a later time and then we often get the potential buyer calling us and have an opinion on the x rays And we give our opinion if we are allowed to do that by the owner of the horse who was paid for the x-rays.

There is a research projects set up in Denmark actually with the aim of reducing or trying to reduce OCD in houses. And we were asked to participate with x-rays. We did send all X-rays to a central place where with our opinions on the x-rays saying OCD here and here. The readon for doing that is that they wanted to be able to find an index for those stallion that was being used as breeding stallion. So for example if I have 40 x-rays of the offspring of one stallion and analyses them and I say that 10 our of 40 shows OCD. Then that should give an indication of 25% risk of getting OCD if you used that stallion. But this is just for research purposes and trying to put up an index on stallions.

I:

Have you heard any results?

M:

No not yet.

I:

Have other hospitals contributed as well?

M:

Yes and the only goal was to create an index of the horses.

I:

OK - so they probably also have a database?

M:

Yes, but they were only allowed to be used for research purposes to try and create an index on the specifics.

I:

So to do that they had to get permission from the client?.

M:

They didn't need permission, since the only thing that could be seen was who was the father of this horse, which is not considered confidential the way it is used. They only needed permission from us.

They didn't really look at the x rays - what they were interested in was in our comments on the x rays. So it's like if I send up a set of x ray, I also had to send out my evaluation saying this has OCD in right hock and the father is.

I:

So these x-rays are probably also labelled in some way.

So thye could be interesting.

M:

Yes, they could be very interesting. There should be quite a number of horses with OCD.

I:

Do you know the name of this research project?

M:

It was carried out by Landsudvalget for Heste situated in Skejby.

I:

Could you just tell me a little about how you approach innovation. Is there many new technical developments and so on going on in your industry and changing the way you do veterinarian. Or is it still sort of a big conservative industry where it's not changing that rapidly.

M:

I think thing that the industry is changing rapidly and I think we are probably the business and major equine hospital changing most rapidly. It is not many years ago we used a developer to develop a

system and now it's third generation. I don't really know the names of these systems, but they are the online systems with very high quality – and in a split second that can send mails all over the worls. We also do scientography, which is a specific way of axaminating horses and MRI scans and CT scans. So I think we are developing just as fast as the human industry, but just a few years behind.

I:

Is there any business you consult when you do those IT investments?

M:

We are in contact with different businesses. When it comes to MRI scans we are in contact with Hallmark. With scientographi we are dealing with a company called MRI from Germany aslo we are deling with a Danish company. The software for x-ray distribution having it stored is a Swiss company. So we have different partners.

I:

So how do you decide that you will go to this company?

M:

We scan the market and then deside.

I:

What is the major risk when a human or veterinarian is diagnosing the X-rays. What is the biggest risk that can happen under such an X-ray detection.

M:

Well, different things can happen. I mean you look at the X-ray and give your best opinion but everyone makes mistakes. So there are different categories of mistakes and some are worse than others. I mean the two major categories of mistakes is that you: 1) miss something - you don't see it. So I'm saying this horse is okay – there is no signs of OCD. And it turns out at a later time maybe years later when the horse is being sold or something that there was something that should have been taken notice of.

That's one major issue and 2) it but your evaluation is wrong. And then again it's it's better to make the more conservative evaluationa saying that to my opinion this here is likely to cause a problem. The problem occurs when your judgment is wrong or the other way: Saying that what you're seeing here is not likely to affect the intended use of the horse and it turns out later that you were wrong.

I:

And how often do you believe these mistakes take place?

M:

The first category where you miss something and don’t see it happens very seldom and as it is obvious mistakes it usually is being covered by our insurance. The other mistake where you see it and the judgment is wrong will happens sometimes – normally our insurance does not cover this because… there has always been a lot of discussion about why this judgment is not correct. Also lawyers are trying to push that limit to have the insurance company pay as much as possible. But it is very few cases – we don’t have such cases ourselfes… but if we made such judgment we should be hold responsible.

I:

The cost could be many million DKK then, right? Has that happendes before?

M:

Yes. Yes. That happened before.

I:

So if you are dealing with x rays on a very expensive horse then you are probably also even more conservative?

M:

Yeah. Yes you are more conservative. Yes you are right because of the amount of money is much larger. It is a much bigger risk - but also horses being worth a lot of money are being used to the limit. They are very good horses and they are pressured to the upper limit. They are being used much harder than horses that are not as good.

So the risk is if you take a very good horse and he just rides in a field once a week and just put it back - nothing will happen. But if you want to compete in the Olympics and play all over the world every week the pressure on the horse is much higher. So the risk of something being significant is much larger.

I:

OK. And then let us talk about the model we are building for your business. What are your expectations to this system?

M:

I like a reliable second opinion…

I:

How would believe that would work in practice?

M:

In our place where x-rays are take by nurses or technicians – I like that they have already been scanned by the system So besides me looking at the whole - I have your evaluation by the side, so I draw my attention to things that need more attention. So that's how I'd like it. OK - so it's like I assume as soon as the x rays are being taken and loaded in the system they're being evaluated, so when I look at the x-rays I also have the evaluation on the same screen as the x-rays.

I:

So when the system becomes more reliable in time because the performance increases with the amount of data and what if it become able to fully automate the x ray diagnoses? How would you feel about that? That would of course reduce some costs, but would you be willing to handle such a task to a machine) Or is there some sort of barrier there that you don't feel totally comfortable with handing over so much responsibility to it?

M:

I think it's like other areas. I mean be ready to drive my car back home from me without touching the driving whell – probably in 10-20 years time. If the car is or systems is reliable I see no problems.

I:

Do you think there are some ethical concerns in relation to for instance employees that maybe one day will be the ones supporting the machines and not the other way around. Like if you sort of imagine where all this is going or is it.

M:

I don't see any ethical problems at all. I mean if this system can help doing 0 mistakes in the evaluation it is an advantage for anyone.

I:

And if we imagine the sort of scenario where machines just do everything. Where is the jobs for the Veterinarian then? Is it the client facing path then?

M

I think the communication part for the machine would be difficult to. I mean I need the system to do a waterproof evaluation of the x rays and then I will have to do the communication. I guess the system will come with the recommendations. Based on the system's experience it recommends to do an operation.

But I reckon that giving that message to the owner and the discussion will have to be a human face to face discussion.

I:

Do you think that is because we're dealing with something that is much more expensive or like it's another communication topic?

M:

I mean you're you're dealing with real live creatures. Because there is an emotional thing for many people.

I:

So let's imagine this scenario that we are able to produce a model, which based on the current data that we have and based on the tests we can predict OCD on x rays with above 90 percent accuracy and we can eliminate the false negatives completely.

Would you think about using this model in practice as sort of like you said yourself - a decision support, but maybe also more sort of like a tool that you could try once in a while to sort of get an idea of how this works. And then get some experience with if this is something that provides any real value at this point. Or in order for you to be able to deploy this very simple model do you need a bit more advanced diagnoses abilities?

M:

I think it is interesting - I think 90 percent is not enough. I mean if I made 10 percent mistakes I would not work in the business and knowledge of this is definitely not enough. Well participating and for test building the system is ok, but for commercial value it needs to be able to recognize OCD in the major

joints including all joins, but at least the major joins: The hock, fetterlock and the steifel. Then I think there is many areas where I can see the system being used. We are probably the ones having the less need of it, because we are very experienced, and we look at x rays for many other Veterinarians. So vetenatians being on their own and not seeing that many horses would have great support in having a system like this supporting their knowledge. And also you have auctions where other horses are being sold and these horses being sold at auctions. So if these auction houses could have the system and have it scanning the the x rays in conjunction - that could be another way of giving the information that is needed when the horse is being sold. So I think there will be need for it. But it has to be more or less waterproof. And if I should recommended it being on the wrong side it should - I would prefer the system to recognize that something it thinks is not being normal then not recognizing something.

I:

Are there many independent veterinarians out there?

M:

There are many Veterinarians out there who does not have the necessary knowledge or experience. I've had many X-rays being mailed to me asking for my opion.

I:

But that takes you out of business?

M:

It is not business – I just do it to keep good relationships. I don’t charge anything.

I:

So in order for you to get any value from using this system then it should reach your accuracy level.

M:

I mean if this system could remove the very few mistakes we do that would be very fine.

Thinking about it I reckon that the insurance companies, my insurance company, whenever someone makes a major mistake, which could be millions of DKK. They would be interested in such a system. If the system works without any problems. They should be asking for this system as if it was approved. Also insurance companies insuring horses - it's hundreds of thousands of horses being insured. They should based on clinical examination and evaluation. And if that could be more safe they would also be interested.

I:

So you say there's two kinds of insurance?

M:

One is insuring my mistake and the other one is insuring the horse.

I:

OK.

M:

If I don't see something and the horse breaks down - there are actually two different insurance companies that can be asked to pay.

I:

Do we have easy access to the X-rays where you have done a wrong diagnosis? It could be really interesting if there are any of these wrongly diagnosed X-rays. If you had access to them and if we could sort of put them into this machine to see what it says.

M:

Yes, I have one here.

I:

And also if you have others?

M:

We don’t have many…

I:

Ok – I think that is it for know. Do you have anything you would like to add?

M:

No, I don’t. You wanted me to give you the costs associated with our x-ray services?

I:

Yes that is correct.

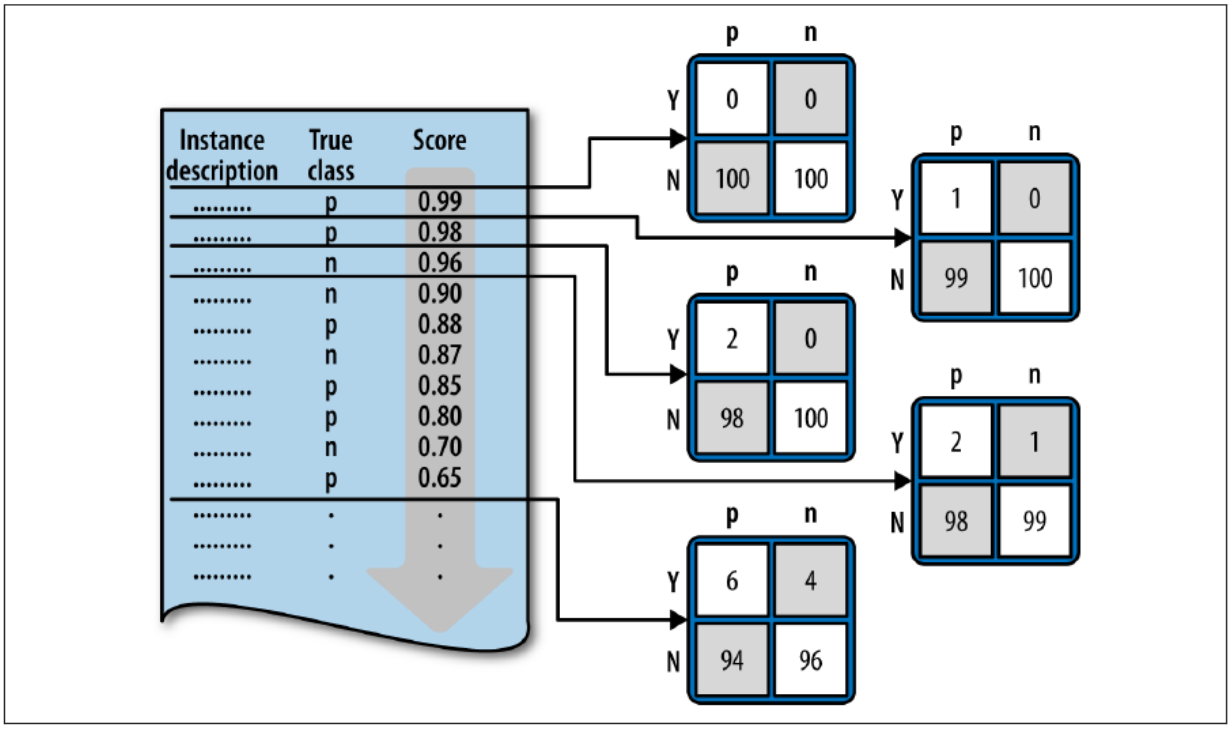
M:

Okay, you will have them here:

The price for 14 x-rays is 3250 DKK incl. VAT. This is the standard package especially in connection with trading horses. The other package for 2-6 x-rays which is 1750 DKK incl. moms. This package is only used if there is a very specific mistake some place on the horse.

If the client follow a surgery recommendation then the price is 11.400 DKK incl. VAT for Atroscopi of one hock and 13.000 incl. Vat for Atroscopi of two hocks.

9.3 Appendix 3: Classifier with a threshold

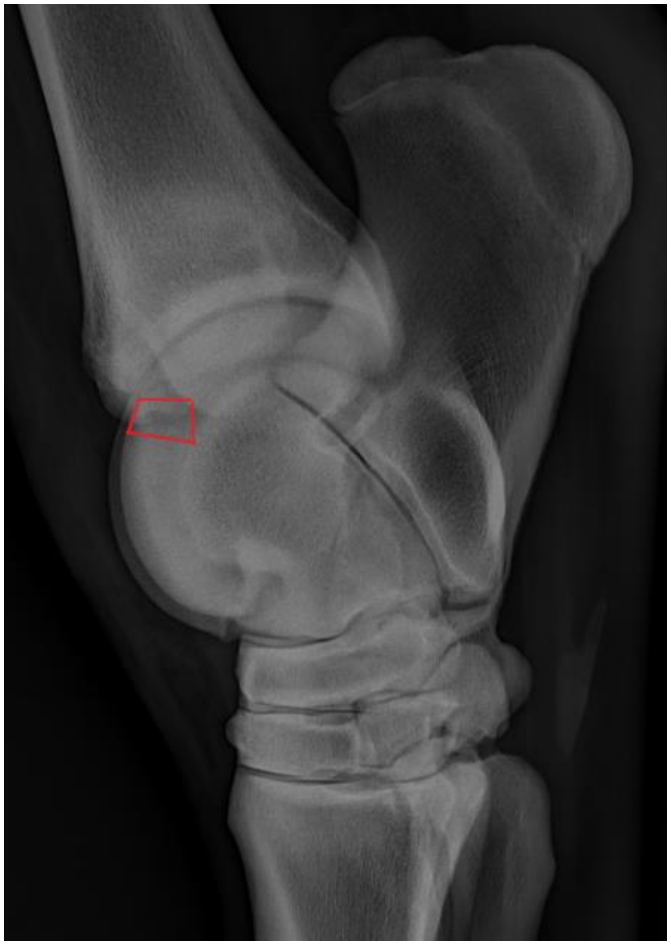


The figure shows the list of instances sorted by its scores. The confusion matrices changes accordingly when the threshold is lowered and produces different classifiers, i.e ranked classifiers. (Provost & Fawcett, 2013)

9.4 Appendix 4: X-rays and metadata

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	Id	Journal	Journal D	Seller	Buyer	Owner	Horse name	Race	Birth	Sex	Hocks OCD Diseas	VBlDor	VBlLat	VBldor	VBlpla	VBloth	HBlDor	HBlLat	HBldor	HBlplantarolate	HBlOth	Conclusion on diag
30	79		09-03-2015	Melberg			E. Caloo				4									Lille tilspidsning dorsoventralt os		2
31	80		21-12-2015				Mon Chevalier				2				Status efter OCD operation creiste intermed					Status efter OCD operation creiste intermed		
32	81		18-12-2015	Dansk Equine			French Match Ask				0											
33	82		12-01-2015	Hasselhøj Hest Aps	Camilla B. Ringstrøm		Winston Sic				4	Fragment Lateral Metalis (NOT OCD)										4 - no significance
34	83		03-12-2015				Champ				1									OCD Crista Intermedia		3
35	84																					
36	85		17-11-2015				Leonardo				3									Status efter OCD		2
37	86	37393	29-10-2015			Dolhe Bak	Slæbækgårds Nabiha				1				Small OCD							2
38	87	37227		Tove Larsen (Munkedal)			Johnsgårds Alanzo				1							OCD Intermediechar				3
39	88	37145	01-10-2015		Anne Sofie Jensen		Fire Fly - Edition - Matcha - AA				4				Ostrolgt dorsiprox i dist intarsalled							
40	89	36750	21-08-2015			Allan Nisse	Helium				1				OCD fragment crista int.							2
41	90					Søren Niss	O'Leary				4									Inragimotan crista int.		2
42	91	35512	15-04-2015	Jørn Borch			Chanel Line				3									Status efter OCD Operation på crista intermedia		
43	92	35552	20-04-2015	Christina Nissen			Elmegaard's Romario				1		OCD Intermedia					OCD Intermedia				3
44	93	35306	13-08-2014		Maria Seerup Sørenser		Skovgaards Leginezzen						Osteofyt med glideled									3
45	94	35215	17-08-2015	Winnie Jakobsen			Wendy				1							Lille fragment col tilmad crista int				
46	95	35081	23-02-2015	Niels Madsen			Treldegårds Sabia				1				Small OCD crist interdel + fragment distalt					Lille OCD crist. Intermed.		
47	96	35005	17-02-2015	Stutteri Hønnerup	Susanne Dam		Hønnerups Jadada				3				Status efter OCD							2
48	97	39724	03-06-2016	Munkedal Aps			Damgårdens Soldier				1							OCD Crista Intermedia				3
49	98	39635		Tanja Thorhauge			Dayton Dawns				4				Obs pro forandring. OCD medial madleous							3
50	99	39629	15-06-2016	Magne Olsen			Mojos Star				3		Status efter OCD operation intermediakan					Status efter OCD operation intermediakan				2
51	100	39483	03-06-2016	Leif Nørbak			Don Cruzado				2		Status efter OCD operation intermediakan					Status efter OCD operation intermediakan				2
52	101	39470		Tonny Nørgaard Bang	Susanne og Anders Hy		Raluc				1				OCD crista intermedia					OCD crista intermedia		4 - no significance
53	102	39176	22-03-2016				Dodo Laser(Lasen)				1							OCD Fragment				3
54	103	39056	27-04-2016				Black Beauty				1			OCD dist. Lateral throclea								2
55	104	39007			Else of Nis Ravn		Søndervangs Cool Cash				2									OC Mediale teochlea		2
56	105		28-10-2016				Agamemnon				4		Otcochondred modelig dist intertarsd joigt									4
57	106		20-10-2016				Donatar				1				OCD crista intermed					OCD crista intermed		
58	107		21-09-2016			Helle Thøg	Miss Lucy				4									Opklaring Crista Intermedia		2
59	108	29316	01-09-2016				Svalegaarden's Jack				1				Mindre fragment v. crista intermedia + osteofyt dorso-prot MT III, Adamstap							3
60	109	28225	05-04-2016	Claire Josephine Knudsen			Maggini				1							Small OCD with intermediakam (3*4mm)				2
61	110	26841	17-03-2016				Bjørnmose Savorage				4		FEJL DIAGNOSE fragment distalt for trochlea -									2
62	111		22-12-2016	Hesselhøj			Hesselhøj Shakaloveles				1		Fragment løs i karse..									
63	112	37503	29-10-2016				Don Ophino				1				Stor OCD defekt crista infermed					OCD crista infermed		3
64	113	38814	07-04-2016	Tozene Dalum			Danzell Overskovlund				4		Røntgent fragment distal for trochlea (muligt fragment)					Røntgent fragment distal for trochlea (muligt fragment)				2
65	114	38590	18-07-2016	Jørgen Helus			Pugemalles Feistos				1									OCD-fragment ved proxmale inte		3
66	115	38285	08-02-2016	Munkedal			Lærke Stensvang				1							OCD fragment vedsustemtacleom tadi				2
67	116	38256	03-02-2016				Felicia Guldbjerg				1									osteochrondrolis fragment i inten		3
68	117	38103	15-01-2016		Hanne Bach		Abkars Heart of Gold				2				OC dist throdia lat.							2
69	118	40530					Nørregårds Faith				1									Fragment v. insta intermedia		3
70	119	41816	29-12-2016	Susanne Lanen	Christian Thorvald		Holsegaards Le Charmeur				4		Alle røntgentest fragment dosalt for glideled									2
71	120	41222	29-10-2016				Famous Birkehøj				1				Stort OCD Crista Media							3
72	121		03-11-2016	Pernille Ravn	Tina Karlskov		Suteika				4		Osteochondral (MISTAKE) forandringer dist intarselled									3
73	122	22425	25-07-2011	Dennis Steffenauer	Isabella Mary Wadstrør		Møllegårds Bibi				1		OCD Crista Int					OCD Crista Int				2
74	123	22193	28-06-2011			Hans Lorer	Ronaldinio				1									OCD fragment hidrørende fra cri		3

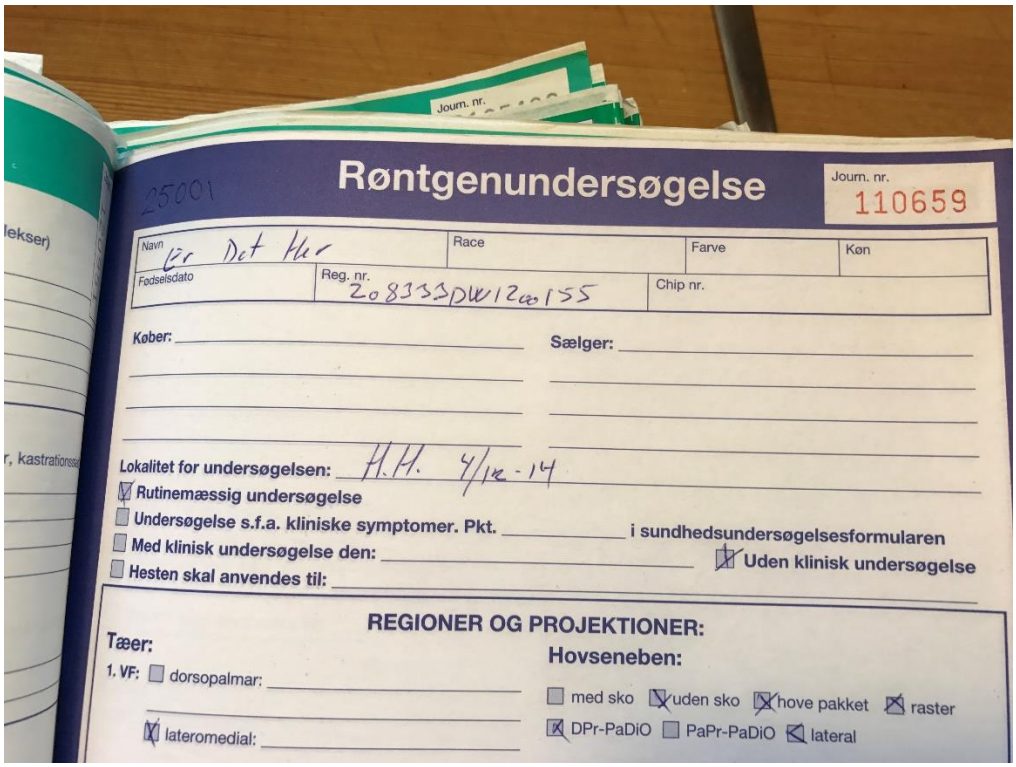
9.5 Appendix 5: X-rays with OCD (difficult cases)



9.6 Appendix 6: Cost-benefit calculations

Business case for x-ray examination:						
Assumptions						
Revenue streams				Costs		
X-ray packages (sunk costs):				**Salary:**		
14 x-rays	3250	DKK incl. VAT		One Nurse/technician:	27500	DKK per month
2-6 x-rays	1750	DKK incl. VAT		One vetenarian:	42500	DKK per month
Extra clinical examination		DKK incl. VAT		Assume 37 hour workweek		
Surgery:				**False negatives:**		
Atroscopi of one hock	11400	DKK incl. VAT		Insurance self-payment	5000	DKK per incident
Atroscopi of two hocks	13000	DKK incl. VAT		Loss of reputation and earnings	N/A	
				Insurance covers 10 mio DKK		
Probability that positives will:				**Time spent on OCD surgery:**		
Affect the use of the horse	50%			2 nurses/technicians giving the horse sedation and taking x-rays:	1,5	hours
And followed up with a recommendation to do surgery on the horse	37,50%			One vetenarian doing the surgery:	0,75	hours
And client will accept the recommendatio	33,75%					
Accumulated costs & Revenues						
Use case:	**Revenues**	**Costs**	**Balance**			
True positive	4117,5	772,8041	3344,696			
False positive	4117,5	772,8041	3344,696			
True negative	0	0	0			
False negative	0	-5000	-5000			

9.7 Appendix 7: Journals with description of disease



Røntgenundersøgelse

25001

Journ. nr. 110659

Navn: Er Det Her | Race | Farve | Køn

Fødselsdato | Reg. nr. 208333DW12ao155 | Chip nr.

Køber: _____ Sælger: _____

Lokalitet for undersøgelsen: H.H. 4/12-14

☒ Rutinemæssig undersøgelse

☐ Undersøgelse s.f.a. kliniske symptomer. Pkt. _____ i sundhedsundersøgelsesformularen

☐ Med klinisk undersøgelse den: _____ ☒ Uden klinisk undersøgelse

☐ Hesten skal anvendes til: _____

REGIONER OG PROJEKTIONER:

Tær:

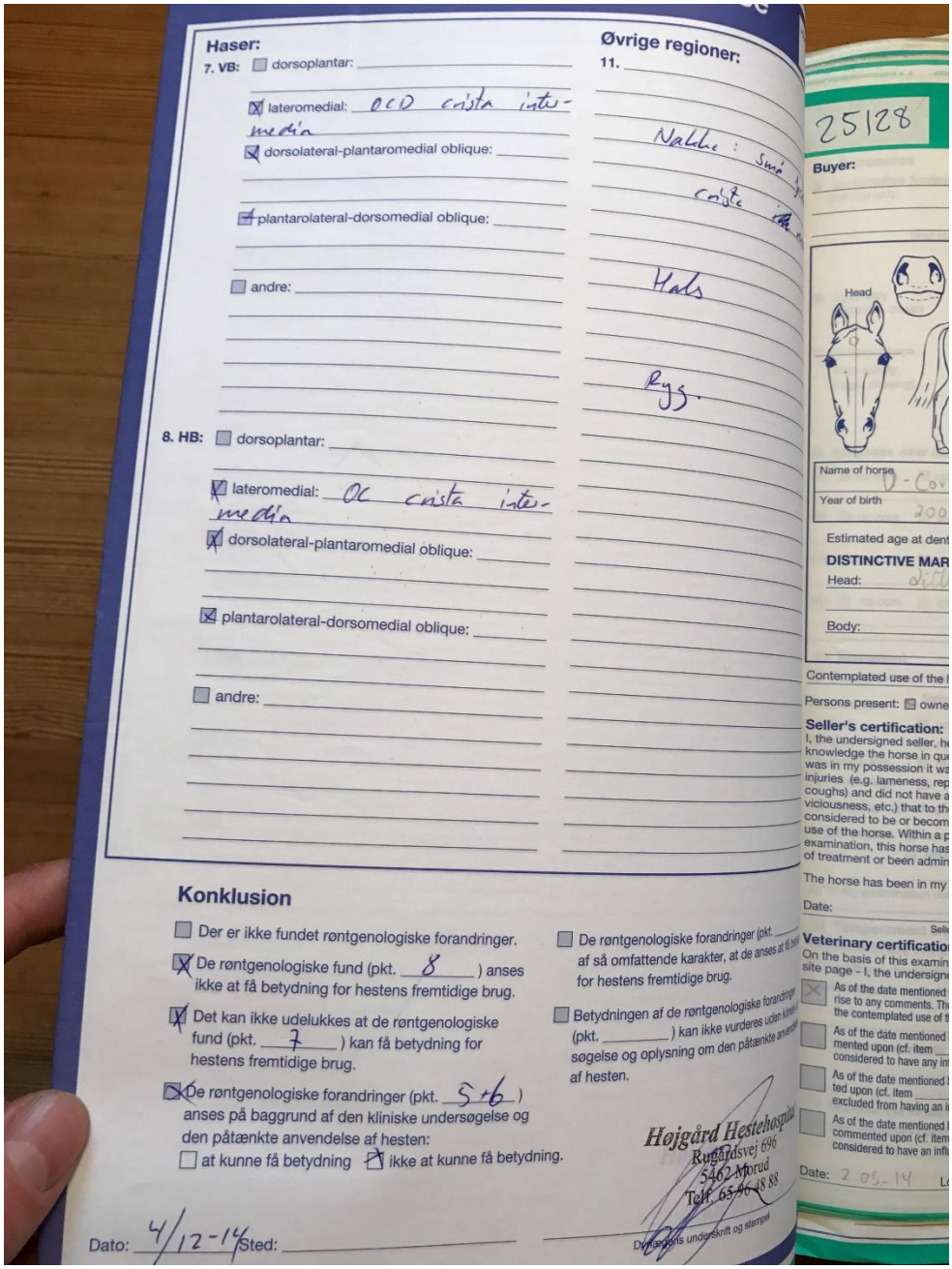
1. VF: ☐ dorsopalmar: _____

☒ lateromedial: _____

Hovseneben:

☐ med sko ☒ uden sko ☒ hove pakket ☒ raster

☒ DPr-PaDiO ☐ PaPr-PaDiO ☒ lateral



Haser:

7. VB: ☐ dorsoplantar:

☒ lateromedial: OCD crista intermedia

☒ dorsolateral-plantaromedial oblique:

☒ plantarolateral-dorsomedial oblique:

☐ andre:

8. HB: ☐ dorsoplantar:

☒ lateromedial: OC crista intermedia

☒ dorsolateral-plantaromedial oblique:

☒ plantarolateral-dorsomedial oblique:

☐ andre:

Øvrige regioner:

11. Nakke: Små [illegible] crista [illegible]

Hals

Rygs.

Konklusion

☐ Der er ikke fundet røntgenologiske forandringer.

☒ De røntgenologiske fund (pkt. 8) anses ikke at få betydning for hestens fremtidige brug.

☒ Det kan ikke udelukkes at de røntgenologiske fund (pkt. 7) kan få betydning for hestens fremtidige brug.

☒ De røntgenologiske forandringer (pkt. 5+6) anses på baggrund af den kliniske undersøgelse og den påtænkte anvendelse af hesten:
☐ at kunne få betydning ☒ ikke at kunne få betydning.

☐ De røntgenologiske forandringer (pkt. ____) er så omfattende karakter, at de anses at få betydning for hestens fremtidige brug.

☐ Betydningen af de røntgenologiske forandringer (pkt. ____) kan ikke vurderes uden klinisk søgelse og oplysning om den påtænkte anvendelse af hesten.

Højgård Hestehospital
Rugårdsvej 696
5462 Morud
Tlf. 65 96 48 88

Dato: 4/12-14 Sted:

Dyrlægens underskrift og stempel

25128

Buyer:

Head

Name of horse

Year of birth

Estimated age at dent

DISTINCTIVE MAR

Head:

Body:

Contemplated use of the

Persons present: ☐ owne

Seller's certification:

I, the undersigned seller, h knowledge the horse in qu was in my possession it wa injuries (e.g. lameness, rep coughs) and did not have a viciousness, etc.) that to th considered to be or becom use of the horse. Within a p examination, this horse has of treatment or been admin

The horse has been in my

Date:

Veterinary certificatio

On the basis of this examin site page - I, the undersign

☒ As of the date mentioned rise to any comments. Th the contemplated use of t

☐ As of the date mentioned mented upon (cf. item considered to have any in

☐ As of the date mentioned ted upon (cf. item excluded from having an i

☐ As of the date mentioned commented upon (cf. item considered to have an infl

Date: 2.05.14

This journal specificly state that OCD is located on the left hock (VB) crista intermedia and can be seen on the LM x-ray projection (lateromedial).

9.8 Appendix 8: X-ray examinations

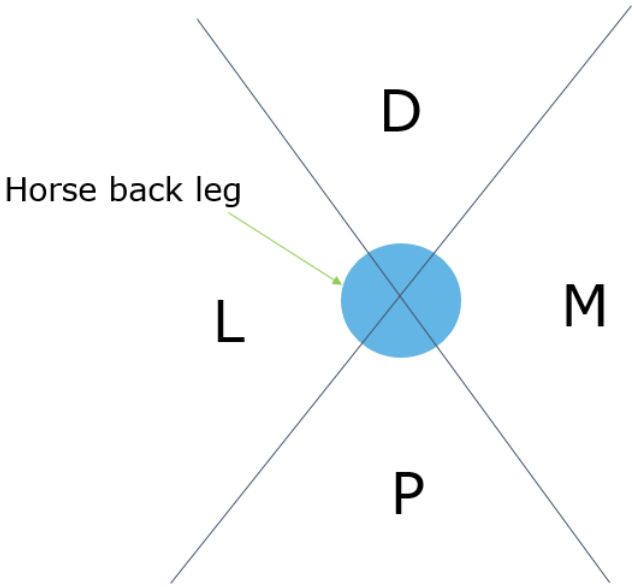


Figure: X-ray projections on the horse's hock

Description of x-ray angles:

D = Forward

P = Behind

L = Taken from the "outside" side of the horse

M = Taken from the "inside" side of the horse

X-ray projections

LM – X-ray projected from the side of the leg

DMPLO – X-ray projected from the oblique and back

DP – X-ray projected from the front

LM DLPMO DP

X-rays are maintained and collected in two databases. One of them is hosted by a Swiss company named easyVET. From this database it is possible to retrieve x-rays.

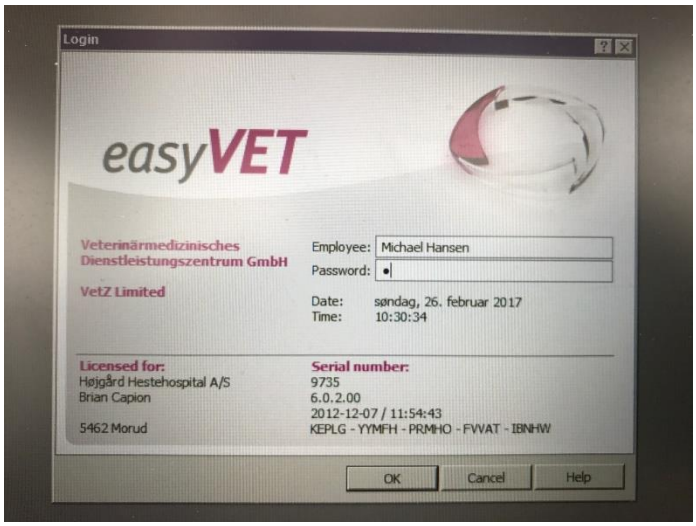


Image: Login to cloud database in order to retrieve x-rays.

9.9 Appendix 9: E-mail correspondence about x-rays and lack of consistent x-ray evaluation

Fra: Frank Kastbjerg
Sendt: 5. december 2016 08:27
Til: Peter Busk
Cc:

[redacted]

Emne: SV: Kvalitetssikring røntgen

Godmorgen [redacted]

Mange tak for dit forslag, som passer glimrende til den meget vigtige opgave, der ligger i, at vi i DH skal være klart bedre til vidensdeling/sparring, end vi har været i en lang periode. Jeg er helt enig i de fordele ved et røntgensamarbejde, som du oplister. Behovet for denne sparring er nok størst for DH-kollegerne i enmandspraksis, men jeg ved ikke, om der er signifikant forskel på "præstationerne" i enmands- vs. flermandspraksis. Vi har på både bestyrelses- og stormøder talt om bl.a. røntgen, som et oplagt sted at bruge hinanden noget mere – fx ved at danne ERFA-grupper for røntgeninteresserede. Dit forslag er meget konkret og det bedste bud, jeg p.t. har set har set desangående. Skal vi ikke bare se at komme i gang?

MVH

[redacted]

[redacted]

Fagdyrlæge i sygdomme hos heste

& certificeret hestekiropraktor

Hald Ege Hestepraksis

Tlf. 8663 8611 (træffes bedst kl. 7 – 9)

www.haldegehestepraksis.dk

Fra: Peter Busk [redacted]
Sendt: 3. december 2016 15:40
Til: Frank Kastbjerg Pedersen
Emne: Kvalitetssikring røntgen

Hej [redacted]

Jeg har spekuleret over en ide, som jeg vil dele med dig som fromand for DH, ikke mindst affødt af de mange sager jeg ser som skønsmand.

Mange af disse sager opstår som følge af forskellige vurderinger af røntgenbilleder og røntgenfund, nogen gange næsten alene baseret på forskelle i terminologi snarere end kvalitative forskelle i fund og vurderinger. Om end jeg udmærket ved, at vi aldrig opnår 100 % ensartethed i beskrivelser og vurderinger, forekommer det mig, at en højere grad af ensartethed kun være med til at begrænse antallet af meget dyre og ubehagelige ansvarssager.

Da jeg for snart mange år siden havde kontakt til kvægpraksis, begyndte alle at dyrke deres egene mælkeprøver og stille mikrobiologiske diagnoser i deres hjemmelaboratorium, og det var langt fra alle, der var lige gode til eller lige omhyggelige med det arbejde, så ret hurtigt viste det sig, at validiteten af de stillede diagnoser var meget svingende, ligesom vi ser det med hensyn til radiologiske undersøgelser og vurderinger hos heste.

For at støtte dyrlægeren blev der lavet et koncept, kaldet Ringtesten. Dette bestod i korthed i, at mastitislaboratoriet rendyrkede nogle bakterier, sendte disse ud til de deltagende dyrlæger, bad dem dyrke, karakterisere og diagnosticere disse test-bakterier. Efter en måned blev identiteten af de fremsendte test-bakterier offentliggjort og man kunne så sammenligene sine resultater se hvordan man selv havde performed. Senere blev dette udbygget, så man indsendte sine dyrkningsresultater på test-bakterierne, og mastitislaboratoriet foretog så en "ranking" af de deltagende dyrlægers præstationer, og offentliggjorde denne til alle deltagende dyrlæger, ikke generelt. Det sidste højnede interessen og omhuen i diagnostikken, husker jeg.

Jeg har tænkt på, om noget lignende ikke var muligt på røntgenområdet: DH nedsætter et udvalg på f.eks tre medlemmer, og dette udvalg udvælger røntgenoptagelser, såvel blandt egne som blandt billeder indsendt fra kolleger til udvalget, som dette udvalg beskriver, vurderer og diagnosticerer som udvalgets medlemmer mener de bør beskrives.F.eks. 4 gange om året udsender DH så et testsæt af billeder til interesserede DH medlemmer, som så har en uge til at bedømme og vurdere disse billeder, indsende deres vurdering til udvalget, der herefter foretager en ranking og offentliggør deres vurdering, så alle kan lære af det.

På den måde kunne vi opnå: (over tid)

Bedre og mere ensartede billeder kvalitativt, både teknisk og projektions/vinklingsmæssigt.

Mere ensartede beskrivelser af billederne (meget vigtigt, tror jeg)

Højere grad af reproducerbarhed i bedømmelserne.

Jeg mener ikke det behøver at koste en million eller være særligt besværligt, for i disse digitale tider kan udvalgets arbejde med at udvælge, bedømme og udsende røntgenbilleder, samt indsamle besvarelser og ranke disse jo ske over nettet, så de behøver ikke bruge tid på at mødes og røntgenbilleder ser vi jo alle flere gange dagligt.

Jeg tror på, at dette kunne blive et kvalitetsløft, der kunne profilere DH medlemmer overfor andre dyrlæger og forebygge, at DH dyrlæger bliver internt uenige om vurderingen af kbalitetetn af eller evt. fund på hinandens røntgenbilleder.

Jeg har kun sendt til dig i første omgang, men du er naturligvis velkommen til at delagtiggøre hvem du vil i ideen.

Venlig hilsen

[redacted]

Fagdyrlæge vedr. sygdomme hos hest

Skyggehusvej 2, 6900 Skjern

Tlf.: [redacted]

Web.: www.hshestepraksis.dk

Mail.: [redacted]