A NEW METHODOLOGY FOR EVALUATION AND BENCHMARKING OF SKIN DETECTOR BASED ON AI MODEL USING MULTI CRITERIA ANALYSIS

QAHTAN MAJEED YAS

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENT FOR DOCTOR OF PHILOSOPHY (ARTIFICIAL INTELLIGENCE)

FACULTY OF ART, COMPUTING & CREATIVE INDUSTRY
UNIVERSITI PENDIDIKAN SULTAN IDRIS

2018

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

ABSTRACT

This study aims to develop a new multi-criteria decision analysis methodology for skin detector evaluation and benchmarking based on artificial intelligence models. Two experiments were conducted. The first experiment comprised two stages: (1) Adaptation of the best previous case of skin detection approach utilizes multi-agent learning based on different color spaces. This stage aimed to create a decision matrix of various color spaces, and three groups of criteria (i.e., reliability, time complexity, and error rate within dataset) to test, evaluate and benchmark the adapted skin detection approaches. (2) Performance of multiple evaluation criteria for skin detection engines, this stage included two key stages. First, the correlation between criteria to investigate their relationship and determine their degree of correlation. Second, the performance analysis of criteria to identify the factors that affect the behavior of each criterion. The second experiment utilized a new multi-criteria decision-making by adopting the integration of TOPSIS and AHP to benchmark the results of skin detection approaches. In the validation process, multi-criteria measurement was used to calculate the trade-off for different criteria. Color spaces assessment were conducted to determine the best color spaces with adaptive skin detection engines. Moreover, mean and standard deviation values for thresholds were calculated to select the best color space. Two groups of findings were provided. First, the overall comparison of external and internal aggregation values in selecting the best color space, that is the norm RGB at the sixth threshold. Second, (1) the process proves that the distribution of color spaces with its threshold values affects the behavior of the criteria determined as a trade-off between the criteria according to their weight distribution. (2) The YIQ color space obtains the lowest value and is the worst case, whereas the norm RGB color space receives the highest value and is the most recommended. (3) The best result achieved at the threshold = 0.9. Thus, the implications of this study benefit individuals, research centers, and organizations interested in skin detection applications. Moreover, it provides benefits to software developers working in industrial companies and institutions in developing different techniques and algorithms with different applications.

METODOLOGI BARU UNTUK PENILAIAN DAN PENYELESAIAN DETEKSI KULIT BERDASARKAN MODEL AI MENGGUNAKAN ANALISIS KRITERIA MULTI

ABSTRAK

Kajian ini bertujuan untuk membangunkan metodologi baharu bagi menilai dan menanda aras pengesanan kulit berdasarkan model kecedasan buatan menggunakan analisis pelbagai kriteria. Untuk tujuan ini, dua eksperimen telah dijalankan. Eksperimen pertama terdiri daripada dua peringkat: (1) Adaptasi kes terbaik terdahulu dalam mengesan kulit menggunakan pendekatan multi-agen berdasarkan ruang warna yang berbeza. Peringkat ini bertujuan untuk membuat matriks keputusan pelbagai ruang warna dan tiga kumpulan kriteria (iaitu, kebolehpercayaan, kerumitan masa, dan kadar kesilapan dalam set data) untuk menilai dan menanda aras pendekatan pengesanan kulit yang telah disesuaikan. (2) Prestasi kriteria pelbagai penilaian bagi enjin pengesanan kulit, di mana peringkat ini melibatkan dua peringkat kekunci. Pertama, korelasi antara kriteria untuk menyiasat hubungan dan menentukan darjah korelasi. Kedua, analisis prestasi kriteria untuk mengenal pasti faktor kriteria yang mempengaruhi kelakuan setiap kriteria. Eksperimen kedua menggunakan pendekatan membuat-keputusan multi-kriteria baharu melalui integrasi antara TOPSIS dan AHP untuk menanda aras keputusan pendekatan pengesanan kulit. Di dalam proses pengesahan, pengukuran pelbagai kriteria digunakan untuk mengira keseimbangan bagi pelbagai kriteria. Penilaian ruang warna dijalankan untuk menentukan ruang warna yang terbaik dengan enjin pengesanan kulit yang telah diadaptasi. Seterusnya, nilai min dan sisihan piawai dikira untuk memilih ruang warna yang terbaik. Hasil dapatan daripada dua kumpulan adalah seperti berikut. Pertama, perbandingan keseluruhan nilai agregasi luaran dan dalaman dalam memilih ruang warna terbaik, iaitu RGB norma pada ambang keenam. Kedua, (1) proses membuktikan bahawa penagihan ruang warna dengan nilai ambangnya mempengaruhi kelakuan kriteria yang ditentukan sebagai keseimbangan antara kriteria berpandukan pengagihan berat masing-masing. (2) Ruang warna YIQ memperoleh nilai terendah dan merupakan kes terburuk, manakala ruang warna norm-RGB memperoleh nilai tertinggi dan paling disyorkan. (3) Dapatan terbaik dicapai pada ambang = 0.9. Oleh itu, implikasi kajian ini memberi manfaat kepada individu, pusat penyelidikan dan organisasi yang berminat dalam aplikasi pengesanan kulit. Kajian ini turut memberi manfaat kepada pembangun perisian yang bekerja di industri dan institusi dalam membangunkan teknik dan algoritma yang berbeza bagi aplikasi yang berbeza.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATION

ANN	Artificial Neural Network
AHP	Analytic Hierarchy Process
ANP	Analytic Network Process
CPU	Central Processing Unit
CIE	Commission International de L'Eclairage
CR	Consistency Ratio
DM	Decision Matrix
EM	Evaluation Matrix
FP	False Positive
FN	False Negative
GH	Grouping Histogram
GDM	Group Decision Making
HAW	Hierarchical Adaptive Weighting
IT	Information Technology
KNIME	Konstanz Information Miner
KEEL	Knowledge Extraction based on Evolutionary Learning
LUT	Lookup Table
MCDM	Multi- Criteria Decision Making
MADM	Multi- Attribute Decision Making
MCDA	Multi-Criteria Decision Analysis

MEW	Multiplicative Exponential Weighting
RI	Random Index
SVM	Support Vector Machine
SAN	Segment Adjacent-Nested
SAW	Simple Additive Weighting
TP	True Positive
TN	True Negative
TOPSIS	Technique for Order Preference by Similarity to Ideal Solution
WEKA	Waikato Environment for Knowledge Analysis
WSM	Weighted Sum Model

CHAPTER 1

INTRODUCTION

1.1 Introduction

This chapter introduces the research direction, research background, and a statement of the problem. This chapter also presents the ambitions, motivations, and objectives of this research are also presented.

Section 1.2 presents a brief background of the research components. Section 1.3 introduces the statement of the problem, which is the basis of the research direction. Section 1.4 discusses the scope of the research. Section 1.5 describes the research objectives. Section 1.6 presents a general view of the research. Finally, Section 1.7 briefly outlines the main structure of the research.

1.2 Research Background

Decades ago, the skin detection approach has been considered an important platform for various fields, such as medical and several scientific disciplines (L. Huang et al. 2015). In other words, skin detection has gained an important function in a wide range of image or video processes for various applications. A few factors that directly impact skin appearance include illumination, background, camera characteristics, and ethnicity (Kakumanu, Makrogiannis, and Bourbakis 2007). Elgammal, Muang, and Hu (2009) defined the skin detection approach as a process of finding skin-colored pixels and regions in an image or a video into a specific region. This process is typically used as a preprocessing step to finding regions in images that potentially detect the human face and limbs. The skin detection approach includes various applications, such as face detection, (Zhipeng, C., Junda, H., & Wenbin 2010), face tracking (Tsai 2012), gesture analysis (Hussain, I., Talukdar, A. K., & Sarma 2014), Internet pornographic image filtering (Lee, Kuo, and Chung 2010), surveillance systems (Zui Zhang 2009), content-based image retrieval systems (Patil, C. G., Kolte, M. T., Chatur, P. N., & Chaudhari 2014), and various human–computer interaction domains (Hollender et al. 2010). The most practical and effective techniques are used in developing skin detector artificial intelligence (AI) algorithms according to the literature on skin detection for skin pixel and non-skin pixel features based on color features. On the contrary, many researchers have applied hybrid algorithms in AI models (Singh Sisodia and Verma 2011; (Shruthi, M. L. J., & Harsha 2013; Zaidan et al. 2014b). However, with the current rapid development of the skin detection approach in various applications, finding an evaluation and benchmarking

methodology that is reliable, effective, and comprehensive has become critical (Jones and Rehg 1999; Phung, Bouzerdoum, and Chai, D. 2005; Gamage, Akmeliawati, and Chow 2009; Taqa and Jalab 2010a).

Considering the basic criteria evaluation of reliability, time complexity, and error rate within the dataset in the design of any skin detector application, (Jones and Rehg (1999) adapted three criteria, namely reliability, computational cost, and error rate of skin detection. In one of the earliest works that highlight the problem of skin detection evaluation and benchmarking, three general requirements for the skin detection approach are reported: adapted reliability (i.e., the obtained skin detection rate and false positives) and datasets (i.e., the obtained equal error rate comparison of AI models) with less time-consuming requirements to process web images.

Despite the importance of the remaining criteria, Phung, Bouzerdoum, and Chai, D. (2005) highlighted the dataset criterion by comparing two algorithms. The dataset is represented by training and testing for skin and non-skin pixels for skin segmentation images. However, the output images created through a classifier are compared pixel-wise with the ground truth of skin segmentation. Gamage, Akmeliawati, and Chow (2009) reported a skin detection algorithm that has been tested with images through independent databases. They investigated the size of the image, which has a significant impact on time complexity. Thus, they proved that increasing image size leads to low accuracy than increase time complexity of the experiment. Finally, Taqa and Jalab (2010a) stated that reliability is a prerequisite for

skin detection evaluation. They highlighted a reliability criterion based on accuracy, precision, and recall of the image color despite the importance of the remaining criteria. However, the quality assessment of skin detection requires attention.

Consequently, two key problems are encountered by skin detection developers. One is the evaluation of skin detection approaches based on the abovementioned evaluation criteria and benchmark new skin detection approach versus existing approaches. Therefore, the evaluation and benchmarking process need to consider these requirements. Despite the tradeoff among various criteria, (Jones and Rehg (1999); Phung, Bouzerdoum, and Chai, D. (2005); Gamage, Akmeliawati, and Chow (2009); Taqa and Jalab (2010a) have adopted each of the proposed criteria. They attempted to evaluate the reliability criterion for a given time complexity based on different datasets. However, the term "reliability" is unclearly defined in the literature. According to the preceding studies mentioned, the percentage of reliability varies depending on different adapted algorithms and thus exhibit an inconsistent level. Meanwhile, Fernandes, Cavalcanti, and Ren (2013) reported time complexity variation between the algorithms, which depend on the CPU time. Consequently, the processing time of an image is affected, but this aspect is excluded in the scope of the present research. Therefore, the calculation should be the highest percentage of reliability compared with the lowest time complexity of the output image. Kawulok (2013) mentioned that the dataset can be divided into two classes, namely training and validation data, to find the minimum detection error. In general, all these studies have

proven the evaluation and benchmarking process of each of these criteria based on independent guidelines.

Therefore, conducting further investigations and developing a clear methodology for testing, evaluation, and benchmarking are necessary to standardize basic and advanced requirements for the skin detection approach. Redefining the problem of evaluation and benchmarking need is also necessary. Moreover, the new evaluation methodology must be flexible to handle the conflicting criteria problem and must have the capability to maintain the current criteria.

1.3 Significance of Study

The evaluation and benchmarking of skin detection approaches are important areas for many researchers and organizations interested in their applications. Many individuals and organizations are interested in the applications of skin detection approaches, such as researchers working in scientific research centers, developers working in industrial companies and institutions, and graduate students enrolled in schools that develop various applications of skin detection approaches. Thus, the importance of the study is the development of multiple applications of skin detection approaches, including face detection, face tracking, gesture analysis, Internet pornographic image filtering, surveillance systems, content-based image retrieval systems, and various human-computer interaction domains. Moreover, this study

covered different techniques and algorithms in image processing to solve various problems in the field (See figure 1.1).

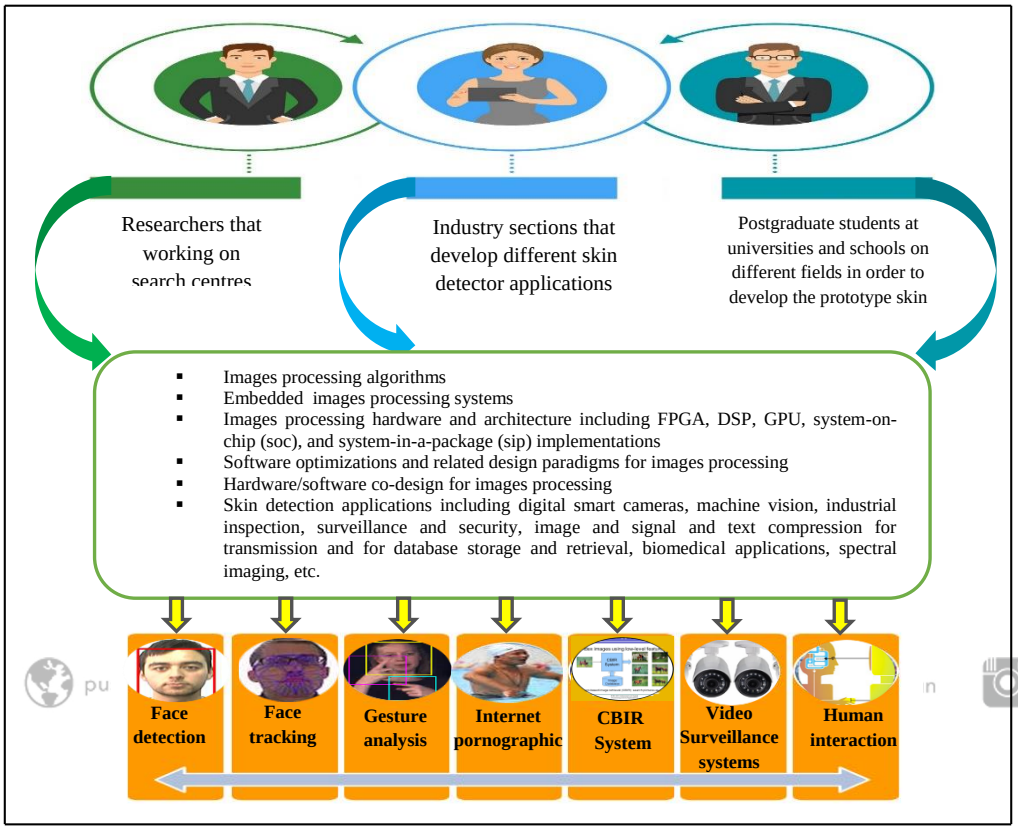


Figure 1.1. Significance of the Study

1.4 Problem Statement

In most scientific studies, the research problem is investigated based on the gap derived from the literature. The determination of any gap in a study is considered an important aspect and a challenging task. Therefore, the research problem is formulated from the gap in this study, which is derived from three sources as shown

in Tables 2.3, 2.4, and 2.5. Moreover, the gap leads to the general problem determined as the selection benchmarking problem.

According to Oxford Dictionaries (Oxford dictionaries. 2013), "benchmarking" is a standard or point of reference against which things may be compared. In the fields of information technology and computer system, benchmarking is the process of comparing the output of different systems for a given set of criteria to ensure quality, improvement, contribution, or performance of a new system (Trentesaux et al. 2013). Benchmarking comparisons have been conducted for the industry software, such as insurance, military, telecommunication, and commercial software. In other domains, "benchmarking" usually refers to the collection of a substantial body of quantitative data. Similarly, skin detection approaches should be tested, evaluated, and accurately benchmarked. The main problem that seriously threatens the future of skin detection, such as requirements, tools, and methodologies, can be defined as the selection problem. By contrast, insufficient benchmarking of the dataset is a critical problem in developing a system (Köhler et al. 2012; Bajcsy et al. 2013; Wang et al. 2014; Gurari et al. 2015). See figure 1.2 highlights the general problem of this study.

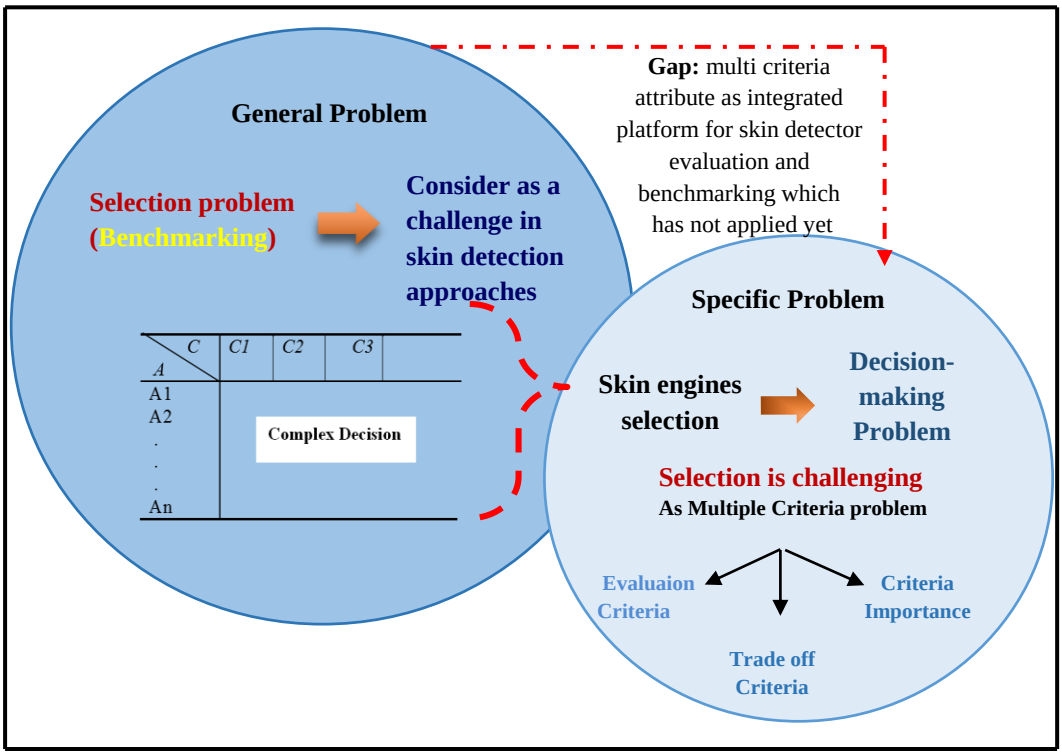


Figure 1.2. Research Problem and Gap

Three main requirements should be measured first to evaluate skin detection approaches: reliability, time complexity, and error rate within the dataset (Yadav and Nain 2016; Ramachandra and Busch 2017; Luo and Guan 2017). Reliability should possess a high rate, a low time complexity for conducting output images, and a low error rate from training datasets, as represented in the corners of the magic triangle shown in Figure 1.3. However, Duffner and Odobez (2014),Szkudlarek and Pietruszka (2015), and Lu and Mandal (2015) proposed the time complexity procedure based on the CPU and complexity of the algorithm. Sanmiguel and Suja (2013), Ballerini et al. (2013), Stergiopoulou et al. (2014), and Mahmoodi and Sayedi (2016) provided the dataset criterion without procedure. Each of the aforementioned researchers also suggested the reliability criterion but without reference to a specific

level for comparison with other criteria. In particular, the main challenge to the development of skin detection is that developers focus on either increasing reliability with low error rate or decreasing time complexity only. Accordingly, this trade-off (conflicting criteria) is reflected in the evaluation and benchmarking processes (See Figure 1.3).

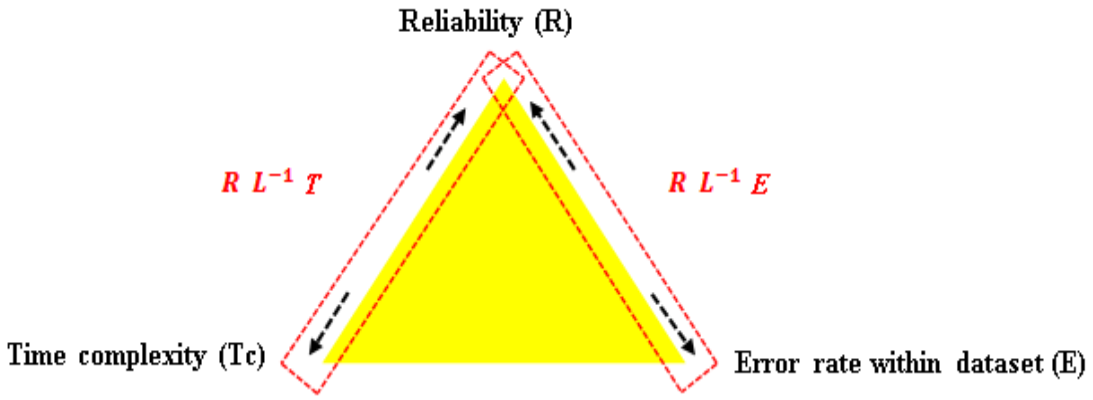


Figure 1.3. Magic Triangle of Skin Detection Requirement (B. B. Zaidan and Zaidan 2018)

In the figure, the magic triangle includes three basic criteria, namely, reliability (R), time complexity (Tc), and error rate within dataset (E). Thus, according to the formula (R L^{-1} Tc), the relationship among these criteria represents reliability and time complexity, whereas the relationship (R L^{-1}E) denotes reliability and error rate within the dataset; hence, both relationships should be inversed.

Current evaluation and benchmarking processes rely on evaluating the approach using available criteria and tools of reliability. For example, the RapidMiner tool neglects the time complexity of criteria and conducts the benchmarking process

by testing the reliability values of skin detection approaches (Ashwin Satyanarayana 2013; Jovic, A., Brkic, K., & Bogunovic 2014; Al-odan and Saud 2015).

However, the limitations of the benchmarking tools that cannot satisfy the overall needs of skin detection benchmarking are as follows: (1) benchmarking between two or more techniques (Wahbeh, A. H., Al-Radaideh, Q. A., Al-Kabi, M. N., & Al-Shawakfa 2011), (2) individual calculation of the error rate (Madeo, R. C., Lima, C. A., & Peres 2017), (3) matching between techniques (Verhoeven, G., Sevara, C., Karel, W., Ressl, C., Doneus, M., & Briese 2012), (4) calculation of the time consumption of the techniques (Pujol and García 2012), and (5) calculation of the overall parameters of the reliability group (Burget et al. 2010).

Notably, all researchers in the fields of skin detection evaluation and benchmarking have used either a criterion or a set of criteria defined in the literature but with different priorities. Consequently, the problem of evaluation and benchmarking processes in skin detection is the multi-criteria problem with conflicting criteria.

The specific problem is derived from three concerns: The first concern is the trade-off criteria mentioned above. The second concern is the evaluation criteria, considered as an important aspect of this study based on the accurate calculation of the criteria values. Thus, the reliability group is based on the matrix of parameters (TP, FP, TN, and FN). Furthermore, some pixels are lost after cropping the

background from the skin images using Adobe Photoshop when the actual class needs to be labeled manually and compared with the predicted class to compute one of the matrices of parameters. This process is debatable because it affects the results from all reliability groups (matrix, relationship, and behavior of parameters) (Liu et al. 2013). The third concern is the importance of criteria, which is a key objective in this study through evaluation and benchmarking, despite the conflict among them. Therefore, the conflict among the criteria is a major challenge during the evaluation process. A suitable procedure should be developed for these objectives when increasing the importance of a particular evaluation criterion and when reducing others. Two main aspects should be considered. 1) The behavior of skin detectors should be understood with particular importance to the design. 2) The evaluation of the approach should consider the trade-off (Rautaray and Agrawal 2015). Thus, the problem can be solved according to the objectives proposed in this study.

1.5 Research Scope

The scope of this research is defined by the following considerations.

A) This research primarily focuses on the development of skin detection evaluation and benchmarking. Therefore, the adapted skin detection approach is not the main issue; hence, the type of skin detection approach in place does not matter.

B) The proposal does not impose any restriction on the type of skin detection application in place.

C) The selected case study has been used based on multi-agent learning Bayesian and neural network; hence, the case study covers the implementation of multi-agent learning Bayesian and neural network models and their evaluation methodologies.

D) Multi-criteria decision making (MCDM), or multi-attribute decision making or multi-criteria decision analysis methods can be classified into three direcaions according to the type of information collected from the decision maker: no information, information on the attribute, and information on the alternative. We focus on information on the attribute and therefore review MCDM techniques that can deal with "information on the attribute."

E) Considering that our data can be classified as cardinal data, we study the algorithms related to MCDM methods with information on the attribute and cardinal data.

1.6 Research Objectives

The objectives of this research are highlighted in the development objectives (main contribution and developments), which are listed as follows:

1- To investigate the existing technology on evaluation and benchmarking approaches of the skin detector model and highlight its weakness.

2- To identify and perform multi-dimensional criteria for skin detector engines.

3- To develop a new methodology for the benchmarking of skin detector based on AI model using MCDM techniques.

4- To validate the identified criteria and determine the effectiveness of color spaces and thresholds in the proposed methodology.

1.7 General View and Scope of the Research

This research has a cross-domain nature, thereby primarily focusing on software testing, evaluation, and benchmarking. The research is designed to solve the problem of skin detection evaluation and benchmarking.

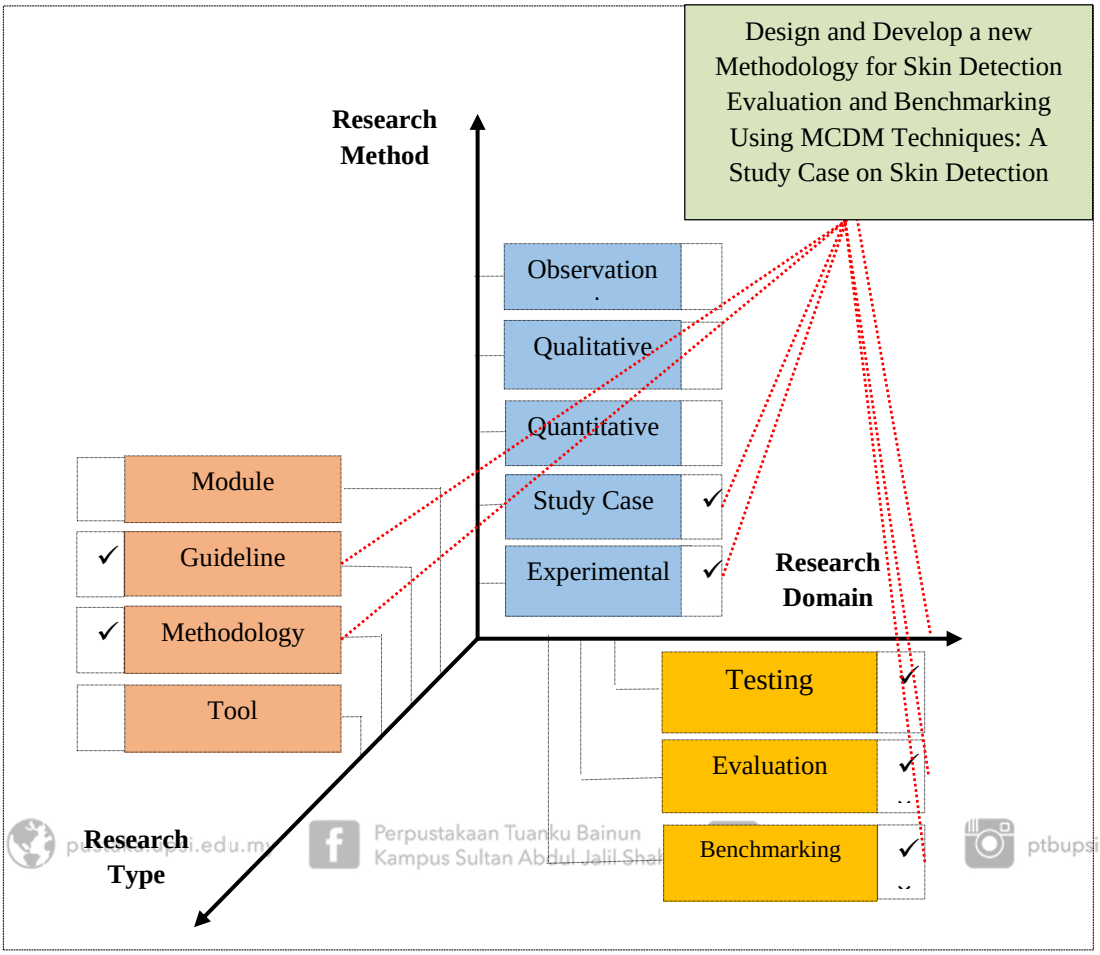


Figure 1.4. General View of the Research

Different research methods are involved in the present research because the problem is classified as an inter-disciplinary problem. Experimental analysis is the first research method used to select, evaluate, and adopt a suitable multi-criteria scoring in skin detection. The second research method is a case study, in which the skin detection approach is selected to perform further experiments to adjust the settings and parameters of the new evaluation and benchmarking approach. Two outputs are expected from this research. The first output is a methodology performed via several experiments. This methodlogy can be used in the process of skin detection evaluation

and benchmarking and can also be adopted in assessing several aspects and approaches in other skin detection applications, such as face detection, face tracking, and video surveillance evaluation. The second output is a complete guideline for testing, evaluation, and benchmarking of skin detection, which can accommodate the development in this area (Figure 1.4).

1.8 Organization of Thesis

The present thesis comprises seven chapters. The background of the skin detection evaluation and benchmarking problem, research objective, and research scope are provided in Chapter 1. The remainder of this thesis is organized as follows.

Chapter 2: An in-depth investigation on skin detection evaluation and benchmarking approaches is presented in this chapter. This chapter investigates four objectives: first, gathering a multi-criteria for evaluation as a reliability group, time complexity group, and error rate within the dataset group; second, identifying the problems and limitations of the evaluation and benchmarking for skin detection approaches; third, investigating the available benchmarking techniques/tools; and fourth, examining the MCDM techniques, context, and parameters, as well as the opinion of researchers on each technique.

Chapter 3: This chapter presents a detailed description of the four research phases designed to develop a new methodology for the evaluation and benchmarking of skin

detection approaches. The first phase investigates the existing technology on evaluation and benchmarking approaches of skin detection models and highlights their weakness. The second phase includes identifying and performing multidimensional criteria for skin detection engines to build a decision matrix. The third phase develops a new methodology for the evaluation and benchmarking of skin detection approaches based on an AI model using MCDM techniques. The last phase achieves validation and determines the effectiveness of color spaces in the proposed methodology according to the results obtained from the third phase.

Chapter 4: The second phase is implemented to achieve the objectives in this chapter. Correlation coefficient is adapted based on Pearson approach to identify the relationship among different criteria. In addition, the performance analysis of the criteria is conducted to evaluate skin detection approaches. Therefore, the evaluation of the criteria is necessary to determine their behavior based on the threshold values for each color space utilized in the case study.

Chapter 5: The third phase is carried out for the integration of the most appropriate MCDM techniques, such as ML-AHP and TOPSIS, to obtain efficient results. The AHP method is applied to derive the weights for different criteria, whereas the TOPSIS method is implemented to select the best alternatives from different configurations. This phase is implemented to identify the results and ranks of various algorithms in skin detection approaches.

Chapter 6: The last phase validates the results collected from the third phase. This phase realizes the final objective based on the discussion of the results by comparing the multi-criteria measurement process and conducting the statistical calculations for color space performance. The two directions of the validation process are as follows: (1) trend analysis of the behavior of multi-criteria measurement using a numerical sequence process and (2) discussion of the behavior of color spaces based on criteria and statistical calculation of the thresholds for the final results obtained, as represented by the external scores.

Chapter 7: The main claims, limitations, and contributions of this research, as well as the directions for future research, are elaborated in this chapter.

1.9 Chapter Summary

This chapter highlights the background of the case study, which is related to the evaluation and benchmarking of skin detection approaches based on AI models. A total of targets are identified in the problem statement. First, the trade-off among multiple criteria is briefly explained based on the conflict among them. Second, the weakness in the techniques/tools during the benchmarking process is utilized to test the reliability values of skin detection approaches. The overall scope of the research is identified in five points. The objectives proposed in this research are highlighted for development. The general perspective of the research is designed to solve the problem of skin detection evaluation and benchmarking.

Thus, we comprehensively surveyed the literature to cover all cases related to our case study, objectives, and research problem in Chapter 2.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter introduces the published academic literature focusing on skin detection based on AI models in terms of evaluation and benchmarking. The following five main sections are explored; section (1) is to investigate, identify and gathering multi-criteria for evaluation and benchmarking for reliability group, time complexity group, and error rate within dataset group. Section (2) is to investigate the available benchmarking for techniques/ tools problems and limitations. Section (3) is to investigate in open issues and challenges for evaluation and benchmarking process of the criteria. Section (4) is to investigate the theoretical background about multi criteria decision making technquies for our research problem. Section (5) is to investigate about selecting the case study according to the literature. The background of Chapter Two is presented in Section 2.1. Various evaluation criteria are discussed in Section 2.2. The benchmarking techniques and literature on these techniques are reported in Section 2.3. The concern for evaluation criteria are discussed in section 2.4. MCDM techniques are elaborated in Section 2.5. Skin detection conducted based on the case

study in Section 2.6. Finally, our claims from the literature review are summarized in Section 2.7. Whereas, Figure 2.1 illustrated the chapter direction of the research based on the basic taxonomy. The details will be explained throughout this chapter.

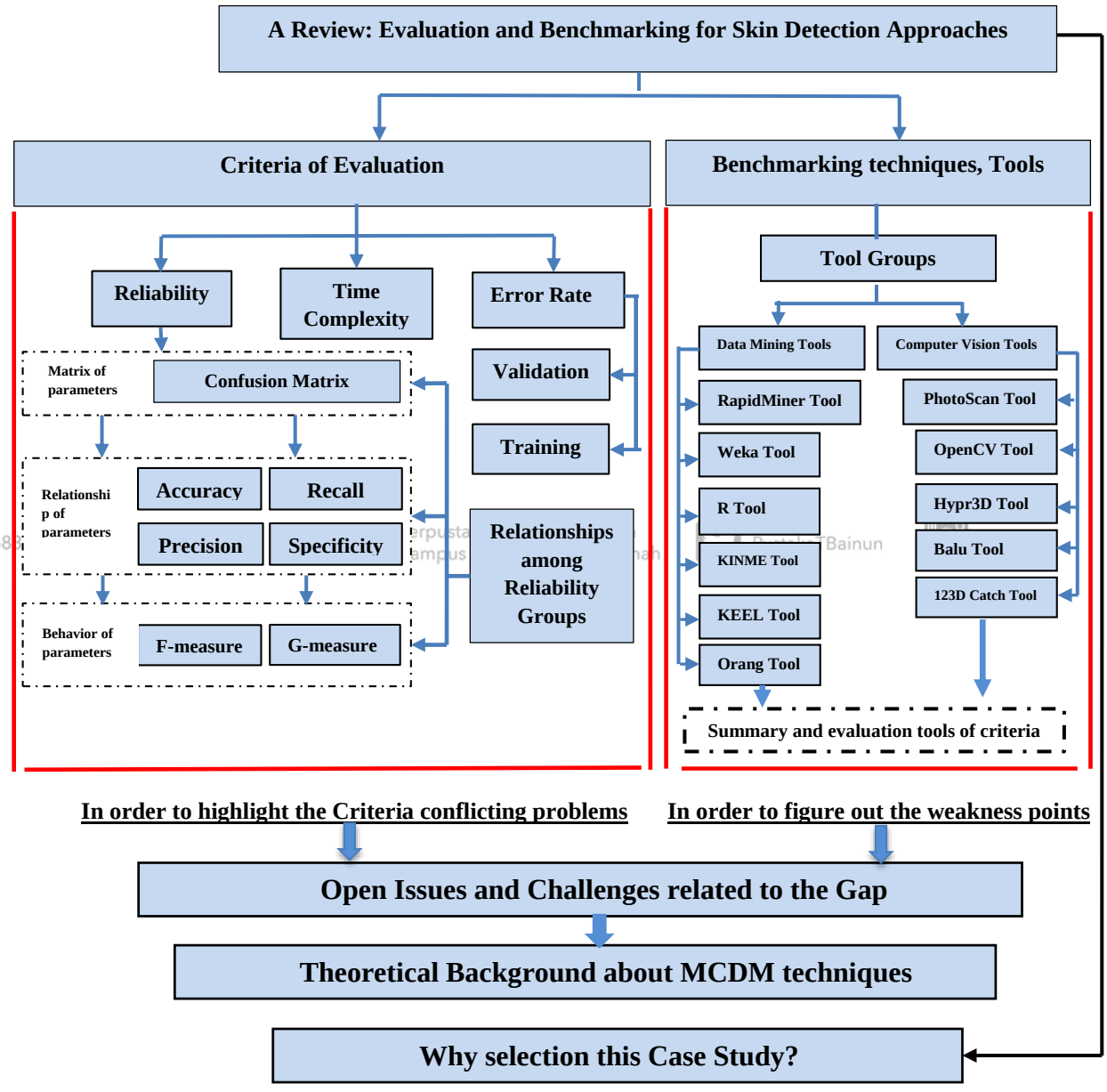


Figure 2.1. Taxonomy for Chapter Direction of the Research

2.2 Evaluation and Benchmarking for Skin Detector Approaches

Evaluation and benchmarking will be reviewed based on two critical directions. First, the conflicts in existing evaluation criteria will be highlighted. Second, the limitations of existing benchmarking techniques/ tools will be determined. Other aspects influencing the process will also be discussed such as open issues and challenges to evaluation and benchmarking will be presented. As well as, the recommend and solution based on several MCDM techniques will be highlighted. Lastly, a summary of the case study will be selected based on identifying the weak points for different of research.

2.2.1 Criteria of Evaluation

Skin detection requires powerful and reliable evaluation and benchmarking tools under diverse circumstances. Numerous criteria for evaluating novel skin detectors have been proposed. Xu, T., Wang, Y., & Zhang (2012) and Huang et al. (2015) discussed two parameters, reliability and dataset of skin detection, and explained their /relationship. The computation of reliability based on a large dataset with a manually defined ground truth by using a receiver operatig characteristic (ROC) curve depends on the testing methodology. According to the Sun (2010), Jensch, Mohr, and Zachmann (2012), Song et al. (2017) studied reliability and time complexity and their effects on each other. Several studies have argued that this relationship is based on the complex background of the algorithm, and other researchers have claimed that high-

quality and accurate skin detection relies on the computation features of the processor. Li and Kitani (2013), Kawulok, M., Kawulok, J., Nalepa, J., & Smolka (2014a), and Zaidan, A. A., Ahmad, N. N., Karim, H. A., Larbani, M., Zaidan, B. B., & Sali (2014a) proposed three criteria for skin detection and explored the influence of each criterion on the output. The present study provides a brief review of the basic criteria for computing reliability (i.e., reliability group), training and testing (i.e., dataset group), and time complexity to evaluate and benchmark skin detection techniques (Phung, Bouzerdoum, and Chai, D. 2005).

The present study shows that testing skin detectors can increase their reliability, reduce time complexity, and minimize error rates within datasets without tradeoff. Moreover, improving one criterion affects other criteria, thus complicating the comparison with previous approaches.

Issues with regard to the evaluation and benchmarking of skin detectors can therefore be considered as a complex problem with conflicting criteria. Table 2.1, presents evaluation and benchmarking processes used in previous studies.

Table 2.1

Literature review of evaluation criteria.

Author	Analysis and View of Points
Sun,2010	This study proposed a new skin detector approach for detecting skin in a single image. **Highlights:** • Both criteria in the local skin colour model have been evaluated • The study included trade-off between criteria. Notes: Generating a local skin colour model requires more processing time without decreasing the accuracy.
Xu, T., Wang, Y., & Zhang. 2012	The study introduced a novel human skin detector method based on the flexible neural tree for pixel-wise skin colour detection. **Highlights:** • Two criteria are evaluated. • Trade-off between the criteria has been investigated. Notes: Skin detector method achieves higher accuracy and lower error rate compared with sate of the art methods.
Li, C., & Kitani, K. M.2013	The study presented a Self-Organizing Mixture Network (SOMN) is modified to improve its computation performance, stability and applicability, and a probability density estimation method to establish colour models for skin and non-skin classes accurately and effectively. **Highlights:** • The author has been evaluated each three criteria of the SOMN method • Trade-off between for three criteria has been investigated. Notes: The proposed method has advantages such as higher estimation accuracy, simple structure and computational form and faster convergence speed compare with EM algorithm.

(Continue)

Table 2.1 (*continued*)

Author	Analysis and View of Points
Kawulok, M. Kawulok, J., Nalepa, J., & Smolka. 2014a	The author proposed a new method for creating self-adaptive seeds for spatial-based skin segmentation. **Highlights:** • In proposed method the criteria is evaluated. • Trade-off between for three main criteria has been done. Notes: An extensive experimental study demonstrated that the DSPF domain outperforms all of the investigated methods both for the ECU and HGR data sets.
Zadain. A.A. et al. 2014a	The author proposed a skin detector that uses a hybrid method involving the technique of a grouping histogram (GH) for the Bayesian method and a back-propagation neural network with an adjacent segment-nested (SAN) technique, to improve skin detection performance. **Highlights:** • Evaluation all criteria are addressed as individual in this study. • The author not addressed the referred for a trade-off between three key criteria. Notes: The main contribution of the proposed multi agent learning system for skin detection is demonstrated a high detection rate.
Huang, et al. 2015	The author proposed a novel method for human skin detector in real-world images. **Highlights:** • The study conducted evaluation for two criteria based on various algorithms. • The study point out trade-off between both criteria Notes: The proposed method outperforms traditional graph cuts with significant accuracy using publicly available dataset.
Jensch, Mohr, and Zachmann. 2015	The author(s) compared the quality of three skin detector and segmentation approaches, RehgJones, HybridClustering, and NeuralGasColorClustering, for real applications. **Highlights:** • For each approach have been evaluated both criteria • The author mentioned the trade-off between the criteria. Notes: The study focused on measure the computation time needed to evaluate their usefulness for three real applications in skin segmentation approaches.

(*Continue)*

Table 2.1 *(continued)*

Author	Analysis and View of Points
Song.W.et al. 2017	The study proposed the motion-based skin ROI detection method that was designed and implemented in GPU- based connected component labelling algorithm to achieve real-time performance. **Highlights:** • In real time of skin ROI detection evaluated two main criteria • The author mentioned the trade-off between the criteria. Notes: The proposed labelling algorithm could be implemented in real time of the monitoring applications; in addition the proposed method provides natural user interfaces for several multimedia applications, such as touch less operation systems.

The findings obtained from studies explored in Table 2.1 are summarized as follows:

1- Most of the benchmarking processes are individually implemented for skin detectors.

2- Skin detectors exhibit tradeoff among different criteria, resulting in problems that should be addressed.

3- Each skin detector requires multiple criteria for application. Therefore, a general solution needs to be developed.

2.2.1.1 Reliability Group

This section discusses the relationship among reliability (R) groups and emphasizes the significance of evaluation and benchmarking criteria. The reliability group includes three key sections, namely, matrix of parameters (i.e., confusion matrix),

relationship of parameters (i.e., accuracy, recall, precision, and specificity), and behaviour of parameters (i.e., F-measure and G-measure). Several techniques and applications for skin detection are evaluated by using the reliability groups that included three stages (Al-Mohair, Mohamad Saleh, and Suandi 2015; Kawulok, Kawulok, and Nalepa 2014b; (Khan et al. 2014b).

2.2.1.1.1 Matrix of Parameters (MP) Section

Taqa and Jalab (2010a) and A. A. Zaidan et al. (2014b) reported that skin detection based on color and texture features is a more efficient and reliable method compared with other techniques. The matrix of parameters (i.e., confusion matrix) is used as a basic evaluation technique of the reliability group. We will discuss each section within the reliability group in detail and highlight all criteria within this group.

A) Confusion Matrix

ROC curves are commonly used to present results for binary decision problems in machine learning. In a binary decision problem, a classifier labels examples as either positive or negative. The decision made by the classifier can be represented in a structure known as a confusion matrix or contingency table. Measures of the quality of classification are built from a confusion matrix, which records correctly and incorrectly recognized examples for each class. The confusion matrix has four

categories. True positives (TP) are examples correctly labelled as positives. False positives (FP) refer to negative examples incorrectly labelled as positive. True negatives (TN) correspond to negatives correctly labelled as negative. False negatives (FN) refer to negative examples incorrectly labelled as negative. This matrix forms the basis for many common metrics. Table 2.2 shows the confusion matrix and equations of several common metrics that can be calculated from it.

Table 2.2

Confusion Matrix

	Actual Parameters	
Predicated Parameters	**TP**	**FP**
	FN	**TN**

These categories calculated by numbers along the major diagonal represent the correct decisions made, and the numbers of this diagonal represent the errors the confusion between the various classes (Davis and Goadrich 2006; Fawcett 2006; Sokolova and Lapalme 2009; W 2011). The TP rate (TPR; also called hit rate and recall) of a classifier is calculated as:

The true positive rate of a classifier is estimated as

$$\mathbf{TPR} = \frac{\boldsymbol{Positives\ correctly\ classified}}{\boldsymbol{Total\ positives}} \qquad \textbf{(2.1)}$$

The FP rate (FPR; also called false alarm rate) of the classifier is calculated as

$$\mathbf{FPR} = \frac{\boldsymbol{Negatives\ incorrectly\ classified}}{\boldsymbol{Total\ negatives}} \quad \textbf{(2.2)}$$

The confusion matrix for skin segmentation can be used as the basis for the evaluation; TP, FP, TN, and FN are computed for all pixels in the testing set through skin detector testing. FP is the proportion of non-skin pixels classified incorrectly as skin, and TP is the proportion of skin pixels classified correctly as skin. TN and FN are the complements of FP and TP, respectively. These parameters are computed for all pixels in the skin classifier testing set through skin segmentation testing. The probability samples are used to distinguish between the actual class manually cropped and the predicted class for positive and negative labels in the classification model (Naji, Zainuddin, and Jalab 2012; Tazaree,A. et al. 2014). Issues arise because some pixels will be lost after the background is cropped from the actual class via Adobe Photoshop. Conflict of data generated occurs because of the increase in TP and TN and the decrease in FP and FN. This phenomenon shows a conflict among the probability criteria (A. A. Zaidan et al. 2014b).

2.2.1.1.2 Relationship of Parameters (RP) Section

Fawcett (2006) and Taqa and Jalab (2010) proposed two types of criteria in the reliability group by constructing three histograms for skin detectors models. They highlighted the criteria into matrix of parameters (i.e., confusion matrix) and the relationship of the parameter (RP) (i.e., accuracy, recall, precision, and specificity).

A) **Accuracy Measure**

The performance accuracy (Acc) of experiments and scientific instruments must be assessed. Vamsi Krishna et al. (2017) stated that accuracy typically refers to the exactness of an analytical method or the closeness of agreement between the measured and accepted values, either as a conventional true value or an accepted reference value. Therefore, the features of this measure should be highlighted based on the performance evaluation of various data.

Accuracy statistically measures how well a binary classification test correctly identifies or excludes certain conditions. Metz (1978) calculated the proportion of true results (TP and TN) among the total number of examined cases. Accuracy also refers to the closeness to the target and the trueness of the closeness to a specific target on average. This measure thus represents a weighted arithmetic mean of precision and inverse precision (weighted by bias) and a weighted arithmetic mean of recall and inverse recall (weighted by prevalence) (Powers 2013). However, these features do not make this measure an evaluation parameter that is generalizable for each instance. Accuracy can be calculated as follows:

$$\textbf{Accuracy}(\textbf{Acc}) = \frac{\textbf{True Positive+True Negative}}{\textbf{True Positive+False Positive+False Negative+True Negtaive}} \qquad \textbf{(2.3)}$$

B) Precision (PR) Measure

Different methods are used to measure the performance of a system. Precision is a common evaluation index employed in most studies. It refers to the number of TPs for different classes (i.e., the number of items correctly labeled as belonging to the positive class) divided by the total number of elements described as belonging to the positive class (i.e., the sum of TPs and FPs, which are items incorrectly labeled as belonging to the class). Therefore, precision refers to the consistency or the ability to group well, which can also be called TP accuracy. The key contribution of precision with their combinations focuses only on positive examples and predictions. Precision can capture information on the rates and kinds of errors obtained. These features imply that this measure is a major criterion in the evaluation matrix (Sokolova and Lapalme 2009; Powers 2013; Elalami 2014). Thus, precision represents several correctly classified positive examples divided by the number of examples labeled by the system as positive:

$$\textbf{Precision}(\textbf{PR}) = \frac{\textbf{True Positive}}{\textbf{True Positive+ False Positive}} \qquad \textbf{(2.4)}$$

C) Recall (RE)Measure

Numerous evaluation and benchmarking measures are used to measure the performance of a system. Recall, also referred to the sensitivity, and is the most commonly used evaluation method for sensitivity. Al-Mohair, Mohamad Saleh, and Suandi (2015) stated that recall is a measure of completeness or quantity. It is the average probability of a complete retrieval referred to as the TPR. Recall is the ratio

of TP components to elements inherently ranked as positive. Thus, recall represents the number of correctly classified positive examples divided by the number of positive examples in the data. Finally, the contribution of the recall also focuses only on the positive examples and predictions. These features can capture information on the rates and kinds of errors generated (Chaouch et al. 2013; Powers 2013; Elalami 2014). Recall can be calculated as follows:

$$\textbf{Recall}(\textbf{RE}) = \frac{\textbf{True positive}}{\textbf{True positive+False nagetive}} \qquad \textbf{(2.5)}$$

D) Specificity (SP) Measure

Measures of accuracy, precision, and recall focus on positive statistics and determine the rate of prediction classes during evaluation. Specificity indicates the capability of a classifier to recognize patterns of a negative class to determine real negative situations correctly labeled as negative. It can be measured between 0 and 1. Specificity, which is sometimes called TN to measure the rate of negative values correctly identified and is complementary to the FPR (Rathore, S., Iftikhar, M. A., Hussain, M., & Jalil 2013). The behavior of this measure is completely contrary to that of recall. Therefore, specificity measure can be calculated with the same confusion matrix used to calculate the rest of the criteria as follows:

$$\textbf{Specificity}(\textbf{SP}) = \frac{\textbf{True Negative}}{\textbf{True Negative + False Positive}} \qquad \textbf{(2.6)}$$

2.2.1.1.3 Behaviour of Parameters (BP) Section

This section highlights two main parameters, namely, F-measure and G-measure. Kokkinos (2010), Kim and Lattimer (2015) used these measures in classification algorithms to achieve highly accurate image processing.

A) F-measure (F)

The F1-measure, defined as the harmonic mean of the precision and recall of a binary decision rule, is a traditional way of assessing the performance of classifiers. As it favors high and balanced values of precision and recall, this performance metric is usually preferred to as label-dependent weighted classification accuracy when classes are highly imbalanced and when the cost of an FP relative to an FN is not naturally given to the problem at hand (Chinchor, N., & Sundheim 1993). Several studies have defined F-measure as F1-measure (Chinchor, N., & Sundheim 1993). The full definition of the F-measure is given as follows (Chinchor, N., & Sundheim 1993).

$$F_\beta = \frac{(\beta^2+1)RP}{\beta^2 P+R} \qquad (0 \leq \beta \leq +\infty) \qquad (2.7)$$

where β is a parameter that controls a balance between P and R. If β = 1, F1 becomes equivalent to the harmonic mean of P and R; if β > 1, F becomes recall-oriented; and if β < 1, F becomes precision-oriented, e.g., F0 = P.

Consequently, F-measure is considered appropriate, especially in cases with the imbalance of other samples. This measure is also used to improve performance

and retrieve information when the negative of the sample largely outnumbers the positive. Additionally, it allows the differential weighting of recall and precision, though these features are commonly given equal weight. Thus, the features of the F-measure present a systematic method. This criterion corresponds to the TP for the arithmetic mean of the predicted positives, and real positives should be reflected to ensure closeness to the ideal value. Thus, comparing F-measure in skin color space is similar to comparing color with non-color (Kokkinos 2010; Rehanullah Khan , Allan Hanbury , Julian Stöttinger 2012; Bilal et al. 2015). F-measure is expressed as follows:

$$\mathbf{F-measure\ (F)} = \frac{\mathbf{2*precision*recall}}{\mathbf{recall+precision}} \qquad \mathbf{(2.8)}$$

B) G-measure (G)

G-measure refers to the geometric mean of precision and recall. This measure can be represented mathematically as the square root of precision multiplied by recall and is typically used to evaluate the performance of algorithms. G-measure can only increase when precision and recall are high. It therefore reflects the general classification of algorithms in terms of the performance and accuracy of the positive sample classification (Kim, J. H., & Lattimer 2015). G-measure is expressed as follows:

$$\mathbf{G-measure\ (G)} = \sqrt{\mathbf{Precision\ \times\ Recall}} \qquad \mathbf{(2.9)}$$

In this section, we discussed a reliability group including three key sections. This group is considered a critical measure in the present study. Most studies used only one of these measures despite their importance in research. By contrast, the present study adopted all these criteria within a unified guideline. The next section will discuss the summary of relationships among these measures and their parameters and contributions.

2.2.1.1.4 Summary of Relationships among Reliability Groups

The reliability group included three key sections as abovmentioned. The relationship among these measurements reinforces the importance of evaluating and benchmarking skin detectors without one key section exceeding others.

First, the matrix of the parameters shows that the confusion matrix comprises TN and FN models, which are complemented by TP and FP, respectively. The confusion matrix is considered an important measure for all cases within the classification model. These models are also regarded as the backbone in mathematical calculations for all parameters in an evaluation matrix. Table 2.3 shows the variations in application of this criterion in different studies. The features and procedures of a confusion matrix are essential to distinguish between positive and negative regions. Accordingly, this measure should be considered a basic criterion in the evaluation and benchmarking process.

Second, relationship parameters include four parameters, namely, accuracy, precision, recall, and specificity. The accuracy parameter represented as TP is the number of correct predictions when an instance is positive, and TN is the number of correct predictions when an instance is negative (ElAlami, M.E., 2014). Accuracy is considered an important measure in the evaluation process. It measures the closeness of the agreement between the measured and accepted values. Accuracy also represents a weighted arithmetic mean between precision and recall according to Eq. (2.3). Sokolova and Lapalme (2009) stated that precision does not depend on TN but generally adapts only with positive examples (TP) and predictions. This measure provides the best perspective on the classification performance of skin detection. Precision measures the predictive value of skin detection as either positive or negative, depending on the class for which it is calculated. Similarly, recall does not depend on TN but adapts only with TP and predictions. TPR cases are those instances correctly predicted as positive to measure different cases on the basis of the class for which they are calculated. Precision and recall constitute the basic mathematical relationship with accuracy within positive examples, TP, and predictions. Jadhav, Nalbalwar, and Ghatol (2011) indicated that the number of correctly classified negatives is equal to the ratio of TN to the sum of TN and FP. Thus, FPR equals (100−specificity). Specificity measure can therefore recognize negative-class patterns to evaluate real negative situations correctly predicted to be negative in different cases.

Finally, behavior parameters include two key measures. (Bilal et al. 2015) reported that F-measure refers to the weighted average of precision and recall when it reaches the optimal value of 1. The worst F-measure score is 0. F-measure is the most popular criterion because it provides a tradeoff between recall and precision. Moreover, (W 2011) stated that G-measure represents an effective normalized version of precision and recall as TP to the geometric mean of predicted and real positives. The information content of the G-measure corresponds to the arithmetic mean information by recall and precision. The G-measure generally represents the geometric mean of precision and recall to evaluate the performance of algorithms. This index also often reflects the accuracy of the positive sample classification for particular cases. These points are summarized in Figure 2.2.

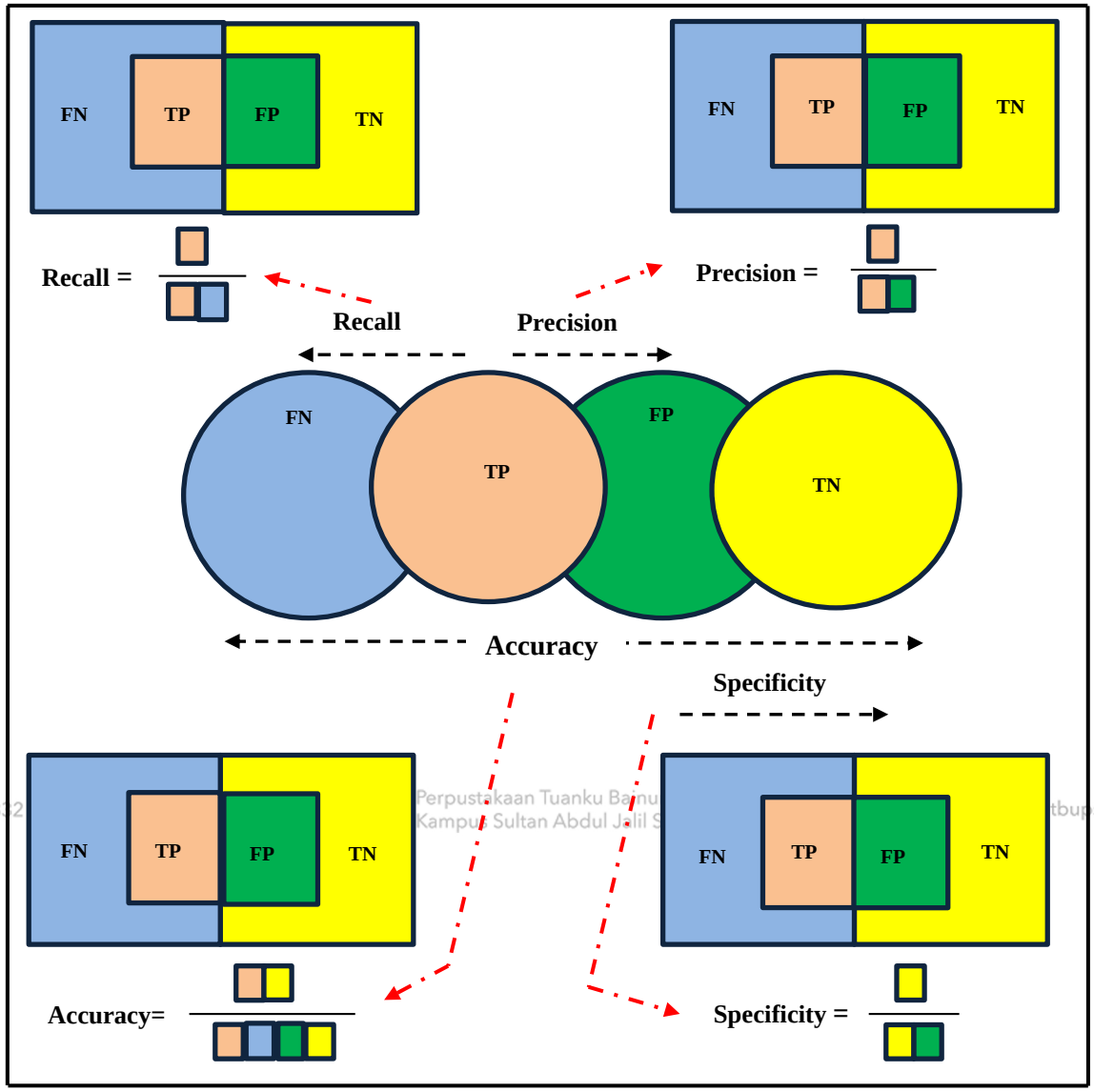


Figure 2.2. Sections Reliability Group

A comprehensive review of literature was performed to determine the limitations of using the aforementioned important criteria for studies on the skin detection. Thus, we noted the varying proportions of the different criteria, and the results are shown in Table 2.3.

Table 2.3

Reliability Group for Skin Detection Approach

No	Author & year	class	Matrix of parameters				Relationship of parameters				Behavior of parameters	
			Confusion Matrix				Accuracy	Precision	Recall	Specificity	F- measure	G-measure
			TP	TN	FP	FN						
1	**Mahmoodi and Sayedi 2015**	Adaboost	√	×	√	×	√	×	×	×	×	×
2	**Mahmoodi and Sayedi 2014a**	Adaboost	√	×	×	√	×	√	√	×	√	×
3	**Priya, Vasuhi, and Vaidehi 2015**	Adaboost	√	√	√	√	√	√	√	×	×	×
4	**Wen, D., Han, H., & Jain 2015**	SVM	√	×	√	×	×	×	×	×	×	×
5	**Thaweekote, V., Songram, P., & Jareanpon 2013**	Adaboost	×	×	√	√	√	×	×	×	×	×
6	**Saxen, F., & Al-Hamadi 2014**	Bayes	×	×	×	×	×	×	×	×	√	×
7	**Lin, H. I., Hsu, M. H., & Chen 2014**	CNN	×	×	×	×	×	×	×	×	×	×
8	**Al-Mohair, H. K., Saleh, J. M., & Suandi 2015**	ANN\ GA	√	√	√	√	√	√	√	×	√	×
9	**Kawulok et al. 2014a**	ANN	√	√	√	√	√	√	√	×	√	×
10	**Szkudlarek and Pietruszka 2015**	Fuzzy	×	×	×	×	√	√	×	×	×	×
11	**Sanmiguel and Suja 2013**	ANN\ Bayes	√	×	√	√	×	√	√	×	√	×
12	**Lin, Leng, and Yu 2013**	Bayesian	√	×	√	×	√	×	×	×	×	×

(Continue)

Table 2.3 *(continued)*

No	Author & Year	class	Matrix of parameters				Relationship of parameters				Behavior of parameters	
			Confusion Matrix									
			TP	TN	FP	FN	Accuracy	Precision	Recall	Specificity	F-measure	G- measure
13	**Soran, B., Hwang, J. N., Lee, S. I., & Shapiro 2012**	SVM	×	×	×	×	√	×	×	×	×	×
14	**Zaidan, A. A., Ahmad, N. N., Karim, H. A., Larbani, M., Zaidan, B. B., & Sali 2014a**	ANN	√	√	√	√	√	×	×	×	×	×
15	**A. A. Zaidan et al. 2014b**	ANN	√	√	√	√	√	√	√	×	×	×
16	**Anghelescu, Serbanescu, and Ionita 2013**	ANN	√	×	√	×	√	×	×	×	×	×
17	**Stergiopoulou, E., Sgouropoulos, K., Nikolaou, N., Papamarkos, N., & Mitianoudis 2014**	Bayes	√	×	√	×	√	×	×	×	×	×
18	**Cheng, J., & Liu 2015**	FSVM	√	×	√	×	√	×	×	×	×	√
19	**Lee, D., Wang, J., & Plataniotis 2014**	Bayesian	√	×	√	√	√	×	×	×	√	×
20	**Yang, Z., Zhu, Y., & Yuan 2014**	AdaBoost	√	×	√	×	×	√	√	×	×	×
21	**Molina, J., Escudero-Viñolo, M., Signoriello, A., Pardàs, M., Ferrán, C., Bescós, J., ... & Martínez 2013**	SVM	√	√	√	√	√	√	√	×	√	×
22	**Khan, Hanbury, Sablatnig, et al. 2014a**	Bayesian	√	√	×	×	×	√	√	×	√	×

(*Continue*)

Table 2.3 (*continued*)

No	Author & Year	class	Matrix of parameters				Relationship of parameters				Behavior of parameters	
			Confusion Matrix				Accuracy	Precision	Recall	Specificity	F-measure	G- measure
			TP	TN	FP	FN						
23	**Khan, R., Hanbury, A., Stöttinger, J., Khan, F. A., Khattak, A. U., & Ali 2014b**	Bayesian	√	√	√	√	×	√	√	√	√	×
24	**Chen et al. 2016**	ANN	√	×	√	×	√	×	×	×	×	×
25	**Chuang, Y., Chen, L., & Chen 2014**	SVM	√	×	√	×	√	√	√	×	√	×
26	**Gor, A. K., & Bhatt 2015**	SVM	√	×	√	√	√	√	√	×	√	×
27	**Mahmoodi, M. R., & Sayedi 2014b**	Adaboost	√	×	×	√	√	×	×	×	×	×
28	**Gupta, A., & Chaudhary 2016**	ANN	√	√	√	√	×	×	×	×	√	×
29	**Z. Li et al. 2015**	Adaboost	√	×	√	×	√	×	×	×	×	×
30	**Pengyu, N., & Jie 2013**	Adaboost	×	×	√	√	×	√	√	×	×	×
31	**Tsitsoulis, A., & Bourbakis 2013**	SVM	√	×	√	√	×	√	√	×	×	×
32	**Xu, T., Wang, Y., & Zhang 2012**	ANN	√	√	√	√	√	×	×	×	×	×
33	**Scherbaum et al. 2013**	Adaboost	×	×	√	×	√	×	×	×	×	×
34	**Esposito, L. G., & Sansone 2013**	SVM	×	×	×	√	×	×	√	√	√	×
35	**Khan et al. 2012**	SVM	×	×	×	×	√	√	√		√	×
36	**Liew and Yairi 2014**	Adaboost	√	×	√	×	√	×	×	×	×	×
37	**Kawulok, Kawulok, and Nalepa 2014**	Bayesian	√	×	√	√	×	×	×	×	×	×
38	**Taimori and Behrad 2015**	Fuzzy	×	×	×	×	√	√	√	×	√	×
39	**Kawulok, M., & Nalepa 2014c**	SVM	√	×	√	×	√	×	×	×	×	×

(*Continue*)

Table 2.3 (*continued*)

No	Author & Year	class	Matrix of parameters				Relationship of parameters				Behavior of parameters	
			Confusion Matrix									
			TP	TN	FP	FN	Accuracy	Precision	Recall	Specificity	F-measure	G-measure
40	**Yan, C. C., Liu, Y., Xie, H., Liao, Z., & Yin 2014**	SVM	√	×	√	×	√	√	×	×	×	×
41	**Yu, Cheng, and Lee 2013**	Naïve Bayes	√	√	√	×	√	√	√	×	×	×
42	**S. Chen et al. 2013**	SVM	×	×	×	×	√	×	×	×	×	×
43	**Zafeiriou, S., Zhang, C., & Zhang 2015**	CNN	√	×	√	×	√	√	√	×	×	×
44	**Tan, W. R., Chan, C. S., Yogarajah, P., & Condell 2012**	Bayesian	√	×	×	√	√	√	√	×	√	×
45	**A.A. Zaidan, Karim, and Ahmad 2010a**	Bayesian	√	√	√	√	√	√	√	√	×	×
46	**Zaidan, A. A., Karim, H. A., Ahmad, N. N., Alam, G. M., & Zaidan 2010**	Fuzzy	√	×	√	√	√	×	×	×	×	×
47	**Nguyen et al. 2014**	SVM	×	×	×	×	√	×	×	×	×	×
48	**Grzejszczak, T., Kawulok, M., & Galuszka 2016**	ANN	√	×	√	√	√	√	√	×	√	×
49	**Xu, Wang, and Zhang 2013**	Bayesian	√	×	√	×	√	×	×	×	√	×
50	**Shoyaib, Abdullah-Al-Wadud, and Chae 2012**	Bayesian	√	√	√	√	√	√	√	×	√	×
	Average		39	10	36	23	37	24	23	3	19	1
	Percentage		88 %	22 %	70 %	46 %	74 %	48 %	46 %	6 %	38 %	2 %

Table 2.3 presents various studies on the skin detection. A comprehensive review of the reliability group during evaluation and benchmarking with various sub criteria is also provided. The ratios of the sub-criteria vary as a result of the conflict among them, which is the challenge to the evaluation process in the present study. This group includes three main stages. The matrix of parameters as confusion matrix comprises four parameters with a ratio of 51.4%. Numerous studies use these metrics to evaluate skin detection, despite the fact that these metrics are heavily criticized in the literature. The relationship of parameters includes four measures with a ratio of 39.5%, and 9.1% of which involved the rest two measures within the behavior of parameters. Table 2.3 shows that the varying rates of sub-criteria led to the difficulty in evaluation and benchmarking because of the absence of a general guideline for evaluating different criteria by researchers. Moreover, each study that used the reliability criterion did not refer to a specific level that can be compared with those of other studies.

2.2.1.2 Time Complexity Group

Time complexity is described as the time needed to address image segmentation in relation to its size, thereby showing a direct correlation among them (L. Sun et al. 2013). Time complexity is a key challenge for numerous studies. Different types of image segmentation algorithms have been proposed with the development of computer and information technology (Comaniciu and Meer 2002; Felzenszwalb and Huttenlocher 2004). These algorithms have been applied in many information

systems. The applications of image segmentation techniques require the creation of an appropriate complexity algorithm with an implementation method to determine whether the algorithm is applicable (Fu 2011; G. Li and Shi 2013). Similarly, time complexity is vital in measuring performance efficiency in image processing (Fu 2011).

Most studies have shown that calculating the processing time for any image mainly depends on the image size. Measurement steps are determined by an algorithm depending on microprocessor features. The microprocessor system typically affects the speed of execution time, such that few operation contents result in a short execution time. (Gamage, Akmeliawati, and Chow 2009) discussed the testing of the algorithm by using images from various databases to achieve high accuracy for small images and low accuracy for large images. Time complexity significantly affects the reliability of an image (i.e., in terms of accuracy) depending on the image size tested. Thus, accuracy and time complexity need to have an inverse relationship. In most studies, a direct correlation among these parameters exists. Therefore, image size and other criteria should be considered in comparing processes (A. A. Zaidan et al. 2015a). These criteria assume a uniform score to facilitate comparison. Figure 2.3 shows the calculation process of time complexity for skin segmentation.

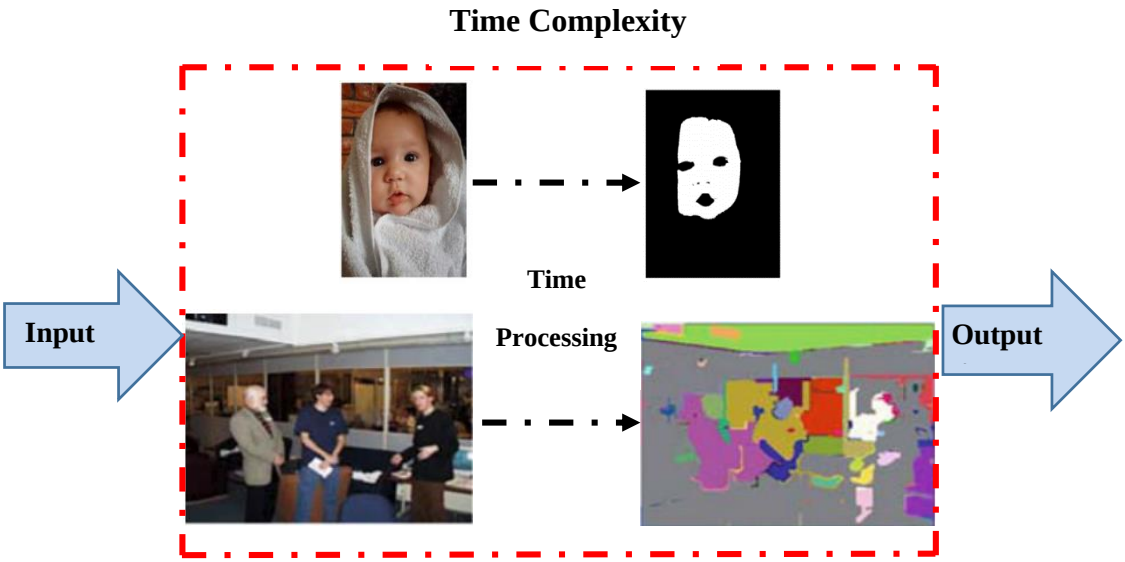


Figure 2.3. Time Complexity Process for Skin Segmentation

In this study, many articles are reviewed to provide a comprehensive view on skin detection. Table 2.4 shows the time complexity of the skin detector approaches.

Table 2.4

Time Complexity group for Skin Detection Approaches

No	Author & year	Class	Time	Note
1	**Mahmoodi and Sayedi 2015**	Adaboost	√	Computation time has been applied in real time
2	**Mahmoodi and Sayedi 2014a**	Bayesian	×	Does not mention time criterion
3	**Priya, Vasuhi, and Vaidehi 2015**	Adaboost	×	Does not mention time criterion
4	**Wen, D., Han, H., & Jain 2015**	SVM	√	Computation time has been applied in real time

(*Continue*)

Table 2.4 (*continued*)

No	Author &year	Class	Time	Note
5	**Thaweekote, V., Songram, P., & Jareanpon 2013**	Adaboost	√	Time complexity has been accounted to algorithm
6	**Saxen, F., & Al-Hamadi 2014**	Bayes	√	Time complexity accounted based on algorithm
7	**Lin, H. I., Hsu, M. H., & Chen 2014**	CNN	×	Does not mention time criterion
8	**Al-Mohair, H. K., Saleh, J. M., & Suandi 2015**	ANN\ GA	√	Time complexity has been applied in real-time
9	**Kawulok et al. 2014a**	ANN	√	Time complexity has been applied on algorithm
10	**Szkudlarek and Pietruszka 2015**	Fuzzy	√	Time complexity accounted based on size image
11	**Sanmiguel and Suja 2013**	ANN\ Bayes	√	Time complexity accounted based on algorithm
12	**Lin, Leng, and Yu 2013**	Bayesian	√	Time complexity accounted based on algorithm
13	**Soran, B., Hwang, J. N., Lee, S. I., & Shapiro 2012**	SVM	×	Dose not mention time criterion
14	**Zaidan, A. A., Ahmad, N. N., Karim, H. A., Larbani, M., Zaidan, B. B., & Sali 2014a**	ANN	√	Time complexity accounted based on algorithm
15	**A. A. Zaidan et al. 2014b**	ANN	√	Time complexity accounted based on algorithm
16	**Anghelescu, Serbanescu, and Ionita 2013**	ANN	√	Time complexity accounted based on algorithm
17	**Stergiopoulou, E., Sgouropoulos, K., Nikolaou, N., Papamarkos, N., & Mitianoudis 2014**	Bayes	√	Time complexity has been applied in real time

(*Continue*)

Table 2.4 (*continued*)

No	Author &year	Class	Time	Note
18	**Cheng, J., & Liu 2015**	FSVM	√	Time complexity has been applied in real time
19	**Lee, D., Wang, J., & Plataniotis 2014**	Bayesian	√	Time complexity accounted based on algorithm
20	**Yang, Z., Zhu, Y., & Yuan 2014**	AdaBoost	√	Time complexity has been applied in real time
21	**Molina, J., Escudero-Viñolo, M., Signoriello, A., Pardàs, M., Ferrán, C., Bescós, J., ... & Martínez 2013**	SVM	√	Time complexity has been applied in real time
22	**Khan, Hanbury, Sablatnig, et al. 2014a**	Bayesian	√	Time complexity accounted based on algorithm
23	**Khan, R., Hanbury, A., Stöttinger, J., Khan, F. A., Khattak, A. U., & Ali 2014b**	Bayesian	√	Time complexity has been applied in real time
24	**Chen et al. 2016**	ANN	√	Time complexity accounted based on algorithm
25	**Chuang, Y., Chen, L., & Chen 2014**	SVM	×	Does not mention time criterion
26	**Gor, A. K., & Bhatt 2015**	SVM	√	Time complexity has been accounted to size image
27	**Mahmoodi, M. R., & Sayedi 2014b**	Adaboost	×	Does not mention time complexity as criterion
28	**Gupta, A., & Chaudhary 2016**	ANN	√	Time complexity accounted based on algorithm
29	**Z. Li et al. 2015**	Adaboost	√	Time complexity accounted based on algorithm
30	**Pengyu, N., & Jie 2013**	Adaboost	√	Time complexity has been applied in real time
31	**Tsitsoulis, A., & Bourbakis 2013**	SVM	×	Does not mention time complexity as criterion
32	**Xu, T., Wang, Y., & Zhang 2012**	ANN	×	Does not mention time complexity as criterion

(Continue)

Table 2.4 (*continued*)

No	Author &year	Class	Time	Note
33	**Scherbaum et al. 2013**	Adaboost	√	Time complexity accounted based on algorithm
34	**Esposito, L. G., & Sansone 2013**	SVM	×	Does not mention time complexity as criterion
35	**Khan et al. 2012**	SVM	×	Does not mention time complexity as criterion
36	**Liew and Yairi 2014**	Adaboost	√	Time complexity has been applied in real time
37	**Kawulok, Kawulok, and Nalepa 2014**	Bayesian	√	Time complexity has been applied in real time
38	**Taimori and Behrad 2015**	Fuzzy	√	Time complexity accounted based on algorithm
39	**Kawulok, M., & Nalepa 2014c**	SVM	√	Time complexity accounted based on algorithm
40	**Yan, C. C., Liu, Y., Xie, H., Liao, Z., & Yin 2014**	SVM	×	Does not mention time complexity as criterion
41	**Yu, Cheng, and Lee 2013**	Naïve Bayes	√	Time complexity accounted based on algorithm
42	**S. Chen et al. 2013**	SVM	√	Time complexity has been applied in real time
43	**Zafeiriou, S., Zhang, C., & Zhang 2015**	CNN	√	Time complexity has been applied in real time
44	**Tan, W. R., Chan, C. S., Yogarajah, P., & Condell 2012**	Bayesian	√	Time complexity accounted based on algorithm
45	**A.A. Zaidan, Karim, and Ahmad 2010a**	Bayesian	×	Does not mention time complexity as criterion
46	**Zaidan, A. A., Karim, H. A., Ahmad, N. N., Alam, G. M., & Zaidan 2010**	Fuzzy	×	Does not mention time complexity as criterion
47	**Nguyen et al. 2014**	SVM	√	Time complexity has been applied in real time
48	**Grzejszczak, T., Kawulok, M., & Galuszka 2016**	ANN	√	Time complexity has been applied in real time

(*Continue*)

Table 2.4 (*continued*)

No	Author &year	Class	Time	Note
49	**Xu, Wang, and Zhang 2013**	Bayesian	×	Does not mention time complexity as criterion
50	**Shoyaib, Abdullah-Al-Wadud, and Chae 2012**	Bayesian	×	Does not mention time complexity as criterion
	Average		35	
	Percentage		70%	

Table 2.4 shows that 70% of the reviewed studies used time complexity. The rest does not use this criterion. Time complexity is a key measure in the evaluation and benchmarking process in the present study. Thus, this criterion should be evaluated, which is a significant challenge to most researchers because of the differences in image sizes. Accordingly, this measure adopted in present study based on image size to extract the output under the same environment. This difference was probably due to the absence of a general guideline for evaluating the criteria. Each study in Table 2.4 that used the time complexity criterion did not specify a particular level to facilitate comparisons among studies.

2.2.1.3 Error Rate within Dataset Group

Error rate refers to the lowest possible error for any classifier of a random outcome that is analogous to the irreducible error. This group constitutes a key criterion during the evaluation and benchmarking process skin detection based on soft computing techniques using specific datasets. However, this criterion includes error rate for

training and error rate for validation as sub-criteria. These error rates should be evaluated in terms of performance. Meanwhile, no standard dataset is available to measure the error rate of skin detection (A. A. Zaidan et al. 2014a). This insufficiency led to a problem in the variation of error rates that resulted from the variation in size for the datasets used in different skin detection experiments. However, many studies collected datasets on the basis of their individual studies, which therefore leads to a waste of effort and time. This study uses the error rate to obtain the minimum error rate of the dataset during training and validation. (Och 2003; Mahdieh and Pournoury 2010; Gijsberts et al. 2014).

2.2.1.3.1 Cross Validation Pattern

Bergmeir, Costantini, and Benítez (2014) showed that the cross-validation process is an estimator used to evaluate prediction errors and is widely used in research. K-fold and leave-one-out cross validation are two popular approaches generally used to evaluate the performance of a classification algorithm on a dataset. The latter involves the random splitting of the dataset into n partitions. At each n-th iteration, n−1 partitions are used as the training set, and the sample left is used as the test set. At each n-th iteration, the entire dataset is used as the training set, whereas the sample left is used as the testing set (Rushing et al. 2015). By contrast, k-fold cross validation has a randomness mechanism, such that the mean accuracy resulting from k-fold cross validation on a dataset is not constant. Therefore, this type of cross validation should be employed for a large amount of data to estimate the accuracy of the model induced

from a classification algorithm (Wong 2015). Cross validation is commonly used to evaluate the predictive performance of a model, which is given a priori or developed by a modelling procedure. Part of the data is basically used to fit each competing model on the basis of data splitting. The rest of the data are used to measure the predictive performance of the models by validating the error rates. The model with the best overall performance is selected based on the minimum error rate (Y. Zhang and Yang 2015).

Zhang and Yang (2015) proposed a framework suitable for high-dimension regression with the possibility of expanding the true dimension of the regression function to reflect the challenge of high dimension and small sample size. This framework is used to investigate the relationship among the performances of cross validation and data splitting ratio in terms of modelling selection, instead of the usual model selection. Cross validation is applied by randomly splitting the data into three disjoint parts, namely, training set, validating set, and testing set. The predictive performance of a model is evaluated by validating its error rate. Multiple data splitting is utilized by either averaging or voting. By contrast, Yang (2007) proposed a procedure that is asymptotically better than others. This procedure is intended for traditional regression setting, which should be generalized to accommodate the high-dimension case depending on the following key points:

1. In the traditional case, the estimator based on the true model is asymptotically better than that based on a model with extra parameters.

2. Traditional parametric regression involves a fixed true model.

2.2.1.3.2 Training Pattern

The training set is the second key parameter in the process of splitting the dataset. Several types of training sets can be used to produce a hypothesis function. Batch method is used for the entire training set and can be applied at once to account for the function. The variation in this method is for the entire training set to modify a current hypothesis recursively until an acceptable hypothesis is obtained. Incremental method is implemented by selecting one member at once for the training set, and this example alone is used to update the current hypothesis. The selection method can also be applied randomly (with replacement) or recursively on the training set (Nilsson 1996). This mechanism is used in artificial intelligence algorithms to obtain results through training and testing. Training is a key step in these algorithms. The algorithm is trained several times to achieve a low error rate for a specific dataset. The training for any algorithm typically involves three cases during implementation. First, the result will be low if the error rate is high and the accuracy is low. Second, the result will be low if the error rate and accuracy are low. Finally, the result will be high when the error rate is low and accuracy is high. This case is considered as the most suitable mechanism because it implies that the datasets on the training set are sufficient in that application and the target is achieved (Zaidan, A. A., Ahmad, N. N., Karim, H. A., Larbani, M., Zaidan, B. B., & Sali 2014a). For example, the output value in ANN algorithm is compared with the associated target output to compute the error rate based on the input data. Gradient steepest descent approach is then used to propagate this error back by adjusting the weights of the network to minimize the sum-of-square error rate. Finally, the entire process is repeated for each training dataset until the

overall error value drops below a certain predetermined threshold (J. Wang, Lin, and Hou 2015). Table 2.5 shows the survey results of various articles collected for the dataset group skin detection with all its sub-criteria.

Table 2.5

Error Rate within Dataset Group for Skin Detection Approach

No	Author & year	Class	Error Rate		Notes
			Turning	Validation	
1	**Mahmoodi and Sayedi 2015**	Adaboost	√	×	Does not mention procedure of error rate in training
2	**Mahmoodi and Sayedi 2014a**	Bayesian	√	×	Does not mention procedure of error rate in training
3	**Priya, Vasuhi, and Vaidehi 2015**	Adaboost	√	×	Does not mention procedure of error rate in training
4	**Wen, D., Han, H., & Jain 2015**	SVM	√	√	Does not mention procedure of error rate in training and validation
5	**Thaweekote, V., Songram, P., & Jareanpon 2013**	Adaboost	√	×	
6	**Saxen, F., & Al-Hamadi 2014**	Bayes	√	√	Does not mention procedure of error rate in training and validation
7	**Lin, H. I., Hsu, M. H., & Chen 2014**	CNN	√	√	Does not mention procedure of error rate in training and validation
8	**Al-Mohair, H. K., Saleh, J. M., & Suandi 2015**	ANN\ GA	√	×	Does not mention procedure of error rate in training
9	**Kawulok et al. 2014a**	ANN	√	×	
10	**Szkudlarek and Pietruszka 2015**	Fuzzy	×	×	Does not use error rate within dataset criteria
11	**Sanmiguel and Suja 2013**	ANN\ Bayes	√	×	Does not mention procedure of error rate in training

(*Continue*)

Table 2.5 (*continued*)

No	Author & year	Class	Error Rate		Notes
			Turning	Validation	
12	**Lin, Leng, and Yu 2013**	Bayesian	√	×	
13	**Soran, B., Hwang, J. N., Lee, S. I., & Shapiro 2012**	SVM	×	√	Does not mention procedure of error rate in validation
14	**Zaidan, A. A., Ahmad, N. N., Karim, H. A., Larbani, M., Zaidan, B. B., & Sali 2014a**	ANN	√	√	
15	**A. A. Zaidan et al. 2014b**	ANN	√	√	
16	**Anghelescu, Serbanescu, and Ionita 2013**	ANN	√	×	Does not mention procedure of error rate in training
17	**Stergiopoulou, E., Sgouropoulos, K., Nikolaou, N., Papamarkos, N., & Mitianoudis 2014**	Bayes	√	×	Does not mention procedure of error rate in training
18	**Cheng, J., & Liu 2015**	FSVM	×	√	Does not mention procedure of error rate in validation
19	**Lee, D., Wang, J., & Plataniotis 2014**	Bayesian	√	√	
20	**Yang, Z., Zhu, Y., & Yuan 2014**	AdaBoost	√	×	Does not mention procedure of error rate in training
21	**Molina, J., Escudero-Viñolo, M., Signoriello, A., Pardàs, M., Ferrán, C., Bescós, J., ... & Martínez 2013**	SVM	√	√	Does not mention procedure of error rate in training and validation
22	**Khan, Hanbury, Sablatnig, et al. 2014a**	Bayesian	√	×	Does not mention procedure of error rate in training

(*Continue*)

Table 2.5 (*continued*)

No	Author & year	Class	Error Rate		Notes
			Turning	Validation	
23	**Khan, R., Hanbury, A., Stöttinger, J., Khan, F. A., Khattak, A. U., & Ali 2014b**	Bayesian	√	×	Does not mention procedure of error rate in training
24	**Chen et al. 2016**	ANN	√	×	Does not mention procedure of error rate in training
25	**Chuang, Y., Chen, L., & Chen 2014**	SVM	√	√	Does not mention procedure of error rate in training and validation
26	**Gor, A. K., & Bhatt 2015**	SVM	√	√	Does not mention procedure of error rate in training and validation
27	**Mahmoodi, M. R., & Sayedi 2014b**	Adaboost	√	×	Does not mention procedure of error rate in training
28	**Gupta, A., & Chaudhary 2016**	ANN	√	√	Does not mention procedure of error rate in training and validation
29	**Z. Li et al. 2015**	Adaboost	√	×	
30	**Pengyu, N., & Jie 2013**	Adaboost	√	×	Does not mention procedure of error rate in training
31	**Tsitsoulis, A., & Bourbakis 2013**	SVM	√	×	Does not mention procedure of error rate in training
32	**Xu, T., Wang, Y., & Zhang 2012**	ANN	√	×	
33	**Scherbaum et al. 2013**	Adaboost	√	√	Does not mention the procedure of error rate in training and validation
34	**Esposito, L. G., & Sansone 2013**	SVM	√	√	Does not mention the procedure of error rate in training and validation
35	**Khan et al. 2012**	SVM	√	√	Does not mention the procedure of error rate in training and validation
36	**Liew and Yairi 2014**	Adaboost	√	√	Does not mention the procedure of error rate in training and validation

(*Continue*)

Table 2.5 (*continued*)

No	Author & year	Class	Error Rate		Notes
			Turning	Validation	
37	**Kawulok, Kawulok, and Nalepa 2014**	Bayesian	√	×	Does not mention the procedure of error rate in training
38	**Taimori and Behrad 2015**	Fuzzy	√	×	Does not mention the procedure of error rate in training
39	**Kawulok, M., & Nalepa 2014c**	SVM	√	√	Does not mention the procedure of error rate in training and validation
40	**Yan, C. C., Liu, Y., Xie, H., Liao, Z., & Yin 2014**	SVM	√	×	Does not mention the procedure of error rate in training
41	**Yu, Cheng, and Lee 2013**	Naïve Bayes	√	√	
42	**S. Chen et al. 2013**	SVM	√	√	Does not mention the procedure of error rate in training and validation
43	**Zafeiriou, S., Zhang, C., & Zhang 2015**	CNN	√	×	Does not mention the procedure of error rate in training
44	***Tan, W. R., Chan, C. S., Yogarajah, P., & Condell 2012***	*Bayesian*	√	×	*Does not mention the procedure of error rate in training*
45	**A.A. Zaidan, Karim, and Ahmad 2010a**	Bayesian	√	×	Does not mention the procedure of error rate in training
46	**Zaidan, A. A., Karim, H. A., Ahmad, N. N., Alam, G. M., & Zaidan 2010**	Fuzzy	√	×	Does not mention the procedure of error rate in the training
47	**Nguyen et al. 2014**	SVM	√	×	Does not mention procedure of error rate in training
48	**Grzejszczak, T., Kawulok, M., & Galuszka 2016**	ANN	×	×	Does not use error rate within dataset
49	**Xu, Wang, and Zhang 2013**	Bayesian	√	×	Does not mention procedure of error rate in training and validation

(*Continue*)

Table 2.5 (*continued*)

No	Author & year	Class	Error Rate		Notes
			Turning	Validation	
50	**Shoyaib, Abdullah-Al-Wadud, and Chae 2012**	Bayesian	√	×	
	Average		46	19	
	Percentage		92 %	38 %	

Table 2.5 also shows the last group of evaluation and benchmarking criteria and sub criteria for skin detectors. Results show 65% percentage ratio for two basic sub-criteria, namely, training and validation. Moreover, the variations in the behaviour for sub criteria in the initial analysis of aforementioned studies are presented. In particular, some studies use criteria following a clear procedure, whereas others do not. Accordingly, the disparity of the ratios and behaviour clearly indicates a challenge during performance evaluation. Therefore, the disparity is primarily due to the absence of a clear and comprehensive evaluation method within a uniform guideline. Moreover, each study that used the error rate criterion did not provide reference to a specific level that can be compared with those of other studies.

2.3 Benchmarking Techniques/ Tools

According to (Oxford dictionaries. (2013), benchmarking is a standard or point of reference against which things may be compared. In particular, it involves the comparison of the performance of an approach or technique against external criteria. Benchmarking in IT and computer systems involves the comparison of the output values of different systems for a given set of criteria to ensure the quality, improvement, contribution, or performance of the new system (Trentesaux et al. 2013).

Exploring several assessments, evaluation, or benchmarking for various applications of skin detection usually results in an incomplete benchmarking process (Han et al. 2013). Numerous skin detectors have been proposed and developed, and these detectors were compared using various skin detection applications on the basis of their target. However, all applications were designed and assessed within a particular domain to measure the reliability of the skin detectors (ZAIDAN et al. 2013). The high percentage of reliability of a skin detection technique does not necessarily imply an optimal skin detector. Other criteria should also be considered (Trigueiros, Ribeiro, and Reis 2013). Accordingly, not all developers focused on the reliability of skin detectors because of the objective in developing these detectors. For example, reliability is desired not on account of time complexity applications, such as video surveillance systems (Hsieh and Hsu 2008; Venetianer and Deng 2010; De-La-Torre et al. 2015). Several of the key factors influencing the performance, importance, and challenges in selecting the appropriate measures and the difficulties and

mechanisms of creating the ground truth have been described. These researchers have attempted to overcome the limitations of standard approaches by providing a suitable environment. Accordingly, the development of various applications in the skin detection domain is probably based on the development of the skin detectors. Meanwhile, similar to time complexity, reliability is important in desktop computation approaches and should be considered in evaluation and benchmarking.

To the best of our knowledge, evaluation and benchmarking techniques have not been implemented for skin detection or other image processing applications. However, tools for simulating the results on different image processing applications are available. For example, skin detectors with different machine learning techniques can be performed, and the results depend on the reliability group for each individual method. These tools are considered as one of the challenges in the benchmarking of the skin detection, because many tools do not meet the requirements of the skin detection. Two groups used to implement skin detection are discussed below.

2.3.1 Data Mining Tools Group

Data mining how various data are included in the best platform is necessary because of the numerous commercially available tools and techniques. Recognizing which tool or approach is suitable is also essential (Han, J., Pei, J., & Kamber 2011; Brown and White 2017). The following benchmarking will explain the full documentation on the basis of the type of platform and application evaluated.

A) **Rapid Miner Tool**

Rapid Miner is a popular and frequently used open-source tool worldwide. The project was established at the University of Dortmund in 2001 and has been developed further by Rapid-I GmbH since 2007. Rapid Miner is a tool for machine learning, data mining (DM), image processing, predictive analytics, and business analytics. This tool has been applied by universities and researchers in various fields (Land, Sebastian 2012; Dwivedi, Kasliwal, and Soni 2016).

Rapid Miner currently covers most of the commonly required DM tasks, especially in structured DM. For example, extension adds the capability for some advanced image mining tasks in image processing. Burget et al. (2010) used this tool and obtained highly reliable results without reference to the time factor. However, other studies have reported acceptable reliability rates because of the time complexity of image detection (Pujol and García 2012; Ashwin Satyanarayana 2013).

B) Waikato Environment for Knowledge Analysis (WEKA) Tool

WEKA is a tool for machine learning and DM. This tool was developed by the Department of Computer Science at the University of Waikato in New Zealand and was first implemented in its modern form in 1997. WEKA uses the GNU General Public License. The software is written in Java language and contains a GUI for

interacting with data files and producing visual results (Ashwin Satyanarayana 2013; Dwivedi, Kasliwal, and Soni 2016).

WEKA is used as a reference system for feature selection. Thus, it assumes complete knowledge of the problem on the basis of available dataset in schemes. The objective of feature selection is to apply to a dataset consisting of the entire image database during image retrieval (Grigorova et al. 2007). The WEKA attribute selection tool is applied separately for both datasets using a correlation-based feature selection among different algorithms. This tool exhibits high reliability when tested on a basic dataset (Ochoa, Yayilgan, and Cheikh 2012; Al-odan and Saud 2015).

C) R Tool

R has been improved in the last 15 years as a statistical language and open-source tool. This tool was originally developed by Bell Labs in the 1970s. The source code of R is written in C++, FORTRAN, and R itself. R is an interpreted language and is mostly optimized for matrix-based calculations. Its performance is comparable with those of commercially available MATLAB and freely available GNU Octave. The main language is extended by a myriad collection of packages for all types of computational tasks. The tool offers a simple GUI with a command-line shell for input. The R tool is not a user-friendly environment because all commands need to be entered in the R language. The key advantage of the R tool is its rapid implementation of numerous machine learning algorithms; the number of algorithms is comparable

with those of RapidMiner, WEKA (from which a large number of algorithms is borrowed), and with that of the full prospect of statistical data visualization methods (Jovic, A., Brkic, K., & Bogunovic 2014; Dwivedi, Kasliwal, and Soni 2016).

D) Konstanz Information Miner (KNIME) Tool

KNIME is a software-platform open-source tool for integration, processing, and analysis of data. This tool is designed to import and transform large datasets for convenient use. KNIME has been used to provide basic image processing, such as image input and output, and standard threshold algorithms. Image segmentation consists of several customized KNIMEs that are combined with standard image processing. This concept allows the processing of numerous image segmentations and automatically saves the results with high accuracy. The processing time per image segmentation is in the range of 1–2 min per segment during computer operation. The processor then implements the registration and processing stages, which are time consuming (Riess et al. 2011; Dwivedi, Kasliwal, and Soni 2016).

E) Orange Tool

Orange is a Python-based tool for DM that is being developed at the Bioinformatics Laboratory of the Faculty of Computer and Information Science at the University of Ljubljana (Jovic, A., Brkic, K., & Bogunovic 2014). Orange is a library of C++ and routines, which also include various standard and non-standard machine learning, DM

algorithms, routines for data input, and manipulation. This tool has a comprehensive and component-based framework for machine learning and DM. Orange is a powerful and easy-to-use visual programming environment (Zupan, B., & Demšar 2004).

Orange is a suitable platform for different classification algorithms. After the datasets are selected, several classification algorithms are chosen to conduct the test. The accuracy of a classifier in a given test set is indicated by the percentage of the test set tuples correctly classified by the classifier. Thus, the tool provides results of varying accuracies for different classifier algorithms under a specific dataset (Wahbeh, A. H., Al-Radaideh, Q. A., Al-Kabi, M. N., & Al-Shawakfa 2011).

F) Knowledge Extraction based on Evolutionary Learning (KEEL) Tool

KEEL tool is non-commercial and supported by Java language. The first version was released in 2004, and the latest version is KEEL 2.0. This tool empowers the user to analyse the behaviour of evolutionary learning for different kinds of DM problems, such as regression, classification, and unsupervised learning (Alcalá-Fdez et al. 2009; Alcalá-Fdez et al. 2011).

The KEEL tool is useful for different types of users and has specific features in the DM software. However, most DM tools either have basic support or no support at all for two types of re-processing, namely, statistical tests and evolutionary algorithms (EAs). EAs have become important techniques for learning and knowledge

extraction, particularly in genetic algorithms. Therefore, EA has been extensively used to solve problems, such as image retrieval. Various machine learning and pattern recognition techniques have been applied to relevance feedback in information retrieval in general. Therefore, the major imbalance between text retrieval and image retrieval is probably in the application of EAs. This issue has been computed over the training images using either retrieval precision or average retrieval rank of positive images. This method is suitable for learning with small sample sizes (Stejić, Takama, and Hirota 2006).

2.3.2 Computer Vision Tools Group

This section discusses various tools operated within the computer vision field. The features of these tools are used to evaluate the type of applications within the computer vision domain. Computer vision includes general and specific types of tools (Shah, S. A. H., Ahmed, A., Mahmood, I., & Khurshid 2011). Several benchmarking tools are presented below.

A) PhotoScan Tool

Scientific research in the last decade has increasingly moved toward automated procedures using computer vision approach to reduce time complexity during data processing. Numerous surveys have been conducted using image processing techniques. PhotoScan package has been used in computer vision interfaces for 3D

web services and low-cost software. It is a low-cost tool for high-quality image processing. This software is based on a multi-view reconstruction technology that can operate with calibrated and uncelebrated images under controlled and uncontrolled conditions. Therefore, PhotoScan has become a promising tool in image processing for investigations with a limited budget. Various processes can be performed using this tool resulting in different levels of accuracy, and many parameters can be set to improve the final result (Shah, S. A. H., Ahmed, A., Mahmood, I., & Khurshid 2011).

B) OpenCV Tool

OpenCV was originally an Intel research initiative. This tool is a cross-platform open-source computer vision framework employed in real-time image processing. OpenCV was developed primarily in C++ but lacks API datasets or integrated analysis utilities. For instance, hand gesture application considers a methodology to recognize gestures. OpenCV is generally used for real-time applications because it requires less computational time for hand gesture processing. However, this method utilizes a training set of images that consists of positive and negative examples (hands and non-hands, in this case) with high accuracy (Shah, S. A. H., Ahmed, A., Mahmood, I., & Khurshid 2011).

C) Hypr3D Tool

Hypr3D is a valuable method for computer vision that works directly with images or videos online. The procedure is implemented by first choosing a file format to upload (images or videos). At least five images are uploaded to create a 3D model. The process computes the camera parameters and produces a point cloud, a wireframe model, and a texturing high-resolution model. The model can be downloaded in different formats and resolutions. Several tests were also performed with different sets of photos to evaluate the image that can obtain a complete and high-resolution 3D model using Hyper3D.

Basically, 3D maps for Hypr3D result in maximum deviations of ± 0.06 m situated in critical areas of the images. The 3D maps show the distribution of deviations among different models through a color scale. The deviations represent the shortest distance from the reference model in 3D. Thus, the distribution of deviations is not uniform for all models (Mery, D., Pedreschi, F., & Soto 2013).

D) Balu Tool

Balu for computer vision, pattern recognition, and image processing was developed by the Group of Machine Intelligence from the University of Catolica, in Chile (Mery, D., Pedreschi, F., & Soto 2013). For instance, feature extraction is a critical step in image classification. A comparison protocol with several features extracted for

techniques under different classifiers is implemented. The feature extraction performance of techniques in the context of image classification for both binary and multi-class classifications is evaluated. Balu is used to evaluate and analyze the performance measures of the results, including classification accuracy rate, precision, recall, F-measure, and G-mean. The classifier systems should split the dataset into training and test sets. The results showed the relevant feature extraction technique that improved the classification accuracy rate under machine learning algorithms (Medjahed 2015).

E) 123D Catch Tool

Developing 3D models from photographic images is an efficient and intuitive way to create 3D digital models of objects. 123D Catch software is practical for desktop systems with web-based software. This tool will overcome the dramatically slow and hardware-heavy nature of computer approaches

123D Catch by Autodesk is a widely used web-based package. It was selected because of its easy operation, visual quality of the reconstructed scene, and the possibility to interact with and develop the results (Santagati and Laura. 2013). Table 2.6 lists the limitations of various tools that fall within Data mining and computer vision.

Table 2.6

Summary of Weaknesses of the Tools

Tools	Weakness
Raped Miner (Land, Sebastian 2012; Burget et al. 2010; Kozielski, Sikora, and Wróbel 2015; Pujol and García 2012; (Jovic, A., Brkic, K., & Bogunovic 2014); (Verhoeven, G., Sevara, C., Karel, W., Ressl, C., Doneus, M., & Briese 2012)	• This tool does not calculate the consumed time during implementation. • Semantic analysis of data mining processes is unavailable. • It focuses on each reliability element only. • Rapid Miner lacks the free version support to connect to the MySQL databases. • It has limited support for deep learning methods and some of the more advanced specific machine learning algorithms.
Weka (Grigorova et al. 2007;Ochoa, Yayilgan, and Cheikh 2012; Wahbeh, A. H., Al-Radaideh, Q. A., Al-Kabi, M. N., & Al-Shawakfa 2011; Kosorus, Honigl, and Kung 2011;Al-odan and Saud 2015b)	• It has worst efficiency connected to Excel spreadsheet and databases. • CSV reader is not robust compared with Rapid Miner. • WEKA is much weaker in classical statistics. • This tool does not have the facility to save parameters for scaling to apply to future datasets. • WEKA has no automatic facility for parameter optimization of machine learning/statistical methods. • It lacks many data survey and visualization methods. • It has limited support for large data, text mining, and semi-supervised learning.

(*Continue*)

Table 2.6 (*continued*)

Tools	Weakness
R (Jovic, A., Brkic, K., & Bogunovic 2014; Torgo 2010; Rangra and Bansal 2014; Al-odan and Saud 2015b)	• This tool is less specialized toward data mining. • Steep learning curve is obtained, unless user is familiar with array languages. • This tool uses challenging language. It is also difficult to learn thoroughly enough to become productive in DM. • R documentation was criticized because the developer did not spend effort on building a knowledge base for beginners but instead focused on advanced users. • R can handle all data sources, except Microsoft Excel.
KINME (Riess et al. 2011; Berthold et al. 2009; Jovic, A., Brkic, K., & Bogunovic 2014; Rangra and Bansal 2014; Al-odan and Saud 2015)	• It has only limited error measurement methods. • It has no wrapper method for descriptor selection. • This tool has no automatic facility for parameter optimization of machine learning/statistical methods.
Orang (Jovic, A., Brkic, K., & Bogunovic 2014; Zupan, B., & Demšar 2004; Wahbeh, A. H., Al-Radaideh, Q. A., Al-Kabi, M. N., & Al-Shawakfa 2011; Rangra and Bansal 2014;Al-odan and Saud 2015)	• This tool is not super polished. • Orange has a large installation file, with limited reporting capabilities. • This tool has a limited list of machine learning algorithms. • Machine learning is not handled uniformly among different libraries. • Orange is weak in classical statistics and provides no widgets for statistical testing. • Orange can handle all data input except sources from Microsoft products.

(*Continue*)

Table 2.6 (*continued*)

Tools	Weakness
KEEL (Alcalá-Fdez et al. 2009; Alcalá-Fdez et al. 2011; Graczyk 2009; Rangra and Bansal 2014)	• Its efficiency is restricted by the number of algorithms. • It has support compared with other tools.
PhotoScan (Brutto and Meli 2012; Verhoeven 2011; Verhoeven, G., Sevara, C., Karel, W., Ressl, C., Doneus, M., & Briese 2012; Koutsoudis, Vidmar, and Arnaoutoglou 2013; Cerrillo-Cuenca and Sepúlveda 2015)	• The automatic procedure control is difficult with this tool. • This tool has insufficient tools to check the 3D reconstruction. • The reconstruction process may fail when the used images do not meet the requirements for the processing under several tests. • A remarkable disadvantage is the absence of editing commands that allow to scale and to geo-reference the 3D model directly in the web services. • An absence of parameters to verify the correct image orientation. • Its accuracy is in the centimeter scale, which is too low for high-accuracy applications. • The 3D maps resulting from the metric evaluation show an uneven distribution of the deviations.
OpenCV (Rautaray, S. S., & Agrawal 2012; Shah, S. A. H., Ahmed, A., Mahmood, I., & Khurshid 2011; Marengoni and Stringhini 2011; Bradski, G., & Kaehler 2008;Anjos, A., El-Shafey, L., Wallace, R., Günther, M., McCool, C., & Marcel 2012)	• It has become hardest only because of the absence of proper documentation and error handling codes for this tool. • It has a small set of machine learning algorithms. • Debugging and visualizing is difficult in any C++ environment. • This tool lacks the dataset APIs or integrated analysis utilities.

(Continue)

Table 2.6 (continued)

Tools	Weakness
Hypr3D (Brutto and Meli 2012; (Gede, M., & Mészáros 2013)	• It has limited accuracy. • It has a problem with straight edges and flat surfaces represented as minor faults. • It has limited scalability. • Hypr3D has a very limited resolution and unwanted.
Balu (Medjahed 2015; (Mery, D., Pedreschi, F., & Soto 2013)	• It has a defect in computational time in the feature extraction and model selection phase. • It needs to process load similar to the other approaches when suitable features and classifier are selected. • It is worth to consider that the intensive exploration process is performed once and offline.
123D Catch (Santagati, C., & Inzerillo 2013; Kersten and Lindstaedt 2012; Brutto and Meli 2012)	• Photos dataset should be structured. • It cannot manage the overlapping between two frames in height. • It is not very reliable from metrical point of view. • It affected by errors exceeding 3-5 times the fault tolerance. • The automatic detection of homologous points on different images causes the creation of scattered point cloud models from which digital surface model (DSM) is derived. • It does not fully respond to the real monument.

Table 2.6 presents various tools reported in the literature. These tools are related to the skin detection. The tools presented above are designed to measure several criteria of skin detection. Data mining and computer vision tools are generally used. In other

words, most skin detection applications have not been benchmarked using existing tools, and only some applications were benchmarked using the tools discussed above. Their limitations were also highlighted. Several studies have presented the areas for improvement of the evaluation tools in relation to criteria variation, especially for a reliable skin detection method. Therefore, the existing tools cannot satisfy the overall needs of skin detectors in the benchmarking process, as follows: (1) benchmarking between two or more techniques, (2) individual calculation of error rates, (3) matching among techniques, (4) calculation of the time consumption of these techniques, and (5) calculation of the overall parameters of the reliability group. Therefore, improve the performance of tools by its development to include all the basic criteria for evaluation and benchmarking as in reliability, time complexity, and error rate in order to meet all the requirements of the skin detector applications.

2.4 Open Issues and Challenge for Evaluation and Benchmarking Process

In our study, there are some limitations and challenges related with the evaluation and bechmarking process according to the research problem (Lan et al. 2013; Liu et al. 2013; Mahmoodi and Sayedi 2016). Benchmarking in skin applications is limited to reliable skin detection. Essentially, benchmarking is based on a comparison of the new generation with others under the conditions and criteria to be considered after the development process for any system development. The main challenge to the development of skin detection is that the developers focus on either increasing reliability while maintaining a low error rate or decreasing time complexity only.

Such approach frequently affects the results of the skin detection system, and high reliability and low rate (time complexity rate or error rate) cannot be obtained simultaneously. Accordingly, this tradeoff is reflected in the benchmarking process. Studies often face conflicts among various criteria during benchmarking, resulting in a major challenge because measuring other criteria creates a set of numbers representing different criteria. In addition, a tradeoff among the criteria creates another problem with which developers would not be able to compare the new approach among other approaches. Therefore, cases affecting the benchmarking process conducted among different criteria results should be avoided (Kozielski, Sikora, and Wróbel 2015; Al-odan and Saud 2015; Cerrillo-Cuenca and Sepúlveda 2015; Rautaray and Agrawal 2015 ;Madeo, R. C., Lima, C. A., & Peres 2017).

2.4.1 Concern for Evaluation Criteria

Evaluation criteria for skin detection received several criticisms in relation to the evaluation matrix. Several criticisms have been made on the evaluation criteria, namely, error rates within dataset metrics of images, particularly in training and validation and in the reliability group of evaluation criterion. In the error rates within dataset criticism, a problem figure exists on the variation of error rate values resulting from the variation in the size of the datasets used in different skin detection experiments. Thus, the lack of a standard dataset causes significant problems when error rates in numerous experiments are considered. However, many studies collect datasets on the basis of their individual studies, which therefore leads to an

unnecessary consumption of effort and time. Given that the criticisms on the reliability group are based on the matrix of parameters (TP, FP, TN, and FN), some pixels will be lost after cropping the background from the skin images using Adobe Photoshop, when the actual class needs to be labeled manually and compared with the predicted class to compute one of the matrices of parameters. This process is debatable because it will affect the results from all reliability groups (matrix of parameters, relationship of parameters, and behavior of parameters). Although these metrics are heavily criticized in the literature, considerable studies still use them to evaluate skin detection and other domains of image processing. Moreover, each study used reliability, time complexity rate, or error rate but without reference to a specific level to be compared with those of other criteria (Kruppa, Bauer, and Schiele 2002; Sajedi and Jamzad 2007; Gasparini, Corchs, and Schettini 2008; Gen Li et al. 2010; Belaroussi and Milgram 2012; Tan, W. R., Chan, C. S., Yogarajah, P., & Condell 2012; Lan et al. 2013; Liu et al. 2013).

2.4.2 Concern for Criteria Trade-off

A tradeoff is a situation where losing one reliability or aspect of something results in gaining another reliability or aspect and vice versa. In our literature review, numerous aspects of tradeoff among different criteria used by researchers were found. Tradeoffs among criteria typically causes confusion for decision makers. The varying ratios among the different criteria collected in our study also showed the effect of the conflict on various criteria used by researchers. The conflict among evaluation criteria

for skin detection clearly indicates a challenge in our intention to develop a skin detection segmentation/classification approach. Basically, this challenge comes from conflicting terms, namely, the conflict among the criteria and among the data. Realizing the advantage and disadvantage of a particular choice is thus crucial in decision making. The term "tradeoff" is widely used in an evolutionary context, in which the selection process acts as the "decision-maker".

Conflicting criteria or tradeoff problems among reliability, time complexity of skin detection, and error rate within the dataset in the evaluation and benchmarking of skin detection are reported in the aforementioned studies. These criteria are the main requirements that should be measured to evaluate skin detection. The reliability should be high, the time complexity for conducting the output images should be low, and the error rate resulting from the training datasets should be low. Conflict in data generated is observed because the section matrix of parameters contains TP, FP, TN, and FN, which indicate the rise in TP and TN when FP and FN are reduced. This phenomenon shows a conflict among the probability criteria. These parameters significantly affect the rest of the criteria values within the reliability group. Therefore, evaluation and benchmarking processes should consider these requirements. All the reviewed studies have proven that evaluation and benchmarking of each criterion is as an independent. A skin detection classification approach should therefore be conducted to standardize the basic and advanced requirements, and an effective methodology for testing, evaluation, and benchmarking should be implemented during research. A new evaluation method should be flexible to handle

the conflicting criteria and data problems. However, to the best of our knowledge, no solutions have been suggested on these particular issues thus far (Sigal, Sclaroff, and Athitsos 2000; Pham, Worring, and Smeulders 2002; Jaiswal 2011; Mahmoodi and Sayedi 2016).

Table 2.7

Trade-off Problem in the Academic Literature

Author	Trade-offs problem?	Provide solution
Kruppa, H., et al. 2000	✓	×
Sigal, L., et al.2000	✓	×
Pham, T.V.et al. 2002	✓	×
Sajedi, H. and Jamzad, M. 2007	✓	×
Gasparini, F., et al. 2008	✓	×
Gen LI.et al.2010	✓	×
Jaiswal, S. 2011	✓	×
Belaroussi, R. and Milgram, M. 2012	✓	×
Tan, W. R., et al. 2012	✓	×
Lan, Z., et al.2013	✓	×
Liu, J.,et al.2013	✓	×
Mahmoodi, M.R. and Sayedi, S.M. 2016	✓	×

2.4.3 Concern for Criteria Importance

Studies on skin detection have numerous objectives during the planning stage. These objectives are reflected in the system design, system evaluation, and system

benchmarking. The importance of criteria is a key objective in this study through evaluation and benchmarking, despite the conflict among them. Therefore, the conflict among the criteria constitutes a major challenge during evaluation. A suitable procedure should be developed for these objectives when increasing the importance of a particular evaluation criterion and reducing others. Two main aspects should be considered. First, the behavior of skin detectors should be understood, giving particular importance to the design. Second, the evaluation of the approach should consider the tradeoff (Rautaray and Agrawal 2015). Nevertheless, the opinions of evaluators might conflict with the objectives of the designer, which can affect the final evaluation of the new approach (Ghaziasgar, Connan, and Bagula 2016; Mahmoodi and Sayedi 2016; Zuo et al. 2017). Technically, skin detection during evaluation and benchmarking processes involves simultaneous consideration of multiple attributes (reliability, time complexity, and error rate within the dataset) and assigns the proper weight for each feature to benchmarking the skin detection techniques.

Approaches with the highest balancing rate should receive the highest priority levels, whereas approaches with the least balancing rate should be given the lowest priority levels compared with the scores of other approaches. However, evaluation and benchmarking processes are difficult and challenging tasks because each approach for skin detection has multiple attributes that should be considered. For example, time complexity and error rate within the dataset have been proven to be critical in skin detection because they provide an objective complement to the

decision of skin detection and optimize inter-rater consistency. Consequently, each decision maker gives different weights for these attributes. On one hand, a developer who aims to give a score for a skin detection approach might give more weight to the feature rather than to other features that gain less interest than these attributes. By contrast, developers who aim to use benchmarking software to solve this problem would probably target different attributes as the most important attribute. Thus, evaluation and benchmarking processes of skin detection approaches has a multi-complex attribute problem, such that each approach is considered an available alternative for the decision maker (Mahmoodi and Sayedi 2016).

2.5 Theoretical Background about Multi Criteria Decision Making Techniques

Useful techniques that deal with multi-criteria decision-making (MCDM) problems are recommended solutions that collectively help decision makers organize the problems to be solved and conduct analyses, comparisons, and ranking (Jadhav, A. S., & Sonar 2009a; Jadhav, A. S., & Sonar 2009b).The goals of MCDM are as follows: (1) help decision maker to choose the best alternative, (2) categorize the viable alternative among a set of available alternatives, and (3) rank the alternatives in descending order of performance (A. A. Zaidan, Zaidan, et al. 2015c; Jadhav, A. S., & Sonar 2009a; Jadhav, A. S., & Sonar 2009b). Accordingly, the suitable alternative(s) will be scored. The fundamental terms in any MCDM ranking should be defined, including the DM or the EM, as well as its criteria (Nedher, A. S., Hassan, S., & Katuk 2014; Whaiduzzaman et al. 2014). An evaluation matrix consists of *m*

alternatives and *n* criteria that need to be created. The intersection of each alternative and criteria is given as x_{ij}. Therefore, we have a matrix (x_{ij})_(*m***n*) expressed as follows:

$$D = \begin{array}{c} \\ A_1 \\ A_2 \\ \vdots \\ A_m \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \dots & C_n \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix} \end{array} \quad \textbf{(2.10)}$$

where $A_1, A_2, \dots, A_m$ are possible alternatives that decision makers have to score (i.e., skin detection approaches). $C_1, C_2, \dots, C_n$ are the criteria against which the performance of each alternative is measured (i.e., reliability, time complexity, and error rate within the dataset). Finally, x_{ij} is the rating of alternative A_i with respect to criterion C_j, and W_j is the weight of criterion C_j. Certain processes should be completed to rank the alternatives, such as normalization, maximization indicator, adding weights, and other processes depending on the method.

For example, suppose that *DM* is the decision matrix used to rank the performance of the alternative A_i, where *i*= {1, 2, 3 and 4} based on C_j (j = {1, 2, 3, 4, 5 and 6}). Table 2.8 is an example of an MCDM problem reported by (Zaidan, B. B., Zaidan, A. A., Abdul Karim, H., & Ahmad 2017).

Table 2.8

Example of Multi-Criteria Problem

Ci / Aj	C_1	C_2	C_3	C_4	C_5	C_6
A_1	2	1500	20000	5.5	5	9
A_2	2.5	2700	18000	6.5	3	5
A_3	1.8	2000	21000	4.5	7	7
A_4	2.2	1800	20000	5	5	5

The data in the chart is not easy to evaluate because of the large numbers of C_2 and C_3 (Figure 2.4).

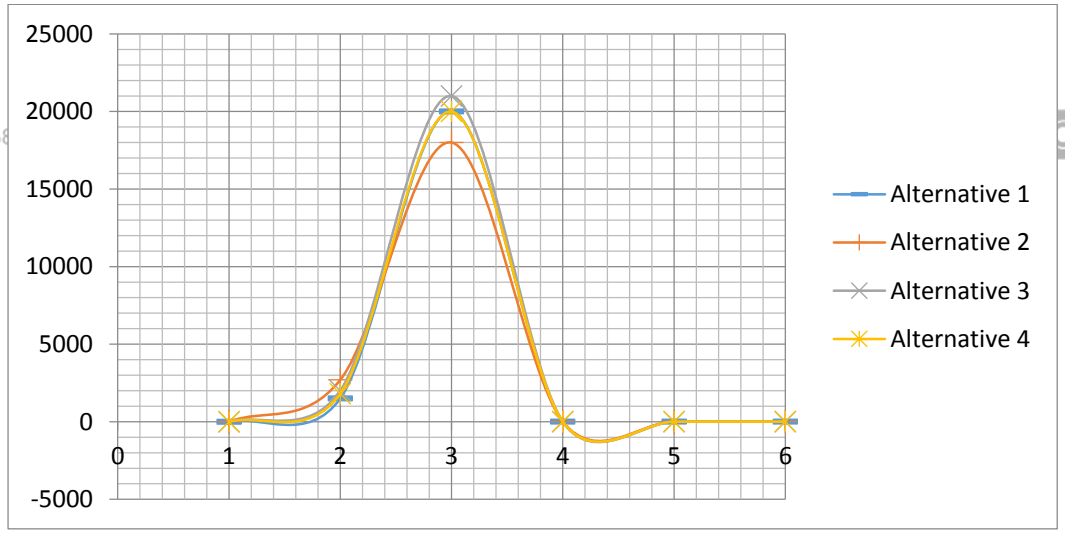


Figure 2.4. Presentation of the Example in Table (2.8)

2.5.1 Analytical study of MCDM Techniques

Several MCDM theories or methods have been investigated (B. B. Zaidan et al. 2017a; B. B. Zaidan et al. 2017b). The most popular methods of MCDM using

different concepts include Multiplicative Exponential Weighting (MEW), Weighted Product Method (WPM), Weighted Sum Model (WSM), Simple Additive Weighting (SAW), Hierarchical Adaptive Weighting (HAW), Analytic Hierarchy Process (AHP), Analytic Network Process (ANP), and TOPSIS (Gayatri, V., & Chetan 2013; Aruldoss, M., Lakshmi, T. M., & Venkatesan 2013). To the best of our knowledge, none of these methods are used to evaluate and benchmark the approaches of skin detection.

Triantaphyllou, E., Shu, B., Sanchez, S. N., & Ray (1998), (Aruldoss, M., Lakshmi, T. M., & Venkatesan (2013), Whaiduzzaman et al. (2014), B. B. Zaidan et al. (2017a), B. B. Zaidan et al. (2017b) summarized the benefits, drawbacks, and recommendations for popular methods of MCDM techniques as follows. WSM and HAW techniques are easy to use and understand, but the weights of the attributes are arbitrarily assigned in these techniques. Thus, both techniques become difficult to use with an increasing number of criteria. Another limitation of these methods is caused by the use of common numerical scaling to obtain the final score. SAW considers all criteria, makes decisions intuitively, and provides simple calculation. All criteria should be maximum and positive. However, SAW does not commonly reflect the actual situation. The strengths of MEW and WPM include the ability to remove any unit of measure and the use of relative values rather than actual ones. Nevertheless, no solution with an equal weight of decision matrices is offered. AHP enables DMs to hierarchically arrange a decision-making problem, which helps in understanding and simplifying the problem. However, this technique is time consuming because of the

required mathematical calculations and number of pairwise comparisons, which increase as the number of alternatives and criteria increases or changes. Ranking in AHP depends on the alternatives considered for evaluation. Adding or deleting alternatives can change the final ranking (rank-reversal problem). TOPSIS is functionally associated with problems of discrete alternatives. This technique is practical for solving real-world problems. TOPSIS is relatively advantageous because it can rapidly identify the most suitable alternative. By contrast, the major weakness of TOPSIS is its lack of provision for weight elicitation and consistency checking for judgments. The use of AHP is significantly limited by the human capacity for information processing. Thus, 7±2 is regarded as the ceiling for comparison (T.L. Saaty and Ozdemir 2003). The performance of ANP provides insights into the level of importance that a criterion can take consistent to its interrelationship with other elements of the model. ANP evaluates the consistency of judgments, which is not feasible when evaluating through the assignment of weights by compromise. Moreover, ANP facilitates the assignment of weights because splitting up the problem into smaller parts allows a group of studies to have a manageable discussion, such that only two criteria can be compared to assign judgments. However, ANP has two disadvantages. First, it does not provide the correct network structure among criteria even for experts, and different structures lead to diverse results. Second, all criteria have to be pairwise to form a super matrix compared with all other criteria. This process is difficult and unnatural (Al-Azab, Ayu, and Ai-azabl 2010; Thomas L. Saaty 2008).

Therefore, further hands-on analysis is required to select one of the available methods. Certain criteria should be set to allow compression among the outcome scores of each algorithm because the accuracy of each algorithm cannot be measured. Selecting the most appropriate method from several feasible alternatives is thus considered challenging. Zanakis et al. (1998) selected SAW as a benchmark for other MACM techniques and measured the following: (1) mean squared error of weights and that of ranks, (2) mean absolute error of weights and that of ranks, (3) Theil's coefficient U for weights and that of ranks, (4) Kendall's correlation tau for weights, (5) Spearman's correlation for ranks, (6) weighted rank crossing 1, (7) weighted rank crossing 2, (8) top-ranked matched count, and (9) number of ranks matched with the number of alternative L. These criteria are used to measure the differences between each algorithm and SAW from different perspectives.

The variety of MCDM techniques also create another challenge in selecting the best technique (B. B. Zaidan et al. 2017a; B. B. Zaidan et al. 2017b; Salman, O. H., Zaidan, A. A., Zaidan, B. B., Naserkalid, & Hashim 2017; Qader, M. A., Zaidan, B. B., Zaidan, A. A., Ali, S. K., Kamaluddin, M. A., & Radzi 2017). MCDM problems occur under various situations in which many DMs have several alternatives and actions or candidates should be selected in reference to a set of attributes (Zavadskas, E. K., Kaklauskas, A., Turskis, Z., & Tamošaitienė 2009). Selecting among algorithms (e.g., SAW, MEW, HAW, TOPSIS, WSM, and WPM) is thus important. Statistical comparison is an unreasonable solution given that these techniques have different scores. The total score of the final ranking should be

normalized for each technique to allow for comparison. At this stage, mean, STD, and paired sample t-tests are used in the comparison. Paired sample t-tests typically consist of the sample of matched pairs of similar units or one group of units tested twice. A paired sample t-test is based on "matched-pair sample" resulting from an unpaired sample subsequently used to form a paired sample. Additional variables measured along with the variable of interest are adopted. Matching is conducted by identifying the pairs of values that consist of one observation from each of the two samples. The pair is similar in terms of other measured variables. This approach is occasionally used in observational studies to minimize or eliminate the effects of confounding factors. In the final stage, the normalized score for the utilized MCDM techniques should be used to describe the closeness and divergence in the curve behavior per algorithm. Comparing alternatives is difficult. Thus, the top- and worst-ranking alternatives should be compared as shown by (Zanakis, S. H., Solomon, A., Wishart, N., & Dublish 1998). Curve behavior should also be studied in terms of curve irregularity, correctness of shape, and behavior.

2.5.2 Analytic Hierarchy Process (AHP) Method

Analytic Hierarchy Process (AHP) is one of the most popular multi criteria decision making methods that had been developed by Thomas L. Saaty in the seventies of the last century based on mathematics and psychology (T L Saaty 1990; T.L. Saaty and Ozdemir 2003). Figure 2.5 illustrates the initial decision in the hierarchical structure.

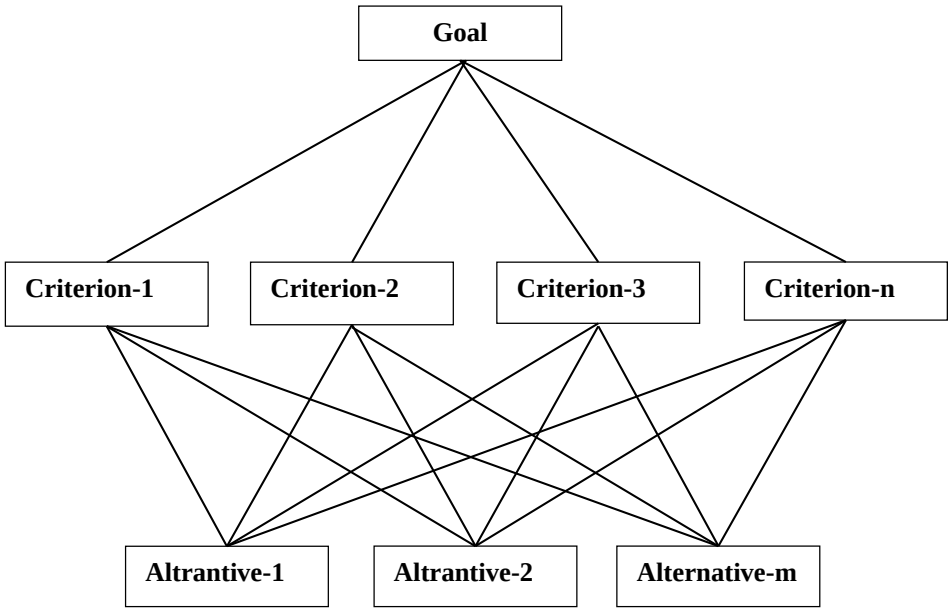


Figure 2.5. Initial Decision in the Hierarchical Structure

AHP is depended on the collection of knowledge from experts based on the phenomena under study. On the other hand, AHP method can be used by experts or teams of people are working on straightforward decisions to solve complex problems, particularly with high risks which depend on perspectives and judgments, which have repercussions on the long-term (Bhushan and Rai 2007). Typically, it adapted with the theory of fuzzy set and concept of hierarchical analysis, which considered a systematic approach to select of alternatives and justification problem. Therefore, decision-makers are more confident when providing final decisions of those fixed-value decisions (Aruldoss, M., Lakshmi, T. M., & Venkatesan 2013). Thus, this method can be applied even though sometimes the user's preferences are ambiguous or is unclear. AHP includes the opinions of experts and multi-criteria evaluation; it is not capable of reflecting human's vague thoughts. Generally, the procedure of the AHP method doing analysis different problems in the MCDM with systematic

hierarchy style, although fuzzy set theory is a capability to explain expert's perspectives through flexible comparison operation (Aldlaigan and Buttle 2002). Typically, the AHP method addresses the components of the matrix structure as m * n (where m represents a number of alternatives, while n represents the number of criteria). It depends on the relative significance of alternatives to create any matrix based on a variety of criteria. AHP method adapted concept of priority theory. Whereas, it solved a complex problems which is included a multi-criteria with different alternatives simultaneously. Thus, the AHP method extracts the weights for each criterion from any source with its features based on the pairwise comparison process. Usually, the pairwise comparison process applied to solve problems for individual decision makers in order to compare between their judgments, as well as used this method to achieve a linguistic rating method to get an absolute judgments (Aruldoss, M., Lakshmi, T. M., & Venkatesan 2013). On the other hand, the requirement of the AHP method often no general rule in the selection of the evaluator's number. At the end, AHP is a methodology based on the decision maker's relative preference to select the best one attribute than another. Due to it is not statistically adapt its methodology with small sample size is enough to make a decision (Herath, Prato, and Prato 2007; Duke and Aull-hyde 2002). Finally, the AHP method can be applied easily and does not need a large sample size (Lam, K., & Zhao 1998).

2.5.3 Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) Method

Deng, H., Yeh, C. H., & Willis (2002) briefly defined the TOPSIS method is an approach to identify an alternative which is closest to the ideal solution and farthest from the negative ideal solution in a multi-dimensional computing space. This technique was developed by Hwang and Yoon in 1981. The TOPSIS method introduces an aggregating function, including the distances from the ideal point and from the negative ideal point without considering their relative importance. However, the reference point can be a major concern in decision making and should be as close as possible to the ideal solution (San Cristóbal 2011; ur Rehman, Z., Hussain, O. K., & Hussain 2012). In addition, TOPSIS method proposed for each criterion which has increasing or decreasing to tendency towards of monotonically, thus leads to easily identify the aspects of positive and negative for the ideal solution. Thus, this method is proposed a Euclidean distance approach is the best solution when evaluating the relative closeness for each alternative at the ideal solution (Aruldoss, M., Lakshmi, T. M., & Venkatesan 2013). There are series of comparisons should be computed these relative distances which provide the preference order during selecting the best alternatives.

The main procedure in the TOPSIS method is a converts the different criteria, dimensions into non-dimensional criteria (Taqa, A. Y., & Jalab 2010). The principle of TOPSIS based on the rules of selecting a best alternative which has the shortest

distance from the positive ideal solution, as well as it has the farthest from the negative ideal solution. In addition, TOPSIS method used to rank alternative and to get the best performance between various alternatives based on multi-criteria decision making. This technique used in various applications, such as supply chain management and logistics, healthcare, design, engineering, manufacturing systems, business, marketing management, environmental management, human resource management, energy management, chemical engineering, and water resource management.

2.6 Survey of Skin Detection Models

Recently, several studies have addressed different problems of the skin detection applications. Since, skin detection considered one of the important techniques widely used in the image processing. Color is a key parameter to determine the skin pixels of the image, which is an important cue for the detection of the presence of humans; usually the process is relatively simple. In addition, the advantage of the skin color during processing that lead quickly to compare with other features (Taqa, A. Y., & Jalab 2010b). Thus, the process of human skin detection based on the color algorithm is to determine the skin pixel and non-skin pixels (Zolfaghari, H., Nekonam, A. S., & Haddadnia 2011). This process is done by providing a measure to determine the distance between the pixel colors in regard to the skin tone. There are various methods used to distinguish between the skin pixel and non-skin pixels in the literature (Vezhnevets and Degtiareva 2003; Kakumanu, Makrogiannis, and

Bourbakis 2007). Thus, different skin detection applications adapted three key types of modeling methods are an explicitly identifying the skin regions, however this method is out of our scope. Whereas, the parametric and non-parametric skin approaches will be discussed in details in the next sections.

2.6.1 Parametric Skin Modelling

Parametric model approaches can be classified based on a single Gaussian model (Kakumanu, Makrogiannis, and Bourbakis 2007), multiple Gaussian clusters (Phung, Bouzerdoum, and Chai, D. 2005), a mixture of Gaussian (MoG) models (Hossain et al. 2012), or an elliptic boundary model (Kwolek 2003). Generally, these methods have characterization speed is slow most frequently during skin segmentation. As a result, they need to process each pixel individually. Additionally, all the described parametric methods operate in a color space chrominance domain, ignoring the luminance information. Given that an explicit distribution model is used, a question of model validation arises. Other words, parametric methods have low detection accuracy, as they rely on approximated parameters rather than authentic appropriate skin colors (Abdullah-Al-Wadud, Shoyaib, and Chae 2009). Furthermore, these methods have various performance depends in the color space that is used (Vezhnevets, Sazonov, and Andreeva 2003).

Parametric skin modeling methods are more suitable for constructing classifiers in cases where there are limited training and expected target datasets

(Shoyaib, Abdullah-Al-Wadud, and Chae 2012). The generalization and interpolation abilities of these methods facilitate the construction of classifiers with acceptable performance even when incomplete training data are used (Paul and Gavrilova 2011). Parametric methods are expressed with a small number of parameters and require minimal storage space; however, compared with non-parametric methods, they require more computation time to establish the skin probability model. A mixture of Gaussians in both training and work, as well as in their performance, depends strongly on the skin distribution shape. Moreover, most parametric models ignore the non-skin color statistics. This aspect, together with the dependence on the skin cluster shapes, result in higher false positive rates compared with the non-parametric models (Vezhnevets and Degtiareva 2003). To overcome the generality of the previous skin detection model classifiers, while the dynamic classifiers which are based on artificial neural networks (ANN) have been proposed in this case (Jadhav, Nalbalwar, and Ghatol 2011). Neural Networks have flexibility and ability to adapt of various image conditions make them a good choice for enhancing classification tasks for human skin pixels (Al-Mohair, Mohamad Saleh, and Suandi 2015). Thus, these methods are considered to be the most suitable method in parametric skin modeling (Doukim, C.A., Dargham, J.A., Chekima, A. and Omatu 2011; Singh Sisodia and Verma 2011;Paul and Gavrilova 2011; Al Abbadi, Dahir, and Abd Alkareem 2013; Borah and Konwar 2014). Neural Network method will review in details in the next section.

2.6.1.1 Neural Network Method

Myllymaki, P., & Tirri (1993) explain that Artificial Neural Networks (ANNs) are developed using large scale inputs and elements. These inputs and elements identified as artificial neurons are much larger than inputs used in traditional architectures. These elements are positioned in an interrelated way in a group that utilizes a mathematical model for processing of information founded on a connection approach to calculation. In order to store them, those neurons are made receptive by the neural networks.

In the past various types of neural network methods have been carried out in relation to skin detection. In some research studies, only a single-layer perceptron consisting of an input layer and an output layer was carried out (Rubegni, P. et al. 2002). This is a simple type of feed-forward network where a series of weights are used to feed the input elements directly to the output elements. In the multi-layer perceptron which includes hidden layers in its structure the classification task are more complicated (Dongare, Kharde, and Kachare 2012; Basheer and Hajmeer 2000).

In some advantages in regard the task of classification of the artificial neural network that not only has the built-in ability to deal and address images with high dimensional features but it can further deal with images that are noisy and which has contradictory data. Another advantageous feature is that there linear speed up in the matching process related to computational elements and this is enhanced when the input values are contrasted with the value of the stored cases in relation to others

(Myllymaki, P., & Tirri 1993). ANNs have some disadvantages include the high cost of computing, high memory usage and complexity in understanding by the ordinary layman. Hence, it is questionable if it would have a positive impact (Sharef, Omar, and Sharef 2014). In addition, neural networks need longer processing time when compared with others such as the Bayesian method (Zulhadi Zakaria 2009; Ardil and Sandhu 2010).

According to Araokar (2005), ANN was employed in the mid 1900's to serve as a simulation of the human mind in computation. From the late 1980s, ANN have been employed widely to the field of pattern recognition because of new innovative and well organized applications in the back-propagation algorithm and used during training of multilayer networks. These are competent in sorting out class areas in advanced distributions (Al-Boeridi, Syed Ahmad, and Koh 2015).

Neural Network has been successfully used in many skin detection classifiers (Zulhadi Zakaria 2009). Initial image training set is recognized by a neural network and the skin pixel identification is created by the trained network. Each neural network is set to be able to differentiate the images that have been trained as input. The neural networks will identify similarities of the target image. The neural networks can be setup quickly but some incorrect results may occur if inappropriate training data is used. On other hand, neural network produces better results in comparison to other complicated methods which are used in skin detection. In skin detection, some

researchers had said that the detection of the skin pixels using neural network method was not accurate, inefficient and sometimes wrong (Taqa and Jalab 2010a).

Back propagation algorithm is still used because it is efficient for training the multilayer networks, which are capable of separating class regions of arbitrarily complicated distributions (Al-Boeridi, Syed Ahmad, and Koh 2015; Al-Mohair, H. K., Saleh, J. M., & Suandi 2015). In order to learn a classifier a set of labelled skin and non-skin pixels are given in the ANN as adopted by (Kakumanu, Makrogiannis, and Bourbakis 2007; Al Abbadi, Dahir, and Abd Alkareem 2013; Duan et al. 2009; Bhoyar and Kakde 2010; Taqa and Jalab 2010a; Doukim, C.A., Dargham, J.A., Chekima, A. and Omatu 2011). In neural network feature vectors were used as input and subsequently trained to modify the weight values between nodes in relation to their feature values and class. Briefly, three layers make up the classifier namely the input, output and several hidden layers. The length of every description is dependent on the number of input nodes and the output nodes is basically single. During the training process, the algorithm extracts features from query images or pixels and subsequently categorized using the neural network (Al-Mohair, H. K., Saleh, J. M., & Suandi 2015).

2.6.2 Non-Parametric Skin Modelling

In skin detection often used skin color distributions from the training data without deriving an explicit skin color model. Consequently, the non-parametric methods

provided a fast performance both in training and classification, independent of distribution shape depends on the histogram model in the skin color domain. The key idea for non-parametric method based on interested colors as a histogram and then to replace the pixel colors by the value of the model histogram for each pixel color in image. (Q. Liu and Min 2010). A few studies (Vezhnevets and Degtiareva 2003, and Kakumanu, Makrogiannis, and Bourbakis 2007) has demonstrated three clear advantages of non-parametric methods such as fast training, classification, and usage. Their performance depends heavily on the training set selection and is theoretically independent of the shape of the skin distribution, which is contrary to the case of explicit skin cluster definitions and parametric skin modelling. These methods typically have high true positive and low false positive rates, indicating that they can locate most of the skin regions when there are few non-skin pixels marked as skin pixels (Kakumanu, Makrogiannis, and Bourbakis 2007).

The disadvantages of non-parametric methods include its need for a huge storage space and the inability to interpolate or generalize the training data. The storage of the (skin|RGB) table, also known as the lookup table (LUT), requires a large amount of memory (Vezhnevets and Degtiareva 2003). Thus, naïve Bayes classifier is considered the most suitable method for non-parametric skin modelling (Phung, Bouzerdoum, and Chai, D. 2005; (Hu et al. 2007; Ma and Leijon 2010; Linderoth, Robertsson, and Johansson 2013; Zhongdong, Saichao, and Zichao 2013). The Naïve Bayesian classifier will review in details in the next section.

2.6.2.1 Naive Bayes Classifier

Naive Bayes classifier is well-known as a simple probabilistic classifier which is rooted in the use of Bayes' Theorem with good self-determining suppositions. The Naïve Bayes classifier is also called as an independent feature model simply because the order of the features is irrelevant. Furthermore, the presence of one feature does not influence other features in categorization tasks (Ren et al. 2009; Taheri and Mammadov 2013; Cho 2014). This is a plus point as they make the calculation of Bayesian classification approach extra competent. The drawback of the Bayesian classification approach is that they strictly curb its applicability. This drawback can be overcome if a small number of training data are used to train data for approximating the parameters which are used for categorization (Sebe et al. 2004). As the self-determining variables are rooted in certain conditions, the resolve of the variances of the variables is done for every class and not the entire covariance matrix (Sharef, Omar, and Sharef 2014).

Whereas, the advantage of the Bayesian approach is able to classify more effectively in several practical situations under particular conditions (Rish, Hellerstein, and Jayram 2001; Zhong 2006). Ultimately, this approach is able to give precise categorization provided that the right group is more probable than the others (Ting, Ip, and Tsang 2011). Another benefit is that severe inadequacies in its basic Naïve probability model are disregarded by the largely tough classifier (Titterington 2008).

Although, there are many benefits of using Naïve Bayes classification approach, there still appears to be some weaknesses especially in relation to previous discerning algorithms, like the support vector machine (SVM) which is better in terms of providing efficient categorization. In comparison with other categorization approaches, Naïve Bayes categorization approach has poor categorization performance (Choi, Chung, and Ryou 2009). Hence, there is a dire need to undertake active research to seek reasons for the let-down of Naïve Bayes classifier in task categorization (Khan, Hanbury, and Stoettinger 2010; Cao and Liu 2012). It is interesting to note that more research is now being undertaken to put right the weaknesses inherent in Naïve Bayes categorization (Flach, P. A., & Lachiche 2004; Kuncheva 2006; Shirali-Shahreza and Mousavi 2008). In the past, Naïve Bayes had given significant results compared to other complicated approaches in terms of skin detection (Ma and Leijon 2010). The main universal mode to do this is by using Bayesian classifiers where the aim is to compute P (skin|c), which is the probability that a pixel with color c be a skin. After learning the P (skin|c), these values are used to generate a skin probability map for the image. In the skin probability map, the probability that each pixel be a skin pixel is put aside. This probability map is employed to generate the skin map which illustrates skin and non-skin pixels (Shirali-Shahreza and Mousavi 2008; Shruthi, M. L. J., & Harsha 2013). The process can be done by collecting measures of skin and non-skin pixel color samples and then arranging them in a normalized of histogram operation. Typically, the histogram operation achieves a probability for each pixel as skin or non-skin, such that a probability map is referred to the entire image. Where the proper threshold can be

applied that lead to the map use to detect whether each pixel is skin or non-skin (Zhongdong, Saichao, and Zichao 2013; Elgammal, Muang, and Hu 2009; Siqueira, Schwartz, and Pedrini 2013; Shruthi, M. L. J., & Harsha 2013; Patravali, Wayakule, and Katre 2014). The threshold method can be constant (Jones and Rehg 1999) or calculated adaptively for each image (M. J. Zhang and Gao 2005).

2.6.3 Why Selected the Case Study?

According to the literature, we carried out a comparison of various studies to find the strengths or weaknesses as below in the Table 2.9:

Table 2.9

Literature Survey for Various studies in Skin Detection Domain

Author & year	Technique \ method	Brief Description	Strengths and Weaknesses
García-Mateos et al. 2015	Adapt non-parametric method is modeling using histograms.	The author proposed a techniques based on color analysis allow classifying accurately and efficiently soil/plant regions in the images.	-The problem of PGC in natural images is addressed, by automatic binary classification. - There is a significant lack of comparative studies to select the optimum color spaces and color representation techniques for the plant segmentation problem.
Hoshyar, Al-Jumaily, and Hoshyar 2014	Pre-processing techniques used depend on the most popular techniques as Gaussian mean and median filters, and speckle noise filters.	The author represented comparison between three processes as Image enhancement, Image restoration and hair remove in the pre-processing techniques for designing the automatic skin cancer detection system.	- addressing a vital issue among researchers to reduce the rate of errors for automatic diagnostics of skin cancer. - The most challenging problems in medical image processing is the Automatic diagnostics of skin cancer.

(*Continue*)

Table 2.9 (*continued*)

Author & year	Technique \ method	Brief Description	Strengths and Weaknesses
Kukolja et al. 2014	The performed experiments through combination of a multilayer perceptron (MLP) with k-nearest neighbor (kNN)	The author presented a comparative analysis of emotion estimation methods in order to find the most suitable methods for the development of a personalized adaptive emotion estimator.	- addressing problem for physiology-based emotion estimation through comparative analysis of popular feature reduction and machine-learning methods. - Inconsistency problem that arise when using different emotion data related for databases are mutually compared.
Han et al. 2013	The illumination component using Gaussian smoothing filter, discrete cosine transform (DCT)	The author presented a comparative study on 12 different illumination preprocessing methods from two novel perspectives in illumination preprocessing for face recognition.	- Evaluation performance of 12 illumination preprocessing approaches with six face matching methods on four public face databases. - Often in the face recognition method, there a better visualization effect after illumination preprocessing does not imply get higher recognition accuracy.
Korotkov and Garcia 2012	using CAD systems that attempt to diagnose a PSL based on its visual similarity to images of skin lesions	The author presented a review in the computerized analysis of dermatological images with emphasis on computer-aided systems for skin cancer detection.	- providing a public dermoscopy dataset, will allow researchers to immediately report performance results for their methods, and thereby boost overall progress in the field. - These studies still do not provide unified results to algorithms, because of the differences in the datasets employed and different evaluation metrics.

(*Continue*)

Table 2.9 (*continued*)

Author & year	Technique \ method	Brief Description	Strengths and Weaknesses
Belaroussi and Milgram 2012	Skin color distribution is estimated using a non-parametric approach.	The author proposed an efficient approach for face detection and tracking.	-comparing two skin color based tracking approaches connected to the component segmentation and coupled Camshaft. -classification problem in the complexity of defining the non-face class in face detection in still images.
Y. H. Chen, Hu, and Ruan 2012	Using the default RGB color space to develop a very efficient skin detection method.	The author proposed an effective skin color model without color transformation, and also a very efficient embodiment of face detection.	-avoid transforming a large number of color information by method directly calculating the difference between each color component, and without performing floating point operations. -In the skin color models usually need to perform color space transformation which is not suitable for direct hardware implementation.
Abbas, Celebi, and García 2011	Evaluation of the hair detection error (HDE) used quantity of statistical metrics and manually used by a dermatologist.	The author presented a comparative study of the state-of-the-art algorithms for automatic detection of hair and restoration of the texture-part of tumors from occluded information.	- This comparative study is essential to reduce undesired segmentation and classification results of melanoma and other pigmented lesions. - In the case of intensity of hair pixel surrounding for tumor areas will decrease the effectiveness of hair segmentation and repairing algorithm.
Ebrahimzadeh and Khazaee 2010	Use number of MLP neural networks with different number of layers and different training algorithms.	The author proposed a number of efficient methods to accurately classify ECG beats for a relatively large set of data.	-Various network architectures were evaluated to find an optimum solution for ECG signal diagnosis problem. - One of the significant issues in ECG beat classification is how to appropriately evaluate the performance of a proposed method.

(*Continue*)

Table 2.9 (*continued*)

Author & year	Technique \ method	Brief Description	Strengths and Weaknesses
Gen Li et al. 2010	They used quantitatively evaluation for three skin color detection methods with mixture of Gaussian-based background subtraction.	The author proposed a method that combines skin color detection with background subtraction.	-reducing falsely detection of skin pixels by combine the skin color detection with background subtraction. -most existing algorithms in the skin color detection produce false positives when non-skin pixels have similar color to skin color.
A.A. Zaidan et al. 2014b	Using the multi-agent learning more efficient than other approaches.	They proposed a hybrid method involving the technique a Bayesian method with grouping histogram (GH) and a neural network with a segment adjacent nested (SAN).	-This study addressed three issues using new technique as multi-agent learning to detect the skin pixels accurately for problems of light-changing conditions, skin-like color, and reflection from glass and the water. - This study, although treated three problems in detecting the skin, but did not solve the problem of black color in the images detected.

As note in the Table 2.9 compares some of the studies in the skin detection domain. These studies are important in most research that highlighted the weakness and strength aspects of various studies. These studies can serve as guidelines for various issues in the comparative study of different studies that examine skin detection approaches

The comparisons of various studies that used individual or hybrid approaches, which are based on their strength and weakness aspects, are discussed above. This study adopts a case study, which is based on multi-agent learning system that combines Bayesian and neural networks (A.A.Zaidan,et al.2014b) solved three problems in skin detection approaches, which related with research problem in our study.

2.7 Chapter Summary

In this chapter, five sections have been identified for evaluation and benchmarking skin detection approaches. These sections achieved as follows;

In order to achieve the section (1), by gathering multiple criteria included three main criteria as a reliability, time complexity, error rate within the dataset has been done.

In order to achieve the section (2), investigate available benchmarking for techniques/ tool problems and limitations are needed. Thus, the issues and limitations of tools for a reliable skin detection method are identified.

In order to achieve the section (3), investigate the open issues and challenges for evaluation and benchmarking process of the different criteria. Four concern are highlighted as; (1) Evaluation criteria for skin detection received several criticisms in

relation to the evaluation metric for concern of the evaluation criteria. (2)A tradeoff is a situation where losing one criterion or aspect of something results in gaining another criterion or aspect and vice versa in regarding of concern of the criteria trade-off. (3)The benchmark process based on a comparison of the new generation with others under the conditions and criteria to be considered after the development process for any system regarding of concern of the benchmarking process. (4)A skin detection has numerous objectives should be developed when increasing the importance of a particular evaluation criterion and reducing others regarding of concern of the criteria importance.

In order to achieve section (4), investigate the proper techniques that deal with multi-criterion decision-making (MCDM) problems based on the recommended solutions that help grouping decision makers and organize the problems to be solved and conduct analyses, comparisons, and ranking.

Finally, in order to achieve the section (5) investigate about selection the case study through conducted a critical analysis of the different study according to the literature.

CHAPTER 3

METHODOLOGY AND DESIGN OF EXPERIMENTS

3.1 Introduction

In this chapter, a review of the methodology to achieve the four main research objectives designed to develop a new approach for skin detection, evaluation, and benchmarking is presented. The methodology comprises four key phases. The preliminary phase highlights the skin detection approach based on the literature and comprises three steps to achieve the first objective. The second phase, which is referred to as identification and performance phase, includes two main steps to achieve the second objective. The main purpose of this phase is to generate the decision matrix. Development phase, which includes three main steps to achieve the third objective, and aims to integrate two important MCDM techniques to obtain the final results. The final phase validate the final results to achieve the fourth objective and aims to explain the sequence of the case study in skin detection approach adapted from the literature. The case study is assessed based on the development of a new

methodology to evaluate and benchmark skin detection approaches. The four phases are described in Sections 3.1 to 3.4. The chapter summary is presented in Section 3.5.

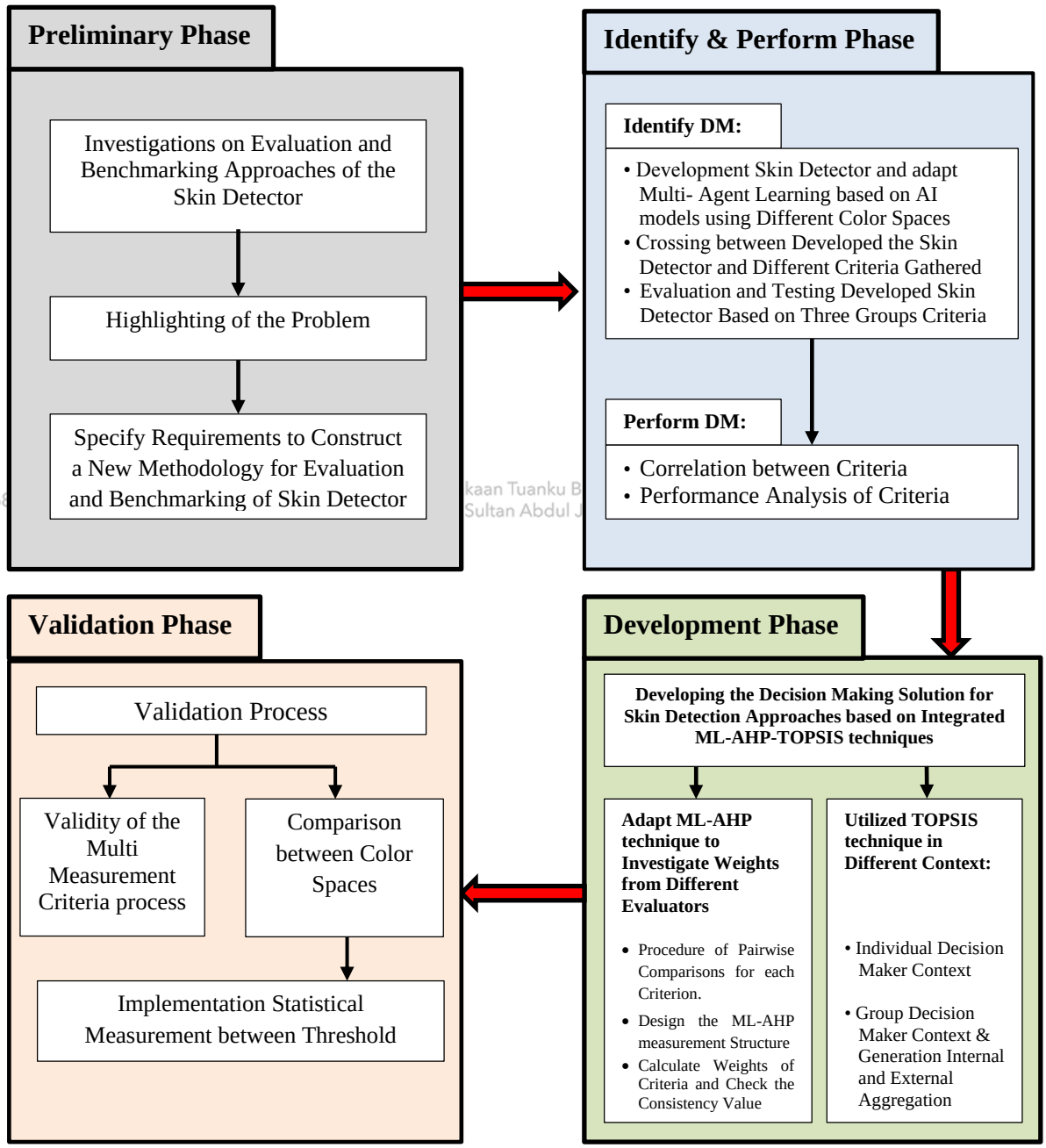


Figure 3.1. Research Methodology of Design Phases

3.2 Preliminary Phase

The preliminary phase highlights evaluation and benchmarking of skin detection approaches and related issues to achieve the first objective. Furthermore, the problem is emphasized through the trade-off between criteria, which require finding an optimum solution based on a new methodology. The new methodology will be implemented based on the MCDM techniques. This phase has been discussed in detail in Chapters 1 and 2.

3.3 Identification and Performance Phase

This section highlights the identification and performance of multi-dimensional criteria for skin detector engines to achieve the second objective of the present study. This phase implements two main stages to evaluate skin detection based on multiple criteria with different developed color spaces. Eventually, the results of the two stages will generate the decision matrix data. The two stages are discussed in detail below.

3.3.1 Identification of the Decision Matrix

Ascertaining the skin detector approach is an important stage in the creation of the decision matrix. This stage comprises three main steps: 1) development of the skin detector using different color spaces, 2) execution of crossing between different

criteria with developed skin detector engines, and 3) evaluation and testing of the developed skin detector based on three groups of criteria. The outcome of this stage is the establishment of the decision matrix from the practical aspect, which is discussed in detail below.

3.3.1.1 Development Skin Detector and adapt Multi-Agent Learning based on AI models using different Color Spaces

This section highlights a case study adopted in the current work according to Table 2.9. The case study will be developed based on selected 14 color spaces. This step is important in completing the identification and performance phase. The development of multi-agent learning technique for the skin detector is discussed in detail.

3.3.1.1.1 Multi-Agent Learning Technique

We adopted a case study using multi-agent learning based on neural network and Bayesian models according to a previous study which solved three problems of the skin detector approaches (A. A. Zaidan et al. 2014b). Researchers of the previous study proposed a new technique based on hybrid multi-agent learning to resolve three key issues in skin detection approches. They used parametric skin modelling (neural network model) with segment adjacent-nested (SAN) technique to solve the skin like problem. The Bayesian model with grouping histogram (GH) technique was also used to address the problem on the lighting condition. After that in parallel two models are

combined to resolve the reflection problem of water and glass. This technique is further discussed in detail below.

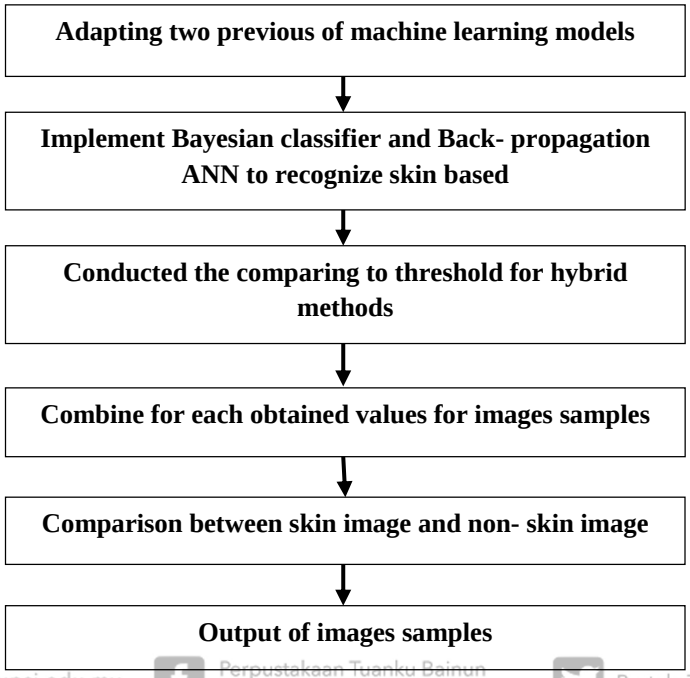


Figure 3.2. Multi-Agent Learning of Skin Detection

3.3.1.1.2 Color Space Adapted

A total of 14 samples of different color spaces collected from the literature are used in the study (Mircea 2012). The procedure applied to the color groups provides a solution to the lighting condition problem based on the removal of the illumination component from these samples using the Bayesian model of the proposed multi-agent learning technique. Color spaces are widely used in the studies of skin detection because they address most of the problems in this field (Zaidan, A. A., Karim, H. A., Ahmad, N. N., Alam, G. M., & Zaidan 2010a,b; A. A. Zaidan et al. 2014a,b). In our

study, each color space is built based on the separation of the illumination element from the Chroma element, which is a key step in the development phase. Therefore, the luminance element is ignored, whereas the Chroma element is retained because it is considered essential in determining skin color during skin detection. Thus, skin detection is conducted because the elimination of the luminance component is considered be a significant aspect in size reduction of skin cluster in the color space (Mircea 2012). Different color spaces are discussed in detail as follows.

1- Normalized RGB

Different color spaces can be easily changed to an RGB representation. RGB components do not only represent color but also luminance; hence, these components are used to represent the skin color in their chromatic color space. The luminance can be removed from the color space by normalization. Chromatic colors, which are known as "pure" colors in the absence of luminance, are defined by:

$$R = R/(R+G+B), \qquad G = G/(R+G+B).$$

The procedure removes B, which represents the luminance component (Al-Mohair, Mohamed-Saleh, and Suandi 2012; J. Yang, Lu, and Waibel 1998).

2- YCbCr

YCbCr is an encoded nonlinear RGB signal generally used by European television studios and utilized for image compression pattern. Basically, this color space is considered a clear choice for skin detection when it efficiently separates luminance and easily transforms from RGB and vice versa:

$$Y = 0.299R + 0.587G + 0.114B,$$

$$Cb = B - Y,$$

$$Cr = R - Y.$$

where Y represents the excluded luminance component while using Cb and Cr only (Vezhnevets, Sazonov, and Andreeva 2003; Chai and Bouzerdoum 2000).

3- YCgCr

YCgCr is a color space derived from YCgCr. This color space has Y channel which provides the luminous component, that is, light intensity; meanwhile, Cg and Cr channels represent the green and red difference of chromaticity components, respectively. The color space is used for digital video encoding. Y is avoided and only the chrominance part is adopted in the proposed integrated approach (Daithankar, Karande, and Rarale 2014; Chaves-González and Vega-Rodríguez 2010). YCgCr color space is generated by transforming the RGB values using the equations below:

$$\mathbf{Y} = 16 + 65.481R + 128.553G + 24.966B,$$

$$\mathbf{Cg} = 128 - 81.085R + 112G - 30.915B,$$

$$\mathbf{Cr} = 128 + 112R - 93.768G - 18.214B.$$

4- YCgCb

YCgCb is another color space derived from YCbCr. The RGB image determines the fitting skin region for each Y as the luminance component and the two chrominances as Cg and Cb. At this stage, the luminescent element is excluded while the chrominance element is retained. If both color components of a pixel are within the

boundary of a fitting skin region, then the pixel is classified as a skin pixel. Cg-Cb color space for skin tone detection is represented by a circular model (Z. Zhang and Shi 2009). Thus, the circular model for the skin tone in the transformed Cg-Cb space is described using the following expression:

$$\frac{(x-Cg)^2 + (y-Cb)^2}{12.25^2},$$

$$\begin{pmatrix} x=12.25\,cos+107 \\ y=12.25\sin\_110 \end{pmatrix}.$$

5- YUV

In this color space, Y is the luminance component while UV is represented as the chrominance component. For YUV, the color space removes the luminance-related component (Y) to improve the performance of the skin detection process. Thus, the definition of the luminance component in this color space is a good step toward obtaining invariant to luminance. YUV is used in the analog television system as PAL or NTSC. Human vision is sensitive to the luminance and chrominance factors in the images. Color space confirms this sensitivity by increasing bandwidth of the luminance to be close to human perception. YUV is derived from the original RGB source. Thus, this color space can be converted to RGB formats based on linear transformations (Abadpour and Kasaei 2005; Chaves-González et al. 2010). YUV color space is generated by transforming RGB values using the following equations:

Y = +0.299R + 0.587G + 0.114B,

U = −0.14713R − 0.28886G + 0.436B,

V = +0.615R − 0.51499G − 0.10001B.

6- YIQ

The YIQ color space is nearly similar to YUV. This color space comprises luminance (Y) and chrominance components (I and Q). The components I and Q can be represented as a second pair located on the axes of the graph; therefore, I and Q represent different coordinate systems on the same plane. Thus, these components can be represented in RGB values, where I is matched to range B, and Q is matched to range G. This color space can also be converted to RGB formats based on linear transformations and is represented by the following expressions (J. Yang, Lu, and Waibel 1998; Chaves-González et al. 2010). To convert the RGB into YIQ model, the following equations are used:

$$\mathbf{Y} = 0.299R + 0.587G + 0.114B,$$

$$\mathbf{I} = 0.595716R - 0.274453G - 0.321263B,$$

$$\mathbf{Q} = 0.211456R - 0.522591G + 0.311134B.$$

7- HSI, HSV and HSL

Perceptual color spaces are considered to be popular samples in skin detection. In these color spaces, I, V, and L represent luminance components, while H and S represent chrominance components. Three color spaces separate the components: hue (H), saturation (S), and luminance (I, V, and L). Essentially, the three color spaces are deformations of the RGB color cube and can be mapped from the RGB space via a nonlinear transformation. Moreover, these color spaces allow users to intuitively specify the boundary of the skin color class in terms of hue and saturation, and this capability of color spaces is considered to be one of the advantages of these samples.

As I, V, or L provide information on brightness, these components are often dropped to reduce illumination dependency of skin color (Manigandan and Jackin 2010; Albiol, Alberto, Luis Torres 2001; Shin, Chang, and Tsap 2002).

8- IHLS

IHLS is also known as improved hue, luminance, and saturation (IHLS) color space. The IHLS model is improved with respect to similar color spaces, such as HSL, HSI, and HSV, using normalization to remove the luminance component. Thus, this method overcomes certain numerical problems limited by the color components, thereby providing a better distribution of the features of space (Khan, Rehanullah 2012; Shin, Chang, and Tsap 2002).

9- CIEXYZ

CIEXYZ is one of the perceptual uniformity systems, which means that a small perturbation to a component value is approximately equally perceptible across the range of the value. The CIE color system is based on the Commission International de l'Eclairage (CIE) primaries established in 1931. The CIEXYZ color space forms a cone-shaped space, with Y as the luminance component and X and Z as chrominance components. The luminance component of each color space is dropped to form a 2D color. Thus, the values of each component of the color spaces are adjusted to the range (0-255) and quantized in 256 levels (Schmugge et al. 2007; Chaves-González et al. 2010).

10- CIELab

The CIELab is a reasonably perceptually uniform color space proposed by the CIE. This color space has two components represented as a and b values in the 1976 CIE Lab color space. Thus, a and b in the color space refer to the chrominance component, whereas L refers to the luminance component. Typically, CIELab does not use the luminance component of the color because it considerably varies across the human skin. Generally, chrominance is reliably used to separate the skin from the surrounding non-skin regions (Kasson and Plouffe 1992).

11- CIELuv

CIELuv is another color space derived from perceptually uniform color space proposed by the CIE. U and V represent the chrominance component, and L represents the luminance component. Generally, as RGB color space is far from being perceptually uniform, non-linear transformation of CIELAB and CIELUV attempts to correct this situation. Therefore, the CIELuv color space is considered to be the best candidate among the other samples when the luminance component is dropped (Xiong and Li 2012; Vezhnevets and Degtiareva 2003).

12- CIELch

CIELch is color space that is also derived from perceptual uniformity systems created by the CIE, in which L represents luminance, and c and h represent the chrominance components. Performing the Utans for the first time often causes dropping of the illumination components that lead to many errors in the light cluster. Therefore,

lighting pixels are dependent on illumination components, while dark cluster (absence of light) performs well even when illumination components are omitted. Furthermore, the minimum necessary components cannot be achieved by using Utans, thereby possibly reducing the features and training and testing on the network (Araban, Farokhi, and Kangarloo 2011). Thereafter, we adopted different color spaces developed using AI models according to the literature. Thus, the aim of this process is to obtain various alternatives (See Figure 3.3).

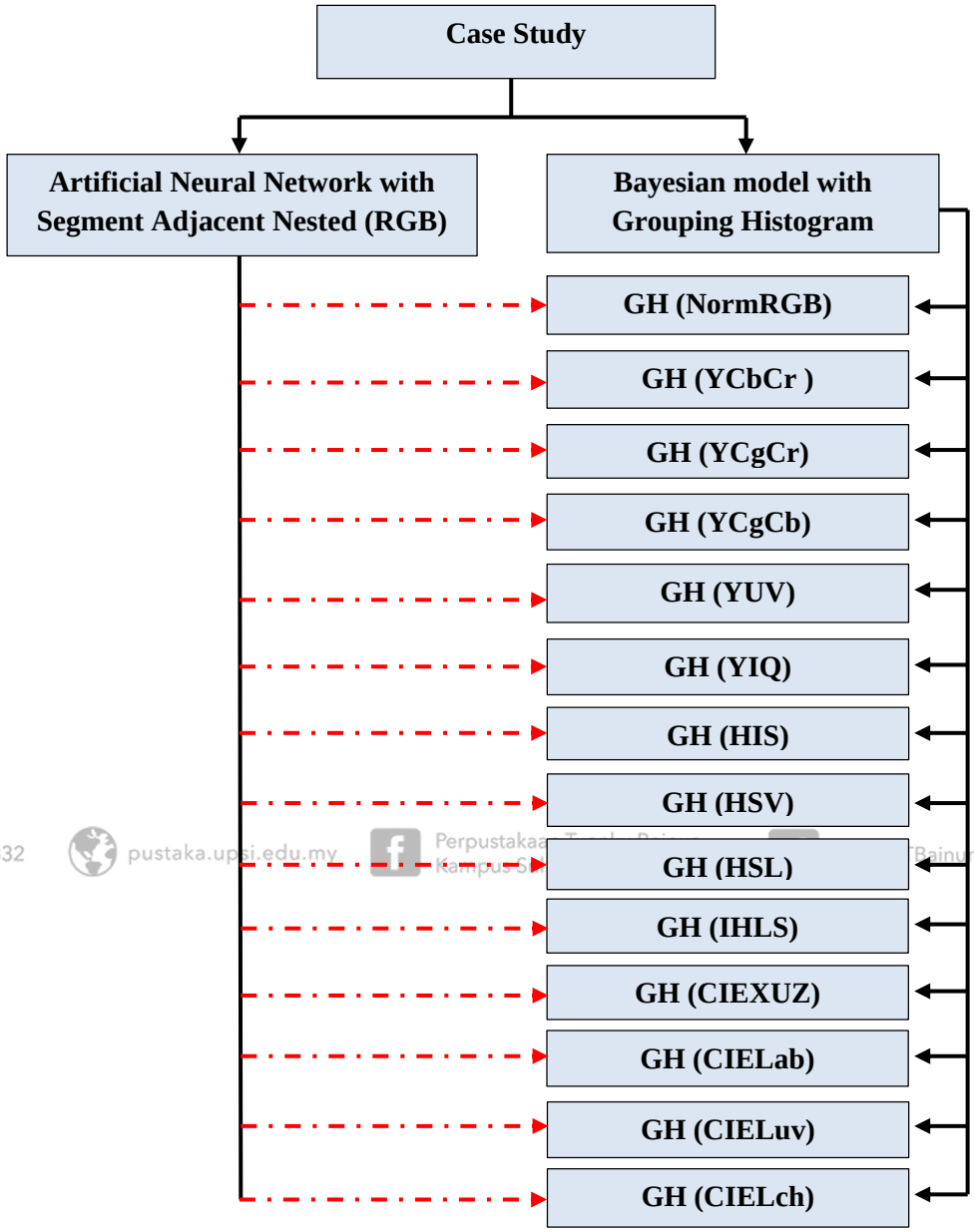


Figure 3.3. Development Case Study using Different Color Spaces

Figure 3.3 illustrates development procedure of different color spaces implements a training process for AI models, such as ANN and Bayesian, according to the literature. The neural network model used the RGB color space only, while the Bayesian model used 14 color spaces. The procedure is discussed in detail below.

3.3.1.1.3 Training operation of Neural Network Model

According to A. A. Zaidan et al. (2014b), the case study applied the neural network model in the proposed multi-agent learning of the skin detector which comprises three main layers, including the input, hidden, and output layers. The architecture of this model comprises nine neurons in the input layer, four neurons in the hidden layer, and a single neuron in the output layer. The model aims to conduct segmentation process for different samples and is implemented using the RGB color space, which is considered to be an essential vector for an image sample. The dataset is distributed into three parts: 1200 samples for training, 300 samples for validation, and 300 samples for testing. A validation process is implemented in the setup of the training process, which is conducted to calculate the main square error (MSE) to represent the performance function of the ANN model.

The training operation for ANN adopts the back-propagation pattern, which is a suitable way to train the feed-forward neural network model. The training operation is achieved through 331,282,971 pixels based on 1200 images from the dataset of the skin and non-skin pixels. The default training function adopted a Levenberg–Marquardt in the back-propagation function (trainlm) (Demuth, Howard 2009). The aim of feed-forward network training is to create a network object. Feed-forward operation requires five steps to generate network object. The first step creates an array for input vectors to implement segment adjacent-nested (SAN) technique. The second step identifies the probability elements as skin and non-skin indexes by creating an array that includes output sample considered to be target vectors. The third step is the

creation of an array, which identifies the hidden layer size in the network. The fourth step determines the cell array, including the names of transfer functions, which are used in two layers.

The default transfer function is used for the hidden and output layers to achieve three layers of the network only. The hyperbolic tangent sigmoid (tansig) is used for the hidden layer, whereas the linear transfer function (purelin) is used for the output layer. The fifth step includs the name of the training function to be used. (Demuth, Howard 2009). By contrast, the activation function is used for three layers. The input layer is deactivated, and the hidden and output layers are non-linear and linear, respectively. In the current study, the back-propagation feed-forward neural network is used based on the reasonable results obtained by this method from previous studies (Bhoyar and Kakde 2010; Doukim, C.A., Dargham, J.A., Chekima, A. and Omatu 2011; Taqa and Jalab 2010a; Zolfaghari, H., Nekonam, A. S., & Haddadnia 2011; A. A. Zaidan et al. 2014b). The RGB color space is only used during the training process of the ANN model.

3.3.1.1.4 Training operation of the Bayesian Model

The Bayesian model is considered to be important in machine learning algorithms. This model is derived from the Bayesian rules and normalization of the lookup table function (LUT). The procedure of the Bayesian model is based on clustering

histogram using 1200 images to calculate the LUT to identify skin and non-skin pixels. Typically, histogram computation is implemented in the post-training, followed by the normalization of the LUT, thus providing the distribution of separate probabilities. The training process of the Bayesian model is illustrated in Figure 3.4.

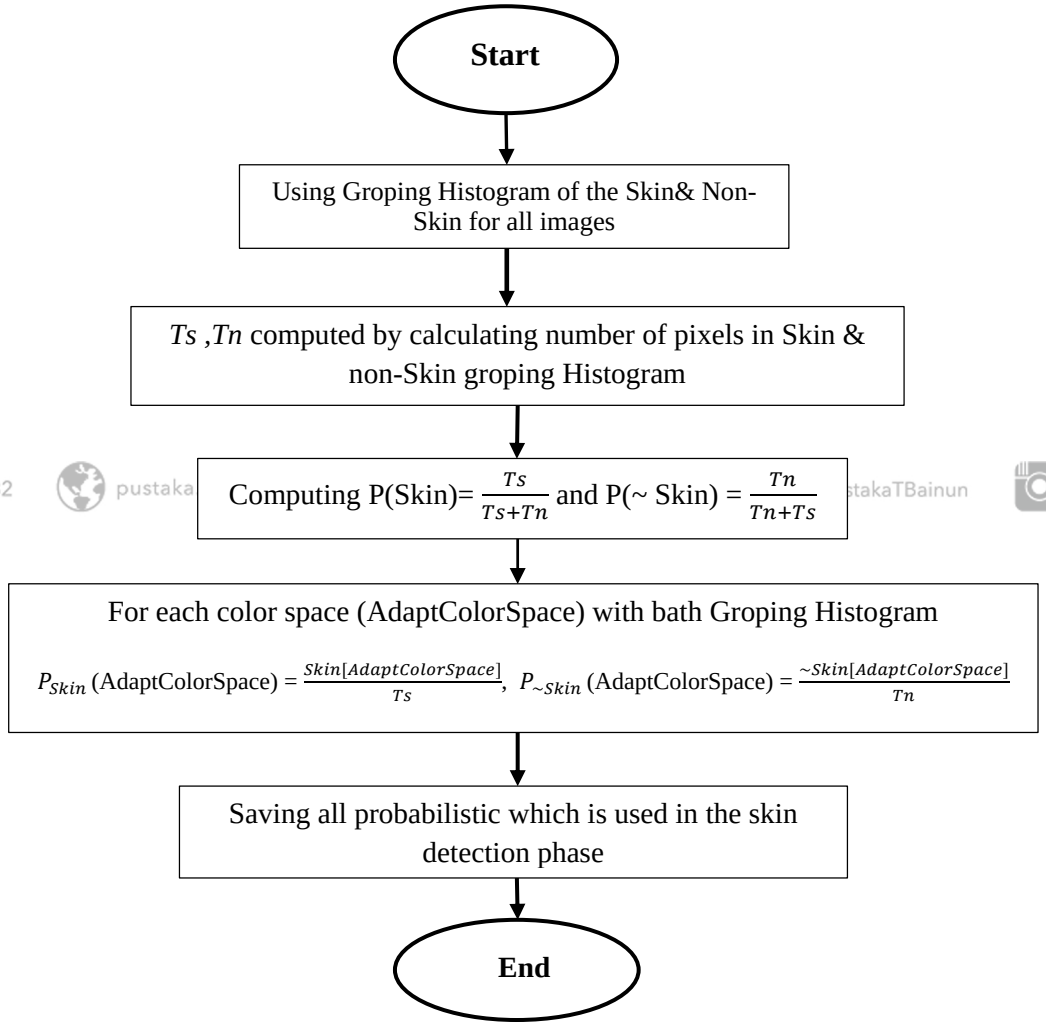


Figure 3.4. Training Process of the Bayesian Model (A. A. Zaidan et al. 2014b)

The probability of different color space pixels calculated as a pixel of color space (AdaptColorSpace) to identify the skin pixel is observed. This probability is denoted by P_{skin} (AdaptColorSpace) and is represented by Equation 3.1.

$$P_{Skin}\,(AdaptColorSpace) = \frac{Skin[AdaptColorSpace]}{Norm} \quad \textbf{(3.1)}$$

where skin[AdaptColorSpace] refers to the histogram value that matches the color vector (AdaptColorSpace). The calculation of grouping histogram values using the normalization coefficient is represented in the parameter Norm, which is a summation of all grouping histogram values. The normalization value of the LUT refers to the color matching probabilities of the skin. Thus, skin detection can be calculated as P(skin|AdaptColorSpace) following the Bayesian rule, which is given in Equation 3.2 (Chai, D. and Bouzerdoum, A., 2000; Flach, P. A. and Lachiche, N., 2004):

$$P(Skin\backslash AdaptColorSpace) = \frac{P(AdaptColorSpace\backslash skin).P(Skin)}{P(AdaptColorSpace\backslash skin).P(Skin)+P(AdaptColorSpace\backslash \sim Skin).P(\sim Skin)} \quad \textbf{(3.2)}$$

The probability values of the P(AdaptColorSpace\skin) and P(AdaptColorSpace\~Skin) are directly calculated for skin and non-skin pixels, respectively, using the grouping histogram process. By contrast, the previous probabilities P(skin) and P(~ skin) can be easily computed by calculating the total number of pixels of the skin and non-skin at the training step, as shown in Equations 3.3 and 3.4, respectively (Chai and Bouzerdoum 2000):

$$P\,(Skin) = \frac{Ts}{Ts+Tn} \quad \textbf{(3.3)}$$

$$P\ (\sim Skin) = \frac{Tn}{Tn+Ts} \quad \textbf{(3.4)}$$

where Ts and Tn are the values of skin and non-skin pixels, respectively. After the training process for the dataset, the resulting probability values must be saved in two files which can be represented by the following:

Ps = FUN (GH(X)) and

Pns = FUN (GH(Y)).

Hence, the multi-agent learning technique is implemented according to the desired goal based on specific functions to achieve the best results.

3.3.1.1.5 Detection Step of the Skin Detector

The detection phase begins after the training phase is completed to gather the required data from various image samples using the proposed technique by A. A. Zaidan et al. (2014b). This phase evaluates the performance of the multi-agent learning technique adapted using 14 different color spaces. Figure 3.5 illustrates the process of segmentation and skin detection.

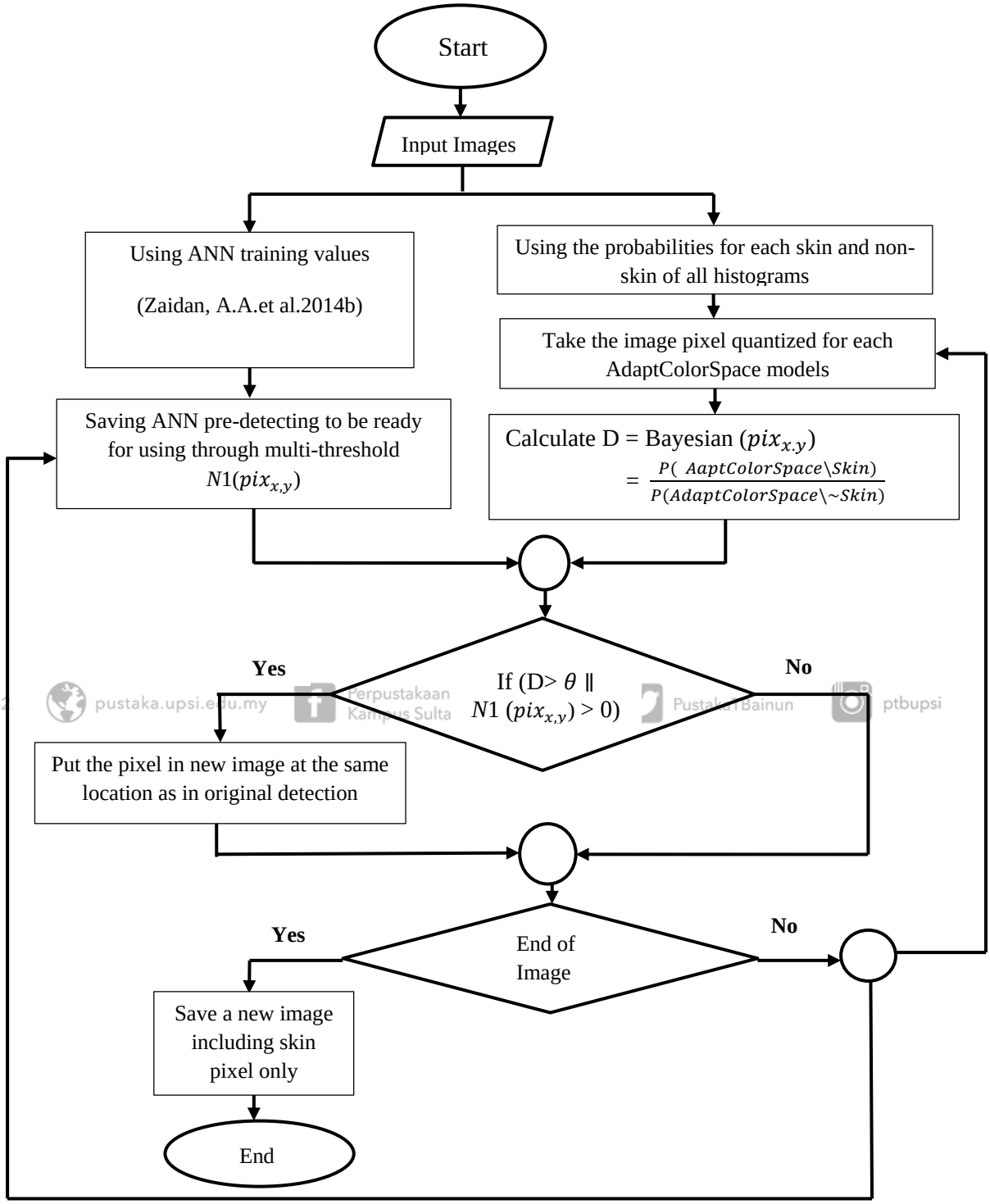


Figure 3.5. Skin Segmentation and Detection Processes

The detection process depends on the parameters obtained from the training process for ANN and Bayesian models. Therefore, these parameters are determined based on the final stage of the proposed system according to the output of the original image. The system generates a new image that represents only the skin pixels, while the non-skin pixels are represented as a white background. According to the abovementioned system, the original image pixels represent the inputs or probabilities represented in two files (Ps and Pns), which indicate a pre-detection phase in the Bayesian model. Thus, these pixels are saved in two variables:

nn = P(AdaptColorSpace \skin) ➔ skin pixels

ww = P(AdaptColorSpace \~skin) ➔ non-skin pixels

If the values of P(ww) and P(AdaptColorSpace \ww) are represented for ww = {skin or non-skin}, then P(ww\ AdaptColorSpace) is determined, which is already an accepted result that allows the usage of the Bayesian model rule:

$$\text{IF } \frac{P(AdaptColorSpace\backslash skin)}{P(AdaptColorSpace\backslash\sim skin)} > \frac{P(\sim skin)}{P(skin)} \quad \textbf{(3.5)}$$

where (AdaptColorSpace) is classified as a skin pixel. Otherwise, it is non-skin pixels.

Thus, the Bayesian rules can be computed at the minimum cost (Chai, D. and Bouzerdoum, A., 2000):

$$\frac{P(AdaptColorSpace\backslash skin)}{P(AdaptColorSpace\backslash\sim skin)} > \theta \longrightarrow \text{AdaptColorSpace} \in \text{Skin} \quad \textbf{(3.6)}$$

$$\frac{P(AdaptColorSpace\backslash skin)}{P(AdaptColorSpace\backslash \sim skin)} < \theta \longrightarrow \text{AdaptColorSpace} \in \sim\text{Skin} \qquad \textbf{(3.7)}$$

According to Equation 3.7, the calculation is such that Bayesian ($pix_{x,y}$) = *ww*/*nn*; therefore, if *nn* is equal to 0 then *nn* will automatically resets to *nn* = 0.00000000001, and if *ww* = 0, then *ww* will automatically resets to *ww* = 0.00000000001.

The threshold variable (θ) is defined in Equation 3.8, as follows:

$$\theta = \frac{P(\sim skin)}{P(skin)} \qquad \textbf{(3.8)}$$

The outcome of the multi-agent technique is distributed on the image pixels of *I* with the threshold value based on Equation 3.9, in which the procedures described apply the same conditions for both models.

For Neural Network:

$$N1\ (pix_{x,y}\ ) = \begin{cases} \sim skin & if \quad N1(pix_{x,y}) < 0 \\ skin & else\ where \end{cases}$$

For Bayesian:

$$\text{Bayesian}\ (pix_{x,y}) = \begin{cases} \sim skin & if \quad Bayesian\ (pix_{x,y}) < \theta \\ skin & else\ where \end{cases}$$

$$\forall Pix_{x,y}, N1(Pix_{x,y}) < 0 \quad OR \quad Bayesian\ (Pix_{x,y}) < \theta \longrightarrow Pix_{x,y} \in \sim skin \qquad \textbf{(3.9)}$$

where $N1(Pix_{x,y})$ is the second parameter for the function FUN (*I*, *NI*, *Ps*, *Pns*). Bayesian ($Pix_{x,y}$) is collected from the previous second steps. The thresholds are considered based on two aspects. First, ($N1\ (Pix_{x,y}) < 0$), represents the neural part, where 0 was selected. The range of training for non-skin pixels was lower than zero,

while that of the skin pixels was larger than zero. From these points, the boundary between the skin and non-skin pixels is considered to be 0.

However, the applied rules of the pixels, which are referred to as the non-skin at [255 255 255], are identified as a white pixel. Otherwise, the system will return the pixels from the original image I as [R (y, x), G (y, x), B (y, x)] for skin. Thus, the system collects skin pixels according to the rules applied. By contrast, the procedure of the Bayesian model is represented as $(pix_{x,y}) < \theta$.

Hence, the nine threshold values (∅) selected are 0.50, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, and 0.95 (A. A. Zaidan et al. 2014b). We also obtained 108 algorithms according to different color spaces and are used in our study. The results collected are based on the individual implementation of nine experiments for each color space. However, the detection process is repeated for each color space used in all experiments.

3.3.1.2 Crossing between Developed Skin Detector and Different Criteria

In this section, a crossover between 14 color spaces is conducted according to Figure 3.3 with 13 of the criteria collected from the literature. The procedure highlights the effect of different criteria on various color spaces. Therefore, we emphasize the obtained multiple criteria to be a basic element in the established decision matrix.

In summary, the crossover between multiple criteria and adopted different color spaces leads to the creation of the decision matrix, which is used for generating the final results in the next phase. Table 3.1 shows the establishment of the decision matrix.

Table 3.1

Establishment of the Decision Matrix

Criteria / Algorithm	**Reliability**	**Time Complexity**	**Error Rate for (Training and Validation)**	
Algorithm 1	RV (A1/ TS)	TcV (A1/ TS)	ERT (A1/ TSt)	ERV (A1/ VS)
Algorithm 2	RV (A2/ TS)	TcV (A2/ TS)	ERT (A2/ TSt)	ERV (A2/ VS)
Algorithm 3	RV (A3/ TS)	TcV (A3/ TS)	ERT (A3/ TSt)	ERV (A3/ VS)
Algorithm 4	RV (A4/ TS)	TcV (A4/ TS)	ERT (A4/ TSt)	ERV (A4/ VS)
Algorithm 5	RV (A5/ TS)	TcV (A5/ TS)	ERT (A5/ TSt)	ERV (A5/ VS)
.	.	.	.	.
.	.	.	.	.
Algorithm n	RV (An/ TS)	TcV (An/ TS)	ERT (An/ TSt)	ERV (An/ VS)

RV: Reliability values
TcV: Time complexity values
ERT: Error Rate for Training
ERV: Error Rate for Validation

A: algorithm
TS: Test Samples
n: number of algorithms
TSt: Training Samples
VS: Validation Samples

Table 3.1 shows three main groups of criteria with different color spaces as alternatives within the proposed decision matrix. The procedures for each criterion is discussed in detail below.

3.3.1.2.1 Procedure for Computation Reliability Group Elements

The reliability group includes three basic sections, namely matrix, relationship, and behavior of the parameters. The relationship among these sections emphasizes the importance of evaluating the skin detectors in our study. The procedure for each sub-criteria within a reliability group is discussed in detail as follows.

Confusion matrix generation is considered to be an important and basic phase which constructs the matrix of parameters that represents the first sub-criteria of the reliability group. The matrix of parameters comprises four key parameters, namely TP, TN, FN, and FP, which are considered to be the backbone for the computation of the remaining criteria within the reliability group. A certain procedure is performed to calculate the values of the basic parameters and their complementary values based on the matching process. The matching process is discussed in detail for a sample of the predicated parameters and the actual parameters to generate the confusion matrix. These parameters represent the results obtained by the previously multi-agent learning technique discussed above. Ultimately, these parameters represent the final results of the decision matrix after conducting the matching process in Figure 3.6.

The procedure for the confusion matrix is performed to calculate the values of the basic and complementary parameters based on the matching process. The matching process is performed by matching the object locations of both images to calculate the image skin pixels. This procedure is implemented to calculate the locations of the two images based on their object location. For example, the figure

below contains six objects representing the base of both images, thus calculating the number of pixels in Object 1 with Object 2 through a pointer that matches each pixel of the actual parameters with each pixel of the predicated parameters. These pointers will calculate only the skin pixels of the actual parameters using the predicated parameters, which represent a standard measure for the number of pixels of the predicated parameters calculated as the TP, whereas the difference between the standard and calculated pixels is represented as the FN. Meanwhile, the rest of the pixels is calculated for other objects. Therefore, the average value of the skin pixels for such example will be calculated to produce the final TP for the entire image and obtain the average FN as the final FN. Similarly, the TN, which represents the background of the image as non-skin pixels, can be calculated, whereas the FP is considered to be a complement of the TN. Therefore, the parameter values are calculated to be TN according to the values of the non-skin pixel objects from the predicated parameters, while the FP is calculated based on the difference between the standard and the calculated values.

By contrast, these parameters are calculated based on different threshold values for each color space to obtain the final result of the decision matrix. According to the literature, nine threshold values, which represent the basis for each criteria value relative to the color spaces identified in our study, are adapted. Thus, the values for each color space will be separately calculated by conducting individual experiments to generate the final parameter values for the decision matrix. The detailed results will be discussed in Chapter4.

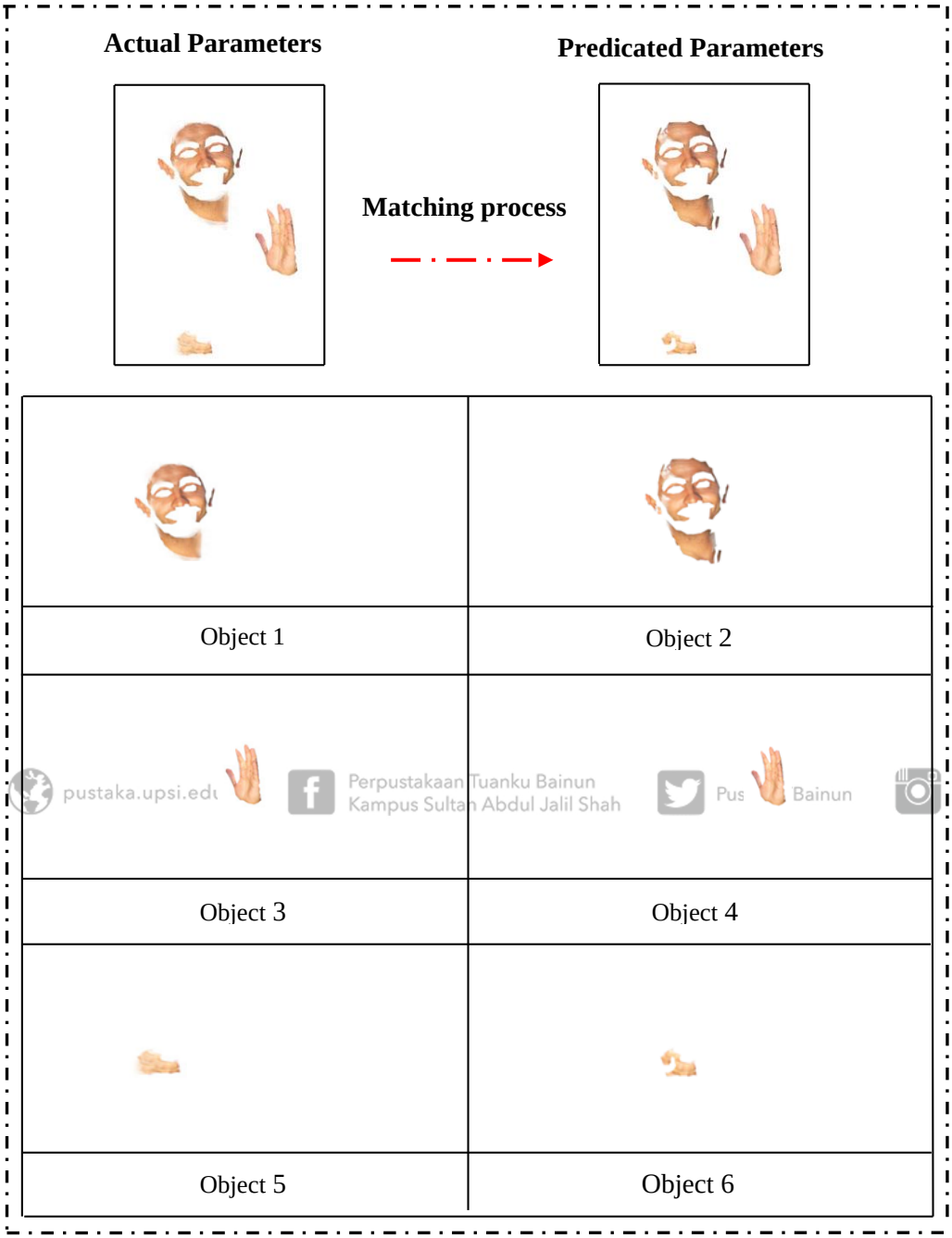


Figure 3.6. Matching Process for Different Objects

The second sub-criteria within the reliability group is a relationship of parameters, which includes accuracy, recall, precision, and specificity. Their parameter values can be computed based on the results of the confusion matrix, which comprises the four

main parameters previously mentioned. According to Eqs. 2.3, 2.4, 2.5, and 2.6, the parameters mainly depend on the values (TP, FP, TN, and FN) to obtain the final results.

The last sub-criteria of the reliability group is the behavior of parameters that comprise F- measure and G-measure. These parameter values can be calculated based on the values of precision and recall through Eq. 2.8 and 2.9 to obtain the final results.

Finally, the identification and performance process is completed by evaluating the decision matrix and obtaining the required dataset to be used in the development stage.

3.3.1.2.2 Procedure Computation for Time Complexity Criterion

Time complexity is an important criterion in our research. The procedure and methodology for calculating the time are based on the time consumption of the input and output sample images. A flow chart is presented below to show the process of computing time complexity.

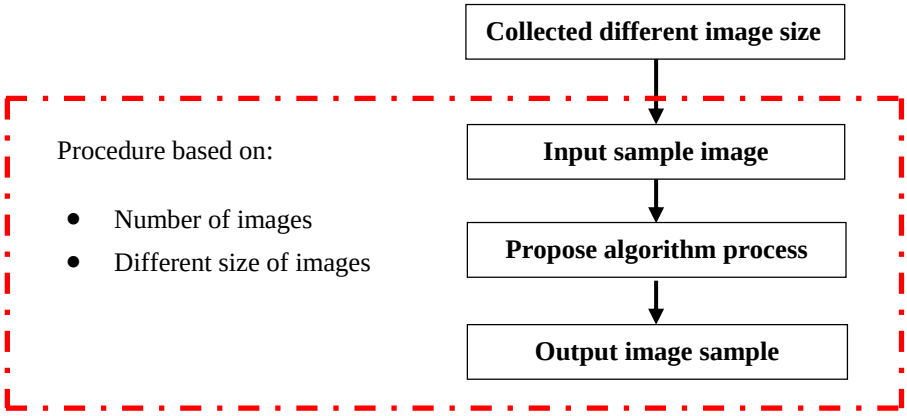


Figure 3.7. Procedure of Time Complexity

The procedure of calculating image process depends on the number and size of image samples as follows:

$$T_{process} = T_o - T_i \quad \textbf{(3.10)}$$

where T_o is output time image process and T_i input time image process

$$T_{total} = \frac{T_{process}}{T_{Avarge}} \quad \textbf{(3.11)}$$

Where $T_{process}$ represent the difference among output and input image samples and $T_{average}$ represents the average process for all samples. These particular equations cover the computation time of different image sizes and different objects of skin in the same image size.

3.3.1.2.3 Error Rate Computation within Dataset Elements

The error rate within the dataset is considered to be the main pillar in the evaluation of numerous studies that rely on machine learning algorithms. These algorithms use a special mechanism in most AI models to obtain the results through two main stages: training and testing. The training procedure is a key step in these algorithms; thus, the dataset is trained for several times to derive a minimum error rate by selecting a specific dataset, such as training and validation.

The training dataset normally uses two-thirds of the data, and the rest is used for validation and testing. The procedures of dataset training and validation are conducted to reach the minimum error rate. However, the reliability values in the testing data are dealt to obtain the final results.

The training process is individually performed for each color space. A total of nine thresholds that produce equal training values are implemented for each color space. The final results of these experiments will vary according to the color spaces used.

3.3.1.3 Evaluation and Testing Developed Skin Detector Based on Three Groups Criteria

The decision matrix is created based on two key parameters: multiple criteria and different color spaces. First, three key criteria (reliability, time complexity, and error

rate within a dataset) are used to evaluate and test the proposed skin detector approaches. Second, 14 color spaces are selected to stand as alternatives in the decision matrix.

The decision matrix, which is evaluated and tested based on the calculation procedure for each color space, is performed from nine experiments according to threshold values for data collection, thereby providing the final results of the decision matrix. Thus, after completing the identification stage of the decision matrix, 108 algorithms are generated according to the different color spaces that have been processed. The calculations are performed for the first color space. The process is then individually repeated for the rest of the color spaces. The decision matrix will be discussed in detail in Chapter 4.

3.3.2 Performance of Decision Matrix

This step is considered to be the second part of the identification and performance phase. The part conducted in two trends; first, to investigate the relationship between the criteria and determine their degree of correlation. Second, the performance analysis is conducted to evaluate and compare the criteria and identify the factors that affect their behavior. This stage is implemented in two steps discussed as follows.

3.3.2.1 Correlation between Criteria

We determine the importance of finding the relationship among the different data criteria at this step. This step is implemented to verify the existing statistical differences between them; otherwise, only one will be used. The case study includes three main groups of criteria, which have been discussed in detail in Chapter 2. These criteria have interconnected physical characteristics. Therefore, proving the relationship among these criteria is necessary. Several software programs and techniques based on mathematical and statistical methods are available for proving the relationships among criteria. We adopt a Pearson's method to find a correlation among the various criteria in our study (Rodgers and Nicewander 1988; Asuero, Sayago, and González 2006; Egghe and Leydesdorff 2009). The method represented by Pearson's formula *r* is shown in the following equation:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})*(Y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X}))^2 * (\sum_{i=1}^{n}(Y_i - \overline{y}))^2}} \quad \textbf{(3.12)}$$

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{N} x_i \in \text{the mean } \mathbf{x} \quad \textbf{(3.13)}$$

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{N} y_i \in \text{the mean } y \quad \textbf{(3.14)}$$

where *n* is the number of input value pairs, *X* and *Y* are two criteria and $\overline{X}$ and $\overline{Y}$ are their average values and *r* is correlation coefficient value.

The procedure to calculate the coefficient value *r* for (x, y) ranges from −1 to 1 and its value has constant to linear transformations between variables. An *r* value

near 0.00 indicates uncorrelated criteria, while an *r* value near or equal to 1 indicates a high level of correlation (D. Wang et al. 2014; Hall 2015). The results of the method will be discussed in detail in Chapter 4.

3.3.2.2 Performance Analysis of Criteria

In our study, various criteria were collected in evaluating the skin detector to determine the behaviour of each criterion at the nine threshold values for all colour spaces used. Therefore, this step will verify the existing difference among the behaviours of criteria to be adopted; otherwise, only one will be used. A total of 13 criteria were investigated despite existing trade-offs among them. On the basis of the literature, nine threshold values were adopted for each test of the colour space. The three main groups of criteria have been discussed in detail in Chapter 2, and the performance analysis of criteria will be thoroughly discussed in Chapter 4.

In summary, the perform decision matrix stage investigated the existing correlation between each criterion according to Pearson's formula. The factors that affected the behavior of these criteria were determined. Further details will be provided in Chapter 4.

3.4 Development Phase

In the development phase, a new methodology will be presented based on MCDM techniques to achieve the third objective. The new methodology is constructed based on the integration of two basic MCDM techniques to achieve its design purpose by ranking and selecting the best alternatives. It is based on the values of the decision matrix created in the previous phase. The implementation of this phase will be discussed in detail below. Figure 3.8 describes the new methodology for skin detector.

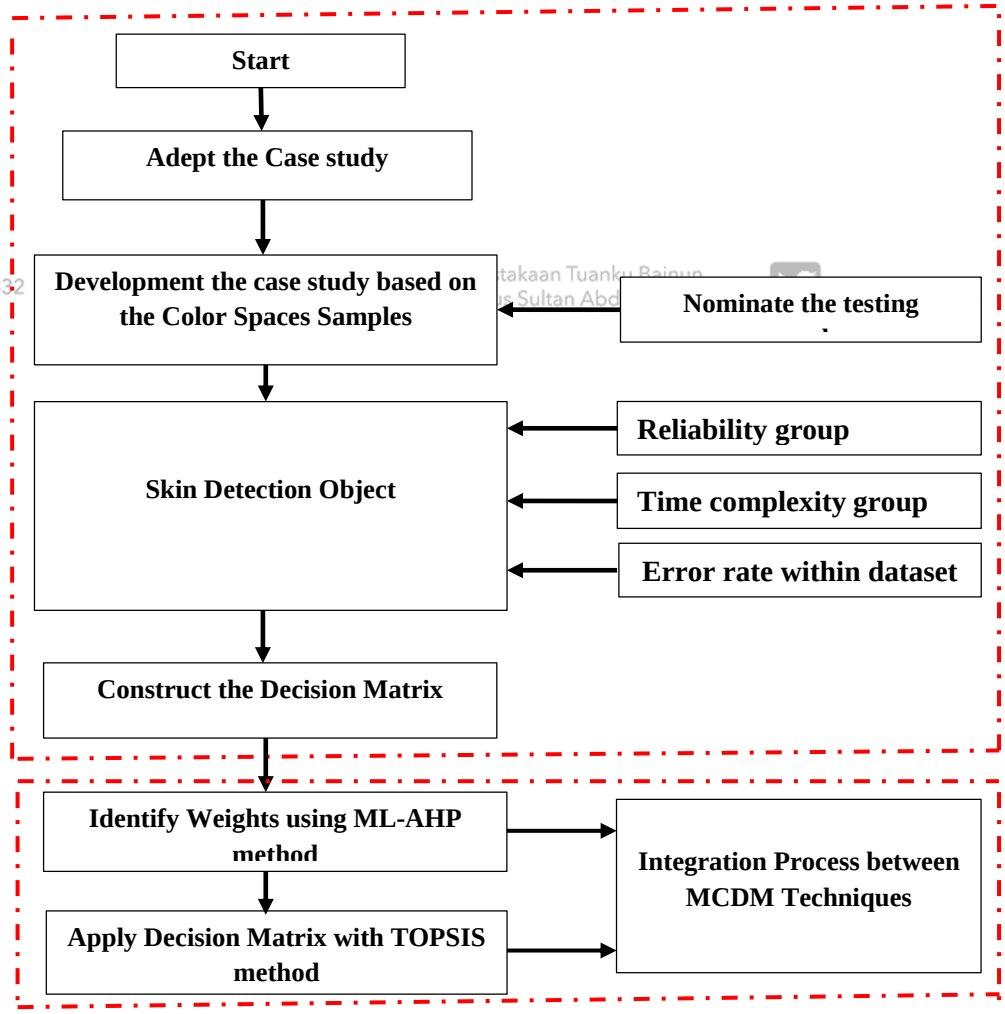


Figure 3.8. New Methodology for Skin Detector

Figure 3.8 shows the stages of constructing the new methodology in details. The new methodology is constructed through adapting and developing the previous technique based on different color spaces compared to various criteria to generate decision matrix. Thus, the integration of two MCDM technologies from ML-AHP and TOPSIS method will implement in order to obtain the final results and choose the best alternative. Further details will discuss the process in the next section.

3.4.1 Development of Decision-making Solution for Skin Detection Approach Based on Integrated ML-AHP&TOPSIS

The integration of the two ML-AHP and TOPSIS methods is widely accepted by many researchers based on the following conditions. First, they are capable of presenting the results of complete ranking and calculating the relative distance based on weights and objective data. Second, their results are satisfactory for random analyses and show a harmony trade-off through using nonlinear relationships, thereby allowing for easy conversion into a programmable format. Thus, in our study, we will adopt integration between ML-AHP and TOPSIS methods to rank numerous color space algorithms in the skin detection approach.

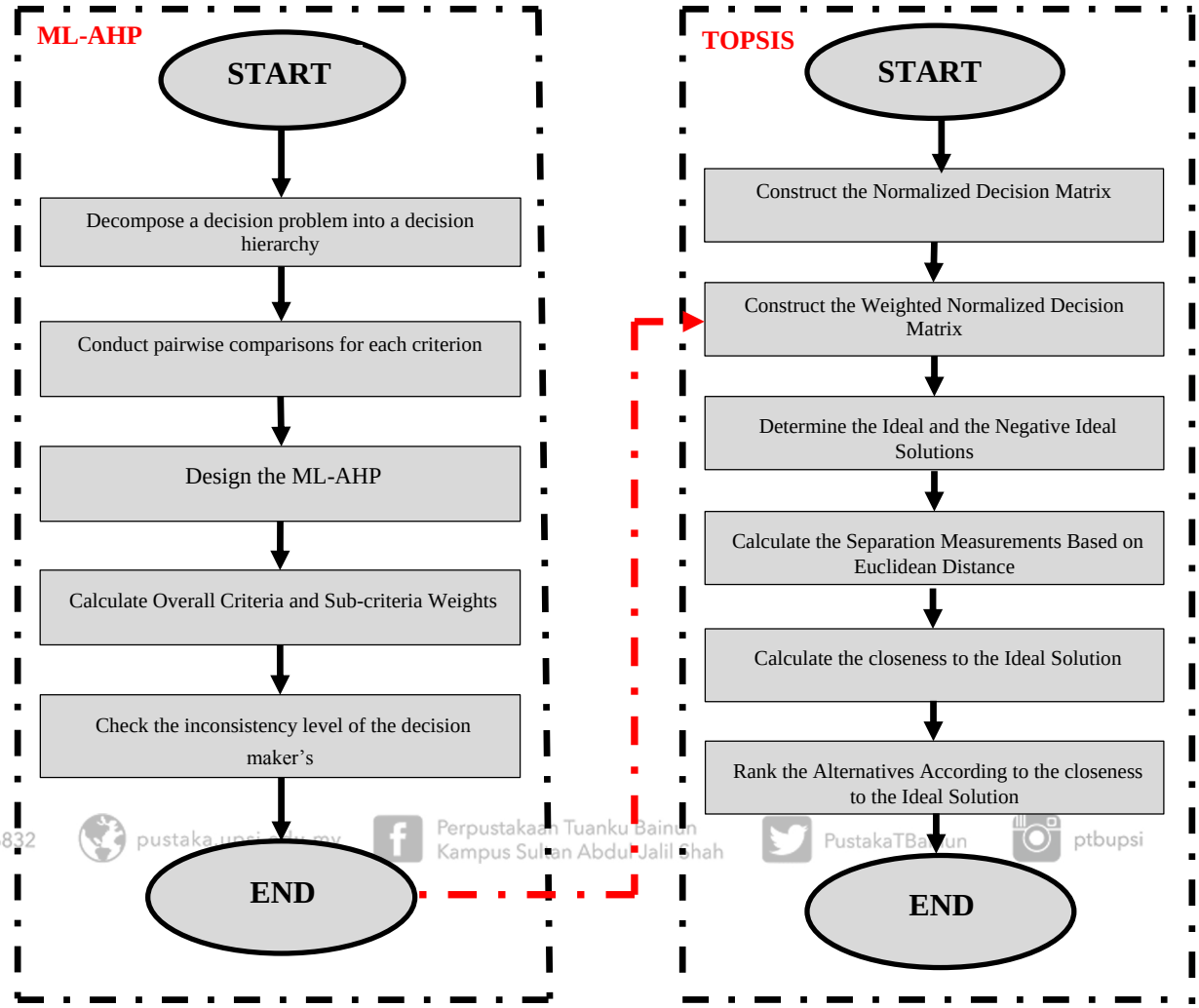


Figure 3.9. Integration of ML-AHP and TOPSIS methods for Skin Detection Approaches

A total of 13 weight settings, which represent three key groups of criteria under different circumstances, are used in the first part. In this step, the weights are assigned according to external evaluator preferences. Thus, the AHP technique is used to measure the weights from the pairwise form, and the outcome of this technique will be used in the TOPSIS method. Different color spaces have been developed as alternatives in the decision matrix. These alternatives must be ranked to configure the

selection of the best one in the second part. Eventually, the TOPSIS method will use the decision matrix to provide the final results.

3.4.2 Adaptation of ML-AHP Technique for Weight Investigation of Different Evaluators

Any approach to skin detection is developed to achieve a few objectives (for example, detection of skin and non-skin in an image). Based on this objective, the developer can assign the weight of each evaluation criteria. The determination of weights is based on the priority preference in MCDM with interval numbers. With regard to the skin detection evaluation and benchmarking problem, assigning weights is difficult. By contrast, weights can be assigned with the help of experts in normal life problems. This difficulty arises because each expert has different opinions on the importance of skin detection criteria, thereby creating a conflict with the objective of the designer. The weights can be assigned in several ways, such as using the ML-AHP algorithm to make a pairwise comparison between the criteria and using fuzzy weights to solve the problem. However, these solutions rely on experts, thereby creating a conflict between the preference of the experts and that of the designer. A total of 13 different weight settings are selected to be assigned in the decision-making process to solve this problem. Thus, our research will use pairwise technique as represented in the ML-AHP method which will be discussed in detail. Figure 3.10 shows the AHP method based on multiple layers.

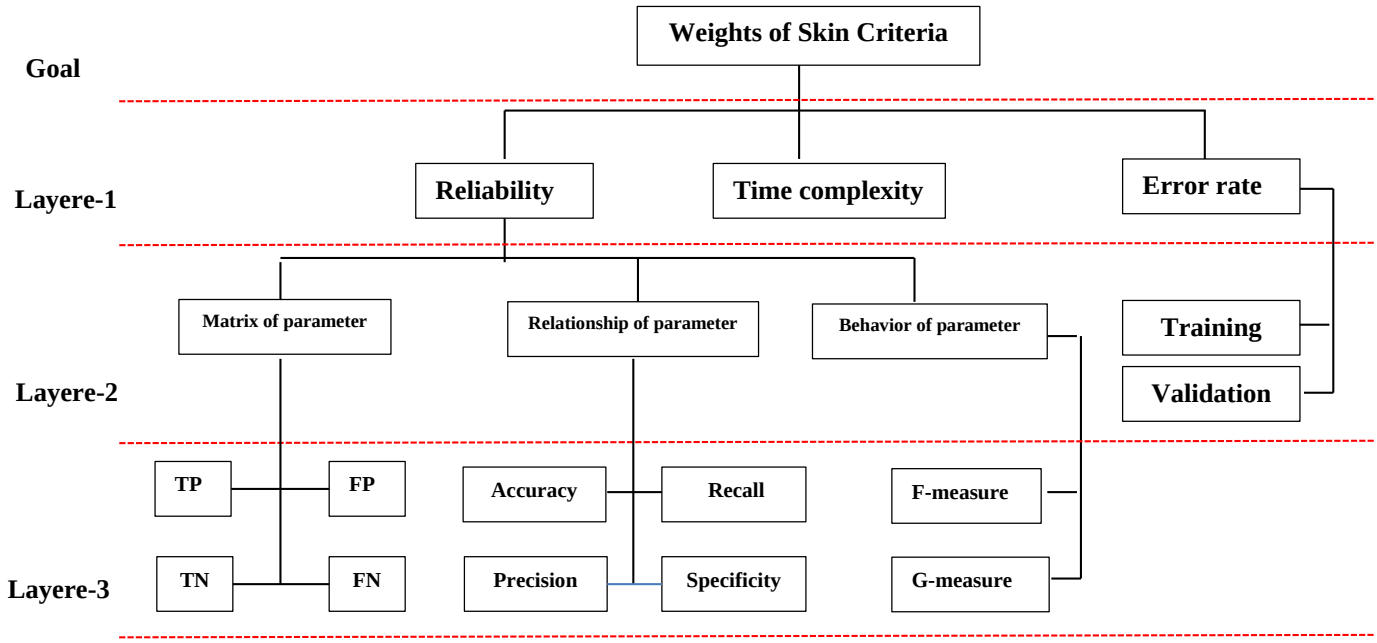


Figure 3.10. AHP method based on Multi-Layer Structure

3.4.2.1 Pairwise Comparisons for Each Criterion

AHP is one of the most popular multi-criteria decision-making methods that was originally developed (T L Saaty 1990; T.L. Saaty and Ozdemir 2003) as a technique to realize ratio scales from paired comparisons. ML-AHP allows a few inconsistencies in judgment because humans are not precisely consistent. The ratio scales are derived from the principal Eigen vectors, and the consistency index is derived from the principal Eigen value. The number of required pairwise comparisons can be represented by the following formula:

$$n*(n-1)/2 \qquad (3.13)$$

where n is the number of criteria that is utilized during the evaluation process. In comparing a set of criteria, n in pairs is based on the amount of relative weights. Criteria and weights can be represented as $(C_1 \ldots . C_n)$ and $(w_1 \ldots . . w_n)$. This comparison can be represented in the matrix as follows:

$$C = \begin{array}{c|ccc} & C_1 & C_2 \ \ldots\ldots\ldots & C_n \\ \hline C_1 & w_1/w_1 & w_1/w_2 \ \ldots\ldots & w_1/w_n \\ C_2 & w_2/w_1 & w_2/w_2 \ \ldots\ldots & w_2/w_n \\ \vdots & \vdots & \vdots & \vdots \\ C_n & w_n/w_1 & w_n/w_2 \ \ldots\ldots & w_n/w_n \end{array}$$

The matrix uses a pairwise ratio whose rows provide the ratio of the weights for each element with respect to all other ratios. This method focuses on extracting weights for various activities according to importance. Typically, the importance is a judgment based on different criteria. Occasionally, these criteria correspond with objectives selected by activities for investigation (T L Saaty 1990). Table 3.2 presents a comparison matrix for priority rating comprising three pairwise elements in the decision matrix.

Table 3.2

Sample Pairwise Comparison Matrix

Criteria	***A***	***B***	***C***
A	**1**	A/B	A/C
B	B/A	**1**	B/C
C	C/A	C/B	**1**

In our study, we adopted six experts from various universities in Malaysia. These experts have sufficient background in image processing using AI methods according to their curriculum vitae. We collected answers from these experts after asking them questions on the evaluation of different criteria according to the questionnaire (See appendix A). The outcomes will be discussed in detail in Chapter 5.

Typically, these answers will be adopted when matched with the degree of consistency according to the rules of hierarchy theory to calculate the weights. Their answers are based on the procedure as follows: compare different criteria based on the priorities identified according to the perspective of the evaluators. Figure 3.11 shows the formula of the questionnaire presented to the experts.

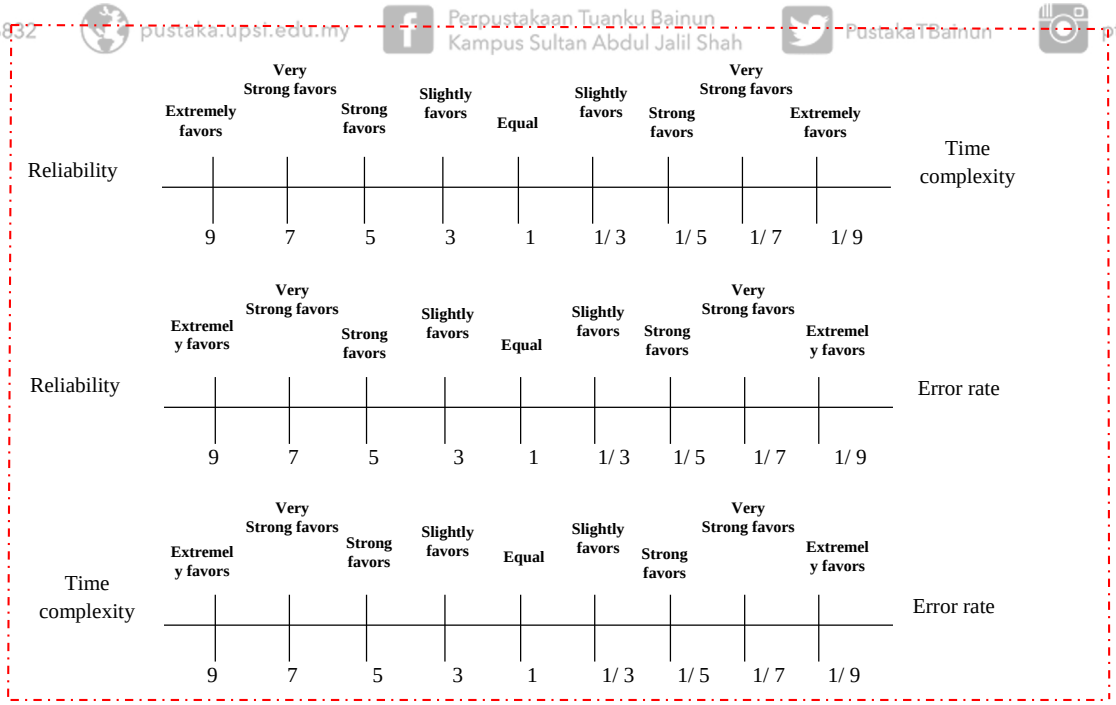


Figure 3.11. Pairwise Answer from Evaluators

Thomas L. Saaty (1977) proposed a new scale to calculate the degree of importance between the different criteria. Basically, the scale used the difference between successive scale values to allow criteria comparison within scale values ranging from 1 to 9. Table 3.3 shows an initial step toward the construction of the intensity scale of importance for activities.

Table 3.3

Intensity Scale of Criteria

Degree of Importance	Description
1	Equal importance
3	Weak importance of one over another
5	Essential or strong importance
7	Demonstrated importance
9	Absolute importance
2,4,6,8	Intermediate values between the two adjacent judgments

3.4.2.2 Design of the ML-AHP measurement Structure

In this step, our work has included multiple layers to distribute the criteria. ML-AHP measurement matrix is implemented to obtain the weights according to the preference of the evaluator. ML-AHP measurement uses mathematical calculations based on pairwise to convert the judgments of experts to generate weights for each criterion. A consistency ratio must also be calculated for judgments that represent the internal

consistency values entered. Thereafter, the answers of evaluators with pairwise comparisons are collected and the reciprocal matrix is created. This matrix provides the sub-criteria values for each main criterion at each level and identifies the importance of each feature compared with its parent. Thus, the main criteria features obtained represent the importance of each feature in relation to the goal.

Figure *3.**12*** shows the weights using ML-AHP measurement based on different evaluators.

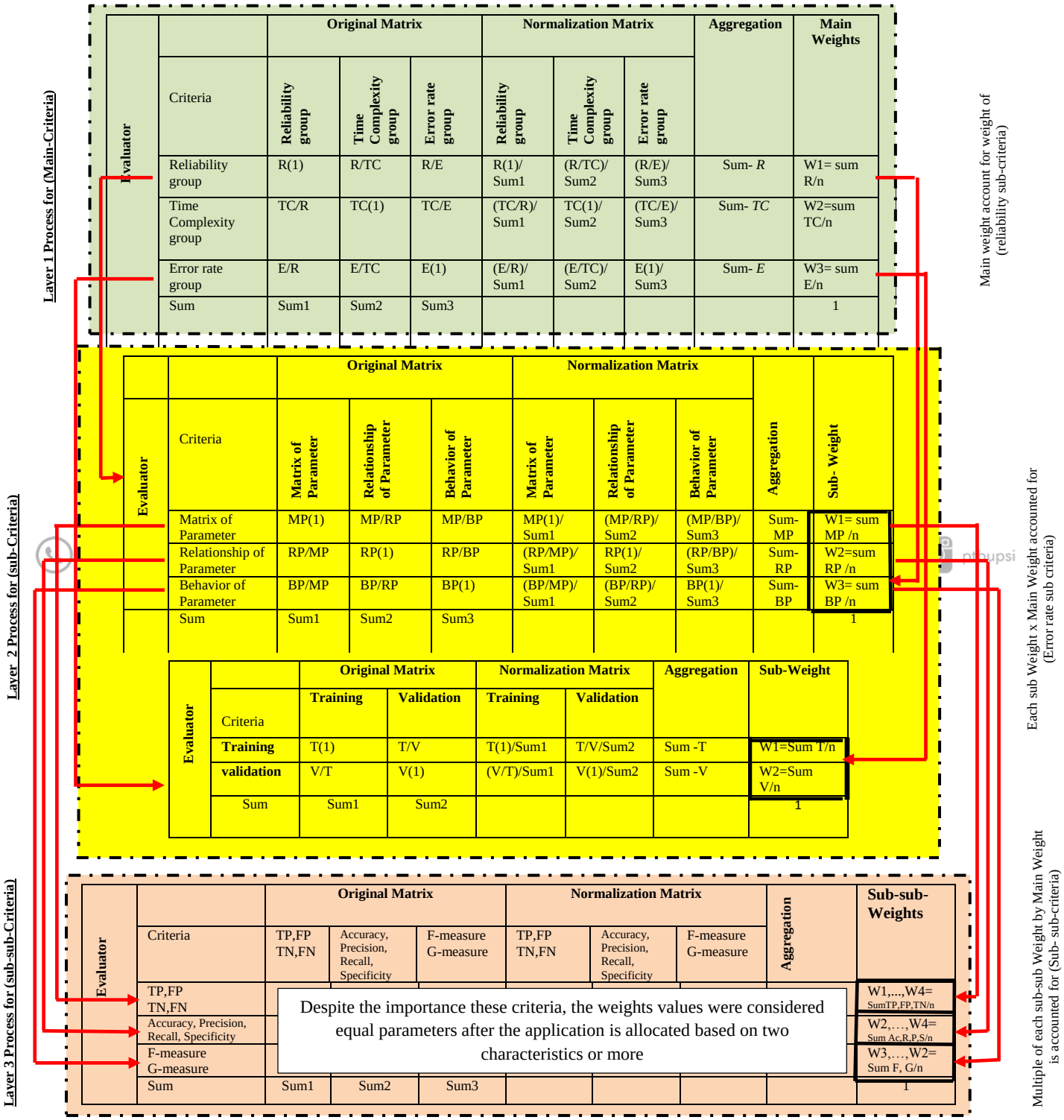


Figure 3.12. ML-AHP Steps Used to Account for Multi-layer Matrix

Figure 3.12 illustrates three layers of the various criteria adopted in this study. According to the rules of the AHP method, the criteria were distributed on the three layers according to its priority to calculate their weights. The weights of the criteria were calculated based on the pairwise relationship in this method. After collecting the answers from six evaluators and determining the weights of criteria. These weights represent the final result that will be used in the TOPSIS method later.

3.4.2.3 Weight Calculation of Criteria and Validation of Consistency Value

Various responses collected by different evaluators must be converted to numerical values in the decision matrix. The decision matrix implements procedures such as normalization and aggregation process for these values. The next stage is determining the weights of the criteria and ranking them. By contrast, the ML-AHP measurement considers an important vector to conduct a consistency test, which is normally required after completing the calculation of the criteria weights. Inconsistency is often observed in the answers obtained by the ML-AHP questionnaire from individual evaluators, thereby affecting the overall consistency of the test. Hence, the consistency ratio must be tested before all responses are collected from the evaluators (Thomas L. Saaty and Vargas 1984).

Finally, consistency ratio (CR) is measured to determine the consistency of pairwise. This procedure is called a consistency index. A CR larger than 0.10 indicates an inconsistency in the pairwise comparison, whereas a CR equal or less to

0.10 indicates that the comparison is reasonable (Al–Azab, F. G. M. and Ayu, M. A., 2010). We can calculate CR using the following formula:

CR= CI / RI **(3.14)**

where CI represents the consistency index obtained from the formula:

CI = (max–n) / (n-1) **(3.15)**

Then Random Index (RI) represent in the Table 3.4.

Table 3.4

Random Index (T.L. Saaty and Ozdemir 2003)

N	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
RI	0.00	0.00	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49	1.51	1.54	1.56	1.58	1.59

3.4.3 Utilization of the TOPSIS Method for Skin Detection Evaluation and Benchmarking

In this section, we utilize TOPSIS, which is favoured among MCDM techniques. This method involves several steps. The TOPSIS method is applied to each alternative based on the geometric distance from positive and negative ideal solutions. Thus, the mechanism is followed to select the best alternative according to the rules of the technique: the alternative with the shortest geometric distance to the positive ideal solution and the longest geometric distance to the negative ideal solution. The procedures of the TOPSIS method are described as follows:

1- Construct the normalized decision matrix

In this process, the various attribute dimensions are transformed into non-dimensional attributes; this process allows a comparison across the attributes. The matrix $(x_{ij})_{m*n}$ is then normalized form $(x_{ij})_{m*n}$ to the matrix,R $= (r_{ij})_{m*n}$ using the normalization method:

$$r_{ij} = \boldsymbol{x_{ij}} / \sqrt{\sum_{i=1}^{m} \boldsymbol{x_{ij}^2}} \qquad \textbf{(3.16)}$$

This process will result in a new Matrix R where R is expressed as:

$$R = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ r_{21} & r_{22} & \dots & r_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ r_{m1} & r_{m2} & \dots & r_{mn} \end{bmatrix} \qquad \textbf{(3.17)}$$

2- Construct the weighted normalized decision matrix

In this process, a set of weights $w = w_1, w_2, w_3, \cdots, w_j, \cdots, w_n$, from the decision maker is accommodated in the normalized decision matrix. The resulting matrix can be calculated by multiplying each column from normalized decision matrix (R) with its associated weight w_j. Notably, the set of the weights is equal to 1,

$$\sum_{j=1}^{m} w_j = 1 \qquad \textbf{(3.18)}$$

This process produces the new matrix V where V is expressed as

$$\text{V} = \begin{bmatrix} v_{11} & v_{12} & \dots & v_{1n} \\ v_{21} & v_{22} & \dots & v_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ v_{m1} & v_{m2} & \dots & v_{mn} \end{bmatrix} = \begin{bmatrix} w_1 r_{11} & w_2 r_{12} & \dots & w_n r_{1n} \\ w_1 r_{21} & w_2 r_{22} & \dots & w_n r_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ w_1 r_{m1} & w_2 r_{m2} & \dots & w_n r_{mn} \end{bmatrix} \qquad \textbf{(3.19)}$$

3-Determining the ideal and negative ideal solutions

In this process, two artificial alternatives A^* (the ideal alternative) and, A^- (the negative ideal alternative) are defined as:

$$A^* = \left\{\left(\left(\max_i v_{ij} \,\middle|\, j \in J\right), \left(\min_i v_{ij} \,\middle|\, j \in J^-\right) \,\middle|\, i = 1,2,\dots,m\right)\right\} \quad \textbf{(3.20)}$$

$$= \{v_1^*, v_2^*, \dots, v_j^*, \cdots v_n^*\} \quad \textbf{(3.21)}$$

$$A^- = \left\{\left(\left(\min_i v_{ij} \,\middle|\, j \in J\right), \left(\max_i v_{ij} \,\middle|\, j \in J^-\right) \,\middle|\, i = 1,2,\dots,m\right)\right\} \quad \textbf{(3.22)}$$

$$= \{v_1^-, v_2^-, \dots, v_j^-, \cdots v_n^-\} \quad \textbf{(3.23)}$$

Notably, J is a subset of $\{i = 1,2,\dots,m\}$, which present the benefit attribute, whereas J^- is the complement set of J, or (J^c), which the set of cost attribute.

4-Separation measurement calculation based on the Euclidean distance

In the process, the separation measurement is conducted by calculating the distance between each alternative in V and the ideal vector A^* using the Euclidean distance which is shown in the following equation:

$$S_{i^*} = \sqrt{\sum_{j=1}^{n}\left(v_{ij} - v_j^*\right)^2}, \; i = (1,2,\cdots m) \quad \textbf{(3.24)}$$

Similarly, the separation measurement for each alternative in V from the negative ideal A^- is given by:

$$S_{i^-} = \sqrt{\sum_{j=1}^{n}\left(v_{ij} - v_j^-\right)^2}, \; i = (1,2,\cdots m) \quad \textbf{(3.25)}$$

In the end of step 4, two values namely S_{i^*} and S_{i^-} for each alternative are counted, and these values represent the distance between each alternative as well as the ideal and negative ideal.

5-Closeness to the ideal solution calculation

In the process, the closeness of A_i to the ideal solution A^* is defined as:

$$C_{i^*} = S_{i^-}/(S_{i^-} + S_{i^*}),\ \ 0 < C_{i^*} < 1\,,\ i = (1,2,\cdots m) \qquad \textbf{(3.26)}$$

Obviously, $C_{i^*} = 1$ if and only if ($A_i = A^*$), similarly, $C_{i^*} = 0$ if and only if ($A_i = A^-$)

6-Ranking the alternative according to the closeness to the ideal solution

The set of the alternative A_i can now be ranked according to the descending order of C_{i^*}, where a high value means better performance.

3.4.3.1 Decision Making Context

Two main decision-making contexts are emphasized based on the individual and group decision makers (GDM). This situation is encountered by individuals when they collectively make a choice among the alternatives presented to them. The decision is then no longer attributable to any individual member of the group because all individuals within social group processes such as social influence contribute to the outcome. GDM techniques systematically collect and combine the knowledge for the judgment of experts from different fields (C.-T. Chen 2000; Y. S. Huang et al. 2013;

Xia and Chen 2015; B. B. Zaidan and Zaidan 2017). Thus, the GDM for each expert or evaluator subjectively provides his or her judgment and weights for the criteria. Figure 3.13 shows the group decision making process.

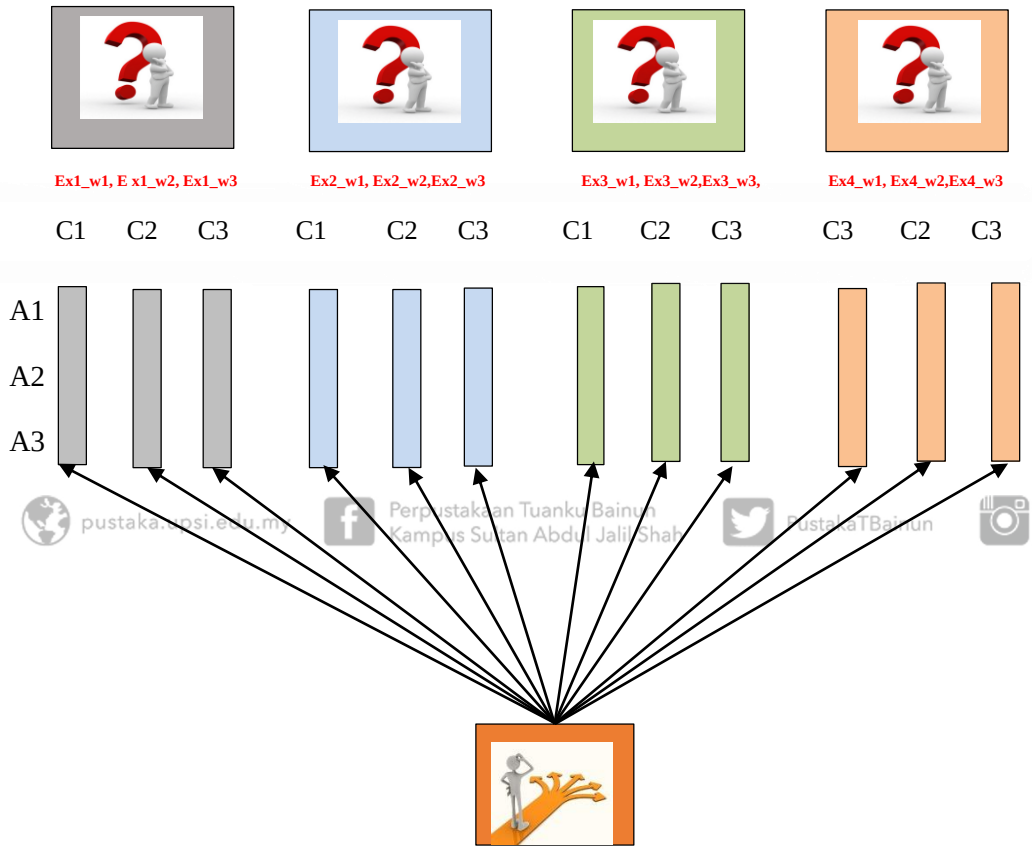


Figure 3.13. Group Decision Maker Process

For example, taking the problem of skin detection evaluation and benchmarking in the context of group decision making, C1 is the reliability, C2 is the time complexity, and C3 is the error rate within the dataset subjectively measured by the evaluators (See Figure 3.14).

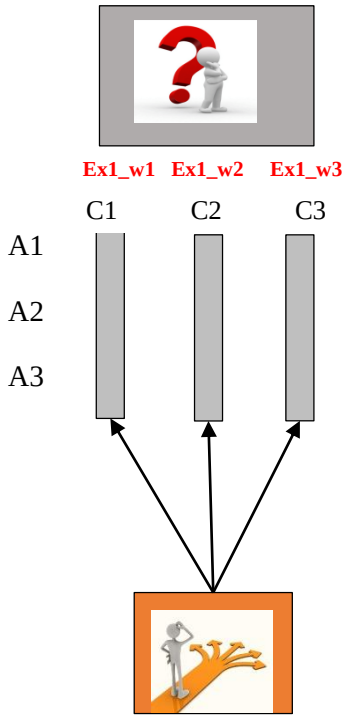


Figure 3.14. Individual Decision Maker Process

In the original context of decision making, two different settings that mainly rely on the problem itself are considered. The first setting is normal decision making where an individual provides subjective judgments and the weights for each criterion. The other setting is group decision making where decision makers provide their subjective judgments and their own weights for each criterion as a group. In our study, the problems of evaluation and benchmarking skin detection for all the data were objectively obtained (numerical numbers). Hence, the context of the decision-making process must be considered.

3.5 Validation Phase

The validation is considered to be an important process in most studies based on the comparison between variables using statistical methods. This phase is implemented to achieve the fourth objective in our study. The multi-criteria evaluation is validated based on the results obtained from the decision matrix. In addition, validation of the final results is obtained from the use of the new methodology using the calculation of mathematical statistics. The results will be discussed in Chapter 6.

3.5.1 Validity of the Multi Criteria Measurement Process

Numerous studies have addressed the concept of trade-off criteria used in various fields, such as management of industrial projects and agricultural and healthcare projects, to evaluate and address cases of uncertainty (Faith, Margules, and Walker 2000; Butler et al. 2013; Jumaah, F. M., Zaidan, A. A., Zaidan, B. B., Bahbibi, R., Qahtan, M. Y., & Sali 2017). This step highlights the problem of the trade-off between the multiple criteria of the skin detection approaches. Therefore, a mathematical process was proposed to prove the trade-off problem between criteria by implementing a multi-criteria measurement process using a numerical sequence to calculate the weights (Jumaah, F. M., Zaidan, A. A., Zaidan, B. B., Bahbibi, R., Qahtan, M. Y., & Sali 2017). The basic principle of this process is a distribution of the weight values from 1 to 0, where the value decreases by 0.1. The results obtained

will be evaluated and compared with the basic criteria in our study. Thus, what is the purpose of implementing the numerical process between criteria in this step?

In addition, these results will be tested using the paired sample t-test method to calculate the t-value and correlation degree based on the significance value for the results (Hedberg and Ayers 2015; Baringhaus and Gaigall 2017). Additional details will be discussed in Chapter 6.

3.5.2 Comparison between Color Spaces

This step highlights the behavior of different color spaces based on the final results obtained from Chapter 5. The color spaces are compared based on nine values of the threshold distributed for each color space. Thus, the main purpose of implement this step to determine the best and worst color space. Further details will be discussed in Chapter 6.

3.5.3 Statistical Measurement for Color Spaces

The last step of this phase is implemented using mathematical statistics for the different color spaces. The final results obtained from the decision-making methods are validated by conducting a comparison between various values of the color spaces. The mean and standard deviation values will be calculated for each color space based

on the values of the external aggregation obtained from Chapter 5 (Bergmeir, Costantini, and Benítez 2014). The results of the comparison show the effect of these values on the ranking of color spaces.

3.6 Chapter Summary

A total of four critical objectives were identified in our research. This chapter highlights the proposed solutions and the ways to achieve the research objectives. Four main phases are designed in this chapter to serve as a guidelin for this research.

The first phase was designed to achieve the first objective. This phase includes conducting a comprehensive survey of the relevant studies in the skin detector approach to gather different criteria and identify their weaknesses. Thus, the research problem is highlighted in relation to the evaluation of criteria and the benchmarking of techniques and tools.

The second phase was designed to achieve the second objective of determining and performing the decision matrix based on the case study adopted in our study. This phase comprises two basic steps: identification and performance decision matrix. The first step develops the case study based on different color spaces. The second step finds the correlation between the criteria and compares them with the developed color spaces. Thus, the results from the two steps generate the parameters representing the values of the final decision matrix to be evaluated and tested.

The development phase is an important step to achieve the third objective in our research which is implemented using MCDM techniques. This phase involves the generation of criteria weights using an AHP technique. Thus, this phase is integrated with the TOPSIS technique based on the decision matrix created in the previous phase. These techniques are implemented to obtain the final results, which are used to rank and select the best alternatives.

In the last phase, the fourth objective is achieved using the validation operation of the research results. The validation for final results has been carried out through conducting mathematical statistics, comparisons and calculate the mean and standard deviation for different color spaces.

CHAPTER 4

MULTI CRITERIA ANALYSIS AND COMPARISON

4.1Background

In this chapter will implement steps that have been proposed in Chapter 3 to obtain the results. These results represented in the decision matrix which obtained after conducting crossover between various criteria and different color spaces. In our study, the decision matrix represents the dataset to be used for determining the rest of the required results according to the proposed methodology. However, numerous methods are used to identify the relationship among different criteria according to their characteristics. Therefore, proving this relationship is necessary to practically determine the effect of one variable on the others. Statistical analysis is an appropriate method to prove the relationship among several parameters. On the other hand, the decision matrix is built from multiple criteria and several color spaces based on the nine threshold values adopted in our study. Therefore, identifying the factors that affect the behavior of the criteria for each color space used is necessary. Thus, a

performance analysis will be conducted to determine and evaluate the behavior of each criterion.

Section 4.2 highlights the decision matrix consisting of multiple criteria and different color spaces. Section 4.3 presents the statistical analysis method based on correlation measurement. In Section 4.4, a performance analysis of criteria is implemented. The chapter summary is presented in Section 4.5. Figure 4.1 illustrates the overview of the results and evaluation of different criteria.

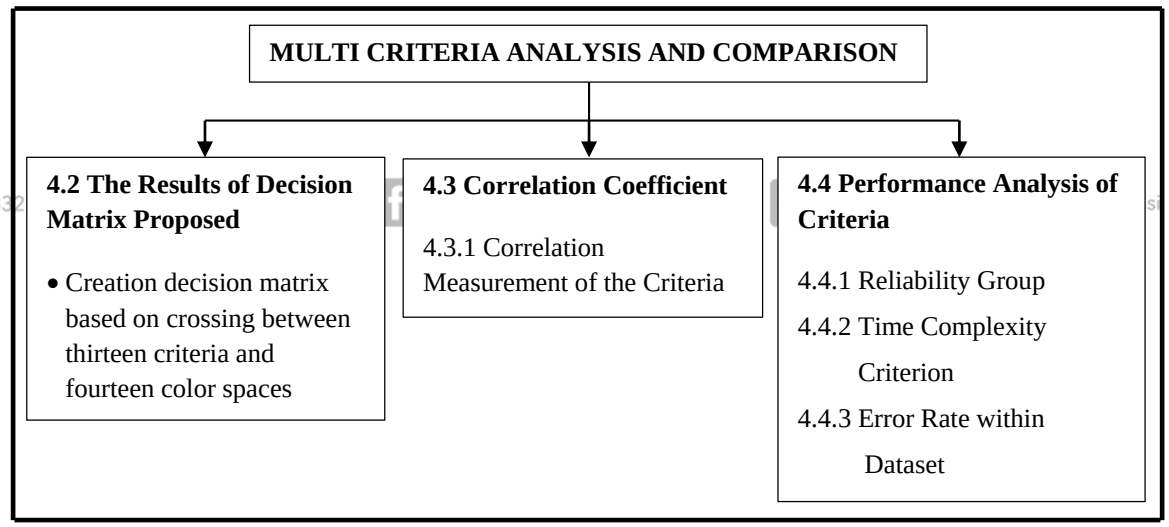


Figure 4.1. Overview of the Results and Evaluation of Different Criteria

4.2 Results of the Proposed Decision Matrix

In this section, the results obtained from the development of a case study were adapted. A total of 14 color spaces were developed based on the multi-agent

technique using two AI models. The outcome of the development process generated parameters which considered fundamental values to calculate the reliability group values. In addition, the values of other basic criteria were calculated according to their respective methodologies. Thus, multiple criteria values with skin detection engine values produced the decision matrix that represents the dataset in our study. The dataset is considered a basic result that will be used in the next phase to generate the final results to select the best alternatives.

The decision matrix was constructed using values from 13 criteria and 14 color spaces. The values of each color space were individually calculated based on the nine threshold values. Thus, the criteria values were calculated based on the threshold values that generated nine different values for each color space. Meanwhile, new values obtained for these criteria according to other color spaces will be implemented. Eventually, 108 algorithms obtained based on 14 color spaces were adopted. However, the final number of color spaces that appeared in the decision matrix is 12 due to the color spaces group of (HSI, HSL, and HSV), where the luminance element is deleted from each color space (I, L and V) and the common Chroma element (HS) between color space group is retained; thus, processing is only applied once. Table 4.1, illustrates the evaluation result for 108 color space samples tested using 13 criteria.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

Table 4.1

Implementation of the Decision Matrix

Color Spaces	TN	TP	FP	FN	Accuracy	Recall	Precision	Specificity	F-measure	G-measure	Tc sec	ERV	ERT
Normalized RGB	79.5	80.06	20.5	19.94	79.9003	0.8006	0.7961	0.7950	0.7984	0.7984	8.10	0.00003	0.00004
	81.32	80.22	18.68	19.78	81.7211	0.8022	0.8111	0.8132	0.8066	0.8066	8.14	0.00003	0.00004
	82.04	79.56	17.96	20.44	82.4378	0.7956	0.8158	0.8204	0.8056	0.8057	8.01	0.00003	0.00004
	83.77	80.91	16.23	19.09	84.1746	0.8091	0.8329	0.8377	0.8208	0.8209	8.08	0.00003	0.00004
	85.84	80.12	14.16	19.88	86.2406	0.8012	0.8498	0.8584	0.8248	0.8251	8.25	0.00003	0.00004
	90.45	81.3	9.55	18.7	90.8565	0.813	0.8949	0.9045	0.852	0.853	8.31	0.00003	0.00004
	91.84	79.99	8.16	20.01	92.24	0.7999	0.9074	0.9184	0.8503	0.852	8.32	0.00003	0.00004
	93.72	78.62	6.28	21.38	94.1131	0.7862	0.926	0.9372	0.8504	0.8533	8.41	0.00003	0.00004
	92.91	79.14	7.09	20.86	93.3057	0.7914	0.9178	0.9291	0.8499	0.8522	8.35	0.00003	0.00004
YCbCr	80.22	81.92	19.78	18.08	80.6296	0.8192	0.8055	0.8022	0.8123	0.8123	9.82	0.0001	0.00007
	85.27	84.37	14.73	15.63	85.6919	0.8437	0.8514	0.8527	0.8475	0.8475	9.84	0.0001	0.00007
	86.49	85.78	13.51	14.22	86.9189	0.8578	0.8639	0.8649	0.8609	0.8609	9.86	0.0001	0.00007
	88.95	86.61	11.05	13.39	89.3831	0.8661	0.8869	0.8895	0.8764	0.8764	9.91	0.0001	0.00007
	91.64	89.27	8.36	10.73	92.0864	0.8927	0.9144	0.9164	0.9034	0.9035	9.95	0.0001	0.00007
	91.9	90.85	8.1	9.15	92.3543	0.9085	0.9181	0.919	0.9133	0.9133	9.85	0.0001	0.00007
	96.55	94.53	3.45	5.47	97.0227	0.9453	0.9648	0.9655	0.9549	0.955	9.96	0.0001	0.00007
	99.55	98.68	0.45	1.32	100.043	0.9868	0.9955	0.9955	0.9911	0.9911	9.9	0.0001	0.00007
	95.3	93.94	4.7	6.06	95.7697	0.9394	0.9524	0.953	0.9458	0.9459	9.94	0.0001	0.00007
YCgCr	81.72	82.6	18.28	17.4	82.133	0.826	0.8188	0.8172	0.8224	0.8224	11.61	0.00015	0.0001
	84.64	85.12	15.36	14.88	85.0656	0.8512	0.8471	0.8464	0.8492	0.8492	11.63	0.00015	0.0001
	85.84	86.47	14.16	13.53	86.2724	0.8647	0.8593	0.8584	0.862	0.862	11.64	0.00015	0.0001
	86.62	88.84	13.38	11.16	87.0642	0.8884	0.8691	0.8662	0.8786	0.8787	11.62	0.00015	0.0001
	89.47	91.81	10.53	8.19	89.9291	0.9181	0.8971	0.8947	0.9075	0.9075	11.71	0.00015	0.0001
	92.56	90.81	7.44	9.19	93.0141	0.9081	0.9243	0.9256	0.9161	0.9162	11.73	0.00015	0.0001
	93.59	97.38	6.41	2.62	94.0769	0.9738	0.9382	0.9359	0.9557	0.9559	11.66	0.00015	0.0001
	98.63	99.66	1.37	0.34	99.1283	0.9966	0.9864	0.9863	0.9915	0.9915	11.74	0.00015	0.0001
	93.49	96.59	6.51	3.41	93.973	0.9659	0.9369	0.9349	0.9512	0.9513	11.71	0.00015	0.0001
YCgCb	82.59	81.87	17.41	18.13	82.9994	0.8187	0.8246	0.8259	0.8217	0.8217	11.94	0.00016	0.00015
	85.08	84.73	14.92	15.27	85.5037	0.8473	0.8503	0.8508	0.8488	0.8488	11.92	0.00016	0.00015
	86.41	85.89	13.59	14.11	86.8395	0.8589	0.8634	0.8641	0.8611	0.8611	11.91	0.00016	0.00015
	88.74	86.68	11.26	13.32	89.1734	0.8668	0.885	0.8874	0.8758	0.8759	11.9	0.00016	0.00015
	91.79	89.53	8.21	10.47	92.2377	0.8953	0.916	0.9179	0.9055	0.9056	11.94	0.00016	0.00015
	90.84	92.64	9.16	7.36	91.3032	0.9264	0.91	0.9084	0.9181	0.9182	11.92	0.00016	0.00015
	97.25	93.65	2.75	6.35	97.7183	0.9365	0.9715	0.9725	0.9537	0.9538	11.92	0.00016	0.00015
	99.61	98.67	0.39	1.33	100.103	0.9867	0.9961	0.9961	0.9914	0.9914	11.94	0.00016	0.00015
	96.52	93.52	3.48	6.48	96.9876	0.9352	0.9641	0.9652	0.9494	0.9496	11.93	0.00016	0.00015

(*Continue*)

Table 2.1 (*continued*)

Color Spaces	TN	TP	FP	FN	Accuracy	Recall	Precision	Specificity	F-measure	G-measure	Tc sec	ERV	ERT
YUV	82.84	81.46	17.16	18.54	83.2473	0.8146	0.826	0.8284	0.8203	0.8203	12.02	0.00018	0.00013
	85.72	84.14	14.28	15.86	86.1407	0.8414	0.8549	0.8572	0.8481	0.8481	12.19	0.00018	0.00013
	86.68	85.37	13.32	14.63	87.1069	0.8537	0.865	0.8668	0.8593	0.8593	12.09	0.00018	0.00013
	87.58	87.73	12.42	12.27	88.0187	0.8773	0.876	0.8758	0.8766	0.8766	12.23	0.00018	0.00013
	89.54	90.61	10.46	9.39	89.9931	0.9061	0.8965	0.8954	0.9013	0.9013	12.25	0.00018	0.00013
	93.27	89.62	6.73	10.38	93.7181	0.8962	0.9302	0.9327	0.9129	0.913	12.35	0.00018	0.00013
	93.62	96.8	6.38	3.2	94.104	0.968	0.9382	0.9362	0.9528	0.953	12.08	0.00018	0.00013
	96.81	98.9	3.19	1.1	97.3045	0.989	0.9688	0.9681	0.9788	0.9788	12.31	0.00018	0.00013
	94.04	95.42	5.96	4.58	94.5171	0.9542	0.9412	0.9404	0.9477	0.9477	12.42	0.00018	0.00013
YIQ	82.51	80.72	17.49	19.28	82.9136	0.8072	0.8219	0.8251	0.8145	0.8145	12.81	0.00019	0.00015
	85.02	84.71	14.98	15.29	85.4436	0.8471	0.8497	0.8502	0.8484	0.8484	12.83	0.00019	0.00015
	84.27	85.38	15.73	14.62	84.6969	0.8538	0.8444	0.8427	0.8491	0.8491	12.85	0.00019	0.00015
	86.31	86.71	13.69	13.29	86.7436	0.8671	0.8636	0.8631	0.8654	0.8654	12.86	0.00019	0.00015
	91.48	88.36	8.52	11.64	91.9218	0.8836	0.9121	0.9148	0.8976	0.8977	12.86	0.00019	0.00015
	88.71	92.01	11.29	7.99	89.1701	0.9201	0.8907	0.8871	0.9052	0.9053	12.88	0.00019	0.00015
	97.82	91.38	2.18	8.62	98.2769	0.9138	0.9767	0.9782	0.9442	0.9447	12.92	0.00019	0.00015
	99.93	95.85	0.07	4.15	100.409	0.9585	0.9993	0.9993	0.9785	0.9787	12.95	0.00019	0.00015
	96.37	92.26	3.63	7.74	96.8313	0.9226	0.9621	0.9637	0.942	0.9422	12.95	0.00019	0.00015
HSI, HSV, HSL	80.55	81.11	19.45	18.89	80.9556	0.8111	0.8066	0.8055	0.8088	0.8088	8.51	0.00005	0.00006
	82.37	81.27	17.63	18.73	82.7764	0.8127	0.8217	0.8237	0.8172	0.8172	8.6	0.00005	0.00006
	83.09	80.61	16.91	19.39	83.4931	0.8061	0.8266	0.8309	0.8162	0.8163	8.53	0.00005	0.00006
	84.82	81.96	15.18	18.04	85.2298	0.8196	0.8437	0.8482	0.8315	0.8316	8.73	0.00005	0.00006
	86.89	81.17	13.11	18.83	87.2959	0.8117	0.8609	0.8689	0.8356	0.836	8.7	0.00005	0.00006
	91.5	82.35	8.5	17.65	91.9118	0.8235	0.9064	0.915	0.863	0.864	8.77	0.00005	0.00006
	92.8	79.63	7.2	20.37	93.1982	0.7963	0.9171	0.928	0.8524	0.8546	8.52	0.00005	0.00006
	92.64	76.56	7.36	23.44	93.0228	0.7656	0.9123	0.9264	0.8325	0.8357	8.55	0.00005	0.00006
	93.95	75.37	6.05	24.63	94.3269	0.7537	0.9257	0.9395	0.8309	0.8353	8.72	0.00005	0.00006
IHLS	81.61	82.37	18.39	17.63	82.0219	0.8237	0.8175	0.8161	0.8206	0.8206	9.12	0.00011	0.00009
	83.75	83.43	16.25	16.57	84.1672	0.8343	0.837	0.8375	0.8356	0.8356	9.18	0.00011	0.00009
	84.92	84.45	15.08	15.55	85.3423	0.8445	0.8485	0.8492	0.8465	0.8465	9.12	0.00011	0.00009
	85.26	85.67	14.74	14.33	85.6884	0.8567	0.8532	0.8526	0.8549	0.8549	9.23	0.00011	0.00009
	87.47	85.28	12.53	14.72	87.8964	0.8528	0.8719	0.8747	0.8622	0.8623	9.31	0.00011	0.00009
	92.25	84.93	7.75	15.07	92.6747	0.8493	0.9164	0.9225	0.8816	0.8822	9.32	0.00011	0.00009
	93.44	87.73	6.56	12.27	93.8787	0.8773	0.9304	0.9344	0.9031	0.9035	9.4	0.00011	0.00009
	96.14	86.28	3.86	13.72	96.5714	0.8628	0.9572	0.9614	0.9075	0.9088	9.37	0.00011	0.00009
	94.35	87.34	5.65	12.66	94.7867	0.8734	0.9392	0.9435	0.9051	0.9057	9.19	0.00011	0.00009

(*Continue*)

Table 2.1 (*continued*)

Color Spaces	TN	TP	FP	FN	Accuracy	Recall	Precision	Specificity	F-measure	G-measure	Tc sec	ERV	ERT
CIE-XYZ	82.54	81.89	17.46	18.11	82.9495	0.8189	0.8243	0.8254	0.8216	0.8216	10.33	0.00016	0.00013
	85.37	84.21	14.63	15.79	85.7911	0.8421	0.852	0.8537	0.847	0.847	10.38	0.00016	0.00013
	86.55	85.82	13.45	14.18	86.9791	0.8582	0.8645	0.8655	0.8613	0.8613	10.39	0.00016	0.00013
	88.97	86.58	11.03	13.42	89.4029	0.8658	0.887	0.8897	0.8763	0.8763	10.41	0.00016	0.00013
	91.73	89.22	8.27	10.78	92.1761	0.8922	0.9152	0.9173	0.9035	0.9036	10.43	0.00016	0.00013
	91.93	90.81	8.07	9.19	92.3841	0.9081	0.9184	0.9193	0.9132	0.9132	10.32	0.00016	0.00013
	96.63	94.32	3.37	5.68	97.1016	0.9432	0.9655	0.9663	0.9542	0.9543	10.37	0.00016	0.00013
	99.61	98.52	0.39	1.48	100.103	0.9852	0.9961	0.9961	0.9906	0.9906	10.44	0.00016	0.00013
	95.25	93.98	4.75	6.02	95.7199	0.9398	0.9519	0.9525	0.9458	0.9458	10.44	0.00016	0.00013
CIE-LAB	82.6	81.79	17.4	18.21	83.009	0.8179	0.8246	0.826	0.8212	0.8212	10.52	0.00017	0.00014
	85.25	84.41	14.75	15.59	85.6721	0.8441	0.8513	0.8525	0.8477	0.8477	10.55	0.00017	0.00014
	86.08	85.98	13.92	14.02	86.5099	0.8598	0.8607	0.8608	0.8602	0.8602	10.57	0.00017	0.00014
	88.91	86.37	11.09	13.63	89.3419	0.8637	0.8862	0.8891	0.8748	0.8749	10.54	0.00017	0.00014
	91.81	89.16	8.19	10.84	92.2558	0.8916	0.9159	0.9181	0.9036	0.9037	10.62	0.00017	0.00014
	90.9	92.49	9.1	7.51	91.3625	0.9249	0.9104	0.909	0.9176	0.9176	10.67	0.00017	0.00014
	97.43	93.36	2.57	6.64	97.8968	0.9336	0.9732	0.9743	0.953	0.9532	10.68	0.00017	0.00014
	99.57	98.57	0.43	1.43	100.063	0.9857	0.9957	0.9957	0.9907	0.9907	10.64	0.00017	0.00014
	96.38	93.63	3.62	6.37	96.8482	0.9363	0.9628	0.9638	0.9494	0.9494	10.66	0.00017	0.00014
CIE-LUV	81.66	82.64	18.34	17.36	82.0732	0.8264	0.8184	0.8166	0.8224	0.8224	10.14	0.00014	0.00011
	84.52	85.25	15.48	14.75	84.9463	0.8525	0.8463	0.8452	0.8494	0.8494	10.16	0.00014	0.00011
	85.99	86.56	14.01	13.44	86.4228	0.8656	0.8607	0.8599	0.8631	0.8631	10.15	0.00014	0.00011
	86.57	88.93	13.43	11.07	87.0147	0.8893	0.8688	0.8657	0.8789	0.879	10.13	0.00014	0.00011
	89.34	91.87	10.66	8.13	89.7994	0.9187	0.896	0.8934	0.9072	0.9073	10.25	0.00014	0.00011
	92.49	90.92	7.51	9.08	92.9446	0.9092	0.9237	0.9249	0.9164	0.9164	10.22	0.00014	0.00011
	93.41	97.42	6.59	2.58	93.8971	0.9742	0.9366	0.9341	0.9551	0.9552	10.28	0.00014	0.00011
	98.59	99.71	1.41	0.29	99.0886	0.9971	0.9861	0.9859	0.9915	0.9916	10.25	0.00014	0.00011
	93.46	96.62	6.54	3.38	93.9431	0.9662	0.9366	0.9346	0.9512	0.9513	10.29	0.00014	0.00011
CIE-Lch	80.02	81.72	19.98	18.28	80.4286	0.8172	0.8035	0.8002	0.8103	0.8103	9.54	0.00013	0.00012
	86.56	84.48	13.44	15.52	86.9824	0.8448	0.8627	0.8656	0.8537	0.8537	9.57	0.00013	0.00012
	87.5	85.62	12.5	14.38	87.9281	0.8562	0.8726	0.875	0.8643	0.8644	9.58	0.00013	0.00012
	88.68	86.53	11.32	13.47	89.1127	0.8653	0.8843	0.8868	0.8747	0.8748	9.62	0.00013	0.00012
	91.34	89.02	8.66	10.98	91.7851	0.8902	0.9113	0.9134	0.9006	0.9007	9.65	0.00013	0.00012
	91.73	90.6	8.27	9.4	92.183	0.906	0.9164	0.9173	0.9111	0.9112	9.55	0.00013	0.00012
	96.05	94.33	3.95	5.67	96.5217	0.9433	0.9598	0.9605	0.9515	0.9515	9.61	0.00013	0.00012
	99.04	98.62	0.96	1.38	99.5331	0.9862	0.9904	0.9904	0.9883	0.9883	9.73	0.00013	0.00012
	95.27	99.84	4.73	0.16	95.7692	0.9984	0.9548	0.9527	0.9761	0.9763	9.78	0.00013	0.00012

Table 4.1 shows the decision matrix obtained according to the proposed methodology. The decision matrix comprises 13 criteria and 14 color spaces, thereby producing 108 skin detector engines. The relationship between these criteria will be determined using Pearson test and performance analysis for each criterion to determine the behavior based on various threshold values that will be discussed in detail below.

4.3 Correlation Coefficient

The present study adopted various parameters used in skin detection approaches. Three main groups of criteria are used, namely reliability, time complexity, and error rate within dataset groups mentioned in detail in Chapter 2. Therefore, the Pearson formula is applied to compute the correlation coefficient between different criteria, which is highlighted in the following section.

4.3.1 Correlation Measurement of the Criteria

In this section, multiple criteria are evaluated and tested based on data of criteria. In the present study, various criteria influencing one another have been independently collected. Therefore, determining the relationship between criteria and verifying the degree of correlation between them are imperative. According to the literature, various mathematical and statistical methods can be used to prove such relationship. One of the most important methods used to statistically measure the degree of

correlation is the Pearson method. This method focuses on finding the value of *r* based on Equation 3.12. Figure 4.2 shows the taxonomy of criteria distribution into three main layers.

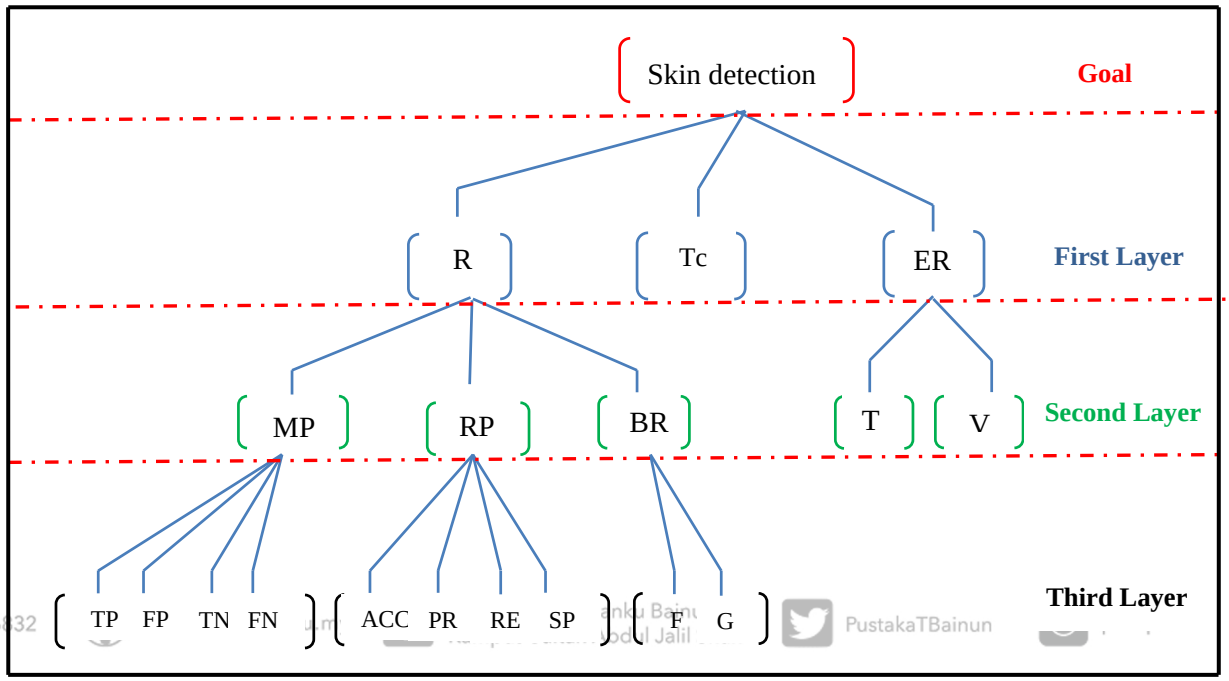


Figure 4.2. Taxonomy of Criteria Distribution into Three Layers

Figure 4.2 illustrates the taxonomy constructed from three layers comprising three major sets of criteria in our study. The first layer includes reliability (R), time complexity (Tc), and error rate within the dataset (ER) groups. Meanwhile, the second layer includes three key sections, such as matrix, relationship, and behavior of parameters, which are derived from the reliability criterion. In addition, the validation and training criteria are derived from the error ratio criterion. The third layer comprises 10 criteria. Among the 10 criteria, four are called confusion matrix, namely true positive (TP), true negative (TN), false positive (FP), and false negative (FN),

which are derived from the matrix of parameters. The other four criteria are accuracy (ACC), precision (PR), recall (RE), and specificity (SP), which are derived from the relationship of parameters. The final two criteria are F-measure (F) and G-measure (G), which are both derived from the behavior of parameters. Details of the correlation tests conducted between these criteria are presented below.

4.3.1.1 Correlation Analysis in Layer 1

This layer includes three independent criteria, namely reliability group, time complexity group, and error rate. The Pearson method is implemented to determine the extent of the relationship and correlation degree among the criteria. After conducting the test and selecting the desired path to determine the correlation among the criteria, we obtain the following results as shown in Table 4.2.

Table 4.2

Comparison of Reliability, Time Complexity, and Error Rate Criteria

Correlation Coefficient				
		R	**Tc**	**ER**
R	Pearson Correlation	1	-.239*	-.260**
	Sig. (2-tailed)		.013	.007
	N	108	108	108
Tc	Pearson Correlation	-.239*	1	.862**
	Sig. (2-tailed)	.013		.000
	N	108	108	108
ER	Pearson Correlation	-.260**	.862**	1
	Sig. (2-tailed)	.007	.000	
	N	108	108	108

*. Correlation is significant at the 0.05 level (2-tailed).
**. Correlation is significant at the 0.01 level (2-tailed).

Table 4.2 illustrate details the relationship and degree of correlation between criteria based on the rules of the correlation coefficient according to the Pearson test. Therefore, the correlation coefficient between the reliability and time complexity group was − 0.239, which indicates that a reverse correlation exists when ($r < 0$) at a degree of significance 0.013 for 108 samples. In addition, a high correlation exists in the negative aspect while the correlation value in the error rate group was −0.260, in which a reverse correlation exists when ($r < 0$) at a degree of significance 0.007 for 108 samples. Meanwhile, the correlation degree between the time complexity and error rate group was 0.892, which indicates that a positive correlation exists when ($r > 0$) at a degree of significance 0.000 for 108 samples; thus, a high correlation exists in the positive aspect. Overall, the existing correlation between each criterion is proven based on the rules of Pearson test results.

4.3.1.2 Correlation Analysis in Layer 2

The second layer includes two groups of the sub-criteria. The first group included three basic sub-criteria generated of reliability group. Whereas, the second group included two sub-criteria derived from error rate group. Pearson test will be used to calculate the correlation coefficient between each group as following in the Table 4.3.

Table 4.3

Comparison among Matrix of Parameters, Relationship of Parameters, and Behavior of Parameter Sub-Criteria

Correlation Coefficient				
		MP	**RP**	**BP**
MP	Pearson Correlation	1	-.973**	-.953**
	Sig. (2-tailed)		.000	.000
	N	108	108	108
RP	Pearson Correlation	-.973**	1	.997**
	Sig. (2-tailed)	.000		.000
	N	108	108	108
BP	Pearson Correlation	-.953**	.997**	1
	Sig. (2-tailed)	.000	.000	
	N	108	108	108
**. Correlation is significant at the 0.01 level (2-tailed).				

Table 4.3 shows the relationship and correlation degree among sub-criteria based on the correlation coefficient rules of the Pearson test. The table highlights the correlation among three sub-criteria, namely the matrix, relationship, and behavior of parameters, which are all derived from the reliability group. The correlation coefficient between the matrix and the relationship of parameters is −0.973, which indicates that a reverse correlation exists when ($r < 0$) at a degree of significance 0.000 for 108 samples; a high correlation exists in the negative aspect, while the correlation value for the behavior of parameters is −0.953, in which a reverse correlation also exists when ($r < 0$) at a degree of significance 0.000 for 108 samples. Meanwhile, the correlation degree between the relationship and the behavior of parameters is 0.997, which indicates that a positive correlation exists when ($r > 0$) at a

degree of significance 0.000 for 108 samples. Overall, the existing correlation between each criterion is proven based on the Pearson test results.

Table 4.4

Comparison Training and Validation Sub-Criteria

Correlation Coefficient			
		T	**V**
T	Pearson Correlation	1	.942**
	Sig. (2-tailed)		.000
	N	108	108
V	Pearson Correlation	.942**	1
	Sig. (2-tailed)	.000	
	N	108	108
**. Correlation is significant at the 0.01 level (2-tailed).			

Table 4.4 depicts the relationship and degree of correlation between the sub-criteria based on the correlation analysis using the Pearson test. The table highlights the correlation between validation and training sub-criteria, which are both derived from the error rate within the dataset group. The correlation degree between the validation and training criterion is 0.942, which indicates that a positive correlation exists when ($r > 0$) at the degree of significance 0.000 for 108 samples. Overall the existing correlation between each criterion is proven based on the Pearson test results.

4.3.1.3 Correlation Analysis in Layer 3

Layer 3 comprises three groups of sub-sub-criteria. The first group includes four parameters generated from the matrix of parameter group, namely TP, FP, TN, and FN. The second group includes four parameters derived from the relationship of parameter group, namely accuracy, precision, recall, and specificity. The third group includes two parameters, namely F-measure and G-measure, which are both generated from the behavior of parameters. The Pearson test is implemented to calculate the correlation coefficient for each group as shown in Table 4.5.

Table 4.5

Comparison among TP, FP, TN, and FN Sub-Sub-Criteria

Correlation Coefficient					
		TP	**FP**	**TN**	**FN**
TP	Pearson Correlation	1	-.744**	.744**	-1.000**
	Sig. (2-tailed)		.000	.000	.000
	N	108	108	108	108
FP	Pearson Correlation	-.744**	1	-1.000**	.744**
	Sig. (2-tailed)	.000		.000	.000
	N	108	108	108	108
TN	Pearson Correlation	.744**	-1.000**	1	-.744**
	Sig. (2-tailed)	.000	.000		.000
	N	108	108	108	108
FN	Pearson Correlation	-1.000**	.744**	-.744**	1
	Sig. (2-tailed)	.000	.000	.000	
	N	108	108	108	108

**. Correlation is significant at the 0.01 level (2-tailed).

Table 4.5 shows the relationship and degree of correlation between sub-sub-criteria based on the correlation coefficient rules according to the Pearson test. The table highlights the correlation among four sub-sub-criteria, namely TN, FP, TN, and FN, which are derived from the matrix of parameters. The correlation coefficient between TP and FP is −0.744, which indicates that a reverse correlation exists when ($r < 0$) at the degree of significance 0.000 for 108 samples. Meanwhile, a high correlation exists at value 0.744 with TN also shows a positive correlation when ($r > 0$) at the degree of significance 0.000 for 108 samples. The correlation degree with FN is −1.000, which indicates that a reverse correlation exists when ($r < 0$) at the degree of significance 0.000 for 108 samples. Meanwhile, the correlation degree between the FP and TN is −1.000, which indicates that a reverse correlation exists when ($r < 0$) at the degree of significance 0.000 for 108 samples, and the correlation with FN is 0.744, thereby indicating that a positive correlation exists when ($r > 0$) at the degree of significance 0.000 108 samples. Finally, the correlation between TN and FN is −0.744, which indicates that a reverse correlation exists when ($r < 0$) at the degree of significance 0.000 for 108 samples. Overall, the existing correlation between each sub-sub-criterion is proven by the Pearson test results.

Table 4.6

Comparison of Accuracy, Precision, Recall and Specificity Sub-Sub-Criteria

Correlations					
		ACC	**PR**	**RE**	**SP**
ACC	Pearson Correlation	1	1.000**	.741**	1.000**
	Sig. (2-tailed)		.000	.000	.000
	N	108	108	108	108
PR	Pearson Correlation	1.000**	1	.744**	1.000**
	Sig. (2-tailed)	.000		.000	.000
	N	108	108	108	108
RE	Pearson Correlation	.741**	.744**	1	.744**
	Sig. (2-tailed)	.000	.000		.000
	N	108	108	108	108
SP	Pearson Correlation	1.000**	1.000**	.744**	1
	Sig. (2-tailed)	.000	.000	.000	
	N	108	108	108	108
**. Correlation is significant at the 0.01 level (2-tailed).					

Table 4.6 shows the details of the relationship and degree of correlation among sub-sub-criteria based on the Pearson test. The table highlights the correlation among four sub-sub-criteria, namely accuracy, recall, precision, and specificity, which are derived from the relationship of parameters. The correlation coefficient between the accuracy and precision is 1.000, which means that a positive correlation exists when ($r > 0$) at the degree of significance 0.000 for 108 samples. A high correlation exists in the positive aspect, while the correlation value with recall is 0.741, where a positive correlation also exists when ($r > 0$) at the degree of significance 0.000 for 108 samples. The correlation degree with specificity is 1.000, which means that a positive correlation exists when ($r > 0$) at the degree of significance 0.000 for 108 samples.

Meanwhile, the correlation degree between the precision and recall is 0.744, which indicates that a positive correlation exists when ($r > 0$) at the degree of significance 0.000 for 108 samples. The correlation with specificity is 1.000, which indicates that a positive correlation exists when ($r > 0$) at the degree of significance 0.000 for 108 samples. Finally, the correlation between recall and specificity is 0.744, which means that a positive correlation exists when ($r > 0$) at the degree of significance 0.000 for 108 samples. Overall, the existing correlation between each sub-sub-criterion is proven based on the Pearson test results.

Table 4.7

Comparison between F-measure and G-measure Sub-Sub-Criteria

Correlation Coefficient			
		F	**G**
F	Pearson Correlation	1	.951**
	Sig. (2-tailed)		.000
	N	108	108
G	Pearson Correlation	.951**	1
	Sig. (2-tailed)	.000	
	N	108	108
**. Correlation is significant at the 0.01 level (2-tailed).			

Table 4.7 depicts the relationship and degree of correlation between sub-sub-criteria based on the Pearson test. The table highlights the correlation between two sub-sub-criteria, namely the F-measure and the G-measure, which are both derived from the behavior of parameters. Thus, the correlation degree between the F-measure and the G-measure is 0.951, which indicates that a positive correlation exists when ($r > 0$) at

the degree of significance 0.000 for 108 samples. Overall, the existing correlation between each criterion is proven based on the Pearson test results.

4.3.1.4 Summery

Overall, the Pearson test generally showed a strong correlation among the criteria based on the aforementioned results. The main objective of using the Pearson test is to determine the effect of one criterion on the other, which is dependent on the degree of correlation between data of ctreria. According to the results, a significant negative correlation exists when the value of ($r < 0$). By contrast, a significant correlation exists on the positive aspect when the correlation value was ($r > 0$). Thus, these results were identical with Pearson rules to calculate the correlation value between data of criteria. Thus, the results investigated the second objective mentioned in chapter 3.

4.4 Performance Analysis of Criteria

In this section, performance analysis is conducted for each criterion to determine the factors that affect their behavior based on the nine threshold values for each color space. The three main groups of criteria used in this study are reliability, time complexity, and error rate within dataset groups.

4.4.1 Reliability Group

In this section, we highlight the performance based on nine threshold values for each color space that affects the reliability group, which includes three key sections: 1) the matrix of parameters (TN, TP, FP, and FN), 2) the relationship of parameters (accuracy, recall, precision, and specificity), and 3) the behavior of parameters (F-measure and G-measure).

4.4.1.1 Matrix of Parameters

The matrix of parameters, which is considered the first and most important section in the reliability group, includes four key parameters, namely TN, TP, FP, and FN. The matrix of parameters is also one of the evaluation techniques for skin detection approaches.

A) True Negative Criterion

In this section, the behavior of the first criterion within the matrix of parameters is discussed and analyzed. Figure 4.3 illustrates the behavior of true negative criterion based on nine threshold values with different color spaces.

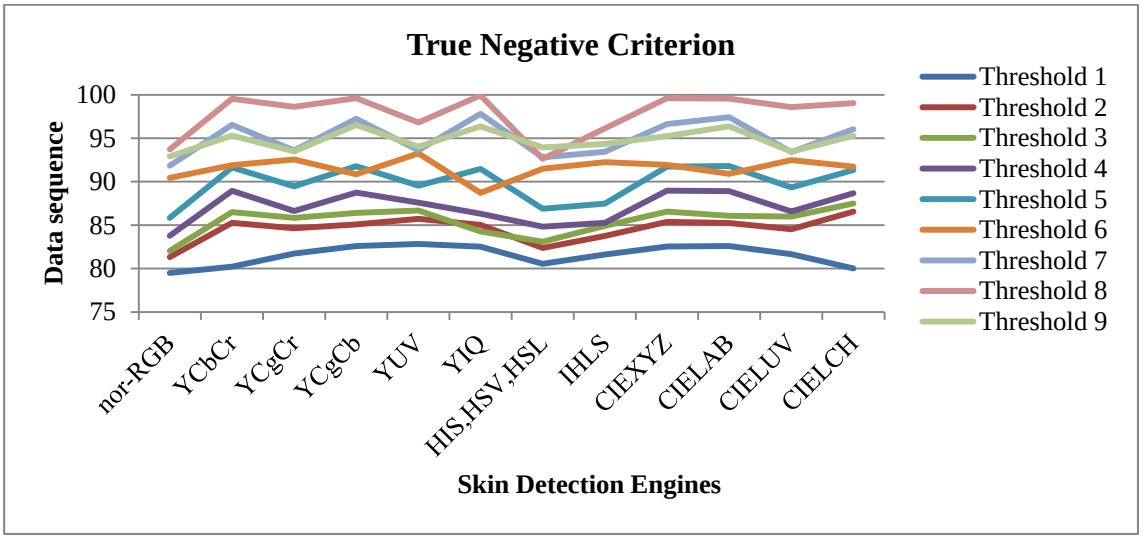


Figure 4.3. Behavior of True Negative Criterion with Different Colors

Figure 4.3 illustrates the behavior of the true negative criterion at different threshold values with various color spaces used as alternatives in the study. The graph shows the behavior of this criterion to appear fairly similar at each threshold value. However, the figure shows the lowest threshold value of the criterion at 79.5%, while the highest value is at 99.04%. The path of each threshold is nearly similar according to the color spaces, except for thresholds 1 and 6. The behavior of the criterion is shown at the value of threshold 1, which starts to slightly increase from Norm-RGB to YIQ. Then, the value slightly drops at the HIS, HSV, and HSL, followed by a slight increase until CIELUV and drops again in CIELCH. By contrast, thresholds 2 and 3 exhibit the same behavior, in which started at rising from Norm-RGB to YCbCr and then stabilizes their track to HIS, HSV, HSL. Such behavior begins to slightly decline then slightly increases until CIEXYZ and then gradually their track rising even CIELCH. Thresholds 4, 5, 7, and 9 starts to slightly increase from Norm-RGB to YCbCr and then begin to drop and increase as well as CIELCH. Meanwhile, threshold

6 has a reverse behavior where in the threshold starts to slightly increase from Norm-RGB to YCgCr and starts to fall and then increases to YIQ. Then, the threshold slightly increases and then stabilizes its track until CIELUV and then falls until CIELCH. Finally, threshold 8 starts to slightly increase from Norm-RGB to YCbCr and then begins to increase, thereby dramatically falling and sharply dropping in HIS, HSV, HSL and starts to slightly increase in CIEXYZ where its track settled down to the end.

B) True Positive Criterion

Figure 4.4, shows the behavior of true positive criterion according to the changes in the threshold values and color space. Notably, the behavior is different from the previous criterion as shown below.

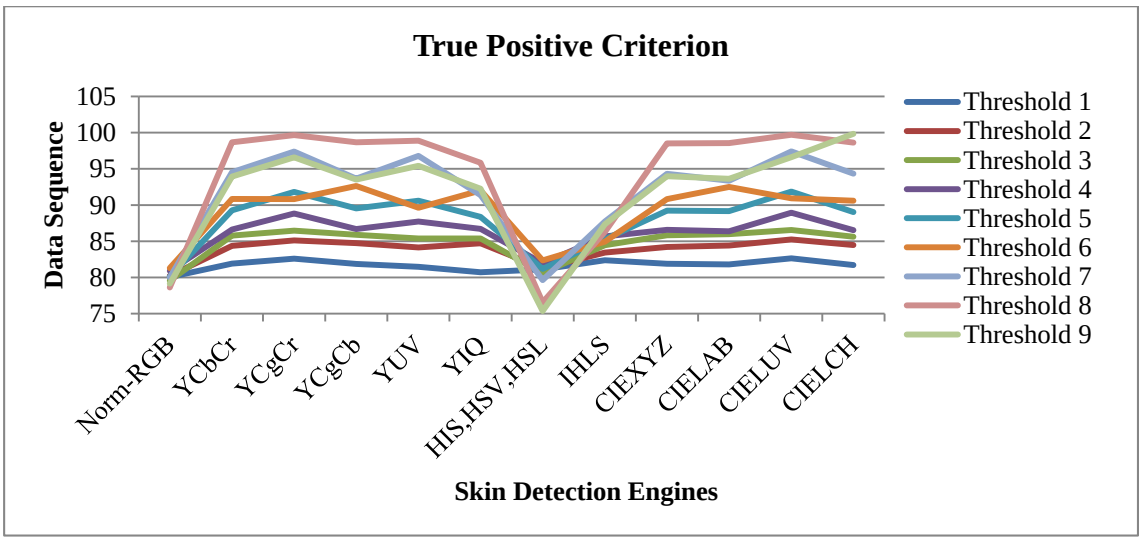


Figure 4.4. Behavior of the True Positive Criterion with Different Color Spaces

Notably, Figure 4.4, illustrates the behavior of the true positive criterion at different threshold values with various color spaces used as alternatives in this research. The figure shows the lowest threshold value of the criterion at 80.06%, while the highest value is 99.84%. Generally, nearly all threshold values start at one point for Norm-RGB. Thus, the behavior of this criterion is affected according to the changes in the threshold track. Threshold 1 starts to slightly increase until YCgCr, its track evenly stabilizes to YIQ, and then slightly increases until IHLS and stabilizes until the end. Thresholds 2 and 3 begin to slightly increase until YCbCr, stabilize their track until YIQ, slightly drop at HIS, HSV, and HSL, and then slightly increase until the end of its track. Thresholds 4, 5, and 7 nearly have similar tracks from start to end. These thresholds start to increase to YCbCr, where tracks change between high and low until YIQ and then drop to a minimum value of HIS, HSV, and HSL. The thresholds increase again until CIEXYZ; thus, their track stabilizes until CIELAB and slightly increase and decrease to the end of their tracks. Threshold 9 has a similar track with the previous threshold but sharply drops at HIS, HSV, and HSL; then, the threshold sharply increases to CIEXYZ, settles in CIELAB, and then increases to its end. However, threshold 6 has a different track, which evenly increases to YCbCr, settles in YCgCr, and changes its track between high and low until YIQ. Then, the track drops to the lowest value at HIS, HSV, and HSL, increases to CIELAB, and then slightly decreases until the end. Finally, threshold 8 represents the highest threshold value, which begins to increase until YCbCr, stabilizes until it sharply drops to the lowest value at HIS, HSV, and HSL, and increases again until CIEXYZ; its track stabilizes until the end.

C) False Positive Criterion

Figure 4.5, shows the behavior of false positive criterion based on different threshold values with various color spaces as in the graph below.

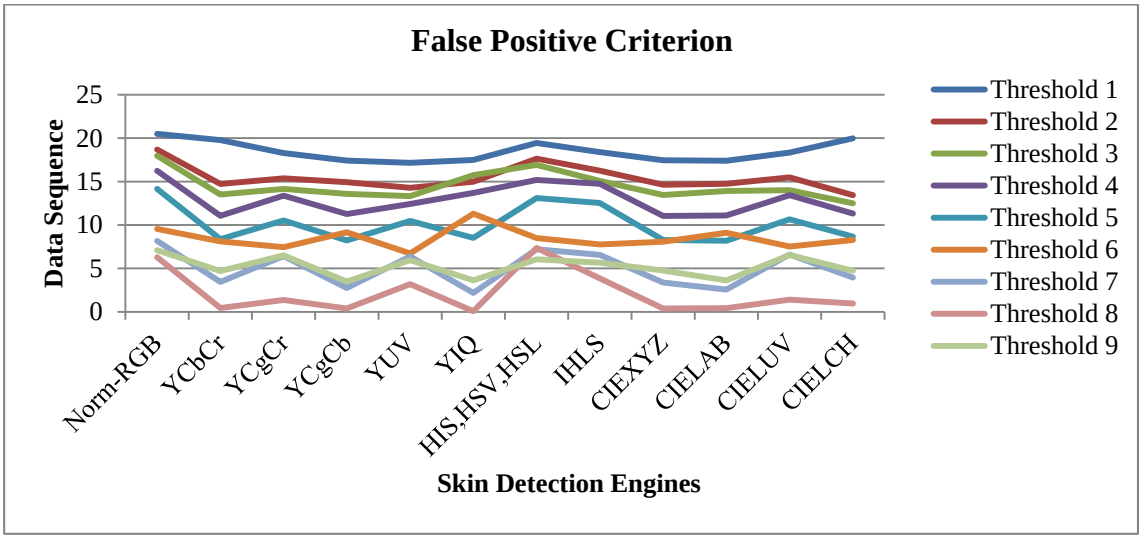


Figure 4.5. Behavior of the False Positive Criterion with Different Color Spaces

Initially, this criterion is considered a complementary to the true negative criterion within probabilistic parameters of the reliability group. However, the figure shows that the lowest threshold value of the criterion is 0.07%, while the highest value is 20.5%. Thus, the track of each threshold is nearly similar according to the color spaces, except for thresholds 1 and 6.

Notably, threshold 1 has the highest value, which slightly declines until YIQ and then slightly increases even in HIS, HSV, HSL, thereby gradually dropping until CIELAB and then slightly increasing again to the end. While thresholds 2 and 3 have a similar track where they slightly decline even in YCbCr and then continue to

increase until YUV, the thresholds slightly increase even in HIS, HSV, HSL and then slightly drops again and then straighten their track until the CIELUV where they drop to the end. In addition, thresholds 4 and 5 have the same track, where they start to slightly drop even in YCbCr and then fluctuates until YIQ. Then, these tracks slightly increase even in HIS, HSV, and HSL, followed by a slight decline. Finally, their tracks settle down to CIEXTZ and then slightly increase and decrease again until the end. Threshold 6 has a different behavior, in which its track drops until YCgCr and gradually increases and decreases, followed by a slight increase in YIQ and then a slight dropping until it stabilizes to the end. By contrast, thresholds 7 and 9 have similar tracks to that of threshold 5. Finally, threshold 8 records the lowest threshold value, where its track slightly drops at YCbCr and then gradually increases and decreases until YIQ. Then, its track sharply increase and then drops even in CIEXYZ until its track remains stable to the end. The general characteristic of this criterion is that its behavior is similar to the true negative criterion behavior but in the opposite direction.

D) False Negative Criterion

Figure 4.6, illustrates the behavior of the false negative criterion using nine thresholds with various color spaces as in the chart below.

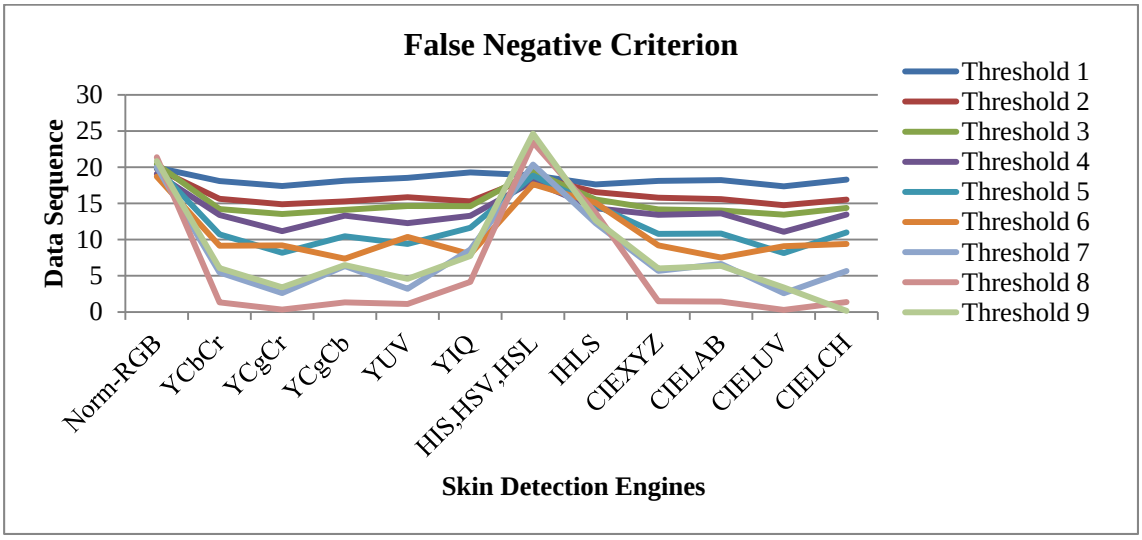


Figure 4.6. Behavior of the False Negative Criterion with Different Color Spaces

Initially, this criterion is also considered a complementary to the true positive criterion within probabilistic parameters of the reliability group. However, the lowest threshold value of the criterion is 0.16%, while the highest value is 24.63%. Generally, Figure 4.6 shows that nearly all threshold values start at one point for Norm-RGB. Thus, the behavior of this criterion is affected according to the changes in the threshold track. Notably, threshold 1 slightly declines until YCgCr and then slightly increases to stabilize even at YIQ and drops to the IHLS. The threshold then slightly rises and declines at CIELUV and then rises again to the end. Thresholds 2 and 3 nearly exhibit a similar trend as threshold 1, but they sharply rise at HIS, HSV, and HSL, followed by a gradually drop until CIELUV and slightly rises to the end afterward. Thresholds 4 and 5 start similarly as the previous thresholds, in which they sharply rise even to HIS, HSV, and HSL, followed by a gradual decline until CIELUV and increase again to the end. Threshold 6 exhibits a behavior that is different from others when it starts from the same point and gradually declines until

YCbCr. Then, this threshold moderates its track at YCgCr, slightly drops to YCgCb, and rises and declines again. Finally, it sharply rises to the top even at HIS, HSV, and HSL afterward, thereby gradually dropping to the end. Thresholds 7 and 9 exhibit similar behavior where they start to decline from the starting point until YCgCr, change its track up and down and rise to the top even at HIS, HSV, and HSL. Then, threshold 7 gradually drops to the end while threshold 9 continues to decline. Finally, threshold 8 records the lowest threshold value, where it starts at the same point and continues to decline even at YCbCr and then stabilizes its track to YUV. This threshold sharply rises to the top until HIS, HSV, and HSL, followed by a sharp decline until CIEXYZ and then gradually continues to the end. We conclude that the behavior of this criterion is similar with that of true positive criterion but in the opposite direction.

4.4.1.2 Relationship of Parameters

In this section, the behavior of the second group of sub-criteria within the reliability group, which includes accuracy, recall, precision, and specificity, is discussed and analyzed.

A) Accuracy Criterion

Accuracy typically refers to the exactness of an analytical method or the close agreement between the measured and accepted values, either as a conventional true value or an accepted reference value.

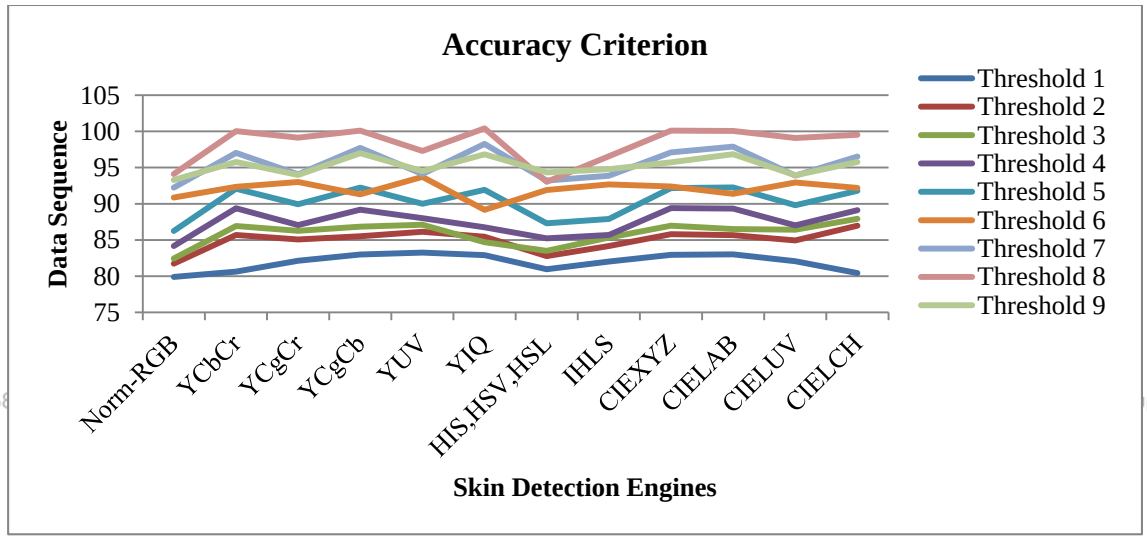


Figure 4.7. Behavior of the Accuracy Criterion with Different Color Spaces

Figure 4.7 shows that the criterion behaves similarly with that of true positive criterion due to the convergence of the values between the two criteria, as shown in the figure below. The accuracy criterion is considered an important measure of the relationship of parameters of the reliability group. However, the figure shows the lowest threshold value of the criterion at 79.90%, while the highest threshold value is 100.40%. Thus, the track of each threshold is nearly similar according to the color spaces, except for thresholds 1 and 6.

Threshold 1 records the lowest value, where it starts to slightly rise from Norm-RGB to YIQ and slightly drops at the HIS, HSV, and HSL afterward. Then, it starts to slightly rise until CIELUV and drops again in CIELCH. Thresholds 2 and 3 exhibit the same behavior, where they start to rise from Norm-RGB to YCbCr and then stabilize their track to HIS, HSV, and HSL. These thresholds begin to slightly decline, followed by a gradual rise until CIEXYZ and a gradual rise of their tracks even to CIELCH. Thresholds 4, 5, 7, and 9 start slightly high from Norm-RGB to YCbCr and then begin to drop and rise even at CIELCH. Threshold 6 demonstrates a reverse behavior, where it starts to slightly rise from Norm-RGB to YCgCr, gradually falls and rises up to YIQ, and then slightly rises and stabilizes its track until CIELUV, thereby falling until CIELCH afterward. Finally, threshold 8 starts at the highest value, followed by a slight rise from Norm-RGB to YCbCr and begins to rise and dramatically fall, thereby sharply dropping at HIS, HSV, and HSL. This threshold starts to slightly rise to CIEXYZ and then its track settles down to the end.

B) Recall Criterion

Recall is considered an important criterion that measures completeness or quantity. This criterion is the average probability of a complete retrieval referred to as the true positive rate. Figure 4.8 shows all threshold values starting at roughly the same point from Norm-RGB. Notably, this criterion behaves fairly similar to that of the true negative criterion due to the matching in the track of thresholds, as shown in the graph.

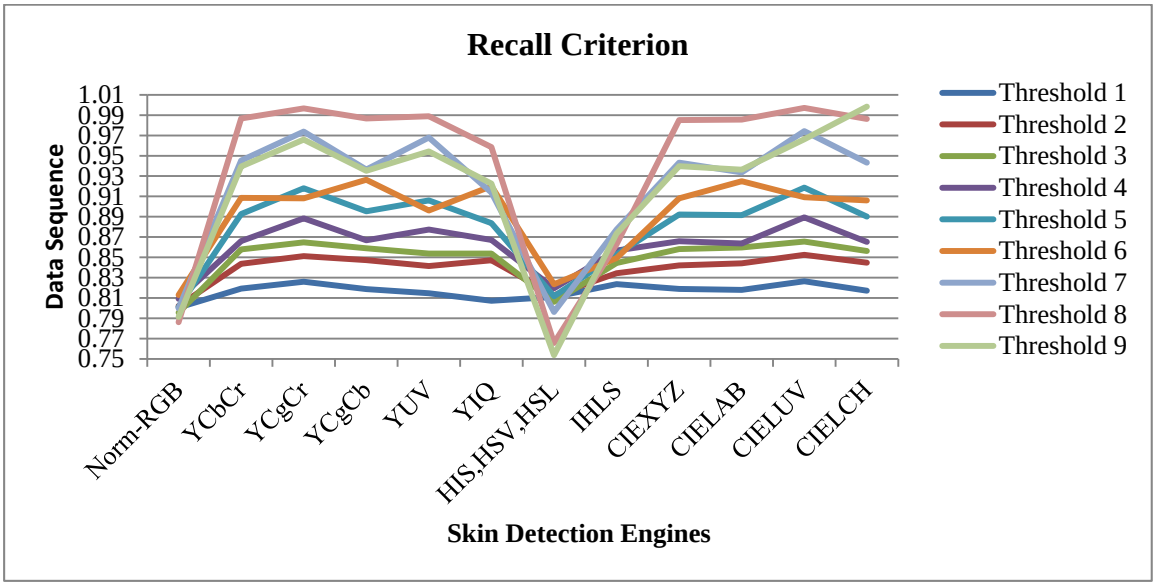


Figure 4.8. Behavior of the Recall Criterion with Different Color Spaces

Figure 4.8, shows that the lowest threshold value of the recall criterion is 0.791%, while the highest threshold value is 0.997%. This figure illustrates the behavior of the recall criterion at different threshold values with various color spaces used as an alternative in the study. The chart generally demonstrates that nearly all threshold values start at one point from Norm-RGB. Thus, the behavior of this criterion is affected according to the changes in the threshold track. Threshold 1 starts to slightly increase until YCgCr; then, its track stabilizes even in YIQ and then slightly increases until IHLS; finally, its track stabilizes to the end. Meanwhile, thresholds 2 and 3 begin to slightly increase until YCbCr and then stabilize its track until YIQ. Then, these thresholds slightly drop at HIS, HSV, HSL and then slightly increases to the end. Thresholds 4, 5, and 7 nearly have a similar track from start to end. These thresholds start to increase to YCbCr. Then, their tracks change between up and down until YIQ and then drops to a minimum value of HIS, HSV, HSL. Finally, their tracks increase

again until CIEXYZ and stabilize until CIELAB, followed by a slight increase and decrease to the end. Threshold 9 has a similar track with the previous threshold. However, its track sharply dropped at HIS, HSV, HSL and then sharply increased to the CIEXYZ. Then, the track remained stable until CIELAB and reached its maximum height to the end. Threshold 6 has a different track, where it increases even in YCbCr and then settles up until YCgCr. Then, its track changes between up and down until the YIQ and then drops to the lowest value at HIS, HSV, and HSL. The track of this threshold increases to CIELAB and then slightly decreases until the end. Finally, threshold 8 records the highest threshold value, which begins to increase until YCbCr and then stabilizes until it sharply drops to the lowest value at HIS, HSV, and HSL. Then, its value increases again until CIEXYZ and its track stabilizes until the end.

C) Precision Criterion

Precision is also considered an important measure and is defined as the number of correctly classified positive examples divided by the number of examples, which is labeled by the system as a positive value.

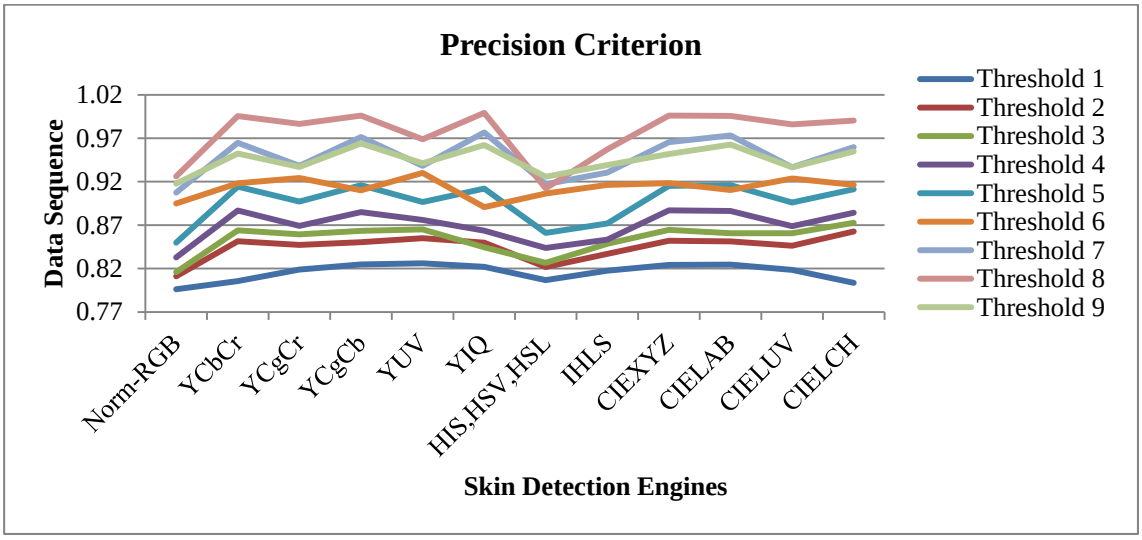


Figure 4.9. Behavior of the Precision Criterion with Different Color Spaces

Figure 4.9 depicts that precision has a similar behavior to accuracy due to the convergence of the respective values, as shown in the figure below. The precision criterion is considered an important measure of the relationship of parameters of the reliability group. However, the figure shows that the lowest threshold value of precision is 0.796%, while the highest value of the threshold is 0.996%. Thus, the track of each threshold is nearly similar according to the color spaces, except for thresholds 1 and 6

Threshold 1 records the lowest value, which slightly starts to increase from Norm-RGB to YIQ. Then, the value slightly drops at the HIS, HSV, and HSL, followed by a slight increase until CIELUV and a decrease again in CIELCH. By contrast, thresholds 2 and 3 have the same behavior, in which their values start to increase from Norm-RGB to YCbCr and then stabilize their track to HIS, HSV, and HSL. Then, their tracks begin to slightly decline, followed by a slight increase until

CIEXYZ and then a gradual increase in their even in CIELCH. Thresholds 4, 5, 7, and 9 start to slightly increase from Norm-RGB to YCbCr and then begin to drop and increase even in CIELCH. Threshold 6 has a reverse behavior, in which it starts to slightly increase from Norm-RGB to YCgCr and starts to decrease and increase to YIQ. Then, this threshold slightly increases and stabilizes its track until CIELUV and then decreases until CIELCH. Finally, threshold 8 starts at the highest value, which slightly increases from Norm-RGB to YCbCr, followed by an increase and a dramatic and sharp drop at HIS, HSV, and HSL. Then, this threshold starts to slightly increase to the CIEXYZ until its track settled down to the end.

D) Specificity Criterion

Specificity is considered an important measure of the relationship of parameters, which represents the capability of a classifier to distinguish patterns of the negative class from 0 to 1. The specificity criterion behavior is shown in Figure 4.10.

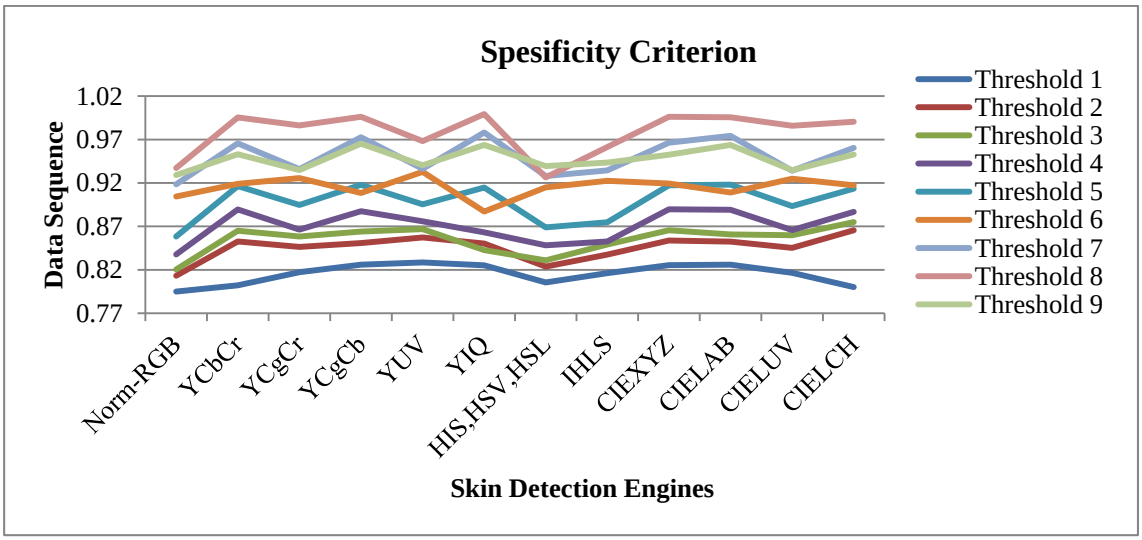


Figure 4.10. Behavior of the Specificity Criterion with Different Color Spaces

Figure 4.10, illustrates that the behavior of specificity is similar to that of precision at the threshold values according to different color spaces.

4.4.1.3 Behavior of Parameters

The behavior of the parameters is the final part of the reliability group which includes two key parameters, namely F-measure and G-measure. This group measures and tests the behavior of the parameters that are closely related to those of the precision and recall.

A) F-measure Criterion

F-measure is the most popular criterion for evaluating classification quality. This criterion is defined as the weighted harmonic mean between recall and precision. In addition, this criterion behaves fairly similar to that of the recall criterion due to the convergence of their values.

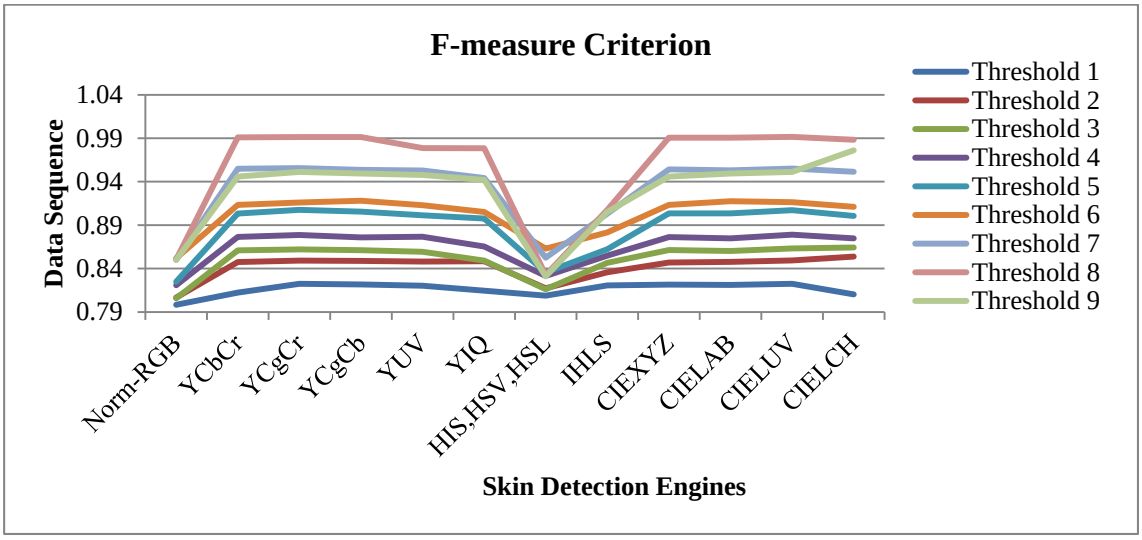


Figure 4.11. Behavior of the F-measure Criterion with Different Color Spaces

Figure 4.11, shows the behavior of F-measure criterion based on nine threshold values with different color spaces as shown below. The figure shows that the lowest threshold value of the criterion is at 0.798%, while the highest threshold value is at 0.991%. Threshold 1 starts to slightly increase at the lowest value from Norm-RGB until YCgCr and then stabilizes its track until the figure shows the lowest threshold value of the criterion at 0.796%. Meanwhile, the highest threshold value is at 0.996%, which slightly increases and stabilizes until CIELUV and then slightly drops until the end track. Thresholds 2 and 3 start at the same point and slightly increase up to YCbCr. Then, their track stabilizes until YIQ and then descends at HIS, HSV, and HSL, thereby slightly increasing up to CIEXYZ and then stabilizes their track again until the end. Thresholds 4 and 5 begin to slightly increase from YCbCr and then stabilize their path until YIQ. Then, these thresholds decline their tracks at HIS, HSV, and HSL and then increase until CIELUV, thus slightly dropping to the end. Finally, thresholds 6, 7, 8, and 9 begin to increase from the same point where threshold 6

starts to increase until YCbCr and then settles down to YIQ. Then, the thresholds sharply fall at HIS, HSV, and HSL, followed by an increase to CIEXYZ and then settle down their track until the end. Thresholds 7 and 9 have the same track as that of the previous threshold. However, a sharp decrease is observed at HIS, HSV, ad HSL, followed by a sharp increase at CIEXYZ. The track of threshold 9 stabilizes to the end while threshold 9 increases from CIELUV. Finally, threshold 8 records the highest value, which starts to increase to YCbCr and stabilize its track until YCGCb. Then, the value slightly drops at YUV and continue to sharply drop at HIS, HSV, and HSL. Finally, its value sharply increases at CIEXYZ and stabilizes its track to the end.

B) G-measure Criterion

G-measure is the last parameter of the reliability group, which refers to the geometric mean of precision and recall, thereby reflecting the general classification of algorithms in terms of the performance and accuracy of the positive sample classification.

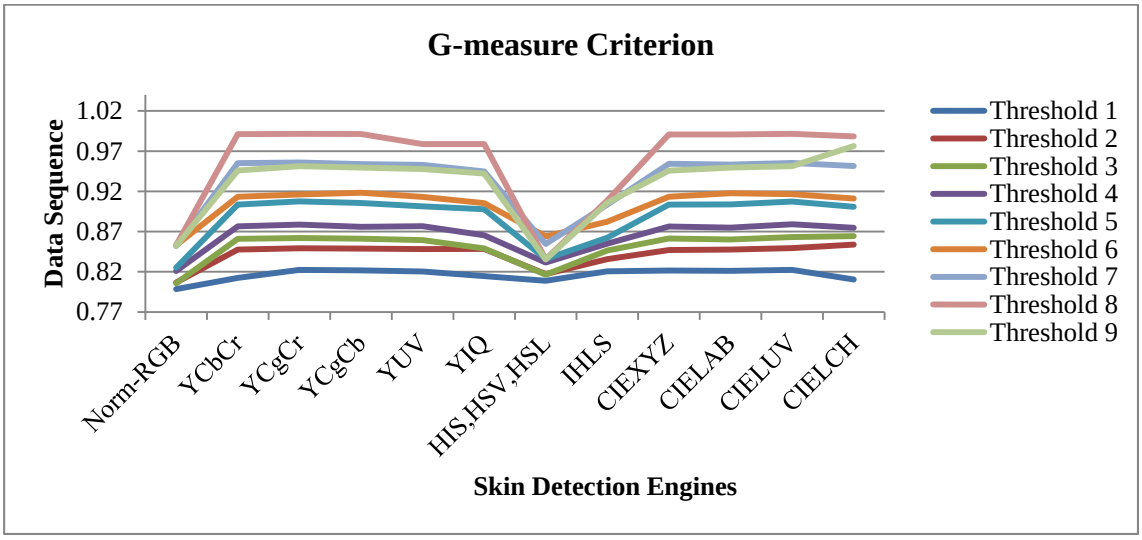


Figure 4.12. Behavior of the G-measure Criterion with Different Color Spaces

Figure 4.12, shows that G-measure exhibits a similar behavior to that of F-measure due to the symmetry in their final values, which impacts the behavior of G-measure according to the threshold distribution values at each color space used.

4.4.2 Time Complexity Criterion

Time complexity is defined as the time needed in addressing the image segmentation related to its size, thereby showing a direct correlation between them. In this section, the key criterion of time complexity is discussed according to the distribution of threshold values with different color spaces.

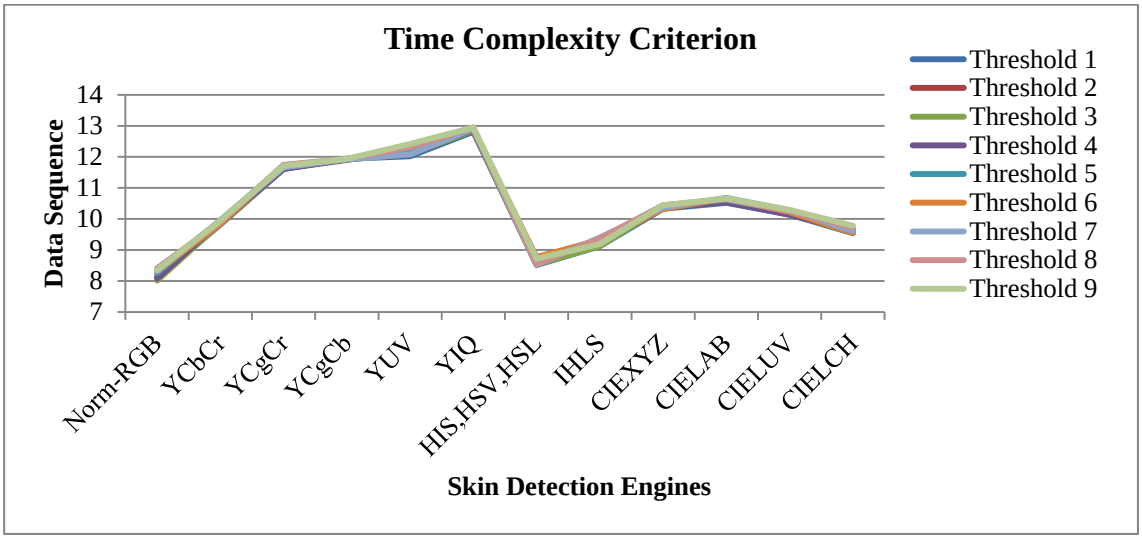


Figure 4.13. Behavior of the Time Complexity Criterion with Different Color Spaces

Figure 4.13, shows that the criterion has the lowest threshold value at 8.01% while the highest threshold value is at 12.42%, which indicates that the threshold values for this criterion are fairly identical. Consequently, according to the matching in its values which clearly influenced the behavior of this criterion. Thus, the track of threshold values appears to be identical from the starting point to the end of its track. The threshold values start at the lowest value, which shows a sharp increase from Norm-RGB until YCgCr and then continues to slightly increase until YIQ. Then, the values sharply drop until HIS, HV, and HSL. Finally, they start to slightly increase until CIEXYZ and their tracks stabilize until CIELAB, followed by a slightly decrease until the end of its track.

4.4.3 Error Rate within Dataset

The error rate is the minimum possible error calculated based on the dataset used by an irreducible classifier which implemented the training and validation process.

A) Error Rate of Validation Criterion

A cross-validation process is usually used to set error rate for the training data, which is widely used in this study. The dataset is divided into three sections, namely the majority of data for training and validation and testing.

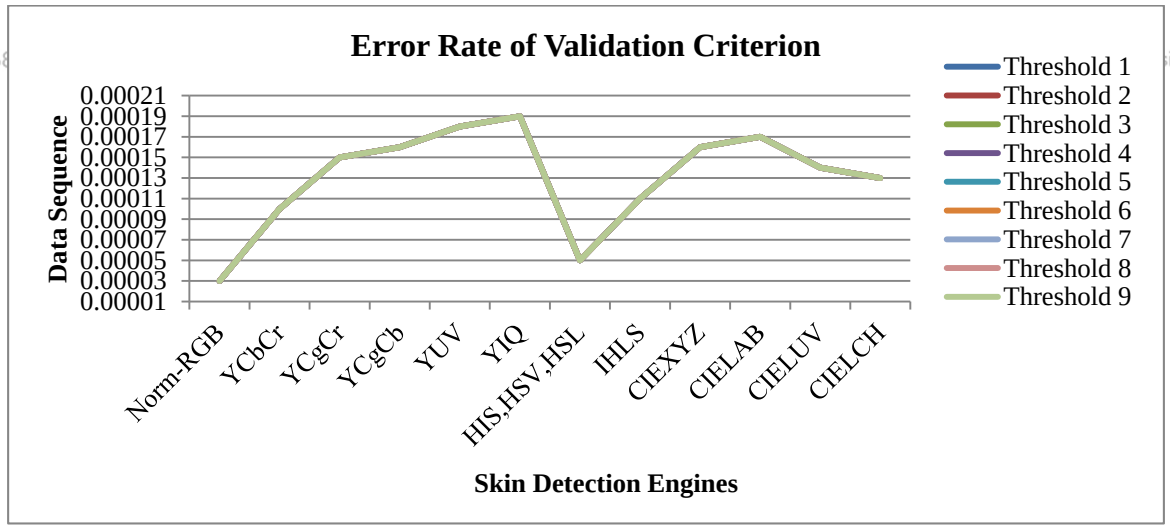


Figure 4.14. Behavior of the Error Rate of Validation Criterion with Different Color Spaces

Figure 4.14, shows the behavior of the error rate of the validation criterion based on nine thresholds according to different color spaces. In this figure, the threshold values are notably identical for each color space. Thus, the track of the threshold values starts

at the lowest value from Norm-RGB to sharply increase until YCgCr. Then, a slight increase continues to YIQ and then a sharp decline until HIS, HSV, and HSL. The sharp increase to CIEXYZ is repeated. Then a slight rise is observed until CIELAB and then slowly drops to the end of the track.

C) Error Rate of Training Criterion

The training set is also considered an important stage in calculating the error rate during the dividing process of the dataset. Figure 4.15 shows the behavior of the training criterion.

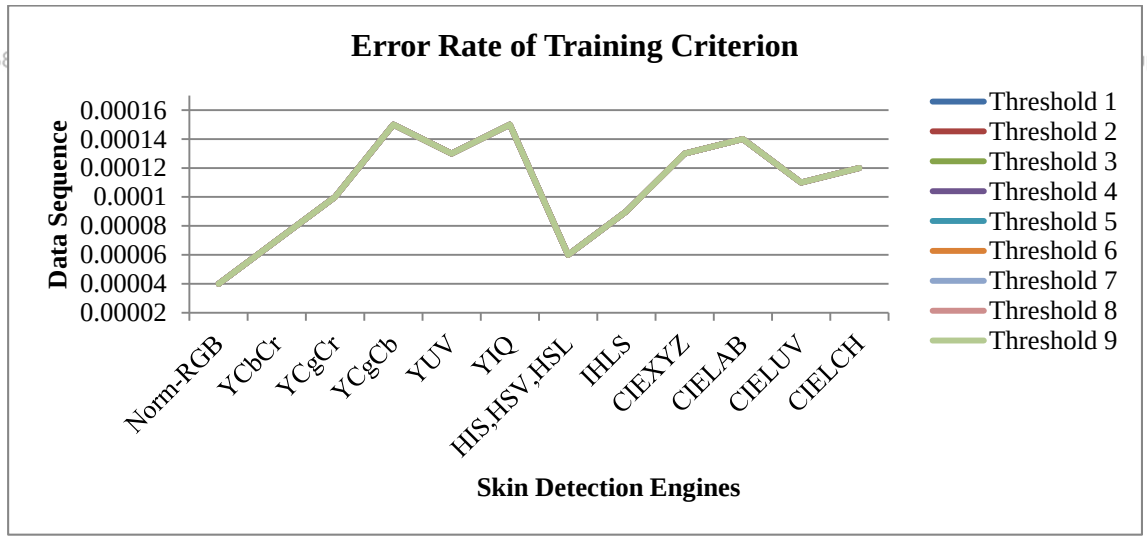


Figure 4.15. Behavior of the Error Rate of Training Criterion with Different Color Spaces

Figure 4.15, shows the behavior of the error rate of training criterion according to the distribution of threshold values at each color space. This figure exhibits similar status

as that of the validation criterion through matching threshold values according to each color space. The values of the threshold start from Norm-RGB to sharply rise until YCgCb and slightly drop until YUV. These values then slightly rise to YIQ, sharply decline to HIS, HSV, and HSL, and sharply rise again to CIEXYZ. They continue to slightly rise to the CIELAB, then slightly decline to CIELUV, and slightly rise until the end of its track.

4.4.4 Summary

In conclusion, the behavior of criteria in all scenarios is affected by the distribution of threshold values for each criterion according to the different color spaces used. Therefore, the reliability group has three sections, and each section has several criteria. The first section represented a matrix of parameters, which include confution matrix that have nearly identical behavior as the following charts. The true negative criterion has a similar behavior to that of the false positive criterion but in the opposite direction due to the values of the false positive criterion, which are considered a complementary to the true negative criterion. Meanwhile, the true positive criterion has a similar behavior to the false negative criterion due to the same reason. The second section also includes four criteria, which represented a relationship of parameters. Notably, the behavior of the accuracy criterion is fairly similar to that of the true negative criterion due to the closeness between their values. Moreover, the accuracy criterion is mainly affected by the values of the matrix of parameters according to its measurement. The recall criterion has a behavior similar

to the true positive criterion according to their threshold track as noted in the charts. By contrast, the precision and specificity criteria have nearly the same behavior due to the convergence of their values. In addition, these criteria have similar threshold track of the true negative criterion as shown in the graph. The third section includes two criteria, namely F-measure and G-measure. Both criteria have a similarity in their behavior due to the convergence of their values. By contrast, the time complexity has a specific behavior as in the diagram. Notably, the behavior of the time complexity is clearly affected by the large convergence between its values, which is distributed according to the threshold values compared to the different color spaces. Finally, the error rate within the dataset has two basic criteria, namely validation and training. Notably, the behavior of the validation criterion is clearly affected by the matching between the threshold values at each color space for this criterion. The behavior of the training criterion is also affected due to convergence between the threshold values for the two criterion. Thus, the variation in the behavior of criteria lead to the truth is necessary using all criteria in the evaluation of the decision matrix.

4.5 Chapter Summery

In this chapter, the second objective is achieved according to the methodology proposed in Chapter 3. We obtained the decision matrix, which includes the final results of the 13 criteria according to the 14 color spaces used. Thus, the following two steps were implemented: 1) the correlation between different criteria was proven

using statistical methods; and 2) performance analysis of criteria was conducted to determine the factors that influence their behavior.

The first step was achieved by using the Pearson method, which proved the existence of correlation between the various criteria. The second step was achieved by conducting a performance analysis of criteria and determining their behavior based on nine thresholds at each color space. Therefore, these two steps achieved the desired goal by evaluating the decision matrix, which will be used in the next phase to generate the final results in Chapter 5.

CHAPTER 5

RESULTS AND DISCUSSION

5.1 Introduction

In this chapter, the process of integration is implemented between the best MCDM techniques according to the proposed methodology. Analytic hierarchy process (AHP) is performed to generate different weights according to the criteria used in our study. This method is based on the preferences of six evaluators from different universities in Malaysia using pairwise question method to provide the results of the final weights. By contrast, the TOPSIS method is employed to generate the final results based on the obtained weights from the method and decision matrix acquired Chapter 4. This method is based on six basic configurations further discussed in Chapter 3 to provide The final results based on the two main contexts, namely, individual and group contexts. Thus, the selection of a suitable context is recommended based on experiment implementation and different aggregation processes conducted to achieve the selection procedure. These contexts are emphasized based on individual decision-making for decision makers and group decision-making for multiple decision makers.

The final results are obtained through the integration process for selecting the best alternatives, which is considered the main objective of this chapter. Consequently, the final results for the six evaluators are presented and discussed in detail below.

This chapter will be organized as follows. Section 5.2 represents a weight measurement process. Section 5.3 provides different weights based on multi-layers weight measurement using AHP. Section 5.4 uses the TOPSIS method to provide different aggregation techniques to test group aggregation measurements. Finally, the chapter summary is presented in Section 5.5. Figure 5.1. Overview of Results and Evaluation of the Skin Detector.

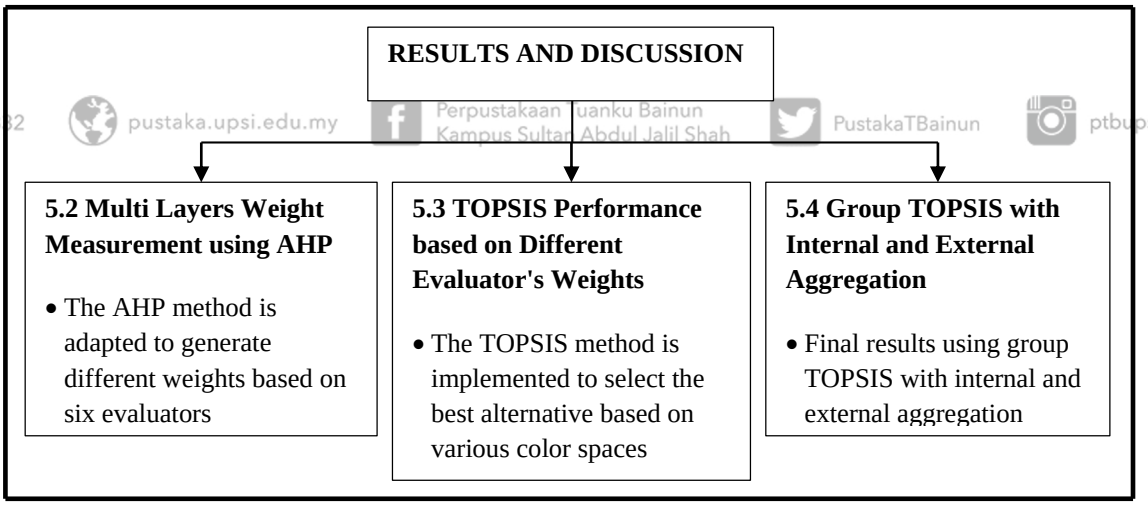


Figure 5.1. Overview of Results and Evaluation of the Skin Detector

5.2 Multi-Layer Weight Measurement using AHP

Table 5.1 shows the results based on the preferences of the six evaluators, as calculated from various Malaysian universities. The questions are presented according

to the rules of the analytic hierarchy process in Figure 3.12. The results are useful for the next stage. Meanwhile, multi-criteria decision-making techniques must weigh the essential criteria to create the decision matrix. Therefore, multi-layer analytic hierarchy process is employed to generate standard weights according to the preferences of evaluators. The first evaluator presents the percentage of the reliability group at 57.3%, whereas the time complexity group percentage is 35.3%, and the error rate is 7.4%. The second evaluator provides the percentage of the reliability group at 28.1%. About 8.1% percentage for time complexity, and the percentage of error rate group is 63.8%. The third evaluator presents the percentage of the reliability group at 33.3%, whereas the percentage of time complexity is 33.3%; by contrast, the percentage of error rate is 33.4%. The fourth evaluator presents the percentage of the reliability group at 62.2%, whereas the percentage of the time complexity group is 30.2%; by contrast, the percentage of the error rate group is 7.6%. The fifth evaluater gives the percentage of the reliability group at 23.9%, whereas the percentage of the time complexity group is 13.8%; by contrast, the percentage of the error rate group is 62.3%. The last evaluator presents the percentage of the reliability group at 21.1%, whereas the percentage of time complexity group is 68.6%; by contrast, the percentage of the error rate group is 10.2%. Finally, the results of criteria weights collected from six evaluators are important to complete the decision matrix that will be used with the decision-making method in the next phase. Table 5.1 represents the ML-AHP measurement for weights of the preferences of the six evaluators.

Table 5.1

ML-AHP measurement for Weights Preferences

	TN	TP	FP	FN	Accuracy	Recall	Precision	Specificity	F-measure	G-measure	Time Complexity	Validation	Training
UKM	0.104	0.104	0.104	0.104	0.032	0.032	0.032	0.032	0.015	0.015	0.354	0.012	0.062
IIUM	0.006	0.006	0.006	0.006	0.021	0.021	0.021	0.021	0.085	0.085	0.081	0.319	0.319
UNITEN	0.028	0.028	0.028	0.028	0.028	0.028	0.028	0.028	0.056	0.056	0.333	0.167	0.167
UPM-1	0.084	0.084	0.084	0.084	0.025	0.025	0.025	0.025	0.093	0.093	0.302	0.008	0.068
UPM-2	0.008	0.008	0.008	0.008	0.034	0.034	0.034	0.034	0.034	0.034	0.138	0.468	0.156
UPM-3	0.036	0.036	0.036	0.036	0.005	0.005	0.005	0.005	0.024	0.024	0.686	0.077	0.026

5.3 TOPSIS Performance based on Different Evaluator's Weights

In this section, TOPSIS will be used to evaluate and benchmark the skin detector approach based on 108 color space algorithms from the perspective of the six evaluators. Thus, this procedure will be applied to select the method that performs best as the appropriate approach. Table 5.1 provides the preference weights that present the features with appropriate color space algorithm from the perspective of the evaluators. In this case, six experts conducted pairwise comparisons to measure the importance of the evaluation criteria from their perspective.

TOPSIS is implemented to identify the best and worst performances of the skin detection approach for each experiment. These experiments are compared based on the ideal and worst performances. S^- represents the approach closest to the worst performance, whereas S^+ represents the approach closest to the ideal performance. Thus, according to the TOPSIS rules, the approach closest to the best performance

and farthest from the worst performance is selected as the ideal approach. Therefore, the preferences of the six evaluators will be discussed in detail as follows.

Table 5.2

First evaluator Result to Evaluate and Benchmark for Different Color Space Algorithms

S-	S+	Rank-1	S-	S+	Rank-1	S-	S+	Rank-1	S-	S+	Rank-1
0.017116	0.023158	**0.424983**	0.009156	0.022128	**0.292674**	0.015896	0.020653	**0.43492**	0.013994	0.017025	**0.451157**
0.01710	0.021884	**0.438591**	0.010468	0.021006	**0.332587**	0.015979	0.018949	**0.457475**	0.016688	0.014636	**0.532751**
0.017451	0.021761	**0.445044**	0.012167	0.019615	**0.382829**	0.016506	0.018181	**0.475855**	0.017815	0.01397	**0.560475**
0.017711	0.01998	**0.469907**	0.015418	0.017485	**0.468584**	0.01851	0.015249	**0.548308**	0.02190	0.01148	**0.656041**
0.017574	0.019217	**0.477674**	0.016609	0.016912	**0.495476**	0.019311	0.016471	**0.539681**	0.025756	0.010115	**0.718008**
0.019251	0.01618	**0.543337**	0.020919	0.014871	**0.584491**	0.018909	0.018659	**0.503319**	0.02140	0.011558	**0.649348**
0.019658	0.016476	**0.544033**	0.024939	0.014025	**0.640054**	0.019207	0.019209	**0.499965**	0.010957	0.021933	**0.333146**
0.020242	0.016829	**0.546028**	0.020382	0.015042	**0.575386**	0.013934	0.021122	**0.397477**	0.012632	0.019034	**0.398909**
0.020012	0.016694	**0.54520**	0.006275	0.024552	**0.20355**	0.014446	0.019271	**0.428442**	0.013728	0.017568	**0.438654**
0.012038	0.022626	**0.347277**	0.008922	0.022309	**0.285672**	0.015226	0.01800	**0.458223**	0.015115	0.016235	**0.482135**
0.013778	0.018305	**0.42946**	0.01020	0.021172	**0.325073**	0.015469	0.017288	**0.472242**	0.017511	0.013695	**0.56115**
0.014687	0.016925	**0.46460**	0.01190	0.020192	**0.370863**	0.015868	0.016178	**0.495166**	0.01860	0.012329	**0.601348**
0.015915	0.015169	**0.51200**	0.014591	0.018471	**0.441309**	0.017956	0.013951	**0.562754**	0.022321	0.010186	**0.686651**
0.01823	0.012524	**0.592768**	0.016206	0.017681	**0.478228**	0.019364	0.01190	**0.619292**	0.026254	0.008256	**0.760767**
0.019224	0.011506	**0.625583**	0.020438	0.015237	**0.572888**	0.020518	0.011857	**0.63375**	0.021894	0.010309	**0.67988**
0.023195	0.008182	**0.739238**	0.02330	0.01499	**0.608563**	0.020138	0.011537	**0.635763**	0.012113	0.022977	**0.345187**
0.026927	0.00640	**0.808013**	0.019744	0.016251	**0.548517**	0.010175	0.022137	**0.31490**	0.014357	0.01766	**0.448432**
0.022262	0.00880	**0.716689**	0.004807	0.026663	**0.152757**	0.011829	0.01950	**0.377605**	0.015118	0.016579	**0.476947**
0.007669	0.023758	**0.244021**	0.008494	0.023973	**0.261619**	0.012976	0.018135	**0.417086**	0.015877	0.015522	**0.505644**
0.009927	0.021132	**0.319624**	0.008573	0.024146	**0.262031**	0.014384	0.016539	**0.465159**	0.018085	0.012969	**0.582367**
0.011158	0.019982	**0.358322**	0.01036	0.022762	**0.312786**	0.016982	0.014108	**0.546218**	0.01915	0.011923	**0.61630**
0.01287	0.018686	**0.407856**	0.014269	0.020216	**0.413765**	0.01800	0.013178	**0.577385**	0.022987	0.008582	**0.728141**
0.015811	0.01652	**0.489042**	0.014818	0.020313	**0.42180**	0.022239	0.010351	**0.682394**	0.026526	0.007128	**0.78820**
0.01690	0.015562	**0.520636**	0.019962	0.018171	**0.52348**	0.025965	0.009292	**0.736451**	0.025135	0.008226	**0.753416**
0.021132	0.013556	**0.60920**	0.023494	0.017284	**0.576149**	0.02120	0.011007	**0.658184**			
0.025116	0.012542	**0.666941**	0.019389	0.01821	**0.515676**	0.00960	0.022484	**0.299276**			
0.020591	0.013824	**0.598319**	0.015769	0.022035	**0.417118**	0.011454	0.019838	**0.366046**			
0.00645	0.0246	**0.207729**	0.015712	0.020789	**0.430455**	0.01246	0.018731	**0.399483**			

According to the ML-AHP results, the weight given by the first evaluator for the reliability group represented in different parameters as (TP, FP, TN, and FN) at 41.5%, (accuracy, recall, precision, and specificity) at 12.8%, and (F-measure and G-measure) at 3%, as well as 35.3% for the time complexity group; the error rate group is represented in training at 6.2% and validation at 1.2%. Each color space algorithm is evaluated using different attributes. Accordingly, TOPSIS ranking results indicate that the first evaluator attained an average of 0.49763 ± 0.13776. The highest rank value is 0.8080, whereas the lowest value is 0.1527. Table 5.2 shows the complete sample results.

Table 5.3

Second Evaluator Result to Evaluate and Benchmark for Different Color Space Algorithms

S-	S+	**Rank-2**	S-	S+	**Rank-2**	S-	S+	**Rank-2**	S-	S+	**Rank-2**
0.046322	0.003036	**0.938494**	0.006707	0.041529	**0.139038**	0.03947	0.007516	**0.84004**	0.005623	0.041159	**0.120201**
0.04632	0.002893	**0.941223**	0.006735	0.041519	**0.139575**	0.039461	0.007442	**0.841341**	0.005724	0.041144	**0.122125**
0.046327	0.002895	**0.941184**	0.006777	0.041507	**0.140361**	0.039465	0.007413	**0.841868**	0.005775	0.04114	**0.123097**
0.046325	0.002657	**0.945756**	0.006877	0.041492	**0.142186**	0.039472	0.007288	**0.844138**	0.005991	0.041128	**0.127153**
0.046316	0.002582	**0.947204**	0.006926	0.041486	**0.143066**	0.039485	0.00732	**0.843613**	0.006256	0.041122	**0.132047**
0.046321	0.002157	**0.955508**	0.00711	0.041473	**0.146338**	0.039479	0.007412	**0.841935**	0.005968	0.041128	**0.126724**
0.046322	0.002178	**0.95509**	0.007328	0.04147	**0.150167**	0.039471	0.007428	**0.841613**	0.015611	0.030939	**0.335358**
0.046319	0.002185	**0.954958**	0.007084	0.041475	**0.145892**	0.024163	0.022414	**0.518773**	0.015627	0.030907	**0.335819**
0.046321	0.002185	**0.954949**	0.00592	0.041274	**0.125443**	0.024163	0.022388	**0.519067**	0.015641	0.030892	**0.336133**
0.029496	0.017708	**0.624859**	0.005954	0.041257	**0.126122**	0.024174	0.022368	**0.519393**	0.015662	0.030877	**0.336534**
0.029505	0.017627	**0.626008**	0.005989	0.041243	**0.126806**	0.024169	0.022359	**0.519447**	0.015693	0.030859	**0.337112**
0.029511	0.017603	**0.62637**	0.00603	0.041238	**0.12757**	0.024168	0.022349	**0.519546**	0.015715	0.03085	**0.337478**
0.029518	0.017579	**0.626756**	0.006121	0.041225	**0.12929**	0.024185	0.022323	**0.520024**	0.015791	0.030837	**0.338656**
0.029537	0.017543	**0.627382**	0.006176	0.041223	**0.13029**	0.024199	0.022303	**0.520388**	0.015894	0.030829	**0.34018**
0.029552	0.017526	**0.62772**	0.006396	0.041195	**0.134395**	0.024211	0.022298	**0.520562**	0.01578	0.030838	**0.338502**
0.029594	0.017499	**0.628418**	0.006548	0.041204	**0.137135**	0.02422	0.022295	**0.520698**	0.015725	0.031081	**0.335961**
0.02965	0.017485	**0.629038**	0.006347	0.041215	**0.133439**	0.00878	0.037766	**0.188638**	0.015748	0.031025	**0.336686**
0.029583	0.017503	**0.628276**	0.00041	0.046401	**0.008748**	0.008802	0.037742	**0.189106**	0.015757	0.031014	**0.336899**
0.016222	0.031263	**0.341627**	0.000891	0.046374	**0.018858**	0.008824	0.03773	**0.189543**	0.015765	0.031005	**0.337073**
0.016238	0.031231	**0.342076**	0.000897	0.046375	**0.018981**	0.008853	0.037718	**0.190102**	0.015799	0.030985	**0.337707**
0.01625	0.031218	**0.342327**	0.001144	0.046364	**0.02407**	0.008923	0.037701	**0.191374**	0.015828	0.030976	**0.338184**
0.016269	0.031202	**0.342717**	0.001658	0.046345	**0.034534**	0.008968	0.037692	**0.192201**	0.015907	0.030958	**0.339418**
0.016308	0.031185	**0.34337**	0.001745	0.046344	**0.036295**	0.009116	0.037678	**0.194811**	0.01599	0.030955	**0.340618**
0.016324	0.03118	**0.343634**	0.002402	0.046331	**0.049292**	0.009274	0.037676	**0.197537**	0.015948	0.030958	**0.339995**
0.016407	0.031159	**0.344934**	0.002908	0.046327	**0.059059**	0.009069	0.037683	**0.193974**			
0.016499	0.031158	**0.34621**	0.002346	0.046333	**0.048198**	0.005497	0.041204	**0.117704**			
0.016395	0.031163	**0.344729**	0.039471	0.007568	**0.839113**	0.005538	0.041181	**0.118539**			
0.006664	0.041555	**0.138195**	0.039466	0.007518	**0.839996**	0.005566	0.041171	**0.119093**			

For the same reason, the weight given by the second evaluator for the reliability group represented in (TP, FP, TN, and FN) is 2.5%, (Accuracy, recall, precision, and specificity) is 8.5%, and (F- measure and G-measure) is 17.1%, as well as 8.1% for the time complexity group, whereas the error rate group is represented in training at 31.9% and validation at 31.9%. According to that weight, the result of the second evaluator reached an average of 0.38137 ± 0.28451. The highest rank value is 0.955, whereas the lowest value is 0.117. Thus, Table 5.3, shows the complete sample results after applying the second evaluator weight.

Table 5.4

Third Evaluator Result to Evaluate and Benchmark for Different Color Space Algorithms

S-	S+	Rank-3	S-	S+	Rank-3	S-	S+	Rank-3	S-	S+	Rank-3
0.028304	0.006507	**0.813078**	0.005219	0.025144	**0.171881**	0.0246	0.007016	**0.778097**	0.008446	0.023141	**0.267388**
0.028245	0.006159	**0.820989**	0.005423	0.025055	**0.17793**	0.024306	0.00678	**0.781895**	0.008637	0.023087	**0.272251**
0.028451	0.006115	**0.823085**	0.005708	0.024953	**0.186161**	0.024377	0.006586	**0.787306**	0.008699	0.023102	**0.273552**
0.028361	0.005617	**0.834675**	0.006236	0.02488	**0.200413**	0.024386	0.006033	**0.801676**	0.009383	0.023003	**0.289712**
0.028111	0.005446	**0.83772**	0.006496	0.024822	**0.207419**	0.024801	0.006057	**0.803702**	0.010199	0.022915	**0.308003**
0.028111	0.004621	**0.858828**	0.00741	0.024719	**0.230638**	0.02473	0.006556	**0.790455**	0.009334	0.022987	**0.288791**
0.028118	0.004702	**0.856745**	0.00831	0.024707	**0.25168**	0.024495	0.006804	**0.782609**	0.011865	0.018274	**0.393676**
0.028025	0.004843	**0.852661**	0.00728	0.02474	**0.227355**	0.017179	0.013419	**0.561448**	0.011953	0.018035	**0.398584**
0.028095	0.004772	**0.8548**	0.004422	0.025342	**0.148573**	0.0171	0.013242	**0.563579**	0.012066	0.017908	**0.402551**
0.018127	0.012275	**0.596238**	0.004497	0.025409	**0.150379**	0.017264	0.013056	**0.569393**	0.012232	0.017787	**0.407481**
0.018193	0.011721	**0.608174**	0.004846	0.025194	**0.16131**	0.017076	0.013062	**0.566609**	0.012254	0.017743	**0.408508**
0.01822	0.011583	**0.611344**	0.004954	0.025324	**0.16362**	0.016959	0.013015	**0.56579**	0.012435	0.017633	**0.413567**
0.018229	0.011459	**0.614026**	0.005464	0.025248	**0.177912**	0.017102	0.012817	**0.571606**	0.012798	0.017588	**0.421174**
0.018342	0.011261	**0.619605**	0.005715	0.025347	**0.183993**	0.017081	0.012725	**0.573074**	0.013423	0.017477	**0.434404**
0.01856	0.011031	**0.627212**	0.006937	0.024839	**0.218305**	0.017237	0.01269	**0.575956**	0.012723	0.017605	**0.419508**
0.018763	0.010957	**0.631335**	0.00742	0.025152	**0.227804**	0.017523	0.012517	**0.583325**	0.013226	0.017923	**0.424595**
0.01923	0.010774	**0.640908**	0.006477	0.025373	**0.203358**	0.009257	0.021659	**0.299425**	0.013334	0.017465	**0.432927**
0.018701	0.010965	**0.630396**	0.00137	0.028812	**0.045399**	0.00929	0.021504	**0.301687**	0.013377	0.017389	**0.43481**
0.00952	0.020394	**0.318237**	0.002365	0.028652	**0.076255**	0.009385	0.021419	**0.304672**	0.013361	0.017342	**0.435172**
0.009656	0.020195	**0.323474**	0.002378	0.028691	**0.076525**	0.009498	0.021336	**0.308049**	0.013517	0.0172	**0.440052**
0.009748	0.020118	**0.326404**	0.002875	0.028616	**0.091302**	0.009784	0.021216	**0.315615**	0.013837	0.017062	**0.447815**
0.009933	0.019988	**0.331974**	0.003973	0.028465	**0.122466**	0.010164	0.021065	**0.325455**	0.014168	0.01695	**0.455303**
0.010164	0.019976	**0.337216**	0.004119	0.028499	**0.126284**	0.010699	0.020988	**0.337661**	0.014417	0.016996	**0.458954**
0.010281	0.019948	**0.340114**	0.005566	0.028442	**0.163675**	0.011207	0.021017	**0.347787**	0.014122	0.017077	**0.452649**
0.010943	0.019722	**0.356856**	0.006562	0.028443	**0.187464**	0.010388	0.021081	**0.330107**			
0.011514	0.019796	**0.367742**	0.005406	0.028489	**0.159485**	0.008008	0.023488	**0.254254**			
0.010813	0.019815	**0.353047**	0.024624	0.007324	**0.770762**	0.008118	0.023326	**0.258177**			
0.004862	0.025351	**0.160935**	0.024486	0.007092	**0.775421**	0.008183	0.023272	**0.260149**			

Similarly, the weight provided by the third evaluator for the reliability group represented in (TP, FP, TN, and FN) is 11.1%, (Accuracy, recall, precision, and specificity) is 11.1%, (F- and G-measures) is 11/1%, and time complexity group at 33.3%, as well as the error rate group represented in training at 16.7% and validation at 16.7%. According to that weight, the result of the third evaluator achieved an average of 0.42512 ± 0.22686. The highest rank value is 0.8588, whereas the lowest value is 0.0453. Table 5.4, show the complete sample results.

Table 5.5

Fourth Evaluator Result to Evaluate and Benchmark for Different Color Space Algorithms

S-	S+	**Rank-4**	S-	S+	**Rank-4**	S-	S+	**Rank-4**	S-	S+	**Rank-4**
0.015111	0.018951	**0.443638**	0.007501	0.018849	**0.284675**	0.013887	0.016931	**0.450615**	0.011582	0.01459	**0.442541**
0.015088	0.017913	**0.45719**	0.00857	0.017968	**0.322941**	0.013908	0.015543	**0.47224**	0.013751	0.012744	**0.519004**
0.015383	0.017816	**0.463347**	0.009955	0.016882	**0.370936**	0.014315	0.01492	**0.489662**	0.014664	0.012234	**0.54518**
0.015563	0.016358	**0.48754**	0.012607	0.015239	**0.452735**	0.015845	0.012523	**0.558556**	0.017979	0.010362	**0.634391**
0.015422	0.01574	**0.494891**	0.013585	0.014791	**0.478745**	0.016501	0.01351	**0.549828**	0.021133	0.009358	**0.69308**
0.016683	0.013252	**0.557317**	0.017104	0.013246	**0.563567**	0.016176	0.015301	**0.513903**	0.017581	0.010417	**0.627937**
0.016989	0.013492	**0.557365**	0.020397	0.01262	**0.617783**	0.01637	0.015748	**0.509683**	0.009343	0.018221	**0.338946**
0.017415	0.013775	**0.558346**	0.016666	0.013376	**0.554754**	0.012002	0.01741	**0.408063**	0.010654	0.015891	**0.401356**
0.017252	0.013667	**0.557967**	0.005215	0.02063	**0.201777**	0.012386	0.015913	**0.437684**	0.011525	0.014716	**0.439191**
0.010599	0.018599	**0.362996**	0.007331	0.018866	**0.279849**	0.013004	0.01488	**0.466357**	0.012633	0.013645	**0.480741**
0.011922	0.01509	**0.441361**	0.008374	0.017952	**0.318071**	0.013174	0.014304	**0.479434**	0.014541	0.011646	**0.555276**
0.012622	0.01397	**0.474639**	0.009754	0.017197	**0.361924**	0.013471	0.013417	**0.501006**	0.015418	0.01058	**0.593044**
0.01357	0.012554	**0.519457**	0.011946	0.01585	**0.429776**	0.01511	0.011624	**0.565188**	0.018425	0.008924	**0.673713**
0.015388	0.01042	**0.596245**	0.013256	0.015256	**0.464927**	0.016221	0.009986	**0.618953**	0.02162	0.007478	**0.743004**
0.016189	0.009587	**0.628063**	0.016729	0.013317	**0.556794**	0.017137	0.009939	**0.632929**	0.018078	0.009021	**0.667119**
0.019347	0.006955	**0.735581**	0.019068	0.013177	**0.591348**	0.016865	0.009675	**0.635454**	0.010315	0.019087	**0.350825**
0.022355	0.005556	**0.800942**	0.016153	0.014161	**0.532846**	0.008592	0.018534	**0.31673**	0.012075	0.01483	**0.448789**
0.018602	0.007439	**0.714324**	0.003902	0.022546	**0.147521**	0.009892	0.016433	**0.375768**	0.012677	0.013969	**0.475746**
0.006627	0.019775	**0.250993**	0.006915	0.02042	**0.252985**	0.010807	0.015351	**0.413134**	0.013274	0.013133	**0.502677**
0.008386	0.017668	**0.321872**	0.00698	0.020556	**0.253498**	0.011933	0.014096	**0.458438**	0.015039	0.011126	**0.57478**
0.009361	0.016749	**0.358514**	0.008443	0.019472	**0.302456**	0.014027	0.0122	**0.534834**	0.015911	0.010301	**0.607016**
0.01073	0.015708	**0.405842**	0.01164	0.017493	**0.399538**	0.014868	0.011466	**0.564588**	0.019	0.007769	**0.709775**
0.01309	0.014004	**0.483137**	0.012095	0.017563	**0.407812**	0.018298	0.009331	**0.662265**	0.021859	0.006734	**0.764473**
0.013968	0.013259	**0.513015**	0.016296	0.015927	**0.505731**	0.021328	0.008568	**0.713395**	0.020726	0.007523	**0.733699**
0.017405	0.011672	**0.598589**	0.019195	0.015259	**0.557114**	0.017445	0.009832	**0.639546**			
0.020642	0.010912	**0.654178**	0.015832	0.015965	**0.497913**	0.008075	0.018903	**0.299324**			
0.016963	0.01189	**0.587903**	0.013795	0.018052	**0.433171**	0.009542	0.016803	**0.362195**			
0.005306	0.020805	**0.203192**	0.013733	0.01704	**0.446267**	0.010345	0.015928	**0.393754**			

Moreover, AHP results show the weight given by the fourth evaluator for reliability group represented in (TP,FP,TN, and FN) at 33.7%, (accuracy, recall, precision, and specificity) at 9.9%, and (F- measure and G-measure) at 18.6%, as well as time complexity group at 30.2%, error rate group represented in training at 6.8%, and validation at 0.8%. According to these weights, the result of the fourth evaluator yielded an average of 0.49409 ± 0.13362. The highest rank value is 0.8009, whereas the lowest value is 0.1475. Table 5.5, shows the complete sample results.

Table 5.6

Fifth Evaluator Result to Evaluate and benchmark for Different Color Space Algorithms

S-	S+	**Rank-5**	S-	S+	**Rank-5**	S-	S+	**Rank-5**	S-	S+	**Rank-5**
0.054019	0.0025915	**0.9542217**	0.0098101	0.0447446	**0.1798211**	0.0470765	0.007376	**0.8645417**	0.0073579	0.0472347	**0.1347784**
0.0540137	0.0024486	**0.9566332**	0.0098271	0.0447366	**0.1801032**	0.04705	0.0073424	**0.8650113**	0.0073887	0.047231	**0.1352752**
0.0540322	0.0024267	**0.9570176**	0.0098528	0.0447272	**0.1805207**	0.0470563	0.0073104	**0.8655346**	0.0073927	0.0472329	**0.135334**
0.0540238	0.002226	**0.9604259**	0.0099019	0.0447212	**0.1812766**	0.0470563	0.0072306	**0.8668082**	0.0075297	0.0472249	**0.1375175**
0.0540015	0.0021466	**0.9617695**	0.0099243	0.0447163	**0.1816288**	0.0470934	0.0072243	**0.8669996**	0.0076909	0.0472182	**0.1400657**
0.0540009	0.0017957	**0.9678173**	0.0100296	0.0447069	**0.1832336**	0.047087	0.0072866	**0.8659899**	0.0075189	0.0472237	**0.1373498**
0.0540015	0.0018107	**0.967558**	0.0101359	0.0447064	**0.1848185**	0.0470658	0.0073187	**0.8654273**	0.0173629	0.036873	**0.3201372**
0.053993	0.0018468	**0.9669265**	0.0100137	0.0447089	**0.1829901**	0.0274547	0.0267998	**0.5060358**	0.0173712	0.0368544	**0.3203501**
0.0539994	0.0018264	**0.9672842**	0.0043737	0.0501777	**0.0801758**	0.0274457	0.0267858	**0.5060842**	0.0173833	0.0368445	**0.3205608**
0.0311852	0.0231853	**0.5735684**	0.0043788	0.0501845	**0.0802523**	0.0274627	0.026771	**0.5063769**	0.0174008	0.0368353	**0.3208343**
0.0311908	0.0231393	**0.5740978**	0.0044378	0.0501665	**0.0812721**	0.0274421	0.0267722	**0.5061787**	0.0173998	0.0368326	**0.3208375**
0.0311927	0.0231285	**0.5742275**	0.00445	0.0501783	**0.0814597**	0.0274297	0.0267684	**0.5061012**	0.0174225	0.0368236	**0.321175**
0.0311931	0.0231186	**0.5743342**	0.0045385	0.0501725	**0.0829539**	0.0274443	0.0267518	**0.5063891**	0.0174581	0.036821	**0.3216355**
0.0312028	0.0231036	**0.574569**	0.0045929	0.0501808	**0.0838522**	0.0274407	0.0267459	**0.5064108**	0.0175339	0.0368124	**0.3226323**
0.0312239	0.0230855	**0.5749254**	0.0048378	0.0501386	**0.0879979**	0.0274576	0.0267419	**0.5066025**	0.0174501	0.0368224	**0.3215271**
0.0312417	0.0230809	**0.5751138**	0.0049455	0.0501654	**0.0897373**	0.0274883	0.0267289	**0.5070031**	0.0202366	0.0341263	**0.3722502**
0.0312864	0.0230672	**0.5756092**	0.0047357	0.0501841	**0.0862291**	0.0105808	0.0437882	**0.1946113**	0.0202466	0.034088	**0.3726287**
0.0312359	0.0230814	**0.5750634**	0.0005089	0.0540611	**0.0093255**	0.0105828	0.0437763	**0.1946838**	0.0202507	0.0340819	**0.3727178**
0.0146269	0.0398701	**0.2683979**	0.0008934	0.054048	**0.0162615**	0.0105949	0.0437697	**0.1948857**	0.0202482	0.0340783	**0.3727127**
0.0146397	0.0398541	**0.2686494**	0.0008902	0.0540516	**0.016203**	0.0106105	0.0437632	**0.1951401**	0.0202635	0.034067	**0.3729672**
0.0146487	0.039848	**0.2688**	0.0010924	0.0540453	**0.0198125**	0.0106501	0.0437541	**0.1957589**	0.0202988	0.0340557	**0.3734524**
0.0146671	0.0398376	**0.2690977**	0.0015483	0.0540323	**0.0278563**	0.0107078	0.0437421	**0.1966542**	0.0203334	0.0340472	**0.3739092**
0.0146898	0.0398376	**0.2694019**	0.0015725	0.0540359	**0.028278**	0.0107886	0.0437365	**0.1978652**	0.0203575	0.0340518	**0.3741548**
0.0147047	0.0398351	**0.2696143**	0.0022029	0.0540308	**0.0391743**	0.0108655	0.0437393	**0.198985**	0.0203204	0.0340585	**0.3736814**
0.014775	0.0398169	**0.2706449**	0.0025906	0.0540316	**0.0457526**	0.0107376	0.0437441	**0.1970862**			
0.014844	0.0398235	**0.271532**	0.0021269	0.0540352	**0.03787**	0.0072835	0.047262	**0.1335302**			
0.0147603	0.0398247	**0.2704089**	0.0470786	0.0074261	**0.8637537**	0.0072996	0.0472494	**0.1338167**			
0.0097826	0.0447633	**0.1793461**	0.0470662	0.0073895	**0.8643028**	0.0073088	0.0472454	**0.1339735**			

Accordingly, AHP results show the weights provided by the fifth evaluator at 3.3% for (TP,FP,TN, and FN), 13.7% for (accuracy, recall, precision, and specificity), 6.8% for (F-measure and G-measure) weights representing the reliability group, and 13.8% for the time complexity group, as well as 15.6% for training and 46.8% for validation parameters representing the error rate group. According to these weights, the result of the fifth evaluator attained an average of 0.37469 ± 0.28847. The highest rank value is 0.9678, whereas the lowest value is 0.0093. Table 5.6, shows the complete sample results after applying the fifth evaluator weight.

Table 5.7

Sixth evaluator Result to Evaluate and Benchmark for Different Color Space Algorithms

S-	S+	**Rank-6**	S-	S+	**Rank-6**	S-	S+	**Rank-6**	S-	S+	**Rank-6**
0.031632	0.008088	**0.796363**	0.007272	0.026229	**0.217082**	0.028762	0.007952	**0.7834**	0.015655	0.018275	**0.461397**
0.031398	0.007668	**0.803709**	0.007536	0.02606	**0.224322**	0.027605	0.008028	**0.774706**	0.015516	0.018449	**0.456813**
0.032177	0.007584	**0.809254**	0.00789	0.025871	**0.233699**	0.027816	0.007707	**0.783048**	0.015386	0.018656	**0.451973**
0.031787	0.006977	**0.820016**	0.008367	0.025914	**0.244064**	0.027565	0.007165	**0.793706**	0.015958	0.018503	**0.463066**
0.030792	0.00686	**0.817802**	0.008729	0.025753	**0.253146**	0.029071	0.006634	**0.814205**	0.016851	0.018187	**0.480931**
0.030567	0.005936	**0.837394**	0.009788	0.0256	**0.276591**	0.028865	0.007386	**0.796258**	0.015985	0.0184	**0.464887**
0.030541	0.006053	**0.834587**	0.010801	0.025658	**0.296246**	0.027909	0.00807	**0.775695**	0.017921	0.016275	**0.524064**
0.030071	0.006363	**0.82536**	0.009609	0.02567	**0.272383**	0.024456	0.010913	**0.691458**	0.017936	0.015928	**0.529646**
0.030397	0.006185	**0.830923**	0.006155	0.027328	**0.183836**	0.024137	0.010733	**0.692188**	0.018092	0.01567	**0.535864**
0.020311	0.014148	**0.589421**	0.005645	0.028051	**0.167534**	0.024549	0.010211	**0.706237**	0.018342	0.01539	**0.543763**
0.020328	0.013471	**0.60144**	0.006401	0.027377	**0.189493**	0.023918	0.010532	**0.694284**	0.017911	0.01573	**0.532408**
0.02029	0.013355	**0.603059**	0.006096	0.028083	**0.178346**	0.023479	0.01068	**0.687337**	0.018214	0.01543	**0.541366**
0.020115	0.013365	**0.600802**	0.006687	0.028053	**0.192497**	0.023602	0.01034	**0.695354**	0.018384	0.015569	**0.541442**
0.020125	0.01325	**0.603009**	0.006768	0.028567	**0.191548**	0.023273	0.010425	**0.690643**	0.019164	0.015263	**0.556647**
0.020801	0.012579	**0.623149**	0.008928	0.026836	**0.249637**	0.023565	0.010267	**0.696533**	0.018266	0.015635	**0.538809**
0.020681	0.012888	**0.616076**	0.009043	0.028147	**0.243145**	0.024568	0.009336	**0.724621**	0.021653	0.013477	**0.616369**
0.021545	0.012418	**0.634359**	0.007634	0.028868	**0.209128**	0.016595	0.01766	**0.484456**	0.021641	0.012607	**0.631895**
0.020667	0.012823	**0.617105**	0.00188	0.032141	**0.055255**	0.016431	0.017533	**0.483765**	0.021645	0.012475	**0.634377**
0.008945	0.024508	**0.26739**	0.003046	0.031997	**0.086925**	0.016476	0.017407	**0.486272**	0.021475	0.012503	**0.632028**
0.0091	0.024326	**0.272238**	0.003046	0.032126	**0.086599**	0.016501	0.017318	**0.487921**	0.021512	0.012296	**0.636295**
0.009217	0.024265	**0.275291**	0.003648	0.032061	**0.102158**	0.016683	0.017161	**0.492934**	0.022203	0.01166	**0.655663**
0.009589	0.024025	**0.285276**	0.004998	0.031853	**0.135615**	0.01744	0.016483	**0.5141**	0.022299	0.01161	**0.657615**
0.009666	0.024349	**0.284157**	0.005178	0.031976	**0.139367**	0.017747	0.016507	**0.518102**	0.022106	0.01213	**0.645706**
0.009792	0.024388	**0.28647**	0.006952	0.032051	**0.178243**	0.017979	0.016812	**0.516767**	0.02163	0.012481	**0.634107**
0.011049	0.023836	**0.316738**	0.008182	0.032166	**0.202783**	0.017203	0.016938	**0.503884**			
0.011699	0.024235	**0.32557**	0.006751	0.032228	**0.173196**	0.015369	0.018877	**0.448779**			
0.010703	0.024145	**0.307133**	0.02887	0.008343	**0.775806**	0.015344	0.018668	**0.451135**			
0.006799	0.026607	**0.203525**	0.028342	0.008177	**0.776096**	0.015321	0.018632	**0.451233**			

Finally, AHP results show that the weight given by the sixth evaluator for the reliability group represented in (TP, FP, TN, and FN) is 14.5%, (accuracy, recall, precision, and specificity) is 1.9%, (F- measure and G-measure) is 4.8%, and time complexity group is 68.6%, whereas the error rate group is represented in training at 2.6% and validation at 7.7%. According to these weights, the result of the color space algorithms achieved an average of 0.49286 ± 0.22406. The highest rank value is 0.8373, whereas the lowest value is 0.0552. Table 5.7 shows the complete sample results after applying the weight by the sixth evaluator.

Further details are discussed in each chart representing the preferences of evaluators ranking depending on internal and external aggregation values.

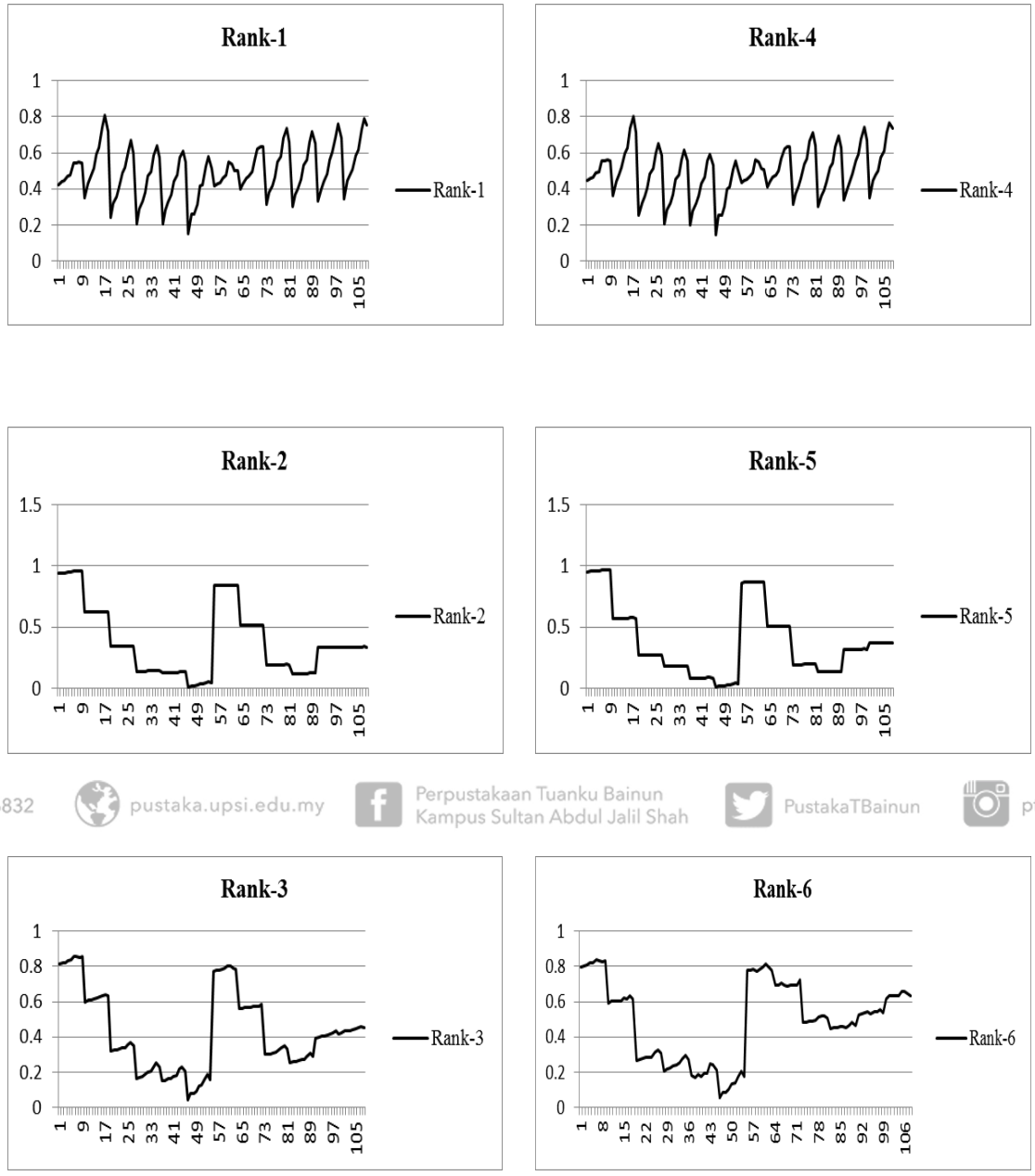


Figure 5.2. Virtualize Ranking for Six Evaluators

According to the tables obtained by applying the TOPSIS method for selecting ideal alternatives based on the 13 criteria and 108 skin detector engines, the final results

varied for internal aggregation. Therefore, these results are be evaluated by each evaluator based on the similarity between their values as shown in Figure 5.2. We observed that the results for the first evaluator are similar to that of the fourth evaluator, given that their final values are shown similar choice in color space as (YCbCr). The results for the second and fifth evaluators yielded similar results by selecting the Normalized-RGB color space. Furthermore, the results of the third and sixth evaluators showed similar results through selecting the Normalized-RGB color space. Therefore, most of the evaluators' results confirmed the selection of the Normalized-RGB color space. Further details in the next section show the final results for the internal and external aggregations within the group decision.

In summary, according to the internal results show the ranks based on individual context. The results presented that four evaluators selected the Norm-RGB as the best color space while others chose the YCbCr color space. Thus, the comparison between the findings of evaluators indicates their lack of consensus to jointly decide that referred to the difficulty and complexity in individual decision-making. Therefore, the drawback of individual decision-maker context compared with compares with group decision maker will discuss according to external results.

5.4 Group TOPSIS with Internal and External Aggregation

Many decision-making problems are resolved through collaborative efforts within organizations. However, according to the two methods mentioned in the literature, the

TOPSIS method is expanded to the group decision environment, either by internal or external aggregations. Internal aggregations aim to apply the aggregation process in the separation stage. In this case, group separation is conducted through aggregating different decision values based on the distance positive and negative ideals followed by the next process. Thus, the internal aggregation calculation based on the summation values of the negative separation is divided by the summation of negative separation values plus the summation of positive separation values for each evaluator as in (Internal aggregation = $\boldsymbol{S}^{-}/(\boldsymbol{S}^{-} + \boldsymbol{S}^{+})$), whereas external aggregation is determined by calculating the averages of all ranked values for all evaluators. Table 5.8, shows the final results of group TOPSIS with the applied internal and external aggregations. Thus, the external aggregation results showed that high values are obtained in the Normalized-RGB color space. Therefore, the final results show somewhat identical internal and external aggregation values.

Table 5.8

Group Decision-maker of TOPSIS method with Internal and External Aggregations

Internal agg	**External agg**	**Internal agg**	**External agg**	**Internal agg**	**External agg**	**Internal agg**	**External agg**
0.75540413	0.728462921	**0.20360009**	0.214195117	**0.715588808**	0.691935664	**0.279624839**	0.312910362
0.765196872	0.736389302	**0.21591285**	0.229576432	**0.724241963**	0.698777958	**0.300911904**	0.339703176
0.767852619	0.739821909	**0.23173244**	0.249084501	**0.731854128**	0.707212241	**0.308457646**	0.348268337
0.782640989	0.753053356	**0.25926195**	0.281543083	**0.756977802**	0.735531853	**0.341683829**	0.384646775
0.78710423	0.756176802	**0.26985588**	0.293246805	**0.754938985**	0.73633807	**0.371135821**	0.412022354
0.816052934	0.786700332	**0.30534859**	0.330809909	**0.736803272**	0.718643343	**0.338953272**	0.382506261
0.813965619	0.785896345	**0.33419751**	0.356791553	**0.729903493**	0.712498735	**0.368212545**	0.37422124
0.810499041	0.784046586	**0.30093977**	0.326460009	**0.515374436**	0.513875895	**0.386732112**	0.397444093
0.812223751	0.785187243	**0.14599179**	0.157226071	**0.524876249**	0.524507387	**0.398296954**	0.41215902
0.528687669	0.515726373	**0.16484707**	0.181634526	**0.53610646**	0.537663155	**0.411358516**	0.428581196
0.556994391	0.546756829	**0.18018585**	0.200337559	**0.537532174**	0.539699114	**0.429681672**	0.452548606

(*Continue*)

Table 2.1 (*continued*)

Internal agg	**External agg**	**Internal agg**	**External agg**	**Internal agg**	**External agg**	**Internal agg**	**External agg**
0.567140466	0.559039624	**0.19160121**	0.213963775	**0.542377533**	0.545824282	**0.441647021**	0.467996274
0.579574405	0.574562659	**0.21609028**	0.242289732	**0.561806889**	0.570219026	**0.467240628**	0.49721201
0.601220114	0.602263027	**0.22822913**	0.255473124	**0.575542102**	0.588126927	**0.495159404**	0.526272567
0.613722011	0.617775486	**0.27251061**	0.303336236	**0.581126593**	0.594388763	**0.464266919**	0.494224097
0.642234098	0.654293665	**0.28922547**	0.316288837	**0.586841973**	0.60114385	**0.402121301**	0.407531084
0.666071374	0.681478298	**0.25760519**	0.285586241	**0.283693552**	0.299792879	**0.4327481**	0.445226292
0.636330937	0.646975467	**0.05761645**	0.069834309	**0.29925125**	0.320435618	**0.440526947**	0.455249427
0.285016588	0.281777803	**0.09911678**	0.118817042	**0.309874324**	0.334265639	**0.447260892**	0.4642178
0.305581424	0.307989038	**0.09953502**	0.118972956	**0.322532775**	0.350801277	**0.467630772**	0.490694814
0.316238506	0.321609669	**0.11937824**	0.142097419	**0.345216693**	0.379455744	**0.480401648**	0.506404736
0.331350588	0.340460308	**0.16104049**	0.188962284	**0.358171157**	0.395063854	**0.510635945**	0.544026889
0.353404239	0.367720779	**0.16590355**	0.193305803	**0.390751637**	0.43218305	**0.52892327**	0.562017886
0.362473843	0.378914072	**0.21495894**	0.243265948	**0.413390184**	0.451820341	**0.516560441**	0.547924446
0.396209899	0.416160982	**0.24540104**	0.271386883	**0.380152917**	0.420463658		
0.420109276	0.438695526	**0.20982833**	0.238723023	**0.238152693**	0.258811181		
0.390775756	0.41025661	**0.70565322**	0.683287128	**0.255374026**	0.281651402		
0.178322097	0.182153553	**0.71283127**	0.68875634	**0.264022127**	0.292947512		

Results proved that for internal aggregation, the ideal value is 0.8160529, whereas 0.7867003 is the ideal value for external aggregations. According to the external aggregation, the result of the color space algorithms reached an average of 0.4442933 ± 0.4827987, whereas the internal aggregation for each color space algorithm achieved an average of 0.432439 ± 0.4776001. Aggregation results indicated similarities between the internal and external rankings of various color space algorithms based on the perspective of the six evaluators. Figure 5.3, shows the tested color space algorithms performance post-aggregation based on the results of the six external evaluators.

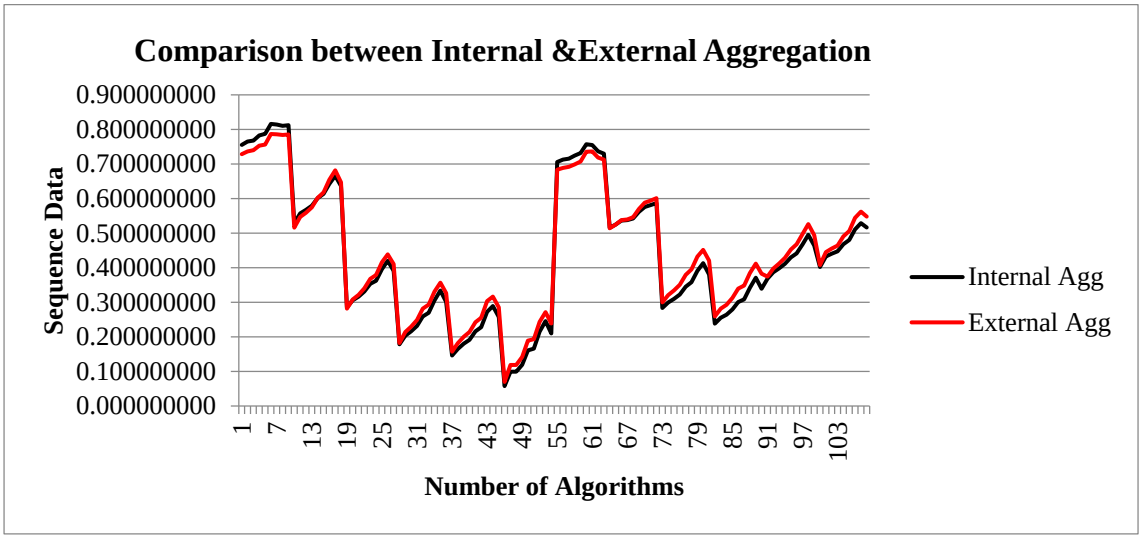


Figure 5.3. Internal and External Aggregation Ranking

5.5 Chapter Summary

In this chapter, we investigate a case study that discusses the reflection of different external evaluators' preferences in skin detector evaluation and benchmarking. Thus, MCDM techniques indicated ease of use, and integrated procedure maintains this strength, which is obtained from multiple sources of knowledge and experience. Testing results are achieved according to six evaluators involved in the evaluation process based on several queries:

- How can evaluators present their perspectives about a case study?
- How are these preferences interpreted for different criteria weights?
- How are these weights reflected in individual and group evaluation?
- How can individual evaluations be aggregated into group analysis?
- What are the possible aggregation functions?
- What are the differences between these aggregation functions?

According to the case study, the final results represent a group of evaluators who provided their opinions on the criteria preference. Thus, we benchmarked 108 color space algorithms in the group context by aggregating these inputs. Internal and external aggregations indicated identical performances. Finally, a critical term must prove the validity of these results in Chapter 6.

CHAPTER 6

VALIDATION AND COMPARISON

6.1 Introduction

This chapter provides an insight into the validation and comparison of the criteria and different color spaces obtained for this research. The validation process is an important measure for various empirical researchers in validating the accuracy of its results. Therefore, there is a need to validate the results obtained according to the MCDM techniques. For this procedure, the initial step is the implementation of the validation process for using the multi-criteria measurement process for three key groups of criteria. This is followed a comparison of the different color spaces based on the final results. Finally, the statistical measurement method is used for the calculation of values for the mean and standard deviation for each of the threshold values used. The following for present the outline of this chapter. Section 6.2, presents a discussion on the validation of multi-measurement criteria process for three main criteria. Section 6.3, makes a comparison of the color spaces. In Section 6.4, analysis the statistical measurements for different color spaces. Finally, Section 6.5

provides a summary of this Chapter. Figure 6.1 presents a diagrammatic view of the design and implementation of the validation process.

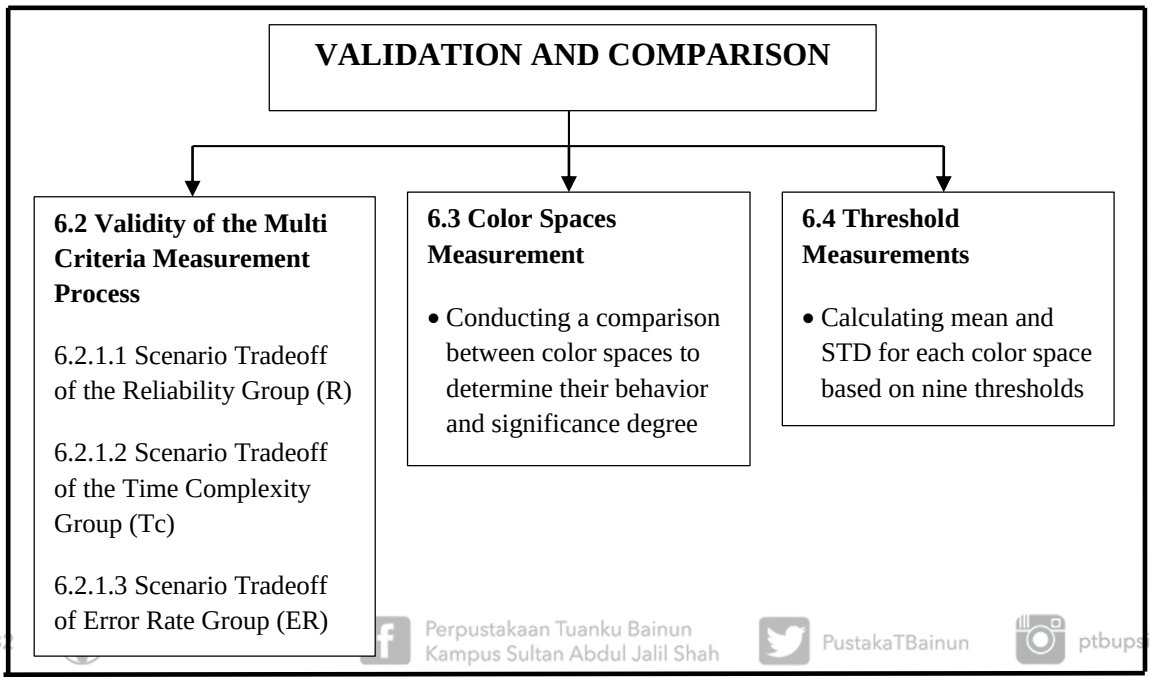


Figure 6.1. Overview of the Design and Implementation of the Validation Process

6.2 Validity of the Multi Criteria Measurement Process

Nowadays multi-criteria problem is taken into consideration as it a complex and crucial issue. So, to investigate in the trade-off problem between multi-criteria is considered an urgent matter. One of the investigation ways are used in the problem of trade-off. This is carried out by calculating the weights using numerical sequence process, as reported in details in Chapter 3. The process is performed based on the distribution of values of criteria that ranges from 1 to 0, then the value is decreased by 0.1, respectively. Thus, the results obtained are used as the ranking orders for these

values and the comparison with others values. Also paired sample t-test method is popular statistical methods for these criteria is applied of emphasizing the validity of the obtained results.

6.2.1 Discussion and Evaluation Trade-off for Multi Criteria Measurement

This study involves data collected from different criteria. The three main groups of criteria are used reliability group, time complexity and error rate group as discussed in Chapter 2. The process applied to obtain the desired results is based on multi-criteria that pose a significant challenge because of the trade-offs between them. As a result, the multi-criteria measurement method has been performed according to a specific methodology and distribution of the values among the criteria within a numerical sequence process that range from 1 to 0 in a descending order. There is a decrease of 0.1 for each weight. The values represent the weight for each criterion that is who the value equals to 1 it is at 100%. Meanwhile, when the value equals 0 it is at 0%. The main contribution to the implementation of such tests is to provide measurements per criterion, which is based on trade-off measurements in the evaluation of the conflict between the measure of criterion and the other. The following is a discussion of three different scenarios for the three main groups of criteria.

6.2.1.1 Scenario Tradeoff for the Reliability Group

The first scenario compares the reliability group with the time complexity and error rate groups to determine the trade-offs among them through a multi-criteria measurement process, thus evaluating the reliability group (See Table 6.1).

Table 6.1

Implementing Numerical Sequence Process for the Reliability Criterion

TN	TP	FP	FN	Accuracy	Recall	Precision	Specificity	F-measure	G-measure	Tc sec	ERV	ERT	
w1	w2	w3	w4	w5	w6	w7	w8	w9	w10	w11	w12	w13	Check
0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.0	0.00	0.00	1
0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.05	0.025	0.025	1
0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.1	0.05	0.05	1
0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.15	0.075	0.075	1
0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.2	0.1	0.1	1
0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.25	0.125	0.125	1
0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.3	0.15	0.15	1
0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.35	0.175	0.175	1
0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.4	0.2	0.2	1
0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.45	0.225	0.225	1
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.5	0.25	0.25	1

Table 6.1 shows the procedure validating the application of the numerical sequence to the distribution weights on the basis of the 13 criteria stated in the literature review. These weights are used to obtain the new results for the three main criteria and thus identify the trade-off among them. Figure 6.2 shows the results from the table.

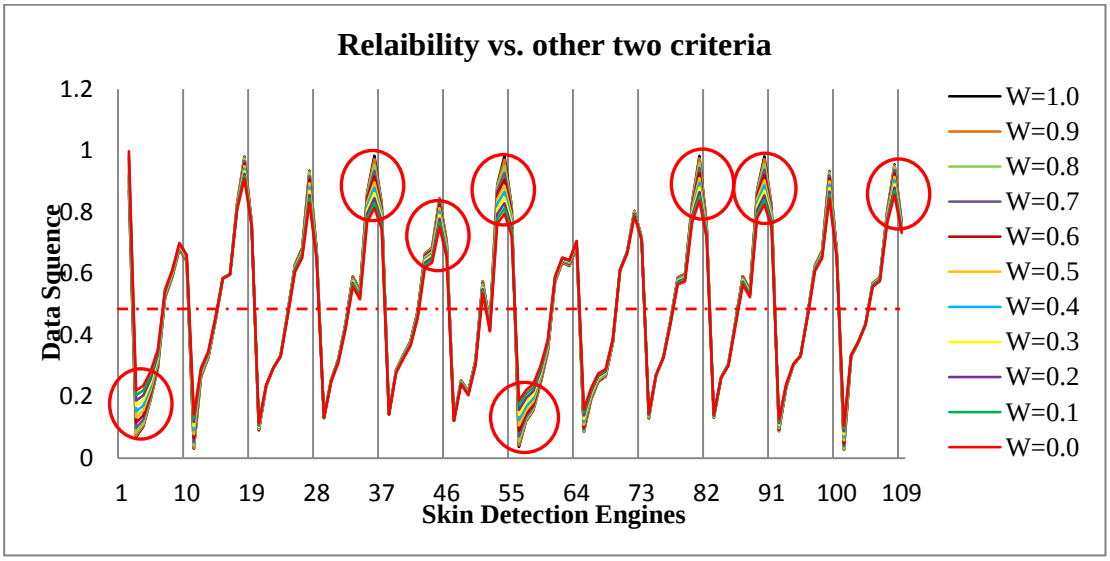


Figure 6.2. Trade-off Scenario between the Reliability Group Compared with other Groups

Figure 6.2 shows the distribution of the weights of the reliability criterion on the basis of the numerical sequence. These weights are distributed clearly at colors 1 and 7, which appear in the lower part of the curve. w = 0.0 is in the upper part of the curve. By contrast, w = 1.0 is below colors 1 and 7. The weights are distributed clearly at colors 4, 5, 6, 9, 10, and 12 in the same order as above, and they are indicators of the upper part of the curve. The graph shows that the weight distributions are affected by the threshold values. The reason is that the color spaces that appear at the top of the curve are the last threshold values, whereas the distribution of the color spaces that appear at the bottom of the curve is at the initial threshold values. The chart indicates that the threshold values influence the weight distribution of this criterion, unlike that of the other criteria. Moreover, regardless of the importance of each criterion, the value with a high threshold is prioritized.

6.2.1.1.1 Employment of Paired Sample Test of Scenario Reliability Group

At this stage, a paired sample test is performed to verify the results of the two samples. This stage mainly aims to highlight the effect of weight distribution of the comparative values. **Error! Reference source not found.**, shows the p-value results for different weights in ranked order of the reliability group.

Table 6.2

P-values for Different Weights of the Reliability Criterion

w=1.0	w=0.9	w=0.8	w=0.7	w=0.6	w=0.5	w=0.4	w=0.3	w=0.2	w=0.1	w=0.0
.218										
.198	.199									
.202	.207	.218								
.201	.206	.213	.211							
.197	.201	.206	.202	.195						
.190	.194	.197	.193	.186	.178					
.183	.186	.188	.184	.177	.170	.163				
.175	.178	.179	.175	.169	.162	.155	.149			
.168	.170	.171	.167	.161	.155	.148	.142	.136		
.104	.104	.100	.092	.082	.069	.054	.036	.016	.001	

Table 6.2 shows the p-values obtained from the comparisons of the different weights. The p-values of the comparisons (w = 0.4 with w = 0.0, w = 0.3 with w = 0.0, w = 0.2 with w = 0.0, and w = 0.1 with w = 0.0) are less than 0.05, which indicates statistically significant differences. The other comparisons do not exhibit the same result. Therefore, the experiment achieves the desired results.

Table 6.3

Results of the Correlation of Different Weights for the Reliability Criterion

w=1.0	w=0.9	w=0.8	w=0.7	w=0.6	w=0.5	w=0.4	w=0.3	w=0.2	w=0.1	w=0.0
1.000										
1.000	1.000									
.999	1.000	1.000								
.998	.999	.999	1.000							
.997	.998	.999	.999	1.000						
.996	.996	.997	.998	.999	1.000					
.994	.995	.996	.997	.998	.999	1.000				
.991	.992	.994	.995	.997	.998	.999	1.000			
.988	.989	.991	.993	.995	.997	.998	.999	1.000		
.985	.986	.988	.990	.992	.994	.996	.998	.999	1.000	

Table 6.3 presents the correlation coefficients obtained by the weight distribution process on the basis of the numerical sequence. The correlation coefficients are less than or equal to 1 at the significance level of less than 0.05. Therefore, the values of the measured samples are correlated.

In summary, weights distribution influences the reliability group somewhat comparison with other criteria, as revealed by the data in the tables

6.2.1.2 Scenario Tradeoff of the Time Complexity Group

The second scenario compares the time complexity group with the reliability and error rate groups to determine the trade-offs among them through a multi-criteria measurement process, thereby examining the time complexity group (See Table 6.4).

Table 6.4

Implementing the Numerical Sequence Process for the Time Complexity Criterion

TN	**TP**	**FP**	**FN**	**Accuracy**	**Recall**	**Precision**	**Specificity**	**F-measure**	**G-measure**	**Tc sec**	**ERV**	**ERT**	
w1	w2	w3	w4	w5	w6	w7	w8	w9	w10	w11	w12	w13	Check
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.0	0.00	0.00	1
0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.9	0.025	0.025	1
0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.8	0.05	0.05	1
0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.7	0.075	0.075	1
0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.6	0.1	0.1	1
0.025	0.025	0.025	0.025	0.025	0.025	0.025	0.025	0.025	0.025	0.5	0.125	0.125	1
0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.4	0.15	0.15	1
0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.3	0.175	0.175	1
0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.2	0.2	0.2	1
0.045	0.045	0.045	0.045	0.045	0.045	0.045	0.045	0.045	0.045	0.1	0.225	0.225	1
0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.0	0.25	0.25	1

Table 6.4 illustrates the implementation of the numerical sequence process on the distribution weights on the basis of the 13 criteria stated in the literature review. Three key criteria are adopted to determine the trade-off among them on the basis of the distribution weights presented in the table. Figure 6.3, shows the performance of the time complexity criterion.

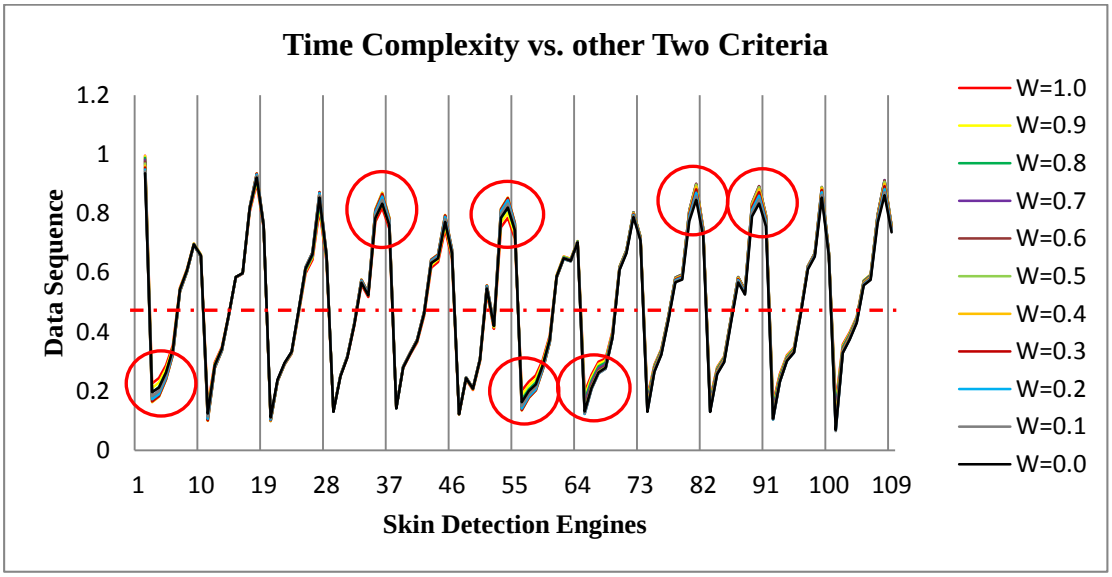


Figure 6.3. Trade-off Scenario for Time Complexity Group Compared with other Groups

Table 6.3, provides the weight distribution of the time complexity group obtained by the numerical sequence process. As shown in Figure 6.3, w = 1.0 is in the upper part of the curve, w = 0.3 is below colors 1 and 7, and w = 0.2 is below color 8. By contrast, the weights are distributed clearly in the top part of the graph at colors 4, 5, and 6. w = 1.0 is at the bottom, whereas w = 0.3 is in the upper part of the curve. In addition, the weights are distributed at colors 9 and 10. w = 0.0 is at the bottom, whereas w = 0.4 is in the upper part of the curve. The graph shows that the weight distributions are affected by the threshold values, given that the color spaces at the top and bottom of the curve are the final and initial threshold values, respectively. The chart implies that the threshold values affect the weight distribution of this criterion, unlike that of the other criteria. Furthermore, regardless of the importance of each criterion, the criterion with a high threshold is prioritized.

6.2.1.2.1 Employment of Paired Sample Test of Scenario Time Complexity Group

At this stage, a paired sample test is implemented to verify the results of the two samples. This stage mainly aims to highlight the effects of the weight distribution of the comparative values. **Error! Reference source not found.**5 shows the p-values for different weights in ranked order of the time complexity group.

Table 6.5

P-value for Different Weights of the Time Complexity Criterion

w=1.0	w=0.9	w=0.8	w=0.7	w=0.6	w=0.5	w=0.4	w=0.3	w=0.2	w=0.1	w=0.0
.114										
.631	.447									
.962	.304	.182								
.642	.174	.083	.027							
.380	.076	.025	.004	.000						
.181	.022	.004	.000	.000	.000					
.060	.003	.000	.000	.000	.000	.000				
.011	.000	.000	.000	.000	.000	.000	.000			
.001	.000	.000	.000	.000	.000	.000	.000	.001		
.000	.000	.000	.000	.000	.000	.000	.001	.003	.000	

Table 6.5 shows the p-values at the freedom level of 107. Most of the obtained values exhibit a statistically significant difference with the mean of the comparison samples at the significance level of less than 0.05. The comparisons of w = 1.0 and w = 0.3 with w = 0.2, w = 0.1, and w = 0.0 and those of w = 0.9 and w = 0.4 with w = 0.3, w = 0.2, w = 0.1, and w = 0.0 exhibit statistically significant differences. By contrast, the comparisons of w = 0.9 and w = 0.4 with w = 0.3, w = 0.2, w = 0.1, and w = 0.0; w =

0.8 and w = 0.5 with w = 0.4, w = 0.3, w = 0.2, w = 0.1, and w = 0.0; w = 0.7 and w = 0.6 with w = 0.5, w = 0.4, w = 0.3, w = 0.2, w = 0.1, and w = 0.0; w=0.2 with w = 0.1 and w = 0.0; and w = 0.1 with w = 0.0 exhibit statistically significant differences. Thus, the experiment achieves the desired results.

Table 6.6

Results of the Correlation of Different Weights for the Time Complexity Criterion

w=1.0	**w=0.9**	**w=0.8**	**w=0.7**	**w=0.6**	**w=0.5**	**w=0.4**	**w=0.3**	**w=0.2**	**w=0.1**	**w=0.0**
1.000										
.999	1.000									
.998	.999	1.000								
.997	.999	1.000	1.000							
.997	.998	.999	1.000	1.000						
.996	.998	.999	.999	1.000	1.000					
.996	.997	.998	.999	1.000	1.000	1.000				
.996	.997	.998	.999	.999	1.000	1.000	1.000			
.996	.997	.998	.999	.999	.999	1.000	1.000	1.000		
.996	.997	.998	.998	.999	.999	.999	.999	1.000	1.000	

Table 6.6, tabulates the correlation coefficients of the comparative samples obtained by the weight distribution process on the basis of the numerical sequence process. The correlation coefficient values are less than or equal to 1 at the significance level of less than 0.05. This finding confirms that the values of the measured samples are correlated.

In summary, the results of this criterion differ from those obtained from the reliability group. Thus, the weight distribution clearly affects the time complexity group compared with the other criteria.

6.2.1.3 Scenario Tradeoff of Error Rate Group

The third scenario compares the error rate group with the reliability and time complexity groups to determine the trade-offs among them through a multi-criteria measurement process, thereby examining the error rate group (See Table 6.7).

Table 6.7

Implementation of Multi Criteria Measurements in Error Rate Criterion

TN	**TP**	**FP**	**FN**	**Accuracy**	**Recall**	**Precision**	**Specificity**	**F-measure**	**G-measure**	**Tc sec**	**ERV**	**ERT**	
w1	w2	w3	w4	w5	w6	w7	w8	w9	w10	w11	w12	w13	Check
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.5	0.5	1
0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.05	0.45	0.45	1
0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.1	0.4	0.4	1
0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.15	0.35	0.35	1
0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.2	0.3	0.3	1
0.025	0.025	0.025	0.025	0.025	0.025	0.025	0.025	0.025	0.025	0.25	0.25	0.25	1
0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.3	0.2	0.2	1
0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.35	0.15	0.15	1
0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.4	0.1	0.1	1
0.045	0.045	0.045	0.045	0.045	0.045	0.045	0.045	0.045	0.045	0.45	0.05	0.05	1
0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.5	0.00	0.00	1

Table 6.7 presents the application of the numerical sequence process to the distribution weights on the basis of the 13 weights stated in the literature review. Three main criteria are adopted to determine the trade-off among them on the basis of the distribution weights in the table. Figure 6.4 shows the performance of the error rate within the dataset criterion.

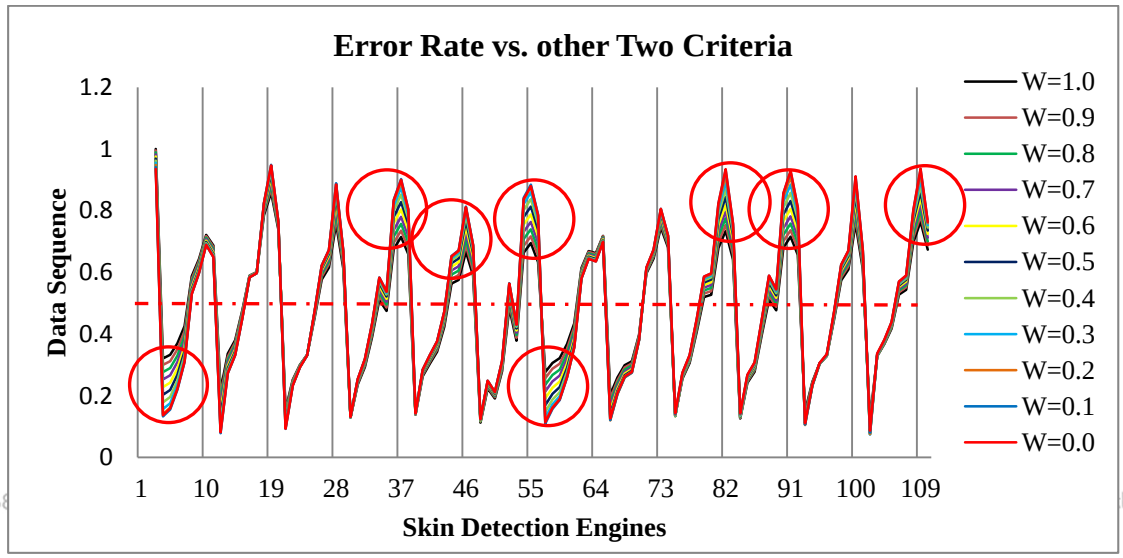


Figure 6.4. Tradeoff Scenario for Error Rate Group Compared with other Groups

Figure 6.4, shows the weight distribution of the error rate group obtained by the numerical sequence process. These weights are distributed clearly at colors 1 and 7, which appear in the lower part of the curve. As shown in figure 6.4 w = 1.0 is on the upper part of the curve, whereas w = 0.1 is below colors 1 and 7. By contrast, the weights are distributed clearly in the top part of the graph at colors 4, 5, 6, 9, 10, and 12. w = 1.0 is at the bottom, whereas w = 0.1 is at the upper part of the curve. The graph shows that the weight distributions are affected by the threshold values, given that the color spaces at the top and bottom of the curve are the middle and initial threshold values, respectively. The chart shows that the threshold values affect the

weight distribution of this criterion, unlike the other criteria. In addition, regardless of the importance of each criterion, the criterion with a high threshold is prioritized.

6.2.1.3.1 Employment of Paired Sample Test of Scenario Error Rate Group

At this stage, a paired sample test is performed to verify the results of the two samples. This stage mainly aims to highlight the effect of the weight distribution of comparative values. **Error! Reference source not found.**8 provides the p-values for different weights in ranked order of the error rate group.

Table 6.8

P-value for Different Weights of the Error Rate Criterion

w=1.0	w=0.9	w=0.8	w=0.7	w=0.6	w=0.5	w=0.4	w=0.3	w=0.2	w=0.1	w=0.0
.000										
.002	.007									
.003	.007	.007								
.003	.007	.007	.007							
.004	.007	.007	.007	.007						
.004	.007	.007	.007	.007	.006					
.004	.006	.006	.006	.006	.005	.004				
.004	.005	.005	.005	.005	.004	.003	.002			
.003	.004	.004	.004	.003	.002	.002	.001	.000		
.002	.003	.003	.002	.002	.001	.000	.000	.000	.005	

Table 6.8 shows the p-values on the basis of the comparisons of different weights at the freedom level of 107. The mean value for all samples shows a significance value

of less than 0.05 has statistically significant differences. Thus, the experiment achieves the desired results.

Table 6.9

Results of the Correlation between Different Weights for Error Rate Criterion

w=1.0	**w=0.9**	**w=0.8**	**w=0.7**	**w=0.6**	**w=0.5**	**w=0.4**	**w=0.3**	**w=0.2**	**w=0.1**	**w=0.0**
.999										
.996	.999									
.992	.996	.999								
.986	.992	.997	.999							
.979	.987	.993	.997	.999						
.973	.982	.989	.995	.998	1.000					
.967	.977	.985	.992	.996	.998	1.000				
.962	.973	.982	.989	.994	.997	.999	1.000			
.959	.971	.980	.987	.993	.996	.998	.999	1.000		
.959	.971	.980	.987	.993	.996	.998	.999	1.000	1.000	

Table 6.9 illustrates the correlation coefficients of the comparative samples obtained by the weight distribution process on the basis of the numerical sequence. The correlation values are less than or equal to 1 at the significance level of less than 0.05. These results show that the values of the measured samples are correlated.

In summary, the results of this criterion differ from those of the reliability and time complexity groups. Thus, the weight distribution clearly affects the error rate group, compared with the other criteria.

6.2.1.4 Summary of Scenarios

A multi-criteria measurement process is implemented on the three main groups, namely, the reliability, time complexity, and error rate groups, by applying the numerical sequence process adopted in this section. Three basic scenarios are selected for the weight distribution through three experiments to determine the trade-offs among the criteria. The experimental results are plotted on three charts that show the distribution of the color spaces and their threshold values, which influence the behavior of the criteria. Findings reveal that a trade-off exists among these criteria on the basis of their weight distribution. Thus, the numerical sequence process was used as a general alternative rather than a subjective strategy that relies on the perspective of evaluators; this approach can be used for any application.

Paired sample tests are performed to ensure the validity of the results of each criterion. This method generates different results according to the comparative samples used in the various experiments. The reliability criterion exhibits statistically significant relationship for some of the number of the measured samples. A statistically significant relationship is revealed by the mean values of most of the measured samples for the time complexity samples. In contrast, the error ratio criterion shows a statistically significant relationship for all the error rate samples. The statistically significant differences found by the paired sample test indicate that the experiments on the values of the comparative samples achieve the desired goal. The results show a correlation among the measured samples for each criterion at significance levels of less than 0.05.

6.3 Color Spaces Measurement

As mentioned in Chapter 5, TOPSIS method is used to select the best alternative based on the different criteria in skin detection. As such two categories of results were collected namely internal aggregation and external aggregation. External aggregation is our goal in this study. For this research study, the goal was to obtain results from external aggregation as it includes all values of comparison between the criteria and alternatives based on the calculation of average between different ranking values. For that reason, the graph shows the behavior of different color spaces according to the criteria at specific threshold values. (See Figure 6.5).

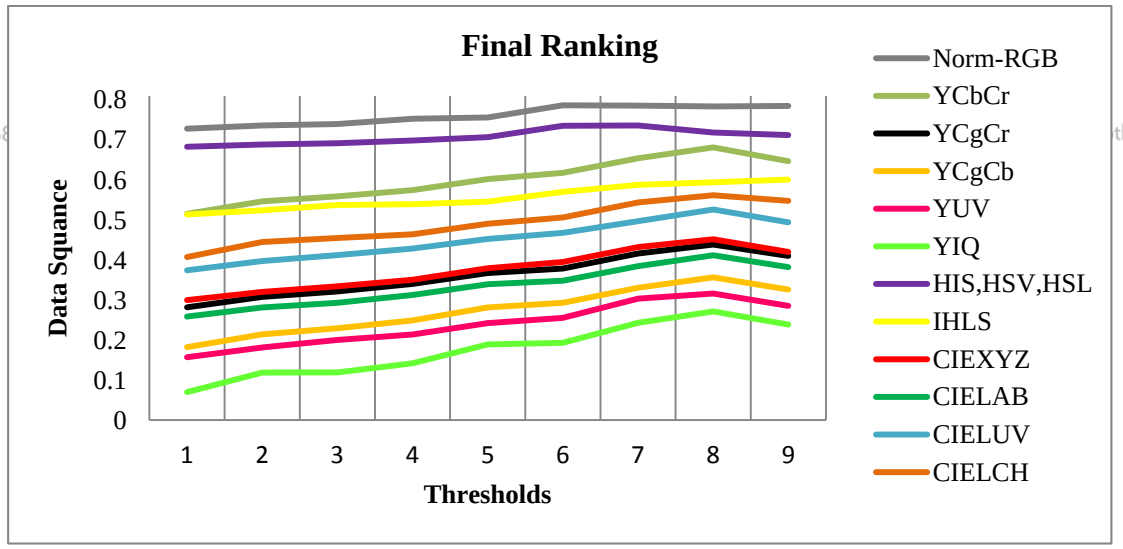


Figure 6.5. Color Space Measurement

Figure 6.5, shows that the behavior of all the color spaces is somewhat identical based on their original values. The behavior of YIQ starts at the lowest value at threshold 1 after which it increases slightly until threshold 2. Then it stabilizes to the threshold 3 followed by the slight rise to threshold 5. It continues stabilizes until

threshold 6. This is followed by another slight rise to threshold 8 before it drops to threshold 9. Meanwhile, the behavior of YUV starts with a slight rise from threshold 1 to threshold 4. This upgrade trend continues to rise to the threshold 6 and to the threshold 7, after which it stabilizes until threshold 8 before it declines slightly until threshold 9. As for the behavior, of YCgCb and CIELAB they start from threshold 1 to threshold 2 before a slight rise to threshold 5. Then they stabilize until threshold 6 before they continue to rise to threshold 8. After that, they decline slightly to threshold 9. While the YCgCr and CIEXYZ show them identical behavior, they start a gradual rise from threshold 1 to threshold 5. Then they stabilize slightly to threshold 6 before they rise slightly to threshold 8. Finally, they drop slightly to threshold 9. In contrast, the CIELUV shows a distinct behavior, it starts gradual rise from threshold 1 to threshold 8. Then it declines slightly to the threshold of 9. While the behavior of CIELCH starts from threshold 1 to threshold 2 before a slight rise to threshold 5. Then it stabilizes slightly to threshold 6 before it rises slightly to threshold 8. Then, it drops slightly to threshold 9. Whereas the IHLS shows a distinct behavior, it starts gradual rise from threshold 1 until threshold 9 at the same level. Meanwhile, the YCbCr starts from threshold 1 to threshold 2 before a slight rise to threshold 5. Then it stabilizes slightly to threshold 6 before it rises slightly to threshold 8. Then, it declines slightly to threshold 9. As for the behavior of the HIS, HSV, and HSL they start from threshold 1 to rise slightly until threshold 5. After that, they continue with a slight rise to threshold 6. Then they stabilize until threshold 7, after that they decline slightly until threshold 9. In the last color space, the behavior of Norm-RGB starts from

threshold 1, and then gradually rising until threshold 5, followed by another slight rise to threshold 6. After that, it stabilizes its track until threshold 9.

In short, from the graph, it can be concluded that the behavior of all the color spaces was differentiated according to their values obtained from the external aggregation. While the YIQ has recorded the werst color space, and the Norm-RGB has recorded the best color space among other. As for, the rest of the colors they come sequentially as shown in the chart above.

6.4 Threshold Measurements

This section, presents the calculation of the mean and the standard deviation value of the threshold values for all color spaces. Nine threshold values were adopted according to the case study in this research. The threshold values are distributed as follows (0.5, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9 and 0.95) for each color spaces. See Table 6.0.

Table 6.10

Mean and Stander Division for Threshold values

	Mean	***±SD***
Ø = 0.5	0.372021	0.256188
Ø = 0.6	0.392262	0.249095
Ø = 0.65	0.400940	0.247265
Ø = 0.7	0.413236	0.244608
Ø = 0.75	0.432983	0.234217
Ø = 0.8	0.446975	0.241581
Ø = 0.85	0.473835	0.223924
Ø = 0.9	0.491004	0.210286
Ø = 0.95	0.468699	0.223906

Table 6.0 highlights the results obtained for each threshold value. The threshold of (0.5) was recorded at 0.372021 for the mean value and 0.256188 for the ±SD for all color spaces. As for the threshold of (0.6) the mean value was recorded at 0.392262 and 0.249095 for the ±SD for all color spaces. The mean value for the threshold of (0.65) was recorded at 0.400940 and 0.247265 for the ±SD for all color spaces. While the mean value for the threshold of (0.7) was recorded at 0.413236 and 0.244608 for the ±SD for all color spaces. The mean value for the threshold of the threshold of (0.75) was recorded at 0.432983 and 0.234217 for the ±SD for all color spaces. As for the mean value for the threshold of (0.8) was recorded at 0.446975 and 0.241581 for the ±SD for all color spaces. For the mean value for the threshold of (0.85) was recorded at 0.473835 and 0.223924 for the ±SD for all color spaces. As for the mean value for the threshold of (0.9) was recorded at 0.491004 and 0.210286 for the ±SD for all color spaces. Finally, the mean values for the threshold of (0.95) was recorded at 0.468699 and 0.223906 for the ±SD for all color spaces.

In summary, from the results above it can be concluded that the mean values for each threshold value start from the lowest value of threshold 0.5, following by a gradual increase until they reach the highest value at the threshold of 0.9 before they return to decrease again at the threshold 0.95. The findings also indicate that the best result of the threshold values was at the threshold of (0.9) for all color spaces.

6.5 Chapter Summary

The results obtained require validation. To facilitate this process, the validation process was conducted in two trends. The first trend is the implementation of the numerical sequence process for the distribution of the weights collected for three main criteria in order to identify the tradeoff of the criteria. The findings revealed that this process produced good results based on the figures. For the second trend involves the use statistical measurements for calculating the threshold values. This procedure calculation of threshold values is to determine the best value among them. It is noted that validation operation has achieved the results which are tabulated in Chapter 5 of the research. The recommendations, objectives and future works are discussed as in Chapter 7.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

7.1 Introduction

This chapter presents a summary of research contributions, limitations, and future work of the research.

Section 7.2, describes the conclusion. Section 7.3 reports the research limitations and issues encountered in the course of the research. Section 7.4 addresses some directions to be adapted for future work studies. Section 7.5, presents the conclusion of this research.

7.2 Conclusions

This section, highlights the contributions of this research. The contributions include three basic outcomes based on the evaluation and benchmarking of the skin detection approaches.

1- Problem analysis in evaluation and benchmarking of skin detection

Skin detection approaches are an important research area. However, despite existing the tradeoff between different criteria which created the gap in this study. Thus, the main goal of the study established the decision matrix in order to identify the weights based on different skin detection engines. According to the case study, it is necessary creating the decision matrix that will be used to generate the final results. This contribution fulfills and meets the second objective which is mentioned in Chapter 3.

2- Adapt ML-AHP and TOPSIS techniques

MCDM includes several techniques used to achieve different processes in various environments. In this research, the integration between two AHP and TOPSIS methods are implemented for the purpose of ranking and selects the best alternative in the evaluation and benchmarking of skin detection approaches. This contribution related is to meet the requirements of the third objective which has been proposed in

the third phase of research methodology. Further details have been discussed in Chapter 5.

3- Validation of parameters

The validation of different results is performed in order to select the best technique. Also, the validation of carried out solutions to identify the relationships between criteria and color spaces based on statistical methods. This validation of parameters is necessary to fourth objective four has been proposed for fourth phase of this research methodology where details are mentioned in Chapter 6.

7.3 Research Limitation and Issues

In most research, there is a need to overcome some limitations for improvements in future research work. One of the limitations of this research is the scope of the experiment. The following is a discussion of the limitations of this research.

1- Statistical Complicated

There are several ways to collect data during the image processing. Basically, the skin detection approach relies on the image segmentation process in this study. Thus, this study used the manual image segmentation process. The manual process often gives inaccurate results when the segmentation operation of the image is carried out. Thus,

it affects on the calculation of the ratio of true positive and false negative due to the loss of the number of pixels during image segmentation operation.

2- Dataset Measure

The dataset is considered a major challenge for many researchers within the scope of machine learning. However, the dataset is created based on the case study requirement, according to the researcher's conditions which meet the requirements of the study. Due to an absence of a standard measure of the dataset, this limitation is not within the scope of this research.

3- Subjective Judgment Techniques

The research highlighted the most important decision-making technique that relies on subjective judgments. Such this technique is a hierarchical analysis process (AHP) which is based on subjective judgments and pair-wise comparisons. This method has some limitaions due to depending on the expert's perspective that often is incorrect. On the other hand, the AHP method is considered accurate in calculating the weights of the criteria we need in the TOSIS method.

4- Selection Process of the Skin Detection

The skin detection approaches included different applications. It depends on different criteria for selection and classification process. Basically, each application depends on specific criteria when conducting an assessment and comparison under certain circumstances. Therefore, the selection process of the skin detection is conducted based on the type of application.

5- Two targets adapted in the case study

In this research, two targets were implemented during the segmentation process which related to the case study of skin detection. Despite there existing other targets such as multi-class, multi-label, and hierarchy that can be used with another study. Thus, in this study used segmentation process was performed only on the skin and non-skin pixels.

7.4 Future work

The limitations identified in this research are out of this research scope. Hence, these limitations should be addressed to establish a new methodology for skin detection approaches. The following is a review of some of the future research proposal.

1- It is necessary to adopt more case studies to confirm the requirement of the research. These requirements can be achieved through developing and enhancing the case studies proposed within the new methodology for evaluation and benchmarking of skin detection approaches.

2- It is useful for application of different color spaces with both luminance and chrominance the different cases studies. These parameters are considered important for verification of color spaces during the implementation process.

3- The implementation of the methodology of evaluation and benchmarking is suitable for most adapt research activities and requirements of markets. Therefore, in order to the solution that is adopted a comprehensive analysis is required for the evaluation of criteria which is a suitable solution in the development of academic studies which is used extensively in the evaluation and benchmarking of skin detection approaches.

4- It has been found that the proposed new methodology is suitable for using different applications of skin detection. So, this methodology is compatible for these applications because of their reliance on the criteria used in this approach.

5- The new methodology of skin detection approaches can be applied in the real-time applications.

6- The new methodology can be applied for skin detection approaches with applications that involves the use of three targets such as skin cancer application.

7- The multi-objective decision-making technique can be performed on different algorithms such as genetic algorithm and swarm.

REFERENCES

Abadpour, A, and S Kasaei. 2005. "Pixel-Based Skin Detection for Pornography Filtering." *Iranian Journal of Electrical & Electronic Engineering* 1 (3): 21–41. http://ijeee.iust.ac.ir/browse.php?a_code=A-10-3-51&slc_lang=en&sid=1.

Abbas, Qaisar, M. E. Celebi, and Irene Fondón García. 2011. "Hair Removal Methods: A Comparative Study for Dermoscopy Images." *Biomedical Signal Processing and Control* 6 (4). Elsevier Ltd: 395–404. doi:10.1016/j.bspc.2011.01.003.

Abdullah-Al-Wadud, M., Mohammad Shoyaib, and Oksam Chae. 2009. "A Skin Detection Approach Based on Color Distance Map." *EURASIP Journal on Advances in Signal Processing* 2008 (1): 814283. doi:10.1155/2008/814283.

Al-Azab, Fadwa Gamal Mohammed, Media A Ayu, and Fadwa Gamal Mohammed Ai-azabl. 2008. "Decision Making with the Analytic Hierarchy Process." *IInternational Journal of Services Sciences.* 1 (1): 83–98. doi:10.1109/ICT4M.2010.5971886.

Al-Boeridi, Omar N., S. M. Syed Ahmad, and S. P. Koh. 2015. "A Scalable Hybrid Decision System (HDS) for Roman Word Recognition Using ANN SVM: Study Case on Malay Word Recognition." *Neural Computing and Applications* 26 (6). Springer London: 1505–13. doi:10.1007/s00521-015-1824-0.

Al-Mohair, H. K., Saleh, J. M., & Suandi, S. A. 2015. "Hybrid Human Skin Detection Using Neural Network and K-Means Clustering Technique." *Applied Soft Computing* 33: 337–47.

Al-Mohair, Hani K., Junita Mohamad-Saleh, and Shahrel Azmin Suandi. 2014. "Color Space Selection for Human Skin Detection Using Color-Texture Features and Neural Networks." In *2014 International Conference on Computer and Information Sciences (ICCOINS)*, 1–6. doi:10.1109/ICCOINS.2014.6868362.

Al-Mohair, Hani K., Junita Mohamad Saleh, and Shahrel Azmin Suandi. 2015. "Hybrid Human Skin Detection Using Neural Network and K-Means Clustering Technique." *Applied Soft Computing Journal* 33. Elsevier B.V.: 337–47. doi:10.1016/j.asoc.2015.04.046.

Al-Mohair, Hani K., Junita Mohamed-Saleh, and Shahrel Azmin Suandi. 2012. "Human Skin Color Detection: A Review on Neural Network Perspective." *Internation Journal of Innovative Computing, Information and Control2* 8 (12): 8115–31.

Al-odan, Hussah A, and Ahmad A Al-daraiseh King Saud. 2015. "Open Source Data Mining Tools." In *In Electrical and Information Technologies (ICEIT), 2015 International Conference ,IEEE*, 369–74. doi:10.1109/EITech.2015.7162956.

Al Abbadi, Nidhal K., Nizar Saadi Dahir, and Zaid Abd Alkareem. 2013. "Skin Texture Recognition Using Neural Networks." In *2008 International Arab Conference on Information Technology (ACIT 2008)*, 3–6.

Albiol, Alberto, Luis Torres, and Edward J. Delp. 2001. "Optimum Color Spaces for Skin Detection." In *IEEE International Conference*, 122–24.

Alcalá-Fdez, J., A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera. 2011. "KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework." *Journal of Multiple-Valued Logic and Soft Computing* 17 (2–3): 255–87. doi:10.1007/s00500-008-0323-y.

Alcalá-Fdez, J., L. Sánchez, S. García, M. J. del Jesus, S. Ventura, J. M. Garrell, J. Otero, et al. 2009. "KEEL: A Software Tool to Assess Evolutionary Algorithms for Data Mining Problems." *Soft Computing* 13 (3): 307–18. doi:10.1007/s00500-008-0323-y.

Aldlaigan, Abdullah H., and Francis A. Buttle. 2002. "SYSTRA-SQ: A New Measure of Bank Service Quality." *International Journal of Service Industry Management* 13 (4): 362–81. doi:10.1108/09564230210445041.

Anjos, A., El-Shafey, L., Wallace, R., Günther, M., McCool, C., & Marcel, S. 2012. "Bob: A Free Signal Processing and Machine Learning Toolbox for Researchers André." In *In Proceedings of the 20th ACM International Conference on Multimedia . ACM.*, 1449–52. doi:10.1109/SIBGRAPI-T.2011.11.

Araban, Sepideh, Fardad Farokhi, and Kave Kangarloo. 2011. "Determining Effective Colour Components for Skin Detection Using a Clustered Neural Network." In *2011 IEEE International Conference on Signal and Image Processing Applications, ICSIPA 2011*, 547–52. doi:10.1109/ICSIPA.2011.6144144.

Araokar, Shashank. 2005. "Visual Character Recognition Using Artificial Neural Networks." *Arxiv Preprint cs0505016*, 1–7. http://arxiv.org/abs/cs/0505016.

Ardil, Ebru, and Parvinder S Sandhu. 2010. "A Soft Computing Approach for Modeling of Severity of Faults in Software Systems." *International Journal of Physical Sciences.* 5 (2): 74–85.

Aruldoss, M., Lakshmi, T. M., & Venkatesan, V. P. 2013. "A Survey on Multi Criteria Decision Making Methods and Its Applications." *American Journal of Information Systems* 1 (1): 31–43. doi:10.12691/ajis-1-1-5.

Ashwin Satyanarayana. 2013. "Software Tools for Teaching Undergraduate Data Mining Course." In *Proceedings of the ASEE-2013 Mid-Atlantic Fall Conference, University of the District of Columbia*, 1–15.

Asuero, A. G., A. Sayago, and A. G. González. 2006. "The Correlation Coefficient: An Overview." *Critical Reviews in Analytical Chemistry* 36 (1): 41–59.

doi:10.1080/10408340500526766.

Bajcsy, Peter, Antoine Vandecreme, Julien Amelot, Phuong Nguyen, Joe Chalfoun, and Mary Brady. 2013. "Terabyte-Sized Image Computations on Hadoop Cluster Platforms." In *Proceedings - 2013 IEEE International Conference on Big Data, Big Data 2013*, 729–37. doi:10.1109/BigData.2013.6691645.

Ballerini, Lucia, Rb Fisher, Ben Aldridge, and Jonathan Rees. 2013. "A Color and Texture Based Hierarchical K-NN Approach to the Classification of Non-Melanoma Skin Lesions." In *Color Medical Image Analysis-Springer, Dordrecht*, 6:63--86. doi:10.1007/978-94-007-5389-1_4.

Baringhaus, Ludwig, and Daniel Gaigall. 2017. "Hotelling's T2tests in Paired and Independent Survey Samples: An Efficiency Comparison." *Journal of Multivariate Analysis* 154. Elsevier Inc.: 177–98. doi:10.1016/j.jmva.2016.11.004.

Basheer, I A, and M Hajmeer. 2000. "Artificial Neural Networks : Fundamentals , Computing , Design , and Application." *Journal of Microbiological Methods.* 43 (1): 3–31.

Belaroussi, Rachid, and Maurice Milgram. 2012. "A Comparative Study on Face Detection and Tracking Algorithms." *Expert Systems with Applications* 39 (8). Elsevier Ltd: 7158–64. doi:10.1016/j.eswa.2012.01.076.

Bergmeir, Christoph, Mauro Costantini, and José M. Benítez. 2014. "On the Usefulness of Cross-Validation for Directional Forecast Evaluation." *Computational Statistics and Data Analysis* 76 (2009). Elsevier B.V.: 132–43. doi:10.1016/j.csda.2014.02.001.

Berthold, Michael R., Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Kilian Thiel, and Bernd Wiswedel. 2009. "KNIME - the Konstanz Information Miner." *ACM SIGKDD Explorations Newsletter* 11 (1): 26. doi:10.1145/1656274.1656280.

Bhoyar, K K, and O G Kakde. 2010. "Skin Color Detection Model Using Neural Networks and Its Performance Evaluation Department of Computer and Information Technology , Yeshwantrao Chavan College of Engineering , Nagpur , India Department of Electronics and Computer Science , Vishweshwarayya." *In Journal of Computer Science.* 6 (9): 963–68.

Bhushan, Navneet, and Kanwal Rai. 2007. *Strategic Decision Making: Applying the Analytic Hierarchy Process*. doi:10.1007/b97668.

Bilal, Sara, Rini Akmeliawati, Momoh Jimoh E. Salami, and Amir A. Shafie. 2015. "Dynamic Approach for Real-Time Skin Detection." *Journal of Real-Time Image Processing* 10 (2): 371–85. doi:10.1007/s11554-012-0305-2.

Borah, S, and S Konwar. 2014. "ANN Based Human Facial Expression Recognition

in Color Images." In *2014 International Conference on High Performance Computing and Applications (ICHPCA)*, 1–6. doi:10.1109/ICHPCA.2014.7045337.

Brown, Capt Gregory E, and Edward D White. 2017. "An Investigation of Nonparametric DATA MINING TECHNIQUES." *Defense Acquisition Research Journal: A Publication of the Defense Acquisition University*, 24 (2): 302–32. doi:10.22594/dau.16-756.24.02.

Brutto, M. Lo, and P. Meli. 2012. "Computer Vision Tools for 3D Modelling in Archaeology." *International Journal of Heritage in the Digital Era* 1 (1): 1–6. doi:10.1260/2047-4970.1.0.1.

Burget, Radim, Jan Karasek, Zdenek Smekal, Vaclav Uher, and Otto Dostal. 2010. "RapidMiner Image Processing Extension:A Platform for Collaborative Research." In *In The 33rd International Conference on Telecommunication and Signal Processing, TSP*, 114–18.

Butler, James R.A., Grace Y. Wong, Daniel J. Metcalfe, Miroslav Honzák, Petina L. Pert, Nalini Rao, Martijn E. van Grieken, et al. 2013. "An Analysis of Trade-Offs between Multiple Ecosystem Services and Stakeholders Linked to Land Use and Water Quality Management in the Great Barrier Reef, Australia." *Agriculture, Ecosystems and Environment* 180. Elsevier B.V.: 176–91. doi:10.1016/j.agee.2011.08.017.

Cao, Xinyan, and Hongfei Liu. 2012. "A Skin Detection Algorithm Based on Bayes Decision in the YCbCr Color Space." *In Applied Mechanics and Materials , Trans Tech Publications.* 121: 672–76. doi:10.4028/www.scientific.net/AMM.121-126.672.

Cerrillo-Cuenca, Enrique, and Marcela Sepúlveda. 2015. "An Assessment of Methods for the Digital Enhancement of Rock Paintings: The Rock Art from the Precordillera of Arica (Chile) as a Case Study." *Journal of Archaeological Science* 55. Elsevier Ltd: 197–208. doi:10.1016/j.jas.2015.01.006.

Chai, Douglas, and Abdesselam Bouzerdoum. 2000. "A BAYESIAN APPROACH TO SKIN COLOR CLASSIFICATIONIN YCBCR COLOR SPACE." In *In TENCON 2000. Proceedings IEEE.*, 421–24.

Chaouch, Melek, Moez Mhadhbi, Emily R. Adams, Gerard J. Schoone, Sassi Limam, Zyneb Gharbi, Mohamed Aziz Darghouth, Ikram Guizani, and Souha BenAbderrazak. 2013. "Development and Evaluation of a Loop-Mediated Isothermal Amplification Assay for Rapid Detection of Leishmania Infantum in Canine Leishmaniasis Based on Cysteine Protease B Genes." *Veterinary Parasitology* 198 (1–2). Elsevier B.V.: 78–84. doi:10.1016/j.vetpar.2013.07.038.

Chaves-González, Jose M., and Miguel A. Vega-Rodríguez. 2010. "Detecting Skin in Face Recognition Systems: A Colour Spaces Study." *Jose M. Chaves-González ∗ , Miguel A. Vega-Rodríguez, Juan A. Gómez-Pulido, Juan M. Sánchez-Pérez* 20

(3). Elsevier Inc.: 806–23. doi:10.1016/j.dsp.2009.10.008.

Chaves-González, Jose M., Miguel A. Vega-Rodríguez, Juan A. Gómez-Pulido, and Juan M. Sánchez-Pérez. 2010. "Detecting Skin in Face Recognition Systems: A Colour Spaces Study." *Digital Signal Processing: A Review Journal* 20 (3). Elsevier Inc.: 806–23. doi:10.1016/j.dsp.2009.10.008.

Chen, Chen-Tung. 2000. "Extensions of the TOPSIS for Group Decicion-Making under Fuzzy Environment." *Fuzzy Sets and Systems* 114: 1–9.

Chen, Wei, Ke Wang, Haifeng Jiang, and Ming Li. 2016. "Skin Color Modeling for Face Detection and Segmentation: A Review and a New Approach." *Multimedia Tools and Applications* 75 (2): 839–62. doi:10.1007/s11042-014-2328-0.

Chen, Yen Hsiang, Kai Ti Hu, and Shanq Jang Ruan. 2012. "Statistical Skin Color Detection Method without Color Transformation for Real-Time Surveillance Systems." *Engineering Applications of Artificial Intelligence* 25 (7). Elsevier: 1331–37. doi:10.1016/j.engappai.2012.02.019.

Chinchor, N., & Sundheim, B. 1993. "MUC-4 Evaluation Metrics." In *In Proceedings of the 5th Conference on Message Understanding . Association for Computational Linguistics.*, 69–78. Elsevier. doi:10.1016/j.firesaf.2015.02.007.

Cho, Kyunghyun. 2014. *Introduction to Machine Learning*.

Choi, Byeongcheol, Byungho Chung, and Jaecheol Ryou. 2009. "Adult Image Detection Using Bayesian Decision Rule Weighted by SVM Probability." In *2009 Fourth International Conference on Computer Sciences and Convergence Information Technology*, 659–62. doi:10.1109/ICCIT.2009.43.

Comaniciu, Dorin, and Peter Meer. 2002. "Mean Shift : A Robust Approach toward Feature Space Analysis 1 Introduction." *IEEE Transactions* 24 (5): 603–19. doi:10.1109/34.1000236.

Daithankar, Mrunmayee V, Kailash J Karande, and Avinash D Rarale. 2014. "Analysis of Skin Color Models for Face Detection." In *In Communications and Signal Processing (ICCSP), 2014 International Conference on IEEE.*, 533–37. doi:10.1109/ICCSP.2014.6949899.

Davis, Jesse, and Mark Goadrich. 2006. "The Relationship between Precision-Recall and ROC Curves." In *Proceedings of the 23rd International Conference on Machine Learning - ICML '06*, 233–40. doi:10.1145/1143844.1143874.

De-La-Torre, Miguel, Eric Granger, Robert Sabourin, and Dmitry O. Gorodnichy. 2015. "Adaptive Skew-Sensitive Ensembles for Face Recognition in Video Surveillance." *Pattern Recognition* 48 (11). Elsevier: 3385–3406. doi:10.1016/j.patcog.2015.05.008.

Demuth, Howard, and Mark Beale. 2009. *Neural Network Toolbox*.

Deng, H., Yeh, C. H., & Willis, R. J. 2002. "Inter-Company Comparison Using Modified TOPSIS with Objective Weights." *Computers & Operations Research*, 27 (10): 963–73.

Dongare, A D, R R Kharde, and Amit D Kachare. 2012. "Introduction to Artificial Neural Network." *International Journal of Engineering and Innovative Technology (IJEIT)* 2 (1): 189–94.

Doukim, C.A., Dargham, J.A., Chekima, A. and Omatu, S. 2011. "Combining Neural Networks for Skin Detection." *Signal & Image Processing : An International Journal (SIPIJ)* 1 (2): 1–11. doi:10.5121/sipij.2010.1201.

Duan, Lijuan, Zhiqiang Lin, Jun Miao, and Yuanhua Qiao. 2009. "A Method of Human Skin Region Detection Based on PCNN." *Lecture Notes in Computer Science* 5553: 486–93. doi:10.1007/978-3-642-01513-7.

Duffner, Stefan, and Jean Marc Odobez. 2014. "Leveraging Colour Segmentation for Upper-Body Detection." *Pattern Recognition-Elsevier* 47 (6): 2222–30. doi:10.1016/j.patcog.2013.12.014.

Duke, Joshua M, and Rhonda Aull-hyde. 2002. "Identifying Public Preferences for Land Preservation Using the Analytic Hierarchy Process." *Ecological Economics* 42: 131–45.

Dwivedi, Shraddha, Paridhi Kasliwal, and Suryakant Soni. 2016. "Comprehensive Study of Data Analytics Tools (RapidMiner, Weka, R Tool, Knime)." In *In Colossal Data Analysis and Networking (CDAN), Symposium on IEEE.*, 1–8. doi:10.1109/CDAN.2016.7570894.

Ebrahimzadeh, Ataollah, and Ali Khazaee. 2010. "Detection of Premature Ventricular Contractions Using MLP Neural Networks: A Comparative Study." *Measurement: Journal of the International Measurement Confederation* 43 (1). Elsevier Ltd: 103–12. doi:10.1016/j.measurement.2009.07.002.

Egghe, Leo, and Loet Leydesdorff. 2009. "The Relation between Pearson' S Correlation Coefficient R and Salton' S Cosine Measure." *Journal of the Association for Information Science and Technology* 60 (5): 1027–36. doi:10.1002/asi.

Elalami, M. E. 2014. "A New Matching Strategy for Content Based Image Retrieval System." *Applied Soft Computing Journal* 14. Elsevier B.V.: 407–18. doi:10.1016/j.asoc.2013.10.003.

Elgammal, Ahmed, Crystal Muang, and Dunxu Hu. 2009. "Skin Detection -a Short Tutorial Skin Detection - a Short Tutorial †." In *Encyclopedia of Biometrics-Springer*, 1–10. doi:10.1007/978-0-387-73003-5.

Esposito, L. G., & Sansone, C. 2013. "A Multiple Classifier Approach for Detecting Naked Human Bodies in Images." In *In International Conference on Image*

Analysis and Processing . Springer, Berlin, Heidelberg., 389–98. doi:10.1109/ICCV.2013.354.

Faith, Daniel P, C R Margules, and P A Walker. 2000. "A Biodiversity Conservation Plan for Papua New Guinea Based on Biodiversity Trade-Offs Analysis Published Version : Faith , D . P ., Margules , C . R . and Walker , P . A ., (2001) A Biodiversity Conservation Plan for Papua New Guinea Based on Biodivers." *Pacific Conservation Biology*, 6 (4): 304–24.

Fawcett, Tom. 2006. "An Introduction to ROC Analysis." *Pattern Recognition Letters* 27 (8): 861–74. doi:10.1016/j.patrec.2005.10.010.

Felzenszwalb, Pedro F., and Daniel P. Huttenlocher. 2004. "Efficient Graph-Based Image Segmentation." *International Journal of Computer Vision* 59 (2): 167–81. doi:10.1023/B:VISI.0000022288.19776.77.

Fernandes, Bruno José Torres, George D C Cavalcanti, and Tsang Ing Ren. 2013. "Lateral Inhibition Pyramidal Neural Network for Image Classification." *IEEE Transactions on Cybernetics* 43 (6): 2082–92. doi:10.1109/TCYB.2013.2240295.

Flach, P. A., & Lachiche, N. 2004. "Naive Bayesian Classification of Structured Data." *Machine Learning*. 57 (3): 233–269.

Fu, Tak Chung. 2011. "A Review on Time Series Data Mining." *Engineering Applications of Artificial Intelligence* 24 (1). Elsevier: 164–81. doi:10.1016/j.engappai.2010.09.007.

Gamage, Nuwan, Rini Akmeliawati, and Kuang Ye Chow. 2009. "Towards Robust Skin Colour Detection and Tracking." In *In Instrumentation and Measurement Technology Conference, I2MTC IEEE*, 895–99. doi:10.1109/IMTC.2009.5168576.

García-Mateos, G., J. L. Hernández-Hernández, D. Escarabajal-Henarejos, S. Jaén-Terrones, and J. M. Molina-Martínez. 2015. "Study and Comparison of Color Models for Automatic Image Analysis in Irrigation Management Applications." *Agricultural Water Management* 151 (March). Elsevier B.V.: 158–66. doi:10.1016/j.agwat.2014.08.010.

Gasparini, Francesca, Silvia Corchs, and Raimondo Schettini. 2008. "Recall or Precision-Oriented Strategies for Binary Classification of Skin Pixels." *Journal of Electronic Imaging* 17 (2): 23017. doi:10.1117/1.2916715.

Gayatri, V., & Chetan, M. 2013. "Comparative Study of Different Multi-Criteria Decision-Making Methods." *International Journal of Advanced Computer Theory and Engineering (IJACTE)*. 2: 9–12. http://www.irdindia.in/journal_ijacte/pdf/vol2_iss4/2.pdf.

Gede, M., & Mészáros, J. 2013. "Digital Archiving and On-Line Publishing of Old

Relief Models." *The Cartographic Journal,* 50 (3): 293–99. doi:10.1179/1743277413Y.0000000064.

Ghaziasgar, Mehrdad, James Connan, and Antoine B. Bagula. 2016. "Enhanced Adaptive Skin Detection with Contextual Tracking Feedback." In *In Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech), IEEE.*, 1–6. doi:10.1109/RoboMech.2016.7813194.

Gijsberts, Arjan, Manfredo Atzori, Claudio Castellini, Henning Müller, and Barbara Caputo. 2014. "Movement Error Rate for Evaluation of Machine Learning Methods for sEMG-Based Hand Movement Classification." *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 22 (4): 735–44. doi:10.1109/TNSRE.2014.2303394.

Graczyk, Magdalena. 2009. "Computational Collective Intelligence. Semantic Web, Social Networks and Multiagent Systems." In *In International Conference on Computational Collective Intelligence . Springer, Berlin, Heidelberg.*, 800–812. doi:10.1007/978-3-642-04441-0.

Grigorova, Anelia, Francesco G.B. De Natale, Charlie Dagli, and Thomas S. Huang. 2007. "Content-Based Image Retrieval by Feature Adaptation and Relevance Feedback." *IEEE Transactions on Multimedia* 9 (6): 1183–92. doi:10.1109/TMM.2007.902828.

Gupta, A., & Chaudhary, A. 2016. "Robust Skin Segmentation Using Color Space Switching." *Pattern Recognition and Image Analysis* 26 (1): 61–68.

Gurari, Danna, Diane Theriault, Mehrnoosh Sameki, Brett Isenberg, Tuan A. Pham, Alberto Purwada, Patricia Solski, et al. 2015. "How to Collect Segmentations for Biomedical Images? A Benchmark Evaluating the Performance of Experts, Crowdsourced Non-Experts, and Algorithms." In *Proceedings - 2015 IEEE Winter Conference on Applications of Computer Vision, WACV 2015*, 1169–76. doi:10.1109/WACV.2015.160.

Hall, G. 2015. Pearson ' s correlation coefficient, 1 1–4.

Han, J., Pei, J., & Kamber, M. 2011. *Data Mining: Concepts and Techniques*.

Han, Hu, Shiguang Shan, Xilin Chen, and Wen Gao. 2013. "A Comparative Study on Illumination Preprocessing in Face Recognition." *Pattern Recognition* 46 (6). Elsevier: 1691–99. doi:10.1016/j.patcog.2012.11.022.

Hedberg, E. C., and Stephanie Ayers. 2015. "The Power of a Paired T-Test with a Covariate." *Social Science Research* 50. Elsevier Inc.: 277–91. doi:10.1016/j.ssresearch.2014.12.004.

Herath, Gamini, Anthony Prato, and Tony Prato. 2007. "Using Multi-Criteria Decision Analysis in Natural Resource Management." *Environmental Sciences* 4

(2): 3–5. doi:10.1080/15693430701317637.

Hollender, Nina, Cristian Hofmann, Michael Deneke, and Bernhard Schmitz. 2010. "Integrating Cognitive Load Theory and Concepts of Human-Computer Interaction." *Computers in Human Behavior* 26 (6). Elsevier Ltd: 1278–88. doi:10.1016/j.chb.2010.05.031.

Hoshyar, Azadeh Noori, Adel Al-Jumaily, and Afsaneh Noori Hoshyar. 2014. "The Beneficial Techniques in Preprocessing Step of Skin Cancer Detection System Comparing." *Procedia Computer Science* 42. Elsevier Masson SAS: 25–31. doi:10.1016/j.procs.2014.11.029.

Hossain, Foisal, Mousa Shamsi, Mohammad Reza Alsharif, Reza A Zoroofi, and Katsumi Yamashita. 2012. "Automatic Facial Skin Detection Using Gaussian Mixture Model Under Varying Illumination." *International Journal of Innovative Computing, Information and Control* 8 (2): 1135–1144.

Hsieh, Jun Wei, and Yung Tai Hsu. 2008. "Boosted String Representation and Its Application to Video Surveillance." *Pattern Recognition* 41 (10): 3078–91. doi:10.1016/j.patcog.2008.03.026.

Hu, Weiming, Ou Wu, Zhouyao Chen, Zhouyu Fu, and Steve Maybank. 2007. "Recognition of Pornographic Web Pages by Classifying Texts and Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29 (6): 1019–34. doi:10.1109/TPAMI.2007.1133.

Huang, Lei, Wen Ji, Zhiqiang Wei, Bo Wei Chen, Chenggang Clarence Yan, Jie Nie, Jian Yin, and Baochen Jiang. 2015. "Robust Skin Detection in Real-World Images." *Journal of Visual Communication and Image Representation* 29. Elsevier Inc.: 147–52. doi:10.1016/j.jvcir.2015.02.004.

Huang, Yeu Shiang, Wei Chen Chang, Wei Hao Li, and Zu Liang Lin. 2013. "Aggregation of Utility-Based Individual Preferences for Group Decision-Making." *European Journal of Operational Research* 229 (2). Elsevier B.V.: 462–69. doi:10.1016/j.ejor.2013.02.043.

Hussain, I., Talukdar, A. K., & Sarma, K. K. 2014. "Hand Gesture Recognition System with Real-Time Palm Tracking." In *In India Conference (INDICON)-IEEE*, 1–6.

Jadhav, A. S., & Sonar, R. M. 2009. "Evaluating and Selecting Software Packages: A Review." *Information and Software Technology*, 51 (3): 555–63. doi:10.1109/ICETET.2009.33.

Jadhav, Shivajirao M, Sanjay L Nalbalwar, and Ashok A Ghatol. 2011. "Artificial Neural Network Based Cardiac Arrhythmia Disease Diagnosis." In *2011 International Conference on Process Automation, Control and Computing*, 1–6. doi:10.1109/PACC.2011.5979000.

Jaiswal, S. 2011. "Frontal Face Detection Methods–Neural Networks and Aggressive Learning Algorithm." *International Journal of Global Research in Computer Science (UGC Approved Journal)* 2 (7): 119–48.

Jensch, Dennis, Daniel Mohr, and Gabriel Zachmann. 2012. "A Comparative Evaluation of Three Skin Color Detection Approaches." *Journal of Virtual Reality and Broadcasting* 12 (1): 1–14.

Jones, M, and J Rehg. 1999. "Statistical Color Models with Application to Skin Detection 2 Histogram Color Models." *Computer Vision and Pattern Recognition* 46 (1): 1–23. doi:10.1023/A:1013200319198.

Jovic, A., Brkic, K., & Bogunovic, N. 2014. "An Overview of Free Software Tools for General Data Mining." In *In Information and Communication Technology, Electronics and Microelectronics (MIPRO),IEEE 2014 37th International Convention*, 1112–17.

Jumaah, F. M., Zaidan, A. A., Zaidan, B. B., Bahbibi, R., Qahtan, M. Y., & Sali, A. 2017. "Technique for Order Performance by Similarity to Ideal Solution for Solving Complex Situations in Multi-Criteria Optimization of the Tracking Channels of GPS Baseband Telecommunication Receivers." *Telecommunication Systems*, 1–19.

Kakumanu, P., S. Makrogiannis, and N. Bourbakis. 2007. "A Survey of Skin-Color Modeling and Detection Methods." *Pattern Recognition* 40 (3): 1106–22. doi:10.1016/j.patcog.2006.06.010.

Kasson, James M., and Wil Plouffe. 1992. "An Analysis of Selected Computer Interchange Color Spaces." *ACM Transactions on Graphics* 11 (4): 373–405. doi:10.1145/146443.146479.

Kawulok, M., & Nalepa, J. 2014. "Hand Pose Estimation Using Support Vector Machines with Evolutionary Training." In *In Systems, Signals and Image Processing (IWSSIP), 2014 International Conference On. IEEE.*, 87–90. doi:10.1007/s11042-014-2204-y.

Kawulok, Michal. 2013. "Fast Propagation-Based Skin Regions Segmentation in Color Images." In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013*, 1–7. doi:10.1109/FG.2013.6553733.

Kawulok, Michal, Jolanta Kawulok, and Jakub Nalepa. 2014. "Spatial-Based Skin Detection Using Discriminative Skin-Presence Features." *Pattern Recognition Letters* 41 (1). Elsevier B.V.: 3–13. doi:10.1016/j.patrec.2013.08.028.

Kawulok, Michal, Jolanta Kawulok, Jakub Nalepa, and Bogdan Smolka. 2014. "Self-Adaptive Algorithm for Segmenting Skin Regions." *Eurasip Journal on Advances in Signal Processing* 2014 (1): 1–22. doi:10.1186/1687-6180-2014-170.

Kersten, Thomas P., and Maren Lindstaedt. 2012. "Image-Based Low-Cost Systems for Automatic 3D Recording and Modelling of Archaeological Finds and Objects." In *In Euro-Mediterranean Conference . Springer, Berlin, Heidelberg.*, 1–10. doi:10.1007/978-3-642-34234-9_1.

Khan, R., Hanbury, A., Stöttinger, J., Khan, F. A., Khattak, A. U., & Ali, A. 2014. "Multiple Color Space Channel Fusion for Skin Detection." *Multimedia Tools and Applications* 72 (2): 1709–30. doi:10.1007/s11042-013-1443-7.

Khan, Rehanullah, et al. 2012. "COLOR BASED SKIN CLASSIFICATION." *Pattern Recognition Letters-Elsevier* 33 (2): 157–63.

Khan, Rehanullah, Allan Hanbury, Robert Sablatnig, Julian Stöttinger, F. Ali Khan, and F. Alam Khan. 2014. "Systematic Skin Segmentation: Merging Spatial and Non-Spatial Data." *Multimedia Tools and Applications* 69 (3): 717–41. doi:10.1007/s11042-012-1124-y.

Khan, Rehanullah, Allan Hanbury, and Julian Stoettinger. 2010. "SKIN DETECTION : A RANDOM FOREST APPROACH IRF , Vienna , Austria." In *In Image Processing (ICIP), 2010 17th IEEE International Conference on IEEE.*, 4613–16. doi:10.1109/ICIP.2010.5651638.

Khan, Rehanullah, Allan Hanbury, Julian Stöttinger, and Abdul Bais. 2012. "Color Based Skin Classification." *Pattern Recognition Letters* 33 (2). Elsevier B.V.: 157–63. doi:10.1016/j.patrec.2011.09.032.

Kim, J. H., & Lattimer, B. Y. 2015. "Real-Time Probabilistic Classification of Fire and Smoke Using Thermal Imagery for Intelligent Firefighting Robot." *Fire Safety Journal* 72: 40–49.

Köhler, Rolf, Michael Hirsch, Betty Mohler, Bernhard Schölkopf, and Stefan Harmeling. 2012. "Recording and Playback of Camera Shake: Benchmarking Blind Deconvolution with a Real-World Database." In *In European Conference on Computer Vision -Springer*, 7578:27–40. doi:10.1007/978-3-642-33786-4_3.

Kokkinos, Iasonas. 2010. "Boundary Detection Using F-Measure- , Filter- and." In *In European Conference on Computer Vision . Springer, Berlin, Heidelberg.*, 650–63.

Korotkov, Konstantin, and Rafael Garcia. 2012. "Computerized Analysis of Pigmented Skin Lesions: A Review." *Artificial Intelligence in Medicine* 56 (2). Elsevier B.V.: 69–90. doi:10.1016/j.artmed.2012.08.002.

Kosorus, Hilda, Jurgen Honigl, and Josef Kung. 2011. "Using R, WEKA and RapidMiner in Time Series Analysis of Sensor Data for Structural Health Monitoring." *2011 22nd International Workshop on Database and Expert Systems Applications*. doi:10.1109/DEXA.2011.88.

Koutsoudis, Anestis, Blaž Vidmar, and Fotis Arnaoutoglou. 2013. "Performance

Evaluation of a Multi-Image 3D Reconstruction Software on a Low-Feature Artefact." *Journal of Archaeological Science* 40 (12): 4450–56. doi:10.1016/j.jas.2013.07.007.

Kozielski, M, M Sikora, and Ł Wróbel. 2015. "DISESOR - Decision Support System for Mining Industry." *2015 Federated Conference on Computer Science and Information Systems (FedCSIS)* 5: 67–74. doi:10.15439/2015F168.

Kruppa, Hannes, Martin A. Bauer, and Bernt Schiele. 2002. "Skin Patch Detection in Real-World Images." *In Joint Pattern Recognition Symposium , Springer, Berlin, Heidelberg.* 2449: 109–16. doi:10.1007/3-540-45783-6_14.

Kukolja, Davor, Siniša Popović, Marko Horvat, Bernard Kovač, and Krešimir Ćosić. 2014. "Comparative Analysis of Emotion Estimation Methods Based on Physiological Measurements for Real-Time Applications." *International Journal of Human Computer Studies* 72 (10–11): 717–27. doi:10.1016/j.ijhcs.2014.05.006.

Kuncheva, Ludmila I. 2006. "On the Optimality of Naïve Bayes with Dependent Binary Features." *Pattern Recognition Letters* 27 (7): 830–37. doi:10.1016/j.patrec.2005.12.001.

Kwolek, B. 2003. "Face Tracking System Based on Color, Stereovision and Elliptical Shape Features." In *Advanced Video and Signal Based Surveillance, 2003. Proceedings. IEEE Conference on*, 21–26. doi:10.1109/AVSS.2003.1217897.

Lam, K., & Zhao, X. 1998. "An Application of Quality Function Deployment to Improve the Quality of Teaching." *International Journal of Quality & Reliability Management.* 15 (4): 389–413.

Lan, Zhangli, Danmei Wang, Fangfang Bao, and Minglan Sheng. 2013. "An Algorithm for Face Determination Based on Convolution and Average Face." *In Image and Signal Processing (CISP), 2013 6th International Congress on IEEE.* 2: 893–98.

Land, Sebastian, and Simon Fischer. 2012. *RapidMiner 5: RapidMiner in Academic Use*.

Lee, Jiann-Shu, Yung-Ming Kuo, and and Pau-Choo Chung. 2010. "Detecting Nakedness in Color Images." *InIntelligent Multimedia Analysis for Security Applications-Springer* 282: 225–36. doi:10.1007/978-3-642-11756-5_10.

Li, Cheng, and Kris M. Kitani. 2013. "Pixel-Level Hand Detection in Ego-Centric Videos." In *In Computer Vision and Pattern Recognition (Cvpr), IEEE Conference*, 3570–77. doi:10.1109/CVPR.2013.458.

Li, Gen, Jae Kyu Suhr, Dongik Kim, Ho Gi Jung, and Jaihie Kim. 2010. "Minimizing False Detection of Skin Color by Using Background Subtraction." In *International Conference on Electronics, Informations and Communications,*

514–16.

Li, Guojun, and Lingsheng Shi. 2013. "Edge-Disjoint Spanning Trees and Eigenvalues of Graphs." *Linear Algebra and Its Applications* 439 (10). Elsevier Inc.: 2784–89. doi:10.1016/j.laa.2013.08.041.

Liew, Chun Fui, and Takehisa Yairi. 2014. "Generalized BRIEF: A Novel Fast Feature Extraction Method for Robust Hand Detection." In *Proceedings - International Conference on Pattern Recognition*, 3014–19. doi:10.1109/ICPR.2014.520.

Linderoth, Magnus, Anders Robertsson, and Rolf Johansson. 2013. "Color-Based Detection Robust to Varying Illumination Spectrum." *2013 IEEE Workshop on Robot Vision, WORV 2013*. doi:10.1109/WORV.2013.6521924.

Liu, Jun, Y Liu, Y Cui, and YQ Chen. 2013. "Real-Time Human Detection and Tracking in Complex Environments Using Single RGBD Camera." In *In Image Processing (ICIP), 2013 20th IEEE International Conference on IEEE*, 3088–92. doi:10.1109/ICIP.2013.6738636.

Liu, Qiong, and S M a X Min. 2010. "A Robust Skin Color Based Face Detection Algorithm." In *International Asia Conference on Informatics in Control, Automation and Robotics*, 1–4. doi:10.1109/CAR.2010.5456614.

Lu, Cheng, and Mrinal Mandal. 2015. "Automated Analysis and Diagnosis of Skin Melanoma on Whole Slide Histopathological Images." *Pattern Recognition* 48 (8). Elsevier: 2738–50. doi:10.1016/j.patcog.2015.02.023.

Luo, Yong, and Ye-Peng Guan. 2017. "Adaptive Skin Detection Using Face Location and Facial Structure Estimation." *IET Computer Vision* 11 (7): 550–59. doi:10.1049/iet-cvi.2016.0295.

Ma, Zhanyu, and Arne Leijon. 2010. "Human Skin Color Detection in Rgb Space With Bayesian Estimation of Beta Mixture Models." In *In Signal Processing Conference, 2010 18th European*, 1204–8.

Madeo, R. C., Lima, C. A., & Peres, S. M. 2017. "Studies in Automated Hand Gesture Analysis: An Overview of Functional Types and Gesture Phases." *Language Resources and Evaluation* 51 (2): 547–579. doi:10.1201/b10328.

Mahdieh, M. H., and M. Pournoury. 2010. "Atmospheric Turbulence and Numerical Evaluation of Bit Error Rate (BER) in Free-Space Communication." *Optics and Laser Technology* 42 (1). Elsevier: 55–60. doi:10.1016/j.optlastec.2009.04.017.

Mahmoodi, M. R., & Sayedi, S. M. 2014. "Boosting Performance of Face Detection by Using an Efficient Skin Segmentation Algorithm." In *In Information Technology and Electrical Engineering (ICITEE), 2014 6th International Conference on IEEE.*, 1–6.

Mahmoodi, Mohammad Reza, and Sayed Masoud Sayedi. 2014. "A Face Detector

Based on Color and Texture." In *2014 6th International Conference on Information Technology and Electrical Engineering (ICITEE)*, 1–6. doi:10.1109/ICITEED.2014.7007952.

Mahmoodi, M. R., & Sayedi, S. M. 2015. "A Face Detection Method Based on Kernel Probability Map." *Computers and Electrical Engineering* 46. Elsevier Ltd: 205–16. doi:10.1016/j.compeleceng.2015.02.005.

Mahmoodi, M. R., & Sayedi, S. M. 2016. *A Comprehensive Survey on Human Skin Detection. International Journal of Image, Graphics and Signal Processing*. Vol. 8. doi:10.5815/ijigsp.2016.05.01.

Manigandan, M, and I M Jackin. 2010. "Wireless Vision Based Mobile Robot Control Using Hand Gesture Recognition through Perceptual Color Space." In *2010 International Conference on Advances in Computer Engineering, ACE 2010*, 95–99. doi:10.1109/ACE.2010.69.

Medjahed, S. A. 2015. "A Comparative Study of Feature Extraction Methods in Images Classification." *International Journal of Image, Graphics and Signal Processing* 7 (3): 16–23. doi:10.5815/ijigsp.2015.03.03.

Mery, D., Pedreschi, F., & Soto, A. 2013. "Automated Design of a Computer Vision System for Visual Food Quality Evaluation." *Food and Bioprocess Technology* 6 (8): 2093–2108. doi:10.1260/2047-4970.1.0.1.

Metz, Charles E. 1978. "Basic Principles of ROC Analysis." *Semin Nucl Med* 8 (4): 283–98. doi:10.1016/S0001-2998(78)80014-2.

Mircea, Ioan-gabriel. 2012. "An Evaluation of Color Spaces Used in Skin Color Detection." *Studia Universitatis Babes-Bolyai, Informatica* 57 (3): 24–34.

Molina, J., Escudero-Viñolo, M., Signoriello, A., Pardàs, M., Ferrán, C., Bescós, J., ... & Martínez, J. M. 2013. "Real-Time User Independent Hand Gesture Recognition from Time-of-Flight Camera Video Using Static and Dynamic Models." *Machine Vision and Applications* 24 (1): 187–204. doi:10.1109/ECAI.2013.6636188.

Myllymaki, P., & Tirri, H. 1993. "Bayesian Case-Based Reasoning with Neural Networks Dapt T Case-Based." *Myllymaki, Petri, and Henry Tirri. "Bayesian Case-Based Reasoning with Neural networks." In Neural Networks, 1993., IEEE International Conference*.

Naji, Sinan A., Roziati Zainuddin, and Hamid A. Jalab. 2012. "Skin Segmentation Based on Multi Pixel Color Clustering Models." *Digital Signal Processing: A Review Journal* 22 (6). Elsevier Inc.: 933–40. doi:10.1016/j.dsp.2012.05.004.

Nedher, A. S., Hassan, S., & Katuk, N. 2014. "On Multi Attribute Decision Making Methods: Prioritizing Information Security Controls." *Security Controls. Journal of Applied Sciences* 14 (16): 1865–1870.

Nilsson, Nils J. 1996. *Introduction to Machine Learning*. doi:10.1016/j.neuroimage.2010.11.004.

Och, F. J. 2003. "Minimum Error Rate Training in Statistical Machine Translation." *In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics* 1: 160–67.

Ochoa, V. M. T., S. Y. Yayilgan, and F. A. Cheikh. 2012. "Adult Video Content Detection Using Machine Learning Techniques." In *2012 Eighth International Conference on Signal Image Technology and Internet Based Systems*, 967–74. doi:10.1109/SITIS.2012.143.

Oxford dictionaries. (2013). "Definition of benchmark in English." Retrieved 10-Oct., 2013, from oxforddictionaries.com. 2013 10–Oct.

Patil, C. G., Kolte, M. T., Chatur, P. N., & Chaudhari, D. S. 2014. "Performance Evaluation of CBIR System Based on Object Detection and Evolutionary Computation." In *IEEE Global Conference on Wireless Computing and Networking (GCWCN)*, 65–69.

Patravali, Shruti D, J M Wayakule, and Apurva D Katre. 2014. "Skin Segmentation Using YCBCR and RGB Color Models." *International Journal*. 4 (7): 341–46.

Paul, Padma Polash, and Marina Gavrilova. 2011. "PCA Based Geometric Modeling for Automatic Face Detection." In *2011 International Conference on Computational Science and Its Applications*, 33–38. doi:10.1109/ICCSA.2011.69.

Pham, Thang V., Marcel Worring, and Arnold W.M. Smeulders. 2002. "Face Detection by Aggregated Bayesian Network Classifiers." *Pattern Recognition Letters* 23 (4): 451–61. doi:10.1016/S0167-8655(01)00177-5.

Phung, S.L., A. Bouzerdoum, and Sr. Chai, D. 2005. "Skin Segmentation Using Color Pixel Classification: Analysis and Comparison." In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:148–54. doi:10.1109/TPAMI.2005.17.

Powers, David M W. 2013. "Advances in Brain Inspired Cognitive Systems." *In International Conference on Brain Inspired Cognitive Systems , Springer, Berlin, Heidelberg.* 7366: 145–56. doi:10.1007/978-3-642-31561-9.

Priya, U., S. Vasuhi, and V. Vaidehi. 2015. "Face Detection Using CbCr Color Model in Video." In *2015 3rd International Conference on Signal Processing, Communication and Networking, ICSCN 2015*, 1–5. doi:10.1109/ICSCN.2015.7219912.

Pujol, Francisco A., and Juan Carlos García. 2012. "Computing the Principal Local Binary Patterns for Face Recognition Using Data Mining Tools." *Expert Systems with Applications* 39 (8). Elsevier Ltd: 7165–72.

doi:10.1016/j.eswa.2012.01.074.

Ramachandra, Raghavendra, and Christoph Busch. 2017. "Presentation Attack Detection Methods for Face Recognition Systems." *ACM Computing Surveys* 50 (1): 1–37. doi:10.1145/3038924.

Rathore, S., Iftikhar, M. A., Hussain, M., & Jalil, A. 2013. "Classification of Colon Biopsy Images Based on Novel Structural Features." *In Emerging Technologies (ICET), 2013 IEEE 9th International Conference*.

Rautaray, Siddharth S., and Anupam Agrawal. 2015. "Vision Based Hand Gesture Recognition for Human Computer Interaction: A Survey." *Artificial Intelligence Review* 43 (1): 1–54. doi:10.1007/s10462-012-9356-9.

Rehanullah Khan , Allan Hanbury , Julian Stöttinger, Abdul Bais. 2012. "COLOR BASED SKIN CLASSIFICATION." *Pattern Recognition Letters*, 33 (2): 157–63.

Ren, Jiangtao, Sau Dan Lee, Xianlu Chen, Ben Kao, Reynold Cheng, and David Cheung. 2009. "Naive Bayes Classification of Uncertain Data." In *Proceedings - IEEE International Conference on Data Mining, ICDM*, 944–49. doi:10.1109/ICDM.2009.90.

Riess, Thorsten, Christian Dietz, Martin Tomas, Elisa Ferrando-May, and Dorit Merhof. 2011. "Automated Image Processing for the Analysis of DNA Repair Dynamics." *arXiv Preprint arXiv*, 1101.3391. http://arxiv.org/abs/1101.3391v1.

Rish, Irina, Joseph Hellerstein, and T Jayram. 2001. "An Analysis of Data Characteristics That Affect Naive Bayes Performance." *IBM TJ Watson Research Center* 30: 1–8. http://www.cs.iastate.edu/~honavar/rish-bayes.pdf.

Rodgers, Joseph Lee, and W Alan Nicewander. 1988. "Thirteen Ways to Look at the Correlation Coefficient." *The American Statistician* 42 (1): 59–66. http://www.tandfonline.com/doi/abs/10.1080/00031305.1988.10475524.

Rushing, Christel, Anuradha Bulusu, Herbert I. Hurwitz, Andrew B. Nixon, and Herbert Pang. 2015. "A Leave-One-out Cross-Validation SAS Macro for the Identification of Markers Associated with Survival." *Computers in Biology and Medicine* 57. Elsevier: 123–29. doi:10.1016/j.compbiomed.2014.11.015.

Saaty, T.L., and M.S. Ozdemir. 2003. "Why the Magic Number Seven plus or Minus Two." *Mathematical and Computer Modelling* 38 (3–4): 233–44. doi:10.1016/S0895-7177(03)90083-5.

Saaty, T L. 1990. "How to Make a Decision: The Analytic Hierarchy Process." *European Journal of Operational Research* 48: 9–26.

Saaty, Thomas L. 1977. "A Scaling Method for Priorities in Hierarchical Structures." *Journal of Mathematical Psychology* 15 (3): 234–81. doi:10.1016/0022-2496(77)90033-5.

Saaty, T. L. 2008. "Decision Making with the Analytic Hierarchy Process." *International Journal of Services Sciences.* 1 (1): 83–98. doi:10.1504/IJSSCI.2008.017590.

Saaty, Thomas L., and Luis G. Vargas. 1984. "Inconsistency and Rank Preservation." *Journal of Mathematical Psychology* 28 (2): 205–14. doi:10.1016/0022-2496(84)90027-0.

Sajedi, Hedieh, and Mansour Jamzad. 2007. "A Contourlet-Based Face Detection Method in Color Images." In *2007 Third International IEEE Conference on Signal-Image Technologies and Internet-Based System*, 727–32. doi:10.1109/SITIS.2007.53.

San Cristóbal, J. R. 2011. "Multi-Criteria Decision-Making in the Selection of a Renewable Energy Project in Spain: The Vikor Method." *Renewable Energy* 36 (2). Elsevier Ltd: 498–502. doi:10.1016/j.renene.2010.07.031.

Sanmiguel, Juan C., and Sergio Suja. 2013. "Skin Detection by Dual Maximization of Detectors Agreement for Video Monitoring." *Pattern Recognition Letters* 34 (16): 2102–9. doi:10.1016/j.patrec.2013.07.016.

Santagati, C., & Inzerillo, L. 2013. "123D Catch: Efficiency, Accuracy, Constraints and Limitations in Architectural Heritage Field." *International Journal of Heritage in the Digital Era* 2 (2): 263–289. doi:10.4324/9780203079959.

Schmugge, Stephen J., M. Adeel Zaffar, Leonid V. Tsap, and Min C. Shin. 2007. "Task-Based Evaluation of Skin Detection for Communication and Perceptual Interfaces." *Journal of Visual Communication and Image Representation* 18 (6): 487–95. doi:10.1016/j.jvcir.2007.04.008.

Sebe, N, I Cohen, T S Huang, and T Gevers. 2004. "Skin Detection: A Bayesian Network Approach." In *Proceedings of the 17th International Conference on Pattern Recognition 2004 ICPR 2004*, 2–5. doi:10.1109/ICPR.2004.1334405.

Shah, S. A. H., Ahmed, A., Mahmood, I., & Khurshid, K. 2011. "Hand Gesture Based User Interface for Computer Using a Camera and Projector." In *In Signal and Image Processing Applications (ICSIPA), 2011 IEEE International Conference on (Pp. 168-173). IEEE.*, 168–73. doi:10.1145/2184319.2184337.

Sharef, Baraa, Nazlia Omar, and Zeyad Sharef. 2014. "An Automated Arabic Text Categorization Based on the Frequency Ratio Accumulation." *Int. Arab J. Inf. Technol.* 11 (2): 213–21.

Shin, M. C., K. I. Chang, and L. V. Tsap. 2002. "Does Colorspace Transformation Make Any Difference on Skin Detection?" *Proceedings of IEEE Workshop on Applications of Computer Vision*. doi:10.1109/ACV.2002.1182194.

Shirali-Shahreza, Sajad, and M. E. Mousavi. 2008. "A New Bayesian Classifier for Skin Detection." In *3rd International Conference on Innovative Computing*

Information and Control, ICICIC'08, 172–172. doi:10.1109/ICICIC.2008.54.

Shoyaib, Mohammad, M. Abdullah-Al-Wadud, and Oksam Chae. 2012. "A Skin Detection Approach Based on the Dempster-Shafer Theory of Evidence." *International Journal of Approximate Reasoning* 53 (4). Elsevier Inc.: 636–59. doi:10.1016/j.ijar.2012.01.003.

Shruthi, M. L. J., & Harsha, B. K. 2013. "Non-Parametric Histogram Based Skin Modeling for Skin Detection." In *IEEE International Conference- In Computational Intelligence and Computing Research, (ICCIC)*, 1–6. doi:10.1109/ICCIC.2013.6724287.

Sigal, L., S. Sclaroff, and V. Athitsos. 2000. "Estimation and Prediction of Evolving Color Distributions for Skin\nsegmentation under Varying Illumination." *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662)* 2 (March). doi:10.1109/CVPR.2000.854764.

Singh Sisodia, Dilip, and Shrish Verma. 2011. "Image Pixel Intensity and Artificial Neural Network Based Method for Pattern Recognition." *World Academy of Science Engineering and Technology* 57: 742–45.

Siqueira, FR de, WR Schwartz, and H Pedrini. 2013. "Adaptive Detection of Human Skin in Color Images." *In IX Workshop de Visão Computacional (WVC), Rio de Janeiro-RJ, Brazil* . https://homepages.dcc.ufmg.br/~william/papers/paper_2013_WVC_skin.pdf.

Sokolova, Marina, and Guy Lapalme. 2009. "A Systematic Analysis of Performance Measures for Classification Tasks." *Information Processing and Management* 45 (4). Elsevier Ltd: 427–37. doi:10.1016/j.ipm.2009.03.002.

Song, Wei, Dong Wu, Yulong Xi, Yong Woon Park, and Kyungeun Cho. 2017. "Motion-Based Skin Region of Interest Detection with a Real-Time Connected Component Labeling Algorithm." *Multimedia Tools and Applications* 76 (9): 11199–214. doi:10.1007/s11042-015-3201-5.

Soran, B., Hwang, J. N., Lee, S. I., & Shapiro, L. 2012. "Tremor Detection Using Motion Filtering and SVM BilgeSVM." In *In Pattern Recognition (ICPR), 2012 21st International Conference on IEEE.*, 178–81. doi:10.1109/FG.2013.6553707.

Stejić, Zoran, Yasufumi Takama, and Kaoru Hirota. 2006. "Variants of Evolutionary Learning for Interactive Image Retrieval." *Soft Computing* 11 (7): 669–78. doi:10.1007/s00500-006-0129-8.

Stergiopoulou, Ekaterini, Kyriakos Sgouropoulos, Nikos Nikolaou, Nikos Papamarkos, and Nikos Mitianoudis. 2014. "Real Time Hand Detection in a Complex Background." *Engineering Applications of Artificial Intelligence* 35 (July). Elsevier: 54–70. doi:10.1016/j.engappai.2014.06.006.

Sun, Hung Ming. 2010. "Skin Detection for Single Images Using Dynamic Skin

Color Modeling." *Pattern Recognition* 43 (4). Elsevier: 1413–20. doi:10.1016/j.patcog.2009.09.022.

Sun, Li, Liu Jia, Lv Caixiai, and Li Zheying. 2013. *Proceedings of the 2012 International Conference on Information Technology and Software Engineering*. Vol. 212. doi:10.1007/978-3-642-34531-9.

Szkudlarek, Michal, and Maria Pietruszka. 2015. "Fast Grid-Based Clustering Method for Automatic Calculation of Optimal Parameters of Skin Color Classifier for Head Tracking." In *Proceedings - 2015 IEEE 2nd International Conference on Cybernetics, CYBCONF 2015*, 119–24. doi:10.1109/CYBConf.2015.7175917.

Taheri, Sona, and Musa Mammadov. 2013. "Learning the Naive Bayes Classifier with Optimization Models." *International Journal of Applied Mathematics and Computer Science* 23 (4): 787–95. doi:10.2478/amcs-2013-0059.

Tan, W. R., Chan, C. S., Yogarajah, P., & Condell, J. 2012. "A Fusion Approach for Efficient Human Skin Detection." *IEEE Transactions on Industrial Informatics* 8 (1): 138–47.

Taqa, A. Y., & Jalab, H. A. 2010. "Increasing the Reliability of Fuzzy Inference System-Based Skin Detector." *American Journal of Applied Sciences* 7 (8): 1129.

Taqa, Alaa Y, and Hamid a Jalab. 2010. "Increasing the Reliability of Skin Detectors." *Scientific Research and Essays* 5 (17): 2480–90.

Ting, S. L., W. H. Ip, and Albert H.C. Tsang. 2011. "Is Naïve Bayes a Good Classifier for Document Classification?" *International Journal of Software Engineering and Its Applications* 5 (3): 37–46.

Titterington, Xue and. 2008. "Comment on 'On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes.'" *Neural Processing Letters.* 28 (3): 1–33.

Trentesaux, Damien, Cyrille Pach, Abdelghani Bekrar, Yves Sallez, Thierry Berger, Thérèse Bonte, Paulo Leitão, and José Barbosa. 2013. "Benchmarking Flexible Job-Shop Scheduling and Control Systems." *Control Engineering Practice* 21 (9). Elsevier: 1204–25. doi:10.1016/j.conengprac.2013.05.004.

Triantaphyllou, E., Shu, B., Sanchez, S. N., & Ray, T. 1998. "Multi-Criteria Decision Making: An Operations Research Approach." *Encyclopedia of Electrical and Electronics Engineering.* 15: 175–86. http://univ.nazemi.ir/mcdm/Multi-Criteria Decision Making.pdf.

Trigueiros, P, F Ribeiro, and L P Reis. 2013. "A Comparative Study of Different Image Features for Hand Gesture Machine Learning." *Proceedings of the 5th International Conference on Agents and Artificial Intelligence. ICAART 2013*, 51–61. doi:10.1007/978-3-662-44440-5_10.

Tsai, Guo-Shiang Lin and Tung-Sheng. 2012. "A Face Tracking Method Using Feature Point Tracking." In *2012 International Conference on Information Security and Intelligent Control*, 210–13. doi:10.1109/ISIC.2012.6449743.

Tsitsoulis, A., & Bourbakis, N. 2013. "Towards Automatic Hands Detection in Single Images." In *In International Conference on Image Analysis and Processing , Springer, Berlin, Heidelberg.*, 469–78. doi:10.1007/s11042-013-1703-6.

ur Rehman, Z., Hussain, O. K., & Hussain, F. K. 2012. "Iaas Cloud Selection Using MCDM Methods." In *In E-Business Engineering (ICEBE), 2012 IEEE Ninth International Conference on IEEE.*, 246–51.

Vamsi Krishna, M., Rajendra N. Dash, P. Venugopal, B. Jalachandra Reddy, P. Sandeep, and G. Madhavi. 2017. "Development of a RP-HPLC Method for Evaluation of in Vitro Permeability of Voriconazole in the Presence of Enhancers through Rat Skin." *Journal of Saudi Chemical Society* 21 (1). King Saud University: 1–10. doi:10.1016/j.jscs.2013.07.003.

Venetianer, Péter L., and Hongli Deng. 2010. "Performance Evaluation of an Intelligent Video Surveillance System - A Case Study." *Computer Vision and Image Understanding* 114 (11): 1292–1302. doi:10.1016/j.cviu.2010.07.010.

Verhoeven, G., Sevara, C., Karel, W., Ressl, C., Doneus, M., & Briese, C. 2012. "Undistorting the Past: New Techniques for Orthorectification of Archaeological Aerial Frame Imagery." *In Good Practice in Archaeological Diagnostics . Springer, Cham.* 1 (6): 31–67. doi:10.1109 / ICDSE.2016.7823946.

Verhoeven, G. 2011. "Taking Computer Vision Aloft–archaeological Three-dimensional Reconstructions from Aerial Photographs with Photoscan." *Archaeological Prospection* 18 (1): 67–73.

Vezhnevets, Vladimir, and Anna Degtiareva. 2003. "Robust and Accurate Eye Contour Extraction." In *Proceding of the Conference {GraphiCon}*, 81–84.

Vezhnevets, Vladimir, Vassili Sazonov, and Alla Andreeva. 2003. "A Survey on Pixel-Based Skin Color Detection Techniques." *Proceedings of GraphiCon 2003* 85 (0896–6273 SB–IM): 85–92. doi:R. Khan, A. Hanbury, J. Stöttinger, and A. Bais, "Color based skin classification," Pattern Recognition Letters, vol. 33, no. 2, pp. 157-163, Jan. 2012.

W, David M. 2011. "Evaluation: From Precision, Recall and F-Measure To Roc, Informedness, Markedness & Correlation." *Journal of Machine Learning Technologies* 2 (1): 37–63. doi:10.1.1.214.9232.

Wahbeh, A. H., Al-Radaideh, Q. A., Al-Kabi, M. N., & Al-Shawakfa, E. M. 2011. "A Comparison Study between Data Mining Tools over Some Classification Methods." *Journal of Advanced Computer Science and Applications* 8 (2): 18–26. doi:10.14569/SpecialIssue.2011.010304.

Wang, Deqing, Hui Zhang, Rui Liu, Weifeng Lv, and Datao Wang. 2014. "T-Test Feature Selection Approach Based on Term Frequency for Text Categorization." *Pattern Recognition Letters* 45 (1). Elsevier B.V.: 1–10. doi:10.1016/j.patrec.2014.02.013.

Wang, Juite, Yung I. Lin, and Shi You Hou. 2015. "A Data Mining Approach for Training Evaluation in Simulation-Based Training." *Computers and Industrial Engineering* 80 (1). Elsevier Ltd: 171–80. doi:10.1016/j.cie.2014.12.008.

Wang, Yi, Pierre Marc Jodoin, Fatih Porikli, Janusz Konrad, Yannick Benezeth, and Prakash Ishwar. 2014. "CDnet 2014: An Expanded Change Detection Benchmark Dataset." In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 393–400. doi:10.1109/CVPRW.2014.126.

Whaiduzzaman, Md, Abdullah Gani, Nor Badrul Anuar, Muhammad Shiraz, Mohammad Nazmul Haque, and Israat Tanzeena Haque. 2014. "Cloud Service Selection Using Multicriteria Decision Analysis." *TheScientificWorldJournal* 2014: 10. doi:10.1155/2014/459375.

Wong, Tzu Tsung. 2015. "Performance Evaluation of Classification Algorithms by K-Fold and Leave-One-out Cross Validation." *Pattern Recognition* 48 (9). Elsevier: 2839–46. doi:10.1016/j.patcog.2015.03.009.

Xia, Meimei, and Jian Chen. 2015. "Multi-Criteria Group Decision Making Based on Bilateral Agreements." *European Journal of Operational Research* 240 (3). Elsevier B.V. doi:10.1016/j.ejor.2014.07.035.

Xiong, Wei, and Qingquan Li. 2012. "Chinese Skin Detection in Different Color Spaces." In *2012 International Conference on Wireless Communications and Signal Processing (WCSP)*, 1–5. doi:10.1109/WCSP.2012.6542853.

Xu, T., Wang, Y., & Zhang, Z. 2012. "Towards Independent Color Space Selection for Human Skin Detection." In *In Pacific-Rim Conference on Multimedia. Springer, Berlin, Heidelberg.*, 7674:337–46. doi:10.1007/978-3-642-34778-8_31.

Xu, T., Wang, Y., & Zhang, Z. 2013. "Pixel-Wise Skin Colour Detection Based on Flexible Neural Tree." *IET Image Processing* 7 (8). Elsevier Inc.: 751–61. doi:10.1016/j.jvcir.2013.05.009.

Xu, Tao, Yunhong Wang, and Zhaoxiang Zhang. 2013. "Pixel-Wise Skin Colour Detection Based on Flexible Neural Tree." *IET Image Processing* 7 (8): 751–61. doi:10.1049/iet-ipr.2012.0657.

Yadav, Shalini, and Neeta Nain. 2016. "A Novel Approach for Face Detection Using Hybrid Skin Color Model." *Journal of Reliable Intelligent Environments* 2 (3). Springer International Publishing: 145–58. doi:10.1007/s40860-016-0024-8.

Yan, C. C., Liu, Y., Xie, H., Liao, Z., & Yin, J. 2014. "Extracting Salient Region for

Pornographic Image Detection." *Journal of Visual Communication and Image Representation* 25 (5): 1130–35.

Yang, Jie, Weier Lu, and Alex Waibel. 1998. "Skin-Color Modeling and Adaptation." In *In Asian Conference on Computer Vision , Springer, Berlin, Heidelberg.*, 687–94.

Yang, Yuhong. 2007. "Consistency of Cross Validation for Comparing Regression Procedures." *Annals of Statistics* 35 (6): 2450–73. doi:10.1214/009053607000000514.

Zaidan, A. A., Karim, H. A., Ahmad, N. N., Alam, G. M., & Zaidan, B. B. 2010. "A New Hybrid Module for Skin Detector Using Fuzzy Inference System Structure and Explicit Rules." *International Journal of Physical Sciences* 5 (13): 2084–97. http://www.researchgate.net/publication/228698045_A_novel_hybrid_module_of_skin_detector_using_grouping_histogram_technique_for_Bayesian_method_and_segment_adjacent-nested_technique_for_neural_network/file/d912f5107afc0eea5e.pdf.

Zaidan, B. B., Zaidan, A. A., Abdul Karim, H., & Ahmad, N. N. 2017. "A New Approach Based on Multi-Dimensional Evaluation and Benchmarking for Data Hiding Techniques." *International Journal of Information Technology & Decision Making*, 1–42. doi:10.1007/978-3-642-48318-9.

Zaidan, A. A., N. N. Ahmad, H. Abdul Karim, M. Larbani, B. B. Zaidan, and A. Sali. 2014b. "Image Skin Segmentation Based on Multi-Agent Learning Bayesian and Neural Network." *Engineering Applications of Artificial Intelligence* 32. Elsevier: 136–50. doi:10.1016/j.engappai.2014.03.002.

Zaidan, A. A., Ahmad, N. N., Karim, H. A., Larbani, M., Zaidan, B. B., & Sali, A. 2014a. "On the Multi-Agent Learning Neural and Bayesian Methods in Skin Detector and Pornography Classifier: An Automated Anti-Pornography System." *Neurocomputing* 131. Elsevier: 397–418. doi:10.1016/j.neucom.2013.10.003.

ZAIDAN, A. A., H. ABDUL KARIM, N. N. AHMAD, B. B. ZAIDAN, and A. SALI. 2013. "An Automated Anti-Pornography System Using a Skin Detector Based on Artificial Intelligence: A Review." *International Journal of Pattern Recognition and Artificial Intelligence* 27 (4): 1350012. doi:10.1142/S0218001413500122.

Zaidan, A. A., B. B. Zaidan, Ahmed Al-Haiqi, M. L.M. Kiah, Muzammil Hussain, and Mohamed Abdulnabi. 2015. "Evaluation and Selection of Open-Source EMR Software Packages Based on Integrated AHP and TOPSIS." *Journal of Biomedical Informatics* 53. Elsevier Inc.: 390–404. doi:10.1016/j.jbi.2014.11.012.

Zaidan, A A, H A Karim, N N Ahmad, B B Zaidan, and M L M Kiah. 2015. "Robust Pornography Classification Solving the Image Size Variation Problem Based on Multi-Agent Learning." *Journal of Circuits Systems and Computers* 24 (2): 37.

doi:10.1142/s0218126615500231.

Zaidan, B. B., and A. A. Zaidan. 2018. "Comparative Study on the Evaluation and Benchmarking Information Hiding Approaches Based Multi-Measurement Analysis Using TOPSIS Method with Different Normalisation, Separation and Context Techniques." *Measurement: Journal of the International Measurement Confederation* 117: 277–94. doi:10.1016/j.measurement.2017.12.019.

Zaidan, B. B., A. A. Zaidan, H. Abdul Karim, and N. N. Ahmad. 2017. "Software and Hardware FPGA-Based Digital Watermarking and Steganography Approaches: Toward New Methodology for Evaluation and Benchmarking Using Multi-Criteria Decision-Making Techniques." *Journal of Circuits, Systems and Computers* 26 (7): 1–27. doi:10.1142/S0219622017500183.

Zanakis, S. H., Solomon, A., Wishart, N., & Dublish, S. 1998. "Multi-Attribute Decision Making: A Simulation Comparison of Select Methods Stelios." *European Journal of Operational Research* 107 (3): 507–29.

Zavadskas, E. K., Kaklauskas, A., Turskis, Z., & Tamošaitienė, J. 2009. "Multi-Attribute Decision-Making Model by Applying Grey Numbers." *Informatica.* 20 (2): 305–320. doi:10.1016/S0377-2217(97)00147-1.

Zhang, Ming Ji, and Wen Gao. 2005. "An Adaptive Skin Color Detection Algorithm with Confusing Backgrounds Elimination." In *Proceedings - International Conference on Image Processing, ICIP.*, 390–93. doi:10.1109/ICIP.2005.1530074.

Zhang, Yongli, and Yuhong Yang. 2015. "Cross-Validation for Selecting a Model Selection Procedure." *Journal of Econometrics* 187 (1): 95–112. doi:10.1016/j.jeconom.2015.02.006.

Zhang, Zhengzhen, and Yuexiang Shi. 2009. "Skin Color Detecting Unite YCgCb Color Space with YCgCr Color Space." In *Proceedings of 2009 International Conference on Image Analysis and Signal Processing, IASP 2009.*, 221–25. doi:10.1109/IASP.2009.5054575.

Zhipeng, C., Junda, H., & Wenbin, Z. 2010. "Face Detection System Based on Skin Color Model." In *2010 International Conference on Networking and Digital Society*, 664–67. doi:10.1109/ICNDS.2010.5479392.

Zhong, Mingjun. 2006. "A Variational Method for Learning Sparse Bayesian Regression." *Neurocomputing* 69 (16–18): 2351–55. doi:10.1016/j.neucom.2006.03.008.

Zhongdong, Wu, Wang Saichao, and Han Zichao. 2013. "A Bayesian Approach to Skin Detection in YCbCr Color Space." In *2013 International Joint Conference on Awareness Science and Technology and Ubi-Media Computing: Can We Realize Awareness via Ubi-Media?, iCAST 2013 and UMEDIA 2013*, 606–9. doi:10.1109/ICAwST.2013.6765511.

Zolfaghari, H., Nekonam, A. S., & Haddadnia, J. 2011. "Color-Base Skin Detection Using Hybrid Neural Network & Genetic Algorithm for Real Times." *International Journal of Computer Science and Information* 9 (10): 67.

Zui Zhang, Hatice Gunes and Massimo Piccardi. 2009. "HEAD DETECTION FOR VIDEO SURVEILLANCE BASED ON CATEGORICAL HAIR AND Zui Zhang , Hatice Gunes and Massimo Piccardi Faculty of Engineering and Information Technology , University of Technology , Sydney (UTS), Australia." In *In Image Processing (ICIP)-IEEE International Conference*, 1137–40.

Zulhadi Zakaria, Nor Ashidi Isa and Shahrel A. Saundi. 2009. "Combining Skin Color And Neural Network For Multiface Detection In Static Images." In *Symposium on Progress in Information & Communication Technology,Conference.*, 147–54.

Zuo, H, H Fan, E Blasch, and H Ling. 2017. "Combining Convolutional and Recurrent Neural Networks for Human Skin Detection." *IEEE Signal Processing Letters* 24 (3): 289–93. doi:10.1109/LSP.2017.2654803.

Zupan, B., & Demšar, J. 2004. "From Experimental Machine Learning to Interactive Data Mining."

LIST OF PUBLICATION

1- MULTE-DIMAENIONAL EVALUATION AND BENCHMARKING FOR DATA MINING APPLECATIONS.**PI2017000692-2017** (PATENT).

2- **ITEX 2017 GOLD MEDAL** (A SMART SOFTWEAR FOR VIRTUALIZATION THE EVALUATION AND BENCHMARKING OF DATA MINING REAL TIME APPLECATIONS BASED ON MULTI-DIMATIONAL CRITERIA) at the 28th International invention, Innovation &Technology Exhibition , ITEX 2017 ,ICT Multimedia, From 11th - 13th May 2017, Kuala Lumpur Convention Centre (KLCC), Malaysia.

3- Towards on Develop a Framework for the Evaluation and Benchmarking of Skin Detectors Based on Artificial Intelligent Models using Multi-Criteria Decision-Making Techniques, International Journal of Pattern Recognition and Artificial Intelligence, Vol. 31, No. 2 (2016) 1759002 (24 pages),Q3,World Scientific Publishing Company , DOI: 10.1142/S0218001417590029.Published.

4- Comprehensive Insights into Evaluation and Benchmarking of Real-time Skin Detectors: Review, Open Issues & Challenges, and Recommended Solutions, Measurement, Volume 111, Pages 167-172, Q1, Elsevier, Published.

5- A Systematic Review on Smartphone Skin Cancer Apps: Coherent Taxonomy, Motivations, Open Challenges and, Recommendations and New Research Direction, Journal of Circuits, Systems, and Computers (JCSC) Vol. 27, No. 5 (2017) 1830003 (40 pages), Q4, World Scientific Publishing Company, DOI: 10.1142/S0218126618300039, Published.

6- Technique for Order Performance by Similarity to Ideal Solution for Solving Complex Situations in Multi-Criteria Optimization of the Tracking Channels of GPS Basedand Telecommunication Receivers. Telecommunication Systems,(2017), Vol.(67), No.176, Pages,1-34,Springer Publishing.

APPENDIX A

PAIRWISE COMPARISON QUESTIONNAIRE PROCESS

UNIVERSITI PENDIDIKAN SULTAN IDRIS
SULTAN IDRIS EDUCATION UNIVERSITY

Toward Conducting Questionnaire

Dear Expert,

The aim of this questionnaire is to compare between criteria for specifying the importance for each of which against others in order to evaluate the skin detection approaches. This questionnaire is a part of the research activities towards Ph.D. degree for Qahtan Majeed Yas at

University Pendidikan Sultan Idris (UPSI)/Malaysia.

Background:

Name:

Your current position title:

University name:

Years of experience:

E-Mail:

Before to answering the questions, it is important to understand the criteria assessed in arriving at a decision. At the high level, these criteria are referred to as main criteria. Each of the main criteria is further refined into sub-criteria. The following figure illustrates the levels:

The questioner to implement the Analytical Hierarchy Process (AHP) method:

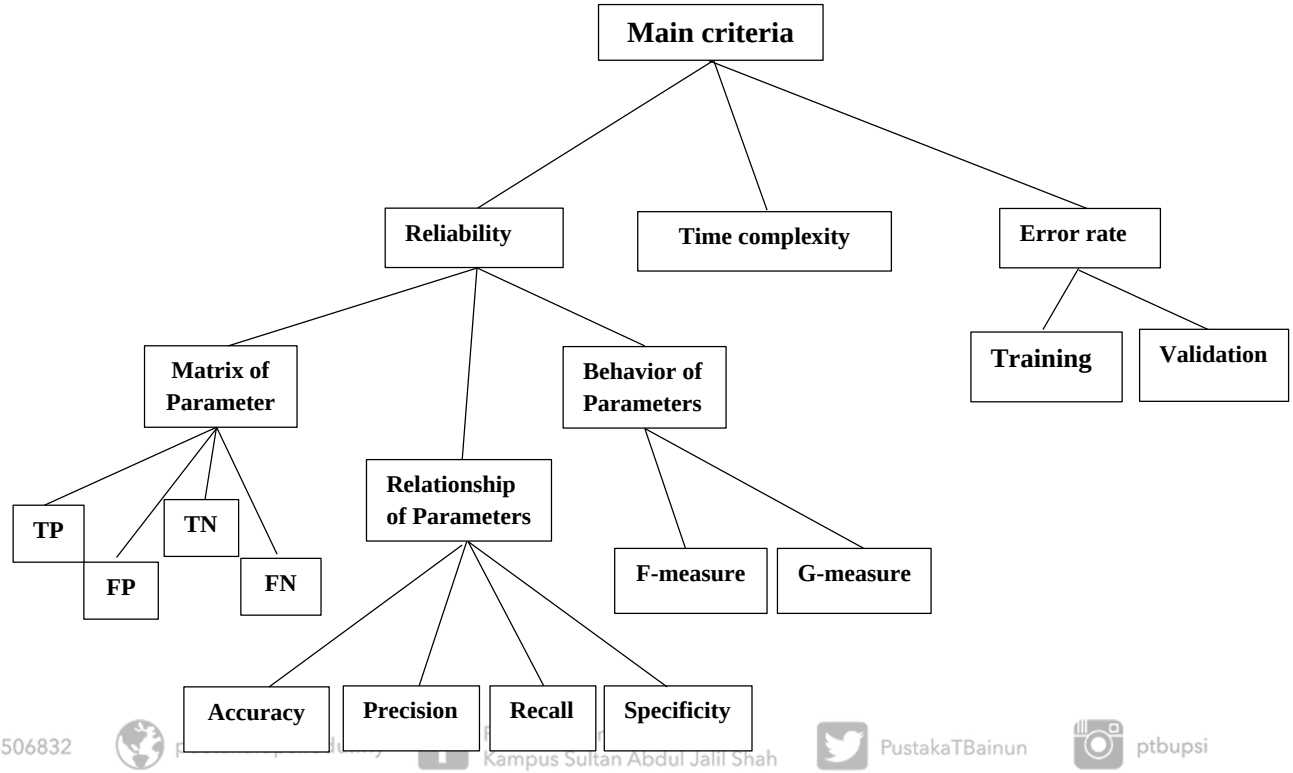


Figure shows the hierarchical distribution criteria

The questioner has several stages will discuss in details as below:

Stage-1: There are three main criteria had been adopted to evaluate in our study. We will highlight these main criteria in detail as follows:

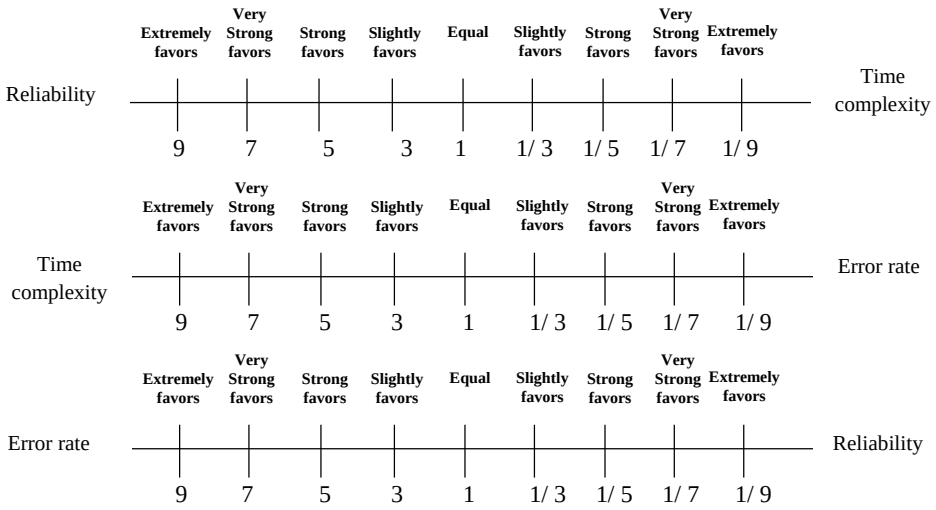
1-Reliability: the degree of quality or state of being fit to be reliable value for any parameter. It is considered one of the main criteria in our study. This criterion includes three subsections will discuss in the next stage.

2. Time complexity: The time complexity is the second main criterion in our study. Time account primarily based on the image size is selected through the difference between the time of input and the time of the output.

3. Error rate within dataset: Basically, the procedure of dataset is to obtain the minimum error rate of the data during the implementation process of the training and validation applied in machine learning. The error rate within the dataset considered is the last main criterion chosen evaluation process for this study. The criterion included on the tow key subsections will discuss in the next stage.

On the other hand, the form shows the procedure how can select/distribute the suitable weight between these criteria.

in the



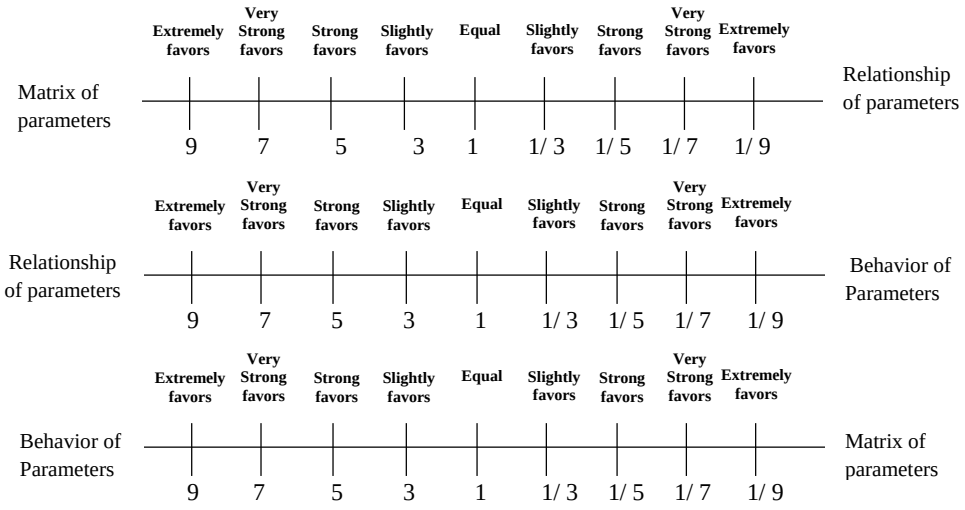
Stage-2: This stage includes the subsections of the main criteria are discussed as follows:.

1- Reliability includes three key subsections will discuss in details:

A) Matrix of parameters included four key parameters called a confusion matrix as backbones for any measure within reliability criterion will discuss in the next stage.

B) Relationship of parameters also included four parameters that are more important criteria typically used to measure the quality ratio for any case will discuss in the next stage.

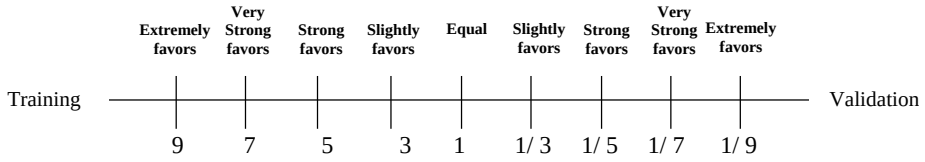
C) Behavior of parameters the last subsections includes two main parameters that are to measure average harmonic mean and geometric for precision and recall perimeter will discuss in the next stage.



2- Error rate within dataset included two main subsections will discuss in details:

A) Validation procedure of the dataset is conducted to set data training at a low error rate.

B) Training often dataset is trained for several times to be set at a lower error rate of the dataset.



Stage-3: this stage includes sub-subsection of stage-2 will discuss in details as follows:

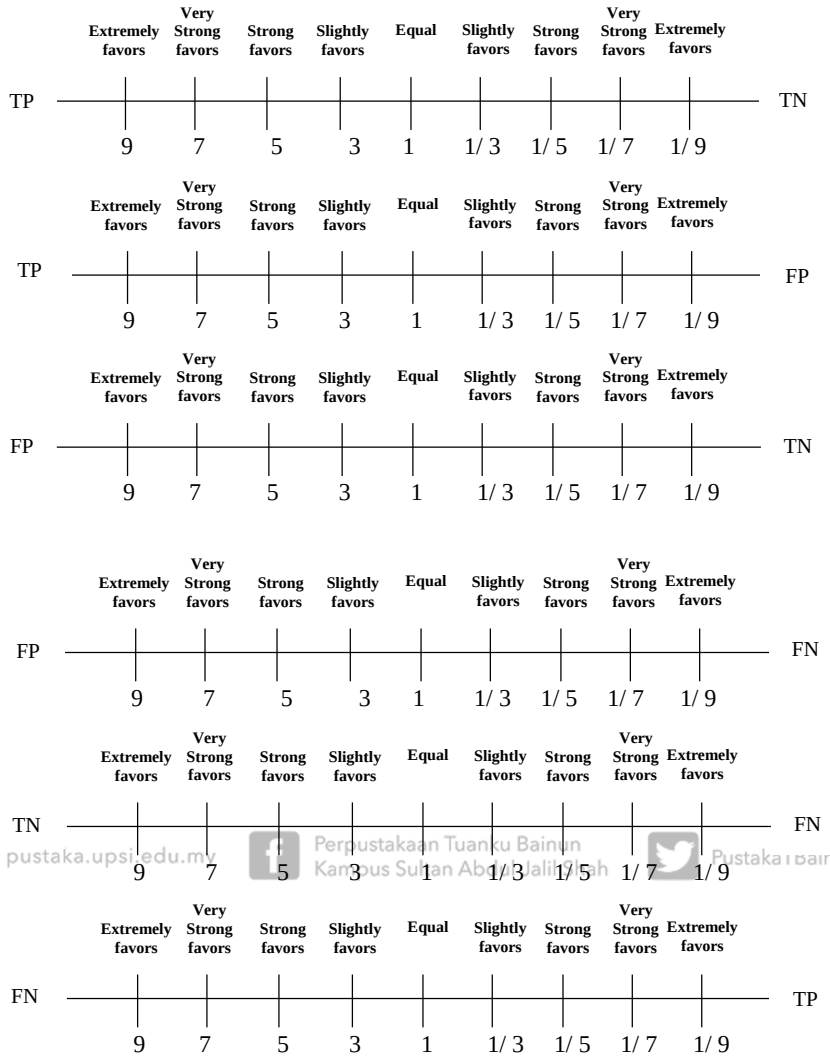
1- Matrix of parameters included four parameters will discuss in details.

A) TP: is the proportion of skin pixels classified correctly as skin.

B) FP: is the proportion of skin pixels classified incorrectly as skin.

C) TN: is the proportion of non-skin pixels classified correctly as skin

D) FN: is the proportion of non-skin pixels classified incorrectly as skin.



2- Relationship of parameters included four parameters will discuss in details.

A) Accuracy measure typically refers to the exactness of an analytical method or the closeness of agreement between the measured value and the value that is accepted, either as a conventional true value or an accepted reference value.

$$\textbf{Accuracy} = \frac{\textbf{True Positive} + \textbf{True Negative}}{\textbf{True Positive} + \textbf{False Positive} + \textbf{False Negative} + \textbf{True Negtaive}}$$

B) Precision measure is the number of true positive (TP) for different classes divided by the total number of elements described as belonging to the positive class (i.e., the sum of TPs and FPs,).

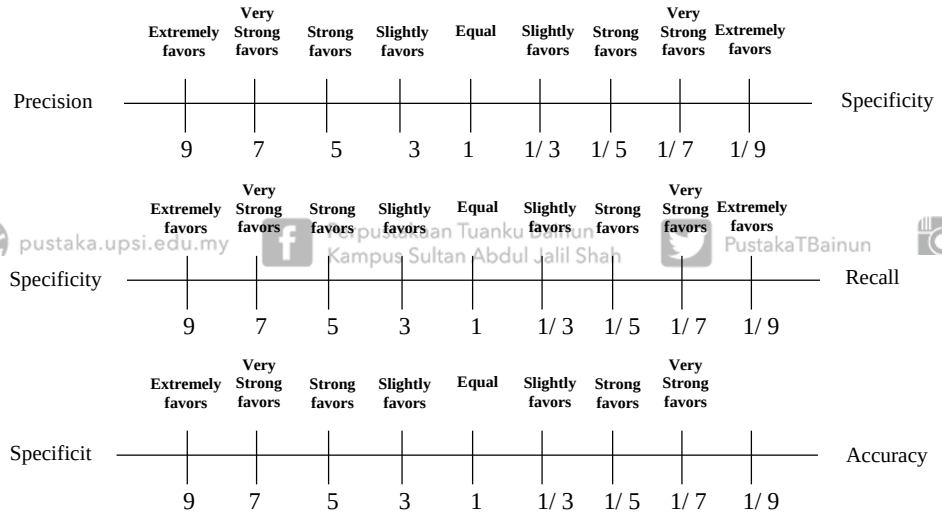
$$\textbf{Precision} = \frac{\textbf{True Positive}}{\textbf{True Positive+ False Positive}}$$

C) Recall is considered as the number of correctly classified positive examples divided by the number of positive examples into the data.

$$\textbf{Recall} = \frac{\textbf{True positive}}{\textbf{True positive+False negetive}}$$

D) Specificity is the ability of a classifier to recognize patterns the negative class to find real negative situations that are correctly predicted as negative.

$$\textbf{Specificity} = \frac{\textbf{True Negative}}{\textbf{True Negative + False Positive}}$$



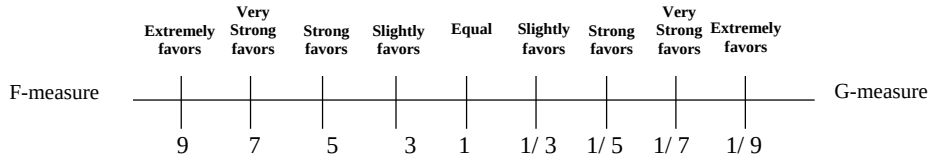
3-Behaviors of parameters included two main parameters will discuss in details.

A) F-measure or called a weighted mean of recall and precision also to make a trade-off between recall and precision and is widely used to evaluate different classifiers.

$$\textbf{F} - \textbf{measure} = \frac{\textbf{2} * \textbf{precision} * \textbf{recall}}{\textbf{recall} + \textbf{precision}}$$

B) G-measure refers to the geometric mean of precision and recall can be represented mathematically as the square root of precision multiplied by the recall and is typically used to evaluate the performance of algorithms.

$$\mathbf{G-measure} = \sqrt{\mathbf{Precision} \times \mathbf{Recall}}$$



Should you have any inquiry or wish to know the result please contact:

Qahtan Majeed Yas

Email: yahoophd@gmail.com

Mobile phone: 00601127184829

……. Thanks for Your Time …….

APPENDIX B

IMPLEMENTATION PAIRED SAMPLE FOR CRITERIA

Ranking order 1

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	W1.0 - W0.9	.0006711	.0056243	.0005412	-.0004018	.0017440	1.240	107	.218
Pair 1	W1.0 - W0.8	.0008818	.0070722	.0006805	-.0004672	.0022309	1.296	107	.198
Pair 1	W1.0 - W0.7	.0015529	.0125654	.0012091	-.0008440	.0039498	1.284	107	.202
Pair 1	W1.0 - W0.6	.0023068	.0186367	.0017933	-.0012482	.0058619	1.286	107	.201
Pair 1	W1.0 - W0.5	.0031379	.0250941	.0024147	-.0016490	.0079247	1.299	107	.197
Pair 1	W1.0 - W0.4	.0040362	.0318087	.0030608	-.0020314	.0101039	1.319	107	.190
Pair 1	W1.0 - W.03	.0049911	.0386862	.0037226	-.0023885	.0123707	1.341	107	.183
Pair 1	W1.0 - W0.2	.0059924	.0456549	.0043931	-.0027165	.0147013	1.364	107	.175
Pair 1	W1.0 - W0.1	.0070320	.0526566	.0050669	-.0030125	.0170765	1.388	107	.168
Pair 1	W1.0 - W.0.0	.0093995	.0596084	.0057358	-.0019711	.0207701	1.639	107	.104

Ranking order 2

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	W0.9 - W0.8	.0005798	.0046587	.0004483	-.0003089	.0014685	1.293	107	.199
Pair 1	W0.9 - W0.7	.0012509	.0102417	.0009855	-.0007028	.0032046	1.269	107	.207
Pair 1	W0.9 - W0.6	.0020048	.0163887	.0015770	-.0011214	.0051310	1.271	107	.206
Pair 1	W0.9 - W0.5	.0028359	.0229100	.0022045	-.0015343	.0072061	1.286	107	.201
Pair 1	W0.9 - W0.4	.0037342	.0296791	.0028559	-.0019272	.0093956	1.308	107	.194
Pair 1	W0.9 - W0.3	.0046891	.0366036	.0035222	-.0022932	.0116714	1.331	107	.186
Pair 1	W0.9 - W0.2	.0056904	.0436130	.0041967	-.0026290	.0140098	1.356	107	.178
Pair 1	W0.9 - W0.1	.0067300	.0506507	.0048739	-.0029319	.0163918	1.381	107	.170
Pair 1	W0.9 - W0.0	.0090975	.0576471	.0055471	-.0018990	.0200940	1.640	107	.104

Ranking order 3

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	W0.8 - W0.7	.0006711	.0056243	.0005412	-.0004018	.0017440	1.240	107	.218
Pair 1	W0.8 - W0.6	.0014250	.0118268	.0011380	-.0008310	.0036810	1.252	107	.213

Pair 1	W0.8 - W0.5	.0022561	.0184069	.0017712	-.0012551	.0057673	1.274	107	.206
Pair 1	W0.8 - W0.4	.0031544	.0252334	.0024281	-.0016590	.0079678	1.299	107	.197
Pair 1	W0.8 - W0.3	.0041093	.0322123	.0030996	-.0020354	.0102539	1.326	107	.188
Pair 1	W0.8 - W0.2	.0051106	.0392724	.0037790	-.0023808	.0126020	1.352	107	.179
Pair 1	W0.8 - W0.1	.0061502	.0463572	.0044607	-.0026927	.0149930	1.379	107	.171
Pair 1	W0.8 - W0.0	.0085177	.0534103	.0051394	-.0016706	.0187060	1.657	107	.100

Ranking order 4

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	W0.7 - W0.6	.0007539	.0062238	.0005989	-.0004333	.0019411	1.259	107	.211
Pair 1	W0.7 - W0.5	.0015850	.0128354	.0012351	-.0008635	.0040334	1.283	107	.202
Pair 1	W0.7 - W0.4	.0024833	.0196981	.0018955	-.0012742	.0062408	1.310	107	.193
Pair 1	W0.7 - W0.3	.0034382	.0267146	.0025706	-.0016578	.0085341	1.337	107	.184
Pair 1	W0.7 - W0.2	.0044395	.0338123	.0032536	-.0020104	.0108893	1.364	107	.175
Pair 1	W0.7 - W0.1	.0054791	.0409337	.0039389	-.0023292	.0132874	1.391	107	.167
Pair 1	W0.7 - W0.0	.0078466	.0480335	.0046220	-.0013160	.0170092	1.698	107	.092

Ranking order 5

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	W0.6 - W0.5	.0008310	.0066240	.0006374	-.0004325	.0020946	1.304	107	.195
Pair 1	W0.6 - W0.4	.0017294	.0135062	.0012996	-.0008470	.0043058	1.331	107	.186
Pair 1	W0.6 - W0.3	.0026843	.0205460	.0019770	-.0012350	.0066035	1.358	107	.177
Pair 1	W0.6 - W0.2	.0036856	.0276690	.0026625	-.0015924	.0089636	1.384	107	.169
Pair 1	W0.6 - W0.1	.0047252	.0348167	.0033502	-.0019163	.0113666	1.410	107	.161
Pair 1	W0.6 - W0.0	.0070927	.0419528	.0040369	-.0009100	.0150954	1.757	107	.082

Ranking order 6

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	W0.5 - W0.4	.0008984	.0068900	.0006630	-.0004160	.0022127	1.355	107	.178
Pair 1	W0.5 - W0.3	.0018532	.0139427	.0013416	-.0008064	.0045129	1.381	107	.170
Pair 1	W0.5 - W0.2	.0028545	.0210817	.0020286	-.0011669	.0068759	1.407	107	.162
Pair 1	W0.5 - W0.1	.0038941	.0282472	.0027181	-.0014942	.0092824	1.433	107	.155
Pair 1	W0.5 - W0.0	.0062617	.0354113	.0034075	-.0004932	.0130165	1.838	107	.069

Ranking order 7

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	W0.4 - W0.3	.0009549	.0070580	.0006792	-.0003915	.0023012	1.406	107	.163
Pair 1	W0.4 - W0.2	.0019562	.0142060	.0013670	-.0007537	.0046660	1.431	107	.155
Pair 1	W0.4 - W0.1	.0029958	.0213831	.0020576	-.0010832	.0070747	1.456	107	.148
Pair 1	W0.4 - W0.0	.0053633	.0285688	.0027490	-.0000863	.0108129	1.951	107	.054

Ranking order 8

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	W0.3 - W0.2	.0010013	.0071519	.0006882	-.0003630	.0023655	1.455	107	.149
Pair 1	W0.3 - W0.1	.0020409	.0143356	.0013794	-.0006937	.0047755	1.479	107	.142
Pair 1	W0.3 - W0.0	.0044084	.0215386	.0020726	.0002998	.0085170	2.127	107	.036

Ranking order 9

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	W0.2 - W0.1	.0010396	.0071866	.0006915	-.0003313	.0024105	1.503	107	.136
Pair 1	W0.2 - W0.0	.0034071	.0144052	.0013861	.0006593	.0061550	2.458	107	.016

Ranking order 10

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	W0.1 - W0.0	.0023675	.0072405	.0006967	.0009864	.0037487	3.398	107	.001