# FEATURE EXTRACTION AND MACHINE LEARNING TECHNIQUES FOR MELANOMA RECOGNITION BY USING DERMOSCOPY IMAGES

by

**MUHAMMAD FARID BIN OMAR**
**(1830612847)**

A thesis submitted in fulfillment of the requirements for the degree of
Master of Science in Mechatronic Engineering

**Faculty of Electrical Engineering Technology**
**UNIVERSITI MALAYSIA PERLIS**

2021

## ACKNOWLEDGEMENT

ALHAMDULILLAH, thank you Allah S.W.T for His grace for allowing me to make it to this point of my life and give me the energy and time to prepare the thesis final report of the project tasks successfully.

Firstly, I want to dedicate this award speech to my supervisor Dr. Muhammad Naufal bin Mansor of the Faculty of Engineering Technology at University Malaysia Perlis for giving guidance and perfect illumination to me from the very beginning since this research was given to me. Dr. Muhammad Naufal reliably enabled this paper to be my very own work and controlled me the correct way at whatever point he thought I required it. Without his assistance I would not achieve the last phase of my research.

I would also like to thank the Unimap Centre for Graduate Studies Staff for their help throughout the course of my research. Every call and e-mail I made regarding the problems I had are answered politely and with great patience from their side.

Finally, I must express my very profound gratitude to both my father Mr. Omar bin Othman and my mother Mrs. Norshamshinar binti Ghazali and also my siblings for putting up with support and encouragement to me. Besides that, they always leave their prayers for blessing me here. And they also helped me a lot spiritually and financially to produce this thesis. Once again, I wish to express my sincere appreciation and thanks to the supervisor, friends and parents involved in my thesis report

TABLE OF CONTENTS

# LIST OF TABLES

**LIST OF FIGURES**

## LIST OF ABBREVIATIONS

| | |
|---|---|
| ANOVA | Analysis of Varience |
| ANN | Artiificial Neural Network |
| CNN | Convolutional Neural Networks |
| FkNN | Fuzzy k-Nearest Neighbor |
| GLCM | Gray Level Co-Occurrence Matrix |
| kNN | k-Nearest Neighbor |
| LBP | Local Binary Pattern |
| LDA | Linear Discriminant Analysis |
| LDP | Local Directional Patterns |
| NN | Neural Network |
| SFTA | Segmentation-based Fractal Analysis |
| SVM | Support Vector Machines |

## LIST OF SYMBOLS

| | |
|---|---|
| Mk | Weighting Factors |
| $d_E(x, y)$ | Euclidean Distance on (x, y) Coordinate |
| $\upsilon_i$ | Number of Pixels |
| $\Psi$ | Small Neighborhood |
| $\Omega$ | Image Domain |
| $\lambda$ | Smoothness Constraint |
| $m_L$ | Local Mean |
| $\Phi_i$ | Vector Image |
| $\omega_k$ | Eigenvector |
| $\delta_k$ | Output Neuron |
| $\delta_j$ | Hidden Neuron |
| $\mu^2$ | Uniform Patterns |
| $v_{ij}$ | Weight between Input and Hidden Layer |
| $w_{jk}$ | Weight between Hidden and Output Layer |
| $Y_K$ | Output Layer Neuron |
| $Z_j$ | Hidden Layer Neuron |
| $n_{pq}$ | Normalized Central Moment |
| $\mu_{pq}$ | Central Moment |
| nt | Number of Threshold |

# Teknik Penyarian Sifat Dan Pembelajaran Mesin Untuk Pengecaman Melanoma Terhadap Gambar Dermoskopi

## ABSTRAK

Melanoma adalah sesuatu neoplasma atau dikenali sebagai kanser kulit yang mengancam nyawa kepada seseorang dengan hanya wujud di permukaan kulit sahaja. Melanoma ini boleh dikenalpasti dengan melihat kesan luka, berbentuk ganjil, berbentuk tidak sekata dan mempunyai pelbagai warna dalam kes-kes tertentu. Melanoma ini terlalu kecil dengan hanya bersaiz beberapa milimeter dalam diameter serta berbentuk seperti benign nevi dan tidak dapat dikesan dan dilihat dengan ujian mata kasar. Cara untuk mendiagnosis melanoma ini adalah dengan menggunakan dermoskopi sebagai alat. Pengesanan awal melanoma dan melakukan pembedahan serta merta dapat menyembuhkan pesakit. Bantuan teknologi seperti pengumpulan data dan pembelajaran tentang teknologi ini boleh meningkatkan ketepatan diagnosis. Sesetengah kajian telah dijalankan dengan menggunakan pengkalan data MIAS. Peruasan Berasas Penanda Analisis (SFTA), Skala Kelabu Sekejadian Matriks (GLCM), Penapis Gabor and Momen vektor Hu memberikan ciri pengembangan yang baik. Walau bagaimanapun, dengan gabungan ini dapat memberikan hasil yang lebih tinggi iaitu lebih daripada 90% empat pengelasan yang berbeza dengan pelbagai algoritma seperti Jiran Terdekat K (KNN), Fuzi Jiran Terdekat K (FkNN), Pembezalayan Lelurus Analisis (LDA) and Masukan Terus Rangkaian Neural (FFNN) dengan data yang berbeza seperti Kepekaan, Ketentuan, Ketepatan, Kawasan Bawah Lengkung (AUC), Cohen's kappa (k), Kepersisan dan Pengukur F.

**Feature Extraction and Machine Learning Techniques for Melanoma Recognition by Using Dermoscopy Images**

## ABSTRACT

Melanoma is a potentially life-threatening neoplasm or a skin cancer that only exist on skin surface of human body. It was manifested by the growing, unusual-looking skin lesion, of the odd-shaped, uneven, or uncertain borders and multiple colors in advanced cases. Thin melanomas a few millimeters in diameter can mimic benign nevi and cannot be detected by the normal eye's examination. The only possibility to diagnose them is using the dermoscopy as a tool. Early recognition and surgical excision can be curative for the patient. The help of technology such as data mining and machine learning can improve the diagnosis accuracy. Thus, in this work a preprocessing of dermatology cancer recognition is proposed. Several experiments were performed on MIAS databases. Single Segmentation Based Fractal Analysis (SFTA), Gray Level Co-Occurrence Matrix (GLCM), Gabor filter and Hu Moment vector feature extraction information yielded a good recognition result. However, by a combination of features gives higher results of more than 90% on five different classifiers with various algorithms such as *k*-Nearest Neighbors (kNN), Fuzzy k-Nearest Neighbors (FkNN), Linear Discriminat Analysis (LDA) and Feed Forward Neural Network (FFNN) with different performance measurement such as Sensitivity, Specificity, Accuracy, Area under Curve (AUC), Cohen's kappa (*k*), Precession and F-Measure.

# CHAPTER 1 : INTRODUCTION

## 1.1 Project Background

Melanoma is a probably life-threatening neoplasm, and it is miles a form of most cancers that gives in the skin's pigment cells that is referred to as melanocytes. Melanocytes produce melanin which is to help defend the pores and skin from ultraviolet (UV) radiation such as sunlight. It was manifested by the uncommon-searching skin lesion, shaped, growing, uneven, or uncertain borders and more than one color in several cases. Benign nevi can mimic with thin melanomas a few millimeters in diameter and cannot be detected with normal eyes examination. Melanoma is a skin cancer, and it is growing fast if not untreated. It can spread to the lower a part of skin which is called dermis and enter the lymphatic system and arts of the body which is lungs, brain, liver or bone. Immunosuppression, sun sensitivity, and exposure to ultraviolet radiation are additional risk factors (Miller, A. J., & Mihm, M. C., 2006).



One of the subjects of object detection and image recognition is cancer detection. It is more challenging to learn and prove about cancer detection and a complex assignment for computer system. Using the dermoscopy as a device is the only possibility to diagnose. Dermoscopy widely used non-invasive technique for diagnosing skin tumor especially for melanoma (Junji Kato et al, 2019). The patient can use early recognition and surgical excision as their curative. Using type of technology such as machine learning and data mining can substantially improve the diagnosis accuracy.

There is a developing enthusiasm for growing programmed frameworks which guide the dermatologists in early recognition of cancer which is include composition of few primary steps along with features extraction, image segmentation, selection and lesion class. The paper (C Grana et al., 2013) proposed new mathematical descriptors for the border of pigmented skin lesion images, like lesion slope and lesion slope regularity. The different works proposed incredible techniques to cancer segmentation and characterization. Ganster has completed 87% sensitivity and 92% specificity for a large statistic set with greater than 5300 dermoscopy images (Ganster et al., 2001). Recent research results from P Sabouri show 90.05% specificity and 83.06% sensitivity of cascade classifiers in tenfold pass-validation mode for recognition of melanoma in medical images (P Sabouri et al, 2014). Early texture kind methods are used for exploring the statistical evaluation of image. This studies work investigates the performances of pre-processing levels with one of a kind aggregate of functions extraction and classifiers. This will allow the most effective and efficient combination of features extraction and classifiers to be identified for this project.



The images are needed to be pre-processed. Different type of feature extractions will be employed which is segmentation-based fractal texture analysis (SFTA), gray-level co-occurrence matrix (GLCM), hu moment vector and gabor image. Trial for performance of a single and combination of these features will be study. The results obtained from the extracted functions are then used in the artificial intelligent classifiers. Artificial intelligence is universal, and it is also approximated an extremely flexible function that was used for the first time in the field of cognitive and engineering sciences. In this thesis, to classify the skin cancer image, the use of artificial intelligent methods is employed. A Matlab program will be developed for the purpose of extracting the features and developing the classifiers of artificial intelligent. Since then, using appropriate feature extraction techniques, feature vectors are derived and these methods can be utilized to accomplish elite of learning calculations that at last improves prescient precision of classifier (Khalid et al, 2014). The details of the extracted features will be used in artificial intelligent methods. The developed system will be tested to validate the results.

## 1.2 Problem Statement

Learning and detecting melanoma proves quite a complete task and more challenging for computer systems. This is due to the dimensions that cannot be detected by using normal eyes examination. The way to diagnose them is the use of the dermoscopy as a device. Early popularity and surgical excision can be healing for the affected person. The help of technology which include data mining and machine learning knowledge can extensively improve the prognosis accuracy. Some researcher came out with single features approach by P. Shahi et.al, 2018 that used GLCM wavelet and Tamura in their research with 100% accuracy by using SVM,87.5% accuracy by using KNN, 87.5% accuracy by using ensemble and 75% accuracy of decision tree. However, single features give some limitation in term of accuracy by A. Ismail et.al, 2019, which is limited to capture the dominant features of a large-scale structures. Thus, some other researchers decide to fusion the features to get a better result suggested by C. Barata et.al,2017, and proved that the fusion method seems the best approach. However, some fusion acquires a large number of training samples and takes time to execute by (A. A. Ross and R. Govindarajan, 2005). Mostly some researchers only focus on term of accuracy in their research result. Thus, in this thesis, a study of the best single features will be compared. Then, a suitable fusion feature also will be defined in term of several performance analysis such as sensitivity (SE), accuracy (ACC), precision (Pre), kappa, F-measure (Fmea), area under curve (AUC) and specificity (SP).

## 1.3 Objectives

The objective of this research is as stated:

1. To compare segmentation-based fractal texture analysis (SFTA), gray-level co-occurrence matrix (GLCM), hu moment vector and gabor image for each trial from the dermatology MIAS database.

2. To propose combination of features from the dermatology MIAS database (I. Giotis,2015) with linear discriminant analysis (LDA), K-nearest neighbor (kNN), fuzzy K- nearest neighbor (FkNN) and neural network (NN).

3. To evaluate the performance of dermatology detection system with various classifiers of sensitivity (SE), accuracy (ACC), precision (Pre), kappa, F-measure (Fmea), area under curve (AUC) and specificity (SP)



## 1.4 Scope

This research will be carried out by detection of melanoma image cancer by using classifier technique. This work is only limited with two types of cancer image that cannot be distinguish which is melanoma and naevus. The databases of images in this work only consist of upfront images and not deal with different poses. This image will go through features extraction process that consist of type of features which is Segmentation-based

Fractal Texture Analysis (SFTA), Gray-Level Co-occurrence Matrix (GLCM), Hu moment Vector and Gabor Image. This feature suitable to extract the image based on recent research by T.T. Htay which is using GLCM as features extraction to detect early-stage breast cancer on Mammography image (T.T. Htay et al,2018) while Xi Li is using hu moment invariant for time-frequency analysis of Partial Discharge (PD)-induced Ultra High Frequency (UHF) signal in gas insulated switchgear (GIS) as features extraction (Xi Li et al,2018). Single feature and fusion with only two types of features are discussing to get a different result. From then, Linear Discriminant Analysis (LDA) as a linear classifier, K- nearest neighbor (kNN) as a non-linear classifier, fuzzy K- nearest neighbor (FkNN) as a hybrid classifier and Neural Network (NN) as a complex classifier. This classification method will be employed to test the sensitivity, accuracy, precision, kappa, F-measure, area under curve and specificity. Finally, the highest percentage of the various performance, should confirm which feature produces the best results.

### 1.5 Thesis outline

The chapters of this dissertation largely follow the order in which the work was done. The scope and objective of the work is presented in this chapter. The second chapter is a literature review encompassing most of infant monitoring research. This chapter also describes some existing applications and research related that have been developed using some of the most common techniques and some critical analysis has been made on them. In addition, the algorithm for all features and classifiers are examined.

The third chapter introduces overview of image with different features extraction and explanation on image built up approached. This chapter will provide a walkthrough of the design and development of image preprocessing and feature extraction method of cancer images recognition system. Chapter four presents best of performance analysis and experimental results from the dermatology MIAS database with linear discriminant analysis (LDA), K-nearest neighbor (kNN), fuzzy K- nearest neighbor (FkNN) and neural network (NN) under different performance. Finally, conclusion remarks and future work are provided in chapter five of the dissertation.

# CHAPTER 2 : LITERATURE REVIEW

## 2.1 Introduction

In this section, this research is about melanoma recognition in dermoscopy images using machine learning techniques. Melanoma is a type of a skin cancer that cannot see with normal eyes. Therefore, there is a growing interest in designing programmed frameworks to support dermatologists in the early detection of melanoma. (SM Goldsmith, 2014; R Garnavi et al, 2012; H Ganster et al, 2001).

In these cases, image classification can be improved by result based on different factors. This literature review recommended that by using many types of features and a selected suitable classification method improving the producing and the accuracy of melanoma recognition in dermoscopy images using classifiers techniques. This section will highlight on major methods involved in melanoma recognition in dermoscopy images using classifiers techniques

However, classifying and recognizing images quite more challenging. Therefore, the result determined which condition allows the program to extract the features with the highest accuracy. The result can be determined by using suitable classification. In these cases, this research uses Segmentation-based Fractal Texture Analysis (SFTA), Gray-Level Co-occurrence Matrix (GLCM), Hu moment Vector, Gabor Image for feature extraction. The extracted feature is then further used in the artificial intelligence algorithm classifiers which is Fuzzy k Nearest-Neighbor (FkNN), Neural Network (NN), K Nearest-Neighbor (KNN), and Linear Discriminant Analysis (LDA) classifiers. The accuracy results gathered by using the combination of different feature extraction method and artificial intelligence classifiers are discussed later in chapter 4.

## 2.2 Features Extraction Method

Extraction of applicable features from the images is an important part of the recognition (Brahnam et al, 2004). Therefore, choosing the proper feature extractor is crucial when designing a system with high recognition rate. The different features of an image include color, texture, shape or domain specific features (J.C. Kavitha et al,2017). From paper (Hiam Alquran et al,2017), they used features extraction method by extract the skin cancer image by using Gray Level C0-Occurance Matrix (GLCM). Technically speaking, a set of variables that contain discriminating and characterizing information about a skin cancer image are defined as features. Consequently, a selection from such features to construct an array is called a feature vector. Feature extraction is indeed a mathematical transformation applied on the data to generate the feature vector.



## 2.3 Gray Level Co-Occurrence Matrix (GLCM)

One feature extraction method used in this study is the Gray Level Co-occurrence Matrix (GLCM) feature extraction method. It is proposed by Haralick (Haralick et al, 1973). This method contains two steps in order to extract features. The first method is where spatial digital image information is extracted using a co-occurrence matrix calculated in a pixel neighborhood that is determined by a window of specific size. Within the window, the matrix consists of frequencies of any combination of gray levels existing between pairs of pixels that are separated by a specific distance and angular relationship. While the second steps of the GLCM methods are to compute and compute the statistics of the co-occurrence matrix of the gray level. It then describes the spatial information according to the relative position of the array elements. Some examples of the most widely used spatial measures are entropy, contrast, moment of inverse difference, and correlation.

$$\begin{matrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{2} \\ \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{1} \\ \mathbf{2} & \mathbf{2} & \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{1} \end{matrix}$$

Figure 2.1: Example of Matrix

$$\boldsymbol{C} = \frac{\mathbf{1}}{\mathbf{16}} \begin{bmatrix} \mathbf{4} & \mathbf{2} & \mathbf{1} \\ \mathbf{2} & \mathbf{3} & \mathbf{2} \\ \mathbf{0} & \mathbf{2} & \mathbf{0} \end{bmatrix}$$

Figure 2.2: Getting the Gray Level Co-Occurance matrix

Consider the figure 2.1. If the operator position 1 pixel to the right and 1 pixel to down is used, this will result in obtaining the gray level co-occurrence matrix as shown in the figure 2.2 on the right above, where C is the count of the number of times F (x, y) = i.e., F (x + 1, y + 1) = j. As an example, the first inputs are the result of the fact that 4 times 0 appears below and to the right of another 0. Meanwhile, factor 1/16 is the fact that there are 16 pairs entering this matrix, thus normalizing the matrix. to be estimates of co-occurrence probabilities.

The difference statistics are then normalized by dividing each element of the vector by the number of possible pixel pairs. Let **d**= (dx, dy) be the displacement vector between two image pixels, and g(d) the graylevel difference at distance **d**.

$$g(d) = f(i, j) - f(i + dx, j + dy) \qquad (2.1)$$

pg (g, d) is the histogram of the gray-level differences at the specific distance, **d**. One distinct histogram exists for each distance **d**. Several texture measures can be extracted from the histogram of graylevel differences:

- Mean

$$\mu_d = \sum_{k=1}^{N} g_k p_g(g_k, d) \qquad (2.2)$$

Small mean values $\mu_d$ indicate coarse texture having a grain size equal to or larger than the magnitude of the displacement vector.

- Entropy

$$H_d = \sum_{k=1}^{N} p_g(g_k, d) Iog p_g(g_k, d) \qquad (2.3)$$

This is a measure of the homogeneity of the histogram. It is maximised for uniform histograms where $H_d$ is the entropy value.

- Variance

$$\sigma_d^2 = \sum_{k=1}^{N} (g_k - \mu_d)^2 \, p_g(g_k, d) \tag{2.4}$$

The variance, $\sigma_d^2$ is a measure of the dispersion of the gray-level differences at a certain distance, **d**.

- Contrast

$$c_d = \sum_{k=1}^{N} g_k^2 \, p_g(g_k, d) \tag{2.5}$$

The initial increase in grey level between adjacent pixels is measured. High spatial frequencies are present in high contrast images where $c_d$ is the contrast value.

N is the number of gray levels and $p_g(g_k, d)$ is the entry $(g_k, d)$ in the GLCM.

## 2.4 Hu Moment Invariant

An essential issue in the field of pattern analysis is the recognition of objects and characters, regardless of their position, size and orientation. The idea of using moments in character recognition gained eminence when a set of invariant moments is derived using algebraic invariants (Muharrem Mercimek et al, 2005). Hu defines seven values, calculated by normalizing the central moments through order three, which are invariant to the object's scale, position and orientation (Ming Kuei Hu, 1962) In terms of the central moments, the seven moments are given as

$$M_1 = (\eta_{20} + \eta_{02}) \tag{2.6}$$

$$M_2 = (\eta_{20} + \eta_{02})^2 + 4\eta_{11}^2 \tag{2.7}$$

$$M_3 = (\eta_{30} + \eta_{12})^2 + (3\eta_{21} + \eta_{03})^2 \tag{2.8}$$

$$M_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \quad (2.9)$$

$$M_5 = (\eta_{30} + 3\eta_{12})^2(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \quad (2.10)$$

$$+(3\eta_{21} + \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]$$

$$M_6 = (\eta_{20} + \eta_2)[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \quad (2.11)$$

$$+4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})$$

$$M_7 = (3\eta - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \quad (2.12)$$

$$-(\eta_{30} + 3\eta_{12})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2]$$

Where, $\eta_{pq}$ is the normalized central moment of order (p+q)

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{w}},\ w = \frac{p+q}{2} + 1 \quad (2.13)$$

and $\mu_{pq}$ is the central moment of order (p+q). Hu moment equation can be explained as:

$M_1$= The sum of horizontal and vertical directed variance, more distributed towards horizontal and vertical axes, the values are enlarged.

$M_2$= The covariance value of vertical and horizontal axes when the variance intensity of vertical axis and horizontal axis were similar.

$M_3$= The result emphasizing the values inclined to left/right and upper/lower axes.

$M_4$=The result emphasizing the values counterbalancing to left/right and upper/lower axes.

$M_5$,$M_6$,$M_7$= The extraction of values invariant against size, rotation, and location.

## 2.5 Segmentation Fractal Texture Analysis (SFTA)

SFTA is occur in two main steps. First the input image is decomposed into binary image using two threshold binary decomposition method (TTBD), and then mean, area is calculated from each binary image and fractal dimension are calculated from each border image. In TTBD algorithm, first a set of threshold value T is computed using multilevel otsu algorithm. The value of threshold is choosen that it minimizes the intraclass variance and maximize inter class variance. Then otsu algorithm is applied to each image until the number of thresholds, nt is obtained, where nt is a user defined parameter. When the value of nt is large the number of redundant features is also increases, and it increase the complexity and memory requirement. The next step is to decompose the input gray scale image into binary images using otsu algorithm. First select a pair of threshold values from T and apply to the input image.

$$\boldsymbol{I}(\boldsymbol{x},\boldsymbol{y}) = \begin{cases} \boldsymbol{1} \; \boldsymbol{if} \; \boldsymbol{t_L} < \boldsymbol{ROI} < \boldsymbol{t_U} \\ \boldsymbol{0} \end{cases} \tag{2.14}$$

Thus 2*nt binary images are obtained. From each binary image mean and area are calculated, and fractal measurements are calculated from border image. The boundaries of binary images are calculated using the given formula.

$$\Delta(\boldsymbol{x},\boldsymbol{y}) = \begin{cases} \boldsymbol{1} \; \boldsymbol{if} \; \boldsymbol{\varepsilon N8}(\boldsymbol{x},\boldsymbol{y}); \\ \quad \boldsymbol{Ib}(\boldsymbol{x}',\boldsymbol{y}')=\boldsymbol{0} \\ \quad \boldsymbol{Ib}(\boldsymbol{x},\boldsymbol{y})=\boldsymbol{1} \\ \boldsymbol{0}, \; \boldsymbol{else} \end{cases} \tag{2.15}$$

Fractal dimension is calculated using box counting method. It is similar to perimeter method. First cover the input image with grid and then count how boxes of the grid are