# Breast Cancer Detection in Mammograms using Deep Learning Techniques

Plan & Context Survey

University of St Andrews - School of Computer Science

Supervisor: Dr David Harris-Birtill

Adam Jaamour

15th June, 2020

# 1 Thesis Outline

**Introduction** Sets the tone for the entire dissertation by presenting the subject and the problem being tackled, while also laying out the plan for the rest of the report. Subsections include the motivation & problem description and the project aims.

**Context Survey** Explores the literature surrounding breast cancer detection using deep learning techniques, including state of the art techniques recently used, by covering the background of both deep learning techniques and their applications to the problem of breast cancer detection. See Section 3 for this part's subsections.

**Requirements** Establishes and prioritises the functional and non-functional properties of the code.

**Ethics** Considers the ethical issues considered for this project.

**Design** Details the general structure of the system, offering an analysis of different parts of the deep learning pipeline without going into technical code-related or mathematics-related details. For example, this includes the choice of input method, data pre-processing steps, output format and programming languages and libraries.

**Implementation** Covers the specifics followed when implementing the deep learning pipeline, explaining the mathematical and software-related reasoning behind each decision (this will include equations and code snippets). Additionally, covers any testing done to validate that the system works as expected.

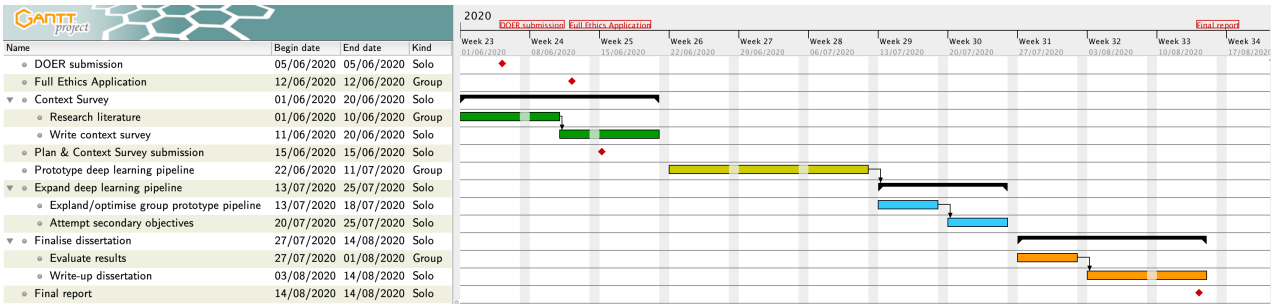
**Evaluation** Compare the final results with the basic pipeline's results developed in common, the final results achieved by the two other group members and the results from papers researching breast cancer detection identified in the context survey.

**Conclusions** This sections will summarise the project as a whole, from the initial objective to the results obtained. Further discussions will be included to objectively assess what could have improved, as well as any potential future work.

# 2 Dissertation Time Plan

Figure 1 showcases the estimated timeline for this dissertation, separating activities into four parts corresponding to the four main objectives established in the DOER document. These main activities are broken down into subactivities. Milestones are included in the Gantt chart as well. The chart was created using GanttProject[1].

[1]GanttProject: `https://www.ganttproject.biz/`




| Name | Begin date | End date | Kind |
|---|---|---|---|
| DOER submission | 05/06/2020 | 05/06/2020 | Solo |
| Full Ethics Application | 12/06/2020 | 12/06/2020 | Group |
| Context Survey | 01/06/2020 | 20/06/2020 | Solo |
| Research literature | 01/06/2020 | 10/06/2020 | Group |
| Write context survey | 11/06/2020 | 20/06/2020 | Solo |
| Plan & Context Survey submission | 15/06/2020 | 15/06/2020 | Solo |
| Prototype deep learning pipeline | 22/06/2020 | 11/07/2020 | Group |
| Expand deep learning pipeline | 13/07/2020 | 25/07/2020 | Solo |
| Expland/optimise group prototype pipeline | 13/07/2020 | 18/07/2020 | Solo |
| Attempt secondary objectives | 20/07/2020 | 25/07/2020 | Solo |
| Finalise dissertation | 27/07/2020 | 14/08/2020 | Solo |
| Evaluate results | 27/07/2020 | 01/08/2020 | Group |
| Write-up dissertation | 03/08/2020 | 14/08/2020 | Solo |
| Final report | 14/08/2020 | 14/08/2020 | Solo |

Figure 1: Gantt Chart ranging from 01/06/2020 to 14/08/2020.

# 3 Draft Context Survey

*Note: According to the dissertation time plan expressed in Section 2, the context survey will be finished by the end of week 3 (19/06/2020). Consequently, this context survey is still in a draft format. The majority of time has been invested in researching and taking notes on state-of-the-art papers (over 20 papers so far) and coming up with a detailed and logical plan. Sections with bullet points still need to be properly written while further information/references will be added to the existing parts.*

## 3.1 Breast Cancer Detection

### 3.1.1 Early Breast Cancer Detection Systems

The detection of breast cancer using mammograms, and any form of cancer using medical imagery, relies on the conventional diagnoses of expert radiologists [1]. These diagnoses rest on the correct interpretation of the mammograms, which may be subject to errors due to the difficulty of correctly interpreting them [2]. Indeed, mammograms are 2D images of 3D breasts that correspond to the superposition of breast tissue, which increases the difficulty for a radiologist to correctly analyse patterns as masses often naturally form due to this superposition [2].

To assist radiologists in their interpretations of mammograms, Computer-Assisted Detection/Diagnosis (CAD) software have been employed since the 1970s. However, pre-1990s CAD systems were very primitive and did not offer much more knowledge than the expert radiologists'. These unsophisticated "expert" systems consisted of manually processing and modelling pixels to construct rule-based systems that mainly used *if-else-then* statements [3].

### 3.1.2 Evolution towards Supervised Machine Learning-based Systems

Supervised machine learning models have been applied to the problem of breast cancer detection since XXXX. However, these models could not accurately operate on purely raw data such as the full-sized mammogram images. Indeed, of the many machine learning models tested against the task of breast cancer detection such as k-Nearest Neighbour (kNN), Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), Naive Bayes (NB) and Multi-Layer Perceptrons (MLP), these all require relevant bits of information to be extracted first in order to solve the given task. These important pieces of information correspond to

features, and need to be extracted by humans before being fed to the aforementioned models for training. These features can range from colours, edges and corners to shapes and textures.

The next logical step in the evolution of breast cancer detection systems is for the selected model to learn these features on its own directly from the data rather than being fed hand-extracted features [4]. Deep learning models, which corresponds to neural networks with hundreds of hidden layers, are based on this concept. However, these models have not been successfully implemented until recent years as they require powerful computers (usually equipped with Graphical Processing Units) to be efficiently trained.

## 3.2 Machine Learning models and breast cancer detection applications

### 3.2.1 Machine Learning tasks

#### Supervised & unsupervised learning algorithms

Machine learning algorithms come in many flavours, ranging from supervised and unsupervised learning to reinforcement learning. The two main types correspond to supervised and unsupervised learning [5]. In supervised learning, a label $y$ is predicted based on input features $x$ by using a training dataset where the labels are known. During supervised training, the models' optimal parameters are determined to best predict the label.

On the other hand in unsupervised learning, the data is unlabelled, requiring the machine learning model to find patterns in the data by using known methods such as clustering and Principal Component Analysis (PCA) [3].

- Different types of learning model: supervised VS unsupervised (briefly mention semi-supervised and reinforcement learning). *In this thesis, supervised learning used with labelled mammograms.*

#### Machine learning tasks applied to medical imagery analysis

- Classification
- Detection
- Segmentation
- Out-of-scope Machine Learning tasks in medical imagery analysis: registration, CBIR, image generation, image enhancement.

### 3.2.2 Machine Learning models

- k-Nearest Neighbours
- Naive Bayes

- Support Vector Machines
- Decision Trees
- Random Forest
- Artificial Neural Networks
  - Multi-Layer Perceptrons (shallow artificial neural networks)
  - Deep neural networks

For each, mention pros and cons based on existing papers and compare different approaches. Naturally lead towards deep neural networks (e.g. CNNs) for the final section.

## 3.3 Deep Learning techniques & Convolutional Neural Networks

### 3.3.1 Convolution Neural Networks

- Deep Neural Networks
- Convolutional Neural Networks

Explore the techniques used in deep learning (techniques, databases, processing, libraries, output metrics, etc.)
Explore the deep learning model that will be explored by each dissertation

### 3.3.2 Rise of Deep Learning in Medical Imagery Analysis

- Evolution of CNNs (first CNN used, first CNN applied to medical imagery analysis, famous architectures e.g. LeNet, AlexNet, GoogLeNet, ResNet, etc.)
- Hardware advances (GPUs)
- Software advanced (libraries)

# References


[1] Alireza Osareh and Bita Shadgar. Machine learning techniques to diagnose breast cancer. In *2010 5th International Symposium on Health Informatics and Bioinformatics, HIBIT 2010*, pages 114–120, 2010.

[2] Matthias Elter and Alexander Horsch. CADx of mammographic masses and clustered microcalcifications: A review. *Medical Physics*, 36(6):2052–2068, jun 2009.

[3] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis, dec 2017.

[4] Adam Yala, Constance Lehman, Tal Schuster, Tally Portnoi, and Regina Barzilay. A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction. *Radiology*, 292(1):60–66, jul 2019.

[5] Aurelien Géron. *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow*. O'Reilly Media, 2nd edition, 2019.