A model for the detection of breast cancer using machine learning and thermal images in a mobile environment

by

Nicholas Lennox

A model for the detection of breast cancer using machine learning and thermal images in a mobile environment

by

Nicholas Lennox

Dissertation

submitted in fulfilment
of the requirements
for the degree

Master of Information Technology

in the

Faculty of Engineering, the Built Environment and Information Technology

of the

Nelson Mandela University

Supervisor: Dr. B. Haskins

April 2020

Declaration

I, Nicholas Lennox 212231081, hereby declare that:

- The work in this dissertation is my own work.
- All sources used or referred to have been documented and recognised.
- This dissertation has not previously been submitted in full or partial fulfilment of the requirements for an equivalent or higher qualification at any other recognised educational institute.

Nicholas Lennox

Abstract

Breast cancer is the most common cancer amongst women and one of the deadliest. Various modalities exist which image the breasts, all with a focus on early detection; thermography is one such method. It is a non-invasive test, which is safe and can be used for a wide variety of breast densities. It functions by analysing thermal patterns captured via an infrared camera of the surface of the breast. Advances in infrared and mobile technology enable this modality to be mobile based; allowing a high degree of portability at a lower cost. Furthermore, as technology has improved, machine learning has played a larger role in medical practices by offering unbiased, consistent, and timely second opinions. Machine learning algorithms are able to classify medical images automatically if offered in the correct format.

This study aims to provide a model, which integrates breast cancer detection, thermal imaging, machine learning, and mobile technology. The conceptual model is theorised from three literature studies regarding: identifiable aspects of breast cancer through thermal imaging, the mobile ecosystem, and classification using machine learning algorithms. The model is implemented and evaluated using an experiment designed to classify automatically thermal breast images of the same quality that mobile attachable thermal cameras are able to capture. The experiment contrasts various combinations of segmentation methods, extracted features, and classification algorithms. Promising results were shown in the experiment with a high degree of accuracy obtained. The successful results obtained from the experimentation process validates the feasibility of the model.

Acknowledgements

Foremost, I would like to express my sincere gratitude to my supervisor **Dr. Bertram Haskins** for your continuous support during my Master's study and your willingness to provide feedback at any moment. Furthermore, I would like to express my gratitude for your patience, motivation, enthusiasm, and immense knowledge. Your guidance assisted me throughout the research and writing of this dissertation. I could not have imagined having a better supervisor for my Master's study.

I would like to thank my parents, **Anita** and **Edward**. You raised me to always push myself; you have supported me through every step of my journey and have always been there to rely on. I could not ask for better parents than the ones I have been given.

To my wife, **Melene**, I could not have completed this without your unending support and patience. You were my rock during those long, stressful nights and without you, I would not have been able to pursue this journey.

Furthermore, I would like to thank and acknowledge the Nelson Mandela Metropolitan University's Post Graduate Research Scholarship (PGRS) for its financial support.

Finally, I would like to thank **Ricky Woods** for providing her proofreading services. Your quick turnaround time and flexible work-style allowed me to continue working on chapters while having others proof read.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction

Amongst women, breast cancer is the most commonly diagnosed cancer and one of the deadliest. The chance of survival is strongly linked to early detection and it is vital that efforts be made in order to ensure better, more widespread early detection. Low-resource areas often do not have the medical infrastructure in place to enable early detection programs. The advancement of mobile technology and attachable thermal cameras affords the ability to implement cancer detection programs in low-resource areas owing to its highly portable and cost effectiveness. Machine learning has played a large role in medical diagnosis for many years; with advancements in technology enabling machine learning to have an even larger role. Advanced machine learning algorithms offer medical professionals second opinions, which are fast, objective, and consistent. This study is an integration of thermal breast imaging, breast cancer detection, machine learning, and mobile technology through a proposed model.

1.2 Breast cancer and thermography

Cancer is a collective term for a set of related diseases, all associated with irregular cell growth. This leads to tumours growing uncontrollably and spreading throughout the body through a process called metastasis (Kandlikar et al., 2017). According to Tortora and Derrickson (2017, p. 100), the most common type of cancers are carcinomas, which are cancers arising from the cells. The most commonly found carcinomas are from the skin, lungs, breasts, pancreas, and other organs and glands.

A healthy breast consists of glands connected to the surface by skin ducts, surrounded by connective tissue with blood vessels, lymph nodes, lymph channels, and embedded nerves

(Kandlikar et al., 2017). Breast cancer commonly begins either in the lobules or in the milk carrying ducts of the breast, termed lobular carcinoma and ductile carcinoma respectively (Kandlikar et al., 2017). Cancer needs nutrients in order to proliferate, it does this by promoting angiogenesis; the formation of new blood vessels (Y. Ng, Ung, Ng, & Sim, 2001). The body responds to cancer in various ways, two of which are: sending white blood cells to fight the cancer, and making the surrounding area inflamed (Anbar, Brown, Milescu, Babalola, & Gentner, 2000). The white blood cells release nitric oxide, which, coupled with the aggressiveness of cancer, dilates blood vessels, resulting in a heating of the surrounding tissue (Kakileti, Manjunath, Madhu, & Ramprakash, 2017).

Amongst women, breast cancer is the most commonly diagnosed cancer and one of the deadliest (Yassin, Omran, El Houby, & Allam, 2018). An estimated one in eight women will, throughout their lifetime, be diagnosed with breast cancer (Bray et al., 2018). Breast cancer is staged from 1-4, depending on the size of the tumour and how much it has spread. The stage of detection has a large impact on the survival chances, with 100% chance at stage 1 and a 22% chance at stage 4 (Saslow et al., 2007). It is critical, therefore, that breast cancer be detected as early as possible.

There are many imaging techniques used when it comes to breast cancer detection. The gold standard of these is mammography with ultrasound and magnetic resonance imaging playing a supportive role (Kandlikar et al., 2017). Tabar et al. (1985) found that mammography reduced mortality from breast cancer by 34%. This reduction comes at the cost of incredible patient discomfort as the patient's breasts are clamped between two cold plates with a considerable force (Collins et al., 2010). The need for an early detection method, which is non-invasive, has led to the discovery of new modalities. One such modality is thermography.

Thermography is a straightforward, non-invasive, non-contact skin surface temperature screening method that is economic, quick and does not inflict any pain on the patient (E. Ng, 2009). According to Head and Elliott (2002), a thermogram shows the normal and abnormal functioning of the body's vascular system, sympathetic and sensorial nervous system, and inflammatory processes.

Breast thermography is being researched and applied all around the world (Etehadtavakol, Ng, Chandran, & Rabbani, 2013). Owing to abnormality being linked to a temperature difference between left and right breasts, asymmetry analysis is the most common method used for detecting breast cancer using breast thermogram images (Borchartt et al., 2013). The left and right breast images are examined and graded according to a categorisation system; the Marseille system for example (Gautherie, 1985). The improvement of infrared

cameras, the introduction of standards, and the use of machine learning have greatly improved the consistency and accuracy of breast thermography (Pavithra, Ravichandran, Sekar, & Manikandan, 2018).

1.3 Machine learning in breast thermography

Machine learning is defined as the ability to learn and improve from experience without being explicitly programmed (Samuel, 1959). Machine learning is applied to many problem areas where the data is too complex for humans or the task is too time consuming (Shalev-Shwartz & Ben-David, 2014, p. 22).

One such task is medical diagnosis, where computers are used in order to assist physicians with their decisions. As far back as 1960, computers were proposed to be of assistance with medical diagnosis, under the term Computer-Aided Diagnosis (CAD) (Moghbel & Mashohor, 2013). One of the first CAD systems for thermography, called the Computerized Breast Thermographic Interpreter (CBTI), developed by Negin, Ziskin, Piner, and Lapayowker (1977), had an accuracy rating of 79%, which was very promising for its time. Since then, improvements to software and image capturing techniques have improved the accuracy of CAD systems to where they are now on par or even better than other methods of diagnosis or screening (Yassin et al., 2018).

Computers and machine learning are mainly applied to breast thermography by way of classification. The process of classifying a breast thermogram is explained by Pavithra et al. (2018) and summarised here as:

1. Segment the region of interest (ROI) from the breast thermogram. This ROI comprises the breast tissue only.

2. Quantify the ROI of each breast separately into features, for asymmetry analysis. This is normally done by texture analysis (Schaefer, Závišek, & Nakashima, 2009).

3. Train classification algorithms on the extracted features and make predictions on unseen images. Commonly found classifiers include: k-Nearest Neighbour (k-NN), naïve Bayes (NB), Support Vector Machine (SVM), Adaptive Boosting (AdaBoost), and Artificial Neural Network (ANN) (Vardasca et al., 2018).

The process of computer-aided diagnosis simply requires images, captured in the correct format, to be processed by machine learning algorithms. This means that any device can be used to capture the images as long as it is in a format and a quality that the machine learning algorithms expect. Advances in thermal cameras present the possibility of using

mobile devices to diagnose breast cancer (Aubakir, Nurimbetov, Tursynbek, & Varol, 2016; Min et al., 2017). These attachable thermal cameras are a highly portable and affordable offering, whereas the use of laptops and traditional cameras have a higher cost and are less easily concealed making them a risk in high theft areas.

1.4 Mobile healthcare

With the increasing access to mobile phones worldwide and more affordable access to wireless technologies in developing countries, mobile health (mHealth) has risen (Clifford & Clifton, 2012). mHealth is the use of mobile phone technology for health-related purposes (Leon, Schneider, & Daviaud, 2012). mHealth can be carried out with the use of smartphones, devices capable of more advanced tasks than making calls and sending texts (Baig, GholamHosseini, & Connolly, 2015). In Southern Africa, mHealth can improve and reduce the cost of patient monitoring, medication adherence, and healthcare worker communication, especially in rural areas by helping standardise, store, analyse, and share patient information (Betjeman, Soghoian, & Foran, 2013; Kahn, Yang, & Kahn, 2010; Bloch, 2010; Ranck, 2011). It has shown great promise in early studies due to high mobile penetration and the ability to provide access to remote areas with the improvement of mobile communication bandwidth (Betjeman et al., 2013).

As handheld thermal cameras have improved, efforts have been made to increase portability. FLIR has led the way in this by manufacturing thermal cameras that can attach directly to a smartphone. Their newest model FLIR ONE has been used to evaluate the viability of mobile based thermography. Some studies have been conducted to evaluate the viability of FLIR ONE for inflammation detection (Kanazawa et al., 2016), monitoring burn wounds (Jaspers, Carrière, Meij-de Vries, Klaessens, & van Zuijlen, 2017), lower extremity ischemia (Lin & Saines, 2017), vital sign monitoring (Aubakir et al., 2016), and diabetic foot monitoring (Boguski et al., 2019). These studies all found the FLIR ONE to be inside the acceptable threshold for capturing medical thermograms. A study by Kockara, Halic, Hudson, Loney, and Crawford (2014) looked at using a smartphone to identify cancerous skin lesions. This study was conducted before the FLIR ONE was on the market, but showed that the device used was more than capable of performing the desired task. The use of smartphone based thermography is in its infancy and shows promise but more research needs to be done for it to become widely accepted.

1.5 Problem area

Breast cancer survival depends significantly on the stage at which is it diagnosed. Early detection is the best method of reducing breast cancer related deaths. For urban living this problem is largely solved by modern techniques, but rural areas suffer. For example, in South Africa, rural areas lack proper access to medical resources, resulting in later stage diagnosis on average (Vorobiof, Sitas, & Vorobiof, 2001; Pillay, 2002; Somdyala, Bradshaw, Gelderblom, & Parkin, 2010). Attachable thermal cameras are small, offer great portability, are relatively cheap, and are easily concealed making them suitable for rural environments (Kakileti et al., 2017). The above mentioned studies have looked at the quality of attachable thermal cameras compared to traditional thermal cameras and provided no insight into how to integrate mobile technology with thermal imaging to provide solutions.

1.6 Problem statement

The problem statement addressed by this study is as follows:

There is no widely accepted process to facilitate the integration of breast thermography and machine learning techniques with a mobile platform.

1.7 Thesis statement

The study is guided by the following thesis statement:

A model for the automatic classification of thermal breast images using machine learning in a mobile environment would ease the development of a mobile application, which integrates these technologies.

1.8 Primary objective

To address the problem statement, the primary objective for this study is to:

Develop a model for the application of machine learning for the automated classification of thermal breast images in a mobile environment.

1.9 Secondary objectives

To support the primary objective, the following secondary objectives need to be met:

1. Identify which aspects of breast cancer are detectable by thermal imaging using mobile devices.

2. Contrast segmentation methods for the extraction of the breast region from thermal images.

3. Contrast machine learning techniques for the classification of thermal breast images.

1.10 Scope and delineation

No mobile-phone based prototype was built as the techniques for performing the required processing and classification would function in the exact same manner on the mobile device as in a desktop environment owing to the machine learning algorithms being trained before being used on the device. In terms of the full patient procedure, the model is only designed for classification and does not contain any steps to be taken after a result has been presented. The experimentation undertaken is done in a Python 3 environment using common libraries found in machine learning solutions. The images used for analysis are taken from publicly available data sets where standards were enforced to ensure accurately captured information.

1.11 Research process

The approach this study undertakes is a positivist one, which (Olivier, 2009) explains as a verifiable approach to research. This approach is objective which reduces the impact of the researchers interpretation. The following research methods are used in this study; literature review, argumentation, and experimentation. Figure 1.1 shows the research process outlined for this study.

1.11.1 Literature review

A literature review is an iterative process where one searches for information pertaining to one's topic area, works through the information, and then decides on what to keep or discard (Olivier, 2009). The literature reviews in this study cover various aspects of breast cancer, thermal imaging and its application to breast cancer detection with mobile devices, breast image segmentation methods, and machine learning classification. Literature reviews will be conducted, in part, to meet secondary objective one, secondary objective two and secondary objective three.

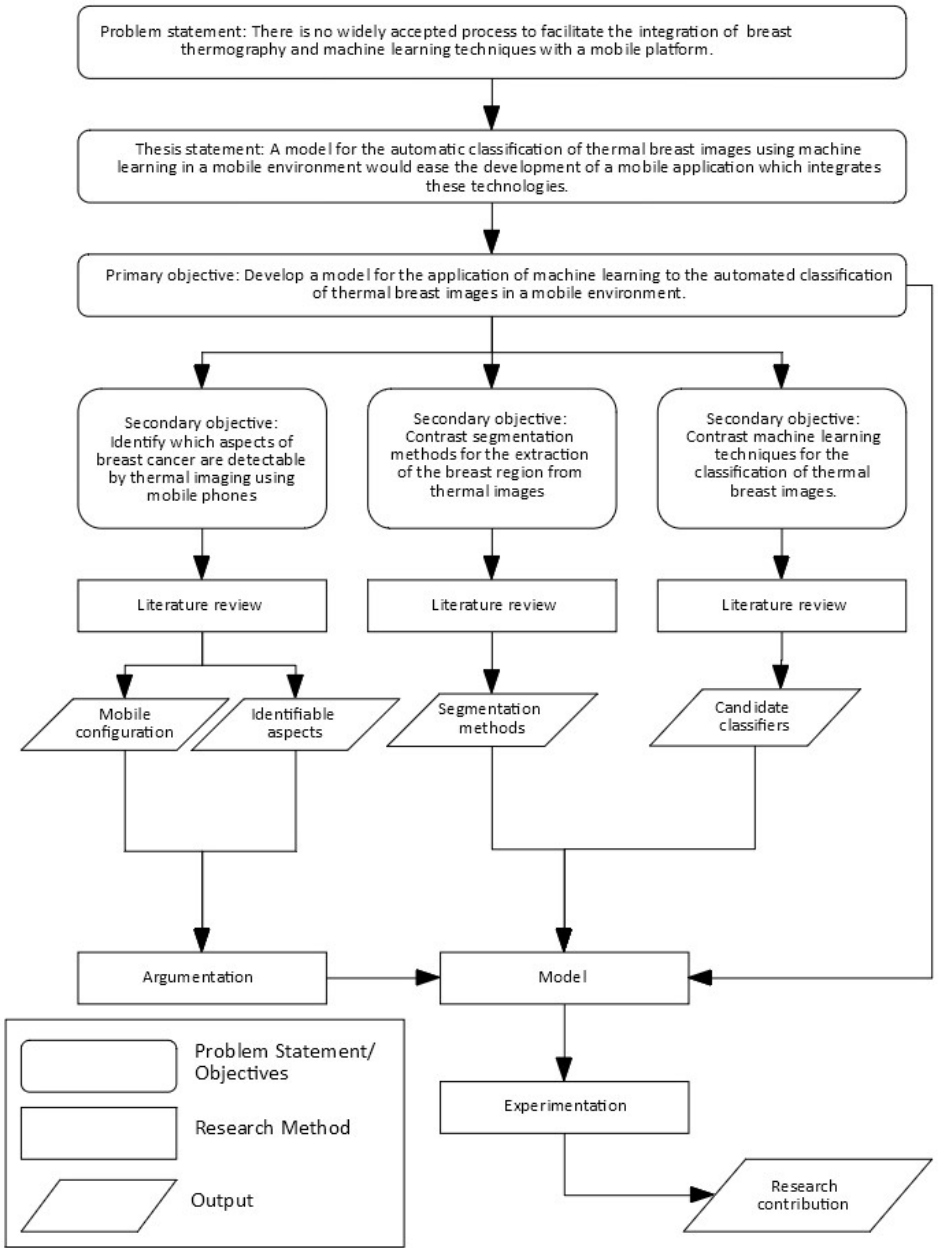


Figure 1.1: Research process undertaken in this study

1.11.2 Argumentation

According to Olivier (2009), an argument uses facts to derive new facts. It takes facts as a premise and makes conclusions based on that. This study uses argumentation to form the initial model from literature to address partially the primary objective. Argumentation is once again used, in conjunction with experimentation, to identify the best suited segmentation method for secondary objective two and the best classifier for secondary object three.

1.11.3 Experimentation

Experimentation will be used in part to meet secondary objective two and three. Olivier (2009) states that one undergoes experimentation when one wants to find something out by comparing cases. In the context of this study, the cases for comparison are machine learning techniques, different sets of features extracted, and segmentation methods performed on thermal breast images.

1.11.4 Model development

Olivier (2009) states that a model captures the essential aspects of a system while ignoring the non-essential aspects, it can be used as blueprint for a new system or to evaluate an existing system. The model is the main driving force of the study, and is the primary output. Its creation is done initially after sub objectives one and two have been completed. The model is validated through an experiment.

1.12 Chapter layout

Chapter 1 of this study provides the necessary background information of the domain. It establishes a research problem as well as objectives to solve this problem. Research methods used in this study are discussed, a research process is outlined, and a chapter layout is presented.

Chapter 2 is a literature review that provides background information on thermal imaging and the type of detectors available, mobile thermal imaging, and on the link between temperature and health.

Chapter 3 is a literature review of breast cancer and the application of thermography towards its detection. This is to understand the physiological signatures detectable via thermal imaging as well as the standards involved while taking a thermal image of a patient for medical use.

Chapter 4 is a literature review of machine learning, with a focus on classification. This is to understand how machine learning can be applied to thermal image analysis.

Chapter 5 is where the conceptual model is proposed based on the background provided by Chapters 2 to 4. The creation of the individual components is detailed as well as how they fit together.

Chapter 6 is partly a literature review chapter where the experimental process is outlined. The specific application of machine learning analysis of thermal images equivalent to those

captured by mobile attachable cameras is detailed. This serves as an implementation of the model proposed in Chapter 5. It provides details on the various stages of the thermal breast image classification process, including applicable segmentation methods and classification algorithms.

Chapter 7 analyses the results of the experimental procedure. It provides insight into the impact that various segmentation methods, choice of features, and classification algorithms have on the diagnostic accuracy of a system designed to be an implementation of the proposed model.

Chapter 8 concludes the study and ensures that the objectives stated in *Chapter 1* are met. Limitations of the study as well as future work are discussed. Figure 1.2 provides a graphical representation of the chapter layout of this study.

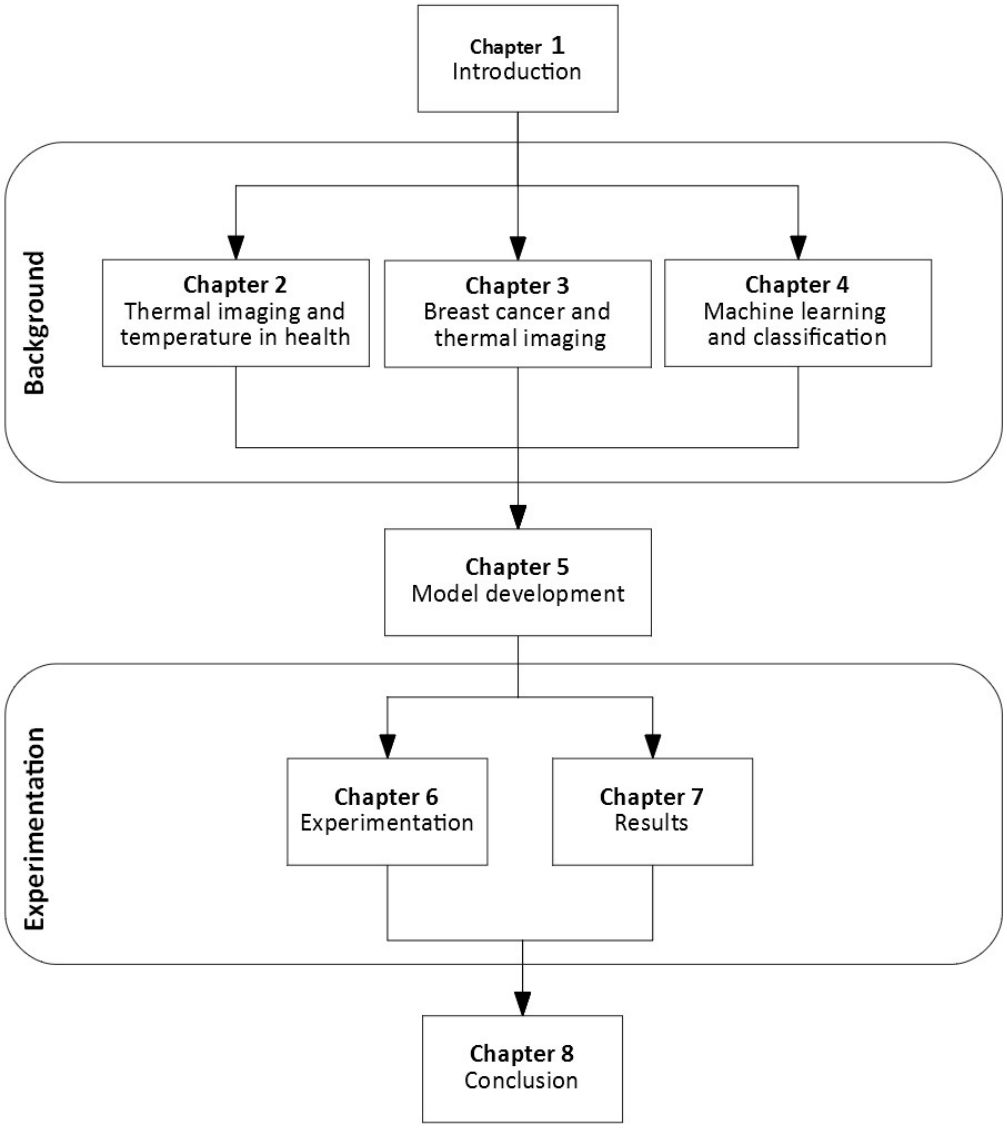


Figure 1.2: Chapter layout for this study

1.13 Ethical considerations

All datasets used for training and testing will be taken from publicly available resources where permission was given by the participants. Therefore no patients are imaged in this research and consequently no ethical clearance is required.

1.14 Conclusion

Breast cancer is a common and deadly disease amongst women. Early detection is key for improving the chance of survival. There are many existing modalities for the early detection of breast cancer, one being thermography. The advancement of technology has allowed the use of mobile devices in medical practices as well as the use of machine learning algorithms to aid in the decision-making process regarding medical diagnosis. This study proposes a model to integrate these technologies, providing a guide on how to classify thermal breast images automatically using a mobile device with an attachable thermal camera.

The next chapter discusses how temperature is measured using an infrared camera, how the human body regulates its temperature, and how anomalous temperature differentials are an indication of an underlying problem.

Chapter 2

Thermal imaging and temperature in health

2.1 Introduction

Infrared (IR) imaging is an imaging technique that absorbs infrared radiation emitted by a body and creates images of that radiation. To do this, specialised sensors and lenses are used. Originally used by the military, it is now used in many areas such as science, law enforcement, industry, and medicine.

This chapter begins with the discovery of infrared light, and how the first infrared cameras where constructed. Various types of thermal cameras are explored, with an emphasis on their inner workings and how they form part of full camera systems. The smartphone environment are details relating to application creation and how thermal imaging can be done on smartphones. The link between body temperature and health is explained with a focus on the thermal regulation of the human body and how measuring the surface temperature is a strong indication of an underlying pathology.

2.2 Discovery of infrared light

For much of human history we only knew about light that was within our visible range. In 1671 Sir Isaac Newton experimented with a glass prism where he separated white light into its separate colours, called a spectrum. This was a phrase coined by Newton in his treatise, *Opticks* (Newton, 1730).

In 1800 the visible spectrum was more precisely defined when light outside of the visible range was discovered by Sir William Herschel, a German-born British astronomer, who exper-

imented with the measurement of the different colour temperatures. He replicated Newton's experiment to separate white light into its full spectrum. Then he measured each colour with a thermometer and found the temperatures to be different, based on the colour, also that all were higher than that of the control. Furthermore, he noticed that the peak temperature was not from within the spectrum, but from just below the red end. This led to the discovery of infrared light, *infra* meaning *below* in Latin (Rogalski, 2012).

2.3 Infrared detectors

An infrared camera converts thermal radiation into an image in much the same way that a camera converts visible light into an image. To do this, infrared cameras have special sensors and lenses. This section aims to explain what infrared detectors are and how they work, the importance of correct lenses, and how they have changed over the course of history. Put very simply, an infrared detector is a detector that reacts to infrared radiation. It acts as a transducer converting radiation into electrical energy (Vollmer & Möllmann, 2017, p. 73). Infrared detectors can be categorised into two classes: photon detectors (cooled) and thermal detectors (uncooled).

2.3.1 Photon detectors

Photon detectors use single-step transduction, which allows absorbed electromagnetic radiation to change the electronic energy in a semi-conductor, by a process called the internal photoelectric effect(Gade & Moeslund, 2014; Vollmer & Möllmann, 2017, p. 87). This change is due to the fact that the free charge carrier distribution is changed. The change means that a semiconductor exhibits a typical electronic band structure with allowed electron energies, called energy bands, and forbidden electron energies, called band gaps.

To have electrons flow through the material, and thus create a current, the quantum photon energy ($E = hf$, from Section **??**) of the incident radiations photons must exceed a certain energy threshold, ΔE (Vollmer & Möllmann, 2017, p. 83). Vollmer and Möllmann (2017, p. 83) go on to explain that this energy threshold is the excitation energy of an electronic transition between the valence (upper) and the conduction (lower) band; in other terms, the energy needed to change state from the energy gap to the energy band. Figure 2.1 demonstrates this band structure.

Owing to ΔE being a threshold, and the fact that wavelength is related directly to frequency (from Section **??**), there is a cut-off wavelength that a photon detector simply cannot detect. This means that the responsivity and detectability are strongly wavelength

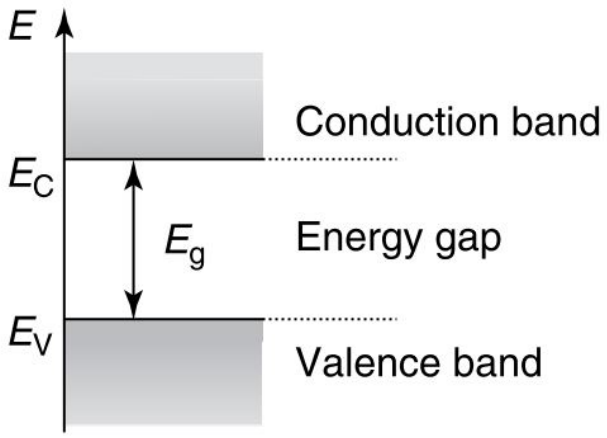


Figure 2.1: Energy band structure (Vollmer & Möllmann, 2017, p. 83)

dependent and that photon detectors operate in the MW infrared band where the thermal contrast is high, leading to photon detectors being very sensitive to changes in environmental radiation (Gade & Moeslund, 2014).

Photon detector noise is mostly made up of two groups of noise; dark current noise and radiation-induced noise (Vollmer & Möllmann, 2017, p. 85). The detector will be exposed to the signal it is measuring and to some background radiation.

In order to reduce background radiation, and overall noise, the detector needs to be cryogenically cooled (typically to 77K), which decreases the free charge carrier concentration (Vollmer & Möllmann, 2017, p. 90). Depending on operation principles, photon detectors can be divided into different types. These types can either be classical semiconductor detectors, such as photoconductors and photodiodes, or novel semiconductor detectors, such as photo emissive Schottky barrier and band gap engineered quantum well infrared photodetectors (QWIPs) (Vollmer & Möllmann, 2017, p. 90). These semiconductor detectors are commonly made of mercury cadmium telluride, Indium arsenide, and Indium antimonide (Kandlikar et al., 2017).

2.3.2 Thermal detectors

Thermal detectors are examples of two-step transducers. The first step is the absorbing of incident radiation which changes the temperature of a material. This then changes a physical property of the heated material. The changed physical property is used to generate an electrical output (Vollmer & Möllmann, 2017, p. 73).

Thermal detectors have an advantage of being wavelength independent; thus, they can be used on the LW infrared range, but have been traditionally slower and less sensitive than

photon detectors (Rogalski, 2012). Thermal detectors have many thermal mechanisms, but only a few have shown great utility, resistive bolometric effect, and the pyroelectric effect (Gade & Moeslund, 2014; Kruse, 2001).

The pyroelectric effect occurs when a material exhibits a spontaneous change in polarisation when exposed to changes in temperature, measured as transient electrical charge. These materials are normally ferroelectric, meaning that their natural electrical polarization is reversible (Gade & Moeslund, 2014).

The ferroelectric effect in these materials disappears after a certain temperature is reached, known as the Curie temperature, and is the strongest just before this temperature is reached (Kruse, 2001). To keep the detector at an optimum state, thermoelectric temperature stabilisers are used to keep the temperature just below the Curie temperature. Some materials, such as barium strontium titanate (Gade & Moeslund, 2014), allow the Curie temperature to adjusted to room temperature by varying the levels of each element (Kruse, 2001).

The bolometric effect results in a change of resistance when the temperature changes (Vollmer & Möllmann, 2017, p. 78). It is typically a thin metal or semiconductor film, which is placed in a cavity in order to isolate it thermally, so that the only change in temperature is a result of the object being detected (Kruse, 2001). Typically, a microbolometer is made of either amorphous-silicon (a-Si) or vanadium oxide (VOx). VOx has become the more popular option and holds a majority market share (Gade & Moeslund, 2014).

Owing to the initial lack of sensitivity and the slow response of thermal detectors, they were ignored for many decades until thermal detectors were placed into large arrays, called focal plane arrays (discussed further in Section 2.4.1), resulting in higher sensitivity and better frame rates matching those of photon detectors (Rogalski, 2002). The integration of detectors into focal plane arrays has led to uncooled infrared detectors being developed that are small enough to be handheld and do not require cooling, dramatically lowering their price (Kakileti et al., 2017).

Aside from the arrangement into FPAs, microbolometers have become the go to uncooled detector method owing to them having higher accuracy, having higher resolution because of smaller pixel sizes on the detector, exhibiting far better noise equivalent temperature difference than other uncooled methods, and not suffering from the halo effect as other uncooled methods do (Gade & Moeslund, 2014).

As of 2010, they hold more than 95% of the uncooled IR detector market share, but will be challenged by newer silicon-based materials (Rogalski, 2012). Uncooled cameras are smaller than cooled ones, and are less expensive, making them the preference if performance is comparable (Kandlikar et al., 2017).

2.3.3 Lenses

Glass is not a suitable material for IR camera lenses because it does not allow most of the thermal radiation to be transmitted. To allow for a better transmittance, other materials need to be used (Gade & Moeslund, 2014). According to Kakileti et al. (2017), the materials used for lenses are typically made from either Germanium (Ge), Zinc Selenide (ZnSe), Zinc Sulphide (ZnS), or Chalcogenide glass.

A lens has an f-number, which is simply a ratio of focal length to diameter, and the cost of a camera decreases with a higher focal length (Gade & Moeslund, 2014).

2.4 Full camera systems

As stated in Section 2.3, the purpose of an infrared camera is to convert the infrared radiation of an object into a visual representation of its temperature. To become a full camera system, all the previously discussed components must be used together with an interface for a user to interact with (Vollmer & Möllmann, 2017, p. 101). Infrared cameras can be used either as scanning devices or as a staring array (Gade & Moeslund, 2014).

In scanning systems, the image is generated row by row, and in staring systems, the image is projected to all pixels simultaneously (Vollmer & Möllmann, 2017, p. 102). A staring system is a focal plane array, briefly mentioned in Section 2.3.2, and is the dominant technology owing to it having no moving parts, being faster, and having better resolution than a scanning system (Gade & Moeslund, 2014).

2.4.1 Focal plane arrays

According to Rogalski (2012), a focal plane array (FPA) is an arrangement of individual detector elements, known as pixels, which are located at the focal plane of the system. For infrared imaging, an FPA consists of the infrared sensor, made from an infrared radiation-sensitive material, and the readout integrated circuits (ROIC), made from silicon (Vollmer & Möllmann, 2017, p. 104).

An FPA can either be classified as hybrid or monolithic and the choice depends on the technical requirements and costs involved (Rogalski, 2012). Lau (1996) describes a hybrid system as a system that consists of two parts, which are manufactured separately to be joined later, typically via flip-chip bonding. Monolithic arrays are slightly different in the sense that all the manufacturing is done on a single piece of equipment. Firstly, the ROIC is made using silicon and the sensors are built on it via lithography, etching and thin film deposition (Vollmer & Möllmann, 2017, p. 104).

The trend is towards using a monolithic system because of improved performance, strength, and lower costs (Kruse, 2001, p. 2). The improvements to FPA technology have led to a revolution in IR imaging (Rogalski, 2012). These improvements follow Moore's Law and their growth is thus limited by it. The improvements in IR technology have led to the following outcomes (Vollmer & Möllmann, 2017, p. 105):

- The number of pixels in the array is increasing with cameras up to 1024x1024 commercially available.
- The size of the pixels is decreasing. For bolometers, $17\mu m$ x $17\mu m$ pixels are possible, and for photon detectors, it gets as small as $15\mu m$ x $15\mu m$. This change in size leads to a decrease in cost because of a lower f-number (as indicated in Section 2.3.3).
- The noise equivalent temperature difference (NETD) is decreasing. NETD is the temperature resolution the smallest difference in temperature detectable by the sensor. At a baseline of $30^{\cdot}C$ microbolometers achieve NETDs of 45mK and photon detectors achieve 10mK.
- The fill factor is increasing. The fill factor is a ratio of the infrared sensitive area versus the entire area of the array. Microbolometers achieve fill factors of above 80% and photon detectos can achieve over 90%.
- Microbolometers have higher frame rates and better resolution owing to the time constraint decreasing to approximately 4ms.

Over the decades, the sensitivity of infrared cameras has been improving. Older cameras were as bad as 0.3K, which is not sensitive enough to capture fine details in heat patterns. They have improved to 10mK using cooling. Table 2.1 details the change in sensitivity over the years.

Table 2.1: Change in IR camera sensitivity (Kandlikar et al., 2017)

Year	**Camera**	**Sensitivity (mK)**
1972	AGA 750	230
1980	ISI Videotherm	150
1987	Inframetrics 500M	100
1995	Inframetrics 600M	50
2000	Amber PM	39
2005	FLIR A8300	20

Table 2.1: Change in IR camera sensitivity (Kandlikar et al., 2017)

Year	Camera	Sensitivity (mK)
2010	FLIR SC6000	< 20

At these sensitivities small localised temperature changes can be clearly seen. These cooled cameras are very expensive and bulky, but there has been development of uncooled microbolometers, which are small, handheld and cheaper (Kakileti et al., 2017).

2.5 The potential of smartphones in thermography

Advancements in technology and user demand has transformed mobile devices from simple communication devices to handheld computers with a wide array of augmented functionality called smartphones (Fling, 2009). Smartphone devices have all the general functionality found in standard mobile devices with the addition of larger screens, faster processors, mobile operating systems and high speed data connectivity options (Ciaramitaro, 2012).

2.5.1 Smartphone platforms and applications

The smartphone environment is dominated by platforms provided by Google and Apple, Android and iPhone respectively (Sørensen, De Reuver, & Basole, 2015). These platforms are home to the two main operating systems, Android and iOS, which hold a large majority share (GlobalStats, 2019).

Having generalised platforms such as Android and iPhone enable the development of mobile applications that are guaranteed to function on any device running the chosen platform. Modern smartphones come bundled with numerous applications, for example: calendar, email, and web browser. Applications that are not bundled with the device can be downloaded through content delivery networks such as app stores. Apples's App Store (iOS) and Google's Play Store are leaders in this field (SensorTower, 2019). They offer coordinated application development through through software development kits (SDKs), application programming interfaces (APIs), and various quality assurance processes (Sørensen et al., 2015). Applications can either be developed natively to the devices by making use of the provided SDKs and development platforms, or create hybrid applications designed to function through the browser and can work on any smartphone (Adobe, 2016).

The applications installed on mobile devices are able to use the devices internal hardware as well as any attachable devices via native SDKs. In the pursuit of mobility in thermal imaging, there have been IR cameras developed for smartphone use. Table 2.2 details the currently available IR cameras for smartphone use.

Table 2.2: Available IR cameras for smartphones (FLIR, 2018; cc Globaltech, 2018; ThermApp, 2018; Seek, 2018)

Name	**Price (USD)**	**Sensitivity (mK)**	**Resolution**
FLIR One Pro	$399	70	160x120
Therm-App Basic	$999	70	384x288
Therm-App Pro	$3499	30	640x480
Seek CompactPRO	$499	70	320x240
Fortric 225 Pro	$2999	60	320x240
Fortric 228 Pro	$9499	50	640x480

These smartphone attachable cameras are designed to be used within applications developed on all major platforms. They offer the ability to conduct thermal exams while on the move at a relatively low cost.

2.6 Temperature and health

The link between body temperature and health is not a new discovery, going as far back as 480BC. Hippocrates conducted the first thermo-biological diagnosis by smearing mud on patients and observing the drying patterns (E. F. J. Ring, 2004). It was later concluded by Hippocrates that whenever an excess of heat or cold was present on a part of the human body, there was a disease to be discovered (Vardasca et al., 2018).

The next step towards temperature measurement was when Galileo Galilei developed his simple thermoscope in 1595 (E. F. J. Ring, 2004). It was only in 1868 that Carl Wunderlich developed the clinical thermometer to systematically record the temperature of patients. His findings were noted in his treatise *On the Temperature in Disease* (Wunderlich, 1871). He concluded that the normal human body temperature is between $36.3°C$ to $37.5°C$ and that any temperature outside of this range should be a warning of disease.

Humans are warm-blooded animals (endotherms). This means that our bodies are kept

at a relatively constant temperature by internal processes, which is often different from the surrounding temperature (Jones, 1998). This state of relatively constant temperature is achieved by a complicated process called thermoregulation (Ghafarpour et al., 2016) and it is only one aspect of a far more complicated process called homeostasis.

Homeostasis

Homeostasis is defined as the state of steady internal physical and chemical conditions maintained by living systems (Betts et al., 2017). It occurs because of the ceaseless interplay between the body's many regulatory systems. It is also a very dynamic condition where small disruptions in the equilibrium state can result in changes to the body's parameters (Tortora & Derrickson, 2017, p. 8).

The regulation of the body's internal environment is achieved by several feedback systems. A feedback system, or feedback loop, is defined by Tortora and Derrickson (2017, p. 10) as "a cycle of events in which the status of a body condition is monitored, evaluated, changed, remonitored, and re-evaluated".

2.6.1 The integumentary system

The integumentary system comprises the hair, skin, nails, oil and sweat glands, as well as the sensory receptors (Tortora & Derrickson, 2017, p. 145). It has several functions, namely to regulate body temperature, to detect cutaneous sensations, to store blood, to excrete and absorb substances, to protect the body from the external environment, and to synthesise Vitamin D (Tortora & Derrickson, 2017, p. 145). This research focuses on measuring skin temperature; thus, only the skin will be discussed.

2.6.1.1 The skin

The skin is known as the cutaneous membrane and is the largest organ in the human body. Its thickness is typically between 1 and 2 mm, and it is made up of a thinner superficial portion called the epidermis and a deeper, thicker portion called the dermis (Jones, 1998).

The epidermis is made up of keratinocytes, intraepidermal macrophages, epithelial cells, and melanocytes. Approximately 90% of the cells in the epidermis are made up of keratinocytes, arranged in four or five layers called strata (Tortora & Derrickson, 2017, p. 147). Keratinocytes produce a tough fibrous protein called keratin which the body uses for protection.

Marks and Miller (2017, p. 2) define these strata, going from deepest to shallowest as: *stratum basale, stratum spinosum, stratum granulosum. stratum lucidum, and stratum coreum.* Figure 2.2 illustrates these layers of the skin.

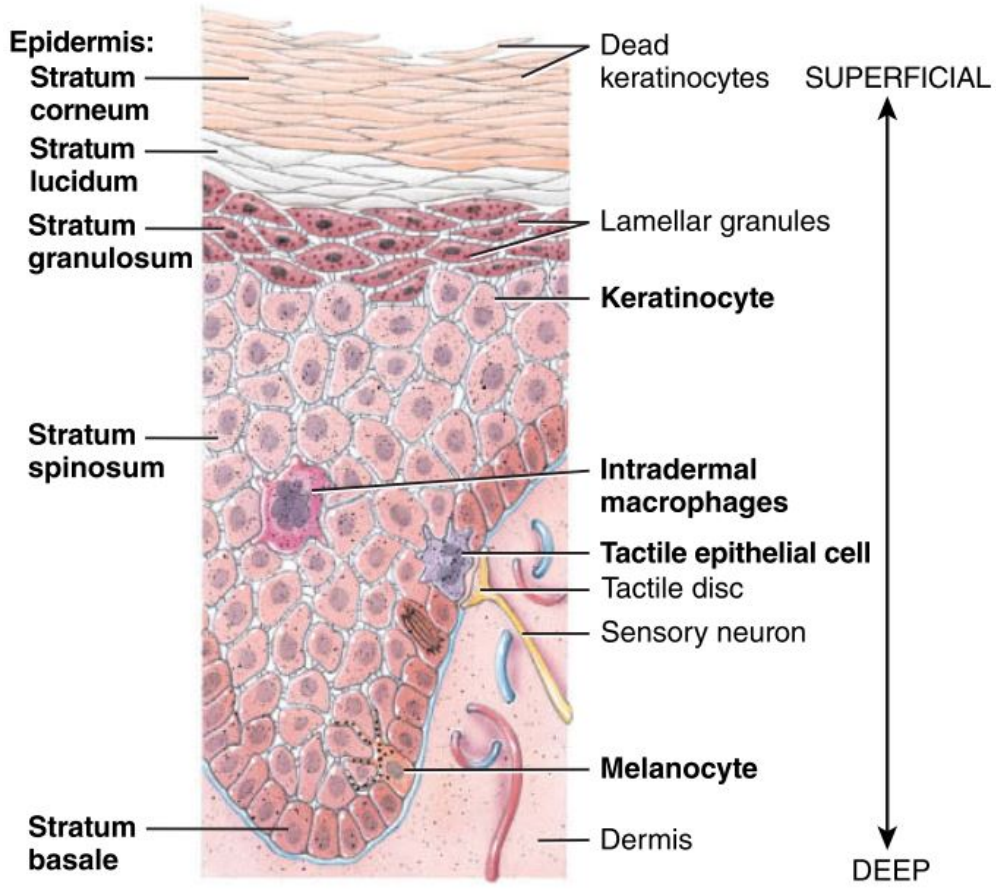


Figure 2.2: The various strata of the epidermis (Tortora & Derrickson, 2017, p. 147)

About 8% of the cells are melanocytes which are responsible for the production of the pigment melanin. Melanin, in the context of human skin, is made up of smaller molecules and is responsible for our skin colour (Jones, 1998) as well as for absorbing almost all the incidental ultraviolet (UV) radiation (Meredith & Riesz, 2004). Melanin is produced by a chemical process called melanogenesis, and our skin darkens when more melanin is present, which leads to freckles and age spots (Tortora & Derrickson, 2017, p. 151).

At the start of Section 2.6 it was mentioned that thermoregulation is the homeostatic regulation of body temperature. Skin contributes to thermoregulation in two ways: by sweating on the surface and by regulating blood flow to the dermis. Sweating results from high environmental temperatures or from exercise. The evaporation of sweat helps to lower body temperature. To regulate blood flow to the skin, blood vessels in the dermis will either

dilate (vasoconstriction) or constrict (vasodilatation). This results in either more or less heat loss to the environment (Tortora & Derrickson, 2017, p. 159).

2.6.2 The maintenance of normal body temperatures

Balancing heat generated with heat loss is how our body regulates its temperature. Tortora and Derrickson (2017, p. 981) define two types of temperature when speaking about our body, core and shell temperature. The shell temperature is the temperature of the skin and the subcutaneous layer, whereas core temperature is the temperature of the deep structures of the body.

The shell temperature varies and can be up to $6°C$ warmer than the core temperature depending on the environmental conditions, but the core temperature is highly regulated to within a narrow range (Jones, 1998). Hyperthermia occurs when the core body temperature is higher, in excess of $38°C$, than it should be under normal thermoregulation (Laupland, 2009). Hypothermia is the opposite of hyperthermia, that is, when the core temperature is lower, below $35°C$, than it should be under normal thermoregulation (Danzl & Pozos, 1994).

2.6.2.1 Metabolism

Metabolism is a term for all the chemical reactions that take place in the body; it comprises anabolism and catabolism (Jones, 1998). Tortora and Derrickson (2017, p. 954) explains anabolism as an endergonic (consumes more energy than it produces) chemical reaction that forms complex components by combining simple molecules, and catabolism as an exergonic (produces more energy than it consumes) chemical reaction that breaks down complex organic molecules into simpler ones. All chemical reactions in the body depend on the efficient transfer of energy between molecules. The molecule adenosine triphosphate (ATP) performs this transfer (Knowles, 1980).

Whenever a catabolic reaction takes place, about 60% of the energy is released as heat (Jones, 1998). The rate at which these reactions make use of energy is termed the metabolic rate. This metabolic rate is influenced by many factors: hormones, exercise, nervous systems, body temperature, ingestion of food, age, gender, sleep, climate, and malnutrition (Tortora & Derrickson, 2017, p. 979). When a body is at rest, the measured metabolic rate is known as the basal metabolic rate (BMR) and can be measured in joule per hour per kg body mass (McNab, 1997). The higher the metabolic rate, the more heat the body produces.

2.6.2.2 The skin's role in heat transfer

Heat is defined by Incropera, Lavine, Bergman, and DeWitt (2011, p. 2) as "the thermal energy in transit due to a spatial temperature difference". The skin is the dynamic interface between the body and its surroundings (E. F. J. Ring, 2010). It transfers heat in four ways: via conduction, convection, radiation, and evaporation. Figure 2.3 shows how these heat transfer methods help maintain thermal equilibrium for humans.

Conduction

Conduction is the transfer of heat between two bodies in contact with one another (Tortora & Derrickson, 2017, p. 981). In this scenario, the heat flows from the warmer object to the cooler one, via energy diffusion, at a certain rate called heat flux (Incropera et al., 2011, p. 4). Approximately 3% of body heat is lost to conduction.

Convection

Convection is the transfer of heat owing to a fluid (air or water) moving between two areas of differing temperatures (Tortora & Derrickson, 2017, p. 981). On the microscopic scale, the moving fluid of differing temperatures interacts with the object via conduction at the boundary layer. The accumulation of this diffusion and the bulk motion of the fluid result in convection (Incropera et al., 2011, p. 4). Convection can either be forced (a blowing fan or wind) or free and about 15% of the body's heat is lost this way.

Radiation

Incropera et al. (2011, p. 8) define radiation as the energy emitted by a material at a non-zero temperature transferred by electromagnetic waves (Section **??**). This allows heat to be transferred from a warmer body to a cooler one without any contact. At room temperature about 60% of body heat is lost owing to radiation (Tortora & Derrickson, 2017, p. 981).

Evaporation

Evaporation is the heat lost in the conversion of liquid to a gas. Incropera et al. (2011, p. 413) explain that the molecules at the surface of the water collide with a gas and have their energy increased such that they overcome the binding energy. The energy needed to sustain evaporation is known as the latent heat of vaporisation and this required energy must come from the internal liquid. This energy needs to be replenished from, in this case, the body. Tortora and Derrickson (2017, p. 981) states that about 22% of heat is lost via

evaporation, approximately 700mL per day (300mL from exhaled breath and 400mL from sweat). More active individuals can lose up to 1 litre of sweat per hour (Jones, 1998).

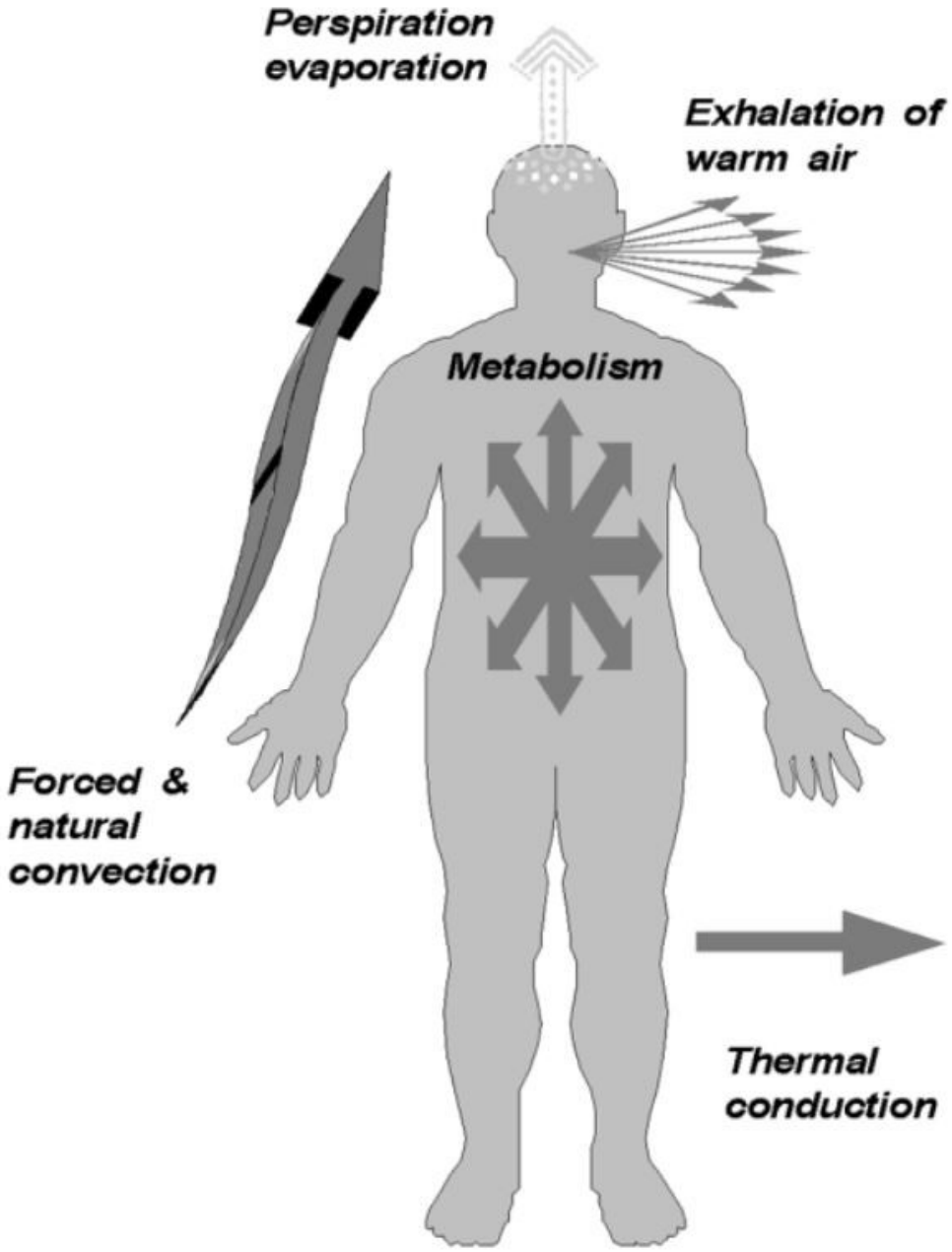


Figure 2.3: Thermal equilibrium maintenance for humans (Ghafarpour et al., 2016)

2.6.2.3 Cardiovascular system

The cardiovascular system contributes to homeostasis by transporting blood around the body (Tortora & Derrickson, 2017, p. 737). Blood consists of blood cells suspended in plasma. It

has three main functions: transportation, regulation, and protection (Tortora & Derrickson, 2017, p. 669). Blood can flow through the body via blood vessels, of which there are five main types: arteries, arterioles, capillaries, venules, and veins (Tortora & Derrickson, 2017, p. 738).

Each type of blood vessel has a specific purpose, which results in varying structures. The heart pumps the blood from the arteries, which then transport it to the capillaries, and finally the blood is returned to the heart through the veins (Jones, 1998).

The dermis contains many blood vessels and is seen as a blood reservoir (Tortora & Derrickson, 2017, p. 159). It plays a vital role in thermoregulation by controlling the dilation of the blood vessels to alter the flow of blood to the surface. This allows more heat to be lost to the surroundings and thus lowers the body temperature (Vollmer & Möllmann, 2017, p. 535).

2.6.3 Emissivity of skin

In 1934 Hardy was the first to describe the physiological role of the IR emission of the skin, owing to the temperature of the skin being influenced by many factors (Sections 2.6.1 - 2.6.2). These factors change when a disease is present; therefore, IR imaging can be used for diagnostic purposes (Faust, Rajendra Acharya, Ng, Hong, & Yu, 2014).

In order to determine the temperature of the skin accuratley, the emissivity must be determined in order to calibrate the cameras. Hardy (1934) estimated it to be approximately 0.989 ± 0.01, and it was independent of wavelength. This was later confirmed by Watmough and Oliver (1969). Steketee (1973) conducted a more thorough test to determine the differences in skin tone with better equipment in order to rectify the differences found in literature. They used a monochromator to determine the emissivity and found it to be the same for white or black skin. Whether it was measured in vivo or in vitro did not matter. Table 2.3 details their results.

Table 2.3: Emissivity values for skin, from (Steketee, 1973)

Tissue	**Infrared range (μm)**	**Emissivity**
Black skin	3 - 12	0.98 ± 0.01
White skin	3 - 14	0.97 ± 0.02
Burnt skin	3 - 14	0.97 ± 0.02

The effect of curvature on skin emissivity owing to inconsistencies in temperature mea-

surement was investigated by Watmough, Fowler, and Oliver (1970). They concluded that it was due to the viewing angle and that viewing angles under $\frac{\pi}{4}$ gives the lowest error. From this, studies were undertaken to measure the normal symmetry of temperature distributions accurately. Uematsu (1985) undertook one such study where it was identified that the difference in temperature for corresponding areas was on average, $0.24°C$. This meant that, when the environment and patient were adequately controlled, the measurement of the skin temperature was able to give an indication of health.

2.7 Conclusion

The discovery of electromagnetic waves and its spectrum allowed measurements to be made outside of our normal visible spectrum. This allowed the measurement of temperature by observing the infrared radiation emitted using specially designed camera systems, revolutionising many industries.

Temperature has long been known as an indicator of health as sick patients were abnormally warm. This is because the body tries to remain at a regular temperature via a process called homeostasis, and any disease disrupts this state. The heat generated by normal bodily functions needs to be balanced by losing heat to the surroundings. This heat loss is via the skin by various means, with blood the main transporter of heat to this boundary where heat can be lost.

This understanding led to infrared cameras being used in order to measure body shell temperature on a large scale, where many experiments were conducted in order to identify the mechanics of thermoregulation and how the presence of disease changes this state. It was found that the presence of disease altered the homeostatic state measured as control. This meant that infrared imaging could be used for diagnostic purposes.

Skin emissivity is an important factor needed for the accurate measurement as it is used to calibrate infrared cameras. Research was done in order to understand skin emissivity and it was found that human skin, regardless of colour, was a near perfect black body. Calibrating for emissivity coupled with the vast improvement in camera systems over the decades has led to infrared cameras being highly accurate in their measurement of skin temperature. The advancement of mobile technology has allowed for the measurement of temperature to be done via a smartphone application using attachable cameras.

The next chapter discusses how measuring the emitted infrared radiation from a patient enables the detection of breast cancer. It also discusses the various forms of breast cancer as well as the related treatments available.

Chapter 3

Breast cancer and thermal imaging

3.1 Introduction

This chapter covers breast cancer and how infrared imaging is applied for its detection. It begins with an explanation of cancer, what causes cancer, and how we have learnt to treat it. It then goes on to define breast cancer, its various types, staging, detection methods, and treatment.

Infrared imaging's role in breast cancer detection is then discussed, as well as how it has changed over the years and what standards have been put in place in order to attain the most accurate results possible. Advances in breast cancer detection software by means of making use of machine learning has vastly improved the diagnostic accuracy of the modality. See Chapter 4 for a review of common machine learning techniques used in pattern recognition and classification.

3.2 Cancer

Cancer is a term for a collection of related disorders, which all are associated with irregular cell growth which leads to the formation of tumours, or neoplasms, that invade the surrounding tissue known as malignancy, and can spread throughout the body known as metastasis (Kandlikar et al., 2017). Not all tumours are malignant, benign tumours are tumours that do not metastasize - to undergo metastasis, and can be surgically removed with little to no risk of them returning (Tortora & Derrickson, 2017, p. 100).

3.2.1 Categories of cancer

There are various broad categories of cancer, each having their own types, these broad categories are defined by Tortora and Derrickson (2017, p. 100):

- The most common type of cancer is carcinoma. It is cancer that develops from epithelial cells. It can form from the skin, the breasts, the lungs, and other organs or glands.
- Sarcoma is a cancer that forms in the the connective tissues.
- Leukemia is a cancer of the blood in which rapid anomalous blood cell growth is exhibited.
- Lymphoma is a cancer of the immune system, it develops from a type of white blood cell.

Bray et al. (2018) states that, "in the 21st century, cancer is expected to rank as the leading cause of death and the single most important barrier to increasing life expectancy in every country of the world". There are 36 recognized types of cancer for adults, the most common types, across the world and both sexes, are: lung (11.6%), breast (11.6%), prostate (7.1%), and colon (6.1%). Lung cancer has the highest mortality rate at 18.6% of total cancer deaths, followed by female breast cancer at 15% (24.2% incidence rate for females), colorectum at 9.2%, and stomach and liver both having 8.2% of total cancer deaths in 2018 Bray et al. (2018).

3.2.2 Events that enable cancer

Cancer research has been able to categorise the events which enable cancer, these events are both cellular and molecular in nature Blanpain (2013). Hanahan and Weinberg (2011) detail these events fully, some of which include: the uncontrolled proliferation of cells, the ability to evade tumour suppression methods, apoptosis, promoting angiogenesis in order to create a suitable environment, and the ability to metastise (spread).

3.2.3 Factors for cancerous growth

Several factors can cause a normal cell to become cancerous, it can be due to one or a combination of the factors. Tortora and Derrickson (2017, p. 101) states these factors as:

- Carcinogenic factors: any factor, substance or radiation, that alters our DNA and produces cancer. Examples include hydrocarbons found in cigarettes, x-rays, and ultraviolet (UV) radiation in sunlight. Most human cancers are caused by carcinogens.

- Oncogene factors: cancer causing genes that can transform normal cells into cancer causing cells (Stratton, Campbell, & Futreal, 2009; Stratton, 2011; Blanpain, 2013; Alexandrov et al., 2013)
- Oncogenic Virus factors: viruses that induce abnormal proliferation of cells, which causes cancer.

The cancer genome

The multistep process of cancer development is called carcinogenesis and can have as many as 10 distinct somatically acquired mutations that need to occur before a normal cell is turned cancerous (Stratton et al., 2009; Tortora & Derrickson, 2017, p. 101). Sir Michael Stratton started the Cancer Genome Project in 2000, which is dedicated to identifying sequence variants/mutations critical in the development of human cancers, he and his peers have done extensive research on how cancer changes our genes. In 2011 he published a paper detailing the progress made and their goal to create a complete catalog of somatic mutations (Stratton, 2011). Stratton, alongside many other peers worked to analyze and extract 20 distinct mutational signatures (Alexandrov et al., 2013).

Cancer is not a single disease and the cells in a tumour population rarely behave the same way, this make cancer incredibly difficult to treat. The rapid proliferation of cancer cells leads to a diverse population of abnormal cells, thus further complicating the treatment process. The treatment depends on the type of cancer and how advanced it is, it can be one treatment method or a combination of methods. In order to provide the best treatment, the disease needs to be properly detected. The detection and diagnosis of cancer has many methods called modalities. Detection and treatment for breast cancer are discussed in Sections 3.3.3 and **??**.

3.3 Breast cancer

This section begins with the anatomy of a healthy breast, then moves on to define how breast cancer forms and its types. The various stages of breast cancer are then explained. Screening and detection methods are covered as well as its diagnosis.

A breast typically consists of glands and ducts surrounded by connective tissue with embedded lymph nodes, illustrated by Figure 3.1 (Kandlikar et al., 2017).

As discussed in Section 3.2, cancer is formed by the abnormal mutation of healthy cells, simply put, breast cancer refers to a malignant tumour that has developed in the breast. There are various types of breast cancers which depend on the location of the malignancy

and its histological (tissue) structure. A breast cancer can be invasive, non-invasive, or metastatic in nature. When breast cancer metastases (spreads through the body) it does so mainly through the lymph nodes, at this point there is a large risk of mortality. The various stages of breast cancer are discussed in more detail in Section 3.3.2.

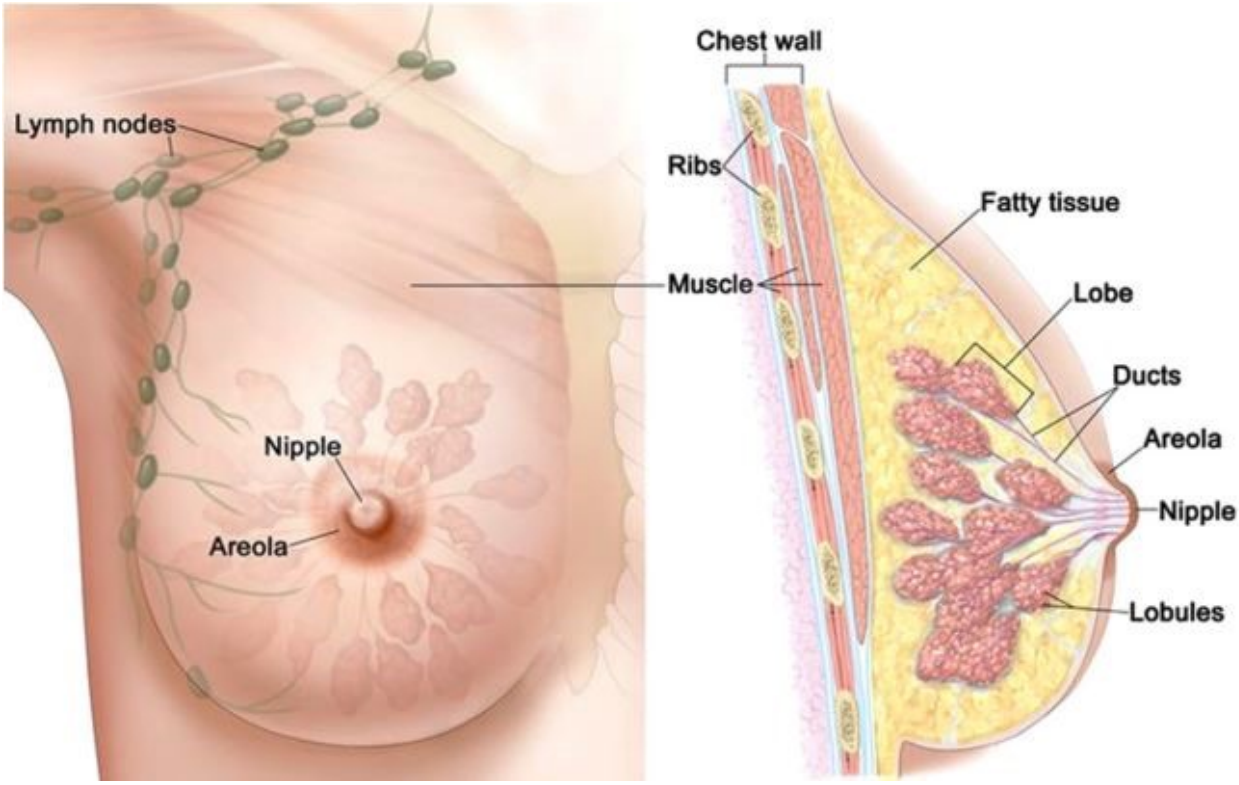


Figure 3.1: Anatomy of female breast (Borchartt et al., 2013)

3.3.1 Types of breast cancer

When cancer begins to form, various physiological processes occur, these processes function to serve the abnormal cells in their proliferation and potential metastasis through the body. Cancers disruption of normal angiogenesis allows the tumour to grow faster by providing the needed nutrients (Y. Ng et al., 2001).

There are many types and subtypes of breast cancer, the more common ones are summarized from (Weigelt, Peterse, & Van't Veer, 2005; Weigelt, Geyer, & Reis-Filho, 2010; Breastcancer.org, 2016):

- Ductile carcinoma: cancer beginning in the milk ducts.
- Lobular carcinoma: cancer beginning in the milk producing lobules of the breast.

- Tubular carcinoma: a smaller and less aggressive subtype of ductile carcinoma made up of tube-shaped structures.
- Cribriform carcinoma: cancer that forms in the connective tissue between the duct and lobules.
- Medullary carcinoma: rare subtype of ductile carcinoma that is small and does not spread outside the breast.
- Papillary carcinoma: a subtype of ductile carcinoma which encompasses a morphologically heterogeneous group of lesions (damaged tissue), all of which share a growth pattern characterized by the presence of arborescent (tree like) fibrovascular (having both fibrous and vascular tissue) stalks lined by epithelial cells (Pal et al., 2010). It is typically present in post-menopausal woman.
- Metaplastic carcinoma: cancer with the presence of multiple cellular types, typically epithelial and mesenchymal (Shah, Tseng, & Martinez, 2012).
- Adenoid cystic carcinoma: a cancer that has luminal and basaloid cells arrange in a specific structure when growing (Miyai, 2014).
- Apocrine carcinoma: cancer of the glands characterized by the apocrine cells with abundant eosinophilic (white blood cell) and granular cytoplasm, centrally to irregularly located nuclei with prominent nucleoli and distinctive cell borders (Vranic, Feldman, & Gatalica, 2017).
- Neuroendocrine carcinoma: slow-growing tumors derived from neuroendocrine (hormone and nervous system) cells, which are present throughout the body (Ogawa et al., 2008).
- Secretory carcinoma: distinct type of breast carcinoma with a good prognosis, typically found in young children and is normally found in the upper-outer quadrant of the breast (Vasudev & Onuma, 2011).

There are many more types and subtypes of breast cancer, but the cases of them are below a dozen. The most common types of breast cancer are ductal carcinoma, and lobular carcinoma (Kandlikar et al., 2017).

The body's response

Kakileti et al. (2017) state that the cancer is able to dilate blood vessels in order to attain more nutrients due to the releasing of nitric oxide as well as the general aggressiveness of the cell growth.

The body responds in various ways to the growth of cancer, it makes the area surrounding the tissue inflamed, and sends white blood cells who release nitric oxide, to combat the cancer cells (Thomsen et al., 1995; Anbar et al., 2000). This increased blood flow, release of nitric oxide, and inflammation all result in a net heat gain for the immediate area. The amount of heat emitted by the surface of the cancer cells and surrounding blood flow can be calculated from Pennes bio-heat equation (Pennes, 1948), which was later refined in 1998 by Pennes himself and now is the standard model for predicting temperature distributions in living tissues (Incropera et al., 2011, p. 178).

3.3.2 Staging breast cancer

A patient's prognosis is the chance that patient has of recovering from breast cancer. A doctor will assign a stage to a patient after analyzing the cancer, the stage depends on many factors and the survival rate is strongly linked to the stage of diagnosis. The earliest stage breast cancers are stage 0, it then ranges from stage I (1) through IV (4). The stages are summarised below, from (American Cancer Society, 2018; Breastcancer.org, 2017; Cancer Treatment Centers of America, 2019).

Stage 0

A difficult to detect stage where the tumour is non-invasive and has not spread to other parts of the body. It is often seen as a precancerous stage that requires observation and not treatment. The 5-year survival rate for this stage is 100%.

Stage I

This is the earliest stage of invasive cancer, the tumour can be up to 2cm big (stage 1A) or small clusters of tumours (1B) with no lymph nodes involved. At this stage the cancer has not spread outside of the breast. The 5-year survival rate for this stage is 100%.

Stage II

The tumor in this stage measures between 2 cm to 5 cm, or the cancer has spread to the lymph nodes under the arm on the same side as the breast cancer. Whether the stage is

determined to be 2A or 2B depends on the size of the tumour and its spread. The 5-year survival rate for this stage is 93%.

Stage III

Known as locally advanced breast cancer, the tumor in this stage of breast cancer is more than 5 cm in diameter across and the cancer is extensive in the underarm lymph nodes or has spread to other lymph nodes or tissues near the breast. Stage 3 can either be categorized as 3A, 3B, or 3C depending on its size and if its spread further than the lymph nodes. The 5-year survival rate for this stage is 72%.

Stage IV

Known as metastatic breast cancer, the cancer in this stage has spread beyond the breast, underarm and internal mammary lymph nodes to other parts of the body near to or distant from the breast. The 5-year survival rate for this stage is 22%.

It can clearly be seen that the early detection of breast cancer is key in reducing mortality. Patient comfort is an important factor in any method of detection. Therefore, there is a need for the development of non-invasive methods to improve early diagnosis and screening of suspicious breast lesions (Morais et al., 2016).

3.3.3 Detection methods

There are various established methods to detect breast cancer, but only imaging techniques are discussed here. Herranz and Ruibal (2012) state that for breast cancer detection, diagnosis as well as management, breast imaging is mostly used.

The early signs of breast cancer can only be detected using modern techniques. These early signs are small clusters of calcium deposits, called microcalcifications, and masses (Mehdy, Ng, Shair, Saleh, & Gomes, 2017). Modern techniques for breast cancer detection are most commonly: mammography, magnetic resonance imaging (MRI), positron emission tomography (PET), and ultrasound (Herranz & Ruibal, 2012). Due to needs for cheaper and less evasive techniques, the development of newer techniques has happened, these include: electronic palpation imaging (EPI), electrical impedance scanning, and infrared thermography (IRT) (Kandlikar et al., 2017).

3.3.3.1 Mammography

Mammography is the most commonly used technique which uses low energy x-rays to image the internal anatomy of the breast. It is the gold standard for breast cancer detection with a 31% reduction in mortality from regular screening (Tabar et al., 1985), as well as having high diagnostic accuracy. Due to the nature of the imaging technique, it underperforms with denser breasts (younger women), implants, and scars from surgery as it has a tougher time seeing through it (Herranz & Ruibal, 2012), which is why mammography is mainly applicable to women over the age of 40 (Kakileti et al., 2017). There is a concern of radiation exposure, due to the fact the breast is bombarded by ionizing radiation that could cause cancer. This concern is not as real as feared as (Sardanelli et al., 2017) clearly outlined the risk as 1 in 100 000, and is 100 times lower than the probability of avoiding a breast cancer related death. Which means the concern of radiation exposure with mammography is unfound statistically speaking. There is a more valid concern around the invasiveness of a mammography, the act of getting a mammogram is a painful and uncomfortable experience as the breast is clamped with approximately 8 kg of force (Kakileti et al., 2017).

3.3.3.2 Magnetic resonance imaging (MRI)

Kandlikar et al. (2017) explains that MRI uses both strong magnetic fields as well as radio waves to create an image of the internal anatomy of the breast. MRI plays more of a supportive role in breast imaging where it is used to differentiate between solid or cystic masses, evaluate palpable masses and used for needle core biopsy (Morais et al., 2016).

In recent years MRI has improved and has better sensitivity than mammography, but lower specificity (Herranz & Ruibal, 2012). It is also very expensive, both in monetary terms and time, to do. For these reasons, MRI is used to screen very high risk woman or to better image tumours found in other tests (Saslow et al., 2007).

3.3.3.3 Positron Emission Tomography (PET)

PET uses cameras that detect the radioactive emissions of an injected radionuclide, it is an additional imaging technique that provides for physiologic information. It is mainly used in the staging of cancer and monitoring response to therapy (Herranz & Ruibal, 2012).

3.3.3.4 Ultrasound

Ultrasound, also known as sonography, uses sound waves which it bounces on the surface of the tissue to create an image of the underlying anatomy (Kandlikar et al., 2017). Much like

MRI, the role of ultrasound has mainly been used to distinguish between solid masses and cysts, the evaluation of palpable masses, and for needle core biopsy (Herranz & Ruibal, 2012). Due to it not being affected by the density of breasts, it is used to supplement mammography in certain situations (Morais et al., 2016).

The other additional techniques were not discussed as they are relatively new, infrared thermogrpahy is one such technique. For a full review of infrared thermography see Section 3.4. It is important to note that no imaging technique can 100% diagnose breast cancer, that is done via a biopsy. A biopsy is an examination of tissue removed from a living body to discover the presence, cause, or extent of a disease (Stevenson, 2010). Interestingly, Morais et al. (2016) stated that two thirds to four fifths of all performed biopsies are benign lesions, thus there is a need for a technique which can easily distinguish between a benign and malignant lesion.

3.4 Thermography in breast cancer

According to Head and Elliott (2002), a thermogram shows the normal and abnormal functioning of the body's vascular system, sympathetic and sensorial nervous system, and inflammatory processes. Not all temperature changes on the surface result from cancerous growth. Chojnowski (2017) provides alternatives, these include: infection, trauma, inflammation, malignancy, or ischemia.

Classical imaging only observes anatomical abnormalities, there are physiological alteration which precede these anatomical abnormalities, such as an increase in heat generation (Gautherie, 1983). This allows thermography to be a potential tool for the early detection of these changes (Araujo et al., 2017).

As discussed in Section 3.3.1, there are physiological processes that result in localised temperature changes, these include angiogenesis and the release of nitric oxide. Cancer is linked to both of these processes (Thomsen et al., 1995; Carmeliet & Jain, 2000), this means that measuring the surface temperature gives an indication of the presence of a tumour.

Sathish, Kamath, Prasad, Kadavigere, and Martis (2017) describe thermography as a simple imaging technique which is non-invasive, painless as well as completely safe from ionizing radiation which can be used on woman of all ages, having all breast densities, and is not affected by breast implants.

History

The implications of temperature differentials of the skin with regard to breast cancer diagnosis and screening was first proposed by Ray Lawson in 1956 by analyzing temperature distributions of malignant tumours (Lawson, 1956). In 1963 Lawson with the help of Chughtai made use of infrared cameras to measure the outer surface of the region surrounding the tumour. They looked at the difference in temperatures between the unhealthy breast area and the contralateral healthy area and found the unhealthy region to be approximately $2°C$ higher.

Thermography was then proposed as a research area of interest (Williams, 1964). IR thermography was widely accepted until a systematic review in 1977 of breast cancer modalities showed that thermography, at the time, was the worst of all the imaging modalities at accurately predicting instances of breast cancer (Feig et al., 1977). Thermography was hastily added, with poor design with untrained personal and violated protocols, which skewed results to be more inconsistent than they are when well controlled (Amalu, 2003). Despite the conclusions of the 1977 study, the Food and Drug Administration (FDA) approved thermography as an adjunct to mammography in 1982 (Kandlikar et al., 2017). The low accuracy of thermography was attributed to: poor camera quality, poor image capture standards, lack of environmental controls, lack of proper training for physicians in the capturing and analysis of thermograms (Kakileti et al., 2017; Kandlikar et al., 2017).

Even during the early stages thermography, it showed promise for early detection. Gautherie and Gros (1980) showed than an abnormal thermogram led to cancer diagnosis in a third of patients five years later. Guidi and Schnitt (1996) looked at pre-invasive breast cancer patients, they reported that patients with increased micro vessels in the breasts have up to seven times greater risk of developing breast cancer compared with normal blood vessel density. Head, Wang, and Elliott (1993) studied deceased and living cancer patients as well as a healthy control group. They found abnormal thermograms to be highly correlated to sickness and the most relevant prognostic feature from this study is that breast cancer patients with abnormal thermograms have fast growing tumors.

The reappraisal of thermography

As seen from Table 2.1 cameras have improved drastically since the initial years of thermography, this, along with all the new findings and research led to a reappraisal in the late 1990s (Jones, 1998; Keyserlingk, Ahlgren, Yu, & Belliveau, 1998).

Head and Wan, Fen, Charles A. Lipari (2000) looked at stages of diagnosis compared with mammography and found that in 70% of the cases thermography was able to detect signs

of early breast cancer up to 1 year earlier than mammography. Furthermore, a persistent abnormal thermogram holds great significance. Compared to familiy history of breast cancer, this persistent abnormal thermogram is 22 times more risky and 10 times more significant (Head & Elliott, 2002).

Due to the fact that a rise in local skin temperature can be from any abnormality, such as malignant tumor, fibrosis, an infection or an inflammation, thermography tends to have high false positives, and thus relies on well-defined features that need to be considered.

Kandlikar et al. (2017) describes these features as: high asymmetry in temperature measurement distributions between breasts, hyperthermic vascular patterns, unusual complexity in the vascular patterns, areolar and peri-areolar temperature patterns, and an average temperature difference of more than $2°C$ between breasts. The features and standards created from them is discussed in more detail in Section 3.4.1.

Since the reappraisal there has been a larger interest in thermography from both research and clinical practice, nowadays many clinics and health services all over the world have used thermography for different prognostic objectives (E. Ng & Etehadtavakol, 2017, p. 46). As far as thermography has come it is still not good enough to replace mammography in the real world, but rather be a supportive role like ultrasound or MRI (Sardanelli et al., 2017).

Due to the increase in quality of the infrared cameras, there is great benefit in creating standards for image production, interpretation, storage, enhancement, and analysis (Amalu, 2003). Research still continues to try improve thermography as much as possible through better cameras, better controls, better interpretation, and larger trials. The largest improvement of thermography came when computers were used to aid in the process via pattern recognition and machine learning, for a full overview of machine learning and pattern recognition see Chapter 4. The remainder of this chapter discusses the image capture and storage standards that have been established for thermography as well as the various standards and measures for the interpretation of thermograms.

3.4.1 Image capture standards

The quality of an IR image depends on many factors, and a lack of control of those factors can lead to thermal artifacting in the image which severely hinders the reliability of the image. E. F. J. Ring (2004) states that “the usage of standard procedures reduces the amount and influence of variables, promotes understanding and knowledge exchange, and enforces reliability”.

In this imaging modality in particular the previoulsy poor results were largely attributed to a lack of standard procedures (Ammer, 2003; Kandlikar et al., 2017). Before the reappraisal

studies would independently adhere to their own control protocols for the capturing and storage of medical thermal images. This was not ideal as many studies did not take adequate control over their images and would produce very negative results.

Since then, however, there has been strict standardizations made for thermography not just breast thermography but all facets of thermography. The standards that should be followed before a patient undergoes an examination, the testing procedure and environment during the examination and the post processing of the obtained thermograms are outlined by E. F. Ring (1990) and E. F. J. Ring and Ammer (2000). Ring and Ammer later refined their initial standards in 2006 (Nowakowski, 2006, Chap. 36). This work was built on and refined once more by E. Ng (2009). The standards encompass many aspects such as:

- Preparation of patient.
- Examination environment.
- Standardization of thermal imager systems.
- Image capture protocol.
- Image analysis protocol.
- Reporting, archiving, and storing.

The above mentioned points are expanded out with brief explanations of the impact of each aspect.

3.4.1.1 Patient preparation

Proper preparation and control of the patient is vital due to the fact that uncontrolled aspects may lead to thermal patterns being present that should not be there. There are many ways to prepare a patient for a thermal exam, below a list summarises these factors (Gershon-Cohen & Habennan, 1968; Mannara, Salvatori, & Pizzuti, 1993; E. F. J. Ring & Ammer, 2000; IACT, 2002; Qi, Teja Kuruganti, & Snyder, 2008, Chapter 27).

- Large meals, coffee, tea, and alcohol consumption should be avoided before the examination due to the affect these factors have on the metabolism of the patient.
- Exercise affects the bodies metabolism as well as the blood vessels in the skin dilating in order to cool down. No physical therapy should be done 24 hours prior to the examination and no physical exercise should be done 4 hours prior to the examination. The patient should not bath 1 hour prior to the examination as the blood vessels will dilate in order to cool the patient down.

- If any part of the body needs to be shaved, this should be done at least 4 hours prior to the examination but the night before is preferred. Lotion, cream, makeup, or deodorants should not be used on the area to be imaged on the day of the examination.
- No sunbathing should be done at least 5 days prior to the examination and avoid wearing tight fitted clothing.
- The practitioner should use proper designed patient data form. Past examinations, diagnosis, and surgeries should be documented.
- It is vital the patient is allowed to acclimate to the environment in order to reduce thermal artefacts. The patients should be allowed 15 minutes to acclimate to the environment with the examined area completely uncovered. In the case of breast exams, the patient is not to fold their arms.

If the patient is controlled for these variables, there will be less erroneous thermal patterns in the generated image allowing for better detection.

3.4.1.2 Examination environment

The conditions in the enviroment need to be strictly controlled in order to avoid thermal artefacting and ensure that all the thermal patterns captured are from the patients being examined. There are strict conditions the room must adhere to outlined by E. F. Ring (1990).

- The room should have minimum dimensions of 2 x 3 meters, but 3 x 4 meters is preferred, and should be carpeted. This is to allow the room to maintain a homogeneous temperature.
- Equipment used generates heat, therefore an air conditioning system is needed to maintain optimum temperature, which should be easily visible at all times by the operator. The temperature of the room should be between 18 and 25 °C, in order to prevent shivering or perspiration, and its relative humidity between 40 and 75%. It is important no airflow is directed towards the patient, this mean that the ducts for the air conditioning unit should not be facing the patient.
- The computer equipment should not be near the patient as it may disturb the temperature reading.
- A sink with cold and hot water is needed for a stress test (cold test).

- External sources of infrared radiation are the largest influences of thermal artefacts. All windows must be shielded in order to avoid sunlight leaking in, incandescent lights or any other heat producing lights should be avoided, florescent lights are recommended. No windows or doors should be open as this creates airflow over the patient. A plain, non-reflective background should be used.

If the environment is adequately controlled the patient can be measured more accurately.

3.4.1.3 Standardization of thermal imager system

The thermal camera itself needs to be properly calibrated and allowed to warm up. The process is detailed by E. F. Ring (1990).

- In order to measure the camera drift a black body is needed to check for any drift in the temperature sensitivity settings, this should only be done if routine services have not been done. Threshold temperature should also be calibrated, due to it being a reference point to differentiate normal temperature from evaluated temperature.

- The camera can be mounted on any studio camera stand, they are preferred to photographic stands due to height adjustment ability and counter balance weights.

- The camera should be set up 15 minutes prior to the examination, this may be longer than what is recommended by the camera manufacturer. The scanner should be allowed to stabilize for about 30 seconds before screening.

IF the camera is calibrated properly and allowed to warm up, much clearer images will be produced.

3.4.1.4 Image capture protocol (ICP)

The image capture protocol essentially is a set of instructions on how to take the images. It is an essential part of a successful thermographic exam, as everything can be controlled for but the operator does not take the correct images, the exam is worthless. The goal is to capture the area of interest relevant to patients symptom, along with any anatomically and physiologically related areas. A complete set of images should be devised with the camera at 90° to the patient parallel to the ground, and mounted on a parallax free stand. These images should be the same size for all patients, even if their bodies are different sizes. Nowakowski (2006, p. 737-741) fully details the ICP, it is summarised below.

- The size of the image is dependent on 2 parameters; distance between the patient and the camera, and the camera's focal length. A workable target plane, where you want to fit as much of the area of interest in as possible, is typically $\frac{1}{3}$ of the width and $\frac{2}{3}$ of the height of the imager's working plane.

- The patient's position during the examination affects the surface temperature of the exposed area. For a breast exam, multiple images of the breasts can be taken as anatomy may obscure parts of the breast which are important. During this process the patient must not move as it will disturb the radiant infrared radiation being reflected and skew results.

- The detectors themselves need to be of a certain quality to qualify for medical use, (E. F. J. Ring & Ammer, 2000) details these requirements and they can be summarised as:

  - A response between 5 and 15 microns with a spectral bandwidth in the 8 to 10 region.
  - An accuracy of $\pm 2\%$ with a temperature calibration of 5 - 150°C, or an accuracy of $\pm 1\%$ if the calibration is < 100°C.
  - A thermal resolution better than 100mK at 30°C.
  - Limiting temperature range of 30 to 40°C.
  - A maximum scanning time of 4 seconds.
  - A resolution of at least 120 x 120 pixels (pixel resolution of $10\mu m$).
  - Maximum drift to be 0.3°C between self correction.
  - Ability to adjust emissivity between 0.94 and 0.99 (to reflect the values for human skin).
  - Minimum detectable temperature difference (MDTD) is less than 0.4°C.

If the patient and environment are controlled properly and the camera is set up appropriately, capturing the proper images will result in an accurate thermographic examination. From this point on the image is analysed in order to be categorised into classes where a diagnosis can be made.

3.4.2 Image interpretation standards

Early methods for the interpretation of thermal breast images were purely subjective and were read for variations in normal vascular patterns with no regards for the differences between

breasts, leading to large variations in outcomes (E. Ng, 2009). Since then, however, many standards for image interpretation have been established in order to narrow the variations in results. These methods include: Hobbins protocol (Hobbins, 1983), Gautherie protocol (Gautherie, 1983), and the Hoesktra protocol (Deborah, Tanya, & Dugald, 2009).

Most of these protocols were first established in the 1970s to 1980s, where there was a search for a standard method for thermovascular analysis proposed (Nowakowski, 2006, p. 551). The mentioned protocols, like all thermovascular protocols, follow a set grading of 1 to 5 TH in order to classify them thermobiologically. They are then often simplified to a binary grading of healthy or not healthy based on these criteria. The gradings depended on 20 discrete vascular and temperature attributes when looking at the breasts and comparing them contra laterally (Gautherie, 1983).

These attributes can be categorized into vascular and non vascular criteria. Kakileti et al. (2017) summarised Gautherie (1983) and Hobbins (1983) criteria as the following:

Vascular criteria

- Asymmetry
- Increased density
- Number and caliber of vessels
- Abnormal direction of clusters of vessels
- Abnormal location of vascularity
- Vascular anarchy

Non vascular criteria

- Global increase in temperature compared to contra lateral breast
- Differences in temperature of contra lateral regions
- Focal increase in temperature
- Abnormal location of focal temperature increases
- Abnormal physical observations

Depending on which attributes were present and which combinations were present, a grading would be assigned, the grading generally goes as follows:

- TH1 - Normal non vascular
- TH2 - Normal vascular
- TH3 - Equivocal
- TH4 - Abnormal
- TH5 - Severely abnormal

According to Nowakowski (2006, p. 551), the use of standardized methods has significantly improved the sensitivity, specificity, positive predictive value and negative predictive value of thermography. However, this method of analysis is highly subjective depending on the thermographer, resulting in many false positives (Kakileti et al., 2017). The use of advanced software to analyse the images has made a leap for all medical imaging techniques, due to its consistency of results and lack of subjectivity. Chapter 4 discusses machine learning and pattern recognition, two concepts that have severely improved the performance of thermal imaging systems.

3.5 Conclusion

Cancer is the uncontrolled growth of cells which invade surrounding tissue. There are many categories of cancer, the most common being carcinomas, which are malignant tumors. Breast cancer is the most commonly diagnosed cancer among women, and has one of the highest mortality rates.

There are many factors than enable cancerous growth, mostly from eternal sources - chemicals and radiation, but also from cancer causing genes in the body. Breast cancer typically forms either in the lobes or the milk ducts and eventually spreads throughout the body. Cancer promotes angiogenesis, the creation of new blood vessels, in order to provide the needed nutrients. The body responds by sending white blood cells to fight, which release nitric oxide, as well as making the surrounding area inflamed, all of which contribute to a localised heating of the tissue.

Breast cancer is staged between 0 and 4 depending on the extent to which the cancer has grown and spread. The earlier it is detected the higher the chances are of survival. Common imaging detection methods include; mammography, MRI, PET, ultrasound, and thermography. Mammography being the gold standard and the main modality, the others normally are in a supportive role.

The focus of this research is the application of thermography to breast cancer detection. Temperature differentials of the skin with regards to breast cancer were first proposed in 1956, but the poor camera quality of the time as well as a lack of standardised processes made thermography fall out of favour with the medical community. Thermography continued to develop and as cameras improved so did the accuracy of this modality, until there was a reappraisal in the late 1990's.

Thermography has come to rival other modalities in terms of accuracy since better equipment and a full standardised procedure have been implemented. One major advantage of thermography is that it in non-invasive and completely safe in terms of radiation exposure, it simply measures emitted radiation and does not bombard the patient with radiation in order to measure a response.

One downside of the modality is the difficulty of interpretation of the resulting thermal images, this problem has been vastly improved on with the advent of machine learning in the medical industry. In this case, machine learning algorithms are trained to recognise specific patterns which showcase disease. The inclusion of machine learning has vastly improved the accuracy of all imaging modalities, especially thermography. The next chapter discusses the types of machine learning algorithms that are used in the classification of thermal breast images.

Chapter 4

Machine learning and classification

4.1 Introduction

This chapter discusses machine learning and which methods are applicable to breast thermography. It begins with a discussion of the various types of machine learning systems, focusing on supervised learning. Classification algorithms are then introduced and explored as they directly relate to the problem that the research is addressing. Computer vision is then introduced and its relation to breast thermography is explained, focusing on transforming raw thermal images into feature sets capable of being given to a machine learning algorithm for classification. The various features that can be calculated from thermal images are also discussed.

4.2 What is machine learning?

Machine learning is a field of computer science where algorithms are built in order to learn from example data in order to make predictions from underlying patterns and requiring no explicit programming. It is a subset of artificial intelligence (Bishop, 2007, Chap. 1). The term "machine learning" was first coined by Arthur Samuel in 1959 when he applied machine learning to the game of checkers (Samuel, 1959). A strict definition of machine learning was established by Mitchell (1997, p. 2) where he states "A computer program 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 by P, improves with experience E".

The goal of any machine learning system is to predict outputs based on some inputs (James, Witten, Hastie, & Tibshirani, 2013, p. 8), in other words, to convert experience into knowledge (Shalev-Shwartz & Ben-David, 2014, p. 19). This input or experience is the

training data,it comes in many forms, normally digitized human-labeled training sets, or other types of information obtained via interaction with the environment. Lack of data in our modern world is not a problem as we are overwhelmed with data, and there is no end in sight (Witten, Frank, Hall, & Pal, 2016, p. 3). Along with these inputs we observe the outcomes, and together they form a data set that can be used to train a prediction model that will learn from the training data and be able to make predictions on unseen examples.

Learning

The ability to predict unseen examples is key in machine learning. If the predictor could not do that it would simply be learning by memorisation. What makes it a proper learning algorithm is its ability to create generalisations based on the training data by a process called inductive reasoning (Shalev-Shwartz & Ben-David, 2014, p. 20). Although machine learning algorithms show much similarity between humans and how we attain knowledge, speaking philosophically, there are a few subtle but key differences. Learning can be achieved in five ways (Witten et al., 2016, p. 8):

- Gaining knowledge of something by being taught, through experience, or study.
- Becoming aware from observation.
- Committing to memory.
- Becoming informed of.
- Receiving instruction.

The first two points fall short when speaking about computers. It is virtually impossible to know if the computer has gained the knowledge of something. If you could ask it, you would not be testing its ability to learn but its ability to answer questions (Witten et al., 2016, p. 8). Similarly, how would you know if a computer has become self-aware? Consciousness and awareness of computers is a philosophical issue that is not discussed in this research. Instead this research focuses on the practical aspects of machine learning.

It is very important for machine learning algorithms to have as much quality data as possible as they lack common sense and are unable to filter out meaningless data; thus, leading to false conclusions when the algorithm makes generalisations (Shalev-Shwartz & Ben-David, 2014, p. 20). Machine learning algorithms also rely on prior knowledge for an inductive bias or learning bias to be established in the data to learn faster, but if the bias is too strong the algorithm becomes very rigid and won't move from these assumptions;

thus, it is important to account for the learning bias when creating data (Shalev-Shwartz & Ben-David, 2014, p. 21). Machine learning methods are data-driven combining fundamentals from computer science with ideas from statistics, probability, and optimisation (Mohri et al., 2018, p. 2).

Where is it used?

According to Shalev-Shwartz and Ben-David (2014, p. 22), machine learning can be used when there is a task that needs performing that is too complex for humans, such as: astronomical data, weather predictions, converting medical archives into data, and performing web searches. Machine learning also needs to be used where programs need to adapt to their data, since traditional programs are very rigid.

Machine learning algorithms can complete many tasks, which makes them very useful for several applications. Some examples adapted from Mohri et al. (2018, p. 2) are listed below:

- Document classification - spam filters, identifying whether website content is explicit, topic modelling.
- Natural language processing (NLP) - part-of-speech tagging, named-entity recognition, context-free parsing, and dependency parsing.
- Speech processing - speech recognition, speech synthesis, speaker verification, and speaker identification.
- Computer vision - Object recognition, object identification, face detection, Optical character recognition (OCR), content-based image retrieval, or pose estimation.
- Recommendation systems - Netflix, YouTube, and Amazon.
- Medical diagnosis - Aiding physicians or radiographers in diagnosing disease.
- Self driving cars - companies like Tesla motors are making headway in this field. At the moment, it is purely assisted driving but a push to full automation is planned.
- Search engines - Google has revolutionised the way we access data with the use of machine learning to match our questions to content.

This is not an exhaustive list, as almost all aspects of modern life can benefit from the use of machine learning.

These applications fall into standard task categories. These categories exist to serve as focal points for research and development. Common categories include:

- Classification - Assigning an input into one of a finite number of discrete categories (Bishop, 2007, p. 3). An example would be to classify email as spam or not spam.
- Regression - Predicting a continuous variable from an input (Bishop, 2007, p. 3). An example would be to predict the stock price given its history.
- Ranking - Ordering instances by their relevance to the given task (Shalev-Shwartz & Ben-David, 2014, p. 238). An example is Google's search engine: the most relevant information is displayed given the query.
- Clustering - Grouping a set of objects together based on similarity, called clusters (James et al., 2013, p. 26). An example would be using customer market information to identify groups who buy similar products.
- Dimensionality reduction - Taking data in high-dimensional space and mapping it to space with lower dimensionality (Shalev-Shwartz & Ben-David, 2014, p. 323). High-dimensional data has a high computation cost and will lead to algorithms having poor generalization.

Most machine learning problems can be placed within one of these categories. This research focuses on classification and Section 4.3 is dedicated to expanding this topic.

4.2.1 Machine learning stages

When undertaking a problem using a machine learning approach, several steps need to be taken and certain terminology is used. This sections aims briefly to explain each of these steps and terms commonly found in machine learning.

- Examples - An example is an individual, independent instance of the concept to be learned. It is characterised by the values of predetermined attributes (Witten et al., 2016, p. 42). This is often the raw data associated with the observed phenomena.
- Features - A vector of a set of attributes associated with a particular example (Mohri et al., 2018, p. 4). In other words, a feature is a measurable characteristic of the observed phenomena.
- Hyperparameters - The parameters of the learning algorithm, which are set before the learning process begins (Géron, 2019, p. 30). They are exactly the same as traditional parameters to a computing method, but with a different context.

- Training sample - A portion of the data used to train the algorithm by providing examples for it to tune its parameters (James et al., 2013, p. 21). The amount of data you use to train depends on the scenario.
- Test sample - A portion of the data used to evaluate the algorithm by providing it with examples it has not yet seen (Mohri et al., 2018, p. 4).
- Loss function - Also called the cost function, this is the measure of the difference (or loss) between predicted labels and the true label (Bishop, 2007, p. 41). The goal of the algorithm is to minimise this loss function.
- Hypothesis set - The set of functions mapping inputs to output labels.

Mohri et al. (2018, p. 4) explain that these steps are done in a particular order. Firstly, you begin with the data collected and split it into training, validation, and testing data. How you split the data can be done in many ways and depends heavily on the given scenario. Sampling is explained in more detail in Section 4.4.

Once the data is split, there need to be features associated with the data. The choice of features is critical in the success of a machine learning model as poor features can mislead the algorithm (Mohri et al., 2018, p. 5).

Once the features have been selected, the algorithm can now be trained. The algorithm will have its hyperparameters tuned while it learns to predict the given training data better. The algorithm selects a hypothesis from its hypothesis set to test on the validation set. The best performing hypothesis is chosen to predict the testing set. Finally, the performance is evaluated using the loss function. There are many ways to evaluate the performance of an algorithm, Section 4.4 details some other methods of performance evaluation.

4.2.2 Types of machine learning systems

The type of machine learning system depends on the data; its availability, how much is labelled, how it is delivered to the system, and how it is evaluated. The most common types of machine learning systems are: supervised learning, semi-supervised learning, unsupervised learning, and reinforcement learning.

Supervised learning

Supervised learning is when your input data is labelled in such a way that you know the output, given a certain input (Bishop, 2007, p. 3). Supervised learning is most commonly found in classification, regression, and ranking problems (James et al., 2013, p. 26).

Semi-supervised learning

Semi-supervised learning is when your input contains labelled and unlabelled data, leading to some instances where you know the output and some where you do not (Burkov, 2019, p. 4). According to Mohri et al. (2018, p. 6), "the hope is that the distribution of unlabeled data accessible to the learner can help him achieve a better performance than in the supervised setting".

Unsupervised learning

Unsupervised learning is when you have data which is unlabelled, meaning that there is no corresponding output vector you can use to train with (Grus, 2019, p. 142). It is difficult to create a prediction on unlabelled data, but it has its uses. You can use clustering methods in order to find groups of data which are similar (James et al., 2013, p. 27), or to visualise the distribution of data throughout the set which may lead to further insight (Bishop, 2007, p. 3).

Unsupervised learning is very similar to data mining (a field within machine learning), where meaningful patterns are searched for in structured data in order to gain an advantage (Witten et al., 2016, p. 5).

Reinforcement learning

According to (Sutton & Barto, 1998), reinforcement learning involved a software agent that is concerned with finding suitable actions to perform in a given situation in order to maximise reward. Unlike supervised learning, the agent is not given examples of optimal outputs. It needs to discover those optimal outputs by trial and error (Bishop, 2007, p. 3). These software agents are also faced with the exploration versus exploitation dilemma, where they need to choose between exploring unknown actions or exploiting known actions (Mohri et al., 2018, p. 7).

4.2.3 Bias-variance trade-off

The generalisation error of a model can be expressed using three terms: bias, variance, and random noise, called the bias-variance decomposition (Bishop, 2007, p. 149).

In order to understand bias and variance, the concept of fitness needs to be defined. A model can overfit or underfit the training data. Everitt and Skrondal (2010) explain that a model overfitting is when the model corresponds to the data too exactly. This means that the model will end up fitting to the noise in the data. Underfitting, on the other hand, is

when the model cannot capture the underlying structure of the data (Everitt & Skrondal, 2010).

Bias is due to wrong assumptions made during the training process. A high bias leads to a model under fitting the data (Burkov, 2019, p. 10). Variance is the error owing to the model's sensitivity to outliers in the dataset. High variance leads to the model over fitting the data (Géron, 2019, p. 136).

The goal is to minimise both bias and variance, but it is not simple as models with low bias have high variance and vice versa (Hastie, Tibshirani, & Friedman, 2001, p. 38). As the complexity of the model increases, the variance increases while the bias decreases. The choice of model complexity is made to trade bias and variance in order to minimise testing errors (Bishop, 2007, p. 149).

There are many methods used to balance bias and variance of models. They are discussed throughout this chapter when relevant to the model being discussed. One general method to reduce variance is regularisation. According to (Burkov, 2019, p. 12), "regularisation is an umbrella-term that encompasses methods that force the learning algorithm to build a less complex model" - the most common types being L1 and L2 regularisation.

4.2.4 Curse of dimensionality

One aspect that contributes to overfitting is having many features, especially if you have more features than training examples. When there are a high number of features, working in a high dimensional space, computation time is drastically increased and there is the risk of the previously mentioned overfitting. This is known as the curse of dimensionality, first coined by Bellman (1961).

A simple way to reduce the dimensions of the input space is to make use of a technique called principal component analysis (PCA), which selects the dimensions that capture as much of the variation in the data as possible (Grus, 2019, p. 134). PCA helps eliminate noisy dimensions and group correlated dimensions, but should not be used on datasets with low dimensionality as it causes the model to have high bias. Other common methods are uniform manifold approximation and projection (UMAP), and outlier detection (Burkov, 2019, p. 13).

A model with a reduced dimensionality yields a more compact, more easily interpretable representation of the target concept, focusing the attention of the user on the most relevant variables (Witten et al., 2016, p. 308).

4.3 Classification

Classification problems deal with qualitative, or categorical, data. Predicting a qualitative response for an observation can be called classifying that observation (James et al., 2013, p. 129). There are many techniques one can use to classify observations. These techniques are known as classifiers. This section details what classification is, how it is performed, the common terms encountered, as well as discussing some popular classifiers.

Bishop (2007, p. 179) provides the formal goal of classification as "to take an input vector x and to assign it to one of K discrete classes C_k where $k = 1, ..., K$". These classes are taken to be disjoint, meaning that an input is only assigned to one class. Owing to the disjointed nature of the classes the data is divided into decision regions with decision boundaries grouping them together. If the decision boundary can be a straight line which fully separates the classes involved, then the data set is said to be linearly separable (Bishop, 2007, p. 179).

Classification problems can be binary, only having two classes, or multiclass (Géron, 2019, p. 102). Binary classification problems are far more common than multiclass. They are typically a pass/fail test and are much simpler than multiclass classification. This research is only concerned with the binary classification problem and thus does not discuss multiclass classification.

Sections 4.3.1 to 4.3.6 introduce and explain some basic concepts relating to common classifiers found in binary classification.

4.3.1 Artificial Neural Networks

An artificial neural network (ANN) is a computation model inspired by the structure of the brain. Where the brain has biological neurons and synapses for communication and calculation, an ANN has artificial neurons, which can transmit data to one another through connections (Shalev-Shwartz & Ben-David, 2014, p. 268).

McCulloch and Pitts (1943) were the first to create an ANN based on threshold logic. This paved the way for research and the development of ANNs. An ANN has a very basic structure composed of layers of neurons with connections between them. At its simplest form, an ANN has three layers: an input layer, a hidden layer, and an output layer. The number of neurons in these layers depends on the problem being solved and the data available. The data travels from the input layer to the output layer, wherein each neuron receives as input a weighted sum of the outputs of the neurons connected to its incoming edges (Shalev-Shwartz & Ben-David, 2014, p. 268). Figure 4.1 illustrates this description. In this case the

representation is a simple feed-forward neural network as there are no cycles involved.

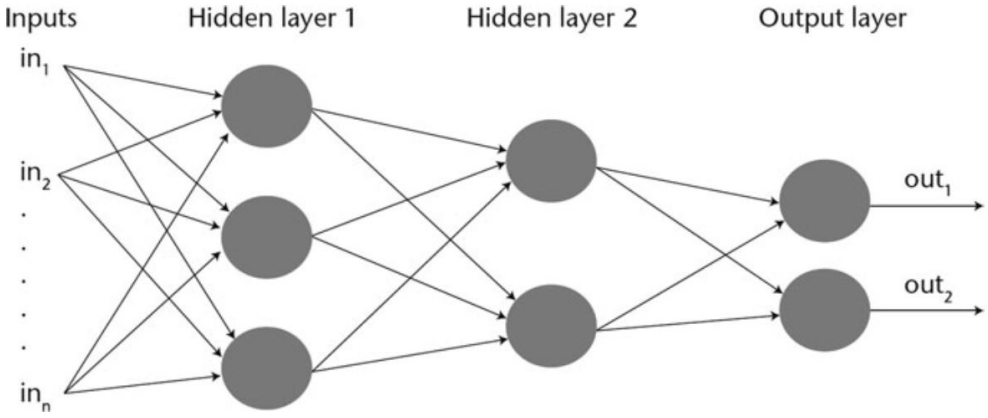


Figure 4.1: Structure of a simple neural network (Vardasca et al., 2018)

Neural networks are simply non-linear statistical models, but are very much black boxes in the sense that you do not know how they are doing what they are doing or why (Grus, 2019, p. 213).

No matter the size of the net, its structure, or how data is propagated through the net, all neural networks have these basic components: an input vector, neurons to receive input and transform it, connections with adjustable weights to transmit the data, and an output scalar.

The neurons themselves contain calculations in order to transform the data. They have a propagation function to compute the input based on the outputs of the previous neurons, a stored threshold, and an activation function that will decide whether the neuron fires and what data to pass to the next neuron. The initial activation for the network is the value of the input data, which then propagates through the network, as discussed, until arriving as an output.

4.3.1.1 Perceptron

The simplest form of any neural network is a perceptron, which is a single neuron that takes m binary inputs and produces a binary output (Grus, 2019, p. 213). It was first introduced by Rosenblatt (1957), where he introduced weights, which would represent the importance of each connection. A perceptron was a very simple machine that simply calculated the sum of all the inputs multiplied by the relevant weights, then underwent an activation function. For a perceptron it is a linear step function, and if that value was greater than a threshold for the neuron, it would fire (Bishop, 2007, p. 193). Figure 4.2 illustrates a perceptron taking inputs and transforming them, via the net input and activation functions, into an output.

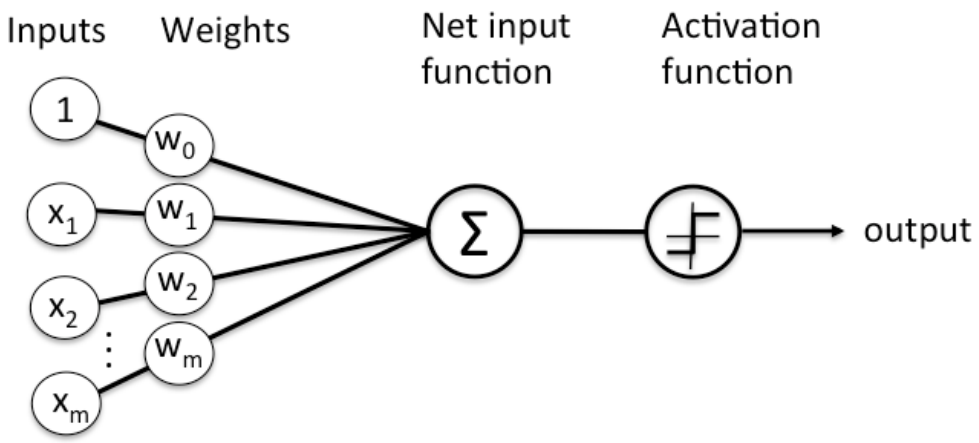


Figure 4.2: Structure of Rosenblatt's perceptron (Singh et al., 2017)

Multiple perceptrons are added together to form a layer of a larger network called a multilayer perceptron (MLP). Section 4.3.1.2 introduces the most basic layered neural network, a feed-forward neural network.

4.3.1.2 Feed-forward network functions

In the case of a simple three-layer feed-forward network, Bishop (2007, p. 227) describes the basic neural network model as a series of functional transformations. Equation 4.1 is M linear combinations of input variables $x_1, ..., x_D$.

$$a_j = \sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)} \tag{4.1}$$

Where $j = 1, ..., M$. Equation 4.1 is for the first layer of the network, hence the (1) superscript. The parameters $w_{ji}^{(1)}$ are the weights, which will be adjusted, and $w_{j0}^{(1)}$ are the biases.

Equation 4.1 is known as an activation and is transformed using some activation function given by Equation 4.2.

$$z_j = h(a_j) \tag{4.2}$$

The function $h(\cdot)$ is typically chosen to be a sigmoidal function, but other options such as: hyperbolic tangent, softmax function, or rectifier function are also used (Hastie et al., 2001, p. 392). The output activations from the first layer are transformed, corresponding to the second layer of the network, given by Equation 4.3.

$$a_k = \sum_{j=1}^{M} w_{kj}^{(2)} z_j + w_{k0}^{(2)} \tag{4.3}$$

Where $k = 1, ..., K$ and K is the number of outputs. Now, because our example only has three layers, the network output needs to be calculated. To do this, once again we need to apply an activation function to give appropriate network outputs y_k. This transformation is similar to Equation 4.2 and is given by 4.4.

$$y_k = \sigma(a_k) \tag{4.4}$$

Where $\sigma(\cdot)$ is the activation function, recall it typically being a sigmoidal function. The sigmoidal activation function is a logistical function transforming all the weighted sums to values between zero and one in a smooth fashion, with negative inputs close to zero and positive inputs are close to one (Shalev-Shwartz & Ben-David, 2014, p. 269). The sigmoid function is given by Equation 4.5. Figure 4.3 shows various activation functions.

$$\sigma(a) = \frac{1}{1 + \exp(-a)} \tag{4.5}$$

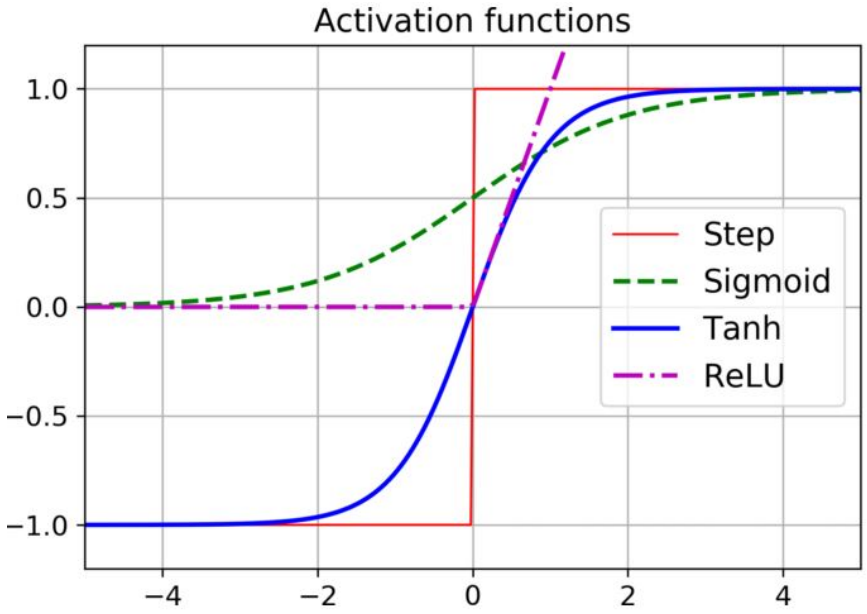


Figure 4.3: Various activation functions (Géron, 2019, p. 289)

When Equations 4.1 to 4.4 are combined, we get the overall network function using a sigmoidal activation function, given by Equation 4.6.

$$y_k(x, w) = \sigma\left(\sum_{j=1}^{M} w_{kj}^{(2)} h\left(\sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)}\right) + w_{k0}^{(2)}\right) \tag{4.6}$$

Where all the weights and biases have been grouped together into a vector w. This gives a nonlinear function that goes from a set of input variables, x_i, to output variables, y_k, controlled by a vector of adjustable parameters, w. Equation 4.6 can be simplified further by defining an additional parameter, $x_0 = 1$. This absorbs the bias found in Equation 4.1. The same can be done for Equation 4.3. This results in Equation 4.7 (Bishop, 2007, p. 229).

$$y_k(x, w) = \sigma\left(\sum_{j=1}^{M} w_{kj}^{(2)} h\left(\sum_{i=1}^{D} w_{ji}^{(1)} x_i\right)\right) \tag{4.7}$$

If the network had more layers, Equation 4.6 would look different. This purpose of this section is to illustrate a simple feed-forward neural network and its associated functions.

4.3.1.3 Training a neural network

In order to train a neural network, the previously mentioned weights vector, w, and the biases have their values altered in order to fit the data well. This is done by using a cost function - also known as the error function or loss function, to estimate the error involved with the current parameters. Then once the cost is known, some method, typically gradient descent, is used to update the weights in order to minimise the cost function for the neural network (Hastie et al., 2001, p. 395).

The neural network has its initial weights and biases chosen at random, but close to zero (Hastie et al., 2001, p. 397), and through the iterative process, called back propagation, it will adapt those values until an acceptable cost is reached.

4.3.1.4 Cost functions

There are many functions that can serve as a cost function. This research is only concerned with classification; thus, only cost functions associated with classification are discussed. Bishop (1995, Chap. 6) and Hastie et al. (2001, p. 395) detail the typical loss functions for neural networks: sum-of-squares error, and cross entropy.

Sum of squares

Typically used for regression, it still can be used with success in classification problems. Bishop (2007, p. 233) describes the sum-of-squares error, given by Equation 4.8.

$$S = \frac{1}{2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{4.8}$$

Where N is the number of training examples, y_i is the true output, and $\hat{y}_i$ is the estimated output. It can be seen that this loss function is very simple, as it takes the difference between true and estimated, then squares it for the entire training set.

It is very similar computationally to the mean squared error (MSE), Shalev-Shwartz and Ben-David (2014, p. 123) provide a definition of the mean squared error, given by Equation 4.9.

$$M = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{4.9}$$

The only difference here is that the MSE gets the average over the entire training set. For the case of multiclass classification, Bishop (1995, p. 229) gives a modified version of Equation 4.8, shown here as Equation 4.10

$$S = \frac{1}{2}\sum_{j=1}^{K}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{4.10}$$

Where K represents the number of classes.

Cross Entropy

Sum-of-squares error was a good cost function initially, but Simard, Steinkraus, and Platt (2003) found that the cross-entropy, or log loss, is faster and generalises better. Originally from information theory, cross-entropy gave an idea of how many bits were lost when transferring them. It works with differences between probability distributions. The cross-entropy error function is defined by Bishop (1995, p. 231) and given by Equation 4.11.

$$C = -\sum_{i=1}^{N}\{y_i \ln \hat{y}_i + (1 - y_i)\ln(a - \hat{y}_i)\} \tag{4.11}$$

Where N is the number of training instances, y_i is the true output, and $\hat{y}_i$ is the actual output. Equation 4.11 can be modified to include multiclass classification in the same way that Equation 4.10 modified Equation 4.8.

There are many other cost functions that can be used, and these same cost functions look different depending on the regularisation used as well as the activation function chosen. This section is aimed at introducing the cost function and giving basic examples of what they are and how they are used, and not at detailing the state-of-the-art exploratory functions being

used today. Section 4.3.1.5 details how the cost functions are used to adjust the weights of the neural network.

4.3.1.5 Back-propagation

Back-propagation is the core algorithm on how neural networks learn, first introduced by Rumelhart, Hinton, and Williams (1985). According to Hastie et al. (2001, p. 395) the generic approach to minimising the cost function is by a process called gradient descent, which is a part of back propagation. This gradient can easily be derived from the cost function using the chain rule for differentiation.

When minimising a function, one either arrives at a local or a global minimum. Stewart (2015) explains these concepts. A local minimum is the minimum value of a function within a given range, or near an arbitrary value, c. The global minimum is the minimum value of the function across its entire domain, or the extreme value of that function. Typically, the global minimum is not wanted as it is computationally hard and likely to over-fit to the data (Hastie et al., 2001, p. 395).

Gradient descent is an optimisation function that is used to find the minimum of a function. This is done by making proportional steps in the negative gradient direction until a minimum is found (Grus, 2019, p. 93).

Many different types of neural networks branch off from a simple feed-forward network. These can have many layers, called deep neural networks, and they can propagate the error back in different ways, but they are out of the scope of this research. The purpose of this section was to introduce neural networks and provide a basic understanding of how they function.

4.3.2 Support Vector Machines

The subject of support vector machines began in the late 1970's by Vladimir N. Vapnik (Burges, 1998). A support vector machine is defined by Cortes and Vapnik (1995) as a supervised learning model for two-group classification problems. It works by non-linearly mapping input vectors to higher dimensional space and creates decision lines (called hyper-planes) between them, making it a linear binary classifier. Originally it was a generalization of the maximal margin classifier, which meant it was limited to linearly separable data sets. It then was expanded on to be able to separate non-linear data sets (Cortes & Vapnik, 1995).

Support vector machines are very popular in many applications, mostly image processing, owing to their computing efficiency and high generalisation performance when compared to other methods (Cortes & Vapnik, 1995).

This section begins with an explanation of a hyperplane, then goes on to detail the maximal margin classifier and the extensions done of it, and finally discussing kernel tricks in order to map non-linear data.

4.3.2.1 Hyperplanes

Understanding a hyperplane is imperative to understanding support vector machines, as their descision boundaries are hyperplanes. James et al. (2013, p. 342) give a formal definition of a hyperplane as "In a p-dimensional space, a hyperplane is a flat affine subspace of dimension $p - 1$". In a more colloquial sense, for a two-dimensional space, found in a typical plot of x vs y, a hyperplane is a straight line, and in three-dimensional space it is a flat plane.

James et al. (2013, p. 342) gives the mathematical definition of a hyperplane in two-dimensional space by Equation 4.12:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0 \tag{4.12}$$

Equation 4.12 can be expanded into p-dimensional space, given by Equation 4.13.

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p = 0 \tag{4.13}$$

How this equation helps to classify a hyperplane is simple. If Equation 4.13 is not zero, it tells us that the point is either side of the hyperplane. If a hyperplane separates data perfectly according to its class labels it is known as a separating hyperplane. The goal of any support vector machine is to create an optimal hyperplane, meaning it is computationally efficient and perfectly separates classes (Cristianini & Shawe-Taylor, 2000, p. 93).

4.3.2.2 Maximal margin classifier

When you have perfectly separable classes there could be an infinite number of possible hyperplanes, this is because they can be moved ever so slightly or rotated. A way to get around this problem is to approach it from an optimisation stance with the maximal margin classifier.

The margin in the name refers to the smallest distance between the decision boundary (hyperplane) and the sample data (Bishop, 2007, p. 326). This classifier creates a hyperplane that maximises this margin, hence the name maximal margin classifier (Cristianini & Shawe-Taylor, 2000, p. 94). Figure 4.4 illustrates a maximal margin hyperplane and its associated margin. The maximal margin classifier only works on data which is linearly separable, and thus has very limited use (Mohri et al., 2018, p. 81). It is, however, the foundation that modern support vector machines were built on.

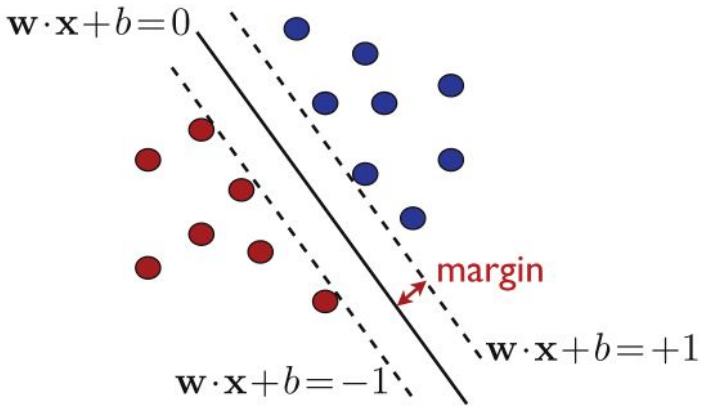


Figure 4.4: Maximal margin hyperplane (Mohri et al., 2018, p. 82)

As seen in Figure 4.4 there are four observations that fall on the dashed line, meaning that are equidistant from the hyperplane and lie on the margin. These observations are known as support vectors. The maximal margin hyperplane will move when these vectors move, this means it is dependent on the support vectors and not the other observations (Cortes & Vapnik, 1995; James et al., 2013, p. 345).

These support vectors are found using optimisation. This optimisation is quadratic in nature and uses Lagrange transforms to identify points that satisfy the Karush-Kuhn-Tucker (KKT) conditions (Cortes & Vapnik, 1995; Bishop, 2007, p. 330). This optimisation removes all the points besides the support vectors.

Once the model is trained, a significant proportion of the data points can be discarded and only the support vectors retained (Bishop, 2007, p. 330). This reliance on a small set of the observations gives this classifier great generalization and computing efficiency (Cortes & Vapnik, 1995).

Non-separable cases

Recall that maximal margin classifiers cannot work unless the data is separable. Support vector classifiers are extensions of maximal margin classifiers, which use soft margins in order to approximate a hyperplane that almost separates the data (James et al., 2013, p. 347). This allows classifications of non-separable data and provides greater robustness to individual observations as well as better classification of most of the observations.

This modification needs to allow the misclassification of observations. This was previously unachievable owing to the error function being infinite when an observation was misclassified. The error function needs to be adapted to be a linear function of the distance away from the

margin (Bishop, 2007, p. 331). This error function has a parameter C, which controls how much of a penalty must be endured. This helps to control the bias-variance trade-off of the support vector classifier (James et al., 2013, p. 351).

4.3.2.3 Kernel functions and non-linear data

Section 4.3.2.2 introduced the foundation on which support vector machines are built. It began with maximal margin classifiers which were limited by non-separability in the data, that was resolved by modifying the optimisation function and error function to give support vector classifiers. There is, however, one last hurdle to cover, a non-linear decision boundary.

To do this, support vector classifiers were once again expanded upon to create support vector machines, which are capable of transforming input observations into a higher feature dimension; they do this using kernels (James et al., 2013, p. 355).

Kernels are, very simply, the functions that allow the mapping of input of data to higher dimensional space. Their full name in the context of machine learning is the positive-definite kernel and were introduced by James Mercer (Cristianini & Shawe-Taylor, 2000, p. 99). The kernel method, a mathematical shortcut, alters the inner products, or dot product, of observation vectors in order to map them to higher dimensions (Cristianini & Shawe-Taylor, 2000, p. 32). The beauty in it is that you can map inputs to infinite dimensions and still be able to find hyperplanes. Kernel functions replace the standard dot product, which is how the data appears in training (Burges, 1998, p. 138).

James et al. (2013, p. 355) define the inner product of two vectors x_i and $x_{i'}$ in Equation 4.14.

$$\langle x_i, x_{i'} \rangle = \sum_{j=1}^{p} x_{ij} x_{i'j} \tag{4.14}$$

Where p is the number of features. This inner product can be generalised in terms of an arbitrary kernel function, k, by Equation 4.15.

$$K(x_i, x_{i'}) \tag{4.15}$$

This allows the use of kernel tricks to map the inputs to higher dimensions. The most common kernel tricks used are; linear, polynomial, and radial-bias function.

Linear kernel

Linear kernels are by far the simplest as the kernel function just gives back the support vector classifier. No mapping needs to be made as a linear decision boundary can already be made.

This means that the kernel has not altered the dot product at all, and the support vector classifier optimises as if there was no kernel function applied.

Polynomial kernel

Polynomial kernels allow decision boundaries to be drawn that bend. The extent of bending depends on the degree of the polynomial (Burges, 1998). What is happening is essentially fitting a support vector classifier in a higher-dimensional space involving polynomials of degree d, rather than in the original feature space.

This is achieved by replacing the normal dot product with Equation 4.16 (Burges, 1998, p. 142).

$$K(x_i, x_{i'}) = (1 + \sum_{j=1}^{p} x_{ij} x_{i'j})^d \tag{4.16}$$

Where d is the degree of the polynomial, $d > 1$. This leads to a much more flexible decision boundary that will more accurately classify data that is not linearly separable. When a support vector classifier has a kernel function performed on it to map it to higher dimensional space it is now known as a support vector machine (James et al., 2013, p. 356).

RBF kernel

The radial bias function, or simply radial, kernel is another popular choice. It is a Gaussian kernel (Bishop, 2007, p. 334) and takes the form of Equation 4.17. Radial kernels demonstrate very local behaviour, meaning that only nearby training observations influence the class label of a test observation (James et al., 2013, p. 357).

$$K(x_i, x_{i'}) = exp(-\gamma \sum_{j=1}^{p} x_{ij} x_{i'j})^2 \tag{4.17}$$

Where γ is a positive constant. Figure 4.5 shows a data set separated using a radial kernel. Radial kernels are unique in that their feature map dimension is infinite (Mohri et al., 2018, p. 131).

Support vector machines are very fast learners and, using special kernel tricks, are able to separate data into classes very accurately. Vapnik (1999) outlines some general properties of SVMs:

- When constructing the SVM using the optimisation problem, a unique solution can always be found.
- When constructing the decision rule, a set of support vectors is obtained.

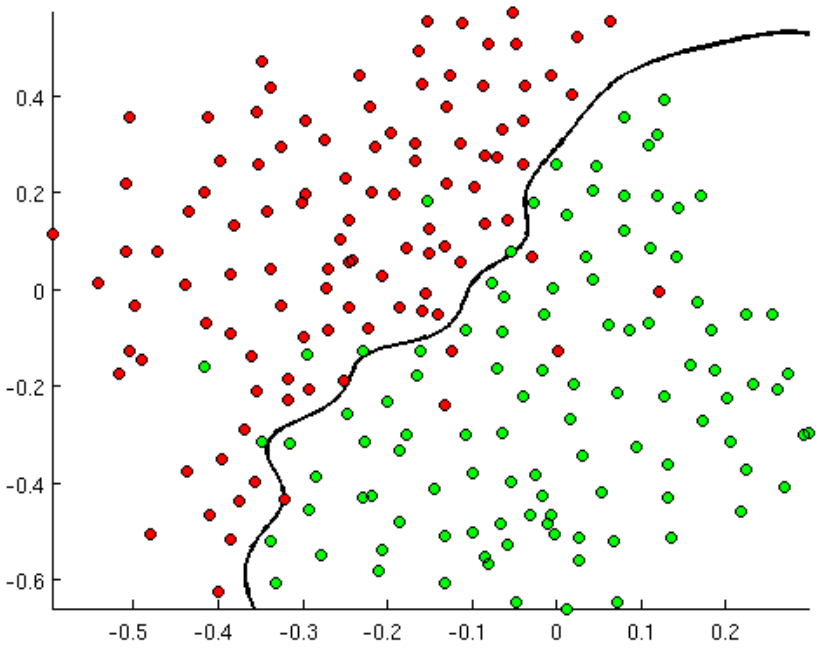


Figure 4.5: Decision boundary drawn by radial kernel with $\gamma = 100$ (A. Ng, 2010)

- SVMs are very fast at constructing the learning process.
- Implementing a new set of decision functions is as simple as changing the kernel function.

SVMs have many applications with classification, they are incredibly adept at classifying images. They are not limited to classification. Regression using SVMs can be done accurately, but it is not in the scope of this research and will therefore not be discussed here.

4.3.3 Naïve Bayes

A naïve Bayes classifier is a probabilistic classifier based on Bayes theorem that assumes that input variables are all independent (Friedman, Geiger, & Goldszmidt, 1997). The naïve Bayes classifier, which looked at text indexing using a probability based approach was first introduced by Maron (1961). Bayes' theorem was first established by Thomas Bayes in his paper "Essay towards solving a problem in the doctrine of chances" in 1764, and was rediscovered by Pierre-Simon Laplace, who showed its broad applicability (Bishop, 2007, p. 21). Bayes' theorem is stated in Equation 4.18 and is the foundation for all probability theory (Vapnik, 1992, p. 119).

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{4.18}$$

Where A and B are two events and the $P(B) \neq 0$. $P(A)$ and $P(B)$ are the probabilities of observing A and B independently of each other. $P(A|B)$ and $P(B|A)$ are conditional properties and are the probability of the event occurring if the other event is true.

Bayesian classifiers use Bayes' theorem in order to assign the most likely class based in the feature vector (Rish, 2001). The assumption that all the input variables are independent may not always be the case but it helps to simplify the estimation and allow it to compete with much more sophisticated classifiers (Hastie et al., 2001, p. 211).

The classifier will create a hypothesis for each class and chose the hypothesis with the highest probability of being the correct mapping, this is known as the maximum a posteriori probability, or MAP (Mohri et al., 2018, p. 297). Rish (2001) defines this function as Equation 4.19. It is calculated by applying Bayes theorem, expanded on using the chain rule and simplified.

$$h(x) = \arg\max_{i} P(X = x|C = i)P(C = i) \tag{4.19}$$

Where i is the class. Naïve Bayes works best for two cases: when the variables are completely independent, and when the variables are functionally dependent. They perform relatively poorly when in between these extremes (Rish, 2001). Owing to the fact that it treats the variables as independent, the naïve Bayes classifier does not suffer as badly from the curse of dimensionality by not requiring the data set to scale exponentially with features. It has the advantage of being very simple with a high accuracy and an optimal Bayes's error (Cruz-Ramírez et al., 2013).

4.3.4 k-Nearest Neighbour

The k-Nearest Neighbour algorithm is a non-parametric method used for both classification and regression (Bishop, 2007, p. 123), using instance based learning, or lazy learning (Witten et al., 2016, p. 78).

It forms part of a larger family of algorithms called Nearest Neighbours. Shalev-Shwartz and Ben-David (2014, p. 258) explain Nearest Neighbour algorithms as the simplest of all machine learning algorithms where the training set is memorised and the prediction is made based on the closest neighbours, from the training set, to the new instance. This memorisation of the dataset means that no model needs to be fit (Hastie et al., 2001, p. 463). This results in a very fast algorithm that works well even with large datasets (Shalev-Shwartz & Ben-David, 2014, p. 258).

The only thing a Nearest Neighbour algorithm needs is a way to calculate distance and having the assumption that points close together are similar (Grus, 2019, p. 151). Computing

distances with numerical values is as easy as subtracting the values, but nominal values, like colour, provide more of a challenge. Some attributes will be more important than others, and this is usually reflected in the distance metric by some attribute weighting (Witten et al., 2016, p. 78). This derivation of suitable attribute weights from the training set is a key problem in instance-based learning.

A more formal definition of the k-Nearest Neighbour classifier is given by Hastie et al. (2001, p. 463) as: Given a query point x_0, we find the k training points $x_{(r)}$, $r = 1, ..., k$ closest in distance to x_0, and then classify using majority vote among the k neighbours. For instances containing continuous variables, Euclidean distance is used. This results in Equation 4.20 being how the distance is calculated. Most applications will convert features into continuous variables and normalise them, so Euclidean distance can be used. However, for the cases when that this cannot happen, other distance measuring methods can be employed, one such method is Hamming distance.

$$d_{(i)} = ||x_{(i)} - x_0|| \tag{4.20}$$

In order to normalise the features, they are standardised to have a mean of zero and a variance of 1. k-Nearest Neighbour does well when the decision boundary is most irregular, and the classes have many possible prototypes (points used for classification). Once the nearest neighbours have been found Bayes' theorem is used to find the most probable class to which the new instance belongs (Bishop, 2007, p. 125).

The special case of k-Nearest Neighbour is when k=1, that is simply known as the nearest-neighbour. In this case the class label chosen for a new observation is simply the closest instance from the training set. Cover and Hart (1967) famously showed that as the number of points in a data set reaches ∞ the error rate never exceeds twice that of an optimal classifier, meaning that it is never more than twice the Bayes error rate. The upper bound error rate is given by Equation 4.21 (Cover & Hart, 1967).

$$R^* \leq R_{kNN} \leq R^*\Big(2 - \frac{MR^*}{M-1}\Big) \tag{4.21}$$

Where R^* is the Bayes' error rate, R_{kNN} is the k-NN error rate, and M is the number of classes in the problem.

The choice of k is an important choice. Bishop (2007, p. 124) explains that a large k leads to over-smoothing (high bias, low variance) and, subsequently, a small k leads to noisy estimates (high variance, low bias). A general rule for binary classifiers is that you choose k to be an odd number, and an optimal k value can be attained using a bootstrap method (Hall, Park, & Samworth, 2008).

4.3.5 Decision tree

A decision tree represents possible decision paths that can be taken and an outcome for each path. It does this in a tree like structure (Grus, 2019, p. 201). When used as a predictor, a decision tree predicts a label associated with an instance by traveling from a root of a node to a leaf (Shalev-Shwartz & Ben-David, 2014, p. 250).

Decision trees can be used for classification and regression. There are some popular frameworks for decision tree learning, namely: CART, ID3, and C4.5 (Bishop, 2007, p. 663).

When the target variable is a discrete value, the tree is a classifier; the leaves represent class labels and the root-to-leaf path represents the conjunctions of features to get to that class label.

Decision trees have some advantages such as: the process to get to a label is completely transparent, they are fast learners, easy to evaluate, and they work well with incomplete data (Mohri et al., 2018, p. 224). The construction of a tree is very important as they can very easily over-fit the data, and thus generalise very poorly (Grus, 2019, p. 202).

4.3.5.1 Tree construction

Firstly, the root node needs to be decided, then each internal node. A node that is not the root or a leaf, holds the value for an input feature. These internal nodes, known as decision nodes, need to have some form of decision made in order to split left or right into the child node.

James et al. (2013, p. 315) explain that growing a tree involves a recursive binary splitting function. The splits are decided by some threshold parameter θ, and the regions are subdivided accordingly. Within each region there is a model that will assign a label. Bishop (2007, p. 664) discusses how, for CART, θ is calculated; for regression you aim to minimise the sum-of-squares error, and for classification there is a choice between the classification error, the Gini index, or the cross-entropy. The problem lies in the computational requirements to minimise these functions. An easier solution, which is nearly as good, is to create each node one at a time based on θ using top-down greedy optimisation (Bishop, 2007, p. 665).

The classification error is a good classification alternative to the sum-of-squares error for regression. James et al. (2013, p. 315) explain that we plan to assign an observation in a given region to the most commonly occurring class of training observations in that region. Simply put, the classification error rate is the fraction of the training observations in that region that do not belong to the most common class. Classification error rate is given by Equation 4.22.

$$E = 1 - \max_k(\hat{p}_{mk}) \tag{4.22}$$

Where $\hat{p}_{mk}$ represents the proportion of training observations in the mth region that are from the kth class. This method is not ideal for growing trees and the Gini index and cross-entropy are preferred in some cases. The Gini index is a measure of the node's purity and is given by Equation 4.23 (Hastie et al., 2001, p. 309).

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{4.23}$$

Where K is the number of classes, because G is an indication of the variance across the classes, a small G is an indication that nodes contain observations mostly from one class. An alternative to the Gini index is the abovementioned cross entropy. It is also a measure of variance and is given by Equation 4.24 (Bishop, 2007, p. 666).

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk} \tag{4.24}$$

James et al. (2013, p. 316) states that when building a classification tree, owing to their node-purity sensitivity, either the Gini index or cross-entropy are typically used. If the goal is prediction accuracy of the final pruned tree, classification error rate is preferable.

4.3.5.2 Tree pruning

Given the top-down greedy approach to creating trees, the next logical line of questioning is when to stop growing the tree because a large tree will over-fit the data and a small tree might not capture the important structure.

Bishop (2007, p. 665) explains that is it common to create a large tree and stop it based on the number of data points with associated leaf nodes, then prune back the tree. This pruning is based on a criterion that balances residual error against a measure of model complexity.

This method of pruning is called cost-complexity pruning, or weakest link pruning, and is explained by Hastie et al. (2001, p. 308), Bishop (2007, p. 665), and James et al. (2013, p. 313). First a subtree, $T \subset T_0$, is defined and it is any tree that can be obtained by pruning T_0, that is, by collapsing any number of its internal (non-terminal) nodes. The cost complexity criterion is given by Equation 4.25.

$$C_\alpha(T) = \sum_{m=1}^{|T|} Q_m(T) + \alpha|T| \tag{4.25}$$

Where $|T|$ indicates the number of terminal nodes of the tree T and $Q_m T$ is the previously discussed measure of the node impurity of the region, which can either be replaced by Equation 4.23 or 4.24 in the case of classification. The tuning parameter α controls a trade-off between the complexity of the subtree and its fit to the training data. As α increases from zero, branches are pruned form the tree in a nested and predictable way. The value of α can be selected using cross-validation.

This section gave a brief overview of some of the core concepts of decision trees using the lens of the CART framework. Rokach and Maimon (2015) provide a far more comprehensive coverage of decision trees and expands the concepts addressed here fully.

4.3.6 Ensemble methods

Ensemble methods are general techniques in machine learning for combining several predictors to create a more accurate one (Mohri et al., 2018, p. 145). This idea was first proposed in the late 1970s by Tukey (1977), but the largest progress was made by Hansen and Salamon (1990) when they suggested an ensemble of similarly configured neural networks to improve the predictive performance of a single one.

A typical ensemble method contains the following components: a training set, a base inducer, which creates the initial generalized relationship between the input and output data, a diversity generator, which is responsible for generating the various classifiers to be used, and a combiner, which combines the classifications of the various classifiers (Rokach, 2010). The two main methods for combining the outputs of the base classifiers are weighting and meta-learning methods. According to Rokach (2010), weighting methods are useful if the base-classifiers perform the same task and have comparable success. Meta-learning methods are best suited for cases in which certain classifiers consistently classify correctly, or consistently misclassify certain instances.

There are many ensemble methods, which depend on the problem, the approach, and the available data. This research is only considering classification; furthermore, only one example each of dependent and independent learning is given. The two examples discussed in this section are: boosting (dependent) and bagging (independent).

4.3.6.1 Boosting

The approach for dependent learning used here is called model-guided instance selection. Provost and Kolluri (1997) discuss that, with this approach to dependent learning, classifiers that were constructed in previous iterations are used for manipulating the training set for the following iteration. They only learn from the misclassified cases. According to (Rokach, 2010)

the most well-known model-guided instance selection is boosting, also known as adaptive resampling and combining (arcing).

Boosting works by running a weak learner, which is a classifier that is slightly better than chance, repeatedly on various training data. The classifiers produced are then combined into a single composite classifier, which has a higher accuracy than the individual weak learning classifiers would have had (Shalev-Shwartz & Ben-David, 2014, p. 130). A popular iterative method of boosting, first introduced by Freund and Schapire (1996), is called adaptive boosting (AdaBoost).

Bishop (2007, p. 657) indicates that the idea behind AdaBoost is to give a higher focus of harder-to-classify patterns. This focus is dictated by some weight assigned to all the patterns in the training set. For the purpose of explanation, the AdaBoost.M1 algorithm is described. In the sequence of weak classifiers $G_m(x), m = 1, 2, ..., M$ is combined through a weighted majority vote given by Equation 4.26 (Hastie et al., 2001, p. 338).

$$G(x) = sign\left(\sum_{m=1}^{M} \alpha_m G_m(x)\right) \tag{4.26}$$

Where α_m is the weight associated with classifier $G_m(x)$. This is done so that a higher influence is given to the more accurate classifiers in the sequence. Hastie et al. (2001, p. 338) details these steps, which consist of applying weights $w_1, w_2, ..., w_N$ to each of the training observations $(x_n, y_n), n = 1, 2, ..., N$.

- Initially, all of the weights are set to $w_n = \frac{1}{N}$, and the training is done in the usual manner.

- For the remaining iterations, $m = 2, 3, ..., M$, the weights are individually modified and the classification algorithm is reapplied to the weighted observations.

- At step m, the misclassified observations from step $m - 1$, by classifier $G_{m-1}(x)$, have their weights increased, whereas the weights are decreased for those that were classified correctly. Through iteration this forces the classifiers to focus on the difficult-to-classify observations.

- Once the iterations are complete, a prediction is made using the final model, given by Equation 4.26.

Bishop (2007, p. 658) gives an algorithmic description of the previously mentioned process. Let the weighted error function be given by Equation 4.27.

$$J_m = \sum_{n=1}^{N} w_n^{(m)} I(y_m(x_n) \neq t_n) \tag{4.27}$$

Where $y_m(x_n)$ is the predicted target value of x_n by classifier y_m, t_n is the target for observation x_n, and $I(y_m(x_n) \neq t_n)$ is an indicator function that gives a 1 or 0 depending on whether the function is true or false. The weighted measures of the error rates of each of the base classifiers on the data set are given by Equation 4.28.

$$\epsilon_m = \frac{\sum_{n=1}^{N} w_n^{(m)} I(y_m(x_n) \neq t_n)}{\sum_{n=1}^{N} w_n^{(m)}} \tag{4.28}$$

From this the updating weighting coefficients, α_m, can be evaluated using Equation 4.29.

$$\alpha_m = \ln \left\{ \frac{1 - \epsilon_m}{\epsilon_m} \right\} \tag{4.29}$$

The weighting coefficients, w_n, can now be updated for the next classifier using Equation 4.30.

$$w_n^{(m+1)} = w_n^{(m)} \exp\{\alpha_m I(y_m(x_n) \neq t_n)\} \tag{4.30}$$

It can be seen here that more weight is given to accurate classifiers. The final model can be used to make predictions once all M classifiers have been iterated. That final classier is given by Equation 4.26.

According to Rokach (2010), AdaBoost produces a combined classifier with a significantly lower variance than the base weak learner. The AdaBoost algorithm has been improved since its initial configuration in a number of ways, some being to reduce computation cost (Friedman et al., 1997), to lower the risk of over fitting (Breiman, 1999), changed in order to make use of parallel processing (Merler, Caprile, & Furlanello, 2007), and to become better at dealing with noise (Zhang & Zhang, 2008) to name a few.

4.3.6.2 Bagging

First introduced by Breiman (1996), bagging, or bootstrap aggregating, is an ensemble algorithm using independent learning designed to improve the accuracy and stability of machine learning algorithms. This method helps reduce variance and helps avoid over fitting. It is typically applied to decision trees.

Just like boosting, bagging improves accuracy by producing a composite model derived from the same inducer and using a majority voting to make decisions. However, Rokach (2010) states that in bagging, each instance is chosen with equal probability and the inducer needs to be an unstable learner. Recall that with boosting, the instances are chosen proportional to their weights and an unstable learner is not needed; only the error rate of every classifier should be kept below 0.5.

According to Breiman (1996), given a training set T consisting of N cases, bagging generates m new training sets of equal size by sampling uniformly from T with replacement. With replacement some observations are repeated, which is good for unstable learners as this creates only a slightly different model. This method of sampling by replacement is called bootstrapping (Efron & Tibshirani, 1994). The classifiers are then trained on these bootstrap distributions and a new observations prediction is based on the majority vote from the trained classifiers, made more reliable as more votes are taken into account (Witten et al., 2016, P. 353).

As already discussed, bagging works well with unstable learners. Breiman (1994) heuristically studied instability and found that neural nets, classification and regression trees, and subset selection in linear regression were unstable, while k-nearest neighbor methods were stable.

Random Forest

Bagging is seen as a special case of a random forest (RF) ensemble, also known as random subspace. First properly introduced by Breiman (2001), RF uses many individual, unpruned decision trees with an input parameter N representing the number of input variables that will be used to determine the decision at a node of the tree. This input N should be much less than the number of attributes in the training set.

One important advantage of the random forest method is its ability to handle a very large number of input attributes as well as it being very fast (Skurichina & Duin, 2002). According to Hastie et al. (2001, p. 587), the performance of random forests is very similar to boosting, and they are simpler to train and tune. Consequently, random forests are popular, and are implemented in a variety of packages. They have lower variance than the individual trees, which are noisy and thus benefit from averaging, and have the same bias as the individual trees. RF also benefits by removing the tendency of decision trees to over fit training data (Hastie et al., 2001, p. 595).

4.4 Model selection and validation

According to Shalev-Shwartz and Ben-David (2014, p. 144), model selection is the process of choosing the best algorithm and parameters for the problem at hand. Incorrect model selection can lead to the data being overfitted or underfitted. Model selection can be done by creating validation sets, which are subsets of the data used for testing the trained model. If there is enough data, a third set can be used to avoid overfitting the validation step with many iterations (Bishop, 2007, p. 32).

An example of how the model selection can impact the prediction accuracy is given by Figure 4.6. This illustrates how the choice of degree of polynomial affects the curve fitting process.

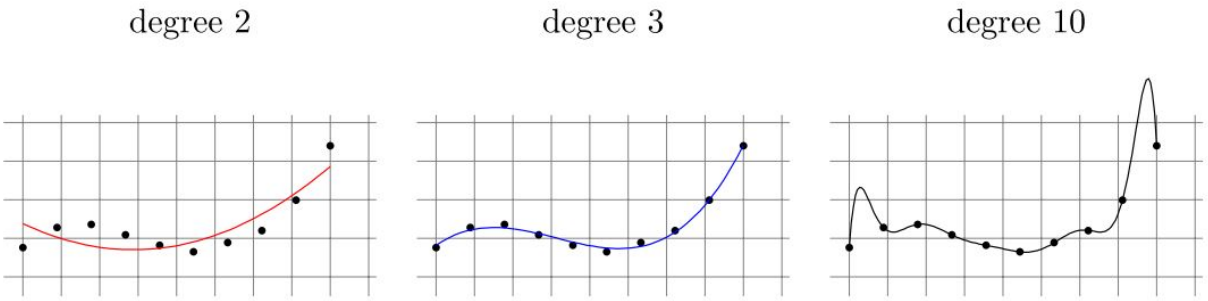


Figure 4.6: How differing degrees of polynomials affect accuracy (Shalev-Shwartz & Ben-David, 2014, p. 145)

It can be seen here that a polynomial of degree three best fits the data, and degree ten is an over-fit which will cause errors when testing unseen data. A validation set is needed because the training error is not a good estimate for the performance of a model as it can be low owing to overfitting, and will thus perform poorly on unseen data (Witten et al., 2016, p. 147). Hastie et al. (2001, p. 220) state that the test error, or generalisation error, is the prediction error over an independent test sample or the validation set. This generalisation performance is a strong indicator of how the predicting capability of the model relates to independent test data.

4.4.1 Validation

The idea behind validation is simple. Split the dataset into at least two parts; training and testing sets, which are each representative of the underlying problem. It is very important that the testing set not be used in any way to train the model, as this nullifies the entire point of splitting the dataset.

As mentioned at the beginning of Section 4.4, if a large enough dataset is available, the ideal validation would involve three datasets: a training set used to create the model(s), a validation set to tune hyper-parameters or select a model, and finally a testing set to measure the error rate of the final optimised model (Witten et al., 2016, p. 149).

Depending on the amount of data available, there are several methods one can use when wanting to do validation. This section details the more common ones found in machine learning problems: hold-out, cross validation, and bootstrapping.

4.4.1.1 Holdout

Holdout is the simplest method of validation. The data is split into two parts, a training set and a hold out (testing) set. Although this is a straight-forward approach, it has a major draw back. Witten et al. (2016, p. 149) explain that in order to find a good classifier, as much of the data as possible needs to be used for training; and to obtain a good error estimate, as much data as possible needs to be used for testing. A general rule of thumb for holdout is to use a third of your dataset for validation (Witten et al., 2016, p. 152).

The sets should be made up of random samples only if you know the observations are independent. If the data is known to be dependent, in the case of stock prices, that should be order chronologically (Shalev-Shwartz & Ben-David, 2014, p. 146). The datasets should be stratified, with even distribution of features between the sets. The only times where a non-stratified approach is taken is when the distribution of the features is expected to be very different from that of the training set (Witten et al., 2016, p. 152).

4.4.1.2 Cross-validation

Although a negative risk of hold out, where the portion selected for testing could not be representative of the problem, is mostly solved with stratification, it can still be better mitigated in other ways. A general way to mitigate this bias is to repeat the process of stratified hold out repeatedly with different random samples (Witten et al., 2016, p. 152).

A very simple and widely used method of doing this is called cross-validation, or more generally, k-fold cross-validation (Hastie et al., 2001, p. 241). This approach does not waste data by holding it separately in a holding set; rather, it uses all the data for training and testing. It was originally defined by Geisser (1975) as consisting of averaging several holdout estimators of the risk corresponding to different data splits, or folds. In other words, the dataset is partitioned into k folds; then, for each fold a model is trained on the union of the other folds and its error is tested on the fold. The average of these errors is then the estimate of the true error (Shalev-Shwartz & Ben-David, 2014, p. 150).

The number of folds is not set in stone as it depends on how much data you have and how computationally intense the model is to train, as higher number of folds increase overall training time. Through extensive tests on numerous different datasets using different learning techniques, the best estimate of error is using ten folds, which is the standard (Witten et al., 2016).

When the data is very sparse, the best choice is to use an exhaustive data-splitting technique. The most popular is a special case of k-fold cross-validation called leave-one-out cross-validation (LOOCV), discussed by Arlot and Celisse (2010) as a technique where k is equal to the number of training examples and each data point is successively left out from the sample and used for validation. This method is computationally very expensive and was used as far back as Picard and Cook (1984). LOOCV has two large advantages: firstly, the greatest possible amount of data is used for training in each case, increasing the chances that the classifier is accurate: and secondly, there is no random sampling involved, meaning that it is completely deterministic (Witten et al., 2016, p. 154).

4.4.1.3 Bootstrapping

Bootstrapping was already touched on before in Section 4.3.6.2, but is discussed in more depth here. Recall that the idea of the bootstrap is to sample the dataset with replacement to form a training set (Efron & Tibshirani, 1994). Previously, when selecting data for sets, once it was selected it was not able to be selected again for another set, but now you take a dataset of n instances and sample n times, with replacement, to give another dataset of n instances. You repeat this until you have m bootstrapped datasets (Hastie et al., 2001, p. 250).

There is always a possibility that a part of the training data does not end up in a boostrapped dataset. These instances are used as a testing set. This probability of a training example being picked is $\frac{1}{n}$, therefore for a large number of samples, n, we have Equation 4.31.

$$C = \left(1 - \frac{1}{n}\right)^n \approx e^{-1} = 0.368 \tag{4.31}$$

Where C is the chance of a sample not being picked for a bootstrapped dataset. Thus, for a reasonably large dataset, the test set will contain about 36.8% of the instances and the training set will contain about 63.2%. To compensate for only 63.8% of the instances being present in the training set, the test-set error rate is combined with the resubstitution (training) error on the instances in the training set. Finally, the procedure is repeated several times, with different replacement samples for the training set, and the results are averaged (Witten et al., 2016, p. 156). It is evident that the bootstrap method is ideal for small

datasets, but because it factors in training error it can underestimate the error rate owing to overfitting (Hastie et al., 2001, p. 252).

4.4.2 Performance metrics

Once proper validation has been done, the accuracy of the model needs to be measured. In order to measure the accuracy and error rates of models you need to know how many predictions made were correct and how many were incorrect. If the model performs well when predicting labels it is said to generalise well.

For regression this is simple as you compare how closely the model predicted the target values. For classification, however, it is slightly more complicated. The metrics used for assessing the performance of a classifier are given by Burkov (2019) as confusion matrix, accuracy, precision, recall, f-measure, and area under ROC curve.

4.4.2.1 Confusion matrix

A confusion matrix, or error matrix, is a way to organise the predictions of the model versus the target classes (Stehman, 1997). Powers (2011) explains that each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class. This makes it easy to see whether the model is confusing the classes. For the sake of simplicity, Table 4.1 shows a confusion matrix for binary classification.

		Predicted class	
		True	**False**
Actual class	**True**	TP	FN
	False	FP	TN

Table 4.1: Confusion matrix for binary classification

From the confusion matrix, the terminology for predictions consists of four categories: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). These are explained by Witten et al. (2016, p. 164) as follow: TP and TN are correct classifications; FP occurs when the prediction was true but the target was false; and a FN occurs when the prediction was false and when the target was true. From these four terms numerous measures of performance can be calculated.

4.4.2.2 Measures of accuracy

Powers (2011) defines some formulas used for assessing the performance of a classifier. They are: precision, recall, accuracy, f-measure, and AUC.

Precision, or specificity, is the portion of the predicted positive cases that is positive. It is given by Equation 4.32 and is known as true positive accuracy.

$$Sp = \frac{TP}{TP + FP} \tag{4.32}$$

Recall, or sensitivity, is the portion of the positive values that is correctly predicted as positive. It is given by Equation 4.33 and is used in ROC analysis as a true positive rate.

$$Sn = \frac{TP}{TP + FN} \tag{4.33}$$

Accuracy is the ratio of the correctly classified instances and the total number of predictions, useful when errors in predicting all classes are equal (Burkov, 2019, p. 16). It is given by Equation 4.34.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.34}$$

Another measure of accuracy is the F-measure, or F_1 score. It makes use of precision and recall. It is the harmonic mean between the two (Powers, 2011) and is given by Equation 4.35.

$$F_1 = 2 \cdot \frac{Sp \cdot Sn}{Sp + Sn} \tag{4.35}$$

James et al. (2013, p. 150) explain that the receiver operator characteristics (ROC) curve is a common method used to assess the performance of classifiers. It uses the true positive rate (recall) and false positive rate (proportion of negative examples predicted incorrectly) to create a graph that gives an image of performance by plotting the true positive rate versus the false positive rate. Burkov (2019, P. 17) notes that ROC analysis can only be done on classifiers that return some confidence score or probability: logistic regression, neural networks, and decision trees.

The overall performance of a classifier is given by the area under this ROC curve, denoted as AUC (Powers, 2011). Witten et al. (2016, p. 150) state that an AUC score of 0.5 is no better than chance and the closer to 1 the AUC, is the better the classifier is.

The choice of performance evaluation metrics depends on the problem at hand as well as the models used. Problems sensitive to false negatives weigh certain metrics higher than others.

4.5 Conclusion

The purpose of this chapter was to establish an understanding of machine learning and the algorithms relating to classification problems. It began with a definition of machine learning, how machine learning algorithms learn, the types of machine learning systems and common terminology found within the paradigm.

The various types of machine learning systems were defined in order to help outline more clearly the problem the research undertakes. Machine learning tasks were discussed with a focus on classification, owing to the fact that the problem being solved is a binary classification problem.

Once the data is collected and features are extracted, the selected features need to be split into training and testing samples via some form of sampling technique. This is done so that the metrics calculated from the resulting confusion matrix accurately represent the ability for the algorithm to generalise. If not done, the result is useless as the model could have severely over fit the data resulting in a perfect score but having poor generalisation ability for real world tasks.

When a model overfits data it is said to have high variance, and when a model underfits the data it is said to have high bias. These two concepts need to be balanced as they have an inverse relationship. The goal is to minimise both as this produces a good generalising model, which still understands the underlying structure of the data.

The algorithms considered for this research are ones that are commonly found in classification problems. They are: artificial neural networks, support vector machines, naïve Bayes, k-nearest neighbour, decision trees, as well as various ensemble techniques. This gives the research a variety of classifiers which work in different ways, allowing for a more rounded examination of the problem.

The following chapter covers the the creation of the conceptual model. The model is created from the literature discussed thus far and contains various components, each representing a different aspect of the model.

Chapter 5

Model development

5.1 Introduction

This chapter describes a model for the automated detection of breast cancer using thermal images and machine learning in a mobile environment. Olivier (2009) states that a model captures the essential aspects of a system while ignoring the non-essential aspects. It can be used as blueprint for a new system or to evaluate an existing system. Sokolowski and Banks (2010, p. 2) describes a system as the construct or collection of elements that produce results not obtainable by the individual elements. It is the subject of the model development. The simplest abstract description of the system is the model.

A conceptual model is typically a graphic depiction of a system where the overall functionality of the system is communicated by means of system concepts, a generalised idea of one or a group of interacting components (Sokolowski & Banks, 2010, p. 148). This research constructs a conceptual model, and for the remainder of the study, the term model will be used for the sake of simplicity.

The model is created using the various standards for thermal image capturing and analysis from literature, within the domain of machine learning classification. This model is then aligned to be implemented in a mobile environment using the technology and processes available. The model encompasses all aspects of the complete system, such as machine learning pipeline creation, machine learning pipeline deployment, camera requirements, in-app operations, and the updating of the application. The model is required to be able to accurately classify thermal breast images that have been captured according to standards on a mobile device.

5.2 Conceptual model overview

This section provides a high-level overview of the model, illustrated in Figure 5.1. The model consists of two components, namely a client-side and a server-side.

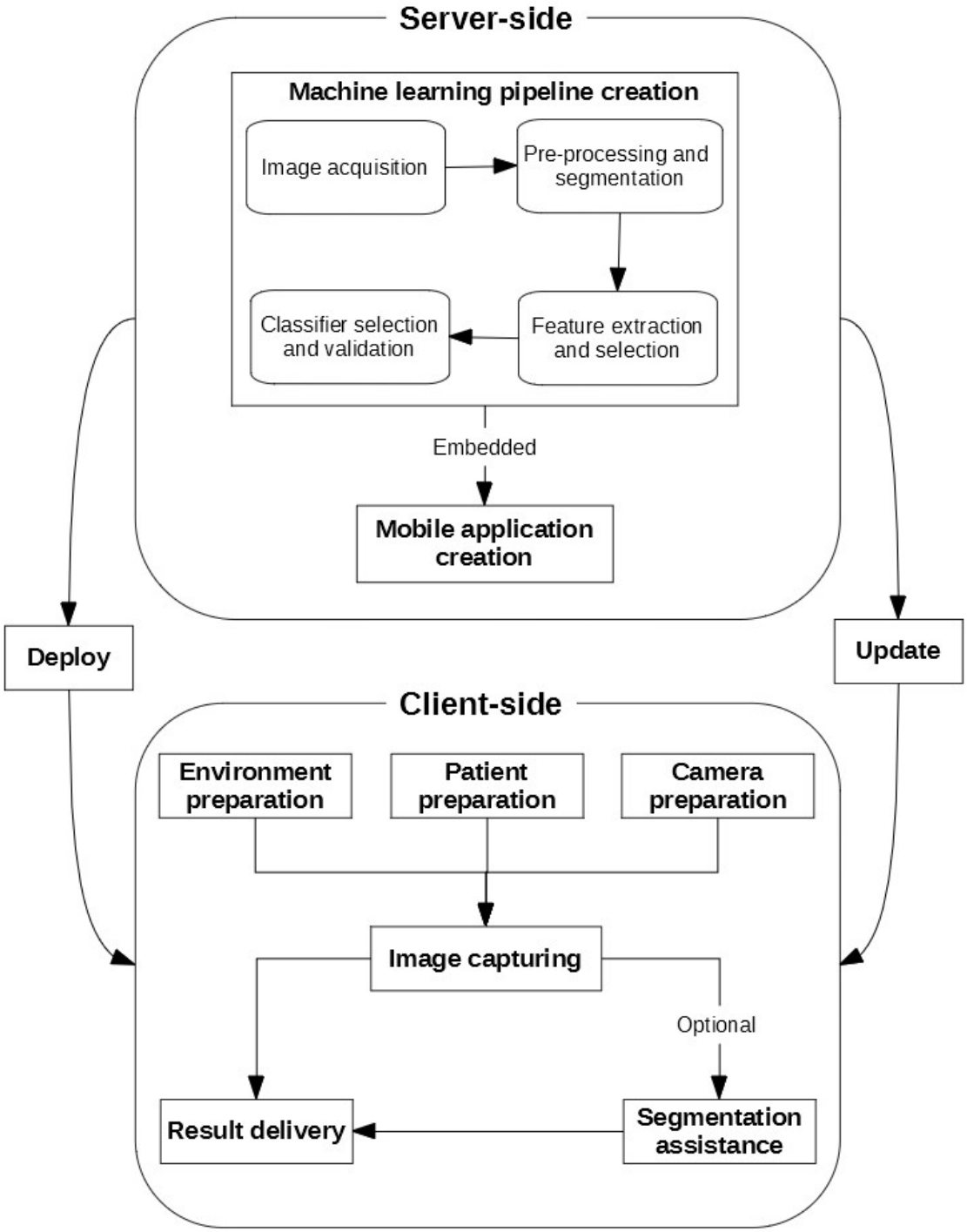


Figure 5.1: Model for the automated detection of breast cancer using thermal images and machine learning in a mobile environment

The server-side component deals with the creation of the machine learning pipeline which is embedded within a mobile application. The mobile application is delivered to the client via a content delivery network.

The client side of the model deals with the everyday use of the mobile application. It covers the required hardware and preparation steps needed in order to perform the intended function.

The two components are linked with two concepts; deployment and update. The deployment concept deals with the initial delivery of the mobile application to the client via a content delivery network. When an update is performed to either the machine learning pipeline used or to the mobile application itself, the new version of the application is delivered to the client via a content delivery network.

The following sections provide a more detailed description of the server side and client side components. They explain the concepts used for each component and how they were derived from literature.

5.3 Server-side component

The server-side functionality has two parts; the creation of the machine learning pipeline and the deployment of the application with the pipeline embedded within it. As mentioned previously, the delivery of the application is done via a content delivery network. The following sub-sections discuss the various facets of the server side aspect of the model in turn, according to the section numbers indicated in Figure 5.2.

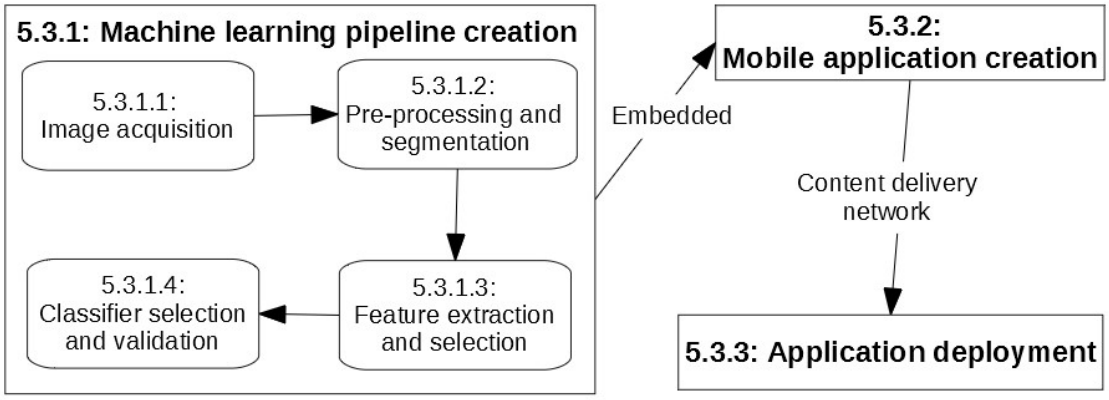


Figure 5.2: Server side component of the model

5.3.1 Machine learning pipeline creation

Chapter 4 detailed how machine learning is applied to problem solving. The task at hand for this research is binary classification of thermal images. This section details the individual steps undertaken in order to achieve this goal. Figure 5.3 details the work flow for this task. The outlined stages are done external to the application on a server, then the resulting machine learning pipeline is used within the application for deployment.

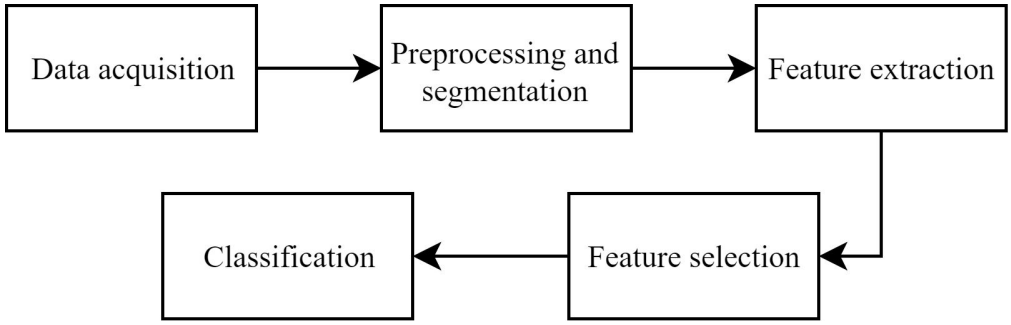


Figure 5.3: Proposed workflow for training a machine learning algorithm on thermal breast images

5.3.1.1 Image acquisition

The first stage in any machine learning problem is data collection. The data collected for this research is thermal breast images of patients where their diagnosis is known. The images in this data set are captured in alignment with standards and best practices at the time of writing, outlined in Section 3.4.1. They dictate the following:

- The preparation of patient.
- The preparation of the examination environment.
- The minimum hardware requirements of the thermal imager.
- Which images to capture.
- Reporting, archiving, and storing of the data.

The latest standards should be consulted when implementing the model. It is imperative that the images used for training adhere strictly to standards as poor data can never yield good results. Therefore, non-adherence to standards will result in a failed implementation of the model.

5.3.1.2 Pre-processing and segmentation

Once the images have been captured and stored on the server, the next step is the pre-processing and segmentation. This stage is vital as its purpose is to remove unwanted tissue, leaving only the breasts behind. There are various methods of segmentation with varying degrees of automation, ranging from fully manual, with large time investment and potential user error, to fully automatic, with no potential user error but potential for computing error in segmentation.

5.3.1.3 Feature extraction and selection

Once the breasts have been isolated and separated into left and right regions of interest, the next step is to extract features from the images. Abnormal breast tissue exhibits temperature distributions which are different to the normal distributions expected. These temperature distributions can be described by analysing the image's texture in order to extract quantifiable information representing the image.

In Section 4.2.4 it was said that highly dimensional data leads to overly complex machine learning models, which result in a poor ability to generalise by modelling the noise in the data. The dimensionality of the data can be reduced by identifying features that are important and removing those that are not. Reducing the features to only the most relevant ones leads to a faster training time and a better performing machine learning model, which can adapt well to unseen cases.

5.3.1.4 Classification and validation

Once the features have been extracted from the data and the redundant features removed, a classifier is chosen so that predictions can be made on new data based on the example data. This is the most important step in the entire process as the choice of classifier has a drastic effect on accuracy. Proper sampling and performance evaluation should be done in order to gauge performance of the chosen classifier accurately.

5.3.2 Mobile application

Once a machine learning model has been trained, it, along with its pipeline, is embedded into a mobile application. This mobile application comprises the machine learning pipeline as well as instructions for the proper capturing of patient images.

5.3.3 Application deployment

Once the application has been created, it needs to be deployed to the client. This deployment is done via a content delivery network in order to allow the user to access the application remotely.

5.4 Client side component

The client side is where most of the daily work is done. This section details the required functionality within the application for the user to adhere to in order to capture adequate thermal images which are used to make predictions. These instructions ensure that the capturing of the images adheres to the same methodology according to which the machine learning pipeline was created, ensuring good data for the predictive algorithm. The following sub-sections discuss the various facets of the client side aspect of the model, in turn, according to the section numbers indicated in Figure 5.4.

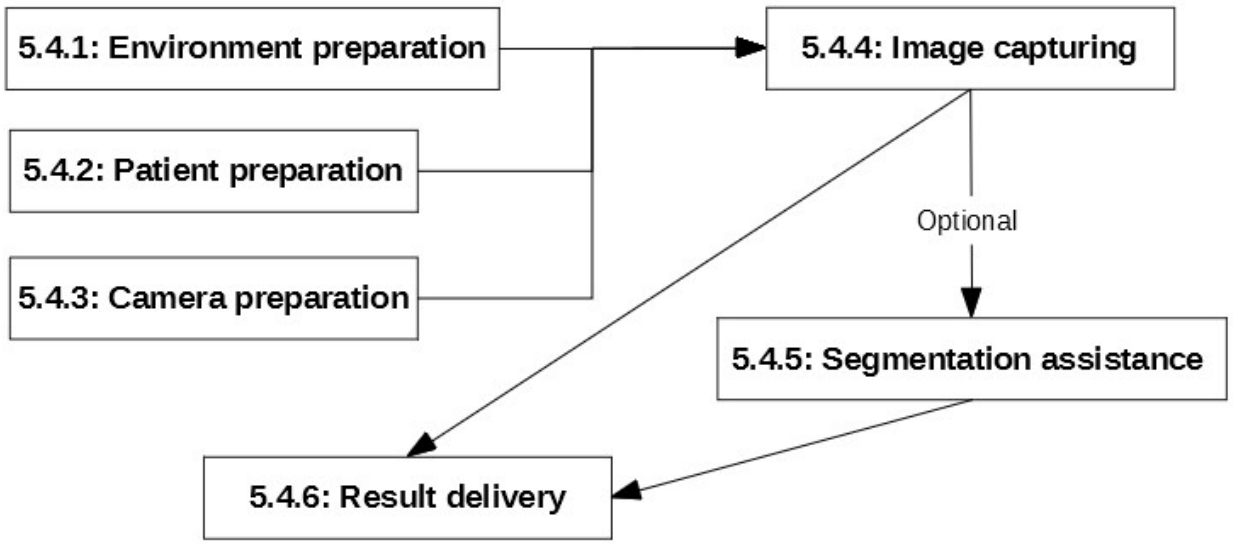


Figure 5.4: Client side component of model

5.4.1 Environment preparation

The correct preparation of the environment is essential to capturing the temperature profile of the patient's skin accurately, as any outside influence creates thermal artefacts in the captured images. Section 3.4.1.2 details the steps to be taken to ensure that the proper environment

is created as per the current standards and best practices. For future implementations of the model, the latest standards and best practices should be consulted.

5.4.2 Patient preparation

To ensure that the captured thermal pattern reflects the patient's rested state and is not altered by substances or exposure to other sources, the patient needs to be properly prepared. The requirements at the time of writing of this preparation, are found in Section 3.4.1.1, but it is recommended that the latest standards be consulted when implementing the model. This is done to eliminate the possibility of abnormal thermal patterns being caused by something other than the underlying physiology.

5.4.3 Camera preparation

The client-side portion of the model begins with a description of the minimum requirements for the attachable thermal camera. The thermal camera needs to be above the minimum requirements outlined by the standards and best practices of the time. The current standards are detailed in Section 3.4.1.4. Table 2.2 provides a list of currently available cameras which are above the minimum requirements set out by these standards.

To reduce thermal noise and artifacting, the thermal imager needs to be calibrated to the current environment, set up properly, and allowed to warm up. The current requirements of this process are found in Section 3.4.1.3.

5.4.4 Image capturing

Once all the preparations have been made, the patient may now be photographed. The images to be taken are critical for a successful prediction to be made. The patent's breasts should be in the centre of the view port and should take up most of it. The standards, at the time of writing, for proper image capturing are outlined in Section 3.4.1.4.

5.4.5 Segmentation assistance

Depending on the choice of segmentation method the user may be required to do some manual segmentation before handing the images over. The degree to which the user is involved in the segmentation is dependent on the choice of implementation made.

5.4.6 Result delivery

Once the images have been captured and properly prepared, they are handed over to the embedded machine learning pipeline. This pipeline reproduces the functionality described in the server-side component with regards to image manipulation, feature extraction, and prediction. The resulting prediction is then made available to the user.

5.5 Conclusion

The purpose of this chapter was to develop the research output, a model for the automated detection of breast cancer using thermal images and machine learning in a mobile environment. The model was developed using the literature covered in previous chapters. It took the form of a graphical representation of the chosen system, a mobile based application that uses machine learning to detect breast cancer with thermal images. The system comprises two components, a server-side component and a client-side component.

The server-side component is responsible for the creation and embedding of a machine learning pipeline into a mobile application, which is then delivered to the client using a content delivery system. The usage of the application is purely client side. This means that diagnosis can be made with no data connectivity. The client side of the model represents the in-app usage, which simply means capturing thermal images in accordance with standards and passing them on to the embedded machine learning pipeline for diagnosis.

The next chapter provides an implementation of the model's server-side functionality by creating a machine learning pipeline capable of accurately diagnosing breast cancer from thermal images. The implementation follows the concepts outlined in the model created here with the goal of identifying the best combination of segmentation, features, and classifiers. The images used for the implementation follow the standards required in the model strictly.

Chapter 6

Experimentation

6.1 Introduction

This chapter details the steps taken to implement the model created in Chapter 5. The implementation is done as an experiment to determine the feasibility of the model. It begins with an overview of the steps, followed by a more focused detailing of the individual steps.

Section 5.3.1 detailed the concepts required to create the machine learning pipeline. These concepts are restated as: data acquisition, pre-processing and segmentation, feature extraction, feature selection, and finally classification.

It is crucial that each of these steps is completed with an adherence to standards and best practices, since failure to do so results in an inaccurate diagnostic system. Sections 6.2 and 6.3 focus on the raw images captured and how to prepare them for further processing. Section 6.4 explains the process around converting real world images into quantifiable representations. It covers the choice of features to extract as well as how these features are reduced to have lower dimensional data. Finally, Section 6.5 covers how the features are used with various machine learning models to make predictions based on examples. It covers the choice of sampling as well as how the performance of the classifiers is evaluated.

6.2 Image acquisition

This section serves as an implementation of the *image acquisition* concept (from Section 5.3.1.1), illustrated in Figure 6.1.

The images taken for analysis need to adhere to a certain set of standards to be considered adequate for use, as indicated in Section 3.4.1. From this configuration, specific images need to be taken of the patient in accordance with the image capture protocol, from Section

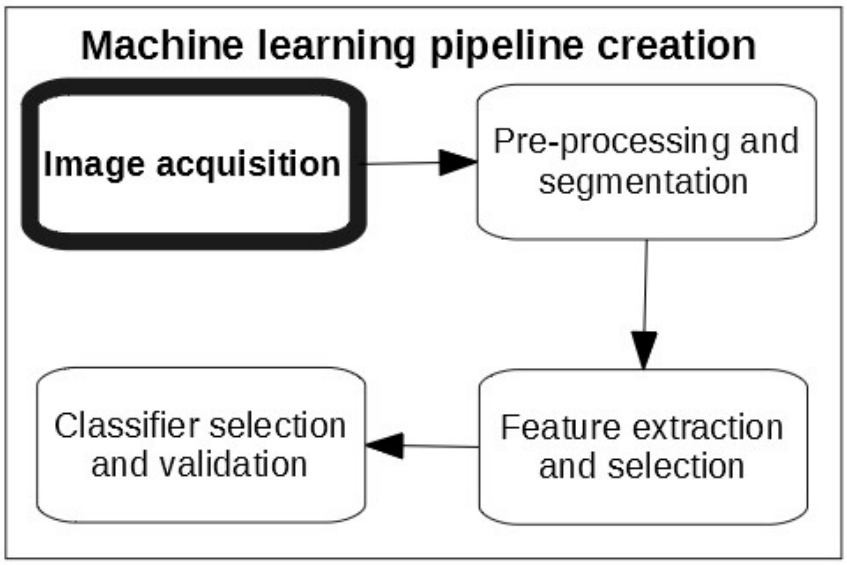


Figure 6.1: Image acquisition step within the machine learning pipeline creation concept

3.4.1.4. Adherence to these standards is imperative for the research to be sound; thus, only data sets which explicitly adhere to these standards were considered. The images need to be of a quality that is both reproducible my mobile attachable thermal cameras and meets the minimum requirements specified in the standards.

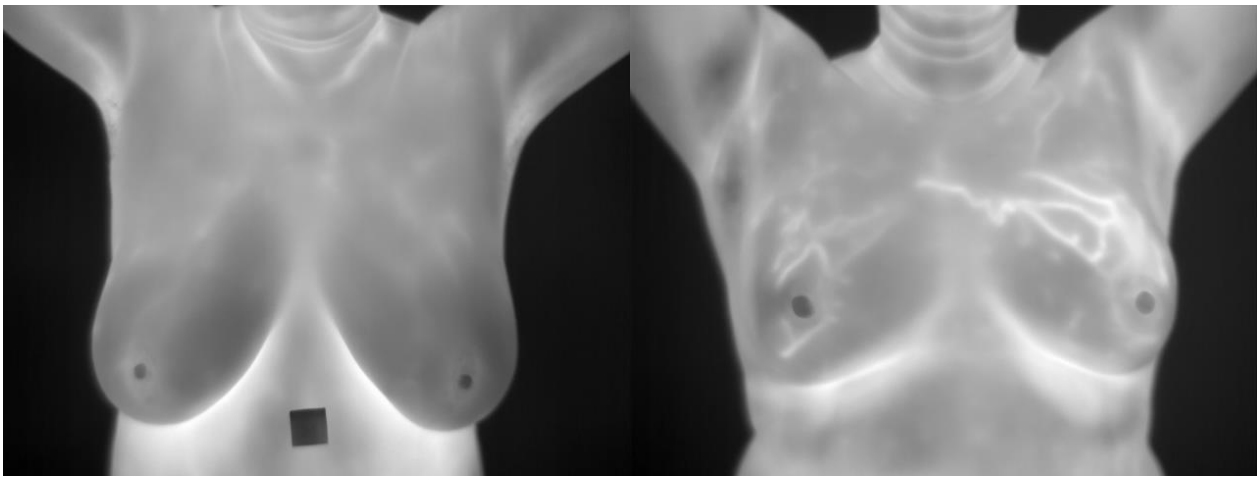
The search for an appropriate database was governed by the following rules:

1. The dataset must be publicly available.
2. The dataset must adhere to the standards outlined in Section 3.4.1.
3. The dataset must contain images with expert diagnosis made beforehand.
4. The dataset must have been used in recent studies relevant to the problem.
5. The quality of the images captured must be reproducible using currently available attachable mobile cameras.

The Database for Mastology Research (DMR) hosted at Visual Lab, Fluminense Federal University, Brazil was chosen for this research (Silva et al., 2014). This was done owing to proper adherence, easy access, accurate labelling, and sufficient number of unhealthy patients. The DMR dataset is also the most commonly used dataset for breast thermography. Many studies report findings on private datasets which are not freely available to use, even after asking permission. Freely available datasets are often small and heavily imbalanced (kanti Bhowmik et al., 2017), or contain very limited frontal views; which are needed for this study

due to automated segmentation. These reasons further strengthen the DMR dataset as the correct choice.

For the DMR, the breast thermograms were captured using a FLIR SC-620 infrared camera with a temperature sensitivity of 0.04C (40mK) and pixel resolution of 640480 pixels; this quality can be replicated using available attachable mobile thermal cameras. The DMR database contains 286 patient images, 42 of which are abnormal. From the set of abnormal images 30 were selected for assessment. The abnormal patient images removed from assessment were due to the patients having one breast, the images being severely out of focus, or the patients not being completely straight on with the capture device causing image artefacts. All these occurrences could skew results. To match the 30 abnormal patient images, 30 healthy patient images were selected for asymmetry analysis. Figure 6.2 shows an example of normal and abnormal thermograms taken from the DMR database.



(a) Normal (b) Abnormal

Figure 6.2: Thermograms taken from DMR dataset

6.3 Pre-processing and Segmentation

This section serves as an implementation of the *pre-processing and segmentation* concept (from Section 5.3.1.2), illustrated in Figure 6.3.

This stage is one of the most important stages, since accurately extracting the region of interest, in this case the breast, results in calculated features that are accurate in the sense that there is no misleading information. This results in classifiers being trained on better quality data; thus, being better performers. All the images had their heads up display (HUD)

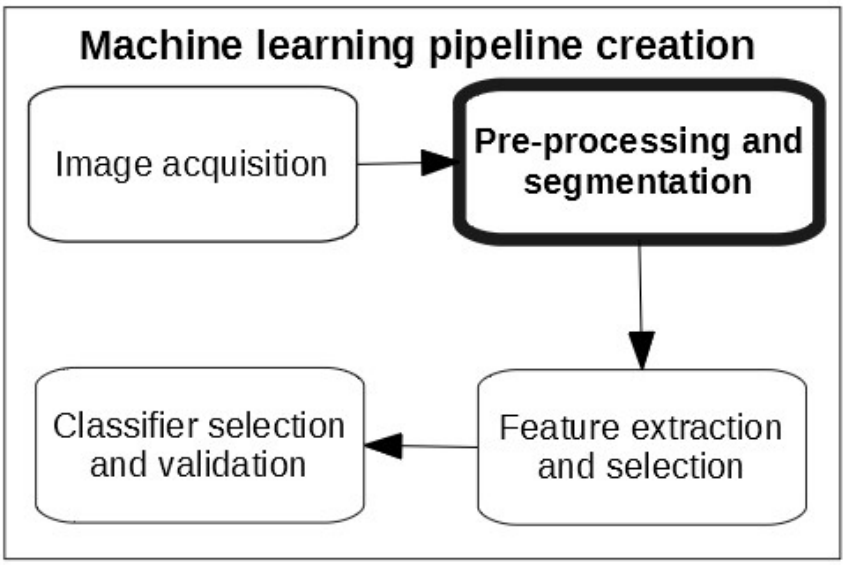


Figure 6.3: Pre-processing and segmentation step within the machine learning pipeline creation concept

elements removed by FLIR software and converted to grey scale, as seen in Figure 6.2. These images were imported into a Python 3 environment using the OpenCV library (Bradski & Kaehler, 2008) and the SimpleITK library (Yaniv, Lowekamp, Johnson, & Beare, 2018).

Different studies approached segmentation in different ways but within a few categories, namely manual, semi-automatic, or automatic. This section outlines the five approaches that this research undertook in order to compare various segmentation methods. These methods were: distance based automatic segmentation, manual region of interest box cropping, semi-automatic segmentation, fully manual segmentation, and fully automatic segmentation.

When performing any form of automatic segmentation, morphological functions are used in order to manipulate the images.

6.3.1 Morphological functions

Morphological functions are mathematical functions used to analyse and process geometrical structures (Beyerer, León, & Frese, 2015, p. 609). The basic methods for morphology are erosion and dilation (Giardina & Dougherty, 1988).

Erosion is defined by Beyerer et al. (2015, p. 611) as a method by means of which the structuring elements are reduced, and dilation achieves the opposite wherein the structuring elements are enlarged.

By combining erosion and dilation, other morphological functions are created depending

on the order in which these functions are performed. Morphological opening is erosion followed by dilation. Opening removes small objects from the image while retaining its shape. Morphological closing is dilation followed by erosion. This fills small holes in the image while preserving its shape (Beyerer et al., 2015, p. 620). Morphological thinning is the difference between the original image and the result of the hit-or-miss operator, which results in the contour of the image being reduced to a width on 1 pixel (Beyerer et al., 2015, p. 627).

6.3.2 Distance-based automatic segmentation

The first method is reliant on the fact that the images are captured in a specific way, outlined in Section 3.4.1.4. This method is taken from Ali et al. (2015) and based on the distance the patient is away from the camera, and on what part of the patient needs to be captured. The authors located four bounds to which the image could be automatically cropped, so that the breasts of the patient were in frame and centred. The algorithm to achieve this cropped image is as follows:

1. Remove the background using binary thresholding.
2. Define two parameters, Y1 and Y2. These parameters represent the upper and lower bounds for the image. The value of Y1 is 0.25 of the image height, M; and Y2 is 0.8M.
3. Crop from Y1 to Y2, removing the unwanted top and bottom portions of the image.
4. Scan all the columns of the image. Those which are all zero, meaning fully black, are removed. This removes the left and right areas of the image in which there is no tissue, resulting in a centred image.
5. Identify the horizontal midpoint of the image and split the image there, resulting in a left and right breast being segmented out.

This method is very light computationally and is thus very fast; however, it does lack generalisation and, as previously discussed, breasts are ambiguous. This will lead to some images being incorrectly segmented.

6.3.3 Manual ROI box segmentation

This method represents the easiest manual approach to segmentation where the expert manually manipulates a box, so that the region of interest is extracted. This method does not remove the tissue below the inframammary folds; thus, it is prone to the influence caused by hot spots in the folds of the skin causing thermal artefacts.

With a manual approach, the user crops the image to specific boundaries, as outlined by Araujo et al. (2017):

1. The background is removed using binary thresholding as before.
2. The upper bound of the region of interest is at the peak of the armpit, or axilla, of the patient.
3. The lowest portion of the inframammary fold represents the lower bound for the region of interest.
4. The outermost left and right edges of the patients breast represent the left and right bounds of the region of interest.
5. The image is then separated at the mid point as before and the left and right breasts are then segmented out.

6.3.4 Semi-automatic segmentation

This method looks to improve on the manual box crop method by introducing an element of automation. This method automatically detects and removes the tissue below the inframammary fold, but it must not remove the fold itself as it contains important information (Gogoi, Bhowmik, Ghosh, Bhattacharjee, & Majumdar, 2017).

The following steps, adapted from Sathish et al. (2017), are taken in order to fit two polynomials to match the inframammary folds:

1. Apply a Gaussian blur to the image, using a 3x3 kernel and $\sigma = 1.4$.
2. Use Canny edge detection with adaptive thresholding in order to detect the edges. The adaptive threshold is set by using Otsu's method, which results in a more accurate extraction of the edges (Fang, Yue, & Yu, 2009).
3. Define a disc-shaped structuring element with radius 4. This is used for the morphological functions.
4. Use morphological closing to join close edges.
5. Use a distance filter to keep only the strongest edges.
6. Thin the edges to 1 pixel.

7. Fit two polynomials to the curves detected, one going from the left to the centre, and the other from the right to the centre.
8. Identify the intersecting (bifurcation) point of the polynomials.
9. Mask what is above the polynomials in order to remove the area below them.
10. Split the image into left and right breasts at the bifurcation point.

6.3.5 Fully manual segmentation

This method follows the same steps as the manual box crop method with the addition of the inframammary folds being traced and segmented out. This is often the preferred method as it does away with the ambiguity of the other methods, which can result in poor segmentation. However, this method is the most time-consuming and has the highest chance of human error.

6.3.6 Fully automatic segmentation

This method of segmentation aims to identify the bounds for the breast tissue as well as remove the lower inframammary tissue automatically. This method is the most complex and prone to error. It is adapted from multiple successful studies doing automatic segmentation (Motta, Conci, Lima, & Diniz, 2010; Pramanik, Bhattacharjee, & Nasipuri, 2015; Lashkari, Pak, & Firouzmand, 2016; Sathish et al., 2017). This method involves five stages:

1. Detect the lower limit.
2. Detect the left and right limits.
3. Detect the upper limit.
4. Detect the inframammary fold.
5. Fit two polynomials to the detected inframammary fold and remove the tissue below it.

Each of these stages requires many other stages in order to be completed. For this reason, these individual stages are discussed, in turn, in the following subsections.

Lower limit detection

For this part of the segmentation only the lower half of the image is considered. There is an assumption that is made for the lower limit to be detected; that is, the region within and immediately surrounding the lower inframammary folds are the regions with the highest temperature.

With this assumption in mind, the following steps are undertaken in order to detect the lower limit:

1. Split the image into vertical halves, focusing only on the lower half.
2. Create a threshold, so that only the pixels with the highest, or very near the highest, intensity remain. These pixels become white while the remainder of the image is set to black.
3. In order to remove smaller regions of temperature spikes and to retain the thickest areas, several morphological functions, making use of a disk shaped structuring element, are used.
   (a) Apply morphological dilation in order to join disconnected pixels.
   (b) Apply a median filter to smooth out any inconsistencies.
   (c) Apply morphological thinning to it in order to reduce the size of the curve to 1px.
4. Scan left-to-right and right-to-left while scanning bottom-to-top in order to identify the set of pixels, which are the lowest of the image (row with the highest value) in each sweep. This lowest of the left and right sweeps is then used as a lower limit, lim_b.

The image is then cropped in order to remove the portion of the image below lim_b

Left and right limit detection

Once again this focuses on the lower half of the image with this lower limit removed. The following steps are followed in order to attain the left and right limits, lim_l and lim_r:

1. Use Canny edge detection with Otsu thresholding in order to identify the edges of the image.
2. Use morphological opening to join smaller edges together.
3. Perform a distance transform in order to remove smaller, unjoined edges.

4. Thin the edges to 1 pixel.

5. To obtain the left limit, lim_l, scan from left-to-right and locate the first white pixel. To obtain the right limit, lim_r, repeat the process but from right-to-left.

Upper limit detection

This limit is detected using the top half of the image. The steps taken in order to identify the top limit, lim_t, are as follows:

1. Extract the border in the same way as was done when detecting the left and right limits.

2. Scan from bottom-to-top at lim_l until a white pixel is found.

3. Scan from bottom-to-top at lim_r until a white pixel is found.

4. The lowest pixel of the two is taken at the upper limit, lim_u.

Once all four limits have been detected, the image is cropped using these limits. The resulting cropped image is subjected to the same steps as in the semi-automatic method to obtain accurately segmented breasts. As mentioned before, this method is the most expensive computationally and most prone to error as the ambiguity of breasts could very easily result in a false segmentation.

6.4 Feature extraction and selection

This section serves as an implementation of the *feature extraction and selection* concept (from Section 5.3.1.3), illustrated in Figure 6.4.

This section is created in alignment with the requirements set out in Section 5.3.1.3. This serves as an implementation of the concept denoted as *feature extraction and selection* in the conceptual model created in Chapter 5.

After proper segmentation has been performed, relevant features need to be extracted from the images in order to train machine learning algorithms. This process replaces the image properties with vectors of numbers representing them.

Abnormal breast tissue exhibits temperature distributions which are different to the normal distributions expected. These temperature distributions can be described using texture analysis techniques (kanti Bhowmik et al., 2017). This analysis leads to a set of features,

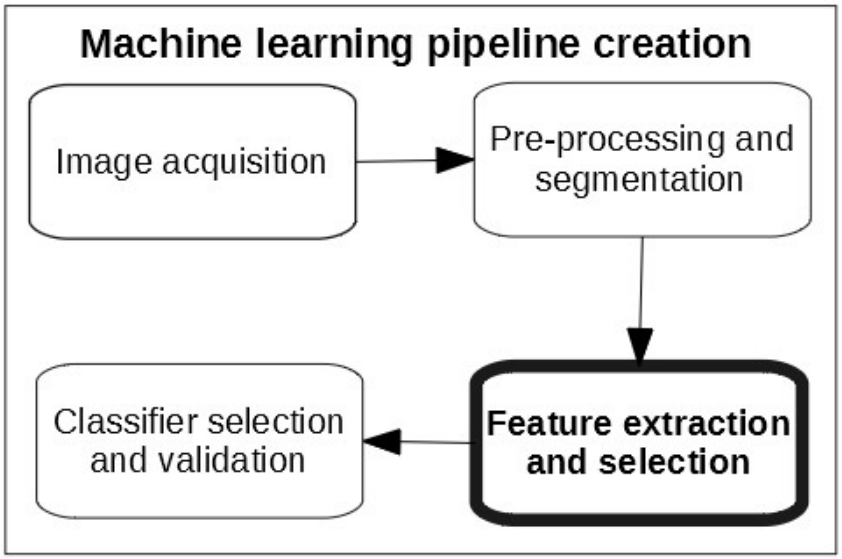


Figure 6.4: The feature extraction and selection step within the machine learning pipeline creation concept

which can represent symmetric and asymmetric thermal patterns quantitatively (Gogoi et al., 2017).

Texture analysis, in the context of medical images, is typically done by a practice called Radiomics. Lambin et al. (2017) define radiomics as "the quantitative mapping, that is, extraction, analysis and modelling of many medical image features in relation to prediction targets, such as clinical end points and genomic features". Radiomics is essentially a high-throughput quantitative image analysis technique that covers every aspect from image acquisition, to segmentation, to feature extraction and selection, to data storage and warehousing (Kumar et al., 2012). For the purpose of this research, only the feature extraction portion is used. These extracted features are known as image biomarkers. There are many different types of these biomarkers, all of which are collated in the image biomarker standardisation initiative, or IBSI (Zwanenburg, Leger, Vallieres, & Lock, 2016).

Radiomics is a relatively new field, as is the use of machine learning in medicine. Despite this, Radiomics is widely used in all computer-aided diagnosis and detection systems and is supported by the National Cancer Institute (NCI), Quantitative Imaging Network (QIN) and other initiatives from the NCI Cancer Imaging Program (Gillies, Kinahan, & Hricak, 2015).

Radiomics is mainly used to offer information on cancer pheno-types as well as on the tumour microenvironment normally imaged, in 2D or 3D, via computed tomography (CT), magnetic resonance (MR), or positron emission tomography (PET) (Lambin et al., 2017).

The field of thermography has made use of radiomic features for breast cancer detection. This is because the extracted features can provide information on the intensity histogram, shape, and texture (Kumar et al., 2012). This is relevant to thermography as these are quantified representations of the aspects that experts had to consider when doing thermovascular analysis, from Section 3.4.2.

These features can be divided into several families, namely: intensity-based statistical, intensity histogram-based, intensity-volume histogram-based, morphological features, local intensity, and texture matrix-based features (Zwanenburg et al., 2016). These features can be categorised into various orders which represent the number of pixel relationships considered in the calculation; first, second, or higher (Gillies et al., 2015). Many different features have been used within breast thermography, the most common being first order histogram and second order texture features (Bhowmik et al., 2016).

6.4.1 First order histogram-based features

These features only consider one-pixel relationships. They give an indication of the distribution of pixels of varying intensity (Sathish et al., 2017). These features are calculated from a histogram of the region of interest. The histogram of an image is a mapping of the intensity of the pixel versus its occurrence and provides insight into the probability of finding a pixel of a specific intensity within an image (Bhowmik et al., 2016).

Nailon, William Henry (2010) explains that for an image consisting of grey levels in the range of $0 \leq i \leq N_G - 1$, with N_G being the number of grey levels and $N(i)$ as the total number of pixels on intensity i with M being the total number of pixels in the image. The histogram, or pixel occurrence probability $P(i)$ of this image is given by Equation 6.1.

$$P(i) = \frac{N(i)}{M} \tag{6.1}$$

The following features are calculated from the histogram of the image: mean, variance, standard deviation, skewness, kurtosis, entropy and energy (Materka & Strzelecki, 1998; Nailon, William Henry, 2010; Zwanenburg et al., 2016; Gogoi et al., 2017).

The mean is the average intensity of the image, given by Equation 6.2.

$$\mu_1 = \sum_{i=0}^{N_G-1} iP(i) \tag{6.2}$$

The variance is the variation of intensity around the mean for the image, given by Equation 6.3

$$\mu_2 = \sum_{i=0}^{N_G-1} (i - \mu_1)^2 P(i) \tag{6.3}$$

The standard deviation is simply the square root of the variance, given by Equation 6.4.

$$\sigma = \sqrt{\mu_2} \tag{6.4}$$

Skewness is a measure of symmetry about the mean, with 0 being perfectly symmetrical, given by Equation 6.5.

$$\mu_3 = \sigma^{-3} \sum_{i=0}^{N_G-1} (i - \mu_1)^3 P(i) \tag{6.5}$$

Kurtosis is a measure of the flatness of the image, given by Equation 6.6.

$$\mu_4 = \sigma^{-4} \sum_{i=0}^{N_G-1} (i - \mu_1)^4 P(i) \tag{6.6}$$

Entropy is a measure of the uniformity of the histogram, given by Equation 6.7.

$$H = - \sum_{i=0}^{N_G-1} P(i) \log_2 P(i) \tag{6.7}$$

Mean, variance, skewness, and kurtosis are known as the first four moments of an image, hence being labelled $\mu_1 - \mu_4$. For a given thermogram, these features are computed for both the left and the right breast and the difference is then used as the feature. These features are commonly used because of their simplicity, severely reducing calculation time for larger datasets.

6.4.2 Second order texture features

These features are calculated using the relationship of a pixel and its neighbours. These pixel pairs are tabulated in something called a grey level co-occurrence matrix (GLCM), which is essentially a mapping of the probability of finding pixel pairs in the image. The GLCM is known as the second order histogram of the image (Sathish et al., 2017).

These second order statistics are very important as they quantify what humans see in a texture, Julesz (1975) found that no texture pair can be discriminated by eye if they agree in their second-order statistics.

Haralick, Shanmugam, and Dinstein (1973) were the first to make use of second order features with a co-occurrence matrix when they analysed aerial photography. Haralick et

al. (1973) define the $N_G x N_G$ gray level co-occurrence matrix, P_d, for a displacement vector $d = (dx, dy)$ as follows: The entry (i, j) of P_d is the number of occurrences of the pair of gray levels i and j which are a distance d and angle α apart, more formally seen as: $P(i, j : d, \alpha)$. This means that while each pixel has eight neighbours, only four angles are used: $\alpha = 0°$, 45°, 90°, and 135°. Figure 6.5 illustrates these angles and the pixel relationships considered.

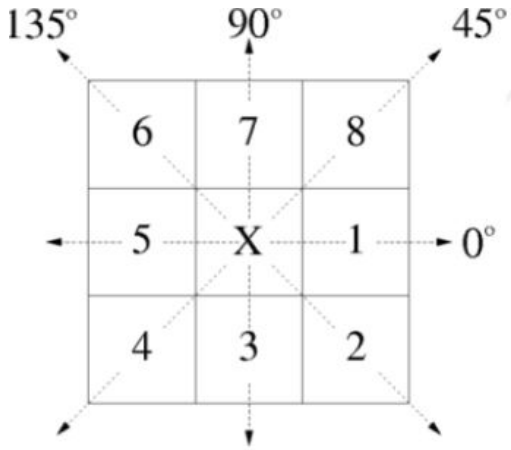


Figure 6.5: The various angles used when calculating second order features from a grey level co-occurrence matrix (Nailon, William Henry, 2010, p. 80)

The features extracted from this matrix provide unique information of the texture being analysed (Nailon, William Henry, 2010, p. 81). Initially, Haralick et al. (1973) proposed 14 features with Soh, Tsatsoulis, and Member (1999), and Clausi (2002) adding to those. This resulted in the GLCM having the following features calculated from it (Zwanenburg et al., 2016, p. 58-79).

- Angular second moment (energy) - a measure of homogeneous patterns in the image.
- Contrast - a measure of the local intensity variation.
- Correlation - shows the linear dependency of gray level values.
- Sum of squares or variance - a measure in the distribution of neighbouring intensity level pairs about the mean intensity level in the GLCM. This differs from the variance mentioned in first order statistics in the fact that it deals with the intensity of the pair of pixels and not with a single pixel.
- Inverse difference moment (homogeneity) - how homogeneous the image is overall.
- Sum average - measures the relationship between occurrences of pairs with lower intensity values and occurrences of pairs with higher intensity values.

- Sum variance - a measure of groupings of pixels with similar gray-level values.
- Sum entropy - a sum of neighbourhood intensity value differences.
- Entropy - a measure of the randomness/variability in neighborhood intensity values.
- Difference variance - a measure of heterogeneity that places higher weights on differing intensity level pairs that deviate more from the mean.
- Difference entropy - a measure of the randomness/variability in neighbourhood intensity value differences.
- Information measure of correlation 1 and 2 - assesses the correlation between the probability distributions of i and j (quantifying the complexity of the texture).
- Autocorrelation - a measure of the magnitude of the fineness and coarseness of texture.
- Dissimilarity - measures the relationship between occurrences of pairs with similar intensity values and occurrences of pairs with differing intensity values.
- Inverse difference normalised - another measure of the homogeneity of the image.
- Inverse difference moment normalised - another measure of the homogeneity of the image.

There are features that are calculated from pixel relationships of three or more using a grey level run length co-occurrence matrix (GLRLM) and wavelet transforms (Haralick, 1979), but those are out of the scope of this research as the focus is on first and second order texture features.

For the purposes of this research, the feature calculation is done with the help of the PyRadiomics library (Hosny et al., 2017). Every feature is calculated for each angle, then averaged, in order to return one value per feature calculated. The full feature set comprises the difference between the left and right breast for each of the features.

6.4.3 Feature reduction

In section 4.2.4 it was stated that a high dimensional input space leads to an overly complex model with high variance. In the feature calculation step, a total of 23 features are calculated. Given the fact that there are 30 images for each class, a model trained on 23 features would be overly complex and have very high variance.

Features which make a negligible contribution to distinguishing between normal and abnormal need to be removed to reduce the complexity of the models. This is done via various methods, namely principal component analysis (PCA) and statistical significance testing (Gogoi et al., 2017). Aside from improved generalisation, the algorithm's performance is improved by using this subset of the features as well as by the fact the algorithm is not confused by redundant features (Araujo et al., 2017). As stated before, machine learning algorithms with lower complexity and lower dimensional feature spaces are better at generalising as they have lower variance, which results in a better and more reliable accuracy.

6.4.3.1 Principal component analysis (PCA)

Principal component analysis (PCA) is defined by Hotelling (1933) as the orthogonal projection of the data onto a lower dimensional linear space, called the principal subspace, so that the variance of the projected data is maximised. In other words, PCA finds a low-dimensional representation of a data set that contains as much of the variation as possible. It has been stated that variance is a representation of the variability, this means that the components with the highest variance are the ones that account for the greatest change in the outcome.

This projection of the data produces principal components. These are the axis, defined so that the first one has the highest variance, the second one has the second highest variance and is orthogonal to the first, and so on (Géron, 2019, p. 223).

In order to identify these components with maximum variance, eigenvectors corresponding to the largest eigenvalues of the covariance matrix need to be calculated (Bishop, 2007, p. 562). There are many methods to calculate the principal components. For the purpose of this research, a standard matrix factorisation technique called singular value decomposition (SVD) is used from the scikit-learn library with varying values of variance required to be maintained. The resulting principal components represent a lower dimensional version of the original feature set. This allows for a machine learning algorithm to be trained with a lower complexity than otherwise would be the case.

6.4.3.2 Tests of statistical significance

There are various ways of deciding which features are significant. Two very common methods, representing parametric and non-parametric methods, are the Student's t-test and the Mann-Whitney-Wilcoxon (MWW) test (Whitney, 1971; Fay & Proschan, 2010).

In these tests a null hypothesis needs to be defined so that it can be rejected. Since asymmetry analysis is done with intensities of pixels that represent temperature readings, the null hypothesis can be written as: The mean of the abnormal group is less than the mean

of the normal group. This mean is not the mean temperature, but the mean of the feature calculated.

This null hypothesis will be rejected when there are differences between the groups where the abnormal is higher, and that disparity is not a likely result of chance. Each test produces a p-value. This p-value represents the confidence that this large difference is due to chance. This means that a lower p-values is indicative of a more significant feature.

This results in a feature set where there are no correlating or redundant features, according to the test of significance chosen. This reduced feature set is then used for training, which results in a less complicated algorithm.

6.5 Classifier selection and validation

This section serves as an implementation of the *classification and validation* concept (from Section 5.3.1.4), illustrated in Figure 6.6.

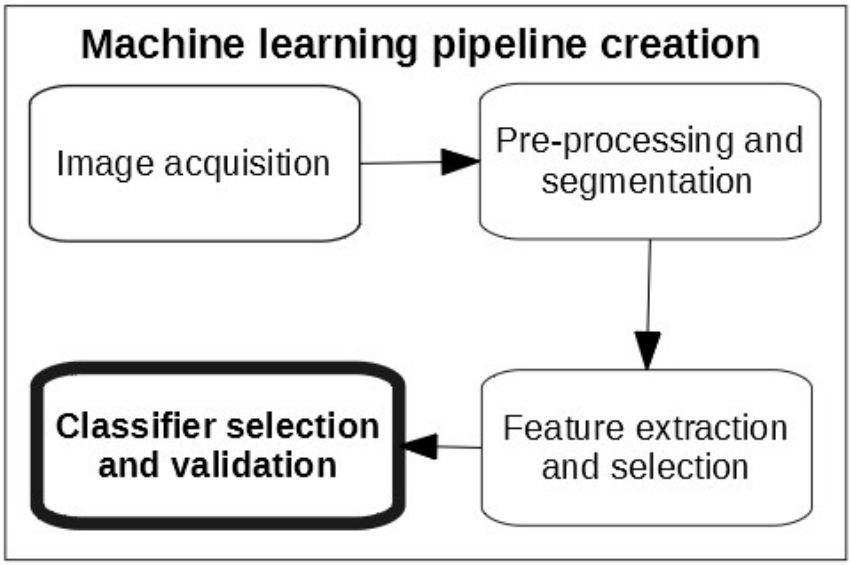


Figure 6.6: Classifier selection and validation step within the machine learning pipeline creation concept

In any breast cancer detection system, classification is the most crucial step. If the breasts are incorrectly segmented with poor features calculated, a high accuracy will never be attained. This study's dataset provides a binary classification problem owing to there only being two classes, normal and abnormal. The feature selection step ensures that the best features are given to the classifiers, as the correct combination of features and classifiers

can yield excellent results.

There are many classifiers, which can be applied to this problem scenario given the structure of the data. Therefore, relevant classifiers need to be identified. This is done by surveying related work. Table 6.1 shows related works. It was compiled by searching common data bases such as Google Scholar, Science Direct, and PubMed for articles relating to asymmetry analysis of thermal images from the last decade. The works that did not include texture features in their asymmetry analysis were excluded. This resulted in 26 papers being captured. In order to reduce its size, Table 6.1 makes use of the following shorthand notation for well-known terms:

- Fos - The first order histogram based features were used.
- GLCM - The texture features calculated from the grey-level co-occurrence matrix were used.
- Hos - Any higher order statistic was used. This includes features from the grey level run length matrix or any spectral analysis.
- SVM - Support vector machines were used. This could be one or several of the various kernel functions available.
- ANN - Any configuration of an artificial neural network was used.
- k-NN - the k-nearest neighbour classifier was used.
- DT - Decision trees were used.
- AdaBoost - The AdaBoost algorithm was used specifically.
- NB - Naïve Bayes classifier was used.
- Various ensemble - The paper contained a combination of boosted and bagged algorithms in its research.
- Acc - the highest accuracy score reported by the authors.
- Sn - the sensitivity of the classifier with the highest accuracy.
- Sp - the specificity of the classifier with the highest accuracy.

It is important to note that studies with higher numbers of patients, 100+, had unbalanced datasets owing to there being a shortage of unhealthy thermograms in available datasets.

Table 6.1: Related works

Author	**Features**	**Classifier**	**Best result**	**Patients**
(Sathish et al., 2017)	Fos, GLCM	SVM	Acc: 90% Sn: 87.5% Sp: 92.5%	80
(Araujo et al., 2017)	Fos, GLCM	SVM	Acc: 94.87%	80
(Gogoi et al., 2017)	Fos, GLCM	ANN, SVM, k-NN, DT	Acc: 87.5% Sn: 95% Sp: 80%	70
(Lashkari et al., 2016)	Fos, GLCM, Hos	AdaBoost, SVM, k-NN, NB, ANN	Acc: 87.42%	67
(Pramanik, Bhattacharjee, & Nasipuri, 2016)	Fos	ANN	Acc: 90% Sn: 95% Sp: 85%	100
(Mohamed, 2015)	Fos, GLCM	ANN	Acc: 96.12% Sn: 78.95% Sp: 97.86%	260
(Schaefer & Nakashima, 2015)	Fos, GLCM, Hos	Various ensemble	Acc: 82.13% Sn: 88.78% Sp: 87.46%	146
(Calderon-Contreras, Chacon-Murguia, Villalobos-Montiel, & Ortega-Maynez, 2015)	Fos	FL	Sn: 82.35% Sp: 92.15%	140

Continued on next page

Table 6.1 – continued from previous page

Author	Features	Classifier	Best result	Patients
(Pramanik et al., 2015)	Fos	ANN	Acc: 90.48% Sn: 87.6% Sp: 89.73%	102
(Gaber et al., 2015)	Fos, GLCM	SVM	Acc: 92.06% Sn: 96.55% Sp: 87.5%	149
(Krawczyk & Schaefer, 2014)	Fos, GLCM	Various SVM based ensemble	Acc: 90.03% Sn: 80.35% Sp: 90.15%	146
(Francis, Sasikala, Bhavani Bharathi, & Jaipurkar, 2014)	Fos, GLCM	SVM	Acc: 83.3%	36
(Schaefer, 2014)	Fos, GLCM, Hos	ACO	Acc: 79.52%	146
(Rodrigues & Conci, 2014)	Fos, GLCM, Hos	SVM, NB	Acc: 61.8% Sn: 62.9% Sp: 61.8%	102
(Francis, Sasikala, & Saranya, 2014)	GLCM	SVM	Acc: 90.91% Sn: 81.82% Sp: 100%	22
(Milosevic, Jankovic, & Peulic, 2014)	GLCM	SVM, k-NN, NB	Acc: 92.5%	40
(Krawczyk & Schaefer, 2013)	Fos, GLCM	Various SVM based ensemble	Acc: 81.37% Sn: 90.59% Sp: 88.76%	146

Continued on next page

Table 6.1 – continued from previous page

Author	Features	Classifier	Best result	Patients
(Francis & Sasikala, 2013)	Fos, GLCM	ANN	Acc: 85.19% Sn: 88.89% Sp: 77.78%	27
(Etehadtavakol et al., 2013)	Fos, GLCM	AdaBoost	Acc: 86%	40
(Kapoor, Prasad, & Patni, 2012)	Fos, GLCM	ANN	Acc: 80%	60
(Zadeh, Haddadnia, Hashemian, & Hassanpour, 2012)	Fos, GLCM	ANN	Acc: 75% Sn: 50% Sp: 75%	200
(Mookiah, Acharya, & Ng, 2012)	GLCM, Hos	DT, FL, k-NN, NB, ANN	Acc: 93.30% Sn: 86.70% Sp: 100%	50
(Acharya, Ng, Tan, & Sree, 2010)	Fos, GLCM, Hos	SVM	Acc: 88.10% Sn: 85.71% Sp: 90.48%	50
(Schaefer et al., 2009)	Fos, GLCM, Hos	FL	Acc: 79.53% Sn: 79.86% Sp: 79.49%	146
(Pak, Lashkari, & Firouzmand, 2009)	Fos, GLCM, Hos	AdaBoost, SVM, k-NN, NB, ANN	Acc: 88.03%	67

From Table 6.1 it can be seen that the most common classifiers are: artificial neural networks, support vector machines, k-nearest neighbour, naïve Bayes, and ensemble methods. In order to create and configure these classifiers, the scikit-learn library is used (Pedregosa et al., 2011). For this experiment, the following classifiers are contrasted:

- Artificial neural network - The structure of the ANN was chosen to have three hidden layers, all with 15 neurons. It is a simple feed forward neural network using a logistic activation function and a stochastic gradient descent solver. This was done by testing many variations in structure, with varying numbers of layers and neurons per layer. This configuration was the most stable and produced the most consistent results based on our datasets.
- Support vector machines - linear, polynomial, and radial-bias function (RBF) kernels were used with varying degrees for the polynomials and various values for gamma.
- Naïve Bayes - a simple straight forward Gaussian naïve Bayes algorithm algorithm was used.
- k-Nearest neighbour - various values of k were tested with uniform voting.
- Random forests - varying numbers of estimators were examined.
- AdaBoost - two different base estimators were used, decision trees and linear support vector machines. Various numbers of estimators and learning rates were also tested.

All of the above mentioned techniques are discussed in greater depth in Section 4.3.

6.6 Performance evaluation

In order to train and test classifiers, the data needs to be split into training and testing subsets. The training set is used to fit the model to the data, and the testing set to evaluate its performance of the model. This subsection details the sampling methods as well as the metrics used to evaluate the performance of the classifiers.

6.6.1 Sampling

In Section 4.4.1, it was stated that the data needs to be sampled into two sets, a training and a testing set. Owing to the dataset being relatively small, either a cross validated or a bootstrapped approach is ideal. For the purpose of this research, k-fold cross-validation was chosen as the sampling method.

The choice of k is generally unfixed, and as a rule of thumb, is generally chosen as 10. Owing to the size and complexity of the dataset used, this study makes use of a special case of cross validation, called leave-one-out cross-validation (LOOCV). This is when the value of k is equal to that of the smallest class, in our case, 30. This method is far more computationally expensive; yet the small sample size allows for this higher computation cost to not be a factor.

6.6.2 Metrics

Section 4.4.2 indicated that the performance of the classifier is normally evaluated using metrics such as accuracy (Acc), sensitivity (Sn) and specificity (Sp) which are generally computed from the parameters of confusion matrix like true positive (TP), false negative (FN), true negative (TN) and false positive (FP). Another common metric is the area under curve (AUC), which is the area under the receiver operating characteristic curve (ROC) - a curve of the true positive rate versus the false positive rate. For each classifier considered, a confusion matrix is generated to extract relevant metrics.

6.7 Conclusion

The aim of this chapter was to establish the experimental method undertaken to create an implementation of the server-side component of the conceptual model outlined. It covered the entire workflow for the experiments. In that process several steps are taken.

The acquisition of images are detailed as they need to comply with the standards for the modality. Various segmentation methods with differing degrees of automation were detailed. This is done so that the most efficient method of segmentation can be chosen for the solution.

From these segmented images, features need to be calculated. In the field of medical image analysis, hundreds of features can be extracted that range from shape descriptors to texture analysis to spectral analysis. For the purpose of this research, only first-order histogram based statistical features and second-order texture features are considered as they are the most common features used in thermography.

Owing to the small dataset being used, a high number of features will result in a high variance model being trained. In order to improve generalisation and reduce complexity feature reduction techniques are applied. Principal component analysis and tests of statistical significance are done in order to identify the best features to use.

The choice of classifier is the most important step; thus, it is important to select a number of classifiers that have been proven to work well in the field. In order to identify

these classifiers, related work was tabulated in order to identify the most commonly used classifiers in works using texture analysis. From this, several candidates were identified; thus, various configurations of these candidates are tested. The classifiers undergo orthodox methods of evaluation with a confusion matrix and accuracy, sensitivity, specificity, f1 score, and AUC being calculated.

The following chapter discusses the results of the experiments. From this the ideal combination of segmentation, features, and classifiers is identified. This is used to refine the implementation of the model.

Chapter 7

Experimental results

7.1 Introduction

This chapter details the results of the implementation of the model. It covers the results of the feature selection and classification for each of the proposed segmentation method. It begins by explaining all the terminology used in the chapter, along with the definition of the various feature sets used. Then, in turn, it discusses the results of each segmentation methods for both feature selection and classification.

7.2 Terminology

This section aims to detail the various terms, abbreviations, and shorthand employed in this chapter. It begins with an explanation of the various feature sets created and the terms that appear in the tables relating to feature extraction. Then it lists and explains the terminology used when indicating the results of the classification stage. The tables detailing extracted features in this chapter all make use of abbreviations for common terms:

- SM1 - Distance-based segmentation
- SM2 - Manual box crop segmentation
- SM3 - Semi-automatic segmentation
- SM4 - Fully manual segmentation
- SM5 - Fully automatic segmentation
- Fo Entropy - First order entropy

- ASM - Angular second moment
- So Entropy - Second order entropy
- IMC1 - Informational measure of correlation 1
- IMC2 - Informational measure of correlation 2
- IDN - Inverse difference normalised
- IDMN - Inverse difference moment normalised
- Sos - Sum of squares

The derivation of the various feature sets, with the number of features contained within, are aided by Tables A.1 and A.2. The feature sets are simplified to the following terms:

- All - All features
- Fo - First-order features only
- So - Second-order features only
- SigT - Significant features for the segmentation method from Student's t-test
- SigU - Significant features for the segmentation method from Mann-Whitney-Wilcoxon test
- MjTtest - Overall significant features from Student's t-test
- MjUtest - Overall significant features from Mann-Whitney-Wilcoxon test
- UnionTU - Union between overall significant features from Student's t-test and Mann-Whitney-Wilcoxon test
- PCAvX - Principle component analysis for the segmentation method with variance X captured

For the tables relating to classification results, they make use of the following shorthand:

- ANN - Artificial neural network with 3 hidden layers of 15 neurons each using stochastic gradient decent with a learning rate of 0.1.
- SVM-L - Support vector machine with a linear kernel.

- SVM-P - Support vector machine with a 3 degree polynomial kernel with a gamma value of 1.
- SVM-RBF - Support vector machine with a radial basis function kernel with a gamma value of 0.5.
- k-NN - k-Nearest neighbour with a k value of 3.
- NB - A Gaussian Naïve Bayes classifier.
- RF - Random forest with 150 estimators.
- Ada - AdaBoost with a 100 Linear SVM estimators and a learning rate of 0.1.

The feature sets MjTtest, MjUtest, and UnionTU all depend on the features deemed significant across all segmentation methods; meaning that they are significant in more segmentation methods than they are not. MjTtest is created from the results of Table A.1. The commonly significant features are: mean, variance, standard deviation, first order entropy, energy, contrast, correlation, second order entropy, sum entropy, difference variance, IMC1, and IMC2.

MjUtest is created from the results of Table A.2. The commonly significant features are: mean, variance, standard deviation, skewness, first order entropy, energy, contrast, correlation, second order entropy, sum entropy, sum variance, IMC1, IMC2, and SoS: variance.

UnionTU is created by identifying common features between MjTtest and MjUtest. This feature set comprises mean, variance, standard deviation, first order entropy, energy, contrast, correlation, second order entropy, sum entropy, IMC1, and IMC2.

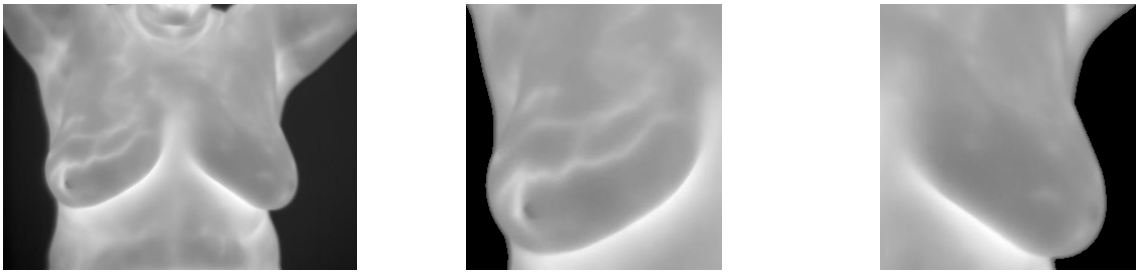
The following sections detail each segmentation method in turn. They each show an example of a patient image segmented by that method, as well as the results of their respective feature significance and classification tests.

7.3 Distance-based segmentation

This segmentation method is by far the simplest. It simply relies on the bulk of the breast tissue being in the centre of the image and within a certain area on the view port. In theory this method should catch the majority of the regions of interest within the box, as it is required by the capture standards to do so.

In practice, however, this method results in poor segmentation. Often, the breasts are cut off towards the lower end as the patient has elongated breasts, or too much of the non

breast tissue is shown; all of which skews results to a high degree. Some patient images were segmented out correctly, resulting in the regions of interest being mostly breast tissue with no breast cut off. Figure 7.1 illustrates an example of a patient image being correctly segmented out using the distance based method.



(a) Before segmentation (b) Left breast (c) Right breast

Figure 7.1: Patient image being segmented by the distance-based method

Even in these correctly segmented patient images there is still a great deal of unwanted tissue, and that cannot be removed by this method. The presence of unwanted tissue in the region of interest affects almost all features that need to be extracted, meaning that the resulting features are describing the tissue we are interested in as well as the tissue that has no value in breast cancer detection.

7.3.1 Feature extraction and selection

This section aims to detail and discuss the features extracted, and their significance, from images segmented using the distance-based segmentation method. Table 7.1 details the average extracted feature values with the associated p-values from the two tests of statistical significance conducted.

Table 7.1: Extracted features from the distance-based segmentation method with associated p-values

Feature	**Healthy average**	**Sick average**	**t-test (p<0.05)**	**MWW (p<0.05)**
Mean	6.182 ± 5.187	10.930 ± 11.007	**0.03997**	0.10578
Variance	85.118 ± 89.978	153.837 ± 144.580	**0.03387**	**0.01958**
Standard Deviation	1.871 ± 1.938	3.118 ± 2.466	**0.03646**	**0.01118**

Continued on next page

Table 7.1 – continued from previous page

Feature	Healthy average	Sick average	t-test (p<0.05)	MWW (p<0.05)
Skewness	0.246 ± 0.226	0.376 ± 0.302	0.06847	**0.03176**
Kurtosis	0.570 ± 0.461	0.738 ± 0.606	0.24036	0.21019
Fo Entropy	0.122 ± 0.124	0.174 ± 0.107	0.09390	**0.01304**
ASM (Energy)	0.030 ± 0.026	0.046 ± 0.027	**0.02486**	**0.00637**
Contrast	0.005 ± 0.003	0.005 ± 0.003	0.72088	0.23213
Correlation	0.005 ± 0.006	0.008 ± 0.005	0.06008	**0.00494**
Dissimilarity	0.004 ± 0.003	0.005 ± 0.003	0.55755	0.20177
Homogeneity	0.002 ± 0.002	0.002 ± 0.001	0.51707	0.16642
So Entropy	0.136 ± 0.130	0.183 ± 0.106	0.13966	**0.01460**
Sum Entropy	0.133 ± 0.129	0.181 ± 0.106	0.13012	**0.01516**
Difference Entropy	0.019 ± 0.013	0.020 ± 0.012	0.70542	0.32602
Sum Variance	0.564 ± 0.605	0.958 ± 0.936	0.06228	0.05116
Difference Variance	0.004 ± 0.003	0.004 ± 0.002	0.77374	0.26975
Sum Average	1.099 ± 1.950	0.925 ± 0.979	0.66854	0.26975
Autocorrelation	4.811 ± 8.868	4.157 ± 4.999	0.73048	0.22321
IMC1	0.012 ± 0.013	0.018 ± 0.012	0.08496	**0.01516**
IMC2	0.005 ± 0.006	0.007 ± 0.004	0.13913	**0.00918**
IDN	0.001 ± 0.001	0.001 ± 0.001	0.65113	0.41513
IDMN	0.000 ± 0.000	0.000 ± 0.000	0.27445	0.89422
SoS: Variance	0.141 ± 0.151	0.240 ± 0.234	0.06290	**0.04813**

There are three important observations to be made here. Firstly, the chosen p-value is 0.05, meaning that there needs to be 95% confidence in the rejection of the null hypothesis to be considered significant. If a p-value of 0.01 was chosen, like the methods which removed unwanted tissue underneath the breast, there would be almost no significant features.

The second thing to note is that there are very few significant features, especially for the Student's t-test. This further illustrates the problem with this segmentation method. In the patient images where part of the breast was cropped out; those parts of the breast could have contained vital predictive information. As previously mentioned, the unwanted tissue could skew the features where they quantify parts of the breast in which there is no interest. This could result in a possible malignancy going undetected because of some hot spot unrelated to the breast on the other side of the patient.

The third, and final, thing to note is that with some features the average for the healthy is higher than the average for the sick. This should not be the case, as all the features

measured should be higher, on average, in the sick group. This results in some features having incredibly poor p-values. All of these factors together result in a very poor feature set, which is used to train the classifiers.

7.3.2 Classification

This section details and discusses the classification results obtained when making use of the features extracted from images segmented with the distance-based method. Table 7.2 details the resulting accuracy obtained with a specific combination of feature set and classifier.

The impact of poor segmentation can clearly be seen here with most of the classifiers being no better than chance. This comes as no surprise as the features extracted were poor and a small portion were deemed significant. The highest performing combination, with an accuracy of 73.33%, was a random forest trained on a feature set comprising the significant features from a Mann-Whitney-Wilcoxon test.

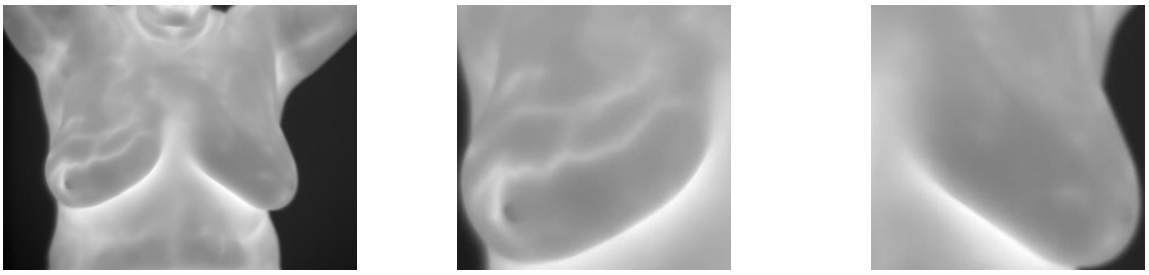
There are other combinations that provide an accuracy of approximately 70%, but overall performance is poor; especially for the AdaBoost algorithm, which achieved a best accuracy of 50%. This segmentation method is not recommended at all, since at best it could match the manual box crop if the images are captured perfectly.

Table 7.2: Accuracy (%) resulting from various combinations of features and classifiers on the distance-based segmentation method

Feature set	**ANN**	**SVM-L**	**SVM-P**	**SVM-RBF**	**k-NN**	**NB**	**RF**	**Ada**
All (23)	65	65	65	61.67	61.67	61.67	70	48.33
Fo (6)	65	65	58.33	65	58.33	61.67	65	48.33
So (17)	53.33	60	61.67	61.67	63.33	48.33	66.67	46.67
SigT (4)	63.33	56.67	60	61.67	50	58.33	65	50
SigU (11)	65	63.33	63.33	70	70	68.33	73.33	48.33
MjTtest (12)	66.67	66.67	60	66.67	66.67	63.33	66.67	48.33
MjUtest (14)	70	71.67	63.33	68.33	61.67	61.67	68.33	50
UnionTU (11)	63.33	66.67	58.33	68.33	65	65	70	50
PCAv99 (11)	60	68.33	53.33	61.67	58.33	66.67	66.67	50
PCAv95 (8)	63.33	70	55	61.67	58.33	66.67	66.67	46.67
PCAv90 (6)	56.67	55	58.33	61.67	60	63.33	56.67	50
PCAv85 (5)	60	56.67	60	63.33	60	56.67	50	46.67

7.4 Manual box crop segmentation

This segmentation method is essentially the manual version of the distance method. The user is required to crop to the region of interest, using bounds defined by experts. Figure 7.2 illustrates an example of a patient image being segmented out using this method.



(a) Before segmentation (b) Left breast (c) Right breast

Figure 7.2: Patient image being segmented by the manual box crop method

This method has a major advantage over the previous method; a guarantee that no breasts will be partially cropped out. With this method there is also a large limit to which unwanted tissue will appear, including most of the tissue under the breasts and the tissue above the armpits; which are areas that would otherwise skew results in cancer diagnosis as they have nothing to do with breast cancer. The centre of the image is far easier to attain as the left and right bounds are where the breast tissue stops, unlike the distance method, which would occasionally place the middle of the image out of alignment with the centre of the patient.

Once again, there is no method to remove the tissue directly under the inframammary folds. The removal of this tissue is vital for the accurate detection of breast cancer. However, this method removes far more unwanted tissue than the distance method and ensures that all the breast tissue is fully visible.

7.4.1 Feature extraction and selection

This section aims to detail and discuss the extracted features, along with their significance, from images segmented using the manual box crop segmentation method. Table 7.3 details these average feature values with their associated p-values from the two tests of statistical significance conducted.

Table 7.3: Extracted features from the manual box crop segmentation method with associated p-values

Feature	**Healthy average**	**Sick average**	**t-test (p<0.05)**	**MWW (p<0.05)**
Mean	6.786 ± 4.761	11.707 ± 10.677	0.02715	**0.04118**
Variance	402.868 ± 263.399	299.821 ± 267.765	0.14497	0.97127
Standard Deviation	5.296 ± 3.304	4.419 ± 3.838	0.35506	0.91881
Skewness	0.290 ± 0.314	0.776 ± 0.751	**0.00216**	**0.00319**
Kurtosis	1.326 ± 1.117	3.057 ± 3.068	**0.00596**	**0.00381**
Fo Entropy	0.127 ± 0.088	0.211 ± 0.135	**0.00650**	**0.00956**
ASM (Energy)	0.020 ± 0.014	0.050 ± 0.037	**0.00021**	**0.00056**
Contrast	0.007 ± 0.005	0.008 ± 0.007	0.46624	0.43250
Correlation	0.003 ± 0.002	0.004 ± 0.003	0.06544	0.06118
Dissimilarity	0.006 ± 0.004	0.007 ± 0.006	0.18762	0.18161
Homogeneity	0.003 ± 0.002	0.004 ± 0.003	0.16894	0.20177
So Entropy	0.143 ± 0.105	0.220 ± 0.144	**0.02273**	**0.01958**
Sum Entropy	0.138 ± 0.099	0.218 ± 0.141	**0.01501**	**0.01460**
Difference Entropy	0.024 ± 0.016	0.030 ± 0.024	0.21739	0.20596
Sum Variance	2.592 ± 1.691	1.795 ± 1.798	0.08717	0.98838
Difference Variance	0.006 ± 0.004	0.007 ± 0.006	0.51484	0.44415
Sum Average	1.009 ± 0.911	1.382 ± 1.267	0.20214	0.14864
Autocorrelation	6.204 ± 4.876	8.829 ± 7.582	0.12227	0.13535
IMC1	0.007 ± 0.006	0.020 ± 0.015	**0.00008**	**0.00004**
IMC2	0.003 ± 0.002	0.006 ± 0.005	**0.00028**	**0.00008**
IDN	0.000 ± 0.000	0.001 ± 0.000	0.29897	0.26976
IDMN	0.000 ± 0.000	0.000 ± 0.000	0.94293	0.64141
SoS: Variance	0.649 ± 0.423	0.449 ± 0.450	0.08594	0.98838

As with the distance-based method, a p-value of 0.05 was chosen. This 95% confidence interval produced more significant features for the Student's t-test and fewer significant features for the MWW test than the distance method. The p-values for this segmentation method vary largely; with some being as low as 0.00021 and some as high as 0.97127.

An important observation is that the first order variance (p=0.97127) and standard deviation (p=0.91881) are higher in the healthy group than in the sick. Although this segmentation method is far better than the distance-based method, in terms of the number of unexpected reversals, this is still an indication that poor segmentation being performed. This reversal

of expected averages is testament to the influence that unwanted tissue has on the resulting features.

Temperature fluctuation present in those unwanted regions influence the distribution of pixel intensities of the histogram. For healthy patients, these unwanted temperature distributions result in the captured image seeming to belong to a sick patient; which in turn influence the resulting statistics taken from said histogram. This manifests as a reversal of expectations when examining the average of the feature groups. It further illustrates the need to remove unwanted tissue from the region of interest accurately.

7.4.2 Classification

This section details and discusses the classification results obtained when making use of the features extracted from images segmented with the manual box crop method. Table 7.4 details the resulting accuracy obtained with a specific combination of feature set and classifier.

Although this segmentation method yielded far better results on average than the distance-based one, the results are mediocre to good at best. The top combinations yielded accuracies from mid 70s to low 80s. The highest performer, with an accuracy of 83.33%, was a random forest classifier trained on a feature set comprising the overall significant features from the Student's t-test.

The poorest performing classifier for this segmentation method was the k-nearest neighbour with the remainder of the classifiers performing relatively close to one another. It can be seen that making use of all the available features leads to an overly complex machine learning model with poor generalisation abilities, and that simply making use of the first order features does not capture enough information. This shows that a subset of the full feature set is best suited for this segmentation method.

Table 7.4: Accuracy (%) resulting from various combinations of features and classifiers on the manual box crop segmentation method

Feature set	**ANN**	**SVM-L**	**SVM-P**	**SVM-RBF**	**k-NN**	**NB**	**RF**	**Ada**
All (23)	70	75	75	71.67	68.33	70	73.33	71.67
Fo (6)	68.33	66.67	70	70	56.67	70	66.67	70
So (17)	73.33	78.33	71.67	70	65	71.67	76.67	76.67
SigT (8)	76.67	81.67	75	70	71.67	70	73.33	76.67

Continued on next page

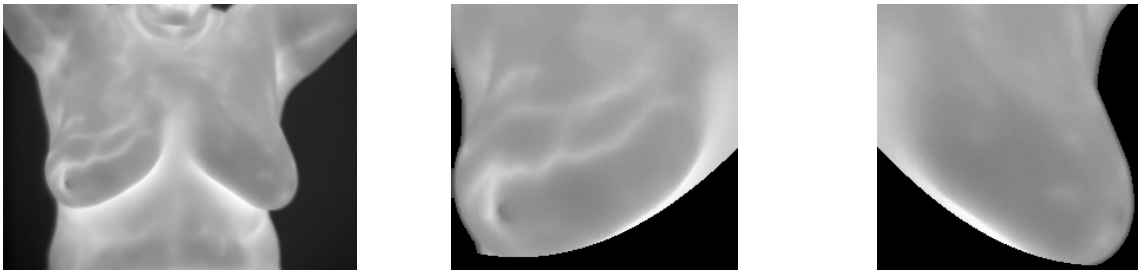
Table 7.4 – continued from previous page

Feature set	ANN	SVM-L	SVM-P	SVM-RBF	k-NN	NB	RF	Ada
SigU (9)	76.67	78.33	71.67	71.67	73.33	73.33	71.67	71.67
MjTtest (12)	76.67	78.33	73.33	73.33	65	75	83.33	78.33
MjUtest (14)	80	76.67	73.33	75	68.33	75	75	81.67
UnionTU (11)	78.33	78.33	75	71.67	68.33	73.33	80	75
PCAv99 (12)	75	75	63.33	71.67	71.67	65	70	73.33
PCAv95 (8)	75	76.67	63.33	76.67	66.67	66.67	60	73.33
PCAv90 (6)	80	78.33	66.67	75	68.33	70	73.33	75
PCAv85 (5)	75	81.67	68.33	78.33	60	70	66.67	80

7.5 Semi-automatic segmentation

Whereas the previous two segmentation methods did little to remove the tissue below the inframammary fold, this method set out to do just that. It is an augmentation of the manual box crop method and thus can be directly compared to it, in order to understand the difference that the removal of unwanted tissue has on the results achieved.

This method took the manually cropped images and identified the inframammary folds by fitting two second degree polynomials to the curves using regression, and removing all the tissue below those polynomials. Figure 7.3 illustrates an example of a patient image being segmented out using this method. Figure 7.4 shows the detected inframammary folds for regular and irregular breasts. For patient images where there was not a sufficient curve detected, the image was left at the manual cropped stage. This is to account for patients with smaller breasts and thus having no clear inframammary fold.



(a) Before segmentation (b) Left breast (c) Right breast

Figure 7.3: Patient image being semi-automatically segmented

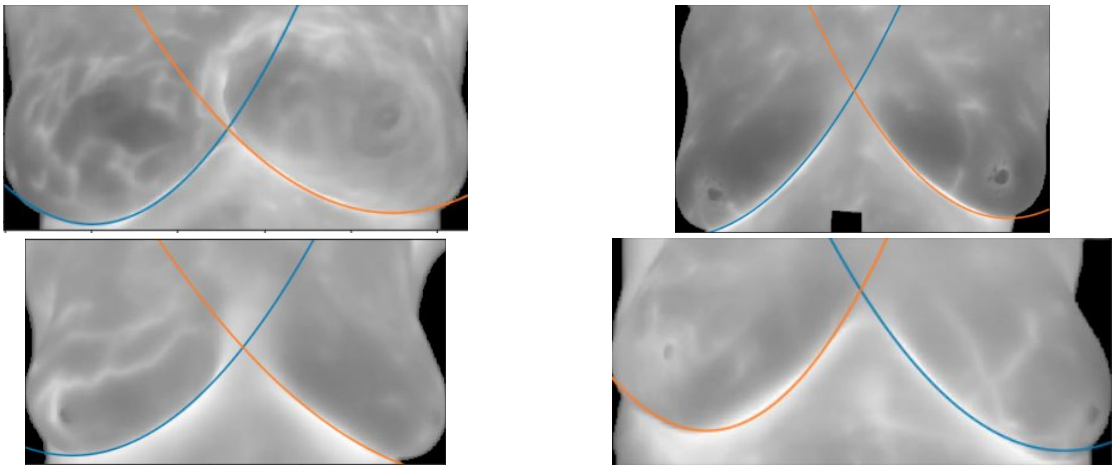


Figure 7.4: The detected inframammary fold for breasts of varying shape

This method greatly reduces the amount of unwanted tissue. This results in features that are far more indicative of the actual breast tissue being examined and are not being skewed by unwanted tissue present in the region of interest.

7.5.1 Feature extraction and selection

This section aims to detail and discuss the extracted features, along with their significance, from images segmented using the semi-automatic method. Table 7.5 details the average feature values along with the associated p-values from the two tests of statistical significance conducted.

Table 7.5: Extracted features from the semi-automatic segmentation method with associated p-values

Feature	**Healthy average**	**Sick average**	**t-test (p<0.01)**	**MWW (p<0.01)**
Mean	5.687 ± 5.223	12.237 ± 11.656	**0.00769**	**0.00561**
Variance	40.944 ± 28.369	161.554 ± 141.334	**0.00003**	**0.00003**
Standard Deviation	0.957 ± 0.661	3.465 ± 2.473	**0.00001**	**0.00001**
Skewness	0.268 ± 0.203	0.446 ± 0.424	0.04615	0.05768
Kurtosis	0.864 ± 0.575	1.625 ± 1.828	0.03668	0.05941
Fo Entropy	0.049 ± 0.047	0.206 ± 0.127	**0.00001**	**0.00001**
ASM (Energy)	0.016 ± 0.013	0.054 ± 0.045	**0.00005**	**0.00005**
Contrast	0.005 ± 0.003	0.009 ± 0.006	**0.00132**	**0.00266**

Continued on next page

Table 7.5 – continued from previous page

Feature	Healthy average	Sick average	t-test (p<0.01)	MWW (p<0.01)
Correlation	0.003 ± 0.003	0.010 ± 0.007	**0.00001**	**0.00001**
Dissimilarity	0.005 ± 0.003	0.008 ± 0.006	**0.00193**	**0.00119**
Homogeneity	0.002 ± 0.001	0.004 ± 0.003	**0.00270**	**0.00152**
So Entropy	0.063 ± 0.050	0.213 ± 0.149	**0.00001**	**0.00003**
Sum Entropy	0.060 ± 0.050	0.212 ± 0.143	**0.00001**	**0.00001**
Difference Entropy	0.020 ± 0.012	0.035 ± 0.024	**0.00237**	**0.00192**
Sum Variance	0.221 ± 0.217	1.031 ± 0.896	**0.00001**	**0.00001**
Difference Variance	0.004 ± 0.003	0.008 ± 0.005	**0.00149**	**0.00333**
Sum Average	0.523 ± 0.487	1.347 ± 1.651	0.01256	**0.00721**
Autocorrelation	3.591 ± 3.349	9.214 ± 10.839	**0.00985**	**0.00956**
IMC1	0.012 ± 0.008	0.024 ± 0.017	**0.00069**	**0.00138**
IMC2	0.003 ± 0.002	0.012 ± 0.009	**0.00001**	**0.00001**
IDN	0.000 ± 0.000	0.001 ± 0.001	**0.00356**	0.01633
IDMN	0.000 ± 0.000	0.000 ± 0.000	0.03711	0.06028
SoS: Variance	0.055 ± 0.054	0.258 ± 0.224	**0.00002**	**0.00001**

The first observation to be made here is that of degrees of confidence. This segmentation method produced features resulting in very small p-values, which indicated that there is a strong degree of confidence that the differences seen between the groups are significant. If the previously used p-value of 0.05 was used, all but one feature would be deemed significant. Therefore, a stricter condition can be safely imposed. For this reason, a p-value of 0.01 is chosen as the threshold for significance; meaning that the confidence threshold is 99%.

The number of features seen as significant in Table 7.5 gives strong indication of proper segmentation. A large number of significant features are achieved with very high confidence values from both of the tests of statistical significance. It is also important to note that the two different tests of significance agree on all but one feature. Another note of the averages of the healthy group are lower than the averages of the sick group. These expected observations are strong indications that proper segmentation is being performed.

7.5.2 Classification

This section details and discusses the classification results obtained when making use of the features extracted from images segmented with the semi-automatic method. Table 7.6 details the resulting accuracy obtained with a specific combination of feature set and classifier.

The results here show the benefit of the removal of tissue under the inframammary fold. The only difference between this method and the manual box crop is the removal of that tissue. This means the classification results can be directly compared to see how this change affects the accuracy.

Overall, the majority of combinations resulted in accuracies upwards of 80% with some combinations reaching as high as 90%. This is far superior to the manual box crop method, which saw the majority of results being in the low to mid 70s. Two combinations resulted in an accuracy of 90%: a random forest trained on the feature set produced by performing principle component analysis with 99% of the variance maintained, and the chosen artificial neural network trained on a feature set comprising the union between the overall significant features from both the Student's t-test and the Mann-Whitney-Wilcoxon test.

The lowest performing classifier for this segmentation method was the support vector machine with a polynomial kernel. It achieved results mostly in the low to mid 70s with one as high as 80%. The remaining classifiers all performed relatively close to one another. Another interesting note is that the full feature set performed adequately with this segmentation method, something not seen with the two previous methods. This could be the result of the removal of more unnecessary tissue, leading to a resulting feature set that is not skewed toward representing unimportant tissue. Another factor that agrees with the previous statement is that the significant features from both the tests of significance resulted in 19 features being deemed significant. This is much higher than the previously discussed methods, and means that more features actually represent a difference between the two groups in a significant way.

Overall, there is no clear weak feature set as they all performed well in combination with some classifiers. The weaker performers were the first order features as well as the significant features resulting from the Mann-Whitney-Wilcoxon test. The rest attained a good to excellent performance.

Table 7.6: Accuracy (%) resulting from various combinations of features and classifiers on the semi-automatic segmentation method

Feature set	**ANN**	**SVM-L**	**SVM-P**	**SVM-RBF**	**k-NN**	**NB**	**RF**	**Ada**
All (23)	86.67	81.67	75	85	83.33	83.33	86.67	80
Fo (6)	73.33	80	75	76.67	83.33	80	80	80
So (17)	88.33	85	78.33	85	83.33	81.67	81.67	85
SigT (19)	86.67	83.33	75	83.33	83.33	83.33	88.33	78.33

Continued on next page

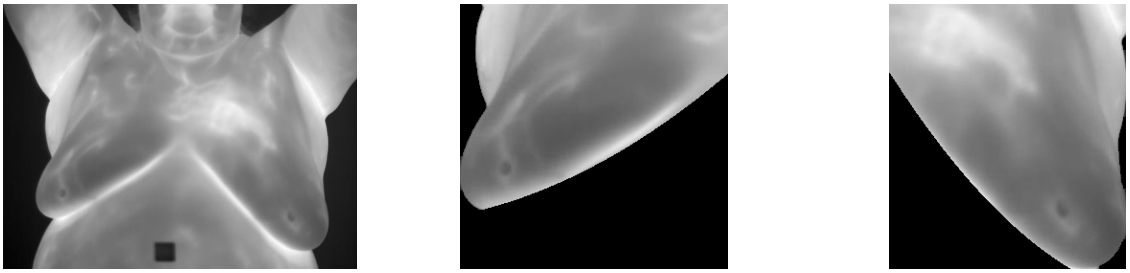
Table 7.6 – continued from previous page

Feature set	ANN	SVM-L	SVM-P	SVM-RBF	k-NN	NB	RF	Ada
SigU (19)	80	80	73.33	81.67	80	81.67	75	75
MjTtest (12)	86.67	86.67	80	85	78.33	85	85	86.67
MjUtest (14)	83.33	85	78.33	81.67	80	81.67	85	85
UnionTU (11)	90	88.33	76.67	83.33	80	83.33	86.67	85
PCAv99 (12)	86.67	81.67	71.67	85	83.33	85	90	80
PCAv95 (7)	76.67	83.33	71.67	85	81.67	85	86.67	83.33
PCAv90 (5)	80	85	71.67	85	81.67	81.67	88.33	80
PCAv85 (4)	81.67	85	73.33	85	86.67	81.67	88.33	76.67

7.6 Fully manual segmentation

The semi-automatic approach solved most of the issues present with the methods that do not remove the inframammary fold and its surrounding tissue. The fully manual approach is an attempt at excluding all unwanted tissue perfectly by having the user create polynomials to match the inframammory folds.

For the most part, this should be indistinguishable from the semi-automated approach, but for some cases this method removes tissue the semi automatic approach does not. It is important to note that this method is incredibly time-consuming and may not result in perfect segmentation all the time. There are cases where there is no clear inframammary fold, which could result in important tissue being removed by accident. Figure 7.5 illustrates an example of a patient image being segmented out using this method.



(a) Before segmentation (b) Left breast (c) Right breast

Figure 7.5: Patient image being manually segmented

7.6.1 Feature extraction and selection

This section aims to detail and discuss the extracted features, along with their significance, from images segmented using the fully manual method. Table 7.7 details the average feature values extracted along with the associated p-values from the two tests of statistical significance conducted.

Table 7.7: Extracted features from the fully manual segmentation method with associated p-values

Feature	**Healthy average**	**Sick average**	**t-test (p<0.01)**	**MWW (p<0.01)**
Mean	5.776 ± 5.134	13.590 ± 14.105	**0.00687**	**0.00637**
Variance	48.077 ± 38.507	217.886 ± 467.432	0.05605	0.00025
Standard Deviation	1.110 ± 0.870	3.421 ± 3.457	**0.00093**	**0.00013**
Skewness	0.303 ± 0.232	0.657 ± 0.611	**0.00506**	**0.00721**
Kurtosis	1.361 ± 1.594	3.658 ± 4.537	0.01272	0.02179
Fo Entropy	0.055 ± 0.051	0.188 ± 0.127	**0.00001**	**0.00001**
ASM (Energy)	0.017 ± 0.014	0.054 ± 0.044	**0.00005**	**0.00001**
Contrast	0.006 ± 0.006	0.023 ± 0.022	**0.00012**	**0.00008**
Correlation	0.005 ± 0.006	0.022 ± 0.028	**0.00210**	**0.00005**
Dissimilarity	0.005 ± 0.003	0.010 ± 0.008	**0.00099**	**0.00138**
Homogeneity	0.002 ± 0.001	0.005 ± 0.003	**0.00115**	**0.00072**
So Entropy	0.067 ± 0.057	0.200 ± 0.145	**0.00002**	**0.00005**
Sum Entropy	0.063 ± 0.057	0.194 ± 0.140	**0.00002**	**0.00006**
Difference Entropy	0.020 ± 0.013	0.040 ± 0.027	**0.00099**	**0.00053**
Sum Variance	0.254 ±0.241	1.310 ± 2.787	0.04674	**0.00007**
Difference Variance	0.005 ± 0.005	0.022 ± 0.021	**0.00014**	**0.00010**
Sum Average	0.531 ± 0.479	1.086 ± 1.155	0.02023	0.02506
Autocorrelation	3.645 ± 3.291	7.571 ± 8.572	0.02490	0.02873
IMC1	0.012 ± 0.008	0.028 ± 0.022	**0.00034**	**0.00029**
IMC2	0.003 ± 0.003	0.013 ± 0.011	**0.00002**	**0.00001**
IDN	0.000 ± 0.000	0.001 ± 0.001	**0.00841**	0.03073
IDMN	0.000 ± 0.000	0.000 ± 0.000	**0.00586**	0.01304
SoS: Variance	0.064 ± 0.060	0.329 ± 0.699	0.04651	**0.00006**

For this segmentation method a p-value of 0.01 was chosen as the threshold. It was for the same reason that it was chosen for the semi automatic method: the tissue remaining

is almost only the tissue we are interested in, resulting in a much higher overall confidence rating with decisions. The decisions of the two significance tests largely agreed with one another once again, with a large portion of the feature set being deemed significant.

7.6.2 Classification

This section details and discusses the classification results obtained when making use of the features extracted from images segmented with the fully manual method. Table 7.8 details the resulting accuracy obtained with a specific combination of feature set and classifier for this segmentation method.

Some interesting results arise from this segmentation method. Most notably, when compared to the semi-automatic segmentation method, these results are somewhat lacklustre. Most of the results fall between the high 70s and low 80s with only a few breaking higher or lower.

The best result for this segmentation method, with an accuracy of 90%, was the artificial neural network trained on a feature set comprising a union between the overall significant features from both the Student's t-test and the Mann-Whitney-Wilcoxon test. In fact, the artificial neural network was clearly the best performer out of all the classifiers, which is a first considering the segmentation methods discussed. Another oddity is that the first order feature set performed much worse than the full feature set or the second order feature set, once again differing from the trend seen thus far.

The feature set extracted seems to be quite linearly separable. This is due to the linear support vector machine and the AdaBoost algorithm performing better among the remaining classifiers. This is the first segmentation method that has produced this phenomenon; where the algorithms best suited to linearly separable data outperform the rest.

This strange behaviour could be due to the level of human interference. There are some patient images where there is no clear inframammary fold; thus, in order to segment them out manually, the user has to estimate where the curves would be; possibly removing important tissue from the image. With with the automated approaches, however, if this fold is not detected, the entire breast region is left in-tact. With all that in mind, this segmentation method performed very well, yet it seems that it is not worth taking the extra time to segment out each breast manually.

Table 7.8: Accuracy (%) resulting from various combinations of features and classifiers on the fully manual segmentation method

Feature set	ANN	SVM-L	SVM-P	SVM-RBF	k-NN	NB	RF	Ada
All (23)	85	85	78.33	81.67	80	81.67	83.33	80
Fo (6)	70	75	68.33	75	71.67	73.33	68.33	76.67
So (17)	83.33	80	80	80	73.33	81.67	80	83.33
SigT (17)	80	85	78.33	80	76.67	81.67	83.33	81.67
SigU (17)	88.33	85	76.67	81.67	76.67	81.67	81.67	88.33
MjTtest (12)	88.33	85	80	83.33	75	81.67	81.67	86.67
MjUtest (14)	86.67	83.33	78.33	81.66	81.67	81.67	78.33	88.33
UnionTU (11)	90	83.33	78.33	81.66	75	81.67	83.33	85
PCAv99 (11)	86.67	83.33	70	81.67	78.33	80	78.33	86.67
PCAv95 (7)	75	80	70	78.33	80	85	76.67	80
PCAv90 (5)	75	80	70	76.67	81.67	78.33	76.67	80
PCAv85 (4)	81.67	80	70	76.67	75	80	75	81.67

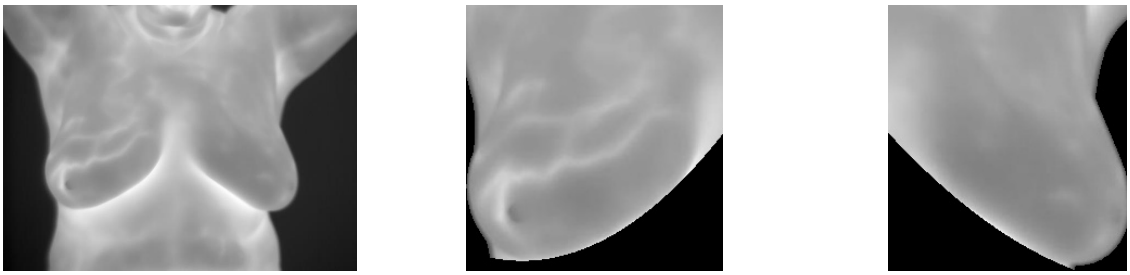
7.7 Fully automatic segmentation

This method attempts to combine the best aspects of all the previous methods; a low reliance on user input and the accurate removal of unwanted tissue. This method is the least time consuming but the most prone to error as the ambiguity of breast makes this process challenging.

In fact, eight patient images had to be excluded from the feature calculation as their segmentation was so poor that it would be useless to try to extract features. This method has many points in the process at which a slight variation could result in poor segmentation. If the boundaries are detected poorly, unwanted hot regions could skew the fitting of the polynomials to the inframammary folds for example. Figure 7.6 illustrates an example of a patient image being segmented out using this method.

7.7.1 Feature extraction and selection

This section aims to detail and discuss the extracted features, along with their significance, from images segmented using the fully automatic method. Table 7.9 details the average feature values extracted along with the associated p-values from the two tests of statistical significance conducted.



(a) Before segmentation (b) Left breast (c) Right breast

Figure 7.6: Patient image being automatically segmented

Once again a p-value of 0.01 was chosen as the threshold. This resulted in a large portion of the features being deemed significant with the two tests largely agreeing with each other. It is important to note that, while the significant features had p-values under 0.01, the semi-automatic and fully manual methods had p-values far lower than those obtained with this method. This could be a result of some less than perfect segmentation being done with the fully automatic approach.

Table 7.9: Extracted features from the fully automatic segmentation method with associated p-values

Feature	**Healthy average**	**Sick average**	**t-test (p<0.01)**	**MWW (p<0.01)**
Mean	5.565 ± 5.177	12.442 ± 12.106	**0.00843**	0.01069
Variance	52.969 ± 32.653	184.434 ± 164.297	**0.00013**	**0.00015**
Standard Deviation	1.279 ± 0.791	3.829 ± 2.790	**0.00002**	**0.00004**
Skewness	0.257 ± 0.258	0.473 ± 0.449	0.03365	0.01390
Kurtosis	0.803 ± 0.710	1.658 ± 1.578	0.01244	**0.00676**
Fo Entropy	0.086 ± 0.082	0.213 ± 0.108	**0.00001**	**0.00001**
ASM (Energy)	0.024 ± 0.023	0.059 ± 0.041	**0.00030**	**0.00016**
Contrast	0.005 ± 0.003	0.009 ± 0.007	**0.00944**	**0.00977**
Correlation	0.004 ± 0.004	0.011 ± 0.008	**0.00011**	**0.00001**
Dissimilarity	0.005 ± 0.003	0.008 ± 0.006	0.01966	0.01718
Homogeneity	0.003 ± 0.002	0.004 ± 0.003	0.02613	0.01945
So Entropy	0.107 ± 0.090	0.222 ± 0.128	**0.00035**	**0.00059**
Sum Entropy	0.103 ± 0.089	0.221 ± 0.123	**0.00019**	**0.00028**
Difference Entropy	0.022 ± 0.013	0.035 ± 0.025	0.01785	0.01867

Continued on next page

Table 7.9 – continued from previous page

Feature	Healthy average	Sick average	t-test (p<0.01)	MWW (p<0.01)
Sum Variance	0.366 ± 0.297	1.161 ± 1.007	**0.00021**	**0.00025**
Difference Variance	0.005 ± 0.003	0.008 ± 0.006	**0.00822**	0.01069
Sum Average	0.660 ± 1.134	0.994 ± 0.983	0.26013	0.04353
Autocorrelation	4.171 ± 6.272	6.928 ± 7.065	0.13828	0.05748
IMC1	0.013 ± 0.008	0.025 ± 0.019	**0.00333**	**0.00977**
IMC2	0.004 ± 0.004	0.012 ± 0.008	**0.00001**	**0.00001**
IDN	0.000 ± 0.000	0.001 ± 0.001	0.05785	0.03369
IDMN	0.000 ± 0.000	0.000 ± 0.000	0.86610	0.43638
SoS: Variance	0.092 ± 0.074	0.290 ± 0.251	**0.00021**	**0.00025**

It is important to keep in mind that eight of the sixty patient images (13%) could not be segmented with this method; meaning that the resulting regions of interest were completely missing the wanted tissue. This exclusion needs to be considered when looking at the results of the classifiers, as those eight were not used for training and testing.

7.7.2 Classification

This section details and discusses the classification results obtained when making use of the features extracted from images segmented with the fully automatic method. Table 7.10 details the resulting accuracy obtained with a specific combination of feature set and classifier for this segmentation method.

When comparing this method to the semi-automatic and fully manual methods this method clearly performs the worst. Relatively few combinations yielded accuracies above 80%, with the majority being in the mid 70s. Four combinations achieved an accuracy of 83.64%. They are an artificial neural network trained using the significant features resulting from the Student's t-test; a linear support vector machine trained on all of the second order features; k-nearest neighbour trained on a feature set comprising a union between the overall significant features from the Student's t-test and the Mann-Whitney-Wilcoxon test; and the AdaBoost algorithm trained on the overall significant features from the Student's t-test.

An observation about those best performers is that their feature sets did not come from specific selected features of the segmentation method. This means that the features deemed significant for this segmentation method are not the best performers. A reason for this is that the automatic segmentation method could be removing tissue that should be present, resulting in a feature set that does not capture the phenomenon properly.

This statement is further supported by the visual inspection of the segmented images. Many have important sections of tissue cut out. This is due to the many layers of automation, which rely on each other, and one mistake cascading down can result in a poor segmentation. It was previously stated that 13% of the patient images were left out owing to improper segmentation. From the results seen here, a fully automatic approach is not recommended.

Table 7.10: Accuracy (%) resulting from various combinations of features and classifiers on the fully automatic segmentation method

Feature set	**ANN**	**SVM-L**	**SVM-P**	**SVM-RBF**	**k-NN**	**NB**	**RF**	**Ada**
All (23)	74.54	74.55	72.72	74.55	76.36	74.55	74.55	76.36
Fo (6)	65.45	72.73	69.09	74.55	72.73	74.55	76.36	74.55
So (17)	80	83.64	67.27	76.36	74.55	80	74.55	80
SigT (14)	74.55	76.36	74.55	76.36	81.82	78.18	76.36	78.18
SigU (13)	83.64	81.82	74.55	80	78.18	80	78.18	81.82
MjTtest (12)	78.18	76.36	72.73	80	81.82	81.82	74.55	83.64
MjUtest (14)	78.18	76.36	70.91	76.36	78.18	78.18	76.36	76.36
UnionTU (11)	76.36	76.36	70.91	80	83.64	80	76.36	78.18
PCAv99 (11)	72.73	74.55	67.27	74.55	74.55	78.18	78.18	76.36
PCAv95 (7)	69.09	74.55	67.27	74.55	80	74.55	80	76.36
PCAv90 (5)	67.27	78.18	63.36	72.72	76.36	78.18	76.36	80
PCAv85 (4)	69.09	80	67.27	76.36	76.36	78.18	76.36	74.55

7.8 Conclusion

This chapter set out to detail the results of the implementation of the model. The goal was to find the best segmentation method, the best feature reduction technique, and the best classifier for the chosen dataset. It began with an overall result for the tests of significance, which are used to create the feature sets.

Overall, five segmentation methods were tested using twelve features sets, which were used to train eight classifiers. The resulting accuracies were tabulated. The distance-based segmentation method performed by far the worst with results no better than simply guessing. The manual box crop segmentation method was an improvement but still not much better. A large difference in accuracy was observed when methods removing the tissue below the inframammary fold were tested.

The fully automatic segmentation method is not recommended as a portion of the patient

images had to be excluded owing to improper segmentation. Resulting in a real-world consequence of constant negative results for patients where their images cannot be segmented out. This leaves the semi-automatic and fully manual segmentation methods.

From the remaining methods, the semi-automatic method is preferred. This is due to two important factors. Less time needs to be invested per patient as they do not have to trace out the inframammary fold, and overall accuracies are higher and far more consistent. The feature set and classifier combination, based on accuracy results, should either combine the configured artificial neural network with the union between the overall significant features from both the Students t-test and the Mann-Whitney-Wilcoxon test, or a random forest trained with a feature set resulting from performing principle component analysis with 99% of the variance maintained. However, a random forest trained using PCA with 99% variance maintained is recommended. This is due to the ease of configuration for both the classifier and the features via PCA. The python script used to train the various classifiers on features from semi-automatic segmentation using PCA can be found in Appendix C.

Comparing the results obtained here with other related works seen in Table 6.1, it can be seen that the configuration used in this study performs relatively well. The best accuracy of 90% obtained in this study is better than most of the studies listed, and matches all but a few. The two highest performing studies, with accuracies of 96.12% and 94.87%, used private data sets with larger number of patients.

The next chapter concludes this research. It restates the objectives stated in Chapter 1 and explains how each objective was met. It then proceeds to provide closing thoughts on the research, on what was learnt, and makes recommendations for future work.

Chapter 8

Conclusion

8.1 Introduction

The early detection of breast cancer is crucial as it impacts survival rates significantly. Urban and high-resource areas are capable of providing early detection with established medical equipment, which is used to perform breast imaging using current techniques. Rural or low-resource areas do not always have access to medical resources and, as such, aren't able to perform early detection. Mobile healthcare and highly portable methods of breast imaging like thermography are able to provide much needed medical resources to these areas with poor medical infrastructure. Furthermore, machine learning applied to thermography improves its diagnostic capabilities significantly by offering an objective second opinion.

Initial research suggested that there was no clear method to integrate mobile technology, machine learning, and thermography in the detection of breast cancer. Therefore the need exists to create a model, which could guide that integration; resulting in the model described in Chapter 5. This study provided an implementation of the model through experimentation, in Chapter 6, and the results of that experiment were discussed in Chapter 7. The experiments conducted showed the validity of performing classification on thermal images, which could be captured by attachable mobile cameras that are available today. High accuracy can be attained with minimal user input in a short period of time. This is highly beneficial in low-resource areas especially because the model allows this process to be done entirely offline with the only connection needed being to download the mobile application.

This chapter begins with an overview of the chapters in this study. The objectives from Chapter 1 are then restated and it was shown how each secondary objective was met. The problem statement is then revisited, and it is shown how the primary objective solves the problem. The limitations of the research are then discussed followed by future research

possibilities.

8.2 Chapter overview

Chapter 1 provided the necessary background information to the identified problem. This chapter established the research problem, and the research objectives that needed to meet the research goals of the study. The research methods were also discussed, and a research process was outlined.

Chapter 2 began with an introduction to infrared light, its physical properties, and how it can be detected. It detailed the types of infrared detectors and how they are arranged to form camera systems. This chapter then introduced the concept of temperature as an indicator of health. Thermal equilibrium mechanics were explored, where it was shown how the skin and blood provide most of the thermal regulation, and that measuring the skins temperature gives an indication of the underlying physiology.

Chapter 3 introduced cancer, more specifically breast cancer. It explained how breast cancer forms, what types of breast cancer exist, and how breast cancer is detected and treated. This chapter then introduced thermography as a method to detect breast cancer. It explained the physiological events that take place as cancer forms, as well as what thermography looks for in its diagnosis. A history of thermography was then presented. The standards that need to be adhered to when performing a thermographic exam were detailed as well as signatures to look for when performing the examination.

Chapter 4 explored machine learning and classification. It began with an introduction to the various methods by which learning occurs, where machine learning is used, the various stages of a machine learning solution, and the types of machine learning systems. It then detailed the bias-variance trade-off and how dimensionality reduction is used to help balance the trade-off. Classification was then introduced and common classifiers used within classification problems were, in turn, discussed with a focus on how they function. Various validation methods were discussed; also, sampling methods, and metrics used to analyse the performance of a classifier.

Chapter 5 focused on the creation of the conceptual model from literature. It gave an overview of the full model then went into detail about how each component was created based on the research done.

Chapter 6 detailed the experimental procedure undertaken to create an implementation of the proposed conceptual model. It detailed what images are used, the methods of segmentation employed, the features extracted, and the classification methods used to diagnose

breast cancer from thermal images of breasts.

Chapter 7 detailed the results of the experimental process and drew conclusions on which combinations work best for the given data and problem. This chapter also served as the remainder of the implementation of the model as it was shown that a high accuracy can be achieved employing the steps proposed in the model.

8.3 Research objectives

This section looks at how the primary and secondary objectives were met in this study.

8.3.1 Secondary objectives

The first secondary objective that needed to be met was to *Identify which aspects of breast cancer are detectable by thermal imaging using mobile devices.* This objective was met with literature reviews in Chapter 2 and Chapter 3. Chapter 2 explained how thermal cameras measure the temperature of the skin, and how the advancement of thermal cameras allows attachable mobile devices to perform on par with traditional thermal cameras. The smartphone environment is then detailed to give an understanding of the available configurations. An explanation of how the human body maintains thermal equilibrium, and that a temperature differential outside of normal distributions is cause for concern, is given. Chapter 3 Section 3.3 then described how breast cancer forms and detailed what physiological traits are present when the cancer is growing. Section 3.4 then described how thermal imaging is used to observe these physiological traits and signals.

The second secondary objective that needed to be met was to *Contrast segmentation methods for the extraction of the breast region from thermal images.* This objective was met, in part, in Chapter 6, by identifying relevant methods from literature . The contrasting of the identified methods was completed in Chapter 7, where the best performing method was identified.

The third secondary objective that needed to be met was to *Contrast machine learning techniques to be used for the classification of thermal breast images.* This objective was met, in part, with a literature review in Chapter 4, which covered classification and how to approach classification problems. Specific candidates for the application of machine learning to thermal breast image analysis were identified though literature in Chapter 6. These identified techniques were contrasted in Chapter 7, where the best performing classifier was identified.

8.3.2 Primary objective

The primary objective of this study was to *Develop a model for the application of machine learning for the automated classification of thermal breast images in a mobile environment.* The primary objective was met in Chapter 5 where a model was outlined in Figure 5.1; consisting of a server-side and client-side component. The server-side component was created using the literature review in Chapter 4. The server-side aspect of the model was implemented as a prototype in Chapter 6 and Chapter 7. The client side of the model was created from literature reviews in Chapter 2 and Chapter 3.

8.4 Problem statement revisited

The main problem addressed in this study is *There is no widely accepted process to facilitate the integration of breast thermography and machine learning techniques with a mobile platform.* Based on this problem statement, a model was presented that provides a possible solution to how mobile technology can be integrated with machine learning and thermography to detect breast cancer. The model was designed with portability and ease of use in mind, and thus provides the ability to perform classification in areas where there may be no signal.

8.5 Research limitations

The extent of the testing of the prototype was limited to a machine learning pipeline using images from a data set with the images captured having the same quality achievable by mobile attachable cameras. The experiments were not conducted in the real world. Meaning that the trained machine learning algorithms were not exposed to images outside of the data set used. In real world testing there may be images taken not adhering to all the standards. This study only focused on texture analysis to extract features, but there are other methods to which features could be extracted, which may result in different outcomes.

The largest limitation was sample size. For thermography, it is difficult to find large data sets that contain good images to be used for training. All large data sets available, consisting of hundreds of images, contain a massive class imbalance; heavily favouring healthy people. This bias reflects the real world as you cannot expect the sick to equal the healthy. The problem lies in the implementation of machine learning algorithms as it may introduce a learning bias, which favours classifying images as healthy. To address this, a pruning activity was required to achieve a balanced data set, which led to the small sample size used in this

study.

Only images of the patient taken from the front, with as little angular deviation as possible were used, this was due to a limitation in the segmentation process. The automatic methods proposed in this study can not be applied to images other than frontal. Many data sets contain images at diagonal angles which provide more tissue that can be analysed. This could help to identify a cancer that has formed on the side or near the rear of the breast tissue that a frontal image could not show.

8.6 Future research

Future research should first and foremost be focused on creating a large data set of images captured according to standards and accurately labelled in order to be used for machine learning training. Only once there is a large data set can definitive statements be made surrounding the viability of the solutions.

This study tested 480 combinations of segmentation methods, extracted features, and classification algorithms. There are many more that could be tested; for instance, different feature extraction techniques aside from texture analysis like spectral analysis or fractal analysis. There are other ways of selecting relevant features, one example is using a genetic algorithm to select the best subset of features.

One area from which thermography can benefit enormously, as mammography does, is classification using deep neural networks; or deep learning. This is only possible once a large enough data set is presented since deep learning results in incredibly complex neural networks, which are excellent at generalising if given enough data to learn from.

8.7 Conclusion

This chapter concluded the study and ensured that all the research objectives stated in Chapter 1 were met. The primary and secondary research objectives were stated and an overview of each chapter in this study was provided. Research limitations that were experienced during the course of the study were also discussed. Future research possibilities were outlined in order to improve the research topic area. The purpose of this study was to create a tool that can be used to create integrated mobile solutions to breast cancer detection. This study identified that there is a need for cheap, portable, easy solutions to the early detection of breast cancer in low-resource areas, and that the integration of mobile technology with machine learning and thermography can be that solution. The model created in Chapter 5

can only be successful if standards are properly adhered to when capturing patient thermal images. If images are captured with a large number of thermal artefacts, the results are poor.

The study looked at various methods of segmentation, feature selection, and classification. There was a range of results, from no better than chance to excellent predictive ability. This showed the importance of removing as much unneeded tissue as possible while selecting the correct combination of features and classifiers. A satisfactory outcome of this study would be the implementation of this model to serve alongside another screening method deployed in an environment where medical resources are low, as this is the case best suited to the output created in this study. This is a step towards a larger acceptance of thermography because, in the past, its limitations caused it disfavour but these limitations have largely been mitigated through hardware improvements and standardisation. The final stage for acceptance would be a large double-blind long-term study evaluating its diagnostic ability. That relies on a very large amount of good data being captured in order to create and train machine learning algorithms capable of excellent generalisation. Hopefully, this model can be used, in part, to work towards that goal. The closing thought for this study are two simple questions: How many undetected breast cancer cases occur every day in areas that do not have the medical infrastructure needed? Could those lives be saved by providing an implementation of this model?

References

Acharya, U. R., Ng, E. Y. K., Tan, J.-H., & Sree, S. V. (2010). Thermography Based Breast Cancer Detection Using Texture Features and Support Vector Machine. *Journal of Medical Systems*, *36*(3), 1503–1510. doi: 10.1007/s10916-010-9611-z

Adobe. (2016). *Adobe PhoneGap - Build amazing mobile apps powered by open web tech.* Retrieved November 26, 2019, from https://phonegap.com/.

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., ... others (2013). Signatures of mutational processes in human cancer. *Nature*, *500*(7463), 415.

Ali, M. A. S., Sayed, G. I., Gaber, T., Hassanien, A. E., Snasel, V., & Silva, L. F. (2015). Detection of breast abnormalities of thermograms based on a new segmentation method. In *Proceedings of the federated conference on computer science and information systems* (pp. 255–261). doi: 10.15439/2015F318

Amalu, W. C. (2003). A Review of Breast Thermography. In *International academy of clinical thermology, ca, usa* (pp. 1–12).

American Cancer Society. (2018). *Breast Cancer Stages.* Retrieved April 23, 2019, from https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/stages-of-breast-cancer.html.

Ammer, K. (2003). Need for Standardisation of Measurements in Thermal Imaging. *Thermography and Lasers in Medicine*(May), 13–18.

Anbar, M., Brown, C., Milescu, L., Babalola, J., & Gentner, L. (2000). The potential of dynamic area telethermometry in assessing breast cancer. *IEEE Engineering in Medicine and Biology Magazine*, *19*(3), 58–62.

Araujo, A. D. S., Conci, A., Resmini, R., Montenegro, A., Araujo, C., & Lebon, F. (2017). Computer aided diagnosis for breast diseases based on infrared images. In *2017 ieee/acs 14th international conference on computer systems and applications (aiccsa)* (pp. 172–177). doi: 10.1109/AICCSA.2017.188

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. In *Statistics surveys* (Vol. 4, pp. 40–79). doi: 10.1214/09-SS054

Aubakir, B., Nurimbetov, B., Tursynbek, I., & Varol, H. A. (2016). Vital sign monitoring utilizing eulerian video magnification and thermography. In *2016 38th annual international conference of the ieee engineering in medicine and biology society (embc)* (pp. 3527–3530).

Baig, M. M., GholamHosseini, H., & Connolly, M. J. (2015). Mobile healthcare applications: system design review, critical issues and challenges. *Australasian physical & engineering*

sciences in medicine, *38*(1), 23–38.

Bellman, R. E. (1961). Dynamic programming treatment of the traveling salesman problem.

Betjeman, T. J., Soghoian, S. E., & Foran, M. P. (2013). mhealth in sub-saharan africa. *International journal of telemedicine and applications*, *2013*.

Betts, J. G., DeSaix, P., Johnson, E., Johnson, J. E., Korol, O., Kruse, D. H., ... Young, K. A. (2017). *Anatomy and physiology*. OpenStax College, Rice University.

Beyerer, J., León, F. P., & Frese, C. (2015). *Machine vision: Automated visual inspection: Theory, practice and applications*. doi: 10.1007/978-3-662-47794-6

Bhowmik, M. K., Gogoi, U. R., Das, K., Ghosh, A. K., Bhattacharjee, D., & Majumdar, G. (2016). Standardization of infrared breast thermogram acquisition protocols and abnormality analysis of breast thermograms. *Thermosense: Thermal Infrared Applications*, *9861*. doi: 10.1117/12.2223421

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford university press.

Bishop, C. M. (2007). *Pattern Recognition and Machine Learning*. Springer.

Blanpain, C. (2013). Tracing the cellular origin of cancer. *Nature Cell Biology*, *15*(2), 126–134. doi: 10.1038/ncb2657

Bloch, A. (2010). The potential of mobile technologies to positively impact access to essential medicines in low and medium income countries. *Washington, DC: mHealth Alliance*.

Boguski, R., Khan, T., Woelfel, S., D'Huyvetter, K., Armstrong, A. A., Armstrong, D. G., et al. (2019). Clinical utility of mobile phone-based thermography and low-cost infrared handheld thermometry in high-risk diabetic foot. *Indian Journal of Vascular and Endovascular Surgery*, *6*(1), 7.

Borchartt, T. B., Conci, A., Lima, R. C., Resmini, R., & Sanchez, A. (2013). Breast thermography from an image processing viewpoint: A survey. *Signal Processing*, *93*(10), 2785–2803. doi: 10.1016/j.sigpro.2012.08.012

Bradski, G., & Kaehler, A. (2008). *Learning opencv: Computer vision with the opencv library*. O'Reilly Media, Inc.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, *68*(6), 394–424. doi: 10.3322/caac.21492

Breastcancer.org. (2016). *Types of Breast Cancer*. Retrieved April 23, 2019, from https://www.breastcancer.org/symptoms/types.

Breastcancer.org. (2017). *What Does Prognosis Mean?* Retrieved April 23, 2019, from https://bit.ly/32bM4HA.

Breiman, L. (1994). Heuristics of instability in model selection. technique report. statistics department. *University of California at Berkeley*.

Breiman, L. (1996). Bagging Predictors. *Machine learning*, *24*(2), 123–140.

Breiman, L. (1999). Pasting small votes for classification in large databases and on-line. *Machine learning*, *36*(1-2), 85–103.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. doi: 10.1007/9781441993267_5

Burges, C. J. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, *2*(2), 121–167.

Burkov, A. (2019). *The hundred-page machine learning book*. Andriy Burkov.

Calderon-Contreras, J. D., Chacon-Murguia, M. I., Villalobos-Montiel, A. J., & Ortega-Maynez, L. (2015). A fuzzy computer aided diagnosis system using breast thermography. In *Proceedings - international symposium on biomedical imaging* (pp. 105–108). doi: 10.1109/ISBI.2015.7163827

Cancer Treatment Centers of America. (2019). *Breast cancer stages.* Retrieved April 23, 2019, from https://www.cancercenter.com/cancer-types/breast-cancer/stages.

Carmeliet, P., & Jain, R. K. (2000). Angiogenesis in cancer and other diseases. *nature*, *407*(6801), 249–257. doi: 10.3949/ccjm.54.1.63-a

cc Globaltech. (2018). *Thermal Imager.* Retrieved October 16, 2018, from https://cc-globaltech.com/product-category/thermal-imager/.

Chojnowski, M. (2017). Infrared thermal imaging in connective tissue diseases. *Reumatologia*, *55*(1), 38–43. doi: 10.5114/reum.2017.66686

Ciaramitaro, B. (2012). *Mobile technology consumption: opportunities and challenges*. Information Science Reference.

Clausi, D. A. (2002). An analysis of co-occurrence texture statistics as a function of grey level quantization. *Canadian Journal of Remote Sensing*, *28*(1), 45–62. doi: 10.5589/m02-004

Clifford, G. D., & Clifton, D. (2012). Wireless technology in disease management and medicine. *Annual review of medicine*, *63*, 479–492.

Collins, K., Winslow, M., Reed, M., Walters, S., Robinson, T., Madan, J., ... Wyld, L. (2010). The views of older women towards mammographic screening: a qualitative and quantitative study. *British journal of cancer*, *102*(10), 1461.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297. doi: 10.1109/64.163674

Cover, M., & Hart, E. (1967). Nearest Neighbor Pattern Classification. *IEEE Transactions*

on Information Theory, *13*(1), 21–27.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

Cruz-Ramírez, N., Mezura-Montes, E., Ameca-Alducin, M. Y., Martín-Del-Campo-Mena, E., Acosta-Mesa, H. G., Pérez-Castro, N., ... Barrientos-Martínez, R. E. (2013). Evaluation of the Diagnostic Power of Thermography in Breast Cancer Using Bayesian Network Classifiers. In *Computational and mathematical methods in medicine* (Vol. 2013, pp. 1–10). doi: 10.1155/2013/264246

Danzl, D. F., & Pozos, R. S. (1994). Accidental hypothermia. *New England Journal of Medicine*, *331*(26), 1756–1760.

Deborah, K., Tanya, L., & Dugald, S. (2009). Comparative review of thermographyas a breast screening technique. *Integrative Cancer Therapies*, *8*(1), 9–16.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Etehadtavakol, M., Ng, E. Y., Chandran, V., & Rabbani, H. (2013). Separable and non-separable discrete wavelet transform based texture features and image classification of breast thermograms. *Infrared Physics and Technology*, *61*, 274–286. doi: 10.1016/j.infrared.2013.08.009

Everitt, B. S., & Skrondal, A. (2010). *The cambridge dictionary of statistics*. New York University.

Fang, M., Yue, G., & Yu, Q. (2009). The study on an application of Otsu method in Canny operator. *Proceedings of the 2009 International Symposium on Information Processing (ISIP'09)*, *2*(4), 109–112.

Faust, O., Rajendra Acharya, U., Ng, E. Y., Hong, T. J., & Yu, W. (2014). Application of infrared thermography in computer aided diagnosis. *Infrared Physics and Technology*, *66*, 160–175. doi: 10.1016/j.infrared.2014.06.001

Fay, M. P., & Proschan, M. A. (2010). Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Survey*, *4*(1), 1–37. doi: 10.1214/09-SS051.Wilcoxon-Mann-Whitney

Feig, S. A., Shaber, G. S., Schwartz, G. F., Patchefsky, A., Libshitz, H. I., Edeiken, J., & Wallace, J. D. (1977). Thermography, mammography, and clinical examination in breast cancer screening: review of 16,000 studies. *Radiology*, *122*(1), 123–127.

Fling, B. (2009). *Mobile design and development: Practical concepts and techniques for creating mobile sites and web apps*. " O'Reilly Media, Inc.".

FLIR. (2018). *Mobile Accessories*. Retrieved October 16, 2018, from https://www.flir.eu/browse/professional-tools/mobile-accessories/.

Francis, S. V., & Sasikala, M. (2013). Automatic detection of abnormal breast thermograms using asymmetry analysis of texture features. *Journal of Medical Engineering and Technology*, *37*(1), 17–21. doi: 10.3109/03091902.2012.728674

Francis, S. V., Sasikala, M., Bhavani Bharathi, G., & Jaipurkar, S. D. (2014). Breast cancer detection in rotational thermography images using texture features. In *Infrared physics and technology* (Vol. 67, pp. 490–496). Elsevier B.V. doi: 10.1016/j.infrared.2014.08.019

Francis, S. V., Sasikala, M., & Saranya, S. (2014). Detection of breast abnormality from thermograms using curvelet transform based feature extraction. *Journal of Medical Systems*, *38*(4). doi: 10.1007/s10916-014-0023-3

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *International conference on machine learning* (pp. 148–156).

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian Network Classifiers. *Machine learning*, *29*(2-3), 131–163. doi: 10.1002/9780470400531.eorms0099

Gaber, T., Ismail, G., Anter, A., Soliman, M., Ali, M., Semary, N., ... Snasel, V. (2015). Thermogram breast cancer prediction approach based on Neutrosophic sets and fuzzy c-means algorithm. In *Engineering in medicine and biology society (embc)* (Vol. 37, pp. 4254–4257). doi: 10.1109/EMBC.2015.7319334

Gade, R., & Moeslund, T. B. (2014). Thermal cameras and applications: A survey. *Machine Vision and Applications*, *25*(1), 245–262. doi: 10.1007/s00138-013-0570-5

Gautherie, M. (1983). Thermobiological assessment of benign and malignant breast diseases. *American Journal of Obstetrics and Gynecology*, *147*(8), 861–869. doi: 10.1016/0002-9378(83)90236-3

Gautherie, M. (1985). New protocol for the evaluation of breast thermograms. *Thermological Methods*, 227–235.

Gautherie, M., & Gros, C. M. (1980). Breast thermography and cancer risk prediction. *Cancer*, *45*(1), 51–56. doi: 10.1002/cncr.2820450110

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American statistical Association*, *70*(350), 320–328.

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow* (2nd ed.; N. Tache, Ed.). O'Reilly.

Gershon-Cohen, J., & Habennan, J. D. (1968). Thermography of smoking. *Archives of Environmental Health: An International Journal*, *16*(5), 637–641.

Ghafarpour, A., Zare, I., Zadeh, H. G., Haddadnia, J., Zadeh, F. J. S., Zadeh, Z. E., ... Nour, S. (2016). A review of the dedicated studies to breast cancer diagnosis by

thermal imaging in the fields of medical and artificial intelligence sciences: Review article. *Biomedical Research*, *27*(2), 377–385.

Giardina, C. R., & Dougherty, E. R. (1988). Morphological methods in image and signal processing. *Engelwood Cliffs: Prentice Hall, 1988*.

Gillies, R. J., Kinahan, P. E., & Hricak, H. (2015). Radiomics: images are more than pictures, they are data. *Radiology*, *278*(2), 563–577. doi: 10.1148/radiol.2015151169

GlobalStats. (2019). *Mobile Operating System Market Share Worldwide.* Retrieved November 20, 2019, from https://gs.statcounter.com/os-market-share/mobile/worldwide.

Gogoi, U. R., Bhowmik, M. K., Ghosh, A. K., Bhattacharjee, D., & Majumdar, G. (2017). Discriminative feature selection for breast abnormality detection and accurate classification of thermograms. In *Innovations in electronics, signal processing and communication (iesc)* (pp. 39–44).

Grus, J. (2019). *Data Science from Scratch.* O'Reilly Media. doi: 10.1017/CBO9781107415324.004

Guidi, A. J., & Schnitt, S. J. (1996). Angiogenesis in preinvasive lesions of the breast. *Breast Journal*, *2*(6), 364–369. doi: 10.1111/j.1524-4741.1996.tb00123.x

Hall, P., Park, B. U., & Samworth, R. J. (2008). Choice of neighbor order in nearest-neighbor classification. *Annals of Statistics*, *36*(5), 2135–2152. doi: 10.1214/07-AOS537

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *cell*, *144*(5), 646–674.

Hansen, L. K., & Salamon, P. (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis & Machine Intelligence*(10), 993–1001.

Haralick, R. M. (1979). Statistical and structural approaches to texture. *Proceedings of the IEEE*, *67*(5), 786–804.

Haralick, R. M., Shanmugam, K., & Dinstein, I. (1973). Textural features for image classification. *Transactions on systems, man, and cybernetics*, *6*(1), 610–621. doi: 10.1109/TSMC.1973.4309314

Hardy, J. (1934). The radiation of heat from the human body. *J. Clin. Invest*, *13*(8), 615–620.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning.* Springer.

Head, J. F., & Elliott, R. L. (2002). Infrared imaging: making progress in fulfilling its medical promise. *IEEE Engineering in Medicine and biology Magazine*, *21*(6), 80–85.

Head, J. F., & Wan, Fen, Charles A. Lipari, R. . E. (2000). The Important Role of Infrared Imaging in Breast Cancer. *IEEE Engineering in Medicine and Biology*, *19*(June),

52–57.

Head, J. F., Wang, F. E. N., & Elliott, R. L. (1993). Breast thermography is a noninvasive prognostic procedure that predicts tumor growth rate in breast cancer patients. *Annals of the New York Academy of Sciences*, *698*(1), 153–158.

Herranz, M., & Ruibal, A. (2012). Optical Imaging in Breast Cancer Diagnosis: The Next Evolution. *Journal of Oncology*, *2012*, 1–10. doi: 10.1155/2012/863747

Hobbins, W. B. (1983). Thermography of the Breast - A Skin Organ. In *Thermal assessment of breast health* (pp. 40–48).

Hosny, A., van Griethuysen, J. J., Parmar, C., Aerts, H. J., Fedorov, A., Beets-Tan, R. G., ... Pieper, S. (2017). Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Research*, *77*(21), e104–e107. doi: 10.1158/0008-5472.can-17-0339

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, *24*(6), 417.

IACT. (2002). *Standards and Protocols in Clinical Thermographic Imaging.* Retrieved 2019-07-29, from http://www.iact-org.org/professionals/thermog-guidelines.html

Incropera, F. P., Lavine, A. S., Bergman, T. L., & DeWitt, D. P. (2011). *Fundamentals of heat and mass transfer* (7th ed.). Wiley.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.

Jaspers, M. E., Carrière, M., Meij-de Vries, A., Klaessens, J., & van Zuijlen, P. (2017). The flir one thermal imager for the assessment of burn wounds: Reliability and validity study. *Burns*, *43*(7), 1516–1523.

Jones, B. F. (1998). A reappraisal of the use of infrared thermal image analysis in medicine. *IEEE transactions on medical imaging*, *17*(6), 1019–1027. doi: 10.1109/42.746635

Julesz, B. (1975). Experiments in the visual perception of texture. *Scientific American*, *232*(4), 34–43.

Kahn, J. G., Yang, J. S., & Kahn, J. S. (2010). Mobile health needs and opportunities in developing countries. *Health Affairs*, *29*(2), 252–258.

Kakileti, S. T., Manjunath, G., Madhu, H., & Ramprakash, H. V. (2017). Advances in Breast Thermography. In *New perspectives in breast imaging* (pp. 92–106). doi: 10.5772/intechopen.69198

Kanazawa, T., Nakagami, G., Goto, T., Noguchi, H., Oe, M., Miyagaki, T., ... Sanada, H. (2016). Use of smartphone attached mobile thermography assessing subclinical inflammation: a pilot study. *Journal of wound care*, *25*(4), 177–182.

Kandlikar, S. G., Perez-Raya, I., Raghupathi, P. A., Gonzalez-Hernandez, J. L., Dabydeen,

D., Medeiros, L., & Phatak, P. (2017). Infrared imaging technology for breast cancer detection Current status, protocols and new directions. *International Journal of Heat and Mass Transfer*, *108*, 2303–2320. doi: 10.1016/j.ijheatmasstransfer.2017.01.086

kanti Bhowmik, M., Gogoi, U. R., Majumdar, G., Bhattacharjee, D., Datta, D., & Ghosh, A. K. (2017). Designing of ground truth annotated DBT-TU-JU breast thermogram database towards early abnormality prediction. *IEEE Journal of Biomedical and Health Informatics*, *22*(4), 1238–1249. doi: 10.1109/JBHI.2017.2740500

Kapoor, P., Prasad, S. V. A. V., & Patni, S. (2012). Automatic Analysis of Breast Thermograms for tumor detection based on Biostatistical feature extraction and ANN. *International journal of emerging trends in engineering and development*, *7*(2), 245–255.

Keyserlingk, J., Ahlgren, P., Yu, E., & Belliveau, N. (1998). Infrared imaging of the breast: Initial reappraisal using high- resolution digital technology in 100 successive cases of stage I and II breast cancer. *Breast Journal*, *4*(4), 245–251.

Knowles, J. R. (1980). Enzyme-catalyzed phosphoryl transfer reactions. *Annual review of biochemistry*, *49*(1), 877–919.

Kockara, S., Halic, T., Hudson, C., Loney, A., & Crawford, A. (2014). Portable malignant lesion detection with low cost mobile infrared thermography. In *2014 ieee innovations in technology conference* (pp. 1–5).

Krawczyk, B., & Schaefer, G. (2013). A pruned ensemble classifier for effective breast thermogram analysis. In *Proceedings of the annual international conference of the ieee engineering in medicine and biology society, embs* (pp. 7120–7123). doi: 10.1109/EMBC.2013.6611199

Krawczyk, B., & Schaefer, G. (2014). Breast thermogram analysis using classifier ensembles and image symmetry features. *IEEE Systems Journal*, *8*(3), 921–928. doi: 10.1109/JSYST.2013.2283135

Kruse, P. W. (2001). *Uncooled Infrared Imaging Arrays, Systems and Applications.* doi: 10.1007/s13398-014-0173-7.2

Kumar, V., Gu, Y., Basu, S., Berglund, A., Eschrich, S. A., Schabath, M. B., ... Gillies, R. J. (2012). Radiomics: The process and the challenges. *Magnetic Resonance Imaging*, *30*(9), 1234–1248. Retrieved from http://dx.doi.org/10.1016/j.mri.2012.06.010 doi: 10.1016/j.mri.2012.06.010

Lambin, P., Leijenaar, R. T., Deist, T. M., Peerlings, J., De Jong, E. E., Van Timmeren, J., ... Walsh, S. (2017). Radiomics: The bridge between medical imaging and personalized medicine. *Nature Reviews Clinical Oncology*, *14*(12), 749–762. Retrieved from http://dx.doi.org/10.1038/nrclinonc.2017.141 doi: 10.1038/nrclinonc.2017.141

Lashkari, A., Pak, F., & Firouzmand, M. (2016). Full intelligent cancer classification of thermal breast images to assist physician in clinical diagnostic applications. *Journal Of Medical Signals And Sensors*, *6*(1), 12–24.

Lau, J. H. (1996). *Flip chip technologies (Vol. 1)* (1st ed.). New York: McGraw-Hill.

Laupland, K. B. (2009). Fever in the critically ill medical patient. *Critical care medicine*, *37*(7), S273–S278.

Lawson, R. (1956). Implications of surface temperatures in the diagnosis of breast cancer. *Cancer Medical Association Journal*, *75*(4), 309–310.

Leon, N., Schneider, H., & Daviaud, E. (2012). Applying a framework for assessing the health system challenges to scaling up mhealth in south africa. *BMC medical informatics and decision making*, *12*(1), 123.

Lin, P. H., & Saines, M. (2017). Assessment of lower extremity ischemia using smartphone thermographic imaging. *Journal of vascular surgery cases and innovative techniques*, *3*(4), 205–208.

Mannara, G., Salvatori, G. C., & Pizzuti, G. P. (1993). Ethyl alcohol induced skin temperature changes evaluated by thermography. Preliminary results. *Bollettino della Societa italiana di biologia sperimentale*, *69*(10), 587–594.

Marks, J. G., & Miller, J. J. (2017). *Lookingbill and marks' principles of dermatology e-book*. Elsevier Health Sciences.

Maron, M. E. (1961). Automatic Indexing: An Experimental Inquiry. *Journal of the ACM*, *8*(3), 404–417. doi: 10.1145/321075.321084

Materka, A., & Strzelecki, M. (1998). Texture Analysis Methods A Review. In *Cost b11 report* (pp. 1–33).

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*(4), 115–133.

McNab, B. K. (1997). On the utility of uniformity in the definition of basal rate of metabolism. *Physiological Zoology*, *70*(6), 718–720.

Mehdy, M. M., Ng, P. Y., Shair, E. F., Saleh, N. I. M., & Gomes, C. (2017). Artificial Neural Networks in Image Processing for Early Detection of Breast Cancer. *Computational and Mathematical Methods in Medicine*, *2017*, 1–15. doi: 10.1155/2017/2610628

Meredith, P., & Riesz, J. (2004). Radiative relaxation quantum yields for synthetic eumelanin. *Photochemistry and photobiology*, *79*(2), 211–216.

Merler, S., Caprile, B., & Furlanello, C. (2007). Parallelizing adaboost by weights dynamics. *Computational statistics & data analysis*, *51*(5), 2487–2498.

Milosevic, M., Jankovic, D., & Peulic, A. (2014). Thermography based breast cancer detec-

tion using texture features and minimum variance quantization. *EXCLI Journal*, *13*, 1204–1215.

Min, S., Heo, J., Kong, Y., Nam, Y., Ley, P., Jung, B.-K., ... Shin, W. (2017). Thermal infrared image analysis for breast cancer detection. *KSII Transactions on Internet & Information Systems*, *11*(2).

Mitchell, T. (1997). *Machine Learning*. McGraw Hill.

Miyai, K. (2014). Adenoid cystic carcinoma of breast: Recent advances. *World Journal of Clinical Cases*, *2*(12), 732. doi: 10.12998/wjcc.v2.i12.732

Moghbel, M., & Mashohor, S. (2013). A review of computer assisted detection/diagnosis (cad) in breast thermography for breast cancer detection. *Artificial Intelligence Review*, *39*(4), 305–313.

Mohamed, N. A. E.-r. (2015). Breast Cancer Risk Detection Using Digital Infrared Thermal Images. *International Journal of Bioinformatics and Biomedical Engineering*, *1*(2), 185–194.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of machine learning* (2nd ed.). MIT Press.

Mookiah, M. R. K., Acharya, U. R., & Ng, E. (2012). Data mining technique for breast cancer detection in thermograms using hybrid feature extraction strategy. *Quantitative InfraRed Thermography Journal*, *9*(2), 151–165. doi: 10.1080/17686733.2012.738788

Morais, K. C., Vargas, J. V., Reisemberger, G. G., Freitas, F. N., Oliari, S. H., Brioschi, M. L., & Neto, C. D. (2016). An infrared image based methodology for breast lesions screening. *Infrared Physics and Technology*, *76*, 710–721. doi: 10.1016/j.infrared.2016.04.036

Motta, L. S., Conci, A., Lima, R. C. F., & Diniz, E. M. (2010). Automatic segmentation on thermograms in order to aid diagnosis and 2D modeling. *Proceedings of 10th Workshop em Informatica Medica*(January), 1610–1619.

Nailon, William Henry. (2010). Texture Analysis Methods for Medical Image Characterisation. *Biomedical imaging*, *84*(3), 32–35. doi: 10.5772/32009

Negin, M., Ziskin, M. C., Piner, C., & Lapayowker, M. S. (1977). A computerized breast thermographic interpreter. *IEEE Transactions On Biomedical Engineering*(4), 347–352.

Newton, I. (1730). *Opticks, or, a treatise of the reflections, refractions, inflections & colours of light* (Vol. 4). doi: 10.1016/0364-9229(81)90009-9

Ng, A. (2010). *Machine Learning Exercise 8: Non-linear SVM classification with kernels.* Retrieved September 18, 2019, from http://openclassroom.stanford.edu/MainFolder/DocumentPage.php?course=MachineLearning&doc=exercises/ex8/ex8.html.

Ng, E. (2009). A review of thermography as promising non-invasive detection modality for breast tumor. *International Journal of Thermal Sciences*, *48*(5), 849–859.

Ng, E., & Etehadtavakol, M. (2017). *Application of Infrared to Biomedical Sciences*. Springer. doi: 10.1007/978-981-10-3147-2

Ng, Y., Ung, L., Ng, F., & Sim, L. S. J. (2001). Statistical analysis of healthy and malignant breast thermography. *Journal of Medical Engineering & Technology*, *25*(6), 253–263. doi: 10.1080/03091900110086642

Nowakowski, A. (2006). *Biomedical engineering handbook (Vol. 2)* (3rd ed.; J. D. Bronzino, Ed.). CRC Press. doi: doi:10.1201/9781420008340.ch7

Ogawa, H., Nishio, A., Satake, H., Naganawa, S., Imai, T., Sawaki, M., ... Miyata, T. (2008). Neuroendocrine tumor in the breast. *Radiation Medicine - Medical Imaging and Radiation Oncology*, *26*(1), 28–32. doi: 10.1007/s11604-007-0182-y

Olivier, M. S. (2009). *Information technology research: A practical guide for computer science and informatics*. Van Schaik.

Pak, F., Lashkari, A., & Firouzmand, M. (2009). Breast thermal images classification using optimal feature selectors and classifiers. *The Journal of Engineering*, *2016*(7), 237–248. doi: 10.1049/joe.2016.0060

Pal, S., Lau, S., Kruper, L., Nwoye, U., Garberoglio, C., Gupta, R., & Somlo, G. (2010). Papillary carcinoma of the breast: An overview. *Breast Cancer Research and Treatment*, *122*(3), 637–645. doi: 10.1007/s10549-010-0961-5.Papillary

Pavithra, P., Ravichandran, K., Sekar, K., & Manikandan, R. (2018). The effect of thermography on breast cancer detection. *Systematic Reviews in Pharmacy*, *9*(1), 10–16.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pennes, H. H. (1948). Analysis of tissue and arterial blood temperatures in the resting human forearm. *Journal of applied physiology*, *1*(2), 93–122. doi: 10.4324/9780203302293_applied_physiology

Picard, R. R., & Cook, R. D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association*, *79*(387), 575–583.

Pillay, A. L. (2002). Rural and urban south african women's awareness of cancers of the breast and cervix. *Ethnicity and Health*, *7*(2), 103–114.

Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63.

Pramanik, S., Bhattacharjee, D., & Nasipuri, M. (2015). Wavelet based thermogram analysis for breast cancer detection. In *2015 international symposium on advanced computing and communication (isacc)* (pp. 205–212). doi: 10.1109/ISACC.2015.7377343

Pramanik, S., Bhattacharjee, D., & Nasipuri, M. (2016). Texture analysis of breast thermogram for differentiation of malignant and benign breast. In *2016 international conference on advances in computing, communications and informatics, icacci 2016* (pp. 8–14). doi: 10.1109/ICACCI.2016.7732018

Provost, F. J., & Kolluri, V. (1997). A survey of methods for scaling up inductive learning algorithms. In *Proc. 3rd international conference on knowledge discovery and data mining.*

Qi, H., Teja Kuruganti, P., & Snyder, W. E. (2008). *Detecting Breast Cancer from Thermal Infrared Images by Assymmetry Analysis.*

Ranck, J. (2011). Health information and health care: The role of technology in unlocking data and wellness–a discussion paper. *Washington, DC: United Nations Foundation & Vodafone Foundation Technology Partnership.*

Ring, E. F. (1990). Quantitative thermal imaging. *Clinical Physics and Physiological Measurement*, *11*(4A), 87–95. doi: 10.1088/0143-0815/11/4A/310

Ring, E. F. J. (2004). The historical development of thermal imaging in medicine. *Rheumatology*, *43*(6), 800–802.

Ring, E. F. J. (2010). Thermal imaging today and its relevance to diabetes. *Journal of Diabetes Science and Technology*, *4*(4), 857–862. doi: 10.1177/193229681000400414

Ring, E. F. J., & Ammer, K. (2000). The technique of infrared imaging in medicine. *Thermology international*, *10*(1). doi: 10.1088/978-0-7503-1143-4ch1

Rish, I. (2001). An empirical study of the naive Bayes classifier. *IJCAI 2001 workshop on empirical methods in artificial intelligence*, *3*(22), 41–46. doi: 10.1039/b104835j

Rodrigues, E., & Conci, A. (2014). Comparing results of thermographic images based diagnosis for breast diseases. In *International conference on systems, signals and image processing (iwssip)* (pp. 39–42). IEEE.

Rogalski, A. (2002). Infrared detectors : an overview. *Infrared Physics & Technology*, *43*(3-5), 187–210.

Rogalski, A. (2012). History of infrared detectors. *Opto-Electronics Review*, *20*(3), 279–308. doi: 10.2478/s11772

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, *33*(1-2), 1–39. doi: 10.1007/s10462-009-9124-7

Rokach, L., & Maimon, O. Z. (2015). *Data mining with decision trees: theory and applications*

(2nd ed.). World Scientific Publishing Co.

Rosenblatt, F. (1957). *The perceptron– a perceiving and recognizing automaton. Tech. Rep. 85-460-1* (Tech. Rep.). Cornell Aeronautical Laboratory.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). *Learning internal representations by error propagation* (Tech. Rep.). California Univ San Diego La Jolla Inst for Cognitive Science.

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, *3*(3), 210–229.

Sardanelli, F., Aase, H. S., Álvarez, M., Azavedo, E., Baarslag, H. J., Balleyguier, C., & Forrai, G. (2017). Position paper on screening for breast cancer by the European Society of Breast Imaging (EUSOBI) and 30 national breast radiology bodies from Austria, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Czech Republic, Denmark, Estonia, Finland, France, G. *European Radiology*, *27*(7), 2737–2743. doi: 10.1007/s00330-016-4612-z

Saslow, D., Boetes, C., Burke, W., Harms, S., Leach, M. O., Lehman, C. D., & Russell, C. A. (2007). American cancer society guidelines for breast screening with MRI as an adjunct to mammography. *Obstetrical and Gynecological Survey*, *62*(7), 458–460. doi: 10.1097/01.ogx.0000269073.50925.38

Sathish, D., Kamath, S., Prasad, K., Kadavigere, R., & Martis, R. J. (2017). Asymmetry analysis of breast thermograms using automated segmentation and texture features. *Signal, Image and Video Processing*, *11*(4), 745–752. doi: 10.1007/s11760-016-1018-y

Schaefer, G. (2014). ACO classification of thermogram symmetry features for breast cancer diagnosis. *Memetic Computing*, *6*(3), 207–212. doi: 10.1007/s12293-014-0135-9

Schaefer, G., & Nakashima, T. (2015). Strategies for addressing class imbalance in ensemble classification of thermography breast cancer features. *2015 IEEE Congress on Evolutionary Computation, CEC 2015 - Proceedings*, 2362–2367. doi: 10.1109/CEC.2015.7257177

Schaefer, G., Závišek, M., & Nakashima, T. (2009). Thermography based breast cancer analysis using statistical features and fuzzy classification. *Pattern Recognition*, *42*(6), 1133–1137. doi: 10.1016/j.patcog.2008.08.007

Seek. (2018). *Powerful thermal imaging cameras designed for your smartphone.* Retrieved October 16, 2018, from https://www.thermal.com/compact-series.html.

SensorTower. (2019). *Worldwide Mobile App Revenue and Downloads.* Retrieved November 23, 2019, from https://sensortower.com/blog/app-revenue-and-downloads-1h-2019.

Shah, D. R., Tseng, W. H., & Martinez, S. R. (2012). Treatment Options for Metaplastic

Breast Cancer. In *Isrn oncology* (Vol. 2012, pp. 1–4). doi: 10.5402/2012/706162

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. doi: 10.1017/CBO9781107298019

Silva, L. F., Saade, D. C. M., Sequeiros, G. O., Silva, A. C., Paiva, A. C., Bravo, R. S., & Conci, A. (2014). A new database for breast research with infrared image. *Journal of Medical Imaging and Health Informatics*, *4*(1), 92–100. doi: 10.1166/jmihi.2014.1226

Simard, P. Y., Steinkraus, D., & Platt, J. C. (2003). Best practices for convolutional neural networks applied to visual document analysis. *Icdar*, *3*(2003), 1–9. doi: 10.1017/CBO9781107415324.004

Singh, S., Srivastava, A., Mi, L., Chen, K., Wang, Y., Caselli, R., ... Reiman, E. (2017, 11). Deep learning based classification of fdg-pet data for alzheimers disease categories. *Proceedings of SPIE–the International Society for Optical Engineering*, *10572*, 84. doi: 10.1117/12.2294537

Skurichina, M., & Duin, R. P. (2002). Bagging, boosting and the random subspace method for linear classifiers. *Pattern Analysis & Applications*, *5*(2), 121–135.

Soh, L.-k., Tsatsoulis, C., & Member, S. (1999). Texture Analysis of SAR Sea Ice Imagery Using Gray Level Co-Occurrence Matrices. *IEEE Transactions on Geoscience and Remote Sensing*, *37*(2), 780–795. doi: 10.1109/36.752194

Sokolowski, J. A., & Banks, C. M. (2010). *Modeling and Simulation Fundamentals: Theoretical Underpinnings and Practical Domains*. John Wiley & Sons. doi: 10.1002/9780470590621

Somdyala, N. I., Bradshaw, D., Gelderblom, W. C., & Parkin, D. M. (2010). Cancer incidence in a rural population of south africa, 1998–2002. *International Journal of Cancer*, *127*(10), 2420–2429.

Sørensen, C., De Reuver, M., & Basole, R. C. (2015). *Mobile platforms and ecosystems*. Springer.

Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote sensing of Environment*, *62*(1), 77–89.

Steketee, J. (1973). Spectral emissivity of skin and pericardium. *Physics in Medicine & Biology*, *18*(5), 686–694.

Stevenson, A. (2010). *Oxford Dictionary of English* (3rd ed.). Oxford University Press. doi: 10.1093/acref/9780199571123.001.0001

Stewart, J. (2015). *Single variable calculus: Early transcendentals*. Cengage Learning.

Stratton, M. R. (2011). Exploring the genomes of cancer cells: Progress and promise. *Science*, *331*(6024), 1553–1558. doi: 10.1126/science.1204040

Stratton, M. R., Campbell, P. J., & Futreal, P. A. (2009). The cancer genome. *Nature*, *458*(7239), 719. doi: 10.1038/nature07943.The

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning*. MIT press.

Tabar, L., Gad, A., Holmberg, L. H., Ljungquist, U., Group, K. C. P., Fagerberg, C. J. G., & Croup, Ö. C. P. (1985). Reduction in mortality from breast cancer after mass screening with mammography: randomised trial from the Breast Cancer Screening Working Group of the Swedish National Board of Health and Welfare. *The Lancet*, *325*(8433), 829–832.

ThermApp. (2018). *Powerful thermal imaging cameras designed for your smartphone.* Retrieved October 16, 2018, from https://therm-app.com/shop/.

Thomsen, L., Miles, D., Happerfield, L., Bobrow, L., Knowles, R., & Moncada, S. (1995). Nitric oxide synthase activity in human breast cancer. *British Journal of Cancer*, *72*(1), 41–44. doi: 10.1038/bjc.1995.274

Tortora, G. J., & Derrickson, B. (2017). *Principles of anatomy and physiology* (15th ed.). John Wiley & Sons.

Tukey, J. (1977). Exploratory data analysis. *Reading: Addison-Wesley*.

Uematsu, S. (1985). Symmetry of skin temperature comparing one side of the body to the other. In *Thermology* (Vol. 1, pp. 4–7).

Vapnik, V. N. (1992). *The Nature of Statical Learning Theory*. doi: 10.1126/science.1246848

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, *10*(5), 988–999. doi: 10.1109/72.788640

Vardasca, R., Vaz, L., & Mendes, J. (2018). Classification and decision making of medical infrared thermal images. *Classification in BioApps. Lecture Notes in Computational Vision and Biomechanics*, *26*, 79–104. doi: 10.1007/978-3-319-65981-7_4

Vasudev, P., & Onuma, K. (2011). Secretory breast carcinoma: Unique, triple-negative carcinoma with a favorable prognosis and characteristic molecular expression. *Archives of Pathology and Laboratory Medicine*, *135*(12), 1606–1610. doi: 10.5858/arpa.2010-0351-RS

Vollmer, M., & Möllmann, K. P. (2017). *Infrared thermal imaging: fundamentals, research and applications*. John Wiley Sons.

Vorobiof, D. A., Sitas, F., & Vorobiof, G. (2001). Breast cancer incidence in south africa. *Journal of clinical oncology*, *19*(18; SUPP), 125s–125s.

Vranic, S., Feldman, R., & Gatalica, Z. (2017). Apocrine carcinoma of the breast: A brief update on the molecular features and targetable biomarkers. *Bosnian Journal of Basic Medical Sciences*, *17*(1), 9–11. doi: 10.17305/bjbms.2016.1811

Watmough, D. J., Fowler, P. W., & Oliver, R. (1970). The thermal scanning of a curved isothermal surface: implications for clinical thermography. *Physics in Medicine & Biology*, *15*(1), 1.

Watmough, D. J., & Oliver, R. (1969). Wavelength dependence of skin emissivity. *Physics in Medicine and Biology*, *14*(2), 201–204. doi: 10.1088/0031-9155/14/2/302

Weigelt, B., Geyer, F. C., & Reis-Filho, J. S. (2010). Histological types of breast cancer: How special are they? *Molecular Oncology*, *4*(3), 192–208. doi: 10.1016/j.molonc.2010.04.004

Weigelt, B., Peterse, J. L., & Van't Veer, L. J. (2005). Breast cancer metastasis: Markers and models. *Nature Reviews Cancer*, *5*(8), 591–602. doi: 10.1038/nrc1670

Whitney, A. W. (1971). A direct method of nonparametric measurement selection. *IEEE Transactions on Computers*, *100*(9), 1100–1103. doi: 10.1109/T-C.1971.223410

Williams, K. L. (1964). Infrared thermogrpahy as a tool in medical research. *Annals of the New York Academy of Sciences*, *121*(1), 99–112.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Elsevier Science. doi: 10.1145/2020976.2021004.

Wunderlich, C. (1871). *On the temperature in diseases: a manual of medical thermometry* (Vol. 49). doi: 10.1016/0003-6870(73)90259-7

Yaniv, Z., Lowekamp, B. C., Johnson, H. J., & Beare, R. (2018). Simpleitk image-analysis notebooks: a collaborative environment for education and reproducible research. *Journal of digital imaging*, *31*(3), 290–303.

Yassin, N. I., Omran, S., El Houby, E. M., & Allam, H. (2018). Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. *Computer Methods and Programs in Biomedicine*, *156*, 25–45. doi: 10.1016/j.cmpb.2017.12.012

Zadeh, H. G., Haddadnia, J., Hashemian, M., & Hassanpour, K. (2012). Diagnosis of breast cancer using a combination of genetic algorithm and artificial neural network in medical infrared thermal imaging. *Iranian Journal of Medical Physics*, *9*(4), 265–274. doi: 10.22038/ijmp.2013.470

Zhang, C.-X., & Zhang, J.-S. (2008). A local boosting algorithm for solving classification problems. *Computational Statistics & Data Analysis*, *52*(4), 1928–1941.

Zwanenburg, A., Leger, S., Vallieres, M., & Lock, S. (2016). Image biomarker standardisation initiative. *arXiv preprint arXiv:1612.07003*. doi: 10.17195/candat.2016.08.1

Appendix A

Significant features summary tables

Table A.1: Significant features for the various segmentation methods obtained using the Student's t-test

	Feature	**SM1**	**SM2**	**SM3**	**SM4**	**SM5**
	Mean	✓		✓	✓	✓
	Variance	✓		✓		✓
	Standard Deviation	✓		✓	✓	✓
	Skewness		✓		✓	
	Kurtosis		✓			
	Fo Entropy		✓	✓	✓	✓
	ASM (Energy)	✓	✓	✓	✓	✓
	Contrast			✓	✓	✓
	Correlation			✓	✓	✓
	Dissimilarity			✓	✓	
	Homogeneity			✓	✓	
	So Entropy		✓	✓	✓	✓
	Sum Entropy		✓	✓	✓	✓
	Difference Entropy			✓	✓	
	Sum Variance			✓		✓
	Difference Variance			✓	✓	✓
	Sum Average					
	Autocorrelation			✓		

Continued on next page

Table A.1 – continued from previous page

Feature	SM1	SM2	SM3	SM4	SM5
IMC1		✓	✓	✓	✓
IMC2		✓	✓	✓	✓
IDN			✓	✓	
IDMN				✓	
SoS: Variance			✓		✓

Table A.2: Significant features for the various segmentation methods obtained using the Mann-Whitney-Wilcoxon test

Feature	SM1	SM2	SM3	SM4	SM5
Mean		✓	✓	✓	
Variance	✓		✓		✓
Standard Deviation	✓		✓	✓	✓
Skewness	✓	✓		✓	
Kurtosis		✓			✓
Fo Entropy	✓	✓	✓	✓	✓
ASM (Energy)	✓	✓	✓	✓	✓
Contrast			✓	✓	✓
Correlation	✓		✓	✓	✓
Dissimilarity			✓	✓	
Homogeneity			✓	✓	
So Entropy	✓	✓	✓	✓	✓
Sum Entropy	✓	✓	✓	✓	✓
Difference Entropy			✓	✓	
Sum Variance			✓	✓	✓
Difference Variance			✓	✓	
Sum Average			✓		
Autocorrelation			✓		
IMC1	✓	✓	✓	✓	✓
IMC2	✓	✓		✓	✓
IDN			✓		

Continued on next page

Table A.2 – continued from previous page

	Feature	SM1	SM2	SM3	SM4	SM5
	IDMN					
	SoS: Variance	✓		✓	✓	✓

Appendix B

Academic publication resulting from this study

This study has resulted in the following conference publication:

Lennox, N., Haskins, B. (2019). Comparison of segmentation methods for the detection of breast cancer using thermal images. In *2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*(p. 1-8).

Appendix C

Python code example

Appendix C contains a python code snippet of training the various classifiers on the features obtained from semi-automatic segmentation and using PCA with various levels of variance retained. This snippet demonstrates how the classifiers were trained based on the feature-sets created. All the scripts containing the training of classifiers follow the same format, the reason for showing this particular combination is that it is the best performing combination of features and classifiers. This snippet gives insight into, not only, the process of training the classifiers, but also the metrics generated for each classifier.

```
import numpy as np
import scipy.stats
import pandas as pd
from matplotlib import pyplot as plt
import os
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn import preprocessing
from sklearn.utils import shuffle
from sklearn.model_selection import train_test_split
from sklearn import svm
from sklearn.model_selection import KFold, cross_val_predict,
cross_val_score, cross_validate
from sklearn import metrics
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.neural_network import MLPClassifier
```

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import AdaBoostClassifier

#Load all feature sets in
df_semi_auto = pd.read_csv("Features/ForResults/SemiAutoSegmentation.csv")

# Split up features
df = df_semi_auto.iloc[:,1:] # from this, column 1 is target
x = df.iloc[:,1:]
y = df.iloc[:,0]
scaler = preprocessing.MinMaxScaler(feature_range=(0, 1))
x = scaler.fit_transform(x)

# Make datasets with principle components that capture
# 99%, 95%, 90%, 85% of the variance
pca = PCA(0.99)
principalComponents_99 = pca.fit_transform(x)
principalDf_99 = pd.DataFrame(data = principalComponents_99)
pca_99_comp = pca.n_components_
pca = PCA(0.95)
principalComponents_95 = pca.fit_transform(x)
principalDf_95 = pd.DataFrame(data = principalComponents_95)
pca_95_comp = pca.n_components_
pca = PCA(0.90)
principalComponents_90 = pca.fit_transform(x)
principalDf_90 = pd.DataFrame(data = principalComponents_90)
pca_90_comp = pca.n_components_
pca = PCA(0.85)
principalComponents_85 = pca.fit_transform(x)
principalDf_85 = pd.DataFrame(data = principalComponents_85)
pca_85_comp = pca.n_components_

# SVM Linear
print('SVM Linear')
```

```
SVMLinear = svm.LinearSVC()

y_pred = cross_val_predict(SVMLinear, x, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
print('All features, Accuracy = ',acc)

y_pred = cross_val_predict(SVMLinear, principalComponents_99, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
text = 'PCA 99%, components = '+str(pca_99_comp)+': Accuracy = '+str(acc)
print(text)

y_pred = cross_val_predict(SVMLinear, principalComponents_95, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
text = 'PCA 95%, components = '+str(pca_95_comp)+': Accuracy = '+str(acc)
print(text)

y_pred = cross_val_predict(SVMLinear, principalComponents_90, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
text = 'PCA 90%, components = '+str(pca_90_comp)+': Accuracy = '+str(acc)
```

```
print(text)

y_pred = cross_val_predict(SVMLinear, principalComponents_85, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
text = 'PCA 85%, components = '+str(pca_85_comp)+': Accuracy = '+str(acc)
print(text)

# SVM RBF
print('SVM RBF')
SVMRBF = svm.SVC(gamma=0.5, kernel='rbf')

y_pred = cross_val_predict(SVMRBF, x, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
print('All features, Accuracy = ',acc)

y_pred = cross_val_predict(SVMRBF, principalComponents_99, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
text = 'PCA 99%, components = '+str(pca_99_comp)+': Accuracy = '+str(acc)
print(text)

y_pred = cross_val_predict(SVMRBF, principalComponents_95, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
```

```
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
text = 'PCA 95%, components = '+str(pca_95_comp)+': Accuracy = '+str(acc)
print(text)

y_pred = cross_val_predict(SVMRBF, principalComponents_90, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
text = 'PCA 90%, components = '+str(pca_90_comp)+': Accuracy = '+str(acc)
print(text)

y_pred = cross_val_predict(SVMRBF, principalComponents_85, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
text = 'PCA 85%, components = '+str(pca_85_comp)+': Accuracy = '+str(acc)
print(text)

# SVM Polynomial
print('SVM Poly')
SVMPoly = svm.SVC(kernel = 'poly', degree=3, gamma=1, C=3)

y_pred = cross_val_predict(SVMPoly, x, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
print('All features, Accuracy = ',acc)
```

```
y_pred = cross_val_predict(SVMPoly, principalComponents_99, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
text = 'PCA 99%, components = '+str(pca_99_comp)+': Accuracy = '+str(acc)
print(text)

y_pred = cross_val_predict(SVMPoly, principalComponents_95, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
text = 'PCA 95%, components = '+str(pca_95_comp)+': Accuracy = '+str(acc)
print(text)

y_pred = cross_val_predict(SVMPoly, principalComponents_90, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
text = 'PCA 90%, components = '+str(pca_90_comp)+': Accuracy = '+str(acc)
print(text)

y_pred = cross_val_predict(SVMPoly, principalComponents_85, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
text = 'PCA 85%, components = '+str(pca_85_comp)+': Accuracy = '+str(acc)
```

```
print(text)

# k-NN
print('k-NN')
kNN = KNeighborsClassifier(n_neighbors=3)

y_pred = cross_val_predict(kNN, x, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
print('All features, Accuracy = ',acc)

y_pred = cross_val_predict(kNN, principalComponents_99, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
text = 'PCA 99%, components = '+str(pca_99_comp)+': Accuracy = '+str(acc)
print(text)

y_pred = cross_val_predict(kNN, principalComponents_95, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
text = 'PCA 95%, components = '+str(pca_95_comp)+': Accuracy = '+str(acc)
print(text)

y_pred = cross_val_predict(kNN, principalComponents_90, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
```

```
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
text = 'PCA 90%, components = '+str(pca_90_comp)+': Accuracy = '+str(acc)
print(text)

y_pred = cross_val_predict(kNN, principalComponents_85, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
text = 'PCA 85%, components = '+str(pca_85_comp)+': Accuracy = '+str(acc)
print(text)

# NB
print('Naive Bayes')
nb = GaussianNB()

y_pred = cross_val_predict(nb, x, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
print('All features, Accuracy = ',acc)

y_pred = cross_val_predict(nb, principalComponents_99, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
text = 'PCA 99%, components = '+str(pca_99_comp)+': Accuracy = '+str(acc)
print(text)
```

```
y_pred = cross_val_predict(nb, principalComponents_95, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
text = 'PCA 95%, components = '+str(pca_95_comp)+': Accuracy = '+str(acc)
print(text)

y_pred = cross_val_predict(nb, principalComponents_90, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
text = 'PCA 90%, components = '+str(pca_90_comp)+': Accuracy = '+str(acc)
print(text)

y_pred = cross_val_predict(nb, principalComponents_85, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
text = 'PCA 85%, components = '+str(pca_85_comp)+': Accuracy = '+str(acc)
print(text)

# ANN
print('Simple MLP')
ann = MLPClassifier(hidden_layer_sizes=(15,15,15), solver='sgd', max_iter=1000, lea

y_pred = cross_val_predict(ann, x, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
```

```
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
print('All features, Accuracy = ',acc)

y_pred = cross_val_predict(ann, principalComponents_99, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
text = 'PCA 99%, components = '+str(pca_99_comp)+': Accuracy = '+str(acc)
print(text)

y_pred = cross_val_predict(ann, principalComponents_95, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
text = 'PCA 95%, components = '+str(pca_95_comp)+': Accuracy = '+str(acc)
print(text)

y_pred = cross_val_predict(ann, principalComponents_90, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
text = 'PCA 90%, components = '+str(pca_90_comp)+': Accuracy = '+str(acc)
print(text)

y_pred = cross_val_predict(ann, principalComponents_85, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
```

```
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
text = 'PCA 85%, components = '+str(pca_85_comp)+': Accuracy = '+str(acc)
print(text)

# RF
print('Random forest')
rf = RandomForestClassifier(n_estimators=150)

y_pred = cross_val_predict(rf, x, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
print('All features, Accuracy = ',acc)

y_pred = cross_val_predict(rf, principalComponents_99, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
text = 'PCA 99%, components = '+str(pca_99_comp)+': Accuracy = '+str(acc)
print(text)

y_pred = cross_val_predict(rf, principalComponents_95, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
text = 'PCA 95%, components = '+str(pca_95_comp)+': Accuracy = '+str(acc)
print(text)
```

```
y_pred = cross_val_predict(rf, principalComponents_90, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
text = 'PCA 90%, components = '+str(pca_90_comp)+': Accuracy = '+str(acc)
print(text)

y_pred = cross_val_predict(rf, principalComponents_85, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
text = 'PCA 85%, components = '+str(pca_85_comp)+': Accuracy = '+str(acc)
print(text)

# AdaBoost
print('AdaBoost svm linear')
svc=svm.SVC(probability=True, kernel='linear')
ada = AdaBoostClassifier(n_estimators=100, learning_rate=0.1, base_estimator=svc)

y_pred = cross_val_predict(ada, x, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
print('All features, Accuracy = ',acc)

y_pred = cross_val_predict(ada, principalComponents_99, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
```

```
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
text = 'PCA 99%, components = '+str(pca_99_comp)+': Accuracy = '+str(acc)
print(text)

y_pred = cross_val_predict(ada, principalComponents_95, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
text = 'PCA 95%, components = '+str(pca_95_comp)+': Accuracy = '+str(acc)
print(text)

y_pred = cross_val_predict(ada, principalComponents_90, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
text = 'PCA 90%, components = '+str(pca_90_comp)+': Accuracy = '+str(acc)
print(text)

y_pred = cross_val_predict(ada, principalComponents_85, y, cv=30)
tn, fp, fn, tp = metrics.confusion_matrix(y, y_pred).ravel()
c = metrics.classification_report(y, y_pred)
acc = metrics.accuracy_score(y, y_pred)
fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
auc = metrics.auc(fpr, tpr)
text = 'PCA 85%, components = '+str(pca_85_comp)+': Accuracy = '+str(acc)
print(text)
```