BAŞKENT UNIVERSITY
INSTITUTE OF SCIENCE AND ENGINEERING
DEPARTMENT OF COMPUTER ENGINEERING
MASTER IN COMPUTER ENGINEERING WITH THESIS

BREAST CANCER DIAGNOSIS FROM THERMAL IMAGES

MASTER OF SCIENCE THESIS

BY

ÇAĞRI CABIOĞLU

ANKARA - 2020

BAŞKENT UNIVERSITY
INSTITUTE OF SCIENCE AND ENGINEERING
DEPARTMENT OF COMPUTER ENGINEERING
MASTER IN COMPUTER ENGINEERING WITH THESIS

BREAST CANCER DIAGNOSIS FROM THERMAL IMAGES

MASTER OF SCIENCE THESIS

BY

ÇAĞRI CABIOĞLU

ADVISOR

PROF. DR. HASAN OĞUL

ANKARA – 2020

BAŞKENT UNIVERSITY

INSTITUTE OF SCIENCE AND ENGINEERING

This study, which was prepared by Çağrı Cabıoğlu, for the program of Computer Engineering Master's Program with Thesis, has been approved in partial fulfillment of the requirements for the degree of MASTER OF SCIENCE in Computer Engineering Department by the following committee.

Date of Thesis Defense: 03 / 02 / 2020

Thesis Title: Breast Cancer Diagnosis From Thermal Images

Examining Committee Members **Signature**

Assoc. Prof. Dr. Hasan Şakir BİLGE, Gazi University

Prof. Dr. Hasan OĞUL, Başkent University

Asst. Prof. Dr. Mehmet DİKMEN, Başkent University

APPROVAL

Prof. Dr. Faruk ELALDI
Director, Institute of Science and Engineering

DATE: … / … / ……..

BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLER ENSTİTÜSÜ
YÜKSEK LİSANS TEZ ÇALIŞMASI ORİJİNALLİK RAPORU

Tarih: 21 / 02 / 2020

Öğrencinin Adı, Soyadı : Çağrı Cabıoğlu

Öğrencinin Numarası : 21710281

Anabilim Dalı : Bilgisayar Mühendisliği

Programı :Bilgisayar Mühendisliği Tezli Yüksek Lisans

Danışmanın Unvanı/Adı, Soyadı : Prof. Dr. Hasan Oğul

Tez Başlığı : Thermal Görüntülerden Meme Kanseri Teşhisi

Yukarıda başlığı belirtilen Yüksek Lisans tez çalışmamın; Giriş, Ana Bölümler ve Sonuç Bölümünden oluşan, toplam 42 sayfalık kısmına ilişkin, 21 / 02 / 2020 tarihinde şahsım tarafından Turnitin adlı intihal tespit programından aşağıda belirtilen filtrelemeler uygulanarak alınmış olan orijinallik raporuna göre, tezimin benzerlik oranı %10'dur.

Uygulanan filtrelemeler:

1. Kaynakça hariç
2. Alıntılar hariç
3. Beş (5) kelimeden daha az örtüşme içeren metin kısımları hariç

"Başkent Üniversitesi Enstitüleri Tez Çalışması Orijinallik Raporu Alınması ve Kullanılması Usul ve Esaslarını" inceledim ve bu uygulama esaslarında belirtilen azami benzerlik oranlarına tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Öğrenci İmzası:…….

Onay

… / 02 / 2020

Öğrenci Danışmanı Unvan, Adı, Soyadı,

ACKNOWLEDGEMENTS

First of all, I would like to thank my thesis supervisor Prof. Dr. Hasan Oğul for his guidance and contributions to the realization of this study.

I would like to present my deepest gratitude to my dear fiancé Yeşim Kalkan who always inspired me and motivates me throughout the thesis.

Also, I would like to express my deepest gratitude to my dear mother Nezahat Cabıoğlu and my father Mehmet Tuğrul Cabıoğlu for their support to me.

ÖZET

Çağrı Cabıoğlu
TERMAL GÖRÜNTÜLERDEN MEME KANSERİ TEŞHİSİ
Başkent Üniversitesi Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı
2020

Meme kanseri, kadınlar arasında en yaygın kanser türlerinden biridir. Göğüs kanserinin erken teşhisi ve tedavisi hastalar için hayati öneme sahiptir. Göğüs kanserine yakalanma oranı gün geçtikçe artar iken, erken teşhis teknikleri sayesinde ölüm oranları azalmaktadır. Gelişen teknoloji ile görüntüleme sistemlerinde birçok önemli gelişmeler yaşanmıştır. Kanserin saptanmasında çeşitli görüntüleme teknikleri kullanılmaktadır. Termal görüntüler, termal kamera tarafından radyasyon verilmeden bölgelerin sıcaklık farkı kullanılarak elde edilir. Bu çalışmada, termal görüntüler kullanılarak meme kanserinin bilgisayar destekli tanısı için yöntemler sunulmaktadır. Bu amaçla, transfer öğrenme metodolojisi kullanılarak çeşitli Evrişimli Sinir Ağı (CNN) modelleri tasarlanmıştır. Tasarlanan ağların performansı, doğruluk, kesinlik, hatırlama, F1 ölçüsü ve Matthews Korelasyon katsayısı dikkate alınarak bir kıyaslama veri kümesinde değerlendirilmiştir. Sonuçlar, önceden eğitilmiş evrişimsel katmanların tutulması ve yeni eklenen tam bağlantılı katmanların eğitiminin en iyi puanları verdiğini göstermektedir. CNN ile transfer öğrenme metodolojisini kullanarak %94.3 doğruluk, %94.7 hassasiyet ve %93.3 duyarlılık elde ettik.

ANAHTAR KELİMELER: Derin Öğrenme; Evrişimsel Sinir Ağları; Termal görüntü; Görüntü İşleme; Kanser; Gögüs Kanseri; AlexNet;

ABSTRACT

Çağrı Cabıoğlu

BREAST CANCER DIAGNOSIS FROM THERMAL IMAGES

Başkent University Institute of Science and Engineering

Department of Computer Engineering

2020

Breast cancer is one of the prevalent types of cancer. Early diagnosis and treatment of breast cancer have vital importance for patients. Various imaging techniques are used in the detection of cancer. Thermal images are obtained by using the temperature difference of regions without giving radiation by the thermal camera. In this study, we present methods for computer aided diagnosis of breast cancer using thermal images. To this end, various Convolutional Neural Network (CNN) models have been designed by using transfer learning methodology. The performance of the designed nets was evaluated on a benchmarking dataset considering accuracy, precision, recall, F1 measure, and Matthews Correlation coefficient. The results show that holding pre-trained convolutional layers and training newly added fully connected layers gives the best scores. We have obtained an accuracy of 94.3%, a precision of 94.7% and a recall of 93.3% using transfer learning methodology with CNN.

KEYWORDS: Deep Learning; Convolutional Neural Network; Transfer Learning; AlexNet; Thermal Image; Image Processing; Breast Cancer

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

°C	Celsius
ACC	Accuracy
ACS	The American Cancer Society
Adam	Adaptive Moment Estimation
ANN	Artificial Neural Network
CNN	Convolutional Neural Network
CPU	Central Processing Unit
DMR	Database for Mastology Research
FN	False Negatives
FP	False Positives
FPA	Focal Plane Array
GPU	Graphical Processing Units
IR	Infrared
K	Kelvin
MCC	Matthews correlation coefficient
MRI	Magnetic Resonance Imaging
SVM	Support Vector Machine
TN	True Negatives
TP	True Positives

1. INTRODUCTION

1.1. Motivation and Background

Cancer is the general name for diseases that starts with the incessant division of certain cells of the body and its spread to surrounding tissues. The starting point of cancer can be from any part of the human body, which consists of trillions of cells. In the normal functioning of the human body, cells grow and divide to form new cells. The death of the cell occurs when it is aged or damaged [1]. Cell division allows the body to fill dead cells with new cells or where it is needed. In areas where cancer develops, this process changes. The cells that should die begin to survive and form new cells without control. Unlike normal cells, tumor cells do not specialize in any work required for the body. So, while normal cells have functionality for the body, cancer cells do not. For example, cancer cells can ignore apoptosis signals and escape the immune system [2]. These cells grow and divide uncontrolled, forming masses called tumors. Many types of cancer form solid tumors like tissue mass, with the exception of blood cancers such as leukemia. Malignant cancerous tissues can spread to other parts of the body. They spread through the blood or the lymphatic system and the new tumor may be seen far away from the point of origin. There are benign tumors as well as malignant tumors. These tumors are called benign because they do not spread to other tissues. The size of such tumors varies. However, depending on the location of the tumor, for example brain tumor, they can be life-threatening. Benign tumors usually do not grow back when removed by certain techniques. There are more than 100 different types of cancer [3]. These species are often referred with by the names of the organs or tissues in which they are formed.

Breast cancer is the most common cancer among women worldwide [4,5]. In 2012, 12% of cases diagnosed with cancer and 25% of women diagnosed with cancer were found to have breast cancer. This corresponds to approximately 1.7 million [6]. The number of breast cancer cases has increased day by day and is the second disease that causes death in women [7]. There are many factors that increase the risk of breast cancer. The first full-term pregnancy age, the age of menopause, the presence of some mutations in the BRCA1 or BRCA2 genes, ionizing radiation to the breast, high alcohol consumption, obesity or higher body mass index are some of these factors [8, 9].

Early cancer detection and appropriate treatment can reduce breast cancer mortality

[10]. The incidence of breast cancer has increased in the last few years, but the mortality rate is decreasing [11]. This reduction is because of new imaging methodologies to diagnose breast cancer. These techniques help doctors to make an accurate diagnosis and localization.

1.1.1. Breast imaging techniques

To date, screening techs such as mammography, magnetic resonance imaging, ultrasound and thermal imaging have been developed for breast cancer diagnosis. These techniques have many advantages under favorable conditions. There are also limitations such as being expensive, inefficient in time and certain age limits.

1.1.1.1. Mammography

Mammography is an imaging technique in which an X-ray image of the breast is taken. The American Cancer Society (ACS) recommends the annual application to women over 40 years. The mammogram is more effective in women aged between 40 and 74 years. Since women under the age of 40 have a high breast density, false-positive and false-negative mammography rates are high in such patients [12]. there are many factors that affect the sensitivity of mammographic images. For example, the patient's age and personal history, the experience of the radiologist and the technical capacity and the adequacy of the device can change the efficiency of the mammogram image. In addition, mammogram sensitivity is low in women with dense breasts or premenopausal women [13]. The mammogram uses ionizing radiation to capture the image of the breast and is harmful to health as it causes changes in human DNA. In addition to a mammogram, CE digital mammography techniques based on tumor angiogenesis are also used to increase sensitivity. Intravenous iodine contrast injections are used in this method, which means greater exposure to radiation [14]. Despite all these limitations, mammography is currently being actively used for the diagnosis of breast cancer.

1.1.1.2. Magnetic resonance imaging (MRI)

MRI is a breast imaging technique using high dose X-ray and radio waves. MRI can take images from different sections of the breast and the sections can be combined to provide more detailed localization of tumors. Imaging with MRI has more detailed information also its sensitivity higher than other techniques. This technique is often used to find small tumors that cannot be detected by other methods or for patients who are at high risk of cancer.

However, MRI is a disadvantageous method according to false-positive rate, cost, inefficient use of time (about 1 hour) and uses high dose X-ray which is harmful to the body. In the MRI guidelines published by ACS, it is recommended only to high-risk patients and patients with BRCA mutations annually [15]. MRI is more effective in terms of sensitivity compared to other imaging techniques, but it is useful to make sure that other methods that are less harmful to health cannot be diagnosed the breast cancer.

1.1.1.3. Ultrasound

Ultrasonography is a low-cost technique for visualizing the chest region using sound waves. The location of the tumor is determined using acoustic properties of sound waves sent to the tissue. It is a technique with a high detection rate in patients at high risk of cancer. It is an effective method for identifying cysts and solid kits. Breast ultrasonography is recommended for pregnant women and patients who are not suitable for mammography because it does not emit radiation [16]. Ultrasonography is a non-invasive and non-radiation method that makes it one step ahead of MRI and mammogram. However, the sensitivity is lower than MRI and mammogram. Particularly because of the same acoustic properties of healthy and tumorous tissues, there are difficulties in detection.

1.1.1.4. Thermography

The use of temperature to diagnose disease is a well-known and old method. They are known to have scanned the body with their hands to determine the temperature change during the period of Ancient Egypt. The Greek doctor Hippocrates applied mud slurry to the patients and observed that the first drying area was a diseased area [17]. But at that time there was a lack of reference point to measure the temperature**.**

Galileo developed the first known temperature measuring sensor. Fahrenheit introduced an approach that fixed the boiling point of water to 212 and the freezing point to 0. Then, with the commonly used unit of celsius, a reference point was formed which equals the boiling point of water to 100 and the freezing point of water to 0.

In the 1800s, Sir William Herschel discovered infrared radiation while trying to measure the color temperature of the rainbow. Through this discovery, he observed that the high-temperature rays were beyond the red color in the spectrum. He was able to obtain an image using the carbon suspension in alcohol and was called a thermogram. The completed electromagnetic spectrum that emerges with the studies after this study is as follows:

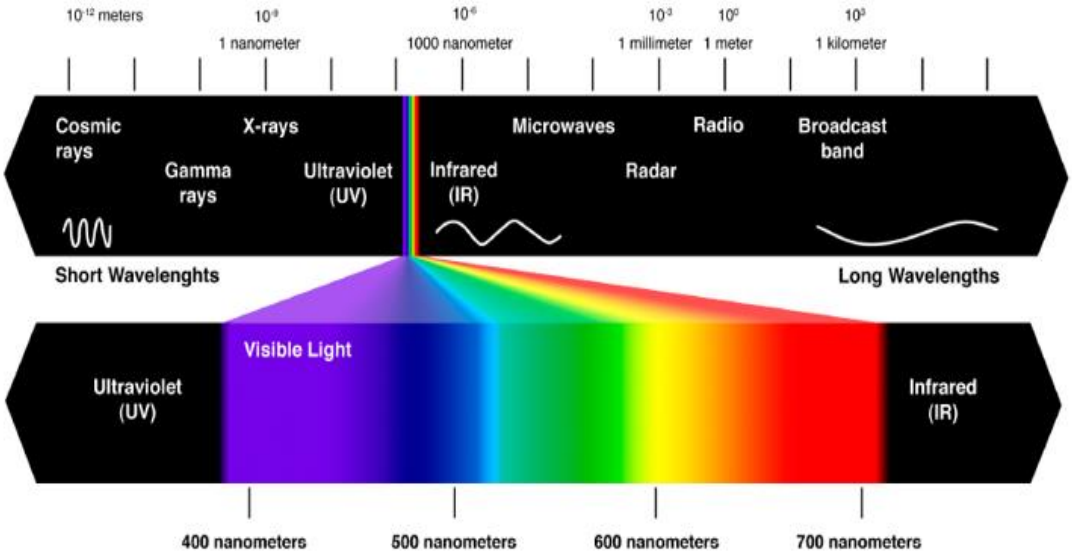

Figure 1.1. Electromagnetic Spectrum

As seen in Figure 1.1, the visible light on which Herschel works is a small area in the entire spectrum. The radiation zone with a wavelength longer than visible light is called the infrared zone. This area is divided into 5 bands (Figure 1.2). Long-wavelength infrared is also called thermal imaging or thermal infrared field. This field is suitable for capturing radiation emitted by heat-producing substances, such as the human body, and is often used for biomedical solutions.

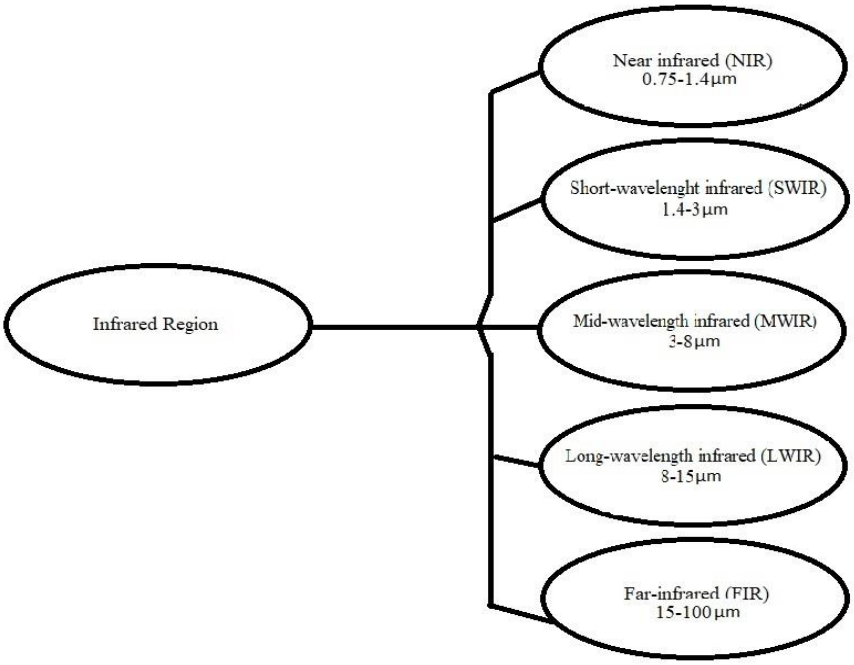

Figure 1.2. Infrared Region

History of Thermal Detectors

Thermal detectors are designed to detect radiation from certain wavelength bands. In this way, they aim to work with high permeability and minimum losses in the selected wavelength.

In 1969, Richard D. Hudson identified the application areas of thermal imaging systems, medical, scientific, military and industrial. Thermal imaging systems for these four areas have different requirements and designs. Today, this technology is mostly used in the military field.

Thermal detectors were first developed in the 1940s with 1 - 2.5 μm wavelength sensitive lead sulfur (PbS). In addition, when the lead sulfide cooling process is applied, radiation sensitivity is observed in the band range of 2 - 4 μm. This means that with this band range can capture near-infrared radiation. With the discovery of indium antimonite (InSb) in 1952, Heinrich Walker paved the way for thermal imaging systems capable of operating beyond the middle band red. After this discovery, Lawson discovered mercury cadmium tellurium (HgTeCd) in 1959 and detectors capable of operating in the far-infrared band (8-14 μm) were introduced.

Thermal systems were first used in military fields. The use of medicine in the field

began in the 1970s. The technology of thermal imaging systems is basically divided into two: liquid crystal contact thermography and non-contact black body. At the beginning, camera performance was not at the desired level due to being the newly developing technology. The liquid crystal thermography used in the medical field in the 1970s was insufficient in terms of thermal resolution (± 0.5°C), the response time (> 60 sec) and spatial resolution.

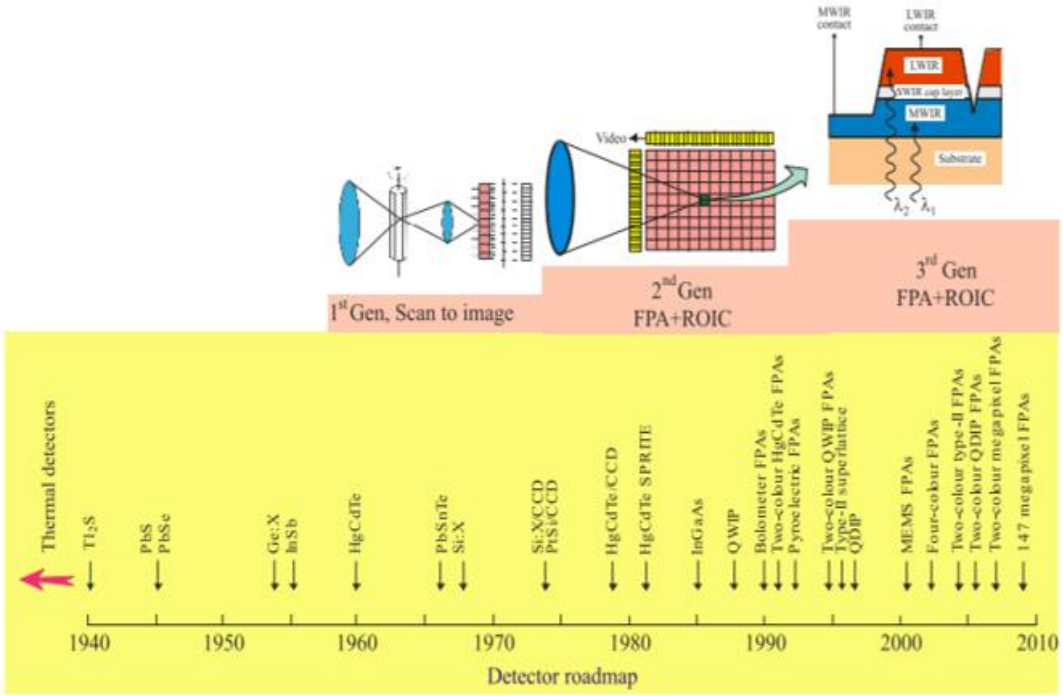

Figure 1.3. History of the development of thermal detectors [18]

With the development of the second and third generation sensors, studies on thermal imaging systems started to increase. With the development of focal plane array (FPA) cameras in the 1990s, images with a resolution of 256x256 and faster than 100 frames / sec began to be acquired. Then came designs with high resolution and temperature sensitivity of 0.01 °C.

There are also sensors that operate at room temperature and do not require cooling. Such thermograms are called "uncooled detectors". These types of systems which are small and easy to use, but have lower sensitivity than cooling systems. However, they have reached levels that can be used in disease diagnosis with developing technology. The FLIR-SC620 thermal imager used in this study is sensitive to 7.5 - 13 μm spectral region and has a sensitivity of >40mK.

Physics of Infrared

Thermal cameras calculate electromagnetic radiation or energy continuously emitted from the source object to measure the temperature. For this process, they capture IR emissions from the source object. Energy movements depend on the surface emission of the source and the wavelength of the radiation emitted. According to Kirchhoff's law, when black bodies or opaque materials are in thermal equilibrium, the energy absorbed and the energy emitted at any wavelength are equal. At equation 1.1, the reflection rate is shown as α and emittance is ε.

$$\alpha(\lambda) = \varepsilon(\lambda) \tag{1.1}$$

Kirchhoff was the first to propose the term black body. According to his law, good absorbers are weak reflectors. Therefore, ideal black objects do not reflect any light and appear black.

In 1879, Stefan-Boltzmann introduced a law defining the radiation emitted on surfaces of objects whose temperature was above absolute zero (-273° C). According to equation 1.2, the total emitted radiation power (P) depends on the surface area of the object (A) and the absolute temperature of the object (T) . ε is the emissivity and σ is the Stefan-Boltzmann constant (J/sm^2K^4) in eq 1.2.

$$P = A\varepsilon\sigma T^4 \tag{1.2}$$

When we evaluate the equation 1.2 according to the radiation emitted from the unit area, the result shows that the emitted radiation depends on emissivity and absolute temperature of the source object.

$$W \sim \varepsilon T^4 \tag{1.3}$$

Where, the total radiation emitted per unit area is shown in W (watts / m2), emission ε and the absolute temperature T (K). The emission value can be maximum 1. It is the ratio of the radiating power of an object to the produced energy by a black body at the identical temperature and wavelength. In the real world, there are gray bodies instead of substances like black bodies. In order to measure the surface temperature of a source object, it must have an emissivity.

In 1901, Max Planck proposed Planck's law, which showed the relationship between the intensity of radiation emitted from the surface of black bodies and the wavelength:

$$B_\lambda(T) = \frac{2hc^2}{\lambda^5} \frac{1}{e^{\frac{hc}{\lambda k_B T}} - 1} \tag{1.4}$$

In equation 1.4, B_λ ($Wsr^{-1}m^{-3}$) represents spectral radiation, h represents the Plack constant, c is the speed of light, λ (m) represents wavelength, k_B (JK^{-1}) represents the Boltzmann constant and T (K) is the absolute temperature.

Metabolism of Tumors and Tumor Growth

Blood flow and metabolic activities are higher in pre-cancerous or breast-tumor tissues than in healthy tissues. The temperature increase is observed in regions where these activities are intense. The diagnosis of cancer by thermogram is based on this principle.

The human body activates the capillaries for healing and reproduction, providing the flow of nutrients and oxygen to the desired site, called angiogenesis. In normal tissues, this procedure is performed with a certain discipline. However, cancerous tissues disrupt this order and secrete proteins that initiate the angiogenesis process. As the capillaries increase, the tumor begins to grow and temperature increases in the area where the tumor is located. According to the studies, the average time required for the tumor to grow from 10 to 20 mm in diameter was 1.7 years. This time is 1-2 months for some tumors and approximately 6 years for others [19]. Tumors of small size may not be detected by X-ray or manual controls. However, capillary growth around the tumor tissue can be observed by IR cameras at an early stage.

IR thermography to detect breast cancer

Breast cancer is the second type of cancer that causes death in women after lung cancer. Therefore, early detection methods are of vital importance as they reduce mortality rates. Thanks to the technological advancement in thermal imaging systems, it has become one of the imaging systems used for early detection of breast cancer. In this section, studies in the field of cancer diagnosis by thermal imaging are given. In these studies, thermal images were studied by elongated people and the diagnosis of cancer was determined by these people. There are also studies using computer-aided diagnostic systems in Chapter 2.

The first study to use thermography for breast cancer diagnosis is by Lawson [20]. In the study conducted in 1956, Lawson investigated in 26 patients diagnosed with breast cancer. According to the results, an average increase of 2.27 °F was observed in where the tumor located.

In a study of 4,000 patients using thermography in 1965 by radiologist Gershon - Cohen, 94% sensitivity and 6% false-positive rates were reported [21].

In 1972, Isard and colleagues showed that the use of thermography and mammography together was an effective method for the diagnosis of cancer [22]. They collected data of 10055 patients with both systems for 4 years. They divided the patients into two groups asymptomatic and symptomatic. Abnormalities were found in approximately 25% of asymptomatic patients who have not got any complaints. As a result, the author says that the diagnosis made by thermal imaging systems increases the accuracy in high-risk patients and increases the detection sensitivity by 10% when used with the mammogram.

The study by Spitalier and colleagues shows the importance of thermogram in early diagnosis [23]. From a 10-year study, they reached 89% accuracy in the diagnosis of breast cancer. Another important point in the study is that 60% of the patients receive cancer stimulation from the first thermograph. According to these results, it shows the effectiveness of thermal imaging systems in determining the risk of breast cancer even in patients without any clinical complaints.

Michel Gautherie and Charles M. Gros in 1980, they concluded that thermal imaging can be used to predict breast cancer risk [24]. In this long-term study, 58000 patients underwent physical examination, x-ray, ultrasound and thermography in groups for 12 years. Of the 1527 women diagnosed abnormally by thermogram and diagnosed normal by other imaging systems, 44% were found to develop breast cancer within 5 years. This study demonstrates that thermography can achieve better results than other imaging systems for early diagnosis.

In 1984, Thomassin et al. In his study, 4000 patients diagnosed with breast cancer were examined [25]. 130 of these patients had a tumor size of 3-5 mm. Mammography, thermography and a combination of these two systems were used in all patients to detect breast cancer. Only 10% of 130 patients who have 3-5 mm tumor were diagnosed with cancer by mammography, and 50% by thermography. The remaining patients were diagnosed using a combination of mammogram and thermography. They reported that 90% of all patients were diagnosed with cancer by thermogram.

In a study conducted by Parisky and colleagues in 2003, thermogram images were

taken from 769 patients who were found suitable for biopsy. Of 875 lesions biopsied, 187 were malignant and 688 were benign tumors [26]. The regions marked on thermal images and the regions where the biopsy was performed were compared. As a result, 97% sensitivity was obtained by thermography.

1.2. Scope of Thesis

With the advancement in the field of use, thermal camera technology is expanding day by day. Thanks to the improvements in the resolution, sensitivity and usability of the thermal imagers that paved the way for this expansion. Thermal cameras which are used in the field of health play an important role in the diagnosis of pain, dry eye and cancer. The differences in the temperature map of the diseased area are at a level that can be easily distinguished by the eye thanks to the thermal cameras used today. Breast cancer is also one of the types of cancer that can be distinguished from thermal appearances. In this type of cancer, there are factors such as temperature difference between the two breasts, deformity on the side with cancer and asymmetry. In this study, we examine the disease diagnosis from the thermograms taken from the chest area with a thermal camera.

1.3. Contribution

This study makes the following contributions. First, we propose a convolutional neural network architecture to classify patients as healthy or sick from IR images. Next, we design CNN models using transfer learning and show how the selection of transferred layers affects performance that evaluates using cross-validation technique. Last, we compare results using five metrics which are accuracy, precision, recall, F1 measure and Matthews correlation coefficient (MCC).

1.4. Outline

This thesis is divided into 5 main chapters:

- Introduction
- Related Work
- Materials and Methods
- Experiments and Results
- Conclusion and Discussion

The first part contains the introduction where the subject, scope and contribution of

the thesis are explained. The second part is about the methods used to solve the problem and the results obtained by the previous studies in this field.

The third part explains our materials and methodology used for classifying IR images. It contains a detailed explanation of deep learning, transfer learning and the dataset. The fourth part is the section where our experimental design and the models we design are explained, and comparative results of the proposed method are shown. The last part is allocated to the conclusion and future work.

2. RELATED WORK

In 2012, Acharya et al. [27] conducted a study for the diagnosis of breast cancer from IR images with a computer-aided diagnosis. In this study, thermal images of a total of 50 patients with 25 normal and 25 cancers from Singapore General Hospital were used. In the pre-processing phase, they applied background subtraction and right-left breast segmentation by cropping the IR images. In the feature extraction phase, they extracted 16 texture features from the co-occurrence matrix and run-length matrices. They gave these feature vectors as input to the support vector machine (SVM) algorithm to classify between diseased and healthy IR images. As a result, they achieved 88.10% accuracy, 85.71% sensitivity and 90.48% specificity.

The work of Sheeja V. Francis and M. Sasikala consists of 3 main sections and has experimented on 27 images [28]. They performed pre-processing, feature extraction and classification for the training of the system that can diagnose breast cancer from IR images, respectively. During the pre-processing phase, they separated the images manually into the right and left breasts. A total of 27 texture features were used for feature extraction. 20 co-occurrence matrix and 7 run-length matrix from a total of 27 units features are used. A feed-forward back propagation network has one hidden layer with 17 neurons and output layer with 2 neuron was used for classification and they created a system that reached 85.19% accuracy.

In 2014, another study for the diagnosis of automatic breast cancer was made by Bartosz Krawczyk and Gerald Schaefer [29]. 146 thermal images (29 malignant and 117 benign) were used in the study and multiple classifier fusion techniques were used. In the pre-processing phase, the Laplacian filter and contrast enhancing were applied. A total of 38 features were extracted to find the asymmetry between the right and left breasts. In the study, multiple classifier fusion with neural network yielded 90.03% accuracy, 80.35% sensitivity and 90.15% specificity.

Araújo et al. [30] developed non-automatic segmentation and three stage feature extraction methodology to classify breast cancer. The first stage is extracting maximum and minimum temperature value from a morphologically processed IR image. The second stage includes extracting interval features and producing continuous features. Using Fisher's criterion continuous features mapped on new feature space for the last stage. Parzen-window, linear discriminant and distance-based classifiers applied on these new feature

spaces. They achieved 85.7% sensitivity and 86.5% specificity.

Krawczyk and Schaefer [31] developed another methodology for detecting breast cancer. Their approach was based on obtaining image features which describes asymmetry of normal and abnormal breast. For feature extraction, they extract basic statistical features, histogram features, image moments, and various texture features. In pattern classification part, they combine multiple classifiers using neural network fusion approach and evolutionary computing. This method was tested on 146 static IR breast image in which 29 of them are sick and 117 of them are healthy. They reported a 90.03% accuracy, 80.35% sensitivity and 90.15% specificity.

Another paper is written by Gaber et al. [32] Their approach has two stages: automatic segmentation and classification. In segmentation part, they used Neutrosophic sets (NS) and optimized Fast Fuzzy c-mean (F-FCM) algorithm. Support vector machine (SVM) used for classification of normal and abnormal breast and tested on 29 healthy and 34 malignant images. They reported an 92.06% accuracy, 96.55% recall and 87.50% precision.

Baffa and Lattari [33] developed another methodology based on convolutional neural networks. They used static and dynamic IR images. Data set tested on their CNN architecture. Data augmentation, mean subtraction and gray to RGB transformation applied on IR images separately. For static protocol, they reported an 98% accuracy, 97% sensitivity and 100% specificity.

3. MATERIALS AND METHODS

3.1. Problem Definition

The main aim of the thesis is the diagnosis of breast cancer from infrared images with convolutional neural networks which have proven themselves in the fields of vision and machine learning area.

The data set used in experiments is prepared by domain experts and all the data belongs to the patients of Antonia Pedro University Hospital. The images are taken with FLIR SC-620 thermal camera and have 640x480 resolution.

Our proposed method is based on a transfer learning methodology. We designed four different CNN models to investigate the effect of the transferred layers from our base model, AlexNet. In addition, we have prepared different types of inputs to examine the effect of the input type.

3.2. Deep Learning

Deep learning has become a popular topic of machine learning amongst researchers in the last decade [34,35]. With the increase of data amount and demanding of solving more complex problems, traditional machine learning algorithms such as support vector machine (SVM), decision tree, and naïve bayes could not perform sufficient solutions in real-world applications. In recent years, deep networks have a significant impact on performance improvement of classification problems in various areas such as speech recognition, data science, robotics, computer vision, audio processing and medical areas.

Deep learning inspired by the human thinking process is a branch of machine learning. Deep Learning uses sequential operations to process input data [36]. These operations are called “layers”. Each of the layers’ output is the input of the next layer. Thus, the information is passed through each layer. Deep networks consist of three layers. The input layer is the first layer and the last layer is called an output layer. Hidden layers located between the input and output layer.

Deep learning architectures have many more layers than other classical neural networks and this feature seems the basis of their good performance. Also, newly found architectural modifications like rectified linear activations (ReLUs) [37] and residual “shortcut” connections [38] have an impact on their performance. Increasing the number of layers increases the number of parameters to be optimized. This problem can overcome with

efficient graphical processing units (GPU) which can compute high-dimensional optimization problems within less time compared to the central processing unit (CPU). Moreover, feature extraction in deep learning generates automatically meaningful features without predefining them manually. With all of the reasons, the popularity of deep learning is increasing day by day.

3.2.1. Deep feedforward networks

Deep feedforward networks are the foundation of most deep learning models. Also, in literature, deep feedforward networks are known as "feedforward neural networks" or "multilayer neural networks". These networks are mostly used for solving supervised machine learning problems and have significant importance at many commercial application areas such as computer vision and natural language processing (NLP).

The main purpose of a deep feedforward network is to define or approximate some regression function f*. This function maps an input x to an output y. A feedforward network defines a mapping $y = f(x; \theta)$ and learns the value of the parameters θ that result in the best function approximation [39]. The x used to define some intermediate functions in the hidden layer sequentially to calculate 'y'. Thus, information flow is provided forward direction. That's why these networks are called feed forward. A combination of many different functions forms these networks and all these functions are composed together. Layers in the feed forward network structure are connected like a chain and the number of layers determines the depth of the network. Hidden layers are between the input layer and the output layers. In the training phase, these layers cannot produce the desired output of training. Hidden layers consist of any number of hidden units known as a neuron. A neuron takes inputs from the neuron of previous layers and processes them with its activation value.

3.2.1.1. Gradient-based learning

There is not much difference between to train a neural network and training other machine learning algorithms with gradient descent. The main goal of the gradient-based techniques is to find the parameters that minimize the error in prediction. Using gradient-based learning guarantees convergence on linear models such as logistic regression or SVMs, since these models have convex loss function. However, deep neural networks produce nonconvex loss functions. In general, iterative gradient-based optimizers are preferred for training a neural network. This means that optimization algorithms like gradient

descent cannot guarantee global convergence and has a high dependence on the values of initial parameters. Assigning small random values for all weights at the beginning is essential. Also, the cost function and representation of the output model should be chosen to train the neural networks.

The cost function, C, shows the difference between the estimated output of the designed network and the actual output that the network tries to reach. This means that it evaluates how good our network predictions. Weight 'w' and bias 'b' vectors of the network are the variables of the function C, i.e; C = C(W) where W = (w_1, b_1, w_2, b_2,….). To find out how the change in the value of weights and biases will affect the cost with

$$\Delta C \approx \nabla C \,.\, \Delta W \tag{3.2}$$

Where ∇ means directional derivative. If ΔW is selected as

$$\Delta W = -\mu \nabla C \tag{3.3}$$

Where μ is called learning rate, and equation 3.2 can be written as

$$\Delta C \approx -\mu \nabla C \,.\, \nabla C = -\mu \|\nabla C\|^2 \tag{3.4}$$

The equation 3.4 shows that C will decrease if we change the value of W according to equation 3.3 because of $\Delta C \leq 0$, $\|\nabla C\|^2 \geq 0$ and μ is positive constant. The equation 3.3 can be used to calculate ΔW and update W as

$$W \longrightarrow W' = W - \mu \|\nabla C\|^2 \tag{3.5}$$

After the update we can do this procedure with new values of W and calculate the cost function again. If all these procedures are performed repeatedly, values of W can be near or equal to a global or local minimum. There are many types of cost function and training phase of feedforward neural networks is profoundly depending on the choice of a cost function.

The main goal of the activation function is to transform an output signal of a node in the network and give it to the next node as an input. These functions are generally a non-linear transformation function. Nonlinear activation functions are more powerful than linear activation functions to model and learn complicated data such as images, speech, video or audio. One of the most popular activation functions is the rectified linear unit (ReLU) function which is computationally efficient and used in this thesis for this reason. It is defined as $\boldsymbol{y = max(0, x)}$.

3.2.2. Overfitting and regularization in deep learning

The primary purpose of the deep neural networks is to produce a model that works well with the data used in training and also new data are unseen before. The good performance on the previously unobserved data is also called generalization. The generalization of the data is the central goal in machine learning. The overfitting model is a model that learning too much the training data and performs well on training data and not good enough on new data. The powerful side of the deep neural networks is that their performance increases as they are continuously trained with large data sets. That's why the deep neural networks have a capacity for overfitting the problem. There are many regularization techniques for decreasing the generalization error. Also, these can lead to faster optimization and better performance overall.

3.2.2.1. L2 and L1 regularizations

Regularization terms are added to the loss function to avoid overfitting and to make the model simpler [39]. L2 is one of the most used regularization terms in deep neural network and basically it is the sum of square of all weights as shown in equation 3.6. The term 'w' is representing the feature weights of the ith neuron. L2 regularization shrinks the weights to small values but not permits them to be zero. The L2 regularization formula shows that small weights have little impact on the model complexity, but outliers have a massive impact because of the square of feature weights.

$$\delta(\theta) = \frac{1}{2}\sum \mathrm{w}_\mathrm{i}^2 \tag{3.6}$$

On the other hand, the L1 regularization term show in equation 3.7 tries to shrink parameters to zero. This means L1 regularization makes feature selection and produces a subset of the feature vector. Also, the L1 term has a sparse solution and is robust to outliers. Therefore, in case of a vast number of features, the L1 regularization term performs better for the feature selection than the L2 term. However, the disadvantages of the L1 term is the model is that produced with this term cannot learn complex patterns.

$$\delta(\theta) = \frac{1}{2}\sum |\mathrm{w}_\mathrm{i}| \tag{3.7}$$

3.2.2.2. Dataset augmentation

Collecting real data might be difficult and time-consuming in some cases i.e. image, video, sound, audio and signal data collection. Training machine learning algorithms with small dataset might cause some problems such as overfitting or underfitting. To reduce overfitting, producing synthetic data using label-preserving transformation from the original dataset is the most common and the easiest methodology [40,41,42]. In theory, enlarging the dataset will result in a model that performs better generalization of training data [40]. There are many ways to apply dataset augmentation, but these methods vary according to a domain. For example, in image data, rotation, noise addition, cropping, lighting condition change, translation and blur making could be applied to increase samples of the training dataset.

3.2.2.3. Early stopping

Hyperparameters play key role to achieve optimal generalization performance when training a deep neural network and number of training epochs is one of them. Basically, one epoch is one forward and backward pass of the full training set. The model trained with a small number of epochs may underfit the problem. Also, while using more epochs the model seems to get better at each epoch but at some point, the error on the validation set increases and the model may overfit the problem.

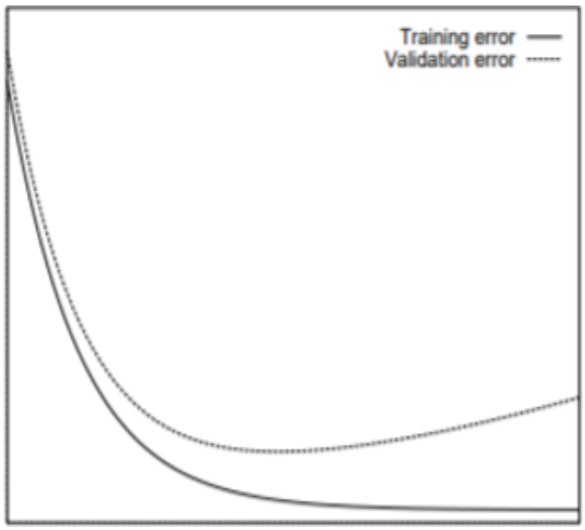

Figure 3.1. Idealized training and validation error curves. Vertical: errors; horizontal: time [43].

Early stopping is one of the ways to determine the optimum number of epochs. In the early stopping approach, while training a neural network, validation error evaluated at each epoch and neural network training stopped when validation error starts to increase [43]. Early stopping is easy to implement and understand the regularization method.

3.2.2.4. Dropout

Dropout is an effective method to solve the overfitting problem in neural networks. What is generally attempted to be done in this technique is to integrate the base network and its derivatives during the training phase [39]. To achieve this combination of network, randomly chosen neurons in the hidden layer are deleted during the feedforward phase. In other words, the input is processed with the modified network. Then backpropagation is applied to the base network, but only the weights of non-deleted neurons are updated. The name of the 'dropout' comes from this temporary neuron deletion. The selection of neurons to be deleted depends on a predetermined and fixed probability *p* where usually p set to 0,5. Dropout method is useful in the training phase. At test time, a single neural network is used which is the combination of the base and its derivatives.

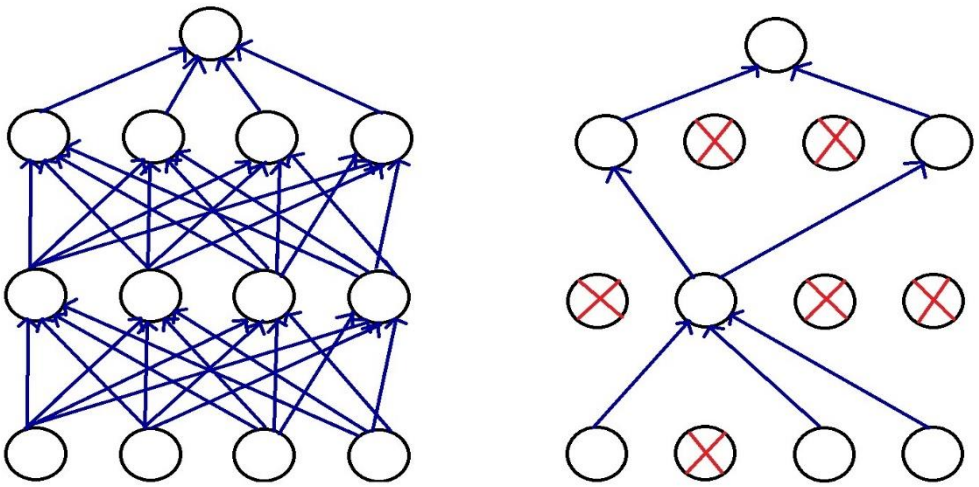

Figure 3.2. Dropout Neural Net Model. (a) A standard neural net, (b) After applying dropout

3.2.3. Optimization for training deep models

3.2.3.1. Challenges in deep neural networks optimization

Optimization is a challenging task for deep neural networks. Generally, optimization problems in machine learning algorithms are carefully designed to be convex objective function. This avoids the complexity of the optimization problem. However, for deep neural network, objective functions are non-convex. Therefore, these are more complex and hard to solve according to convex functions. The main goal of optimization algorithms is finding the global minimum of the objective function. In other words, it is to try to make the cost function equal or close to zero. But, to achieve these, many challenging tasks should be solved.

One of the most common problems is the increase of cost even with minimal steps during the training, also this is known as ill-condition. The problem is learning rate becomes ineffective against growing and big gradients. This situation also slows the training phase. Ill-condition can be seen at both non-convex and convex problems. Newton's method provides a good solution for convex objective functions. On the other hand, Newton's method needs significant modifications to apply to the non-convex objective function.

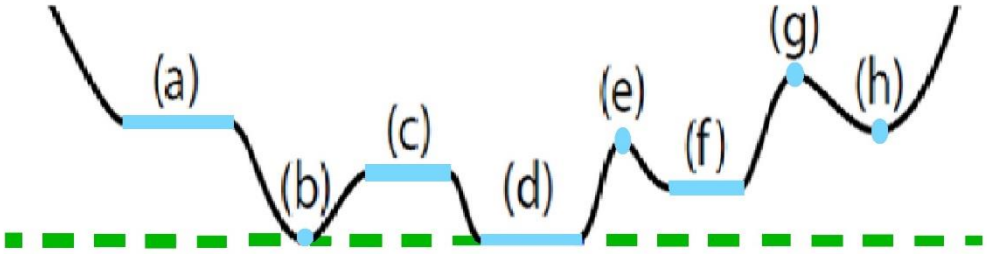

Figure 3.3. Example critical points of a non-convex function. (a,c) Plateaus. (b,d) Global minima. (e,g) Local maxima. (f,h) Local minima.

Another challenging task is objective functions have lots of local minimum points. While any local minimum is guaranteed to be a global minimum for convex problems, but this case is not valid for non-convex tasks. Existing learning algorithms guaranteed to converge to any critical points that have zero gradients [44,45,46,47]. If the learning algorithm stops at a high-cost local minimum, then the performance of the trained network will be lower than desired. However, deep neural networks can produce sufficient results at

low-cost local minimum points [39]. There is not only a local minimum but also saddle points, plateaus, local maximums known as a critical point as illustrated in figure 3.3.

Generally, researchers solve these problems by using empirical methods. Common techniques are random weight initialization, updating weights according to local descents and changing the initialization method if the cost does not decrease sufficient enough [36]. Besides, redesigning network model and building its non-linearity with ReLU increases performance. Despite all these, there is not any silver bullet to guarantee to converge global minimum at non-convex objective functions.

3.2.3.2. Stochastic gradient descent

For many learning algorithms such as linear regression or neural networks, the way of deriving the algorithm was by coming up with a cost function or coming up with an optimization objective. Using an algorithm like gradient descent to minimize the cost function when the training set is large becomes a computationally very expensive procedure. In general, the gradient descent computes all the gradients of the inputs separately and then averages them. In each iteration, basic gradient descent will do this gradient computation. Therefore, it is a time-consuming procedure and for big datasets, we need hardware to store all the samples. The stochastic gradient descent is the modified version of basic gradient descent that allows us to train our learning algorithms with bigger datasets. In stochastic gradient descent, sample sets, i.e $X_1, X_2, X_3, \ldots X_m$, selected inside a randomly shuffled dataset is used to perform each iteration. Assume that we have a cost function like

$$C(w_i, b_i) = \frac{1}{2\mathrm{n}} \sum \|y(x) - a\|^2 \tag{3.8}$$

In equation 3.8, y(x) is equal to the desired output, ‘a’ is representing the produced output and n is the number of samples. To calculate the gradient of the cost function according to the standard gradient descent algorithm, we need to compute all the inputs gradient individually and then average them. However, for the stochastic gradient descent, the formula returns to equation 3.8 where m is number of samples selected into randomly shuffled training set also called mini-batch

$$\nabla C \approx \frac{1}{m}\sum_{i=1}^{m} \nabla C_{X_i} \tag{3.9}$$

According to equation 3.9, update process becomes

$$w_k \longrightarrow w_k^{'} = w_k - \frac{\mu}{m}\sum_{j} \frac{\partial C_{X_i}}{\partial w_k} \tag{3.10}$$

$$b_l \longrightarrow b_l^{'} = b_l - \frac{\mu}{m}\sum_{j} \frac{\partial C_{X_i}}{\partial b_l} \tag{3.11}$$

w_k and b_l denote the weights and biases in our neural network.3.83.8

3.2.3.3. Parameter initialization

The training algorithms used in deep neural networks are usually iterative. Therefore, users need to determine the initial point for starting the training. In addition, the success of training algorithms depends on the choice of initialization. Even if algorithms are successful, the starting point of the iterative learning algorithms determines the cost of computationally.

Strategies for initialization applied these days are simple and heuristic. As neural network optimization is still a problem to be solved, it is difficult to produce advanced strategies. breaking of symmetry is the only well-known feature. In other words, the initial parameters are given in such a way that they differ between units. Because it's known that, if two hidden units in the same layer have the same initial parameters and their activation functions are connected to the same inputs, then learning algorithms will find the same cost for each unit and update weights of these units with the same way. Therefore, different initial parameters must be assigned to such units to avoid this situation.

In general, a Gaussian or uniform distribution is used to make the difference in weights of a model. Moreover, the generalization ability of the network and optimization procedure depends on the scale of the initial distribution. For breaking symmetry and avoiding redundant units, large initial weights should be chosen [39].

3.2.3.4. Adam optimizer

Adam optimizer is a kind of stochastic gradient descent learning methodology.

Classical stochastic gradient descent algorithm has only one parameter (learning rate) to update all the weights and it is constant during the training. The main difference between these two algorithms is Adam optimizer update each parameter with an individual learning rate. The algorithm takes into account the first and second momentums of the gradient to compute these learning rates of each weight [48]. Because of that, the name of this method comes from this approach also known as adaptive momentum estimation.

Mean and uncentered variance are the first and second moment.

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \tag{3.12}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2)\, g_t^2 \tag{3.13}$$

Where m and v are the moving averages and these values updated each iteration, g is mini-batch gradient and beta terms ($\beta_1 \; and \; \beta_2$) are hyperparameter of the optimizer which their default values are 0.9 and 0.999 respectively. If the initial value of m_0 and v_0 set to zero, therefore some bias correction should be used to prevent the results from moving quickly to zero. Finally, with bias correction, the estimator equations 3.12 and

$$\alpha(\lambda) = \varepsilon(\lambda) \tag{1.1}$$

becomes;

$$\widehat{m}_t = \frac{m_t}{1 - \beta_1^t} \tag{3.14}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t} \tag{3.15}$$

And update rule of Adam optimizer is shown at equation 3.16 where w is weights of network, η is step size and ϵ is very small number for prevent equation to be equal undefined.

$$w_t = w_{t-1} - \eta \frac{\widehat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \tag{3.16}$$

3.2.4. Convolutional neural network

Convolutional neural network is another type of neural network that can process data like has grid-like topology. Example of this data are time-series, that can be thought as a 1-D grid taking samples at specific intervals, and image data, as a 2-D grid of pixels. The name of the Convolutional neural network (CNN) comes from convolutional operation applied to at least one of the layers in the neural network. The convolutional process is a linear operation based on matrix multiplication.

David Hubel and Torsten Wiesel's study on the mammalian visual system is the inspiration point for the convolutional neural network [49]. In this study on monkeys, it was observed that some neurons in the early visual system react strongly to specific light patterns. This finding has led to the idea of using a combination of neural network and convolutional operation that can extract patterns from images.

LeChun's LeNet-5 network for recognizing handwritten digits is the first application of the convolutional neural networks [39]. After this study, the popularity of CNN's increased day by day. Convolutional layer and pooling layer are the layer types that differentiate CNN's from other neural networks.

3.2.4.1. Convolution layer

The most general definition of convolution is the mathematical operation between two real-valued functions. It defines the integration of the product of two functions which one of them reversed and shifted.

$$s(t) = \int f(a)\, g(t-a)\, da \tag{3.17}$$

Also another notation for the convolution using with an asterisk:

$$s(t) = (f * g)(t) \tag{3.18}$$

The result of this operation shows the amount of overlap between functions. In convolutional neural network terminology, the **input** represents the first function (f), the **kernel** represents the second function (g), and the **feature map** represents the output function (s). In CNN's, the input and the kernel have more than one dimension. For example if our input data is 2-D image data J and the kernel K is preferred as a 2-D array then the

formulation become :

$$S(i,j) = (J * K)(i,j) = \sum_m \sum_n J(i-m, i-n)\, K(m,n) \qquad (3.19)$$

Many machine learning libraries use cross-correlation instead of convolution and the only difference is cross-correlation do the same operation without flipping.

$$S(i,j) = (J * K)(i,j) = \sum_m \sum_n J(i+m, i+n)\, K(m,n) \qquad (3.20)$$

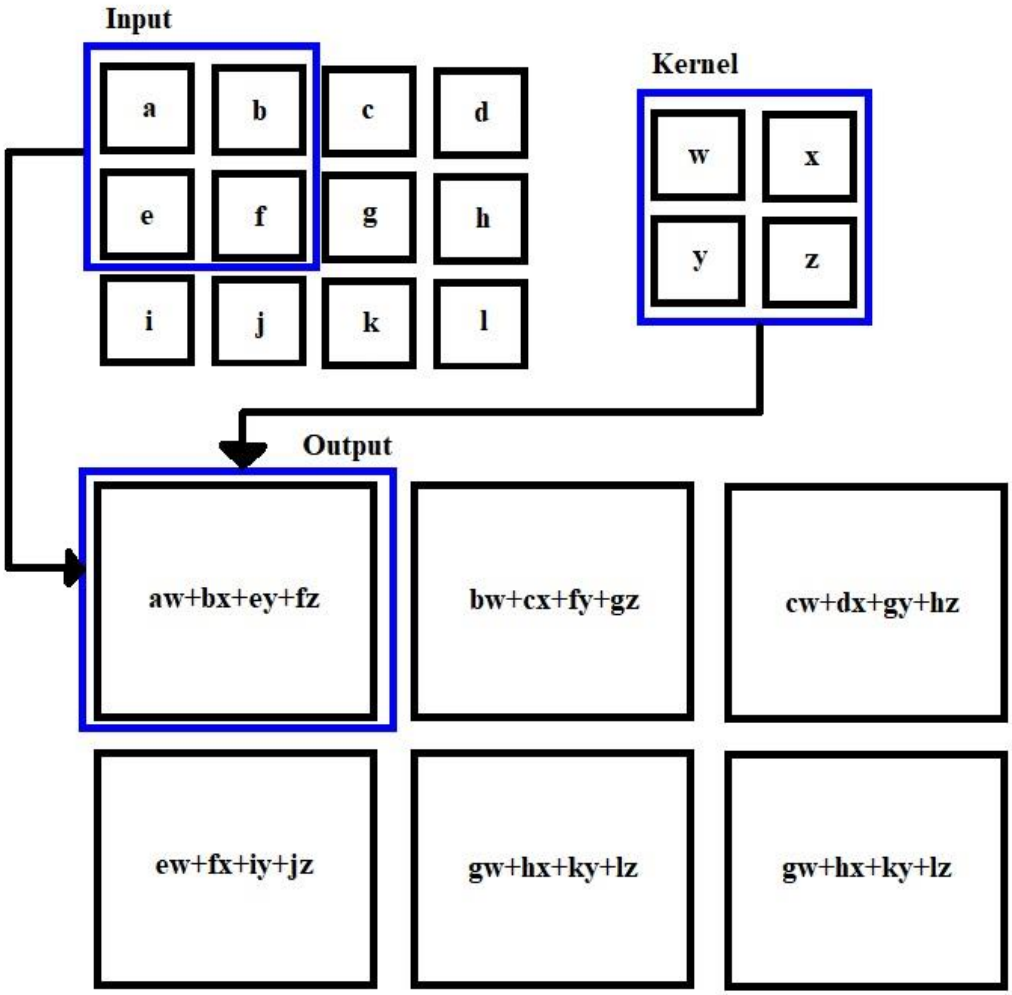

Figure 3.4. An Example of 2-D convolution without flipping [39].

The convolutional layer has two main properties that differentiate it from other neural networks. These are sparse interactions and parameter sharing

The conventional neural network layers attempt to establish a connection with the output unit by multiplying each input with certain parameters. That shows every input unit has a link on every output unit. But CNN has sparse interaction and this is provided by selecting small kernel sizes. Kernels in the convolutional layer detect only meaningful features like edges. This means only some local regions interact with output units. This provides decreasing the parameter size and reduces the memory size to store parameters.

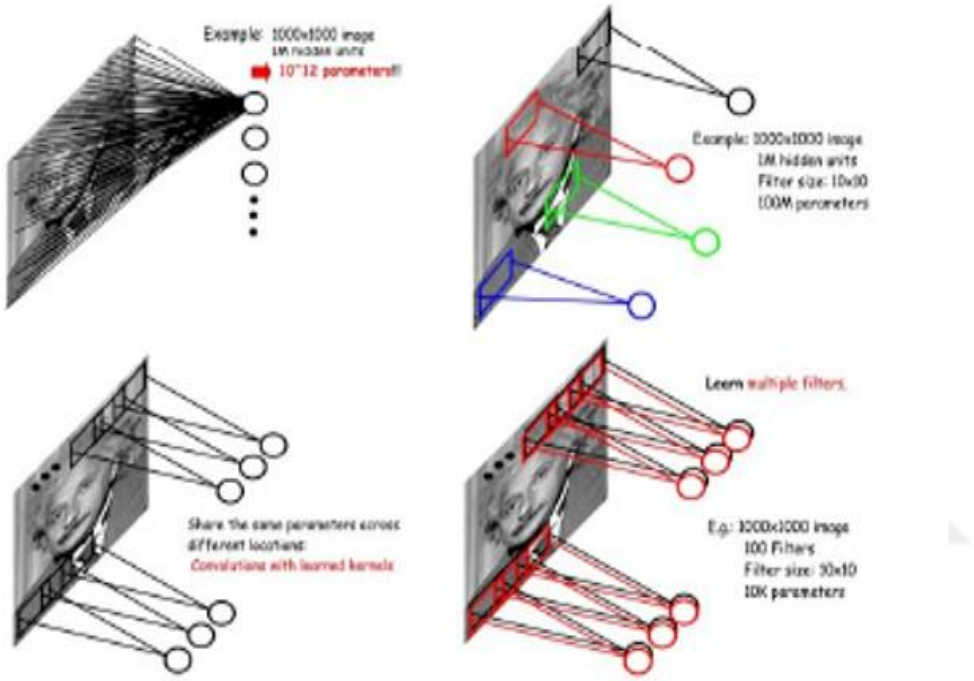

Figure 3.5 Illustrations of sparse representation and parameter sharing principles.

Parameter sharing is another idea that improves the efficiency of CNN. Traditional neural networks each weighs only using once to forward data to the output unit. On CNN, however, each kernel used every location on the input data. That means, instead of learning individual parameters for each location, learning only one set for all location. Convolutional operation is more efficient than dense matrix multiplication according to memory requirements and statistical efficiency.

3.2.4.2. Pooling layer

Another architectural difference that separates convolutional neural networks from other neural networks is the pooling layer. Pooling operation in this layer modifies the output of the previous layer using statistical methods such as selecting the maximum value or average the values of corresponding data. The common advantage of the pooling operation

is that it makes networks invariant to small translations of the input data. Pooling operations focus on which features come out of the input, rather than the exact location of the features and the type of pooling that determines which properties are passed on to the next layer.

Another advantage of the pooling layer is that it downsamples input data, which is usually a feature map. Window size and stride number are the hyperparameters of this layer. The values of the hyperparameters determine the level of downsampling. This process also reduces the memory size required for training a convolutional neural network.

There are many popular pooling operations applied today. Some of them are max pooling, the average of a rectangular neighborhood, the L2 norm of a rectangular neighborhood, and a weighted average based on the distance from the central pixel. Figure 3.6 shows an example of max pooling.

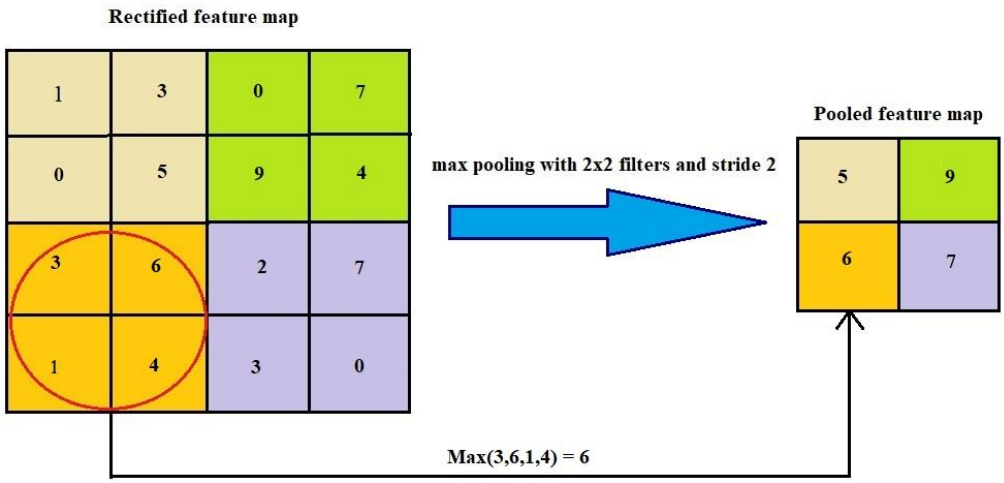

Figure 3.6. An Example of max pooling operation

3.3. Transfer Learning

Machine learning and deep learning methods work according to the feature space and feature distribution. If training data and test data has the same feature space and the same distribution, these methods work very well. But newly collected data has different from an available feature distribution. We have to adjust static models with newly collected data. To obtain new data can be expensive, time-consuming, or impossible. In such a case, transfer learning is beneficial methods [50].

The main purpose of transfer learning is to create a framework for a new feature

distribution data using accumulated data previously to solve different feature distribution problems faster and more effectively [51]. Transfer learning focuses on classification and labeling problems. Learning performance for classification problems increases in direct proportion to the success of knowledge transfer [50].

Transfer learning is a popular approach in deep learning areas where pre-trained deep neural network models used for a starting point. Training a deep neural network and finding an optimum value of the weight of the layers is time-consuming, even with modern hardware. Transfer learning works well if the pre-trained network trained with a large dataset and generalized the dataset sufficient enough. The first layers in the convolutional neural network extract global features, and deeper layers extract domain-specific information. Therefore, it is important at which level and in which domain the layers to be transferred extract features. The general approach is to use pre-trained neural networks as a feature extractor and update their weights with a new dataset. Transfer learning is a useful methodology in case of having a small dataset for training a new network.

3.4. AlexNet

With "ImageNet Classification with Deep Convolutional Neural Networks" publication in 2012, a new approach emerged in the field of computer vision [52]. AlexNet (its name comes from the first author of the paper, Alex Krizhevsky) proved the success of CNN's by winning the challenge of ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Before this achievement, CNN's performance is almost the same as typical feedforward neural networks of the same depth and size. Furthermore, there was no experience in training and hardware optimization of high-resolution images with CNNs. This paper succeeded a solution to all these challenges.

ImageNet dataset is used as dataset in this paper. This dataset consists of about 15 million high-resolution images collected from the web, which has 22,000 categories. Labelling was done by people using Amazon's Mechanical Turk crowd-sourcing tool. In the ILSVRC-2012 competition, a subset of ImageNet was used, with 1000 categories and 1000 images in each category. In total, this database contains approximately 1.2 million training pictures, 50,000 validation images and 150,000 test images.

As the pre-process, the dataset used has been down-sampled to 256 x 256, which is a fixed resolution for use in AlexNet as it has variable resolutions. While performing this process, re-scaling and cropping operations were applied.

AlexNet's architecture consists of eight layers with weights. 5 of these layers are convolutional layers and 3 are fully-connected layers. The first convolutional layer passes the 224x224x3 image through 96 filters of 11x11x3 with a stride of 4 pixels. The second convolutional layer takes the output of the first layer as input and filters with 256 kernels of size $5 \times 5 \times 48$. The other convolutional layers are connected to each other without any max-pooling operation. The third convolutional layer has 384 kernels of size $3 \times 3 \times 256$ connected to the fourth convolutional layer which has 384 kernels of size $3 \times 3 \times 192$ and the fifth convolutional layer has 256 kernels of size $3 \times 3 \times 192$. The fully-connected layers have 4096 neurons each.

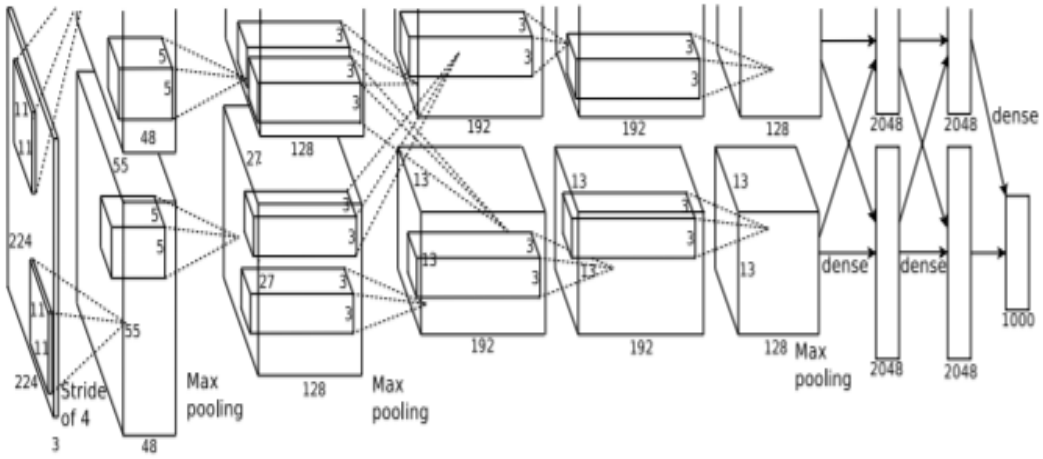

Figure 3.7. An illustration of the architecture of AlexNet [52].

It is not the architecture that makes AlexNet special. The main contribution of AlexNet comes from some features that have not been used before. AlexNet uses ReLU instead of the tanh function. They proved that ReLU is six times faster than standard tanh function. This modification reduces the training time of the network. In addition, multiple GPUs were used during AlexNet's training. In the model, half of the neurons are processed in one GPU, while the other half is processed in the other GPU. This contribution has pioneered the training of larger models and also enables the training of models in a shorter time. Also, another contribution is overlapping pooling. The paper shows that, using overlapping pooling reduces the error rate about 0.5%.

AlexNet has 60 million parameters should learn and update. This huge number of parameter size causes an overfitting problem at training phase. To overcome this problem, data augmentation and dropout methodologies applied on the network training. In the data

augmentation method, image conversion and horizontal reflections were used. In this way, they increased the amount of data in the data set. In addition, they were applied PCA for changing the density of RGB channels. Thus, they showed that they reduced the error rate by 1%.

3.5. DMR Dataset

Database for Mastology Research (DMR) is a database prepared by L. F. Silva and colleagues for use in the diagnosis of breast cancer by infrared imaging [53]. Many of the studies in this area use different data sets, which are collected by different protocols, and therefore it is difficult to compare the studies with each other. The motivation for this study comes from this problem. The authors of this study have created a database with no commercial purpose, open to all, and defined protocols. The images of the DMR set belong to the patients of Antonia Pedro University Hospital. In the data, there are images of patients and healthy people.

Figure 3.8. User Interface of DMR-IR database.

Based on the studies on this subject, they have established two main protocols. These are static and dynamic protocols. In the static protocol, they wait for approximately ten

minutes for patients skin temperature becomes stable. After stabilisation, they took five images from different angles (one frontal, two laterals of his left at 45º and 90º, and two laterals of his right at 45° and 90°) from patients. For the dynamic protocol, they applied a cooling operation to the breast region using alcohol or electrical fan. This protocol consists of images taken for approximately 5 minutes during the skin temperature recovery phase. The database contains twenty photographs for a single individual taken at this stage.

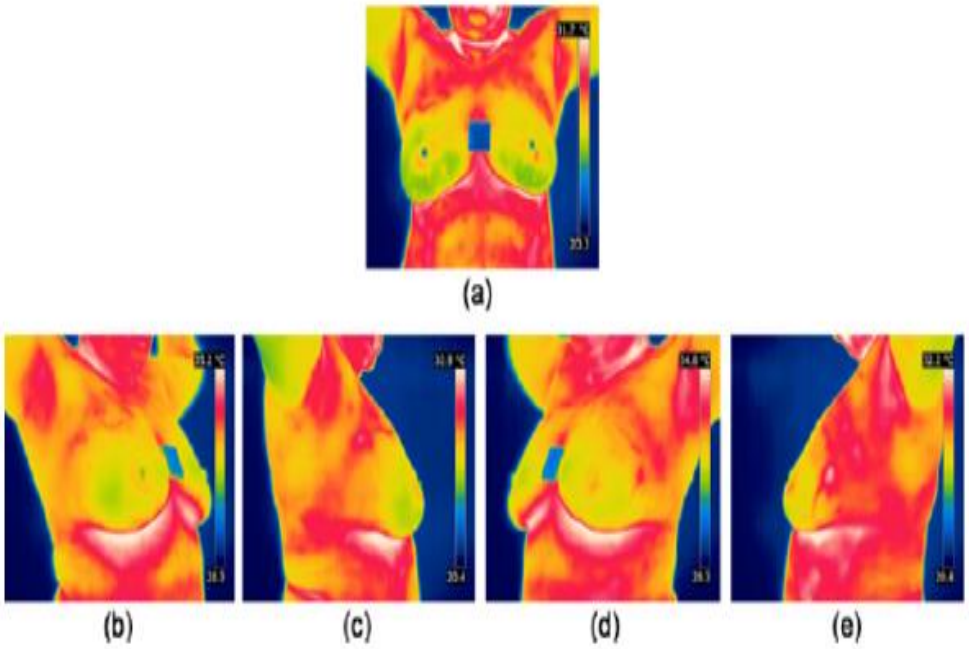

Figure 3.9. Patent Positions of the images. One frontal (a), two laterals of his left at 45º (d) and 90º (e), and two laterals of his right at 45° (b) and 90° (c).

FLIR SC-620 Thermal camera was used for acquisition of thermograms with a spatial resolution of 640 x 480 pixels. In our work, we use 147 thermograms of healthy individuals, and 34 thermograms belong to sick individuals. These are all images that we can download from the database.

4. EXPERIMENTS AND RESULTS

In this section, we explain our experimental setup and results on breast cancer detection from IR images. To comparison, we design and test 4 different CNN based on AlexNet. The performance of the designed nets was evaluated on a benchmarking dataset considering accuracy, precision, recall, F1 measure, and Matthews Correlation coefficient. Also, we compare our result with other studies that use DMR database.

4.1. Data Preparation

In this study, we used data of 181 people in the DMR image library for training and verification. All images in the dataset have a spatial resolution of 640x480 pixels. Both RGB image and thermal matrix versions of the data in the database are available. However, the camera's logo, some unwanted shapes and colour map are on the RGB images. Therefore, it was deemed appropriate to use data containing thermal matrices because it only contains heat values. A total of 181 images belong to 34 patients and 147 healthy individuals.

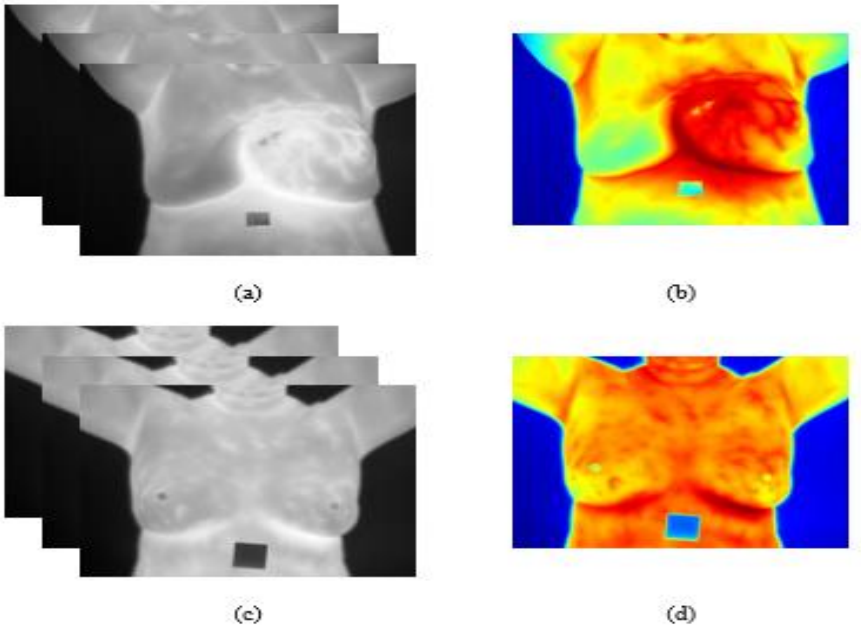

Figure 4.1 Example of prepared images. In (a) shows duplicated grayscale image belongs to a sick patient, (b) represents RGB jet image belongs to a sick patient, (c) shows duplicated grayscale image belongs to a healthy patient and (d) is an example of RGB jet image belongs to a healthy patient.

We used two different input type for training our nets. Pre-processing was applied to thermal matrices separately to produce these inputs. Firstly we generated grayscale images from the thermal matrixes. All values of the thermal matrix were individually scaled to between 0 to 255 to produce grayscale images. In general, the values of the thermal matrix range from 20 to 36 degrees Celsius. Small temperature changes were made more observable by using this scaling method. The input layer of AlexNet designed for three-dimensional coloured images with the size of 227x227 pixel. To fit grayscale images to AlexNet input layer, we duplicated the grayscale image to all three channels and resized to 227x227 (Figure a,c). After all these operations, we had three-dimensional images that all its channel contains grayscale values of thermal matrixes with the same size of the input layer. The second input type is a coloured image (Figure 4.1 b,d). We converted thermal matrixes to RGB image using Matlab Jet colour map. Figure 4.2. is shows the color scheme of the jet colormap.

Figure 4.2. The color scheme of the jet colormap on Matlab.

DMR dataset contains a small amount of data to train a convolutional neural network. Data augmentation techniques are the easiest way to overcome the overfitting problem and to increase the amount of data. We applied two common forms of data augmentation at the training phase, these forms also preserve labels of data. These two forms are translation and reflection. The translation was performed to avoid dependency of the localisation using randomly selected pixel range between -30 to 30 and vertical reflection was applied on only vertical axis because the tumor can be seen on the other breast.

4.2. Network Architectures

In order to classify images and examine the effect of layers, we designed four different Convolutional Neural Networks (CNNs) based on transfer learning. The fact that the DMR dataset is small and the need for a large dataset to achieve good performances in CNN training led us to the transfer learning methodology in this study. The network to which we transfer their layers must be either trained in a similar domain or have a good generalisation feature. In this study, since we could not find a pre-trained network in the same domain, we

chose AlexNet as the base model that performed well a large and demanding image classification task, such as the ImageNet 1000 class image classification competition. All the architectures of models are shown in Fig. 4.3.

Net1 is the first model we designed. While designing this model, we wanted to use AlexNet as both feature extractor and classifier. For this reason, in this model, all the other layers of AlexNet except the last two layers have been transferred. The three layers we extracted are a fully connected layer with 1000 neuron and a softmax layer. These layers are designed for 1000 class classification. However, the classification in our case is divided into two to be patient and healthy. So we added a fully connected layer with two neurons and a softmax layer. During the training, we only trained the newly added layers.

The second model is the Net2. We aim to use AlexNet as a feature extractor. Therefore we only transferred convolutional layers of AlexNet and extracted all the fully connected layers. We added newly four fully connected layers with 4096-4096-1024-2 neurons.

Net3 is the third model. In Net1 and Net2, the convolutional layers of AlexNet are completely taken. While the first convolutional layers in CNNs extracts general features, the last layers extract domain-specific properties. Therefore, the fifth convolutional layer of AlexNet was removed in this model. The effect of the last convolutional layer of AlexNet on our case was measured. Net3 has four fully connected layers with 4096-4096-1024-2 neurons like Net2.

Net4 is the last model in our study. In this model, three fully connected layers with 4096-1024-2 neurons are used. Convolutional layer number and design are the same as Net3. In this model, the effect of fully connected layer number is tried to be measured.

Figure 4.3. The graphical pipeline of the proposed CNN models. The transferred layers are shown in blue and green ones are newly added layers.

In the training phase, we set the learning rate of the newly added layers to 20 according to MATLAB Deep learning toolbox and the learning rate of the transferred layers to 0.0001. This allows us to update the transferred layers to a minimum while accelerating the learning of new additions. Adam optimizer was used for training of all designed networks. Two different methods were used to prevent overfitting. The first of these is data augmentation with random vertical reflection and translation into each epoch. The other one is the dropout application with a rate of %50. We also use the early-stopping approach to stop training of CNN. The training was stopped when if consecutively five iterations had training accuracy was 100%. We also used leave-one-out cross-validation to ensure every image is used in train and test sets.

4.3. Empirical Results

The designed networks were evaluated using leave-one-out cross-validation. 181 images, 34 patients and 147 healthy, were used for testing and training. 180 images were used to train the networks, and an image that is not in the training set was used for testing.

This procedure was repeated until each picture was tested. This means that they have been trained 181 times to achieve the result of a single network.

Also, we produced balanced dataset using RGB images. We increased the amount of sick patients images with data augmentation techniques. We applied random translation and reflection to 34 sick patients images to generate 135 number of image. As a result, we obtained 135 patient images and 147 healthy patient images in the data set.

The results were calculated according to five metrics. All these metrics are calculated by considering the positive (P) as sick and the negative (N) as healthy. These are considering accuracy (Acc), precision, recall, F1 measure (F1), and Matthews Correlation coefficient (MCC). Using the following equations can calculate these metrics:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{3.21}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3.22}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.23}$$

$$\text{F1 Measure} = \frac{2 * (Precision * Recall)}{\text{Precision} + \text{Recall}} \tag{3.24}$$

$$\text{MCC} = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP) * (TP + FN) * (TN + FN) * (TN + FP)}} \tag{3.25}$$

We tested the CNN models with two different input types. One of these input types is duplicated gray image produced from the thermal matrix, and the other is RGB images whose thermal matrix is converted to the jet color map.

We tested Net1 with an augmented dataset and non-augmented dataset. In addition, we tested the effect of input type. Therefore, two different image types are given as an input which is a RGB image and duplicated gray image (Table 4.1-4.2) to Net1. Result of this test show that to train with non-augmented data and to give RGB image as an input achieve %86,74 accuracy, train with augmented data and giving RGB image as an input achieve %87,29 accuracy, train with non-augmented data and giving duplicated gray image as an input achieve %85 accuracy, train with augmented data and giving duplicated gray image as

an input achieve %85,64 accuracy. According to Table 1 and 2 , the augmentation method slightly increases the performance of the network.

Moreover, if we compare the architectures of the network, we can see the effect of layers at the results (Table 4.2). The number of fully connected layer is essential according to performance metrics. Net3-Net4 comparisons are the right choices to observe this effect because Net3-Net4 has the same feature extractor part that transferred from AlexNet. Table 3 and Table 4 shows that 4 fully connected layers with 4096-4096-1024-2 neurons perform well compared to other architectures. Therefore, Net2 has the same fully connected layer design. Another point is the performance of Net3 and Net4 are lower than Net1 and Net2. This shows that the last convolutional layer plays a vital role in the classification of breast cancer from IR images and other convolutional layers feature vectors are not sufficient enough to achieve a suitable generalization of this problem.

Furthermore, giving unbalanced RGB image dataset as input and to train Net2 with augmentation performs the best performance among the other CNN models with %89,5 accuracy (Table 4.3-4.4). Net2 uses AlexNet as feature extractor and we trained fully connected layers according to features that come from pre-trained convolutional layers. All these additions improve its performance because AlexNet trained with RGB images and its parameters were updated according to this input type. Net1 architecture also has pre-trained fully connected layers, and this is the main difference between Net1 from Net2. The weights of the transferred fully connected layers were trained and updated according to ImageNet dataset. Therefore, the performance of Net1 remained low compared to Net2.

At last, we train our designed networks with balanced RGB dataset (Table 4.5). Net2 has the highest score compared to other networks. Results show that training the networks with a balanced dataset increased the performance significantly. Our best result with the unbalanced dataset is %89,5, and with the balanced dataset, Net2 achieves %94,3 accuracy.

We also compared our results with other studies using DMR dataset to detect breast cancer (Table 6). In most of the studies, statistical features were used to extract the features, also SVM, ANN and CNN were used for classification [6,11,18,19,20]. According to the comparison table, convolutional neural network methodology performs better than other traditional machine learning algorithms. These results support why convolutional neural network algorithms are currently very popular. One study performed better than our work [6]. That's because their dataset is more balanced than ours. But it is not fair to make an empirical comparison of these methods since each used a different number of images with an independent experimental setup.

Table 4.1. Result of Net1 trained with the RGB image

Net1	Accuracy	Precision	Recall	F1	MCC
With Aug.	0.8729	0.7037	0.5588	0.623	0.553
Without Aug.	0.8674	0.6786	0.5588	0.6129	0.5375

Table 4.2. Result of Net1 trained with duplicated gray image

Net1	Accuracy	Precision	Recall	F1	MCC
With Aug.	0.8564	0.6429	0.5294	0.5806	0.4984
Without Aug.	0.85	0.6538	0.5	0.5667	0.4886

Table 4.3. Result of CNN models trained with RGB image and augmentation

CNN model	Accuracy	Precision	Recall	F1	MCC
Net1	0.8729	0.7037	0.5588	0.623	0.553
Net2	0.895	0.7143	0.7353	0.7246	0.6599
Net3	0.8177	0.5185	0.4118	0.459	0.3545
Net4	0.7956	0.4286	0.2647	0.3273	0.2233

Table 4.4. Result of CNN models trained with duplicated gray image and augmentation

CNN model	Accuracy	Precision	Recall	F1	MCC
Net1	0.8564	0.6429	0.5294	0.5806	0.4984
Net2	0.8287	0.5429	0.5588	0.5507	0.445
Net3	0.8398	0.619	0.3824	0.4727	0.4
Net4	0.8287	0.5652	0.3824	0.4561	0.3686

Table 4.5. Result of CNN models trained with a balanced RGB image dataset

CNN model	Accuracy	Precision	Recall	F1	MCC
Net2	0.943463	0.947761	0.933824	0.940741	0.886781
Net3	0.915194	0.91791	0.904412	0.911111	0.830122

Table 4.6. Comparison with other studies on detection of breast cancer.

Paper	Features / Classifiers	Acquisition Protocol	Number of Images	Accuracy
Gaber et al. [32]	Gabor Coefficients / SVM RBF	Static	63 (29 healthy and 34 malignant)	88.41%
Lessa and Marengoni [54]	Statistical Features / Artificial Neural Networks	Static	94 (48 normal and 46 abnormal)	85%
Borchartt et al [55]	Statistical Features / Genetic Algorithm	Static	51 (14 abnormal and 37 normal)	88.2%
Acharya et al [27]	Statistical Texture Features / SVM RBF	Static	50 (25 normal and 25 abnormal)	88.10%
Baffa and Lattari [33]	CNN Features / CNN	Static	300 images (126 abnormal and 174 normal)	98%
Our Work	CNN Features / CNN using transfer learning	Static	282 (135 sick and 147 healthy)	94,3%

5. CONCLUSION AND DISCUSSION

Convolutional neural networks have proven to perform better in many areas than other conventional methods. With this study, infrared image analysis was added to these fields. The lack of data set is a well-known concept that is a problem for the training of a CNN. Breast cancer diagnosis from IR images is a challenging problem in this sense. Obtaining a large amount of data set of labelled IR images and the number of data of diseased people is less than the number of data of healthy people poses difficulties for CNN training. However, the small amount of data in the dataset can be overcome by using a pre-trained network, namely the transfer learning approach. Moreover, data augmentation methodology plays a vital role in proper CNN training when working with an unbalanced dataset.

In this study, the diagnosis of breast cancer was studied based on the information provided by IR images of the chest region. The properties of a CNN network that can be used to solve this problem have been extracted. In doing so, different CNN models have been created. While creating these models, transfer learning methodology was used. AlexNet has been identified as the basic model, and from this model the previously trained layers have been transferred to new networks. In order to examine the effect of the transferred layers, layers from different depth were transferred to different networks. Two different input types were produced to examine the effect of the input types as well as the effect of the layers. In addition, the balanced and unbalanced dataset was created, and its contribution to the performance was examined.

As a conclusion, convolutional neural networks have obtained more successful results than other conventional methods in the diagnosis of breast cancer from IR images. The type and number of layers to be transferred are essential. In this study, the best results were obtained only if AlexNet was taken as a feature extractor. Moreover, the results show that AlexNet the feature vectors of the fifth convolutional layer is necessary for improving the performance of classification. In the experiments, it is observed that training fully connected layers according to domain increases performance. The matching of the input type with the input characteristics to which the transferred network is trained affects the performance of the network. In addition, the fact that the classes that make up the dataset are the same number or the numbers close to each other allows the network to generalize better.

In this study, the transferred layers are only taken from AlexNet. Other networks such as VGGNet, GoogleNet, etc. which have proven themselves in the field of computer vision

can be tried and compared the results. In addition, hyperparameter optimization algorithms can be used to measure the effect on performance. The problem discussed in this study is a classification problem. However, the success of CNN's is also obvious in segmentation problems. In this regard, studies can be made to locate the diseased region. These issues can be considered as working areas for the future.

REFERENCES

[1] J. S. Bertram, “The molecular biology of cancer,” Molecular Aspects of Medicine, vol. 21, no. 6, pp. 167–223, 2000.

[2] D. Hanahan and R. A. Weinberg, “The Hallmarks of Cancer,” Cell, vol. 100, no. 1, pp. 57–70, 2000.

[3] P. A. Hall, “Introduction to the Cellular and Molecular Biology of Cancer. 3rd edn. L. M. Franks and N. M. Teich. Oxford University Press, Oxford, 1997. No. of pages: 468. Price: £55.00 (Hardback). ISBN: 0 19 854854 0,” The Journal of Pathology, vol. 186, no. 2, pp. 222–222, 1998.

[4] D. Saslow, D. Solomon, H. W. Lawson, M. Killackey, S. L. Kulasingam, J. Cain, F. A. R. Garcia, A. T. Moriarty, A. G. Waxman, D. C. Wilbur, N. Wentzensen, L. S. Downs, M. Spitzer, A.-B. Moscicki, E. L. Franco, M. H. Stoler, M. Schiffman, P. E. Castle, and E. R. Myers, “American Cancer Society, American Society for Colposcopy and Cervical Pathology, and American Society for Clinical Pathology screening guidelines for the prevention and early detection of cervical cancer,” CA: A Cancer Journal for Clinicians, vol. 62, no. 3, pp. 147–172, 2012.

[5] “Release notice - Canadian Cancer Statistics 2015,” Health Promotion and Chronic Disease Prevention in Canada, pp. 20–20, 2015.

[6] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray, “Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012,” International Journal of Cancer, vol. 136, no. 5, Sep. 2014.

[7] T. Tarver, “Cancer Facts & Figures 2012. American Cancer Society (ACS),” Journal of Consumer Health On the Internet, vol. 16, no. 3, pp. 366–367, 2012.

[8] McTiernan, A., M. Gilligan, and C. Redmond, “Assessing individual risk for breast cancer: Risky business,” J Clin Epidemiol, Vol. 50, pp. 547-556, 1997.

[9] Physician Data Query, Screening for breast cancer. Available: http://www.nci.nih.gov/cancerinfo/pdq/screening/breast/healthprofessional/

[10] A. Migowski, ‘‘Early detection of breast cancer and the interpretation of results of survival studies,’’ Ciencia Saude Coletiva, vol. 20, no. 4, p. 1309, Apr. 2015.

[11] Griffiths C and Brock A. “Twentieth century mortality trends in England and Wales,” Health Statistics Quarterly 18, 5-17. Available: www.statistics.gov.uk/cci/article.asp?ID=1535&Pos=1&ColRank=1&Rank=1

[12] J. Seely and T. Alhassan, “Screening for breast cancer in 2018— what should we be doing today?,” Current Oncology, vol. 25, p. 115, 2018.

[13] P. A. Carney, D. L. Miglioretti, B. C. Yankaskas, K. Kerlikowske, R. Rosenberg, C. M. Rutter, B. M. Geller, L. A. Abraham, S. H. Taplin, M. Dignan, G. Cutter, and R. Ballard-Barbash, “Individual and Combined Effects of Age, Breast Density, and Hormone Replacement Therapy Use on the Accuracy of Screening Mammography,” Annals of Internal Medicine, vol. 138, no. 3, p. 168, Apr. 2003.

[14] M. Lobbes, M. Smidt, J. Houwers, V. Tjan-Heijnen, and J. Wildberger, “Contrast enhanced mammography: Techniques, current results, and potential indications,” Clinical Radiology, vol. 68, no. 9, pp. 935–944, 2013.

[15] E. J. Schneble, L. J. Graham, M. P. Shupe, F. L. Flynt, K. P. Banks, A. D. Kirkpatrick, A. Nissan, L. Henry, A. Stojadinovic, N. M. Shumway, I. Avital, G. E. Peoples, and R. F. Setlik, “Future Directions for the Early Detection of Recurrent Breast Cancer,” Journal of Cancer, vol. 5, no. 4, pp. 291–300, 2014.

[16] R. J. Hooley, L. M. Scoutt, and L. E. Philpotts, “Breast Ultrasonography: State of the Art,” Radiology, vol. 268, no. 3, pp. 642–659, 2013.

[17] Adams, F.; “The Genuine Works of Hippocrates,” Academic Medicine, vol. 14, no. 4, p. 279, 1939.

[18] A. Rogalski, “History of infrared detectors,” Opto-Electronics Review, vol. 20, no. 3, Jan. 2012.

[19] A. Piana and A. Sepper, “Contemporary Evaluation of Thermal Breast Screening,” Pan American Journal of Medical Thermology, vol. 1, no. 2, pp. 93–100, 2014.

[20] R. N. Lawson, “Implications of SurfaceTemperatures in the Diagnosis of Breast Cancer,” Can. Med. Assoc. J. 75, 1956.

[21] J. Gershen-Cohen, J. Haberman, E. E. Brueschke, “Medical thermography: A summary of current status”, Radio Clin North Am vol.3, pp.403-431, 1965

[22] H. J. Isard, W. Becker, R. Shilo, “Breast Thermography after Four Years and 10000 Studies”, Am. J. Roentgenol., pp.811, 1972

[23] H. Spitalier, D. Giraud, R. Amalric, “Does Infrared Thermography Truly Have a Role in Present-Day Cancer Management?”, Biomedical Thermology. Alan R. Liss, Inc., NY, pp. 269-278, 1982

[24] C. Gros, M. Gautherie, “Improved System for the Objective Evaluation of Breast Thermograms”, Biomedical Thermology, pp.897-905, 1982

[25] L. Thomassin, D. Giraud, “Detection of Subclinical Breast Cancers by Infrared Thermography”, Recent Advances in Medical Thermology, Plenum Press, New York, NY. Pp.575-579, 1984

[26] Y. R. Parisky, A. Sardy, “Efficacy of Computerized Infrared Imaging Analysis to Evaluate Mammographically Suspicious Lesions”, Am. J. Roentgenol, 180, pp.263, 2003

[27] U. R. Acharya, E. Y. K. Ng, J.-H. Tan, and S. V. Sree, “Thermography Based Breast Cancer Detection Using Texture Features and Support Vector Machine,” Journal of Medical Systems, vol. 36, no. 3, pp. 1503–1510, 2010.

[28] S. V. Francis and M. Sasikala, “Automatic detection of abnormal breast thermograms using asymmetry analysis of texture features,” Journal of Medical Engineering & Technology, vol. 37, no. 1, pp. 17–21, 2012.

[29] B. Krawczyk and G. Schaefer, “Breast Thermogram Analysis Using Classifier Ensembles and Image Symmetry Features,” IEEE Systems Journal, vol. 8, no. 3, pp. 921–928, 2014.

[30] M. C. Araújo, R. C. Lima, and R. M. D. Souza, “Interval symbolic feature extraction for thermography breast cancer detection,” Expert Systems with Applications, vol. 41, no. 15, pp. 6728–6737, 2014.

[31] B. Krawczyk and G. Schaefer, “Breast Thermogram Analysis Using Classifier Ensembles and Image Symmetry Features,” IEEE Systems Journal, vol. 8, no. 3, pp. 921–928, 2014.

[32] T. Gaber, G. Ismail, A. Anter, M. Soliman, M. Ali, N. Semary, A. E. Hassanien, and V. Snasel, “Thermogram breast cancer prediction approach based on Neutrosophic sets and fuzzy c-means algorithm,” 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2015.

[33] M. D. F. O. Baffa and L. G. Lattari, “Convolutional Neural Networks for Static and Dynamic Breast Infrared Imaging Classification,” 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), 2018.

[34] Y. Bengio, “Learning Deep Architectures for AI,” Foundations and Trends® in Machine Learning, vol. 2, no. 1, pp. 1–127, 2009.

[35] G. E. Hinton, S. Osindero, and Y.-W. Teh, “A Fast Learning Algorithm for Deep Belief Nets,” Neural Computation, vol. 18, no. 7, pp. 1527–1554, 2006.

[36] R. Vidal, J. Bruna, R. Giryes, and S. Soatto. (2017). “Mathematics of deep learning.” [Online]. Available: https://arxiv.org/abs/1712.04741

[37] V. Nair and G.E. Hinton, “Rectified Linear Units Improve Restricted Boltzmann Machines,” Proc. Int’l Conf. Machine Learning, 2010.

[38] K. He, X. Zhang, S. Ren, and J. Sun, “Identity Mappings in Deep Residual Networks,” Computer Vision – ECCV 2016 Lecture Notes in Computer Science, pp. 630–645, 2016.

[39] I. Goodfellow, Y. Bengio, and A. Courville, Deep learning. Cambridge, MA: MIT Press, 2017.

[40] P. Simard, D. Steinkraus, and J. Platt, “Best practices for convolutional neural networks applied to visual document analysis,” Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.

[41] D. Ciresan, U. Meier, and J. Schmidhuber, “Multi-column deep neural networks for image classification,” 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012.

[42] DC Cires¸an, U Meier, J Masci, LM Gambardella, and J Schmidhuber. “High-performance neural networks for visual object classification,” ArXiv e-prints, 2011.

[43] L. Prechelt, “Early Stopping - But When?,” Lecture Notes in Computer Science Neural Networks: Tricks of the Trade, pp. 55–69, 1998.

[44] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. Journal of Machine Learning Research, 11:19–60, 2010.

[45] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” Nature, vol. 323, no. 6088, pp. 533–536, 1986.

[46] Stephen J Wright and Jorge Nocedal. Numerical Optimization, volume 2. Springer New York, 1999.

[47] Y. Xu and W. Yin, “A Block Coordinate Descent Method for Regularized Multiconvex Optimization with Applications to Nonnegative Tensor Factorization and Completion,” SIAM Journal on Imaging Sciences, vol. 6, no. 3, pp. 1758–1789, 2013.

[48] D. P. Kingma and J. L. Ba, “Adam: A method for stochastic optimization,” in Proc. Int. Conf. Learn. Represent., pp. 1–41, 2015.

[49] D. H. Hubel and T. N. Wiesel, “Receptive fields and functional architecture of monkey striate cortex,” The Journal of Physiology, vol. 195, no. 1, pp. 215–243, Jan. 1968.

[50] S. J. Pan and Q. Yang, “A Survey on Transfer Learning,” IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345–1359, 2010.

[51] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, and G. Zhang, “Transfer learning using computational intelligence: A survey,” Knowledge-Based Systems, vol. 80, pp. 14–23, 2015.

[52] A. Krizhevsky, I. Sutskever, and G. Hinton, “Imagenet classification with deep convolutional neural networks,” in Proc. Adv. Neural Inf. Process. Syst., pp. 1106–1114, 2012.

[53] L. F. Silva, D. C. M. Saade, G. O. Sequeiros, A. C. Silva, A. C. Paiva, R. S. Bravo, and A. Conci, “A New Database for Breast Research with Infrared Image,” Journal of Medical Imaging and Health Informatics, vol. 4, no. 1, pp. 92–100, Jan. 2014.

[54] V. Lessa and M. Marengoni, “Applying Artificial Neural Network for the Classification of Breast Cancer Using Infrared Thermographic Images,” Computer Vision and Graphics Lecture Notes in Computer Science, pp. 429–438, 2016.

[55] T. B. Borchartt, R. Resmini, L. S. Motta, E. W. Clua, A. Conci, M. J. Viana, L. C. Santos, R. C. Lima, and A. Sanchez, “Combining approaches for early diagnosis of breast diseases using thermal imaging,” International Journal of Innovative Computing and Applications, vol. 4, no. 3/4, p. 163, 2012.