AUTOMATED MELANOMA DETECTION IN DERMOSCOPIC IMAGES

ERDEM OKUR

Thesis for Ph.D. Program in Computer Engineering

Graduate School

Izmir University of Economics

Izmir

2023

AUTOMATED MELANOMA DETECTION IN DERMOSCOPIC IMAGES

ERDEM OKUR

THESIS ADVISOR: ASSOC. PROF. DR. MEHMET TÜRKAN

A Ph.D Thesis

Submitted to

Graduate School of Izmir University of Economics

Department of Engineering

Izmir

2023

ETHICAL DECLARATION

I hereby declare that I am the sole author of this thesis and that I have conducted my work in accordance with academic rules and ethical behaviour at every stage from the planning of the thesis to its defence. I confirm that I have cited all ideas, information and findings that are not specific to my study, as required by the code of ethical behaviour, and that all statements not cited are my own.

Name, Surname:

Erdem OKUR

Date:

27.12.2023

ABSTRACT

AUTOMATED MELANOMA DETECTION IN DERMOSCOPIC IMAGES

Okur, Erdem

Ph.D. Program in Computer Engineering

Advisor: Assoc. Prof. Dr. Mehmet TÜRKAN

December, 2023

Cancer, with its varying and hard to detect types, became one of the most dangerous diseases for humans. Melanoma is a type of skin cancer that has the most mortality rate among its type. The usual melanoma detection process is based on awareness of the patient and the experience of the visual investigator. Even though the invention of dermoscopes reduce its effects, “subjectivity” problem plays a huge role on the detection accuracy, which creates a need for automated detection. In this thesis, history of automated melanoma detection on dermoscopic images and caveats of present frameworks are studied. Different approaches to overcome these caveats are explored. As a result, a new melanoma detection algorithm based on Bag of Visual Words (BoVW) concept, which combines traditional methods with new age deep learning techniques, is created. The performance of the new algorithm is tested on the popular International Skin Imaging Collaboration (ISIC) Challenge 2017 dataset, which yielded tremendously good results. With 96.2% accuracy and more importantly with 99.8% sensitivity, it surpassed all other entries in the ISIC 2017 Leaderboard. Since, sensitivity represents the algorithm's success on correctly classifying melanoma cases, this success places the algorithm on a special place in the domain. Lastly, future directions on the domain are explored on the terms of increasing the performance of the newly born algorithm further.

Keywords: melanoma detection, bag of visual words, neural networks, ISIC.

ÖZET

DERMOSKOPİK GÖRÜNTÜLERDE OTOMATİK MELANOM TESPİTİ

Okur, Erdem

Bilgisayar Mühendisliği Doktora Programı

Tez Danışmanı: Doç. Dr. Mehmet TÜRKAN

Aralık, 2023

Kanser, çeşitli ve tespit edilmesi zor türleri ile insanlar için en tehlikeli hastalıklardan biri haline gelmiştir. Melanom, türleri arasında ölüm oranı en fazla olan cilt kanseri türüdür. Olağan melanom tespit süreci, hastanın farkındalığına ve görsel muayene eden kişinin deneyimine dayanmaktadır. Dermoskopların icadı ile etkileri azalsa da, "öznellik" sorunu melanom tespit doğruluğunda büyük rol oynamakta ve bu da otomatik algılama ihtiyacını doğurmaktadır. Bu tezde, dermoskopik görüntülerde otomatik melanom tespitinin tarihçesi ve daha önce sunulan sistemlerin açıkları incelenmiştir. Bu açıkların üstesinden gelmek için farklı yaklaşımlar araştırılmıştır. Sonuç olarak, geleneksel yöntemleri yeni çağın derin öğrenme teknikleriyle birleştiren Görsel Kelimeler Çantası (BoVW) konseptine dayalı bir melanom saptama algoritması oluşturulmuştur. Yeni algoritmanın performansı, popüler Uluslararası Cilt Görüntüleme İşbirliği (ISIC) 2017 yarışması veri kümesi üzerinde test edilmiş ve son derece iyi sonuçlar elde edilmiştir. %96,2 doğrulukla ve daha da önemli olarak %99,8 hassasiyetle yeni algoritma ISIC 2017 başarı tablosundaki diğer tüm katılımcıları geride bırakmıştır. Hassasiyet, algoritmanın melanom vakalarını doğru sınıflandırma konusundaki başarısını temsil ettiğinden bu başarı, algoritmayı alanında özel bir yere yerleştirmektedir. Son olarak, yeni doğan algoritmanın performansını daha da arttırmak açısından, alan üzerinde gelecekte izlenebilecek yönler araştırılmıştır.

Anahtar Kelimeler: melanom tespiti, görsel kelimeler çantası, sinir ağları, ISIC.

This thesis work is dedicated to my late grandparents...

ACKNOWLEDGEMENT

This dissertation could not have been completed without the invaluable support and contributions of many individuals. Their assistance was crucial in both the preparation and completion of this study, and I am grateful to everyone who played a role in making this journey successful, both professionally and personally.

Firstly, I extend my deepest gratitude to my advisor, Assoc. Prof. Dr. Mehmet Türkan. His constant support and expertise were vital in navigating the challenging path of this thesis.

I am also deeply grateful to my mentors, Assoc. Prof. Dr. Kaya Oğuz and Prof. Dr. Devrim Ünay, for guiding me through the ups and downs of this process alongside my advisor.

A special thanks goes to the head of our Software Engineering Department, Assoc. Prof. Dr. Senem Kumova Metin, for her unwavering support and encouragement.

The journey of a Ph.D. is demanding, and I was fortunate to have a group of individuals who provided the much-needed distraction and different perspectives. I am immensely thankful to Serhat Uzunbayır, Çınar Gedizlioğlu, Büşra Temuçin, and "the MB 162" for making my time at IUE enjoyable and fruitful.

Particularly, I owe a heartfelt thank you to Hande Aka Uymaz and Duygu Geçkin. Hande has been a fellow traveler in this journey, albeit on different paths. Duygu has been instrumental in helping me see things from various perspectives, often without her realizing it. Their companionship and tolerance of my constant discussions about my Ph.D. and school were invaluable. They played a key role in motivating me daily. Of course these measly sentences are not enough to express my feelings, so again thank you...

Lastly, the most profound gratitude is for my parents. Their unwavering support has been the cornerstone of all my achievements. I am filled with immense gratitude and accomplishment as I say, "Thank you for everything."

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

In the realm of oncology, skin cancer (cutaneous malignancies) remain the most prevalent form of cancer (Mayo Clinic, 2022; American Cancer Society, 2023). According to Republic of Türkiye Ministry of Health, around 22000 new skin cancer incidences occurred just in 2018 (Republic of Türkiye Ministry of Health, 2022). In 2020, the European Union documented in excess of 100000 melanoma skin cancer incidences, with the continent concurrently reporting over 16000 fatalities attributed to this malignancy (Stewart, 2023). Similarly, in the United States, the annual incidence of skin cancer surpasses the cumulative incidence of other prominent cancers such as breast, lung, prostate, and colorectal (American Cancer Society, 2017). Specifically, 197700 new cases diagnosed in 2022 (American Academy of Dermatology (AAD), 2022). Global statistics from the World Health Organization elucidate that annually, non-melanoma skin cancers account for two to three million new cases, while melanoma contributes an additional 132000 (WHO, 2017). It is noteworthy that melanoma statistics are often delineated distinctly. This demarcation stems from melanoma's grave nature, qualifying it as the most lethal variant of skin cancer. Despite comprising a mere 1% of the total skin cancer incidence, it is responsible for the vast majority of skin cancer-related mortalities (American Cancer Society, 2023). A comprehensive review of international data underscores the severe implications of melanoma; for instance, in a 2020 European Commission research indicated that Europe experienced a melanoma-related death almost every 33 minutes (European Commission, 2020). The research further unveils crucial statistics pertaining to melanoma, encompassing aspects of survivability, age demographics, and nation-specific mortality rates within Europe. Australia, grappling with its heightened prevalence, regards melanoma, "Australia's national cancer", as a primary health concern—it ranks as the third leading cancer following prostate (in males) and breast (in females) malignancies. Every 30 minutes, an Australian is diagnosed with melanoma. Projections for the current year suggest that approximately 16800 Australians will receive a melanoma diagnosis (Melanoma

Institute Australia, 2023a). Lastly, according to the Melanoma Research Foundation, it is anticipated that melanoma will result in the demise of 7990 Americans in 2023. Furthermore, melanoma diagnoses are made in approximately 400 American children each year (Melanoma Research Foundation (MRF), 2023).

1.1. What is Melanoma?

Melanoma originates in the melanocytes, or pigment cells (Melanoma Institute Australia, 2023b). During childhood or adolescence, the congregation of these cells in the skin results in the formation of a mole. Melanoma arises when these melanocytes undergo aberrant and unregulated growth. Approximately one-third of all melanomas originate from pre-existing moles. However, it is crucial to note that melanomas can manifest anywhere on the epidermis (Melanoma Institute Australia, 2023b). Due to this fact, in this thesis moles are not separately mentioned and “lesion” will be used as common term. Distinct from other forms of skin cancer, melanoma possesses a rapid propensity to metastasize to other tissues. The mechanism of metastasis can occur via the tissue, lymphatic system, or bloodstream (National Cancer Institute, 2023). When dispersing through tissue, it extends merely to adjacent areas; however, once it invades the lymphatic system or blood vessels, it can proliferate to distant tissues. The affected tissue subsequently manifests a malignant growth, posing significant treatment challenges. A silver lining is that melanoma typically presents on the skin’s surface, allowing for potential early detection through simple visual examination, hence enhancing the prospects of a comprehensive cure if diagnosed in early stages. Regrettably, the definitive staging of melanoma is ascertainable only post the excision or biopsy of a suspicious lesion or mole. For staging, four primary characteristics are considered: tumor thickness (quantified on the Breslow scale (Marghoob et al., 2000)), presence of ulceration, and dissemination to lymph nodes or other bodily regions. Melanoma stages can be named specifically as Stage 0, I (A/B), II (A/B/C), III, and IV, with each stage elucidated in depth in the following section.

1.2. Stages of Melanoma

Staging is critical in selecting appropriate therapies and determining prognosis of a melanoma case. To facilitate staging, a myriad of procedures might be employed, including comprehensive physical examinations; lymph node mapping, where a substance is administered adjacent to the melanoma for tracking through lymphatic channels, potentially leading to a biopsy or lymph node removal surgery; Computed Tomography (CT) scans; Positron Emission Tomography (PET) scans, in which patients receive an injection of radioactive glucose - given that tumor cells consume glucose more voraciously than their normal counterparts, these cells are highlighted more prominently in the scan; Magnetic Resonance Imaging (MRI) using gadolinium, where the tumor cells appear more luminous in magnetic resonance imaging when gadolinium is introduced; and blood chemistry tests, such as measuring Lactate Dehydrogenase (LDH) levels, where elevated LDH could signify the presence of melanoma (National Cancer Institute, 2023). These investigative outcomes, in conjunction with the biopsy from the suspected lesion, culminate in a conclusive stage determination. For individuals diagnosed with Stage III or IV melanoma, merely excising the lesion is not sufficient. As previously emphasized, addressing melanoma at these advanced stages is considerably challenging, necessitating more intensive interventions like chemotherapy (Airley, 2009), radiation therapy (Washington and Leaver, 2016), immunotherapy (Naing and Hajjar, 2017), and targeted therapy (Yan et al., 2011; Siegel et al., 2018). Hence, the prompt evaluation of any suspicious mole or lesion is of paramount importance for potential early-stage identification. Table 1 provides concise and comprehensible definitions of the various stages of melanoma.

Table 1. The stages of melanoma as presented by the PDQ Adult Treatment Editorial Board (last update: June 2023) (National Cancer Institute, 2023).

Stage	**Definition**
Stage 0	At stage 0, atypical melanocytes are located within the epidermis. There is potential for these deviant melanocytes to transform into cancerous cells and infiltrate the adjacent healthy tissue. This stage is also referred to as “melanoma in situ.”

Table 1 continued from previous page

Stage	Definition
Stage I	The presence of cancer is confirmed. This stage is further segmented into stages IA and IB. **Stage IA:** The tumor's thickness does not exceed 1 millimeter, and it may or may not exhibit ulceration. **Stage IB:** The tumor's thickness ranges between 1 to 2 millimeters and is devoid of ulceration
Stage II	This stage is segmented into stages IIA, IIB, and IIC. **Stage IIA:** In Stage IIA, the tumor exhibits one of the following characteristics: -It possesses a thickness greater than 1 millimeter but does not exceed 2 millimeters and presents with ulceration. -Alternatively, its thickness is more than 2 millimeters but remains within 4 millimeters, and it lacks ulceration. **Stage IIB:** In Stage IIB, the tumor exhibits one of the following characteristics: -It has a thickness that exceeds 2 millimeters but is no greater than 4 millimeters, accompanied by ulceration. -Or, the tumor has a thickness surpassing 4 millimeters, but it is devoid of ulceration. **Stage IIC:** The tumor possesses a thickness that exceeds 4 millimeters and is accompanied by ulceration.
Stage III	This stage is segmented into stages IIIA, IIIB, IIIC and IIID. Segmentation is done based on the ulceration status of the primary tumor as well as the degree of its proliferation into adjacent structures, including the lymph nodes, lymphatic vessels, and surrounding skin. Each sub-stage features several number of conditions to consider. For further details on each sub-stage, please refer to National Cancer Institute (2023).
Stage IV	In Stage IV, the cancer has metastasized to distant regions of the body. This includes organs such as the lungs, liver, brain, spinal cord, and bones, as well as soft tissues, encompassing muscles, the gastrointestinal (GI) tract, and distant lymph nodes. Additionally, the cancer might have proliferated to skin areas considerably remote from its initial site of origin.

1.3. Early Detection and Clinical Features

The diagnosis of melanoma in early stages is critical as mentioned previously. With that being noted, early detection hinges on increasing the community awareness at first. Currently, several facilities and initiatives offer evaluation of skin lesions. For example, Türkiye boasts 41 melanoma-specific visual inspection clinics (Euro Melanoma, 2023). Moreover, numerous online platforms facilitate appointment bookings in diverse clinics globally. Comprehensive information is also accessible via the World Wide Web (Turkiye Kanserle Savas Vakfi, 2021; Mayo Clinic, 2023). A prevalent guideline for lesion awareness is the "ABCD(E)'s of Melanoma" (WebMD, 2023; Melanoma UK, 2016). This elementary directive explains the **A**symmetry, **B**order irregularity, multitudinous **C**olor variations, and the **D**iameter characteristics of lesions. The "E" denotes "**E**volution", highlighting the rapid growth of a lesion.

It should be noted that current studies do not incorporate the “E” in the automated melanoma detection paradigm. Individuals suspecting a lesion based on these guidelines are advised to consult a dermatologist or specialized medical professional for a visual examination. Post-examination, the medical practitioner might recommend an excision if deemed necessary. Should this be required, the excision procedure is typically straightforward and brief, potentially offering complete remediation of the suspect.

From the doctors’ and visual inspectors’ point of view, there are algorithms to visually evaluate a lesion’s clinical features to detect melanoma. These features can be classified as either global or local. Global features encompass the entirety of the lesion, whereas local features predominantly manifest in a specific area or cluster within the lesion. Clinically, they can be segregated into three primary categories: *Texture*, *Shape*, and *Color*. The clinical features given in this section are important for detecting melanoma automatically in traditional approaches. In section 2.2, how these features are detected and utilized is explained with example studies.

The clinical features present within a lesion can exhibit diverse patterns, and the methodologies employed by dermatologists during visual inspections lay the groundwork for recognizing these features (Malvehy et al., 2007; Argenziano et al., 2003; Braun et al., 2005). Pigment networks, spots and globules, aberrant and typical networks, star-burst patterns, and vascular structures can be considered as signs of melanoma. Additionally, a standardization approach outlined in Malvehy et al. (2007) delineates two pivotal phases for clinical feature identification. The initial phase involves a comprehensive examination of the lesion, where the features undergo visual assessments, as elucidated in Table 2. If any feature indicative of melanoma is identified as a “melanocytic lesion”, it prompts the visual investigator to proceed to the second phase. This subsequent phase includes four guiding principles, which are elaborated on and contrasted in depth in the following parts of this section.

Table 2. The first step involves identifying key features during the visual inspection. Lesions that raise suspicion based on these characteristics then progress to the second evaluative step. (This table contains minor exceptions, please refer Malvehy et al. (2007).)

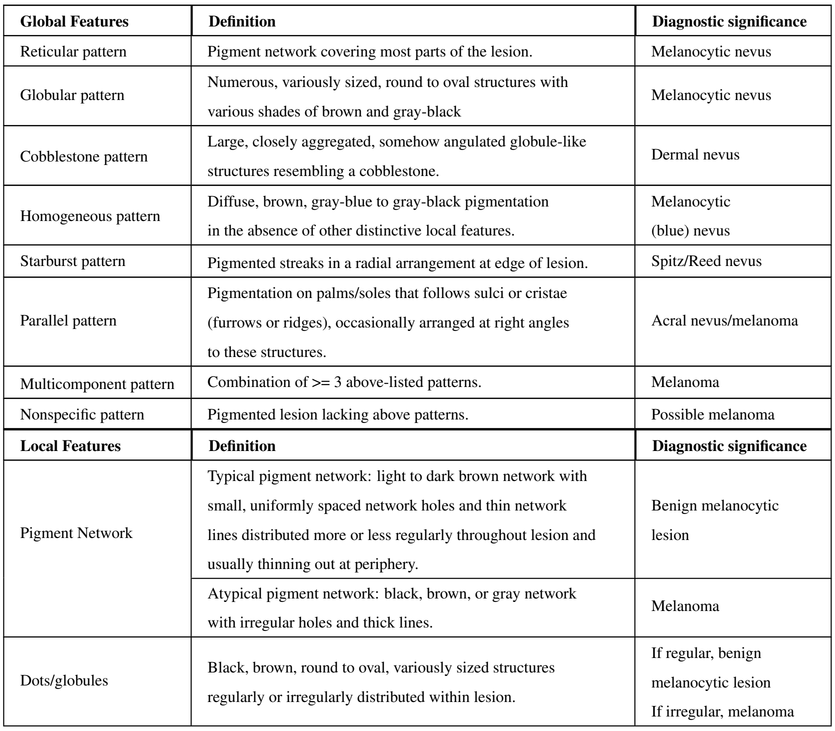
Dermoscopic criterion	**Definition**	**Diagnostic Significance**
Pigment Network-Pseudo-Network	Network of brownish interconnected lines overlying background of tan diffuse pigmentation. In facial skin a peculiar pigment network, also called pseudo-network, is typified by round, equally sized network holes corresponding to preexisting follicular ostia.	Melanocytic lesion
Aggregated globules	Numerous, variously sized, often clustered, round to oval structures with various shades of brown and gray-black. Should be differentiated from multiple blue-gray globules.	Melanocytic lesion
Streaks	These have been previously described separately as pseudopods and radial streaming, but are now combined into one term. They are bulbous and often kinked or finger-like projections seen at the edge of a lesion. They may arise from network structures but more commonly not.	Melanocytic lesion
Homogeneous blue pigmentation	Structureless blue pigmentation in absence of pigment network or other discernible structures.	Melanocytic lesion
Parallel pattern	Seen in melanocytic lesions of palms/soles and mucosal areas. On palms/soles pigmentation may follow sulci or cristae (ie, furrows or ridges) of the dermatoglyphics. Rarely arranged at right angles to these structures.	Melanocytic lesion
Multiple milia-like cysts	Numerous, variously sized, white or yellowish, roundish structures.	Seborrheic keratosis
Comedo-like openings	Brown-yellowish to brown-black, round to oval, sharply circumscribed keratotic plugs in the ostia of hair follicles. Irregularly shaped comedo-like openings are also called irregular crypts.	Seborrheic keratosis
Light brown fingerprint-like structures	Light brown, delicate, network-like structures with fingerprint pattern.	Seborrheic keratosis
Cerebriform pattern	Dark brown furrows between ridges producing brain-like appearance.	Seborrheic keratosis
Arborizing vessels	Tree-like branching telangiectases.	Basal cell carcinoma
Leaf-like structures	Brown to gray/blue discrete bulbous structures forming leaf-like patterns. They are discrete pigmented nests (islands) never arising from pigment network and usually not arising from adjacent confluent pigmented areas.	Basal cell carcinoma
Large blue-gray ovoid nests	Well-circumscribed, confluent or near confluent pigmented ovoid or elongated areas, larger than globules, and not intimately connected to pigmented tumor body.	Basal cell carcinoma
Multiple blue-gray globules	Multiple globules (not dots) that should be differentiated from multiple blue-gray dots (melanophages).	Basal cell carcinoma
Spoke-wheel areas	Well-circumscribed, radial projections, usually tan but sometimes blue or gray, meeting at often darker (dark brown, black, or blue) central axis.	Basal cell carcinoma
Ulceration	Absence of epidermis often associated with congealed blood, not due to well-described recent history of trauma.	Basal cell carcinoma
Red-blue lacunae	More or less sharply demarcated, roundish or oval areas with reddish, red-bluish, or dark-red to black.	Vascular lesion
Red-bluish to reddish-black homogeneous areas	Structureless homogeneous red-bluish to red-black areas.	Vascular lesion
None of listed criteria	Absence of above-mentioned criteria.	Melanocytic lesion

In Table 2, the term “melanocytic lesion” denotes a lesion resulting from the proliferation of melanocytes, which may be potentially precursory to melanoma. “Seborrheic keratosis” encompasses benign, wart-like growths that, although they might resemble precancerous manifestations, are actually benign. “Basal cell carcinoma” represents a form of skin cancer, which, though it very seldom metastasizes, can manifest as red patches, open wounds, or pinkish proliferations. “Vascular lesions” constitute another category of lesions which are typically benign; however, their visual appearance can be diagnostically challenging, contingent on their specific classification.

1.3.1. Pattern Analysis Criteria

The Pattern Analysis Criteria serve as a guideline facilitating the examination of both global and local features within a suspect lesion. In this second step analysis, the characteristics explored are essentially refined sub-categories of those identified in the first step. The diagnosis may suggest melanoma based on the specific attribute, its manifestation, or amalgamations of diverse variations of the same feature. Comprehensive insights into the global and local features and their potential implications can be found in Table 3.

Table 3. Pattern Analysis Criteria - Global and Local Features.



Global Features	**Definition**	**Diagnostic significance**
Reticular pattern	Pigment network covering most parts of the lesion.	Melanocytic nevus
Globular pattern	Numerous, variously sized, round to oval structures with various shades of brown and gray-black	Melanocytic nevus
Cobblestone pattern	Large, closely aggregated, somehow angulated globule-like structures resembling a cobblestone.	Dermal nevus
Homogeneous pattern	Diffuse, brown, gray-blue to gray-black pigmentation in the absence of other distinctive local features.	Melanocytic (blue) nevus
Starburst pattern	Pigmented streaks in a radial arrangement at edge of lesion.	Spitz/Reed nevus
Parallel pattern	Pigmentation on palms/soles that follows sulci or cristae (furrows or ridges), occasionally arranged at right angles to these structures.	Acral nevus/melanoma
Multicomponent pattern	Combination of >= 3 above-listed patterns.	Melanoma
Nonspecific pattern	Pigmented lesion lacking above patterns.	Possible melanoma
Local Features	**Definition**	**Diagnostic significance**
Pigment Network	Typical pigment network: light to dark brown network with small, uniformly spaced network holes and thin network lines distributed more or less regularly throughout lesion and usually thinning out at periphery.	Benign melanocytic lesion
	Atypical pigment network: black, brown, or gray network with irregular holes and thick lines.	Melanoma
Dots/globules	Black, brown, round to oval, variously sized structures regularly or irregularly distributed within lesion.	If regular, benign melanocytic lesion If irregular, melanoma

Table 3 continued from previous page

Streaks (pseudopods and radial streaming)	Streaks are bulbous and often kinked or finger-like projections seen at the edge of lesion. They may arise from network structures but more commonly not. They range in color from tan to black.	If regular, benign melanocytic lesion (Spitz/Reed nevus) If irregular, melanoma
Blue-whitish veil	Irregular, structureless area of confluent blue pigmentation with an overlying white "ground-glass" film. Pigmentation cannot occupy entire lesion and usually corresponds to a clinically elevated part of the lesion.	Melanoma
Regression structures	White scar-like depigmentation and/or blue pepper-like granules usually corresponding to a clinically flat part of the lesion.	Melanoma
Hypopigmented areas (structureless/ homogeneous)	Focal areas devoid of structures with less pigmentation than overall pigmentation of lesion and comprising at least 10% of total area.	Nonspecific
Blotches	Black, dark brown, and/or gray structureless areas with symmetric or asymmetric distribution within lesion.	If symmetric, benign melanocytic lesion If asymmetric, melanoma
Vascular structures	Comma-like vessels.	Dermal nevus
	Hairpin vessels.	If uniformly distributed, seborrheic keratosis If irregularly distributed consider melanoma
	Dotted vessels.	Melanoma
	Linear-irregular vessels.	Melanoma
	Vessels and/or erythema within regression structures.	Melanoma

In Table 3, the term "melanocytic nevus" pertains to benign growths encompassing melanocytes. The "dermal nevus" corresponds to melanocytic growths, predominantly benign in nature, situated within the dermis layer. "Spitz/Reed nevus" are types of melanocytic growths bearing a close resemblance to melanoma, with a subset holding the potential to metastasize. Conversely, "acral nevus" represents benign growths primarily found on the palms or soles, and they are typically more diminutive in size.

1.3.2. ABCD Rule

The renowned ABCD Rule, which serves as a community guideline, is based on four features of a lesion, namely A, B, C, and D, as mentioned in Section 1.3. For professional visual inspectors, a variant of this guideline is prepared. The primary deviation lies in the interpretation of "D". In the public version, "D" refers to the diameter (or dimension) of the lesion. In contrast, for trained investigators, it signifies "dermoscopic structures". These structures include various networks, structureless zones, globules, streaks, and dots. Another significant distinction is present in the evaluative methodology. Each characteristic within the ABCD Rule has a designated score and an associated weight factor. During lesion assessment, the score for a specific characteristic is multiplied by its weight factor, resulting in an individual score for that attribute. The aggregate of these individual scores gives the comprehensive score for the lesion. A score below 4.75 suggests the lesion is benign. A score ranging between 4.75 and 5.45 implies that the lesion might require excision or monitoring over a certain duration. Should the score surpass the 5.45 benchmark, the lesion is deemed necessitating removal. An exhaustive breakdown of these characteristics, alongside their respective scores and weights, can be found in Table 4.

Table 4. ABCD Rule - Feature details with respective scores and weights.

Dermoscopic Criterion	**Definition**	**Score**	**Weight Factor**
A: Asymmetry	In 0, 1, or 2 perpendicular axes; assess not only contour, but also colors and structures.	0–2	1.3
B: Border	Abrupt ending of pigment pattern at periphery in 0-8 segments.	0–8	0.1
C: Color	Presence of up to 6 colors (white, red, light-brown, dark-brown, blue-gray, black).	1–6	0.5
D: Dermoscopic structures	Presence of network, structureless (homogeneous) areas, branched streaks, dots, and globules.	1–5	0.5

1.3.3. Menzies Scoring

The Menzies Scoring method classifies certain lesion features into two distinct categories. The initial category, termed as the negative features group, encompasses merely two attributes: *symmetry* and *singular color*. Conversely, the positive features group consists of nine attributes, given in Table 5. Notably, these positive features serve as unambiguous markers for melanoma. Consequently, if a lesion lacks attributes from the negative group but exhibits at least one feature from the positive group, it is diagnosed as melanoma.

Table 5. Menzies Scoring - Feature Groups and definitions.

Negative Features	**Definition**
Symmetry of pattern	Symmetry of pattern is required across all axes through lesion’s center of gravity (center of lesion). Symmetry of pattern does not require shape symmetry.
Presence of a single color	The colors scored are black, gray, blue, dark brown, tan, and red. White is not scored as a color.
Positive Features	**Definition**
Blue-white veil	An area of irregular, structureless confluent blue pigmentation with an overlying white “ground-glass" haze. It cannot occupy entire lesion and cannot be associated with red-blue lacunae.
Multiple brown dots	Focal areas of multiple brown (usually dark brown) dots (not globules).

Table 5 continued from previous page

Pseudopods	Bulbous and often kinked projections that are found at the edge of lesion either directly connected to the tumor body or pigmented network. They can never be seen distributed regularly or symmetrically around the lesion. When connected directly to the tumor body, they must have an acute angle to the tumor edge or arise from linear or curvilinear extensions. When connected to the network, the width of the bulbous ending must be greater than the width of any part of the surrounding network and at least double that of its directly connected network projection.
Radial streaming	Finger-like extensions at the edge of lesion that are never distributed regularly or symmetrically around the lesion.
Scar-like depigmentation	Areas of white distinct irregular extensions (true scarring), which should not be confused with hypopigmentation or depigmentation due to simple loss of melanin.
Peripheral black dots/globules	Black dots/globules found at or near edge of lesion.
Multiple (5 or 6) colors	The colors scored are black, gray, blue, dark brown, tan, and red. White is not scored as a color.
Multiple blue/gray dots	Foci of multiple blue or gray dots (not globules) often described as “pepper-like" granules in pattern.
Broadened network	Network made up of irregular thicker “cords" of the net, often seen focally thicker.

1.3.4. 7-Point Checklist

The 7-point Checklist serves as an additional diagnostic tool employed by dermatologists during the second step of lesion evaluation, as outlined in reference (Malvehy et al., 2007). This checklist enumerates seven distinct features, combinations of which could signify the potentiality of melanoma in a given lesion. Analogous to the ABCD Rule, each feature within the checklist is allocated a specific score. The cumulative score for a lesion is then derived based on the presence of these enumerated features (elaborated upon in Table 6). A resultant score surpassing the threshold of 3 raises suspicions of melanoma. A visual representation of this evaluation process can be found in Figure 1, showcasing two melanoma instances evaluated via the 7-point Checklist, inclusive of discerned features and their correlated scores.

Table 6. 7-point Checklist - Features with their scores.

Dermoscopic criterion	**Definition**	**Score**
1. Atypical pigment network	Black, brown, or gray network with irregular holes and thick lines.	2
2. Blue-whitish veil	Irregular, structureless area of confluent blue pigmentation with and overlying white “ground-glass" film. The pigmentation cannot occupy the entire lesion and usually corresponds to a clinically elevated part of the lesion.	2
3. Atypical vascular pattern	Linear-irregular or dotted vessels not clearly seen within regression structures	2
4. Irregular streaks	Brown to black, bulbous or finger-like projections irregularly distributed at the edge of lesion. They may arise from network structures but more commonly not.	1
5. Irregular dots/globules	Black, brown, round to oval, variously sized structures irregularly distributed within lesion.	1

Table 6 continued from previous page

6. Irregular blotches	Black, brown, and/or gray structureless areas asymmetrically distributed within lesion.	1
7. Regression structures	White scar-like depigmentation and/or blue pepper-like granules usually corresponding to clinically flat part of the lesion.	1

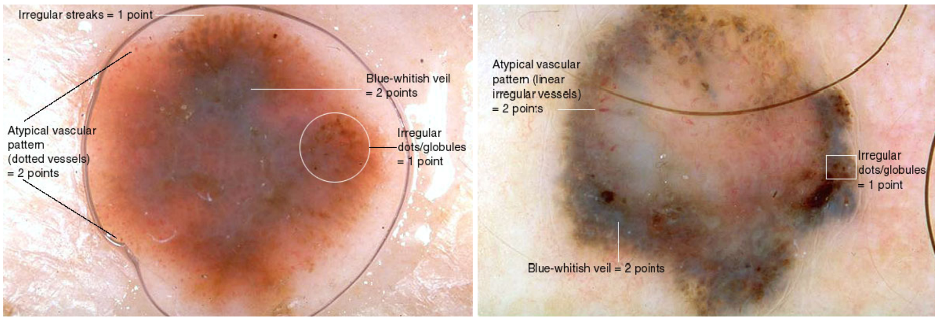

Irregular streaks = 1 point
Atypical vascular pattern (dotted vessels) = 2 points
Blue-whitish veil = 2 points
Irregular dots/globules = 1 point
Atypical vascular pattern (linear irregular vessels) = 2 points
Irregular dots/globules = 1 point
Blue-whitish veil = 2 points

Figure 1. Figure presenting two instances of melanoma assessed using the 7-point Checklist. The discernible features, along with their corresponding scores, are depicted for each case, as detailed in Dermoscopy.org (2003).

1.3.5. Evaluative Comparison of Second-Step Algorithmic Efficiencies

Dolianitis et al. (2005) conducted a study to gauge the efficacy of the four principal methods used during the secondary phase of visual lesion inspection. For this, they trained 61 Australian medical professionals in the techniques of Pattern Analysis Criteria, ABCD Rule, Menzies Scoring, and the 7-Point Checklist. Subsequently, these trained professionals were tasked with evaluating 40 images of melanocytic lesions. The results of this study, captured in Table7, span three main metrics: sensitivity, specificity, and accuracy.

Sensitivity, often termed the true positive rate, indicates the proportion of actual melanoma cases that are correctly identified as such Parikh et al. (2008). This metric is calculated by dividing the number of true positives by the sum of true positives and false negatives. Conversely, specificity or the true negative rate highlights the percentage of non-melanoma cases that are accurately recognized as non-melanoma. It is derived by taking the ratio of true negatives to the sum of true negatives and false positives.

In essence, high values of sensitivity and specificity signify the system's competence in definitively determining the presence or absence of melanoma in a given lesion. As per the data from Table 7, Menzies Scoring outperformed the other methods in terms of sensitivity and accuracy. Pattern Analysis Criteria, however, achieved the highest specificity. Given the critical nature of correctly identifying both melanoma and non-melanoma cases, it is recommended to employ both the Menzies Scoring and Pattern Analysis Criteria for optimal results.

Table 7. Evaluative comparison of four algorithms in the second step of visual inspection Dolianitis et al. (2005).

Method	**Sensitivity**	**Specificity**	**Accuracy**
Pattern Analysis Criteria	68.4%	85.3%	76.8%
ABCD Rule	77.5%	80.4%	79.0%
Menzies Scoring	84.6%	77.7%	81.1%
7-point Checklist	81.4%	73.0%	77.2%

1.4. Problem and Motivation

As stated emphatically with detail in the previous sections, melanoma represents a particularly perilous category of skin cancer, identifiable through visual inspection by trained clinicians. The heterogeneity and irregularity of clinical features in melanoma present significant diagnostic challenges, primarily due to the subjective nature of the visual assessment process. Hence, there exists a compelling imperative for the development of automated detection algorithms, chiefly attributed to this issue of “subjectivity” inherent in human evaluations. The variability in interpretation among examiners, who differ in educational attainment and clinical experience, can lead to inconsistent evaluations of identical lesions, as noted by Haenssle et al. (2018). For a comprehensive analysis of how subjectivity problem affects adult melanoma diagnosis, Dinnes et al. (2018) offer an extensive investigation. Additionally, the diagnostic accuracy is also influenced by the context in which the patient is assessed, with general clinical settings potentially staffed by general practitioners rather than specialists, which could contribute to less accurate evaluations.

A widely adopted measure to improve melanoma detection accuracy is the usage of a dermoscope, a specialized instrument that allows for the magnified examination of skin lesions, as well as the scalp, hair, and nails. The invention of *dermoscopy* can be traced back to 1663 when Kolhaus employed a microscope to study the fine vessels (Katsambas et al., 2015). The methodology underwent significant advancement in 1878 with the introduction of immersion oil, which was applied between the lesion and the lens to enhance the visualization of textures (Senel, 2011). Subsequently, the integration of a light source led to the creation of a specialized microscope tailored for the examination of skin surface lesions or moles, now known as the *dermascope*. Essentially, the dermoscope is a diagnostic tool equipped with a light source and a magnifying lens. This device, typically providing magnifications between 60x and 100x, utilizes a specific gel to enhance the clarity of the lesion's magnified image for subsequent analysis (Katsambas et al., 2015; Senel, 2011). Ultimately, in 2001, 3Gen company emerged as a pioneer with the introduction of a polarized version of the dermoscope, known as DermLite (2001). This innovation marked a turning point, and since its inception, the application of dermoscopy in the detection of melanoma has seen a swift and widespread increase. However, even with dermoscopic examination, Haenssle et al. (2018) report an average sensitivity and specificity of 86.6% and 71.3%, respectively, for melanoma detection by physicians, highlighting the influence of the examiner's experience level on diagnostic outcomes.

Given these findings, it is apparent that reliance on visual assessments can lead to unnecessary medical interventions due to false positives or, conversely, to missed diagnoses with potentially severe consequences, as undetected melanoma may metastasize. This underscores the imperative for the development and integration of sophisticated automated melanoma detection systems. Such systems would provide invaluable support to dermatologists, aiming to bolster the precision and reliability of visual inspections.

1.5. Contributions

In the light of the above, this thesis introduces a novel and computationally efficient framework that synergizes traditional methodologies with the accessibility of state-of-the-art deep learning architectures. This framework is conceived to offer a fresh vantage point within the research arena of melanoma detection. The approach adopts a modular design, characterized by a reduced parameter set and minimal constraints, relying on the Bag of Visual Words (BoVW) concept (Yang et al., 2007) that capitalizes on deep features extracted from a pre-trained neural network. The strategic use of feature extraction from an already trained network bestows the framework with a lightweight and rapid processing capability, circumventing the necessity for protracted deep learning training cycles. This aspect, combined with the framework's modular design, ensures ease of adaptation and updating, thereby positioning our method as a more agile and adaptable option relative to current methodologies. Finally, classification step is executed using the robust technique of support vector machines, which utilize compact, sparse histograms derived from the meticulously assembled BoVW lexicon. Furthermore, our framework is enhanced with two innovative sub-processes tailored specifically for melanoma detection: the refinement of image patches and the application of a bespoke weighting strategy in the creation of histograms. The contributions of this research are specified as follows:

- An innovative methodology for the automated detection of melanoma is described. It combines the time-honored Bag of Visual Words (BoVW) technique with a pre-trained deep learning network.

- Instead of conventional key point descriptors for BoVW, an avant-garde feature extraction strategy utilizing a deep network model is presented. These features are then ingeniously integrated into the BoVW schema, applying K-means clustering for dictionary construction.

- A novel image enhancement algorithm is introduced, that is specifically tailored to augment the clarity and distinction of features extracted from lesion imagery. To the best our knowledge, such an enhancement protocol has not been previously incorporated within a BoVW-oriented framework.

- The framework is robust, yet agile. Its modular architecture not only simplifies the update process but also successfully bypasses challenges associated with prolonged training durations and the burden of heavy computational demands.

1.6. Outline

This thesis is divided into five major chapters. Each chapter and their sections are intended to build on the one before it, resulting in a comprehensive picture of developing an automated melanoma detection algorithm. The following is an outline of each upcoming chapter's structure and key focus.

In Chapter 2 *Literature Review*, an introductory background on the automated melanoma detection domain is initially provided. This is followed by an exploration of the domain's evolution, articulated in two distinct sections. The initial section delves into the array of studies that have employed traditional methodologies for automated melanoma detection, elucidating the algorithmic construction of such traditional approaches specifically tailored to this domain. Subsequently, the final section presents an up-to-date discourse on the state-of-the-art within this sphere, delineating the current status and advancements characterizing the field of automated melanoma detection.

Chapter 3 *Methodology* lays the foundation for the most recent findings of this thesis, tracing the journey from initial methodologies to the innovative concept of integrating traditional techniques with contemporary strategies to develop a high-performing framework for melanoma detection. The sections within this chapter sequentially present each methodology that was examined, detailing the progression and interconnectedness of these approaches to demonstrate the evolutionary pathway leading to the current framework.

Chapter 4 *Experimental Results* provides a comprehensive account of the testing procedures for the newly proposed framework and its comparative analysis against the domain's state-of-the-art. Within its sections, the chapter furnishes meticulous details regarding the dataset employed, the testing conditions, and the comparative outcomes of the framework's performance. Subsequent sections delve into a critical discussion of these results, placing them alongside benchmarks set by competing methodologies. A unique facet of this chapter is the final section, which is devoted to a deeper examination of the framework's predictions. This is conducted through the lens of explainable artificial intelligence (XAI) (Došilović et al., 2018), aiming to shed light on the decision-making processes of the framework, thereby enhancing the interpretability and transparency of its predictive capabilities.

Chapter 5 *Conclusion* is the last chapter and summarizes the key findings, discusses the ramifications for the area, and briefly discusses potentially improvable sections of the framework. It also considers the study's weaknesses and shows possible further research directions for future.

CHAPTER 2: LITERATURE REVIEW

2.1. Background

The development of systems for the automated detection of melanoma involves computational techniques to evaluate skin lesions. These systems are designed to input images of the lesions, analyze them, and then provide a probability score or a categorical determination of whether the lesion has melanoma or not. The inception of such automated systems traces back to 1988, with a range of methodologies having been introduced since that time (Okur and Turkan, 2018), notable examples being the use of artificial neural networks (Marin et al., 2015), decision trees (Zhou and Song, 2013, 2014), and basic thresholding techniques for lesion segmentation (Santy and Joseph, 2015). Initially, the scarcity of data posed significant challenges. The accessible images, which were either photographed using standard cameras or as scanned images of slides, were inadequate in two main ways. Firstly, the quantity of images was too limited to effectively train a diagnostic system that needs to account for the diverse manifestations of the disease. Secondly, the technology of the time was insufficient to accurately identify or discern critical lesion characteristics such as textures and borders. This latter issue was particularly critical, as it impeded accurate diagnosis, even with methods that did not require an extensive training phase.

The introduction of the dermoscope marked a pivotal advancement in the field. Dermoscopy, as outlined in Section 1.4, has empowered dermatologists and visual inspectors to procure images of lesions that are not only well-lit but also significantly magnified. This enhancement has addressed the previously mentioned issue of poor visual quality, enabling the capture of images rich in detail and critical features. Subsequent to the widespread adoption of dermoscopy, the 2000s saw the release of publicly accessible lesion datasets. However, the number of such datasets remains limited, with prominent examples including the PH2, EDRA, DermoFit, HAM10000 and the most importantly the International Skin Imaging Collaboration (ISIC) Archive (Mendonca et al., 2015; Argenziano et al., 2000; The University of Edinburgh, 2013; Tschandl et al., 2018; ISIC Archive, 2022). These repositories have been instrumental in propelling forward research into the domain.

After that, the architectural landscape for automated melanoma detection systems saw considerable diversification. Traditional approaches held the forefront in performance until the surge in application and popularity of convolutional neural networks (CNNs) began to redefine the state of the art. Given that this thesis introduces a framework that integrates a traditional method base with recent enhancements, the subsequent two sections will delve deeply into both the traditional and modern methodologies within this domain separately. This two part detailed exploration aims to provide a clearer context for the framework proposed in this thesis.

2.2. Traditional Approaches

During their peak performance era, designers of traditional melanoma detection systems commonly adhered to an established framework when analyzing dermoscopic images. This framework consisted of three principal stages, each performing the same function across different designs but executed in a manner unique to each system. These stages closely parallel the clinical evaluation processes and lesion characteristics that a seasoned dermatologist would assess during a visual inspection. A block diagram providing a high-level view of this common framework is presented in Figure 2, highlighting the main stages: lesion segmentation, clinical feature extraction, and classification (Mishra and Celebi, 2016). It should be noted that while this approach is widespread among researchers, not all have strictly adhered to or implemented these stages in a rigid sequence; some stages might have been skipped or combined depending on the study.

Following sub-sections will offer a more detailed account of each aforementioned stage, with references to pertinent studies in the field. While the emphasis will be placed on the three core stages depicted in Figure 2, ancillary processes such as “Pre-processing”, “Post-processing” or “Feature generation/selection” will also be discussed in context with the methods employed in the related research.

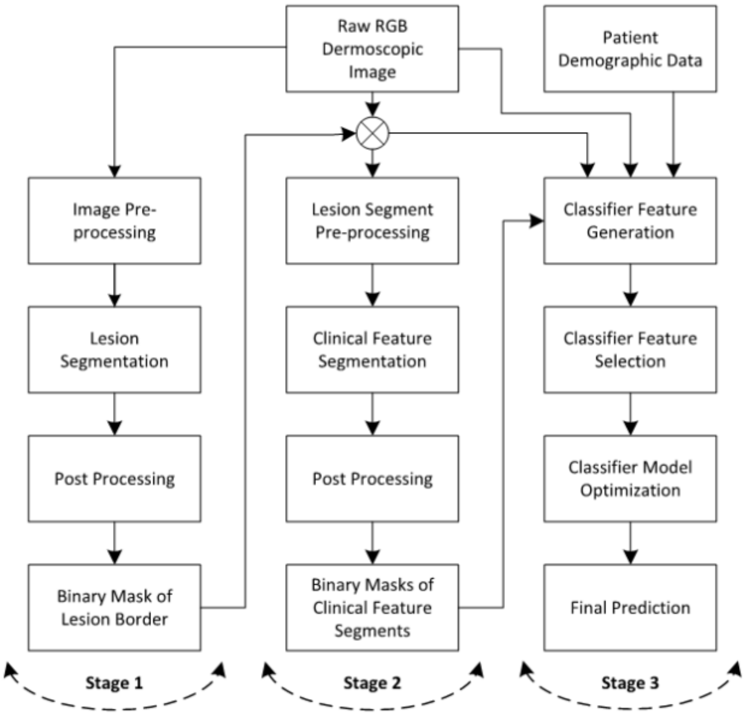

Raw RGB Dermoscopic Image
Patient Demographic Data
Image Pre-processing
Lesion Segment Pre-processing
Classifier Feature Generation
Lesion Segmentation
Clinical Feature Segmentation
Classifier Feature Selection
Post Processing
Post Processing
Classifier Model Optimization
Binary Mask of Lesion Border
Binary Masks of Clinical Feature Segments
Final Prediction
Stage 1
Stage 2
Stage 3

Figure 2. The block diagram illustrating the three principal stages of automated melanoma detection from dermoscopic images (Mishra and Celebi, 2016).

2.2.1. Lesion Segmentation

Traditional system designs for automated melanoma detection begin with the critical stage of lesion segmentation. Despite being conceptually straightforward, its accuracy is foundational for the effectiveness of subsequent processes like clinical feature segmentation and feature extraction for classification. During this phase, the lesion is delineated from the background, which includes the skin and any other extraneous elements. The result is typically a binary image that distinguishes the lesion for further detailed analysis while disregarding the surrounding skin. Successful lesion segmentation is exemplified in Figure 3, where the lesion is clearly isolated from its background. Extraction of clinical features then proceeds exclusively within the confines of the isolated lesion area, enabling the discernment of key global attributes, including the symmetry and the regularity of the lesion’s borders. Conversely, a segmentation approach that is less effective may inadvertently incorporate background pixels into the lesion’s outlined area, particularly around the borders. This can lead to incorrect interpretation of both global and local border features and the extraction of

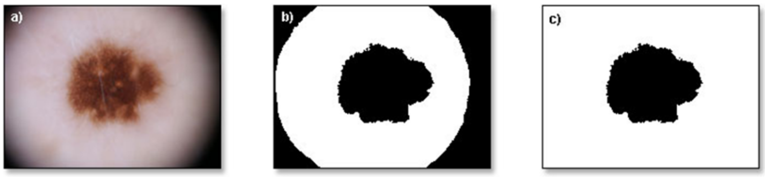


Figure 3. Illustration of an effective lesion segmentation process: (a) the original dermoscopic image, (b) the lesion segmented as a binary mask, and (c) the refined mask post post-processing with the removal of artifacts such as the dermoscope frame (Ogorzalek et al., 2011).

deceitful color features in the subsequent stage of feature extraction. Prior to delving into various systems and methodologies for lesion segmentation, it's imperative to acknowledge two significant challenges in this phase: artifacts in dermoscopic images and the assessment of segmentation quality.

Dermoscopic images often contain a variety of artifacts. Among the most troublesome are those stemming from the inherent constraints of the dermoscopic imaging process itself. Common artifacts include darkened corners, marks from markers, gel bubble effects, color reference charts, ruler demarcations, and hair on the skin (refer to Fig. 4). With the exception of skin hairs, these are typically introduced during the imaging process. Moreover, variations in lighting, along with noise and fluctuating contrast levels, can obscure crucial details in the images. For traditional approaches artifacts "must" be dealt with by removing them as thoroughly as possible either before segmenting the lesion or afterwards. This can be done through pre-processing the original image or post-processing the segmented lesion—or employing both approaches. An effective pre-processing technique for hair removal is the use of software like DullRazor (Lee et al., 1997). Median filtering is another strategy that can aid in noise reduction and image smoothing and can be used either before or after processing to remove hair artifacts (Lee et al., 1997). Additional techniques such as noise filtering, histogram adjustment, color normalization, and contrast enhancement can be integrated into these supportive processes (Quintana et al., 2009; Wight et al., 2011; Abbas et al., 2013). On the post-processing front, methods like region merging (Wong, 2011), border dilation (Iyatomi et al., 2006), and

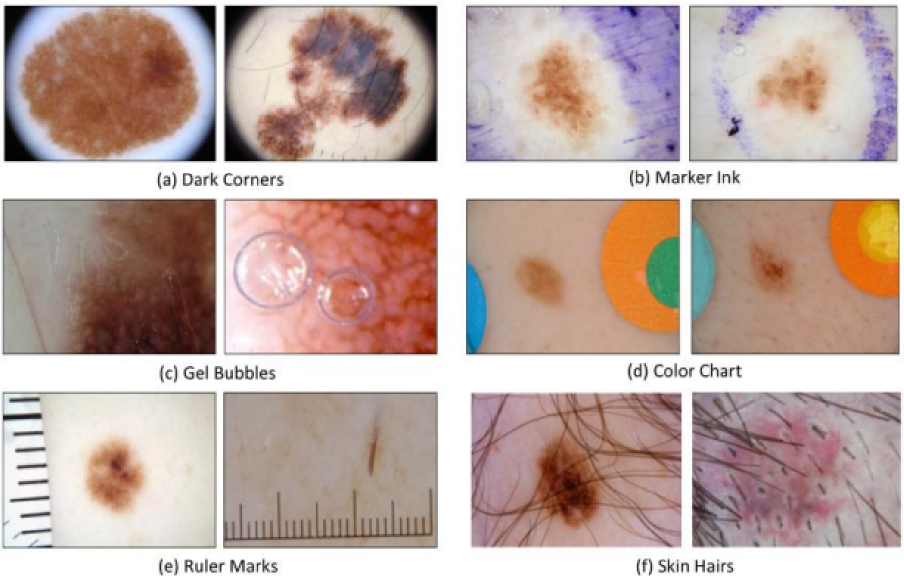


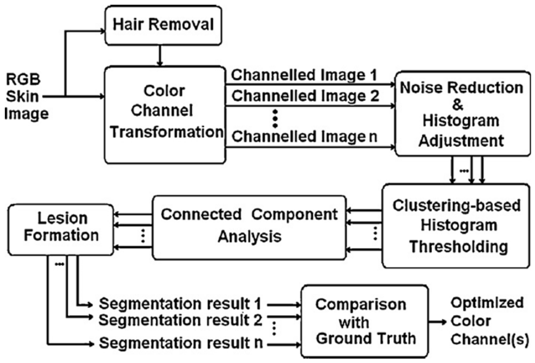
Figure 4. Common imperfections found in dermoscopic photographs include: a) shadowed corners, b) ink from markers, c) air bubbles in gel, d) pigment patches from a color chart, e) measurement lines from a ruler, and f) strands of hair on the skin (Mishra and Celebi, 2016).

smoothing are commonly applied to refine the segmentation outcome.

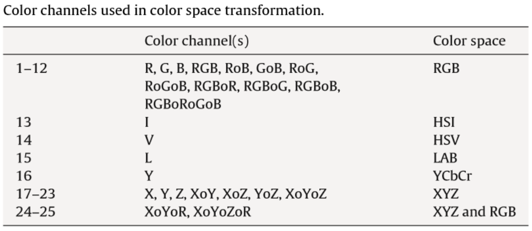
Assessment of segmentation quality is another problem encountered at the lesion segmentation stage. The subjective nature of manual lesion segmentation and its reliance on the expertise of the individual conducting the visual inspection contribute to its inherent challenges. A seemingly optimal strategy for assessment involves comparing manual segmentation against the outcomes of automated techniques. Naturally, to gain a balanced view, manual segmentation should be subject to review across a diverse array of lesions by several experts. For additional insight, one may consult an illustrative study on objective evaluation measures referenced in Celebi et al. (2009).

The literature proposes a variety of techniques for lesion segmentation that also address previously mentioned problems. Notable methods include thresholding (Garnavi et al., 2011; Celebi et al., 2013), clustering (Schmid, 1999; Mete et al., 2011; Melli et al., 2006), fuzzy logic (Baral et al., 2014), supervised learning (Wu et al., 2013), and graph theory (Yuan et al., 2009), with some approaches integrating multiple techniques to enhance segmentation precision (Celebi et al., 2009). Among these techniques thresholding (Sezgin and Sankur, 2004) and its variations are the most widely used.

Thresholding stands out as a straightforward yet prevalent technique in image segmentation and is frequently applied to the segmentation of lesions. It essentially transforms grayscale or color images into binary representations. Numerous studies have adapted thresholding, enhancing it through various modifications or by integrating it with other methodologies for lesion segmentation. Garnavi et al. (2011) devised a hybrid system to delineate lesion borders, combining color optimization with clustering-based histogram thresholding. They scrutinized various color channels across multiple color spaces to achieve a stark contrast between the lesion and the skin. To facilitate this, 30 dermoscopic images were evaluated, with lesions manually encircled by two dermatologists and two dermatology registrars. From this analysis, 25 color channels from six different color spaces —RGB, HSI, HSV, LAB, YCbCr, and XYZ (in combination with RGB)—were examined (Tkalcic and Tasic, 2003), as illustrated in Figure 5.



Hair Removal
RGB Skin Image
Color Channel Transformation
Channelled Image 1
Channelled Image 2
Channelled Image n
Noise Reduction & Histogram Adjustment
Clustering-based Histogram Thresholding
Connected Component Analysis
Lesion Formation
Segmentation result 1
Segmentation result 2
Segmentation result n
Comparison with Ground Truth
Optimized Color Channel(s)



Color channels used in color space transformation.

	Color channel(s)	Color space
1–12	R, G, B, RGB, RoB, GoB, RoG, RoGoB, RGBoR, RGBoG, RGBoB, RGBoRoGoB	RGB
13	I	HSI
14	V	HSV
15	L	LAB
16	Y	YCbCr
17–23	X, Y, Z, XoY, XoZ, YoZ, XoYoZ	XYZ
24–25	XoYoR, XoYoZoR	XYZ and RGB

Figure 5. (Left) The block diagram outlining the color space optimization process integrated with clustering-based histogram thresholding for lesion segmentation, and (Right) the array of color channels employed in the color space transformation (Garnavi et al., 2011).

Subsequently, the segmentation outcomes were benchmarked against the manually defined borders. Following this comparison, the most effective four color channels—X, XoYoR, XoYoZoR (where "o" denotes a logical OR), and R—were selected for a secondary analysis alongside a new composite reference border derived from these four channels. Ultimately, the channels X and XoYoR emerged as the most efficacious and were employed in a two-stage hybrid thresholding process. When the algorithm was tested on an additional set of 85 dermoscopic images, it occasionally surpassed the border marking accuracy of a registrar by 5.3%, using the markings of experienced dermatologists as the ground truth. This study notably highlights the influence of the visual investigator's expertise in the field.

In a different study, Yuksel and Borlu (2009) recommended a thresholding technique that incorporates a type-2 fuzzy logic system. The process begins by transforming the dermoscopic image into a gray-scale version, followed by an analysis of its histogram to determine an optimal threshold level. The procedure starts by selecting a membership function, which is centered at the lowest value of gray-level. This membership function is then systematically shifted along the histogram until a point of maximal ultrafuzziness (Castillo and Melin, 2008) is identified, which is then established as the ideal threshold value. By applying this threshold to the gray-scale image, a binary mask is produced. They evaluated this method against adaptive thresholding and the popular Otsu's method (Nobuyuki, 1979). The authors claim that the Otsu's method tends to understate borders, whereas adaptive thresholding exaggerates them, both introducing artificial irregularities in the border that do not match the true (ir)regularity of the border. This issue is of paramount significance because irregular borders are an indicator of melanoma. They have included some illustrative results, which is presented in Figure 6.

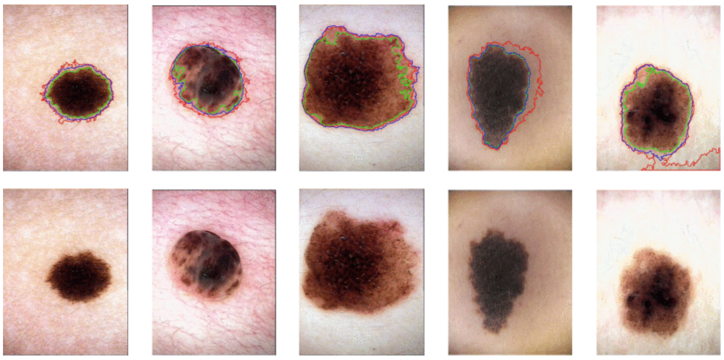


Figure 6. (Lower panel) Source photographs, and (upper panel) outcomes of lesion segmentation: (encircled in red) results from adaptive thresholding, (outlined in green) Otsu's method, and (highlighted in blue) Yuksel and Borlu (2009)'s technique.

In a separate piece of research, Celebi et al. (2013) introduced "Threshold Fusion" as an innovative approach, which employs a combination of various thresholding techniques for the detection of lesion boundaries. They approach the fusion as an optimization problem, aiming to minimize energy, and they construct an ensemble using four established thresholding techniques: the fuzzy similarity method by Huang and Wang (1995), the maximum entropy method by Kapur et al. (1985), the minimum error thresholding method by Kittler and Illingworth (1986), and Otsu's method (Nobuyuki, 1979). The goal of this method is to achieve results that are on par with the top thresholding techniques while remaining unaffected by the peculiarities of different images. They evaluate each thresholding algorithm's effectiveness in border detection, considering the specific attributes of the images used. Acknowledging that Otsu's method may be less precise in certain situations, they pursue a blend of the most efficient thresholding techniques through the threshold fusion strategy proposed by Melgani (2006). Their fusion method demonstrates encouraging outcomes when applied to 90 test dermoscopic images, offering a solution that is both rapid and straightforward to execute. It is also deemed appropriate for images with artifacts such as vessels, skin lines, or fine hairs.

2.2.2. Clinical Feature Extraction

The stage of clinical feature extraction is frequently merged with other stages or, in some instances, completely omitted. The primary rationale for this is its similarity to basic image segmentation problems encountered in the first stage. Once the entire lesion has been segmented in the first stage, typical melanoma cases can be identified by scrutinizing a range of clinical features, which are previously detailed in Section 1.3. The core objective at this stage is either to segment these clinical features further or to detect them.

To automate the recognition of clinical features, various feature extraction methods can be employed. For instance, texture characteristics can be discerned by analyzing the statistics derived from the gray-scale conversion of the input dermoscopic image. Computing a pixel-intensity histogram allows for the estimation of the probability distribution of different pixel intensities (Hayashi et al., 2005). Beyond just considering pixel intensity, statistical methods may also take into account the information from the local neighborhood of pixels (Akram et al., 2015). Additionally, model-based approaches or spatial domain filtering might be used to identify texture features (Jamil et al., 2016).

For extracting shape features, a common practice involves evaluating the lesion's asymmetry by dividing it along its principal axes and comparing the two halves to see if they mirror each other (Ng et al., 2005). With regard to color features, clustering techniques can be utilized to segregate and delineate color classes as defined by dermatologists (Schmid, 1999). One approach might be to count the number of distinct colors present on a lesion and establish a melanoma detection threshold based on this count.

In essence, the primary goal of the clinical feature extraction stage is the precise detection of these aforementioned features. Thus, the methodologies and algorithms applied are very similar to those used in lesion segmentation, with the distinction lying in the output. This stage focuses on segmenting the lesion region into multiple new segments where specific clinical features are isolated. The output from this stage could be a (binary) mask that highlights a feature when superimposed on the lesion area. This mask could then be utilized to extract features for use in the final classification stage. Alternatively, the output could be a simple pixel values correlating to the feature or just a binary indication of the feature's presence or absence.

It's important to note that artifacts present in dermoscopic images may persist into this stage if not completely addressed during the lesion segmentation phase. Therefore, it may be necessary to apply similar pre-processing or post-processing techniques to the feature segmentation as needed.

2.2.3. Classification

In the final stage, the local and global features collected from the previous stages are used in a classification process. This final stage involves synthesizing features based on the border information from lesion segmentation and the feature masks derived from feature segmentation. These generated features are compiled for each sample and serve as input for classification.

Classifiers such as Artificial Neural Networks (ANN) (Priddy and Keller, 2005), Support Vector Machines (SVM) (Steinwart and Christmann, 2008), logistic regression (Tenenhaus et al., 2010), decision trees (Zhou and Song, 2013, 2014), and Bayesian classifiers (Li et al., 2014) are among the most popular choices for traditional approaches to make sense of this data. It is important to recognize that due to the vast and varied range of potential features, it may be beneficial to incorporate an optional feature selection or dimensionality reduction step into the system. This can help reduce computational complexity by selecting the most informative features and discarding redundant or irrelevant ones, thereby streamlining the classification process.

In a methodology advanced by Celebi et al. (2007), subsequent to segmenting the lesion, they proceed to delineate shape features from the lesion's perimeter and partition the lesion into various potential clinical feature zones using the Euclidean Distance Transform. To pinpoint the most pertinent and effective features for the classification task, they apply several feature selection algorithms, including ReliefF (Kononenkoand and Simec, 1995), mutual information-based feature selection (Battiti, 1994), and correlation-based feature selection (Hall, 2000). For the classification itself, they employ a SVM, where they conduct a grid search to fine-tune the hyperparameters of the radial basis function kernel for optimal results (Wang et al., 2004). The efficacy of their system was tested on a collection of 564 dermoscopic images, yielding a specificity of 92.34% and a sensitivity of 93.33%.

Different strategies in this research area have generated notable outcomes by utilizing diverse data sets. For example, Ganster et al. (2001) developed a K Nearest-Neighbor algorithm focusing exclusively on shape and color features. Their system's objective is to classify the identified and extracted features into three categories: benign, dysplastic, and malignant. During the evaluation phase, a dataset consisting of $5,393$ daily clinical lesion images was employed. From the 122 features initially collated and extracted, the feature selection process narrowed this down to 21. The overall results from their experiments indicated a sensitivity of 87% and a specificity of 93% on the dataset used.

In yet another study, Rubegni et al. (2002) trained an ANN on a dataset comprising 588 images, including more than 200 cases of melanoma. They created 48 features categorized into four groups: shape, color, texture, and islands of color. After feature selection, the final 13 features, segmented into shape and color groups, enabled the trained network to achieve a diagnostic accuracy of 94%.

While the studies previously mentioned tend to focus on global feature groups, there are methodologies that incorporate local features as well. Situ et al. (2008) describe a bag-of-features model that employs patches from the segmented lesion region. These patches, measuring 16 × 16 pixels, are systematically extracted using a grid overlay on the lesion. Descriptors for these patches are then developed using wavelet transforms and "Gabor-like" filters (Schmid, 2001). A combination of 10 features is derived from the wavelets and an additional 13 from the Gabor-like filters. These features are then classified using both a Naive Bayes classifier and SVMs. When tested on a set of 100 epiluminescence microscopy skin lesion images—comprising 70 benign lesions and 30 melanomas—the repeated experiments averaged a diagnostic accuracy of up to 82.21% on this test set.

Barata et al. (2014) have also addressed the issue of selecting an appropriate strategy for lesion classification by considering both global and local features. They proposed two distinct systems: the first extracts and classifies global features, while the second identifies regular keypoints on the lesion and extracts local features for classification. For texture features, they analyze gradient amplitude and orientation histograms. For color features, they utilize six different color spaces—RGB, HSV, HSI, LAB, LUV, and Opponent Color Space (Opp) (Bratkova et al., 2009)—characterizing each by concatenating histograms from the three color channels within each color space. The findings from their experiments suggest that both systems with global and local features achieve commendable accuracy levels, yet the local feature-based system has a slight advantage in terms of classification cost-efficiency.

The underlying rationale for using local features is their proven effectiveness in other domains, such as image retrieval (Douik et al., 2016) and object recognition (Guo et al., 2014). Consequently, local features have become increasingly popular for melanoma detection. Despite this, the overall success of such approaches still hinges on the specific methods applied at each stage of the process. Moreover, a direct comparison between most of these traditional methodologies is challenging due to the variation in datasets and image styles used in their evaluations. The forthcoming section is set to explore most recent studies and trends in the domain, shedding light on advanced techniques currently being developed.

2.3. Deep Learning Based Approaches

Automated melanoma detection research commenced nearly three decades ago, gaining significant momentum following the advent of dermoscopy. During this time, an array of solutions has emerged, approaching the challenge from various angles (Okur and Turkan, 2018; Adegun and Viriri, 2021). The most recent methods, particularly the more sophisticated ones, predominantly rely on deep neural networks and their ensemble counterparts. Such reliance brings forth issues related to the considerable time and computational power required. Moreover, these techniques often lack flexibility for updates or retraining when changes or enlargements occur within datasets. Hence, the quest for automated melanoma detection continues unabated in the research community.

One such recent significant contribution to melanoma detection research is outlined by Ain et al. (2019), featuring an inventive application of multi-tree genetic programming coupled with a novel fitness function. This study integrates diverse image feature types such as color and gray-scale local binary patterns, color contrast, and geometric shapes into a sophisticated multi-tree genetic programming framework. This framework is equipped with a set of seven specially selected operations that facilitate crossover and mutation, each operation being meticulously chosen for this specific context. Experimental results of this approach demonstrate its ability to outperform traditional single-tree genetic programming techniques and several widely-used classifiers. These include naive Bayes, K-nearest neighbors, support vector machines, decision trees (J48), random forests, and multilayer perceptrons, all of which were tested using the WEKA Tool (Frank et al., 2005). The data for these investigations were sourced from the Dermofit database (The University of Edinburgh, 2013), containing 1300 images, and the PH2 database (Mendonca et al., 2015), which includes 200 images.

In another notable study by Sharma et al. (2022), an integrated approach is adopted that combines hand-crafted clinical features with convolutional neural network

(CNN) models to enhance melanoma detection for dermatologists. These hand-crafted features are based on color moments—including mean, standard deviation, skewness, and kurtosis—across each color channel and incorporate attributes derived from the gray-level co-occurrence matrix (GLCM) (Singh et al., 2017). The system developed consists of a cascaded, dual-input ensemble deep learning model. The lesion image itself constitutes the first input, which is processed by a CNN. The second input comprises the hand-crafted features extracted from the image, which are then passed through a multilayer perceptron (MLP) that includes four fully connected layers. These two neural network streams converge into a joint fully connected layer, which is followed by a softmax layer responsible for the final classification output. The study's results confirm that the dual-input model outperforms a traditional CNN when both are trained on the HAM10000 (Tschandl et al., 2018) dataset, which consists of 10015 dermoscopic images.

In a very recent study, Ichim et al. (2023) proposed two models of decision support systems. These models are founded on distinct approaches to ensemble neural networks, with the overarching goal being to enhance diagnostic accuracy by consolidating individual decisions from multiple neural network predictions. They have used a dataset that they have assembled using medical images sourced from the HAM10000 (Tschandl et al., 2018) and ISIC (2019) databases. The constructed dataset is consisted four distinct classes: Melanocytic Nevus (NV), Basal Cell Carcinoma (BCC), Benign Keratosis (BKL), and Melanoma (MEL). To address the disproportion in the quantity of NV lesion images compared to the other lesion types, around 20% of the NV images were randomly excluded from the dataset. To compensate for the reduced number and enhance the dataset balance, augmentation techniques were employed for the remaining lesion categories. The first system they have designed utilizes the independent training of three distinct neural networks (MobileNet V2, DenseNet 169, and EfficientNet B2 (Sandler et al., 2018; Huang et al., 2017a; Tan and Le, 2019)) with a weighting system applied to the four lesion categories, influencing the collective prediction outcome. Conversely, the second system comprises a series of six binary models, each corresponding to a different class pairing within each network. Predictive decisions are integrated using a weighted average for each class and model.

Altogether, they claim their system employs 18 binary models. This binary model ensemble achieves a global accuracy of 91.04%.

Although the studies made significant contributions to the field of melanoma detection, direct comparisons or accurately determining their standing within the domain is challenging again, due to the use of different datasets. A solution to this problem emerged with the events like the International Skin Imaging Collaboration (ISIC) Challenges (ISIC, 2022). The first of these challenges, the Skin Lesion Analysis Towards Melanoma Detection (SLATMD) Challenge, took place during the IEEE International Symposium on Biomedical Imaging (ISBI) in 2016 (ISIC, 2016). This annual event has drawn researchers globally, evidenced by the 3300 participants at the most recent ISIC 2020 event. The challenges offer meticulously curated, human-verified training and testing image sets, complete with relevant metadata, and feature thousands of images. Highlighting the ISIC events is essential for an accurate understanding of the current and potential progress in the field.

Initially, the main goal for participants was to differentiate melanoma from benign nevi and other malignancies based on diagnostic accuracy. Later editions introduced new layers to the competition, such as addressing out-of-distribution samples in 2019 and incorporating clinical context in 2020. The datasets provided by ISIC for these challenges, which include a subset from the ISIC Archive (2022), are equipped with both training and test images and their corresponding metadata. The existence of a standardized dataset and evaluation criteria allows for a leader-board at the end of each challenge. This provides a benchmark that researchers in the domain can use to gauge the performance of their own work, even if they did not participate in the challenge itself.

It's important to note that since launching an online evaluation server in 2018, ISIC has decided not to disclose test image data and ground truths, instead releasing data progressively. As of the completion of this thesis, the ISIC 2017 dataset remains the most up-to-date dataset publicly available with ground-truth labels for test images and benchmark segmentation masks. The thesis framework has been tested on the ISIC 2017 data and benchmarked against the performers on the ISIC Challenge 2017 Task 3 leader-board. Hence, remaining part of this section briefly presents the top-5 studies from that leader-board. Task 3 is the “lesion classification only” part of the challenge.

The highest-ranking entry, by Menegola et al. (2017), presented a composite model incorporating seven subsidiary models. Six of these sub-models were based on Inception architectures, with the seventh leveraging ResNet (Szegedy et al., 2015; He et al., 2016a). An SVM classifier layer was employed to integrate the outputs from these sub-models, a process that demanded substantial computational power during the training phase. The team addressed three fundamental challenges related to the efficacy of deep learning approaches: the volume of training data, the complexity of the model architecture, and the need for extensive computational resources. To tackle the issue of data volume, they have combined six datasets including the ISBI-SLATMD 2017 challenge, ISIC Archive, EDRA, PH2, Dermofit Library, and the IRMA Skin Lesion Dataset (which is not publicly listed but available upon request). From this collection, they curated two distinct subsets: the “deploy” set with 9640 images and the “semi” set comprising 7544 images. Their strategy for model development was informed by their prior work (Menegola et al., 2017), which involved utilizing a network pre-trained on the ImageNet database, subsequently refined through fine-tuning for the specific task of lesion classification. Drawing on this experience, they chose to work with more complex models—specifically, ResNet-101 and Inception-v4 (Szegedy et al., 2016)—and proceeded to rigorously evaluate hundreds of models derived from these architectures. Ultimately, they assembled a meta-model that integrates the previously mentioned seven models: three Inception-based models trained on the “deploy” set, another three Inception-based models trained on the “semi” set, and one ResNet-based model also trained on the “semi” set. Additionally, they employed an SVM-based meta-learning layer trained on the validation subset of “deploy”. The team credits their success to the use of deeper network architectures combined with larger datasets, the implementation of data augmentation techniques, per-image normalization procedures (for instance, subtracting the average of the image to enhance Inception-based models), and the strategic fusion of decisions from multiple models, which they argue yields better outcomes than relying on a single optimal model.

Following close behind, the second-place entry by Bi et al. (2017) deployed three variations of ResNet models. Their first model approached the classification challenge as a multi-class problem, distinguishing among melanoma, nevus, and seborrheic keratosis. The second and third models dealt with binary classifications and combined the results of the previous two models into an ensemble, respectively. This team also incorporated additional images from the ISIC Archive to bolster their dataset. They customized the architecture of ResNet by adapting its final layer to contain either 3 or 2 neurons, aligning with the respective number of classes they were targeting. All images in their dataset were resized to have their shorter axis measure 224 pixels. To enhance the model's generalization capabilities, they employed data augmentation techniques that included random cropping and flipping of the images. The training was conducted in batches of 90 images. They claim computational process of training a single ResNet model was carried out over approximately "half a day, utilizing two Titan X GPUs". This is a clear evidence for high computational power needs of CNN training.

In third place, DeVries and Ramachandram (2017) developed a "multi-scale" CNN informed by the Inception architecture, which was pre-trained on the ImageNet dataset and then fine-tuned to process different image resolutions. "Multi-scale" term comes from the fact that, they employ a dual-input strategy, inputting images at two different resolutions to the network. This dual-resolution setup is designed to exploit both the macro and micro features of the lesions: one input provides a coarse-scale image that encompasses the lesion's overall context and shape, enabling the model to grasp the general morphology of the lesion; the other input, at a finer resolution, presents more detailed textural and low-level features that are critical for differentiating between the various lesion classes. By combining insights from both these perspectives, they expected from model to yield a more nuanced and accurate classification. Lastly, they augmented the performance by training multiple models with minor modifications to the original architecture, combining ten such models for their final classification, which also resulted in a considerable demand for computational resources.

The fourth position was secured by Jia and Shen (2017), who utilized a deep CNN with 14 convolutional layers followed by a single fully connected layer. This network, a modified version of VGG-GAP (Zhou et al., 2015), was applied in a two-stage process: first to generate class activation maps and then to classify lesions using these maps. In the first stage, the CNN analyzes the images to generate class activation maps (CAM), which highlight the regions most influential for making a classification decision. Following this, in the second stage, they use the CAM results to identify and crop the significant areas from the original images. These cropped regions, which contain features of interest, are then fed back into the second stage of the CNN. This team also used data augmentation methods to increase the number of samples and they are the only team mentioned here to utilize a pre-processing step for hair removal.

Lastly, occupying the fifth spot, Li and Shen (2017) applied two deep learning methodologies focusing on lesion segmentation, feature extraction, and classification tasks. The two deep learning architectures they introduce are the Lesion Indexing Network (LIN) and the Lesion Feature Network (LFN). The LIN framework is designed to tackle lesion segmentation and classification tasks concurrently. It employs two fully-convolutional residual networks that have been trained on distinct datasets that they have created from originals to output both a segmentation map and a preliminary classification of the lesion. Additionally, they propose a unique component, the lesion indexing calculation unit (LICU), which is used to assess the significance of each pixel in the image with respect to the lesion classification decision. The preliminary classification results are then refined by leveraging the distance map created by the LICU, enhancing the classification's accuracy. The LFN, on the other hand, is a framework aimed specifically at the extraction of dermoscopic features. It is a CNN-based model that has been trained on image patches. These patches are strategically harvested from superpixel masks, which are representations of the image divided into perceptually meaningful atomic regions.

To encapsulate the advancements and research discussed in this section, automated melanoma detection is currently conceptualized predominantly as a binary classification task, typically executed in two primary stages. In the initial stage, within a given dermoscopic image, the aim is to discern and classify individual pixels or clusters of pixels as belonging to either the lesion or non-lesion category. Following this, the second stage involves further classification of those identified as lesion pixels or pixel clusters into malignant (melanoma) or benign (non-melanoma) groupings. Insights from the ISIC challenges and the methodologies developed in response to them suggest a shift in the field towards favoring a two stage approach over the traditional three stage process. It is evident from recent top-performing studies that there is a strong inclination towards integrating some variant of deep learning networks, often pretrained on large datasets, which require extensive computational load and considerable time investment to achieve top-tier results.

Despite the remarkable progress made, there remains room for improvement, and the introduction of innovative algorithms could further advance melanoma detection capabilities. The next chapter of this thesis will delve into the genesis and development of a proposed melanoma detection framework, which aims to address the aforementioned challenges and enhance the effectiveness and efficiency of the detection process.

CHAPTER 3: METHODOLOGY

Developing an automated melanoma detection framework is a complex and evolving process, akin to a journey. For us, it began as an exploratory venture, a proof of concept, where the initial idea of a framework is put to the test using a dataset specifically curated for the task. Based on the outcomes of this initial testing, the foundational concept behind the framework is either refined and developed further, or completely revised.

This chapter meticulously chronicles this journey, laying out the various stages in chronological order and dividing them into distinct sections for clarity and depth of understanding. The following section is an exception to this, which is devoted to describing the dataset prepared for the evaluation of each conceptual iteration of the framework. Subsequent sections are dedicated to the exploration and examination of these different ideas. Each section begins by detailing the idea, setting the stage for what follows. This is succeeded by a discussion of the experiments conducted to test the idea, along with the results obtained from these experiments. The concluding part of each section, aptly titled "Verdict", presents a critical assessment of the idea. It lays out the decision made regarding the viability and potential of the idea based on the experimental results. This could involve advancing the idea to the next stage of development, modifying it in light of new approaches, or discarding it if it proves unfeasible.

3.1. Experimental Dataset

The dataset of 670 dermoscopic images employed for testing the ideas given in this study was sourced from the ISIC Archive. These images, each histopathologically diagnosed as either benign or melanoma lesions, were initially part of a larger set of 3600 images. This collection was then narrowed down to 1643 images, of which 411 were identified as melanoma cases, with the remainder being benign.

However, further refinement of this dataset was necessary. After the exclusion of duplicate samples and those with excessive artifacts, the number of melanoma cases was reduced to 335. To address the issue of class imbalance, an equal number of benign images were selected to match the number of melanoma cases. This deliberate pairing resulted in a balanced dataset comprising a total of 670 samples. This dataset forms the foundational basis for the tests detailed in the subsequent sections of the study. It is crucial to note, however, that this dataset is subject to varied alterations based on the unique requirements of each test. These may involve resizing or pre-processing customized to the experimental circumstances or aims. Each section in which such changes are made, will expressly indicate and describe the alterations.

3.2. Sparse and Redundant Representations Framework

Our research indicated that there were relatively few studies that approached melanoma detection through the lens of sparse and redundant representations, despite the suitability of sparsity-related tools for this type of problem. The rationale for this suitability became evident when examining the nature of the images and the features of interest in melanoma detection.

For instance, the pixels at the border of a lesion that marked the transition from lesion to the background skin were characteristically sparse. This sparsity was in stark contrast to the denser pixel groups found within the lesion itself or in the surrounding skin area. Additionally, the clinical features that were key to identifying and diagnosing melanoma within the lesion area also tended to be sparsely distributed, particularly when compared to the other pixels in the lesion. Therefore, it seemed promising to frame classification of lesions and their clinical features within a sparse representations optimization framework.

Sparse representations involve encapsulating most, if not all, of the information in a signal using a linear combination of a limited number of elements or *atoms* drawn from an overcomplete or redundant basis, often referred to as a *dictionary*. This dictionary comprises a set of atoms, with their count significantly exceeding the dimensionality of the feature space. As a result, any given signal can be represented in an infinite number of ways, but the sparsest representation among these offers valuable insights for a range of signal and image processing applications, as evidenced in various (Elad and Aharon, 2006; Protter and Elad, 2009; Peyre, 2009; Mairal et al., 2008; Mairal et al., 2008;

Bryt and Elad, 2008; Peotta et al., 2006; Fadili et al., 2007; Mairal et al., 2008; Liao and Sapiro, 2008). From a mathematical standpoint, sparse representation involves solving a sparsity-constrained non-convex optimization problem, typically approached through two approximate convex optimization steps: sparse coding, and dictionary update. The process involves iteratively solving these two steps to arrive at a solution. The literature on this subject is rich with a variety of "greedy pursuit algorithms" for sparse coding and diverse "dictionary update techniques". These methodologies are underpinned by robust theoretical foundations and have demonstrated superior performance compared to other signal and image processing tools, as elaborated in Elad (2010). For readers interested in delving deeper into the intricacies of sparse representations and the associated algorithms and techniques, Elad (2010) provides a comprehensive resource.

In the beginning, our very first approach involved leveraging sparse representations by examining the sparsity across different frequencies in dermoscopic images. The fundamental approach was to reconstruct a scaled-down version of a dermoscopic image from its decomposed form at various levels of sparsity. The subsequent subsections of the study delve into the specifics of this method and its impacts in greater detail. The primary objective was to extract feature-rich yet sparse representations from dermoscopic images for the purpose of training an Artificial Neural Network (ANN). This approach aimed to distill the most characteristic and informative parts of a lesion from a dermoscopic image and then apply these extracted features within an ANN framework. The ANN would then be in charge of assessing these characteristics to detect the existence of melanoma.

3.2.1. Decomposition of a Dermoscopic Image

The decomposition process begins with a dictionary and the original image. The first step in this process involves defining a block size, denoted as b. The image is then partitioned into distinct patches, each measuring $b \times b$ pixels. For the purposes of our approach, we have initially selected a block size of 8 which is default for the task. This means that each patch extracted from the original image will be an 8×8 pixel square. To visually represent this step, Figure 7 is provided, illustrating how the image is divided into these smaller patches based on the chosen block size.

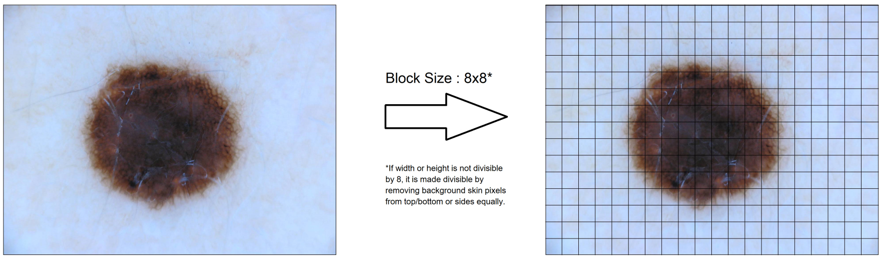

Block Size : 8x8*
*If width or height is not divisible by 8, it is made divisible by removing background skin pixels from top/bottom or sides equally.

Figure 7. The decomposition process involves segmenting the lesion into patches measuring 8 x 8 pixels. This is also done individually for each color channel in the RGB image.

In the process of calculating sparse representation vectors, a Discrete Cosine Transform (DCT) dictionary is employed. The creation of this dictionary follows the methodology outlined in Elad's book (Elad, 2010). A key element in this process is the selection of the number of atoms in the dictionary. This number is predetermined as 1024 by default. Subsequently, the DCT dictionary and the segmented image patches are used together in a specific equation, as depicted in the Figure 8 below. This equation derives sparse vectors that effectively represent each image patch in the calculation.

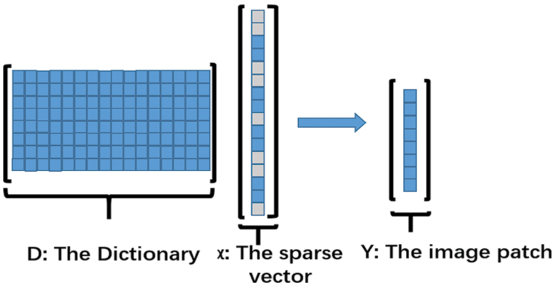


Figure 8. Sparse vectors, each of size 1024 × 1, are computed to represent each image patch. This computation is based on the equation $Dx = y$ and utilizes the Orthogonal Matching Pursuit (OMP) algorithm (Elad, 2010).

In the resultant sparse vectors, each element represents a sparse coefficient corresponding to its respective image patch. These coefficients are arranged in accordance with the frequencies of interest. To construct the decomposed image for a specific frequency, the sparse coefficient with the same index is extracted from each vector and placed back into the spatial location corresponding to the original patch it represents. This process is visually depicted for enhanced clarity using color coding in Figure 9.

Consequently, from this method, a total of 1024 decomposed images are generated, all derived from the sparse coefficients obtained from a single RGB lesion image. An illustrative example of the outcome for two different lesions post-decomposition is presented in Figure 10. This figure showcases the results following the decomposition process.

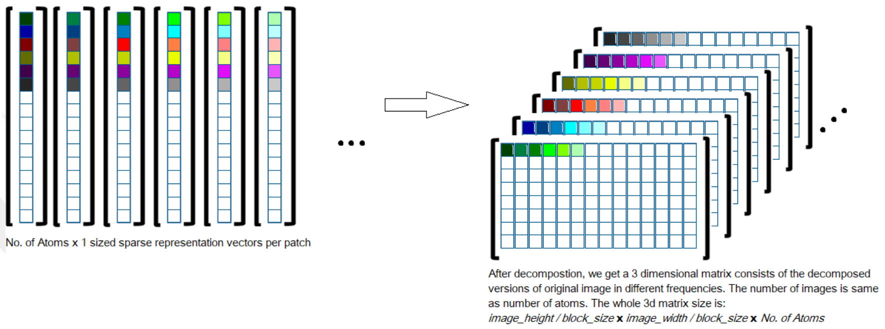

No. of Atoms x 1 sized sparse representation vectors per patch
After decompostion, we get a 3 dimensional matrix consists of the decomposed versions of original image in different frequencies. The number of images is same as number of atoms. The whole 3d matrix size is:
image_height / block_size x image_width / block_size x No. of Atoms

Figure 9. In the representation, each color signifies a distinct frequency, and its shades indicate varying sparse coefficient values. For clarity in the illustration, patches arranged column-wise in the actual implementation are depicted as rows.

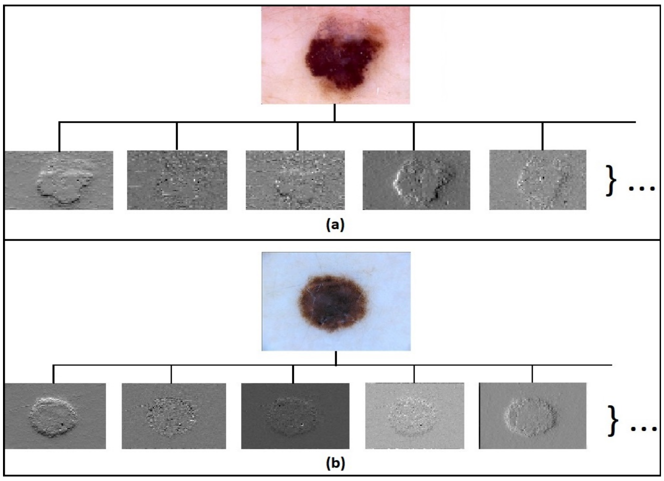


Figure 10. The outcomes of the decomposition process applied to two images are presented, where (a) represents a melanoma case and (b) depicts a benign lesion. For the sake of clarity only five of the total 1024 generated decomposed images are shown.

In the decomposed images, potential indicators of melanoma may exist, yet identifying the specific frequencies that contain these indicators presents a separate challenge. Stemming from this observation, a new way for utilizing decomposed images has been developed. This involves extracting eight new secondary features from each of the 1024 decompositions of a single lesion image. The features obtained from this process are then aggregated to form a comprehensive feature vector, effectively encapsulating a detailed descriptor of that particular image. The secondary features can be described as follows:

- **Energy of Sparse Representation Error** The rationale for employing the energy of sparse representation error stems from the possibility that benign and melanoma cases might differ in how well they can be represented. It remains an open question whether the sparse representations of a lesion's image across various frequencies exhibit discernible differences between melanoma and benign cases. To explore this, the energy levels of these frequencies can be utilized as a feature to investigate and potentially identify any such differences.

$$m_1 = \frac{||x_i||_2}{n} \tag{1}$$

- $l1$ **- norm on Sparse Codes**

$$m_2 = \frac{||x_i||_1}{||x_i||_2} \tag{2}$$

- **Entropy of Sparse Codes** The entropy of sparse codes is considered a viable feature because it is understood that an increase in entropy in a signal corresponds to an increase in uncertainty. This principle is observed in various contexts, such as with ECG signals, where heightened entropy can signify abnormalities. Applying this concept to melanoma detection, a similar inference can be made: an increase in the entropy of sparse codes of a lesion's image might indicate irregularities within the lesion.

$$m_3 = - < \frac{|x_i|}{\|x_i\|_1}, log(\frac{|x_i|}{\|x_i\|_1}) > \tag{3}$$

- **Mean**

$$m_4 = \frac{\sum_1^n |x_i|}{n} \tag{4}$$

- **Variance**

$$m_5 = \frac{\sum_1^n (|x_i| - \mu)^2}{n-1} \tag{5}$$

- **Skewness**

$$m_6 = \frac{1}{n} \sum_1^n [\frac{(|x_i| - \mu)}{\sigma}]^3 \tag{6}$$

- **Kurtosis** Kurtosis, which measures the deviation of a distribution from a normal distribution, is another metric that warrants examination for its effectiveness in melanoma detection. Similar to entropy, both skewness and kurtosis might reveal irregularities in melanoma cases.

$$m_7 = \frac{1}{n}\sum_{1}^{n}[\frac{(|x_i| - \mu)}{\sigma}]^4 \tag{7}$$

- $l0$ **- norm on Sparse Codes** The use of a DCT dictionary in image representation may lead to differing levels of representational accuracy between benign and melanoma cases. If one of these cases (either benign or melanoma) is represented more effectively with the DCT dictionary, the sparsity level of the less effectively represented case will be comparatively lower. In such a scenario, the non-zero elements in the sparse vectors become significant indicators. Here, n represents the total number of elements in the sparse vector, while n_z refers to the count of non-zero elements within that vector. Therefore, a higher number of non-zero elements could indicate a lower level of sparsity, which in turn might be indicative of the nature of the lesion (benign or melanoma) based on its representational accuracy with the DCT dictionary.

$$m_8 = (n - n_z)/n \tag{8}$$

For each individual decomposition out of the total 1024, the aforementioned features are extracted. When this extraction process is applied across all decompositions, it results in a feature vector of size 8192×1, representing a single sample. This comprehensive process of feature extraction and assembly into a large vector is depicted in the accompanying Figure 11.

Subsequently, to evaluate the effectiveness of these features, a simple pattern recognition ANN is trained.

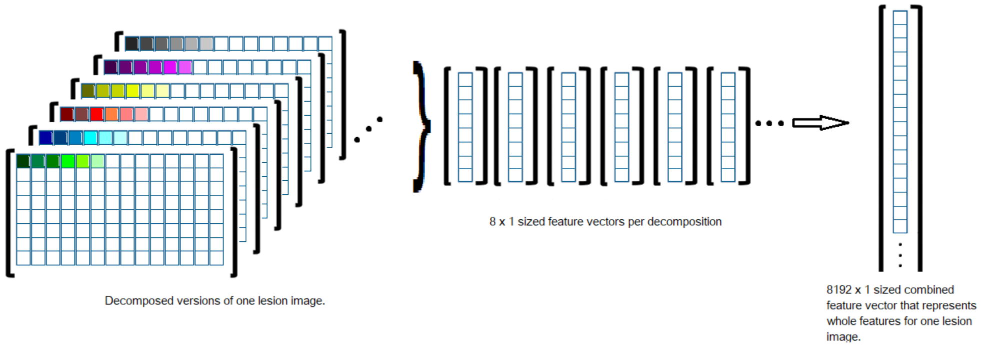

Decomposed versions of one lesion image.
8 x 1 sized feature vectors per decomposition
8192 x 1 sized combined feature vector that represents whole features for one lesion image.

Figure 11. Once the eight features are extracted from a single decomposition, they are concatenated to form a unified feature vector, which has a size of 8192 × 1.

3.2.2. Experiments

Utilizing MATLAB's (The MathWorks Inc., 2022) neural network toolkit, a basic pattern recognition network is trained with the extracted features. This network comprises a single hidden layer, with the number of neurons in this layer set to 5463, which is approximately two-thirds of the total of input and output neurons. The dataset is split in such a way that 70% of the samples are used for training, 15% for testing, and the remaining 15% for validation.

Due to the high resolution of the images, which ranged from 900 × 600 to 4440 × 6666 pixels, a resizing step was necessary before initiating the experiments. The decomposition step became increasingly time-consuming for images with resolutions exceeding 1024 × 768 pixels. Particularly for images larger than 2000 × 1500 pixels, the decomposition of a single image could take almost half a day. To expedite this process, the images were downscaled to near-minimum resolutions while maintaining their aspect ratio.

The training process was repeated 10 times to obtain a reliable evaluation, resulting in an average test accuracy of 56.92%. The average training accuracy was slightly higher, recorded at 62.11%. The detailed results of all the tests conducted are compiled and presented in a Table 8.

Table 8. Training accuracy and test accuracy for each test.

	Training Accuracy	Test Accuracy
Test # 1	76.1%	55.4%
Test # 2	57.5%	54.5%
Test # 3	59.0%	58.4%
Test # 4	64.3%	62.4%
Test # 5	60.0%	59.4%
Test # 6	54.1%	55.4%
Test # 7	56.2%	57.4%
Test # 8	66.5%	51.5%
Test # 9	63.7%	59.4%
Test # 10	63.7%	55.4%
Average	62.1%	56.9%

3.2.3. Verdict

Several key observations emerged from the results of the experiment. Firstly, the accuracy was found to be nearly equivalent to a random guess, akin to a coin toss. Several factors might contribute to this outcome. One possibility is the sheer number of decompositions per image, with a majority of the 1024 decompositions potentially lacking any significant indicators of melanoma. Conceptualizing these decompositions as various “layers” of a lesion, it becomes apparent that many might not contribute meaningful information, leading the network either to learn ineffectively or to be guided towards incorrect conclusions.

Another critical factor could be the structure of the feature set used in the ANN. Although the total number of features was 8192, these essentially comprised groups of 8 features repeated across the decompositions. This redundancy in features might not have provided the diverse and distinct information necessary for effective learning.

Given the overall low accuracy of this approach, it has been decided to abandon this method. However, the insights gained from this experiment were valuable. Building on this learning, a new approach is devised, focusing on the detection of key points within the lesion for feature extraction.

3.3. The Scale-invariant Feature Transform (SIFT) Framework

The next idea in the development of the melanoma detection framework involved conducting a literature review focused on key point detection in images. Through this review, the Scale-Invariant Feature Transform (SIFT) algorithm, as described by David (1999), emerged as a potential candidate. SIFT is renowned for its ability to detect and describe local features in images. It has been extensively applied in the field of object recognition, as evidenced by various studies (Saeed et al., 2018; Li and Wang, 2018; Deshmukh and Bhosle, 2016). The strength of SIFT lies in its invariance to location, scale, and rotational changes, coupled with robustness against affine transformations and variations in illumination. Given that the task of melanoma detection entails the identification of clinical features on lesions, the SIFT algorithm presents itself as a potentially effective tool for pinpointing critical points on a lesion. The forthcoming sub-section is dedicated to providing an in-depth exploration of the SIFT algorithm, including a discussion on how it can be applied specifically to the context of melanoma detection.

3.3.1. SIFT Algorithm

The Scale-Invariant Feature Transform (SIFT) algorithm plays a crucial role in detecting and describing local features within digital images. Fundamentally, SIFT operates by identifying a set of key points within an image and then calculating specific information based on the neighboring pixels around each key point, or directly from the pixel constituting the key point itself.

Once this information is computed, it is attributed to the corresponding key point, resulting in the creation of what are known as descriptors. These descriptors are then utilized in tasks such as object recognition or other similar applications. A significant attribute of these descriptors is their invariance to a variety of transformations. This quality is particularly valuable because it ensures the reliability of the descriptors even when the object of interest appears significantly altered due to changes in perspective, scale, rotation, or lighting conditions. The original SIFT algorithm, which is specifically designed for grayscale digital images, is described in further detail below.

The SIFT algorithm commences with the application of bilinear interpolation to expand the original image, doubling its width and height. Subsequently, a scale space is constructed for the image. This scale space is a conceptual tool that simulates how the image appears at various scales, effectively representing the image at different levels of detail. To create this scale space, the image is convolved repeatedly, progressively reducing its size with each convolution, until it becomes too small to proceed further. The next step involves the identification of candidate key points within this scale space. Conceptually, each image in the scale space can be thought of as a three-dimensional continuum, with the two spatial dimensions being the x and y coordinates of the pixels, and the third dimension corresponding to the standard deviation of the convolution.

At this stage, the scale-space function comes into play. Its role is to allocate gray values to every point within this three-dimensional space. Ideally, calculating the Laplacian of this function and identifying its extrema would yield the desired candidate key points. However, given the necessity to operate in a discrete approximation of this continuous space, the algorithm employs the Difference of Gaussians (DoG) technique as an alternative (Kamaladhas and Abitha, 2012). This technique involves subtracting a blurred version of the image from another less blurred version, effectively filtering out all but a few spatial frequencies present in the original grayscale image. The discrete extrema found in these difference images are then used as reasonable approximations of the Laplacian extrema, thereby identifying the candidate key points for further analysis.

In the subsequent phase of the SIFT algorithm, each key point is assigned a reference orientation, if feasible. There are instances where this step is not possible for certain key points, such as when a key point is located at the image border and lacks a sufficient number of neighboring points, or when a key point does not exhibit a dominant orientation. Key points that fall into these categories are discarded. Conversely, some key points may display more than one dominant orientation. In such cases, these key points are represented multiple times, once for each distinct orientation. The process of assigning reference orientations roughly entails examining the gradients in the immediate vicinity of a point to determine if they share a relatively similar direction.

The final step involves computing descriptors for the key points identified in the previous steps. For each key point, the algorithm creates a histogram based on the distribution of gradient directions within its surrounding area. This surrounding area is conceptualized as a circular neighborhood, and to align with the reference orientation, the coordinate system of this neighborhood is rotated accordingly. A total of sixteen histograms are computed in this manner. Consequently, for each key point, a descriptor vector of size 128 (derived from $4 \times 4 \times 8 = 128$) is generated.

For a more detailed understanding of this process, additional information and explanations can be found in the works of David (1999) and Weitz (2016), which delve deeper into the mechanics and applications of the SIFT algorithm.

3.3.2. Using SIFT in Melanoma Detection

The SIFT algorithm inherently focuses on extracting key points from a grayscale digital image and assigning descriptors to each of these points. However, to effectively adapt and employ this algorithm as a feature extractor for automated melanoma detection task, several challenges needed to be initially addressed.

The primary challenge in adapting the SIFT algorithm for automated melanoma detection lies in preserving the integrity of the data. The original SIFT algorithm is designed for grayscale images, meaning that applying it directly to our dataset of 670 colored images, as mentioned earlier, would result in the loss of valuable color information. Previous sections 1.3 have demonstrated the significance of color data in detecting melanoma, highlighting the need to retain this information in the analysis. To address this challenge, various implementations of SIFT that can handle color images were explored. A particularly useful discovery was the comprehensive comparison of different color descriptors and their invariance properties conducted by van de Sande et al. (2010). Their study methodically examined the uniqueness of color descriptors, providing insights into their performance under various conditions. One standout variant identified through this research is the Transformed Colored SIFT. This version of SIFT normalizes each color channel (RGB) individually before computing the SIFT descriptors for each channel. As a result, the descriptors generated are invariant not just to scale and shift, but also to changes in light color and light color shift. The key distinction of this approach, aside from its compatibility with colored images, is that the descriptors are larger in size, with a dimension of $3(channels) \times 128 = 384$. This expanded descriptor size ensures that no valuable color data is lost in the process, potentially enhancing the robustness and effectiveness of the automated melanoma detection system.

The second challenge in applying the SIFT algorithm for melanoma detection arises from the variable image resolutions within the dataset, leading to a varying number of key points detected by the algorithm for each image. For example, an image labeled as benign might yield 100 key points and their corresponding descriptors, whereas a malignant labeled image could result in 1000 key points and descriptors, or the situation could be reversed. It's important to note that this variability in key point detection is not solely dependent on image resolution; images of the same resolution can also yield differing numbers of key points. This inconsistency presents a two-fold problem. One potential solution could be to impose a limit on the number of key points detected by the algorithm. However, determining the appropriate threshold for this limit poses its own challenge. In our experiments, we observed a wide range in the number of key points detected, with the lowest being 7 and the highest exceeding 12000 in a single image. This leads to another critical issue. As previously discussed in 1.3, melanoma detection relies on the presence of various features on a lesion. However, not every part of a lesion may exhibit signs indicative of melanoma. Therefore, arbitrarily limiting the number of key points could potentially exclude significant areas of the lesion that contain crucial diagnostic information. This situation necessitates a careful approach to ensure that the key point selection process is both efficient and comprehensive, capturing all relevant aspects of the lesion without being overwhelmed by excessive data.

The discussion thus far highlights a third challenge in the application of the SIFT algorithm for melanoma detection: the labeling of data. Our dataset comprises 670 cases, evenly split between benign and malignant melanoma diagnoses, each confirmed through histopathological methods. However, the transformed colored SIFT algorithm identifies a variable number of key points from each image, and these key points are not inherently labeled. Simply assigning the label of the whole image (e.g., melanoma) to all key points extracted from it is not accurate, as not all key points may be representative of the melanoma-indicative region. To effectively utilize our existing labels, the key points and their descriptors must collectively offer meaningful or distinguishing information about the image from which they were extracted. It is essential that these features, in aggregate, accurately characterize the nature of the lesion.

Addressing the last two challenges required a well-thought-out strategy. The following section details a proposed pipeline designed to ensure that the key points can meaningfully represent the images they originate from, thereby providing a viable solution to these challenges in the context of automated melanoma detection.

3.3.3. Bag of Visual Words Pipeline

The Bag of Visual Words (BoVW) is a renowned method for image classification, drawing inspiration from the Bag of Words (BoW) concept in Natural Language Processing (NLP) (Shekhar and Jawahar, 2012; Chen et al., 2017; Malpani et al., 2016; Zhang et al., 2010). In the BoW model, the frequency of each word in a document is counted, and these frequencies are then used to identify potential keywords that characterize the document. A frequency histogram constructed from these word counts serves as a descriptive representation of the document. Using these histograms, specific types of documents can be classified, with each document essentially being treated as a "bag" of words. Similarly, BoVW applies this concept to image features, treating them as the "words" of the image. In the context of lesion images and for this approach, these features are the Transformed Colored SIFT key points and their descriptors. By creating a visual dictionary from these extracted key points and descriptors, a feature frequency histogram can be generated for each image. This histogram can then be instrumental in predicting the class of the image, whether it is benign or malignant.

The entire process of BoVW, from feature extraction to classification, is outlined in the pipeline presented in Figure 12. This figure visually depicts how the BoVW method is adapted and applied to the task of melanoma detection, highlighting each step involved in transforming image features into a useful format for classification.

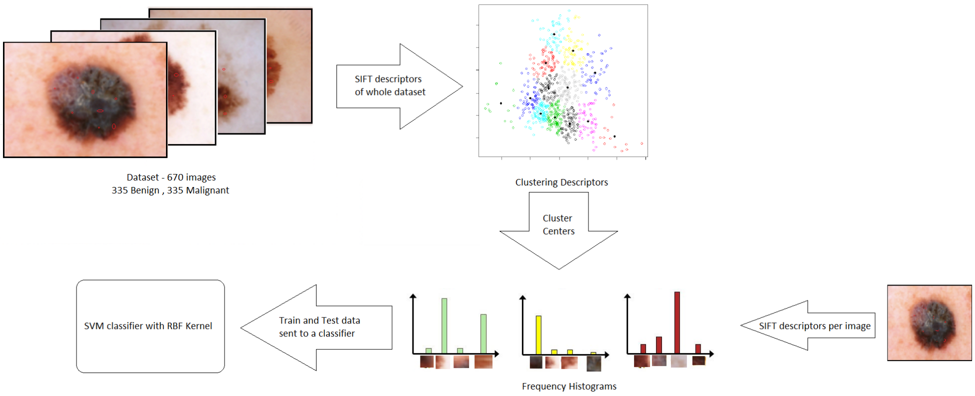


Figure 12. Pipeline of BoVW with SIFT descriptors.

The BoVW pipeline operates through a series of structured steps, beginning with the extraction of key points and their descriptors from the entire dataset. These extracted features serve as the basis for what will become the visual dictionary, representing the whole dataset. The next phase involves clustering these descriptors. This is achieved using a clustering algorithm, which can be chosen based on specific requirements or preferences (Rui and Wunsch, 2005). For this pipeline, K-Means algorithm is used. The clusters formed through this process are integral to the model, with the centroid of each cluster becoming an element of the visual dictionary. These centroids effectively encapsulate the core characteristics of the various groups of features present in the dataset. Following the formation of the visual dictionary, key points and their descriptors are then extracted from each image individually. For each image, a frequency histogram is created based on how its features align with the cluster centers in the visual dictionary. This histogram is a critical component of the model, as it provides a quantifiable representation of the image in terms of the established visual dictionary. Lastly, the classification of an image is determined by a simple SVM classifier with RBF kernel.

3.3.4. Experiments

The experimental procedures followed the outlined pipeline, utilizing MATLAB 2019b (The MathWorks Inc., 2022) for generating frequency histograms and performing the final classification step. However, due to concerns about memory efficiency, Anaconda (Anaconda Inc., 2020) was employed for the clustering process. From the dataset of 670 images, over a million key points and their descriptors were extracted. The Mini Batch K-Means clustering implementation in the Sci-Kit Learn package (Pedregosa et al., 2011) within Anaconda was particularly useful. It features a “partial fit” method that allows for data to be inputted in variably-sized batches for clustering, accommodating the large volume of data efficiently.

In addressing one of the previously mentioned challenges about the variability in the number of key points extracted from each image, a normalization of the frequency histogram for each image is employed. Two different normalization approaches were tested, with the results of both approaches evaluated. The first approach involved using the $l1$-norm, where the frequency histogram is divided by the total number of key points extracted from the corresponding image. The second approach used a similar method but involved dividing each element in the histogram by the $l2$-norm of the histogram itself. These normalization techniques aimed to standardize the histograms, making them comparable across images regardless of the number of key points each image initially produced.

In implementing this pipeline, a new challenge related to determining the optimal number of clusters is emerged. The inherent uncertainty in identifying the precise number of clusters present in the dataset necessitates experimentation with various options. For the current phase of the framework, a few potential cluster numbers are tested. The outcomes of these tests are detailed in the results presented in Table 9 below.

Table 9. Results from the BoVW with SIFT descriptors framework initial experiments.

Normalization Method (right) No. of Clusters (below)	l1 Norm	l2 Norm
K = 10	50.7%	51.7%
K = 100	54.5%	58.2%
K = 200	52.5%	56.75%

3.3.5. Verdict

The experiments conducted and their subsequent results have yielded valuable insights into melanoma detection. However, they have also demonstrated that the SIFT-based approach alone is insufficient for effectively differentiating melanoma from benign lesions. Despite this, the implemented Bag of Visual Words (BoVW) pipeline, with its modular structure comprising four distinct steps, emerges as a promising candidate for the backbone of the final framework.

The modular nature of the BoVW pipeline, encompassing feature extraction, clustering, histogram generation, and classification, offers significant flexibility. Each of these steps can be individually improved or replaced as needed, without necessitating a complete overhaul of the entire pipeline. This modular design allows for targeted enhancements at specific stages of the melanoma detection process. For instance, the results indicated that SIFT descriptors lack the necessary distinctiveness to adequately separate melanoma cases from benign ones. Consequently, we can explore new methods only for feature extraction that yield more discriminative feature vectors suitable for next step, clustering. Likewise, the clustering step itself presents opportunities for refinement. Smart decisions regarding the number of clusters could be made, perhaps by initially clustering data from benign and malignant lesions separately to enhance the uniqueness of the visual dictionary's cluster centers. Alternatively, the introduction of a new clustering algorithm altogether might offer improvements.

In the case of SIFT descriptors, the relatively low accuracies observed may stem from their inherent calculation process. As previously explained, SIFT descriptors are derived from the pixels surrounding a key point, and these key points are identified based on major pixel variations on the image. Such variations can occur not only in clinical features of interest but also along lesion borders and due to imperfections in the background skin. Consequently, some descriptors might not carry any relevant information about melanoma or the lesion itself. Given this limitation and the resulting low accuracies, the use of SIFT descriptor features has been discontinued.

In summary, the exploration and experiments with the SIFT descriptors led to their abandonment. However, this process has been instrumental in the introduction of the new Bag of Visual Words pipeline. The approach following this will focus on enhancing the feature extraction step. It's also crucial to note the importance of normalizing the histograms generated in this pipeline. This normalization is essential due to the varying total number of features represented in each image's histogram. The results in Table 9 indicate that using the $l2$-norm for histogram normalization yields better results. Therefore, moving forward, the $l2$-norm will be the standard approach for normalization in our ongoing development of the BoVW pipeline.

3.4. BoVW with Deep Neural Network Features

In this revised approach to the previous melanoma detection framework, the feature extraction step has undergone a significant transformation. Instead of relying on the SIFT descriptors, the new strategy involves extracting descriptive feature vectors from the activations of pre-trained, well-established deep learning networks. This modification has led to an updated framework, as depicted in Figure 13.

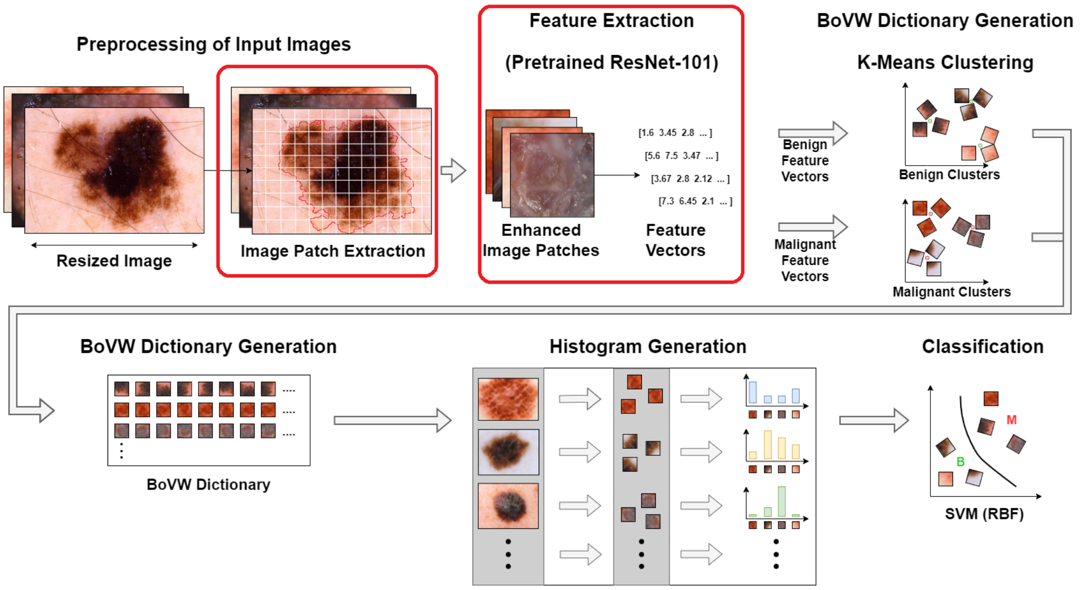


Figure 13. BoVW approach with updated feature extraction and new pre-processing step.

For this proof of concept, a selection of 10 pre-trained models has been chosen: AlexNet, GoogleNet, Inceptionv3, Inception-ResNetv2, Resnet101, VGG19, DarkNet, DenseNet201, ResNet50, and Xception (Krizhevsky et al., 2012; Szegedy et al., 2015, 2016, 2017; He et al., 2016b; Simonyan and Zisserman, 2014; Redmon, 2013; Huang et al., 2017b; Chollet, 2016). These models were selected based on their proven efficacy in object detection and medical image processing tasks, coupled with their availability for direct usage after minor adaptations to our specific requirements. To tailor these pre-trained models to the task of melanoma detection, a few modifications were made, including the introduction of a new pre-processing step. This step not only helps to align the models with the specific needs of our task but also alters the manner in which we represent a single lesion image within the framework.

The integration of the new pre-processing step into the melanoma detection framework is necessitated by two key reasons. Firstly, the various network models selected for this approach each require different input sizes. Directly resizing the original images to these varying dimensions would result in significant data loss. Secondly, the constraint posed by the size of the sample dataset, which comprises only 670 lesion images, also presents a challenge. Utilizing a single feature vector per image would limit the ability to discern distinctive features crucial for accurate melanoma detection. To address these challenges, a strategy was devised where images from the dataset are segmented into multiple overlapping patches with sliding window method, each corresponding to the input dimensions required by the selected deep learning model. These patches are then fed individually into the model, and the feature vector that characterizes each patch is extracted from just before the classification layer of the model.

Consequently, every image in the dataset is transformed into an assemblage of feature vectors, each representing a different segment of the image. This approach bears similarity to the previous SIFT-based method, but it leverages the feature extraction capabilities of deep learning models. Major difference is instead of focusing on specific points on the image, the entire image is utilized. By breaking down each image into multiple patches and analyzing them separately, the framework gains a more granular and comprehensive understanding of the image, potentially enhancing the accuracy of melanoma detection.

3.4.1. Experiments

Different number of experiments were conducted in alignment with the previously outlined workflow. MATLAB 2020b was utilized for generating frequency histograms and for the final classification stage. To maintain memory efficiency, Anaconda was used for the clustering process again. In order to be able to explore more information in a rather quick succession, a smaller dataset is generated from the 670 images previously mentioned in Section 3.1. With that, from the dataset of 670 images, two distinct datasets were created. The first dataset, labeled “full scale”, included each image from the original set. The second dataset, labeled “small scale”, contained a subset of 142 images, selected to expedite the experimental process and allow for quicker iterations. The newly implemented pre-processing step, which involves generating a high number of patches from each original image, ensures that even with the "small scale" dataset, meaningful results can be obtained. It’s important to note before going into the experiments, for both datasets, 70% of the images from each dataset were used for training the models, while the remaining 30% of the images were set aside for testing.

Moreover, two different approaches were employed in the clustering phase, resulting in two variations of the same experiment. In the first approach, all extracted feature vectors were inputted into the K-Means clustering algorithm, and the resulting cluster centers were used as they were. In the second approach, the feature vectors from benign and malignant cases were clustered separately within their respective classes. The cluster centers obtained from these two groups were then combined and utilized together. The rationale behind experimenting with separate clustering for benign and malignant feature vectors, as opposed to clustering them all together, stems from a concern regarding the distinctiveness of the cluster centers. It is hypothesized that clustering all the data together might result in cluster centers that fall into a ’gray area’ between benign and melanoma characteristics. This could potentially diminish the distinctiveness of these centers.

The first experiment was designed to identify which network model yields the most valuable feature vectors. This experiment utilized the “small scale” dataset, encompassing the entire set of 142 images, and was conducted using both of the clustering methods with $K = 100$. The results of this experiment are shown in the Table 10.

Table 10. Initial test results for identifying the most effective network model to extract feature vectors.

Network	**Separate Clustering Accuracy**	**Normal Clustering Accuracy**
AlexNet	59.52%	56.66%
DarkNet	63.19%	59.52%
DenseNet-201	63.80%	59.04%
GoogleNet	65.28%	62.38%
Inception-V3	65.00%	60.80%
Inception-ResNet-v2	70.04%	66.28%
ResNet-50	69.16%	61.90%
ResNet-101	**78.57%**	**67.00%**
VGG19	66.67%	61.90%
Xception	63.67%	59.52%

The results shed light to crucial knowledge beyond choosing the appropriate network model. First of all, the best results came by far from the ResNet-101 model with 78.57% accuracy. Following that, all networks yielded better results if separate clustering is used which proves the hypothesis mentioned previously. Additionally on a side note, it can be deducted that the network models that includes residual layers have an advantage in this framework and yield better results in general.

Following the initial experiment, a subsequent experiment was conducted to investigate whether combining other high-performing feature vectors with those from ResNet-101 could enhance accuracy. Specifically, feature vectors from models that achieved an accuracy of over 65.00% were appended to the ResNet-101 feature vectors. This combined set of vectors was then tested within the framework. This experiment was carried out using the "small scale" dataset again and focused exclusively on the separate clustering approach. Before exploring the results though, it's important to note that the experiments conducted up to this point were primarily for proof of concept, and as such, accuracy was the sole metric used for evaluation. However, moving forward, the experiments aimed at refining the framework will include a broader range of metrics namely “Accuracy, Sensitivity, Specificity, and Precision” to provide more comprehensive insights. The results detailed in Table 11, revealed that while the combination with GoogleNet’s vectors came close, none of the combined feature vector sets outperformed the ResNet-101 vectors when used alone.

Table 11. Combinations of feature vectors and their effects on the results.

Network	Accuracy	Sensitivity	Specificity	Precision
ResNet101 + AlexNet	71.4285%	85.7142%	57.1428%	66.6667%
ResNet101 + GoogleNet	76.1904%	76.1904%	76.1904%	76.1904%
ResNet101 + Inception v3	64.2857%	85.7142%	42.8571%	60.0000%
ResNet101 + Inception-ResNet v2	69.0476%	71.4285%	66.6667%	68.1818%
ResNet 101 + VGG19	71.4285%	90.4761%	52.3809%	65.5172%

In addition to the feature vectors, three key variables were identified that could potentially impact the results of the melanoma detection framework. The first is the resolution of each image. The varying resolutions of the images influence the amount of information extracted when dividing them into patches. Some images, due to their resolution, may yield more representative patches, while others might be underrepresented. To mitigate this discrepancy, standardizing all images to a common resolution without altering the aspect ratio was considered as a potential solution.

The second variable is the “OVERLAP” value, which dictates the degree of pixel overlap when segmenting images using a sliding window technique. The overlap value has an inverse relationship with the amount of information derived from each image, a smaller overlap value results in more information being captured. However, it’s hypothesized that too small an overlap might lead to redundancy and overfitting, while a larger overlap could risk missing melanoma indicators or conflating benign and malignant features.

The third variable is the number of clusters, which directly influences the size of the “dictionary” used in the Bag of Visual Words method. To optimize these variables, a range of values was established for each (as shown in Table 12), and combinations of these values were tested in subsequent experiments.

As indicated in Table 12, the exploration of various combinations of the three

Table 12. Range of possible values for the three variables to be tested.

Variable	**Values**			
Resolution	Between mean and lowest resolution (ML)	Mean of all resolutions (M)	Between mean and highest resolution (MH)	-
OVERLAP	20	50	80	100
No. of Clusters	20	50	80	100

variables amounts to a total of 48 distinct experiments and result tables. To maintain clarity and focus on the most relevant findings, only the tables showcasing the top 3 results from these experiments are presented below.

Table 13. Experimental results where Mean of resolutions (ResNet101_M) and OVERLAP = 50 values are used.

Network / Cluster	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**
ResNet101_M / 20	69.0476%	66.6667%	71.4285%	70.0000%
ResNet101_M / 50	69.0476%	71.4285%	66.6667%	68.1818%
ResNet101_M / 80	**80.9523%**	**85.7142%**	**76.1904%**	**78.2608%**
ResNet101_M / 100	76.1904%	76.1904%	76.1904%	76.1904%

Table 14. Experimental results where between mean of resolutions and maximum resolution (ResNet101_MH) and OVERLAP = 100 values are used.

Network / Cluster	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**
ResNet101_MH / 20	59.5238%	52.3809%	66.6667%	61.1111%
ResNet101_MH / 50	64.2857%	66.6667%	61.9047%	63.6363%
ResNet101_MH / 80	**85.7142%**	**95.2380%**	**76.1904%**	**80.0000%**
ResNet101_MH / 100	**83.3333%**	**80.9523%**	**85.7142%**	**85.0000%**

Table 15. Experimental results where between mean of resolutions and lowest resolution (ResNet101_ML) and OVERLAP = 100 values are used.

Network / Cluster	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**
ResNet101_ML / 20	66.6667%	76.1904%	57.1428%	64.0000%
ResNet101_ML / 50	76.1904%	90.4761%	61.9047%	70.3703%
ResNet101_ML / 80	**83.3333%**	**90.4761%**	**76.1904%**	**79.1667%**

Table 15 continued from previous page

ResNet101_ML / 100	71.4285%	80.9523%	61.9047%	68.0000%

Analyzing the data from Tables 13, 14, and 15, it becomes evident that varying the combinations of the three key variables led to a significant improvement in the results of our melanoma detection framework. Notably, the most effective results were achieved with a specific combination: using 80 clusters, setting the image resolution to a midpoint between high and mean, and dividing the images into patches with 100-pixel intervals.

To conclusively gauge the level of improvement achieved over previous frameworks, the best-performing combination identified from these experiments was also tested on the "full scale" dataset. Remarkably, this framework, with the optimized combination of variables, achieved an accuracy of 75.00%.

3.4.2. Verdict

The results obtained clearly indicate that the modifications made to the feature vector extraction process in our workflow have positively impacted the overall performance. Furthermore, the specific combinations of values for the three key variables have optimized the workflow, leading to the most effective results achieved to date. When compared to the best outcomes of the previous frameworks, there is a notable improvement of nearly 19%.

The outcomes offer several key insights. Firstly, the Bag of Visual Words (BoVW) based framework has proven not only effective but also adaptable for future updates. Secondly, the ResNet-101 model, and residual networks in general, have demonstrated their capability in providing highly representative feature vectors suitable for this framework. Given their effectiveness, ResNet-101 vectors will be the primary choice for feature extraction in future iterations of the framework. Additionally, the approach of clustering feature vectors separately for benign and malignant lesions to construct the BoVW dictionary has shown to be more effective than clustering all vectors together. This separate clustering approach will be adopted as a standard practice moving forward. The optimization of the variables such as resolution, overlap, and the number of clusters has significantly enhanced the framework's performance. The specific combination of using 80 clusters, setting the image resolution to a midpoint between high and average (labelled MH), and segmenting the images into patches with 100-pixel intervals, has been identified as the most effective. These settings will be the default parameters in subsequent versions of the framework.

3.5. BoVW with Neural Style Transfer (NST)

The adaptation of the feature extraction step to incorporate a pre-trained network model marked a significant improvement in the framework. However, further enhancements were necessary for optimal performance. One potential area of improvement identified was the handling of patches extracted from lesion images.

In the current framework, patches are generated from the entire image, including large areas of background skin and potential artifacts. As a result, a substantial portion of these patches may contain information irrelevant to melanoma detection. Under normal circumstances, this might not significantly impact the overall outcome. However, the framework's methodology involves treating each patch as representative of the label assigned to the source image. This means that a patch derived solely from the background skin of a melanoma-labeled image is inaccurately represented as a melanoma feature vector. This misrepresentation can potentially mislead the system.

To overcome this, a new step before the preprocessing was tried: Neural Style Transfer (NST) (Jing et al., 2020). NST is an optimization technique that merges two distinct images: a content image, which is the primary subject, and a style reference image, often you can see this as an artwork by a renowned artist. The goal of this process is to produce an output image that retains the core structure or content of the content image but rendered in the artistic style of the style reference image. This technique is achieved through an optimization process where the output image is iteratively refined to align with the content characteristics of the content image and the stylistic elements of the style reference image. The extraction and matching of these characteristics are accomplished using a CNN.

NST for melanoma detection starts with specifying the two components: the content image and the style reference image. In this context, the content image serves as the foundation, establishing the fundamental characteristics for the output image. For our framework, we have selected two types of content images. The first is a blank canvas, essentially an empty image, while the second is created by averaging the color channels of each lesion image in the "full scale" dataset. Both content images are of the same size, 900 × 900 pixels. This size strikes a balance between ensuring a sufficient number of patches are available for effective feature extraction and reducing the computational load to make the process more time-efficient. The style reference in this case is the lesion images in our "full scale" dataset. NST employs a CNN to extract defining features from the content image and style elements from the reference image. These extracted elements are then combined to produce the output image, which retains the characteristics of the content image but is rendered in the unique style of the reference image. This uniformity in characteristics with individualized styles is the fundamental advantage of NST. Figure 14 illustrates two transformed lesion images using the empty content image. In contrast, Figure 15 demonstrates two transformed lesion images using the averaged content image, a method also referred to as "Guided" NST. The term "guided" is used because the introduction of shapes or styles into the content image influences how the styles from the reference image are applied, thereby guiding the transformation in a specific way.

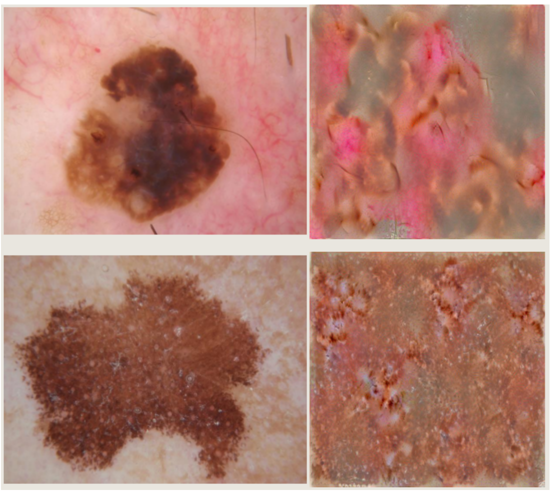
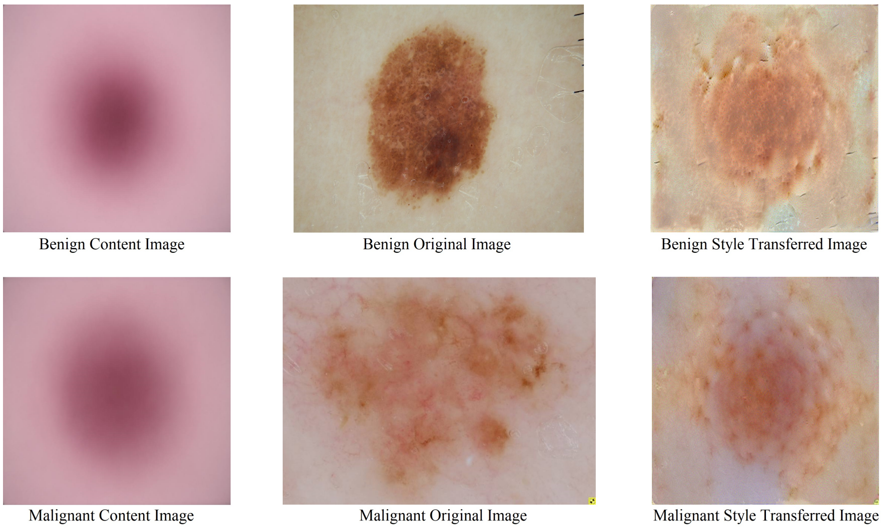


Figure 14. NST - the original lesion images are displayed on the left side, while on the right side, we see the results of transferring their styles onto an empty content image.



Benign Content Image

Benign Original Image

Benign Style Transferred Image

Malignant Content Image

Malignant Original Image

Malignant Style Transferred Image

Figure 15. Guided NST - The top three images show a benign lesion with its styled counterpart and a benign content image derived by averaging RGB channels from all benign images. The bottom three represents the same for a malignant lesion.

3.5.1. Experiments

Two experiments were done to evaluate the effectiveness of the Neural Style Transfer (NST) approach in melanoma detection framework.

The first experiment involved using the NST method to capture the styles from the lesion images in the “full scale” dataset and applying them to an empty content image. This approach focused solely on the style attributes of the lesions, omitting information about shape and location of sub-textures. The hypothesis was that the unique styles of the lesions would enhance the distinguishing features that help classify images as benign or malignant. Despite these expectations, the results were not as anticipated. The accuracy achieved was only 68.95%, which was significantly lower than the results from previous version.

The second experiment employed the Guided NST method. This approach aimed to create a content image that could guide the style transfer process in a way that would highlight common characteristics across all samples, while also enhancing individual features. The content image was generated by averaging the RGB channels of all images within a class. The intent was to bring all samples to a common baseline while emphasizing the textures and colors within the lesions. However, this method, applied to the “full scale” dataset using the stylized images from Guided NST, yielded an accuracy of 70.14%. While this was a slight improvement over the NST experiment, it still fell way short of the previous framework’s performance, which had an accuracy of 75.00%.

3.5.2. Verdict

The results from the experiments with NST and Guided NST suggest that while these techniques offer benefits in enhancing input patches, they fall short of improving the performance of the existing melanoma detection pipeline. Consequently, the use of NST in its current form will be discontinued in future iterations of the melanoma detection framework.

A closer examination of the low performance associated with NST is crucial to understand the decision to cease its use. In the first NST experiment, the default settings involved an empty content image and utilized a pre-trained VGG-19 network. This network was originally trained on everyday images, which differ significantly from medical imaging data. Additionally, the standard NST process extracts activation vectors from four different layers of the VGG-19 network, but this approach proved to be extremely time-consuming in our experiments. Processing just one image took about an hour, indicating that processing the entire dataset would be impractical for a timely diagnosis. To mitigate this, only activations from the last layer of VGG-19 were used in our experiments. This change, however, may have reduced the possible performance boost that employing all four layers may have provided. Given that one of our aims is to develop a melanoma detection framework that is efficient and swiftly adaptable to new data, the lengthy processing time of NST in its default configuration renders it unsuitable for our cause.

In a future research, alternative network models, particularly those with residual capabilities, might be considered for NST. As indicated in Table 10, networks with residual capabilities have previously demonstrated better performance in our framework.

3.6. Bag of Visual Words with Enhanced Deep Features

The exploration of NST revealed that it was not a suitable approach for our purposes. However, the concept of enhancing image patches remains a promising avenue for improving the performance of our melanoma detection framework. Recent advancements have shown that extracting various types of masks from the original image and using them either for classification or as a means of image enhancement can be effective in certain applications (Karakaya et al., 2021). These masks are typically derived from pixel values and statistical analyses between them. Once extracted, they can either be used directly as features or applied back to the image as a form of enhancement, similar to a preprocessing step. In the new iteration of our melanoma detection framework, this method is adopted to enhance image patches. This involves extracting pertinent information from the patches and then reapplying it to them, thereby improving their quality and relevance for the detection process. In the framework, this step is positioned between the steps of dividing images into patches and extracting feature vectors.

Additionally, the issue of patches containing excessive background skin and artifacts is addressed, which has been a persistent challenge. In this updated version, a more selective approach to patch gathering is implemented. This new step, named “Patch Extraction” from now on, is designed to selectively extract patches based on their lesion content. Specifically, it only extracts patches if more than 50% of their pixels are derived from the lesion area. This is made possible by utilizing the segmentation masks provided for each image in the ISIC Archive. According to that, the framework is updated once more as seen in Figure 16.

3.6.1. Patch Enhancement Masks

To enhance the image patches, it is considered to use five different types of masks: Principal Component Analysis (PCA), Well-exposedness, Saturation and Brightness. These masks have been selected based on their demonstrations in similar contexts, as detailed in Karakaya et al. (2021). The following provides an overview of each mask and its intended function. The phrase “applying a feature mask” in this context refers

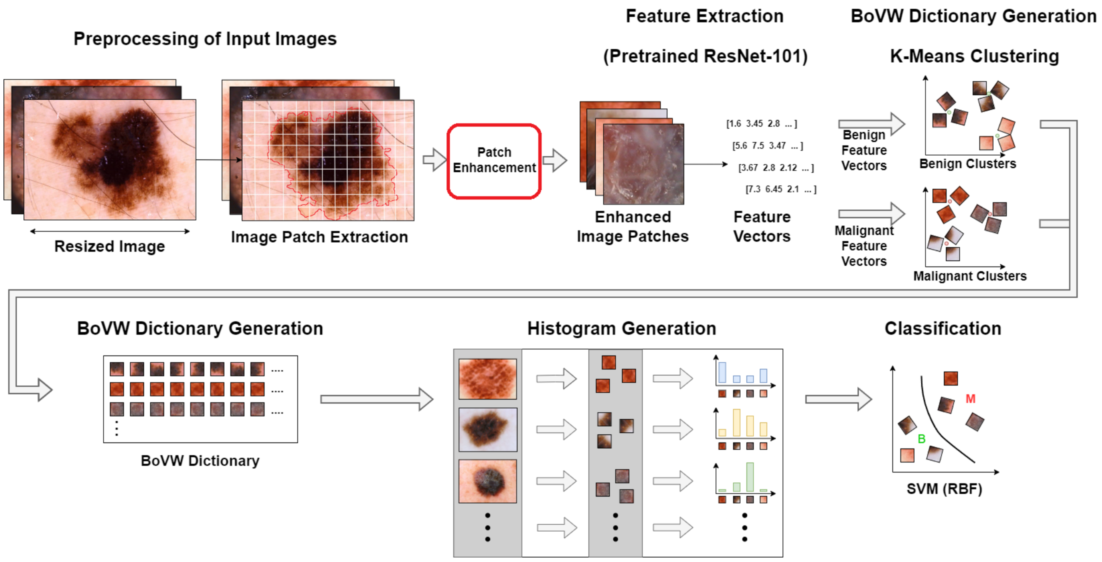


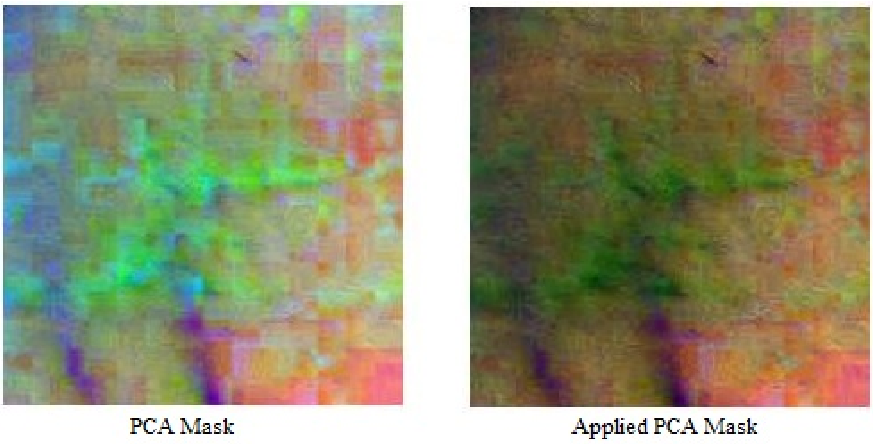
Figure 16. Updated framework with newly added patch enhancement step.

to performing an element-wise multiplication between the original image patch and the feature mask.

PCA. The first mask under discussion utilizes PCA to assign greater weights to dominant pixels in an image, an approach aligned with the core philosophy of PCA. Consider an image patch with its constituent color channels: Red (R), Green (G), and Blue (B). Each of these channels has dimensions of $r \times c$ pixels, where r and c represent the number of rows and columns, respectively, in the patch.

The PCA-based process involves several steps. Each color channel is transformed into a vector. These vectors are then stacked to form a matrix of dimensions $rc \times 3$. PCA is then applied to this matrix to compute observation scores. This analysis highlights the principal components, essentially identifying the dominant features in the color space of the image patch. The resulting score vectors are normalized linearly within the range of $[0, 1]$. These normalized scores are then reshaped back into matrices of dimensions $r \times c$, creating a PCA weight matrix for each color channel of the patch. Lastly, to emphasize the less dominant pixels, which might be crucial in identifying subtle signs of melanoma, the weight matrices are subtracted from 1. This inversion gives larger weights to pixels that were initially less dominant in the PCA analysis.

The decision to focus on less dominant pixels is informed by the nature of melanoma, where key indicators may be dispersed and not immediately conspicuous, especially in areas that appear healthy. This subtlety poses a challenge in melanoma detection, even for experienced dermatologists. The use of PCA in this way aims to enhance the visibility of these subtle but critical signs, as elaborated in the studies by Elder et al. (2020) and Dahiya (2002). The Figure 17, shows a PCA mask and its application.



PCA Mask

Applied PCA Mask

Figure 17. On the left a PCA mask taken from a patch is shown. The same patch after PCA mask is applied is on the right.

Well-exposedness. The “well-exposedness” mask focuses on the optimal exposure level of pixels within an image. It aims to emphasize pixel intensities that are neither too under-exposed (too dark) nor over-exposed (too bright). The exposure level is gauged based on intensity values that fall within the range of $[0,1]$, where 0 represents complete under-exposure and 1 represents full over-exposure.

For this mask, the target is to highlight pixels whose intensity levels are close to the median value of 0.5, indicating a balanced level of exposure. These well-exposed pixels, situated in the mid-range of the intensity spectrum, are deemed to offer the most useful visual information.

The well-exposedness feature for each color channel in an image patch is extracted using a Gaussian curve. This process is applied separately to the Red (R), Green (G), and Blue (B) channels of the image, denoted as I_R, I_G and I_B respectively.

The Gaussian curve is shown in the Eqn. 9 as

$$E_{I*} = exp\left(-\frac{(I_* - 0.5)^2}{2\sigma^2}\right) \tag{9}$$

where E_{I*} denotes the well-exposedness map of the color channel I_* and σ is standard deviation of the image patch pixels.

Saturation. Color variation plays a critical role in identifying melanoma, and saturation (SAT) features are instrumental in assessing the purity and intensity of colors within each image patch. High saturation values are indicative of colors that are more intense and uniform. In the context of melanoma detection, areas with higher saturation are often more significant and thus, SAT features assign greater weights to pixels with higher saturation values.

To compute SAT features, the S-channel (saturation channel) of the HSV (Hue, Saturation, Value) color space is employed. In line with the approach used for PCA matrices, the saturation values are inverted by subtracting them from 1 before they are used as a weight mask. This inversion ensures that the focus is on pixels with higher original saturation values, as these are likely to be more relevant in the context of melanoma detection. The Figure 18, shows a SAT mask and its application.

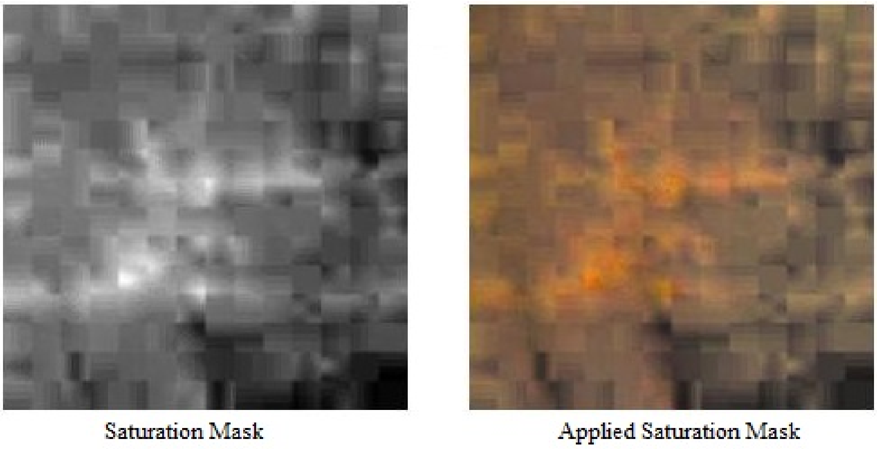


Figure 18. On the left a SAT mask taken from a patch is shown. The same patch after SAT mask is applied is on the right.

Brightness. The brightness (BRI) feature of a pixel is calculated based on its deviation from the mean value of the color channels at the same spatial location. This approach allows for the assessment of how bright or vibrant each pixel is compared to the average color intensity at that point in the image. Once the BRI values are computed with Eqn. 10 below for each pixel in each color channel, then they are normalized to fit within a range of $[0,1]$. The Figure 19, shows a BRI mask and its application.

$$B_{I*} = |I_* - M| \tag{10}$$

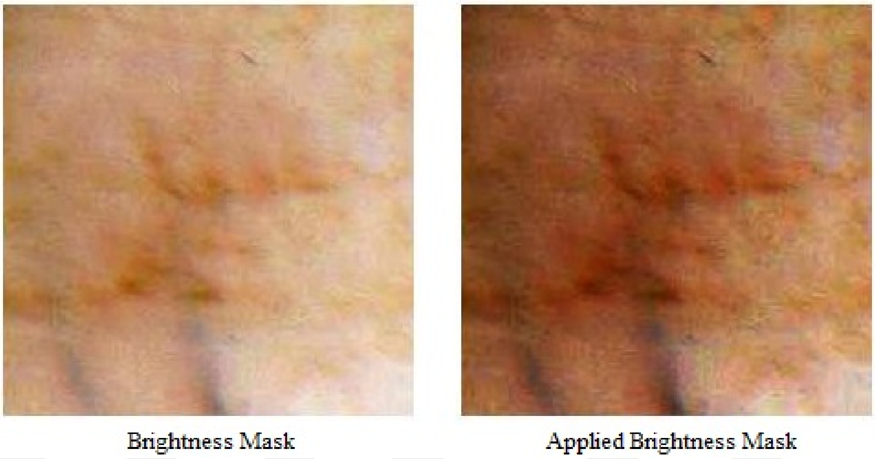


Figure 19. On the left a BRI mask taken from a patch is shown. The same patch after SAT mask is applied is on the right.

3.6.2. Experiments

To comprehensively evaluate the effectiveness of the enhancement masks in the melanoma detection framework, a series of 11 distinct runs were conducted. These runs included tests of each mask individually as well as various combinations of these masks. It's important to note that the patch enhancement step, involving the application of these masks, was incorporated between the patch extraction and feature extraction steps in the workflow, with no other modifications made to the framework.

MATLAB 2020b was again the tool of choice for generating frequency histograms and conducting the final classification step. Meanwhile, Anaconda was utilized for the clustering process, ensuring efficient memory usage and "full scale" dataset is used.

The results of these experiments, which include combinations of different enhancement masks and their impact on the framework's performance, are detailed in Table 16. In this table, "RGB" refers to the original image patch without any mask applied. The "+" symbol is used to represent element-wise multiplication, indicating how each mask was applied to the image patches.

Table 16. Experimental results after applying masks individually or their combinations to the image patches.

	Accuracy	Sensitivity	Specificity	Precision
PCA + RGB	64.45%	60.90%	68.00%	65.55%
SAT + RGB	69.00%	70.00%	68.00%	68.62%
BRI + RGB	70.50%	76.00%	65.00%	68.46%
W_EX + RGB	65.01%	70.00%	60.03%	63.65%
PCA + BRI + RGB	69.50%	72.00%	67.00%	68.52%
PCA + W_EX + RGB	65.50%	73.00%	58.00%	63.47%
PCA + SAT + RGB	63.00%	57.00%	69.00%	64.77%
SAT + BRI + RGB	**71.40%**	**65.00%**	**77.00%**	**73.86%**
W_EX + SAT + RGB	62.50%	70.00%	55.00%	60.86%
PCA + SAT + W_EX + RGB	70.00%	72.00%	68.00%	69.23%
PCA + SAT + BRI + RGB	**72.50%**	**78.00%**	**67.00%**	**70.27%**
PCA + SAT + W_EX + BRI + RGB	68.00%	72.00%	64.00%	66.66%

3.6.3. Verdict

The analysis of the results clearly shows that both the SAT and BRI masks, whether used individually or in combination, tend to improve the results of the framework. Among these, the BRI mask stands out as particularly effective. Conversely, the W_EX (well-exposedness) mask appears to have a detrimental effect on overall accuracy.

In terms of combinations, the best results within the patch-enhanced BoVW framework were achieved when the PCA, SAT, and BRI masks were used together. Despite this improvement, it's notable that the highest accuracy achieved with these combinations still falls short of the previous benchmark of 75.00%. However, it's important to consider that the efficacy of patch enhancement is highly dependent on the number and variability of patches. As such, increasing the number of samples and the diversity of patches could potentially enhance the performance of this method.

Given this possibility, we have decided not to completely abandon the use of patch enhancement. Instead, subsequent variations of the framework will be tested both with and without the patch enhancement step to further evaluate its impact and potential benefits as the dataset scales.

3.7. Weighted BoVW with Enhanced Deep Features for Melanoma Detection

Up to this point, the modifications and experimental efforts have primarily aimed at enhancing feature vector quality. Most changes were responses to the first two challenges highlighted in Section 3.3.2, focusing on improving how feature vectors are derived from image patches. However, there has been no action yet to address the third challenge, which is intrinsic to the nature of melanoma and its representation in the framework.

Despite the substantial changes to the framework, one fundamental issue persists: each original image is still represented by a combination of feature vectors extracted from the image's patches, with these vectors inheriting the same label as the image itself. These labels become problematic in the context of melanoma detection. In a malignant lesion, only a small portion of the lesion may exhibit melanoma-specific features, while the majority of the lesion might display benign, non-informative features. However, under the current labeling scheme, all patches from a malignant lesion are labeled as malignant, irrespective of their actual content.

Melanoma's inherent characteristics contribute to this challenge. A lesion is classified as melanoma if it exhibits certain clinical features indicative of the condition. These features are often scattered and present only in small areas, particularly in early stages. The rest of the lesion may appear healthy, displaying benign features. Conversely, a benign lesion lacks these melanoma indicators. Considering this, the dictionary created in the BoVW framework ends up comprising half benign features and half mixed features mislabeled as malignant.

To address this problem, an update to the histogram generation step in the framework has been thought, based on the following approach.

3.7.1. Weighted BoVW

BoVW technique comes from the Bag of Words as it is mentioned previously. In NLP, this technique can be used to classify text documents with the frequency of each word in the document. When looked carefully, one can see that there is a similar problem in document classification as melanoma detection. If you think the classification of a document in terms of its general topic, the problem seems to be solved relatively easily with BoW because of the number of topic related words in the text. A real life example may be a text about the performance of a processor. When classified this way the BoW technique can easily say that this is a tech related text. However, if the problem changes from “find the topic” to “which PC part’s performance is discussed in the text”, then BoW will have problems. The frequencies of processor related words are much much less than the tech related words. Moreover, words like “a”, “an”, “the” dominate the frequency histograms. Then, the problem becomes to finding small indicators scattered in the text to detect the PC part, which is very similar to melanoma detection.

The solution again comes from the BoW concept. A weighting scheme applied to the histograms can impact the performance immensely. In the case of NLP, one can lower the impact of non-informative words by a smaller weight or some key words may have huge weights to increase their impact. Choi and Han (2013)’s study presents a good evaluation of some of these schemes for BoW concept. The same type of approach can be used for melanoma detection. In the histogram generation step, a weight can be applied to the melanoma features to make them look like their frequency is much higher. This would differentiate the images with patches that are closer to the melanoma cluster centers in the dictionary. Of course this could have been applied to the benign features too but, benign features exist in both parts of the dictionary which makes this not beneficial. The weight scheme is implemented in histogram generation step of the framework.

3.7.2. Experiments

Two experiments were conducted using the “full scale” dataset, utilizing MATLAB 2020b for processing. As in previous instances, Anaconda was employed for the clustering phase. The experiments aimed to test the impact of introducing a weighting scheme in the Bag of Visual Words (BoVW) framework, both with and without the addition of patch enhancement.

The first experiment involved the application of the Weighted BoVW framework without any patch enhancement. In this approach, a default weight of 10 was applied by adding it specifically to the frequency of features identified as melanoma indicators during histogram generation. This weight value has been determined following extensive experiments among a range of options from 1 to 50.

The second experiment used the same Weighted BoVW framework but incorporated the patch enhancement step. Again, a default weight of 10 was applied. Furthermore, the combination of PCA, BRI, and SAT masks, which previously showed the best performance, was used to enhance the patches.

The results of these two experiments, detailing the effectiveness of the weighted approach in the BoVW framework with and without patch enhancement, are presented in the accompanying Table 17.

Table 17. Results of Weighted BoVW approach on “full scale” dataset.

Framework	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**
Weighted BoVW with Deep Features	%91.00	%95.00	%87.00	%87.96
Weighted BoVW with Enhanced Deep Features	%94.50	%96.00	%93.00	%93.20

The outcomes of the experiments provide a compelling insight into the effectiveness of the updated approaches in the melanoma detection framework. Notably, both the Weighted BoVW frameworks outperformed all previous versions in terms of accuracy.

Interestingly, the framework incorporating patch enhancement exhibited slightly better performance compared to its non-enhanced counterpart. This improvement, although modest, is particularly noteworthy given that patch enhancement had not previously demonstrated a clear advantage in earlier experiments. One plausible explanation for this enhanced performance is the introduction of the weighting scheme. By assigning greater significance to features indicative of melanoma, the weighting scheme might have amplified the impact of the melanoma features that were further emphasized through the patch enhancement process.

3.7.3. Verdict

The significant improvement in performance observed with the introduction of a weighting scheme to the BoVW framework strongly suggests that this approach was a beneficial and effective modification. This advancement is a pivotal moment as it demonstrates a level of accuracy and reliability that suggests the framework is ready for a broader evaluation within the field.

With this milestone achieved, the focus now shifts to detailing and illustrating the final version of the framework. The remaining parts of this section will be dedicated to presenting the Figure 20 that visually represents the latest iteration of the framework. Accompanying this figure will be a thorough explanation of each step involved in the process, providing a clear understanding of how the framework functions and the rationale behind each component.

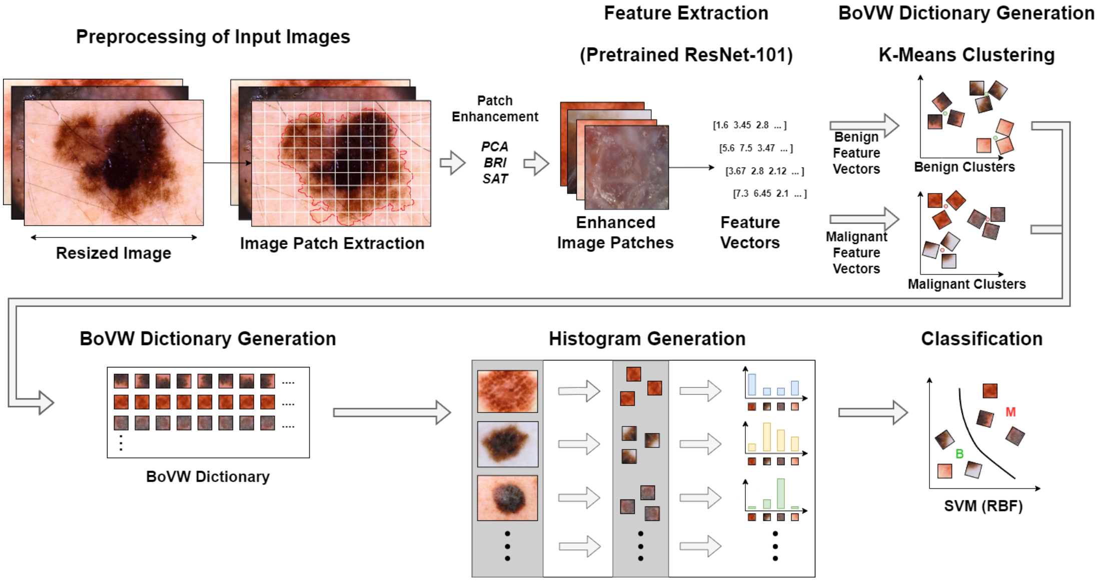


Figure 20. The latest framework: Weighted BoVW with Enhanced Deep Features.

The framework consists of six steps in its finalized form: Preprocessing, Patch Enhancement, Feature Extraction, BoVW Dictionary Generation, Histogram Generation and Classification.

Preprocessing. The preprocessing step is about determining the optimal resolutions for input images and extracting patches from them. It starts by considering three potential common resolutions: "the mean resolution (M)", "the average of the mean and the highest (MH)", and "the average of the mean and the lowest (ML)". Subsequent experiments, building on our previous work, indicated that the MH resolution was the most effective, leading to its selection as the final resolution for the framework. It's important to note, however, that the specific value of MH is dependent on the resolutions of the images in the dataset used, meaning that this value can vary with different datasets.

Once the images are rescaled to the MH resolution, the framework extracts overlapping image patches of $n \times n$ pixels from each image. The size of these patches, denoted by n, is determined by the input requirements of the preselected pretrained network model, which in this case is ResNet-101. Therefore, n is set to 224 pixels in the framework.

Another key parameter in this process is the degree of overlap between successive patches, defined by $m \times m$ pixels. This overlap parameter plays a crucial role in dictating the similarity among the input patches used for feature extraction. Setting m too small may lead to overfitting, as patches might become too similar, while a large m value might cause the system to miss critical features that distinguish benign from malignant lesions. Through empirical testing, the optimal value for m has been established as 100 pixels.

The final phase of this step involves using segmentation masks, where available, to count the number of pixels from both the background skin and the lesion within each patch. Patches with more than 50% of their pixels corresponding to the lesion are selected for feature extraction. Patches that do not meet this criterion are excluded.

Patch Enhancement. The enhancement of image patches extracted from the melanoma detection framework is a crucial step that significantly contributes to the accuracy of the system. This enhancement is achieved through the application of three specific feature maps: PCA, BRI, and SAT. The choice to use these three masks was decided by an extensive analysis of various feature map candidates and their combinations, also building upon insights from previous research (Okur and Turkan, 2022). These masks are calculated as described in Section 3.6.1.

Each of these feature maps is applied to enhance the image patches in every color channel. The enhancement is executed through a direct per-pixel weighting strategy, where the weights derived from these feature maps are applied either individually or in combination to each image patch.

Feature Extraction. Regarding feature extraction, after rigorous testing of various options, ResNet-101, pretrained on the ImageNet dataset (Deng et al., 2009), emerged as the most effective model for our framework. The choice of ResNet-101 is based on its proven efficacy in accurately extracting features relevant to melanoma detection. In the implemented experiments, feature vectors are extracted from the final fully connected layer of the ResNet-101 network. These feature vectors have a fixed size of 1000, corresponding to the fixed patch size of $n = 224$ pixels, which aligns with the input layer requirements of the model.

BoVW Dictionary. In addressing the binary classification problem of distinguishing between benign and malignant lesions in melanoma detection, BoVW needs the feature vectors extracted from the image patches to be strategically organized into clusters based on their labels for dictionary generation.

The step involves separately clustering the feature vectors derived from benign images and those from malignant images. Specifically, the feature vectors from benign images are grouped using (K/2)-means clustering, and the same method is applied to the vectors from malignant images. The objective here is to identify distinct and representative cluster centers for each class—benign and malignant. These centers are expected to efficiently characterize the respective class's feature space while minimizing the presence of outliers. Each cluster center within a class effectively acts as a specific marker, indicating that an image containing this feature likely belongs to that class.

The total number of clusters, denoted as K, plays a critical role in determining the size of the histogram vectors that are created in the next stage of the framework. These histogram vectors are then used in a Support Vector Machine (SVM) for the final classification. It's important for these histograms to have a consistent length and to be sufficiently sparse to effectively distinguish between benign and malignant cases.

The selection of the K value (number of clusters) should be carefully optimized according to the size of the dataset. Larger datasets with more varied features and indicators may require a higher K value to capture the increased diversity. In the case of our "full scale" dataset used in this framework, BoVW dictionary includes 40 key features per class, resulting in a total of $K = 80$ cluster centers.

Histogram Generation. The process of generating histograms starts after the representative BoVW dictionary has been established, the framework proceeds to construct histograms that capture the distribution of feature vectors in relation to the cluster centers.

Each histogram is initialized as a zero vector with a length of K, corresponding to the total number of cluster centers in the BoVW dictionary. For each feature vector extracted from the image patches, the framework identifies the nearest cluster center among the K available. This identification is based on minimizing the Euclidean distance between the feature vector and the cluster centers. Upon determining the nearest cluster center for each feature vector, the histogram is updated accordingly. The index k of the nearest cluster center is identified, and the k-th value in the histogram is incremented. The increment is by a weight, which is either 1 for benign or 10 for malignant, depending on the closest center.

Lastly, given that each image results in a variable number of patches (and thus feature vectors), the histograms produced for each image will have different counts of contributions. To ensure that the histograms can be used to train an unbiased classifier, they need to be normalized. The initial experiments in the development of this framework indicated that normalization using the ℓ_2-norm was more effective than using the ℓ_1-norm. Therefore, each histogram vector is normalized based on its

ℓ_2-norm, bringing them to a common scale and making them suitable for comparison and classification.

Classification. To complete the melanoma detection process, a SVM classifier with a radial basis function (RBF) kernel is employed (Cortes and Vapnik, 1995). The SVM classifier is trained using a 10-fold cross-validation approach. The hyperparameters of this SVM are optimized automatically using the functions available in MATLAB 2020b.

CHAPTER 4: EXPERIMENTAL RESULTS

To accurately assess the standing of our melanoma detection framework within the broader context of the field, a fair and comprehensive comparison with other successful frameworks is essential. The International Skin Imaging Collaboration (ISIC) challenges, as discussed in Section 2.3, provide an ideal platform for such a comparison. Each year, ISIC organizes a challenge accompanied by a specific dataset for that year's competition. Over time, ISIC releases the ground truth data for the test set, along with a leader-board that ranks the participating frameworks based on their performance.

This leaderboard is particularly valuable as it not only ranks the frameworks but also provides detailed results for each competitor. This transparency enables researchers to make fair and informed comparisons with state-of-the-art frameworks in the domain. For this thesis, the ISIC 2017 Dataset (Codella et al., 2017) is utilized and our framework's performance is compared with the top-10 entries on the ISIC Challenge 2017 Leader-board (ISIC, 2017).

4.1. ISIC 2017 Dataset

The ISIC datasets are meticulously organized and provide a standardized platform for melanoma detection research, facilitating fair and uniform comparisons across different studies. The datasets typically consist of separate, clearly defined sets of training, validation, and testing images. This structure ensures that all participants in the ISIC challenges use the exact same dataset for their experiments, maintaining consistency across different frameworks.

For the purpose of our study, the ISIC 2017 dataset was chosen primarily due to the availability of ground-truth labels for the test images (Goyal et al., 2020). Starting from 2018, ISIC began hosting live challenges with an online evaluation server, and consequently, stopped releasing test image data. This change makes the ISIC 2017 dataset the most recent one publicly available that includes both the ground-truth labels for each test image and the gold standard segmentation masks.

The ISIC 2017 dataset comprises 1626 benign and 374 melanoma images for training, and 483 benign and 117 melanoma images for testing. These images are sourced from patients of the contributing facilities and are fully labeled. One notable characteristic of this dataset is the significant class imbalance, with malignant cases being far less frequent than benign ones.

To address this imbalance, we implemented a data augmentation strategy similar to the one used by Menegola et al. (2017), which ranked third in the lesion classification category of the ISIC 2017 Challenge. This augmentation method involves randomly modifying melanoma training images via horizontal and vertical shifts (up to 10%), zoom (up to 20%), and rotation (up to 270 degrees). This process ensures that for each original melanoma image, at least one augmented version is created, effectively balancing the number of melanoma cases in the dataset to match the number of benign cases.

4.2. Environment

The experimental setup for the melanoma detection framework was conducted on a system with modest yet capable hardware specifications. The key specifications of the system used for the experiments are as follows: Intel Core i7-4790*K* processor, 32GB RAM and NVIDIA GTX980. Additionally, MATLAB 2020b is utilized under the Windows 10. Python 3.8 was used for clustering process for memory efficiency.

Even with these specifications, which is far from the latest in terms of technological advancements, the framework demonstrated commendable efficiency in terms of processing time. The most time-intensive part of the experiment, encompassing feature extraction and training, was completed in approximately fourteen hours. This duration is relatively short, considering the complexities often involved in image processing and machine learning tasks.

A key factor contributing to this efficiency is the framework's reliance on feature extraction from a pretrained ResNet-101 model, rather than training a deep neural network from scratch. By utilizing a pretrained model without a change, the framework significantly reduces the computational burden associated with the training phase.

4.3. Evaluation Metrics

In the context of the ISIC 2017 Challenge, the leader-board primarily uses Balanced Multi-class Accuracy (BMA) as the key metric for ranking, considering that the challenge involves two separate binary classification tasks: distinguishing melanoma from nevus and seborrheic keratosis, the latter two being types of benign skin tumors. However, for this research focusing specifically on differentiating between melanoma and benign lesions, additional metrics are needed for a comprehensive evaluation. Therefore, six other metrics provided by the ISIC organization are utilized to assess the effectiveness of the proposed methodology against the leader-board results: Accuracy (ACC=$\frac{TP+TN}{TP+FP+FN+TN}$), Sensitivity (SENS=$\frac{TP}{TP+FN}$), Specificity (SPEC=$\frac{TN}{TN+FP}$), Dice Coefficient (DC=$\frac{2TP}{2TP+FP+FN}$), Positive Predictive Value (PPV=$\frac{TP}{TP+FP}$) and Negative Predictive Value (NPV=$\frac{TN}{TN+FN}$). The symbols TP, TN, FP, and FN represents true positives, true negatives, false positives, and false negatives respectively.

4.4. Parameters and Patch Enhancement

The framework requires careful adjustment of its parameters given below, because these parameters are influenced by the specific characteristics of the dataset in use. For the ISIC 2017 dataset, it is essential to tailor these parameters appropriately to align with the dataset's unique features and ensure the framework operates at its highest efficiency.

The preprocessing step rescales the lesion images based on the average and maximum resolutions of the dataset. This process, which can be automated, computes the MH resolution and then rescales images to match this resolution on their longer sides while preserving aspect ratios. For the ISIC 2017 dataset, the MH resolution standardizes the longer sides to 1430 pixels.

The patch size, n, is determined by the requirements of the chosen feature extractor network. The overlap between patches, m, can typically be set as half the patch size. With ResNet-101 as the feature extractor, n is set to 224, and m is set to 100, as detailed in 3.7.3. Additionally, if segmentation masks are available like in the ISIC 2017 dataset, it allows for the exclusion of patches with more background than lesion pixels more easily. However, in datasets without segmentation masks, methods such as adaptive thresholding can be employed to accomplish this task (Bradley and Roth, 2007).

The total number of clusters in the K-means algorithm (denoted as K) should reflect the diversity of the dataset. Hence, it also needs to be adjusted. For the ISIC 2017 dataset, K is empirically set to 1000 after testing values between 100 to 2500 with increments of 50. This number can also be estimated automatically using techniques like silhouette analysis (Kaufman and Rousseeuw, 2008), keeping in mind that histograms should remain sparse for efficient classification.

For patch enhancement, the effectiveness of patch enhancement masks (PCA, BRI, and SAT) was re-evaluated with the ISIC 2017 dataset and evaluation metrics to confirm their efficacy and determine the best combination. The results, shown in Table 18, reveal that BRI significantly improves performance by 0.043 against no enhancement in terms of accuracy, while SAT has minimal impact in line with previous experiments decreasing the accuracy by an insignificant value, 0.003. It is also important to note that, only SAT falls behind the framework without patch enhancement. Hence, proving the benefits of our enhancement upgrade to BoVW framework. In the meantime, PCA combined with BRI has the best accuracy with increasing the no maps accuracy by 0.057. However, the combination of PCA, BRI, and SAT, despite not being the top performer, was chosen for its superior SENS metric, which is crucial for correctly detecting melanoma.

Table 18. The comparison of PCA, BRI and SAT enhancement maps, when tested on ISIC 2017 dataset.

Maps	**ACC**	**SENS**	**SPEC**	**DC**	**PPV**	**NPV**
No map	0.909	0.985	0.832	0.915	0.854	0.982
PCA	0.923	0.996	0.850	0.929	0.870	0.995
BRI	0.952	0.994	0.911	0.954	0.918	0.993
SAT	0.906	0.996	0.816	0.914	0.844	0.995
PCA×BRI	**0.963**	**0.990**	**0.936**	**0.964**	**0.939**	**0.989**
PCA×SAT	0.912	0.998	0.826	0.919	0.852	0.998
BRI×SAT	0.955	0.998	0.913	0.957	0.920	0.998
PCA×BRI×SAT	**0.962**	**0.998**	**0.925**	**0.963**	**0.931**	**0.998**

4.5. Results

The performance comparison with the top frameworks from the ISIC 2017 Challenge provides a critical benchmark for evaluating the effectiveness of the melanoma detection method developed in this study. The ISIC 2017 Challenge leader-board, featuring twenty successful frameworks for lesion classification, offers a standardized basis for this comparison, as all methodologies were trained and tested using the same dataset without incorporating external data.

In Table 19, the top ten methods from the leader-board are re-ranked based on their success in melanoma detection. The comparison includes the evaluation metrics mentioned previously: ACC, SENS, SPEC, DC, PPV, and NPV. These metrics are ranged from 0 to 1 with higher values indicating better performance. In the table, the highest values for each metric are highlighted in bold for easier comparison.

The "Rank" column in Table 19 shows the relative standing of each framework in melanoma detection, while the "Team Name" and "Approach Name" columns provide the official names of the teams and their methods as listed in the ISIC 2017 Challenge.

Table 19. The assessment of the Weighted BoVW with Enhanced Deep Features framework in contrast to the top-10 frameworks on the ISIC (2017) Leader-board for melanoma detection in terms of statistical performance.

Rank	Team Name	Approach Name	ACC	SENS	SPEC	DC	PPV	NPV
1	RECOD Titans / UNICAMP	release (rc36xtrm) "alea jacta est"	0.872	0.547	0.950	0.624	0.727	0.896
2	USYD-BMIT	EResNet (single scale w/o attributes)	0.858	0.427	0.963	0.541	0.735	0.874
3	University of Guelph - MLRG	Last Minute Submission!!!!	0.845	0.350	0.965	0.469	0.707	0.860
4	CVI	finalv_L2C1_trir	0.843	0.376	0.957	0.484	0.677	0.864
5	Computer Vision Inst. Shenzhen Univ.	task3_final_RQ	0.832	0.308	0.959	0.416	0.643	0.851
6	Inst. of High Performance Comput. and Nat. Skin Center, Singapore	multi-task deep learning model for skin lesion segmentation and classification-3	0.830	0.436	0.925	0.500	0.586	0.871
7	icuff	comb	0.830	0.171	**0.990**	0.282	0.800	0.831
8	Casio and Shinshu Univ. joint team	ResNet ensemble with normalized image	0.828	0.735	0.851	0.625	0.544	0.930
9	Univ. of Debrecen	Ensemble of deep conv. neural networks	0.828	0.470	0.915	0.516	0.573	0.877
10	Univ. Federal de Mato Grosso	Araguaia Medical Vision Lab - GooglAlexNet	0.827	0.521	0.901	0.540	0.560	0.886
*	**Okur and Turkan**	**Weighted BoVW with Enhanced Deep Features**	**0.962**	**0.998**	0.925	**0.963**	**0.931**	**0.998**

From the statistics presented, it is observed that the methodology developed in this study outperforms the other algorithms in all the evaluation metrics, except for SPEC. This achievement underscores the effectiveness of the proposed method in accurately detecting melanoma. The detailed discussion of these results, provided in the following section, offers insights into the strengths and potential areas for refinement of the framework.

4.6. Discussion

The methods featured in Table 19 predominantly rely on deep neural network models and their ensembles (Menegola et al., 2017; Bi et al., 2017; DeVries and Ramachandram, 2017; Jia and Shen, 2017; Li and Shen, 2017; Yang et al., 2017; Vasconcelos and Vasconcelos, 2017; Matsunaga et al., 2017; Sousa and de Moraes, 2017; Harangi, 2018). A very brief overview of the top five methodologies, which were examined in greater detail in Chapter 2, highlights the differences among them as well as BoVW framework of ours.

Menegola et al. (2017)'s method stands out at the forefront with a composite model comprising seven sub-models, six of which are based on Inception and one on ResNet. This complex assembly necessitates substantial computational power, particularly due to its integration via an SVM classifier layer. Following this, Bi et al. (2017)'s method, which ranks second, utilizes three distinct strategies centered around ResNet. They approached the classification task both as a multi-class problem (with labels for melanoma, nevus, and seborrheic keratosis) and as a binary classification problem, culminating in an ensemble model. DeVries and Ramachandram (2017), securing the third spot, introduced a multi-scale CNN model grounded in the Inception network. This model, initially pretrained on ImageNet, was further adapted to various image resolutions, leading to the training of additional models and, subsequently, an ensemble of ten models, imposing a significant computational demand. Jia and Shen (2017)'s methodology, which placed fourth, employs a deep CNN with 14 convolutional layers and a fully connected layer for image analysis and classification. Their approach, utilizing a variant of VGG-GAP (Zhou et al., 2015), involves a two-stage process for generating class activation maps and subsequent classification. Lastly, Li and Shen (2017), ranking fifth, adopted two deep learning strategies for lesion segmentation and feature extraction. Their dual-framework approach comprises two fully-convolutional residual networks for concurrent segmentation and classification, alongside a deep CNN specifically for feature extraction from dermoscopic images.

The Weighted BoVW framework diverges from the common practice of training deep networks or their ensembles. Instead, it leverages a pretrained network to extract features from image patches, a strategy that circumvents the need for extensive computational resources. Impressively, this method surpasses the performance of all leader-board methodologies in almost every metric, barring specificity (SPEC) (Okur and Turkan, 2024).

SPEC is a measure of the true negative rate, reflecting the accuracy in identifying non-malignant cases. The slightly lower SPEC value achieved by our framework could be linked to the construction of the BoVW dictionary. The separate construction of benign and malignant clusters may lead to some overlap between them due to the very close cluster centers, causing the BoVW weighting strategy to exhibit a slight bias towards the malignant class. Despite this, Sensitivity (SENS), which represents the true positive rate or the ability to correctly identify malignant cases, is often deemed more critical than SPEC in melanoma detection. Our method achieves an exceptional SENS value of 0.998, as shown in Table 19, underscoring its effectiveness in detecting a high number of melanoma cases correctly, despite the potential for occasionally misclassifying some benign ones.

Furthermore, PPV and NPV statistics reinforce these observations. Additionally, the ACC and DC performances of our framework outshine those of competing algorithms. This success is attributed to the unique approach taken towards lesion images. While other methods typically treat each image as a single unit for feature extraction and classification, our method adopts a different strategy. It builds a dictionary using image patches from all training images, which is then employed to identify semantically similar patches within a given lesion image, creating a histogram for each. These histograms are crucial for classification, allowing for a more generalized representation of malignant characteristics across multiple image patches in each lesion. This approach contrasts with methods where each image contributes only a single feature vector, leading to a more effective SVM-based detection in our framework.

The proposed approach, while effective, does have its limitations. Presenting these is done according to the framework's modular structure. The first step to focus on is the feature extraction step.

Currently, the feature extraction employs ResNet-101, a pretrained residual neural network. However, it's important to note that ResNet-101 was originally trained on non-dermoscopic images. Despite this, like other neural networks, it has a characteristic where images with similar semantic content yield close representations in the layer preceding classification. This is where the feature vectors for our framework are extracted. The consistency in representations for lesion image patches, even though the network was trained on non-lesion images, provides a basis for their effective use in the framework. The semantic distinctions between these representations are sufficiently clear to be reliable for our purposes. However, they are still extracted based on non-lesion images and this indicates potential for improvement. For instance, incorporating weights from a novel residual architecture specifically trained on lesion images could significantly enhance the framework's performance. As a future enhancement, exploring and testing an alternative, readily available network architecture is a viable strategy. Eventually, designing and training a new architecture tailored to dermoscopic images could be a substantial advancement. This bespoke architecture would be trained only once using publicly available dermoscopic images and could potentially replace the current use of ResNet-101 in the framework.

A closer examination of the dictionary generation step in the framework also reveals certain limitations associated with the current use of K-means clustering. First aspect to look into may be setting the value of K. Instead of experimentally setting the value, a systematic approach can be used, such as Hierarchical Clustering (Nielsen, 2016). Also known as hierarchical cluster analysis (HCA), hierarchical clustering is a cluster analysis method that aims to establish a hierarchy of clusters where we can predict a reasonable K value without experimenting exhaustively. Even though, K-means has shown favorable results and setting this K value may boost this, it is fundamentally a linear clustering algorithm, and there is potential for enhancement by exploring non-linear alternatives. Such exploration could include evaluating different variations of the K-means clustering module such as Kernel K-means (Dhillon et al., 2004), Ensemble K-means (Iam-on and Garrett, 2010), Shift-Invariant K-means (Oktar and Turkan, 2022) and other potential methodologies (Moradi Fard et al., 2020; Kohonen, 1990; Martinetz et al., 1993), also present promising avenues for refining the clustering process. Each cluster center in the BoVW dictionary represents a distinct feature identifiable in dermoscopic images. By adopting a more advanced clustering algorithm, it might be possible to identify a broader range of unique features, leading to the creation of a more comprehensive and descriptive dictionary.

Since one of our contributions includes better computational load and time efficiency, it is important to specify the information regarding complexity of our framework. The ResNet-101 network's complexity is provided by He et al. (2016a) in number of floating point operations as 7.6×10^9. The remaining sections of the framework has $O(n^K)$ complexity in the worst case. n is the number of feature vectors and K represents number of cluster centers in the BoVW dictionary. Lastly, the clustering is done separately in Sklearn library as previously stated. They provide their K-Means implementation complexity as $O(n^{(K+2/p)})$ based on Arthur and Vassilvitskii (2006)'s study. n here represents the number of samples and p is the number of features.

Finally, there's a growing recognition of the importance of eXplainable Artificial Intelligence (XAI) (Adadi and Berrada, 2018; Murdoch et al., 2019). XAI aims to improve the transparency and interpretability of AI systems, enabling users to understand, trust, and effectively manage these systems. This approach is particularly crucial in fields like healthcare, where understanding the decision-making process of such models is vital for clinical validation and user trust. With XAI the framework could significantly enhance its explanatory precision. Hence, the next section of this thesis will be specifically devoted to applying the XAI approach to this framework. This exploration will delve into methodologies and techniques for making internal workings of some steps inside the framework more transparent and understandable.

4.7. eXplainable Artificial Intelligence (XAI)

XAI represents a shift from the traditional "black-box" approach of AI to a more transparent and understandable form (Adadi and Berrada, 2018; Murdoch et al., 2019). The key difference between regular AI and explainable AI lies in the level of clarity and comprehension they offer about how decisions or predictions are made.

In traditional AI systems, particularly those driven by complex machine learning (ML) algorithms, decisions are often made in a way that is not transparent to the users or even for some cases the developers. These systems process input data and produce output (such as predictions or classifications), but the internal decision-making process remains opaque. This lack of transparency is why they are often referred to as "black-box" systems. The architects of these systems may know the inputs and outputs but might not fully understand the detailed workings of the algorithm that leads to a particular result.

XAI aims to make the decision-making process of AI systems clear and understandable. It implements specific techniques and methods to ensure that decisions made during the ML process is traceable and explainable. This approach enhances the ability of users and developers to understand and trust AI systems. In XAI, it's not just about what decision was made, but also about why and how it was reached, including the underlying logic and factors that influenced it, if possible.

In the context of our framework, which uses a SVM for classification, the concept of XAI can still be applied. Although an SVM is not an AI in the conventional sense, it can still be seen as a "black-box" system where the rationale behind its classifications might not be immediately apparent. To make this framework more explainable, identifying and highlighting the features that most significantly influence the SVM's decisions can be focused on. By elucidating which features are key indicators in classifying a case as melanoma, you can provide more transparency and understanding of the framework's functioning.

Explaining classifications in machine learning, particularly in complex models like the one in our framework, can be challenging. Shapley values, a concept from coalitional game theory, provide a powerful method to tackle this challenge (Sundararajan and Najmi, 2019). They offer a way to understand how different features contribute to the final classification decision.

In the context of our melanoma detection framework, applying Shapley values allows for a detailed analysis of how each feature influences the classification of a single test case as melanoma. This method treats each feature value of an instance as a contributor to the outcome, and Shapley values help in fairly attributing the decision (melanoma or not) to these contributors. After calculating Shapley values for each correctly classified melanoma case, the Table 20 presents the most effecting 10 feature indexes.

Table 20. According to the Shapley values from all correctly classified melanoma cases, the top 10 features most effective for a lesion to be classified as melanoma.

1	2	3	4	5	6	7	8	9	10
#818	**#924**	**#651**	#896	#526	#812	#592	#999	#797	#507

From the Table, the top 3 features out of the 1000 can exclusively be focused on, which have the most significant impact on the decisions. They exist nearly %10 of correctly classified melanoma cases. These "top" features represent the number of image patches closest to the same cluster center in the BoVW dictionary, which are assumed to be critical in the classification. It's essential to recognize that these features correspond to specific image patches that have a high resemblance to particular cluster centers, denoting strong indicators of melanoma. The Figure 21 and Figure 22, shows randomly selected 5 image patches per each of these top 3 features which includes the feature they represent.

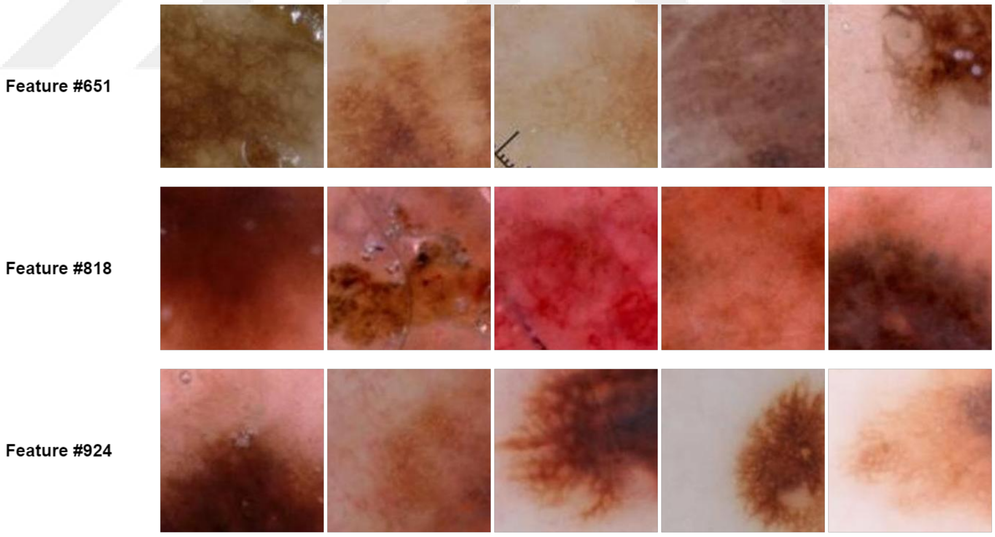


Figure 21. Example image patches for the most effective features. The image patches on the right of a feature means that those patches include the feature.

Evidently, there is a discernible pattern. "Feature 924" primarily targets the margins of lesions. In contrast, "Feature 818" concentrates on elements situated directly on the lesion itself. Although the final image patch for 818 might appear anomalous, it is relevant since over 60% of the patch encompasses the lesion, and

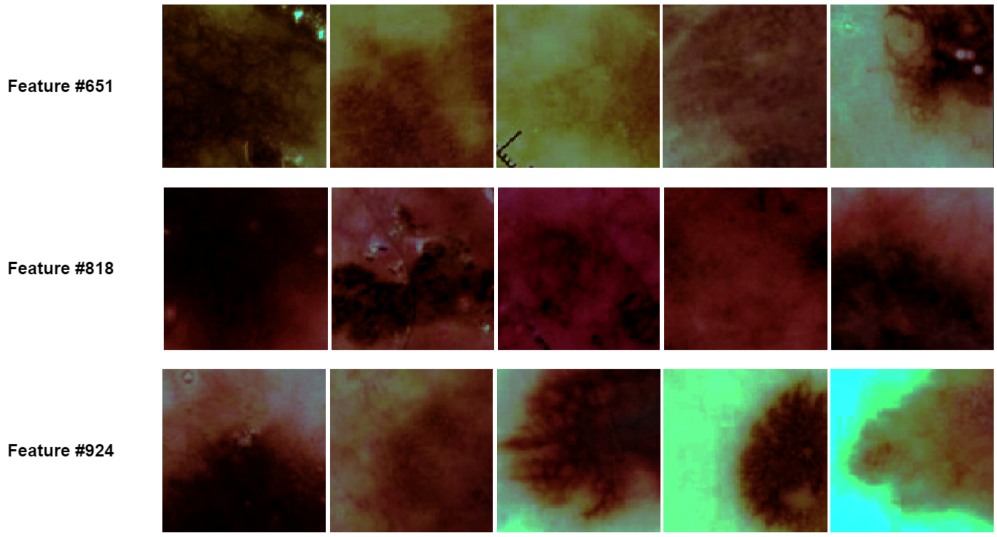


Figure 22. The same example image patches from the Figure 21. The image patches are replaced with their enhancement masks applied versions. The patches are darker but structures inside lesions are more apparent.

it displays discernible structures. "Feature 651" presents an intriguing aspect. At first glance, it appears to target areas similar to those of 818 and 924. Yet, it uniquely focuses on a particular kind of structure where a lesion's light and dark hues intermingle, forming small circular shapes and streaks.

These findings could lay the groundwork for future studies aimed at refining existing features or developing new ones. Firstly, the framework seems to accurately identify melanoma cases by analyzing features derived from the BoVW dictionary's melanoma cluster centers. The last 500 cluster centers in this dictionary are derived from feature vectors of melanoma image patches. Given that Table 20 highlights features originating from indexes above 500, this indicates that the framework leverages melanoma indicative features from the BoVW dictionary.

Furthermore, examining the image patches linked to the top three melanoma indicative features reveals the framework's focus on certain specific clinical characteristics.

In future research, following our framework's methodology could allow for a more detailed analysis of each individual feature. Ultimately, this could lead to a more streamlined BoVW dictionary, with a reduced number of elements, thereby crafting a framework more finely tuned for melanoma detection.

CHAPTER 5: CONCLUSION

In this thesis, a novel skin lesion classification framework, designed for the automated detection of melanoma from dermoscopic images, is presented. This framework, named the “Weighted Bag of Visual Words with Enhanced Features”, is the culmination of a series of iterative improvements. Each iteration involved updating the framework, testing it, and then deciding whether to retain the update or revert and try a different approach. The final product is a robust and high-performing melanoma detection system that leverages a BoVW methodology enriched with enhanced deep learning features.

The framework operates through a six-stage pipeline. It begins with the preprocessing of input images, followed by the generation of enhanced image patches. These patches are then used for feature extraction via the ResNet-101 network. The extracted feature vectors are clustered to create the BoVW dictionary, which in turn is used to produce weighted image histograms. Finally, these histograms are classified using a SVM classifier.

When compared with the top-10 state-of-the-art from the ISIC 2017 Challenge, the framework demonstrated superior performance. It achieved an accuracy value of 0.962, significantly higher than the nearest competitor’s 0.872. This impressive performance is further bolstered by five additional statistical metrics, highlighting the framework’s effectiveness in accurately diagnosing melanoma cases.

An additional advantage of the framework is its relatively low demand for computational resources and time, particularly during the training phase. The entire pipeline can be trained in just a few hours, even on an eight-year-old Intel i7-4790K CPU, making it both time-efficient and practical. This efficiency ensures that the framework can be readily deployed in case of a change in the framework or an update to dataset for classifying test images, offering a timely and effective tool for melanoma detection.

The modular design of the presented framework significantly contributes to its adaptability and potential for future enhancements. This design principle ensures that each step of the framework operates as a distinct unit, with interfaces that interact seamlessly with other components. As a consequence, modifications or improvements to any single step can be made without disrupting the overall pipeline, as long as the input and output parameters remain compatible with the rest of the system.

For instance, the feature extraction step, which currently utilizes ResNet-101 to process image patches and produce feature vectors, can be easily substituted with a different network model. This swap could be executed without affecting other steps of the framework, such as the preprocessing step that supplies the image patches to this step or the BoVW dictionary generation that utilizes the output feature vectors. Newer and potentially more advanced network models, like those mentioned by Liu et al. (2022), could be integrated to explore improvements in feature extraction step. Moreover, the modular nature of the framework also allows for more nuanced refinements within each step. For example, the relatively lower specificity (SPEC) values observed could be addressed by introducing an additional refinement step during the BoVW dictionary generation. This step would involve identifying and resolving overlaps in clusters that have conflicting labels. An automated algorithm could be developed to reassess and adjust benign cluster centers, enhancing the overall accuracy of the framework.

Having being mentioned, the choice of clustering method for generating the BoVW dictionary is also a critical aspect of the framework that could benefit from further refinement. Currently, K-Means, a linear clustering algorithm, is employed. However, considering the high dimensionality of the feature vectors and the fact that they are clustered per class but used in a combined manner, there might be outlier instances where cluster centers from both classes are too closely aligned. This proximity can lead to misclassification, where a feature vector indicative of melanoma might be incorrectly labeled as benign, or vice versa. Implementing new ways to set K value or exploring non-linear clustering methods, as discussed in Section 4.6, could potentially mitigate this issue by accommodating the complex nature of the data more effectively.

In summary, this thesis presents an innovative and robust framework for the automated detection of melanoma from dermoscopic images. Utilizing the BoVW approach, this framework not only competes with but also surpasses existing state-of-the-art methodologies in the ISIC context. Its standout features include combination of traditional and new techniques, a modular structure that facilitates easy modifications and a comparatively lower demand for computational resources and processing time. Additionally, the integration of contemporary concepts like XAI enhances the framework's explanatory precision, adding to its appeal.

This combination of features positions the framework as a significant and refreshing contribution to the field, offering an alternative that holds promise for future development. While the framework achieves exceptionally high classification results, the pursuit of perfection remains ongoing. The ultimate goal is to perfect the model to the extent that it becomes an invaluable tool for dermatologists in clinical practice, aiding in accurate and efficient diagnoses.

REFERENCES

Abbas, Q., Garcia, I. F., Celebi, M. E., Ahmad, W. and Mushtag, Q. (2013) *A perceptually oriented method for contrast enhancement and segmentation of dermoscopy images*, Skin Res. Technol., Vol. 19 (1), pp. 490–497.

Adadi, A. and Berrada, M. (2018) *Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)*, IEEE Access, Vol. 6, pp. 52138–52160.

Adegun, A. and Viriri, S. (2021) *Deep learning techniques for skin lesion analysis and melanoma cancer detection: A survey of state-of-the-art*, Artif. Intell. Rev., Vol. 54.

Ain, Q. U., Xue, B., Al-Sahaf, H. and Zhang, M. (2019) Multi-tree genetic programming with a new fitness function for melanoma detection, *IEEE Cong. Evolutionary Comput.*, pp. 880–887.

Airley, R. (2009) *Cancer Chemotherapy: Basic Science to the Clinic*, Wiley.

Akram, M. U., Tariq, A., Khalid, S., Javed, M. Y., Abbas, S. and Yasin, U. (2015) *Glaucoma detection using novel optic disc localization, hybrid feature set and classification techniques*, Australasian Phys. Eng. Sci. Med., Vol. 38 (4), pp. 643–655.

American Academy of Dermatology (AAD) (2022) *Skin cancer*. [Online]. Available at: https://www.aad.org/media/stats-skin-cancer. (Accessed: 30 October 2023).

American Cancer Society (2017) *Cancer facts and figures in USA 2017*. [Online]. Available at: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2017/cancer-facts-and-figures-2017.pdf. (Accessed: 30 October 2023).

American Cancer Society (2023) *Key statistics for melanoma skin cancer*. [Online]. Available at: https://www.cancer.org/cancer/types/melanoma-skin-cancer/about/key-statistics.html. (Accessed: 30 October 2023).

Anaconda Inc. (2020) *Anaconda Software Distribution*. [Online]. Available at: https://www.anaconda.com/. (Accessed: 13 November 2023).

Argenziano, G., Soyer, H. P., Chimenti, S., Talamini, R., Corona, R., Sera, F., Binder, M., Cerroni, L., Rosa, G. D., Ferrara, G. and Hofmann-Wellenhof, R. (2003) *Dermoscopy of pigmented skin lesions: results of a consensus meeting via the internet*,

J. Am. Acad. Dermatol., Vol. 48 (9), pp. 679–693.

Argenziano, G., Soyer, H. P., Giorgio, V. D., Piccolo, D., Carli, P., Delfino, M., Ferrari, A., Hofmann-Wellenhof, R., Massi, D., Mazzocchetti, G., Scalvenzi, M. and Wolf, I. H. (2000) *Interactive Atlas of Dermoscopy*, Milan: Edra Medical Publishing and New Media.

Arthur, D. and Vassilvitskii, S. (2006) How slow is the k-means method?, *Proceedings of the Twenty-Second Annual Symposium on Computational Geometry*, p. 144–153. https://doi.org/10.1145/1137856.1137880

Baral, B., Gonnade, S. and Verma, T. (2014) *Lesion segmentation in dermoscopic images using decision based neuro fuzzy model*, IJCSIT, Vol. 5 (2), pp. 2546–2552.

Barata, C., Ruela, M., Francisco, M., Mendonca, T. and Marques, J. S. (2014) *Two systems for the detection of melanomas in dermoscopy images using texture and color features*, IEEE Syst. J., Vol. 8 (3), pp. 965–979.

Battiti, R. (1994) *Using mutual information for selecting features in supervised neural net learning*, IEEE Trans. Neural Netw., Vol. 5 (4), pp. 537–550.

Bi, L., Kim, J., Ahn, E. and Feng, D. (2017) *Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks*. [Online]. Available at: https://arxiv.org/abs/1703.04197. (Accessed: 05 November 2023).

Bradley, D. and Roth, G. (2007) *Adaptive thresholding using the integral image*, J. Graph. Tools, Vol. 12 (2), pp. 13–21.

Bratkova, M., Boulos, S. and Shirley, P. (2009) *oRGB: A practical opponent color space for computer graphics*, IEEE Comput. Graph. Appl., Vol. 29 (1), pp. 42–55.

Braun, R. P., Rabinovitz, H. S., Oliviero, M., Kopf, A. W. and Saurat, J.-H. (2005) *Dermoscopy of pigmented skin lesions*, J. Am. Acad. Dermatol., Vol. 52 (1), pp. 109–121.

Bryt, O. and Elad, M. (2008) *Compression of facial images using the K-SVD algorithm*, J. Visual Commun. Image Represent., Vol. 19 (4), pp. 270–283.

Castillo, O. and Melin, P. (2008) *Type-2 Fuzzy Logic: Theory and Applications*, Springer.

Celebi, M. E., Kingravi, H. A., Uddin, B., Iyatomi, H., Aslandogan, Y. A., Stoecker, W. V. and Moss, R. H. (2007) *A methodological approach to the classification of* *dermoscopy images*, Comp. Med. Imag. and Graph., Vol. 31 (1), pp. 362–373.

Celebi, M. E., Schaefer, G., Iyatomi, H. and Stoecker, W. V. (2009) *Lesion border detection in dermoscopy images*, Comp. Med. Imag. Graph., Vol. 33 (2), pp. 148–153.

Celebi, M. E., Wen, Q., Hwang, S., Iyatomi, H. and Schaefer, G. (2013) *Lesion border detection in dermoscopy images using ensembles of thresholding methods*, Skin Res. Technol., Vol. 19 (1), pp. 252–258.

Celebi, M. E., Schaefer, G., Iyatomi, H., Stoecker, W. V., Malters, J. M. and Grichnik, J. M. (2009) *An improved objective evaluation measure for border detection in dermoscopy images*, Skin Res. Technol., Vol. 15 (4), pp. 444–450.

Chen, G., Chang, I. and Yeh, H. (2017) Action segmentation based on bag-of-visual-words models, *2017 10th International Conference on Ubi-media Computing and Workshops (Ubi-Media)*, pp. 1–5.

Choi, S. and Han, S. (2013) Evaluating weighting schemes for adult image detection using bag of visual words, *Int. Conf. ICT Conv.*, pp. 815–816.

Chollet, F. (2016) *Xception: Deep learning with depthwise separable convolutions*, IEEE Conf. Comp. Vis. Pat. Rec., Vol. pp. 1800–1807.

Codella, N., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H. and Halpern, A. (2017) *Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC)*, arXiv: 1710:05006, Vol. .

Cortes, C. and Vapnik, V. (1995) *Support-vector networks*, Mach. Learn., Vol. 20 (3), pp. 273–297.

Dahiya, M. (2002) *The melanocytic proliferations: A comprehensive textbook of pigmented lesions*, Arch. Pathol. & Lab. Med., Vol. 126 (8), pp. 999–1000.

David, L. G. (1999) Object recognition from local scale-invariant features, *Proc. Int. Conf. Comp. Vis. Volume 2*, p. 1150.

Deng, J., Dong, W., Socher, R., Li, L., Li, K. and Fei-Fei, L. (2009) Imagenet: A large-scale hierarchical image database, *IEEE Conf. Comp. Vis. Patt. Rec.*, pp. 248–255.

DermLite (2001) *Who we are*. [Online]. Available at: https://dermlite.com/. (Accessed: 30 October 2023).

Dermoscopy.org (2003) *7-point Checklist*. [Online]. Available at: https://dermoscopy.org/consensus/2d.asp. (Accessed: 30 October 2023).

Deshmukh, J. and Bhosle, U. (2016) Sift with associative classifier for mammogram classification, *2016 International Conference on Signal and Information Processing (IConSIP)*, pp. 1–5.

DeVries, T. and Ramachandram, D. (2017) *Skin lesion classification using deep multi-scale convolutional neural networks*. [Online]. Available at: https://arxiv.org/abs/1703.01402. (Accessed: 05 November 2023).

Dhillon, I. S., Guan, Y. and Kulis, B. (2004) Kernel k-means: Spectral clustering and normalized cuts, *Proc. ACM SIGKDD Int. Conf. Know. Discov. Data Mining*, pp. 551–556.

Dinnes, J., Deeks, J. J., Grainge, M. J., Chuchu, N., di Ruffano, L. F., Matin, R. N., Thomson, D. R., Wong, K. Y., Aldridge, R. B., Abbott, R., Fawzy, M., Bayliss, S. E., Takwoingi, Y., Davenport, C., Godfrey, K., Walter, F. M. and Williams, H. C. (2018) *Visual inspection for diagnosing cutaneous melanoma in adults*, Cochrane Database Syst. Rev., Vol. 12.

Dolianitis, C., Kelly, J., Wolfe, R. and Simpson, P. (2005) *Comparative performance of 4 dermoscopic algorithms by nonexperts for the diagnosis of melanocytic lesions*, Archives of Dermatology, Vol. 141 (8), pp. 1008–1014.

Douik, A., Abdellaoui, M. and Kabbai, L. (2016) Content based image retrieval using local and global features descriptor, *Int. Conf. Adv. Technol. Signal Image Process.*, pp. 151–154.

Došilović, F. K., Brčić, M. and Hlupić, N. (2018) Explainable artificial intelligence: A survey, *2018 41st Inter. Conv. Info. Com. Tech., Elec. Mic. (MIPRO)*, pp. 0210–0215.

Elad, M. (2010) *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*, Springer.

Elad, M. and Aharon, M. (2006) *Image denoising via sparse and redundant representations over learned dictionaries*, IEEE Trans. Image Process., Vol. 15 (12), pp. 3736–3745.

Elder, D. E., Bastian, B. C., Cree, I. A., Massi, D. and Scolyer, R. A. (2020) *The 2018 world health organization classification of cutaneous, mucosal, and uveal melanoma:*

Detailed analysis of 9 distinct subtypes defined by their evolutionary pathway, Arch. Pathol. & Lab. Med., Vol. 144 (4), pp. 500–522.

Euro Melanoma (2023) *Turkiye Euromelanoma Tarama Merkezleri*. [Online]. Available at: https://www.euromelanoma.eu/tr-tr/. (Accessed: 30 October 2023).

European Commission (2020) *Skin melanoma burden in eu-27*. [Online]. Available at: https://ecis.jrc.ec.europa.eu/pdf/factsheets/Melanoma_cancer_en.pdf. (Accessed: 30 October 2023).

Fadili, M. J., Starck, J. L. and Murtagh, F. (2007) *Inpainting and zooming using sparse representations*, Computer J., Vol. 52 (1), pp. 64–79.

Frank, E., Hall, M. A., Holmes, G., Kirkby, R., Pfahringer, B. and Witten, I. H. (2005) *Weka: A machine learning workbench for data mining*, Springer, pp. 1305–1314.

Ganster, H., Pinz, A., Roehrer, R., Wildling, E., Binder, M. and Kittler, H. (2001) *Automated melanoma recognition*, IEEE Tran. Med. Imag., Vol. 20 (3), pp. 233–239.

Garnavi, R., Aldeen, M., Celebi, M. E., Varigos, G. and Finch, S. (2011) *Border detection in dermoscopy images using hybrid thresholding on optimized color channels*, Comp. Med. Imag. Graph., Vol. 35 (2), pp. 105–115.

Goyal, M., Knackstedt, T., Yan, S. and Hassanpour, S. (2020) *Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities*, Comput. Biol. Med., Vol. 127, pp. 104065.

Guo, Y., Bennamoun, M., Sohel, F., Lu, M. and Wan, J. (2014) *3D object recognition in cluttered scenes with local surface features: A survey*, IEEE Trans. Pattern Anal. Mach. Intell., Vol. 36 (11), pp. 2270–2287.

Haenssle, H., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Hassen, A. B. H., Thomas, L., Enk, A., Uhlmann, L., Alt, C., Arenbergerova, M., Bakos, R., Baltzer, A., Bertlich, I., Blum, A., B.-Billmann, T., Bowling, J., Braghiroli, N., Braun, R., B.-Bakhaya, K., Buhl, T., Cabo, H., Cabrijan, L., Cevic, N., Classen, A., Deltgen, D., Fink, C., Georgieva, I., H.-Meibodi, L.-E., Hanner, S., Hartmann, F., Hartmann, J., Haus, G., Hoxha, E., Karls, R., Koga, H., Kreusch, J., Lallas, A., Majenka, P., Marghoob, A., Massone, C., Mekokishvili, L., Mestel, D., Meyer, V., Neuberger, A., Nielsen, K., Oliviero, M., Pampena, R., Paoli, J., Pawlik, E., Rao, B., Rendon, A., Russo, T., Sadek, A., Samhaber, K., Schneiderbauer, R., Schweizer, A., Toberer, F., Trennheuser, L., Vlahova, L., Wald, A., Winkler, J., Wölbing, P. and Zalaudek, I. (2018) *Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists*, Annals of Oncology, Vol. 29 (8), pp. 1836–1842.

Hall, M. A. (2000) Correlation-based feature selection for discrete and numeric class machine learning, *Int. Conf. Machine Learn.*, pp. 359–366.

Harangi, B. (2018) *Skin lesion classification with ensembles of deep convolutional neural networks*, J. Biomed. Info., Vol. 86, pp. 25–32.

Hayashi, A., Suematsu, N., Ishida, Y. and Kanbara, T. (2005) Region-based image retrieval using wavelet transform, *Digital Image Process.*, Springer.

He, K., Zhang, X., Ren, S. and Sun, J. (2016a) Deep residual learning for image recognition, *IEEE Conf. Comp. Vis. Patt. Recog.*, pp. 770–778.

He, K., Zhang, X., Ren, S. and Sun, J. (2016b) Deep residual learning for image recognition, *IEEE Conf. Comp. Vis. Pat. Rec.*, pp. 770–778.

Huang, G., Liu, Z., Maaten, L. V. D. and Weinberger, K. Q. (2017a) Densely connected convolutional networks, *IEEE Conf. Comp. Vis. Patt. Recog.*, pp. 2261–2269.

Huang, G., Liu, Z., Maaten, L. V. D. and Weinberger, K. Q. (2017b) Densely connected convolutional networks, *IEEE Conf. Comp. Vis. Pat. Rec.*, pp. 2261–2269.

Huang, L. K. and Wang, M. J. J. (1995) *Image thresholding by minimizing the measures of fuzziness*, Pattern Recog., Vol. 28 (1), pp. 41–51.

Iam-on, N. and Garrett, S. (2010) *LinkCluE: A MATLAB package for link-based cluster ensembles*, J.Stat. Software, Vol. 36 (9), pp. 1–36.

Ichim, L., Mitrica, R.-I., Serghei, M.-O. and Popescu, D. (2023) *Detection of malignant skin lesions based on decision fusion of ensembles of neural networks*, Cancers, Vol. 15 (20).

ISIC (2016) *ISIC 2016 Challenge*. [Online]. Available at: https://challenge.isic-archive.com/landing/2016/. (Accessed: 05 November 2023).

ISIC (2017) *ISIC 2017 Leader-board*. [Online]. Available at: https://challenge.isic-archive.com/leaderboards/2017/. (Accessed: 19 November 2023).

ISIC (2019) *ISIC 2019 Challenge*. [Online]. Available at: https://challenge.isic-archive.com/landing/2019/. (Accessed: 05 November 2023).

ISIC (2022) *ISIC - Welcome to the ISIC Challenge*. [Online]. Available at: https://challenge.isic-archive.com/. (Accessed: 05 November 2023).

ISIC Archive (2022) *ISIC - ISIC Archive : The International Skin Imaging Collaboration: Melanoma Project*. [Online]. Available at: https://isic-archive.com/#. (Accessed: 03 November 2023).

Iyatomi, H., Oka, H., Saito, M., Miyake, A., Kimoto, M., Yamagami, J., Kobayashi, S., Tanikawa, A., Hagiwara, M., Ogawa, K., Argenziano, G., Soyer, H. P. and Tanaka, M. (2006) *Quantitative assessment of tumour extraction from dermoscopy images and evaluation of computer-based extraction methods for an automatic melanoma diagnostic system*, Melanoma Res., Vol. 16 (2), pp. 183–190.

Jamil, U., Khalid, S. and Akram, M. U. (2016) Dermoscopic feature analysis for melanoma recognition and prevention, *Int. Conf. Inno. Comp. Technol.*, pp. 290–295.

Jia, X. and Shen, L. (2017) *Skin lesion classification using class activation map*. [Online]. Available at: https://arxiv.org/abs/1703.01053. (Accessed: 05 November 2023).

Jing, Y., Yang, Y., Feng, Z., Ye, J., Yu, Y. and Song, M. (2020) *Neural style transfer: A review*, IEEE Trans. Vis. 'I&' Comp. Gra., Vol. 26 (11), pp. 3365–3385.

Kamaladhas, M. D. and Abitha, V. (2012) Fingerprint generation of audio signal using difference of gaussian, *2012 International Conference on Devices, Circuits and Systems (ICDCS)*, pp. 276–279.

Kapur, J. N., Sahoo, P. K. and Wong, A. K. C. (1985) *A new method for gray-level picture thresholding using the entropy of the histogram*, Comp. Vis. Graph. Imag. Process., Vol. 29 (1), pp. 273–285.

Karakaya, D., Ulucan, O. and Turkan, M. (2021) *PAS-MEF: multi-exposure image fusion based on principal component analysis, adaptive well-exposedness and saliency map*, CoRR, Vol. . https://arxiv.org/abs/2105.11809

Katsambas, A., Lotti, T., Dessinioti, C. and D'Erme, A. M. (2015) *European Handbook of Dermatological Treatments*, Springer.

Kaufman, L. and Rousseeuw, P. J. (2008) *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley.

Kittler, J. and Illingworth, J. (1986) *Minimum error thresholding*, Pattern Recog., Vol. 19 (1), pp. 41–47.

Kohonen, T. (1990) *The self-organizing map*, Proceedings of the IEEE, Vol. 78 (9).

Kononenkoand, I. and Simec, E. (1995) Induction of decision trees using relieff, *W. Math. Stat. Methods in AI*, pp. 199–220.

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012) Imagenet classification with deep convolutional neural networks, *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pp. 1097—1105.

Lee, T., Ng, V., Gallagher, R., Coldman, A. and McLean, D. (1997) *Dullrazor: A software approach to hair removal from images*, Comp. Bio. Med., Vol. 27 (6), pp. 533–543.

Li, L., Zhang, Q., Ding, Y., Jiang, H., Thiers, B. H. and Wang, Z. J. (2014) *Automatic diagnosis of melanoma using machine learning methods on a spectroscopic system*, BMC Med. Imag., Vol. 14 (1).

Li, Q. and Wang, X. (2018) Image classification based on sift and svm, *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*, pp. 762–765.

Li, Y. and Shen, L. (2017) *Skin lesion analysis towards melanoma detection using deep learning network*, Sensors, Vol. 18.

Liao, H. Y. and Sapiro, G. (2008) Sparse representations for limited data tomography, *IEEE Int. Symp. Biomed. Imag.*, pp. 1375–1378.

Liu, S., Wang, Y., Yu, Q., Liu, H. and Peng, Z. (2022) *CEAM-YOLOv7: Improved YOLOv7 based on channel expansion and attention mechanism for driver distraction behavior detection*, IEEE Access, Vol. 10, pp. 129116–129124.

Mairal, J., Elad, M. and Sapiro, G. (2008) *Sparse representation for color image restoration*, IEEE Trans. Image Process., Vol. 17 (1), pp. 53–69.

Mairal, J., Sapiro, G. and Elad, M. (2008) *Learning multiscale sparse representations for image and video restoration*, SIAM Multiscale Model. Simul., Vol. 7 (1), pp. 214–241.

Mairal, J., Bach, F., Ponce, J., Sapiro, G. and Zisserman, A. (2008) Discriminative learned dictionaries for local image analysis, *IEEE Comp. Vis. Pattern Recog.*, pp. 1–8.

Malpani, S., Asha C S and Narasimhadhan, A. V. (2016) Thermal vision human classification and localization using bag of visual word, *2016 IEEE Region 10 Conference (TENCON)*, pp. 3135–3139.

Malvehy, J., Puig, S., Argenziano, G., Marghoob, A. A. and Soyer, H. P. (2007) *Dermoscopy report: proposal for standardization: results of a consensus meeting of the international dermoscopy society*, J. Am. Acad. Dermatol., Vol. 57 (1), pp. 84–95.

Marghoob, A. A., Koenig, K., Bittencourt, F. V., Kopf, A. W. and Bart, R. S. (2000) *Breslow thickness and Clark level in melanoma*, Cancer, Vol. 88 (3), pp. 589–595.

Marin, C., Alferez, G., Cordova, J. and Gonzalez, V. (2015) Detection of melanoma through image recognition and artificial neural networks, *World Cong. Med. Physics Biomed. Eng.*, pp. 832–835.

Martinetz, T., Berkovich, S. and Schulten, K. (1993) *'neural-gas' network for vector quantization and its application to time-series prediction*, IEEE Transactions on Neural Networks, Vol. 4 (4).

Matsunaga, K., Hamada, A., Minagawa, A. and Koga, H. (2017) *Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble*, ArXiv, Vol. abs/1703.03108.

Mayo Clinic (2022) *Video: Skin cancer — how skin cancer develops*. [Online]. Available at: http://www.mayoclinic.org/diseases-conditions/skin-cancer/multimedia/melanoma/vid-20084739. (Accessed: 30 October 2023).

Mayo Clinic (2023) *Melanoma*. [Online]. Available at: https://www.mayoclinic.org/diseases-conditions/melanoma/symptoms-causes/syc-20374884. (Accessed: 30 October 2023).

Melanoma Institute Australia (2023a) *Melanoma facts*. [Online]. Available at: https://melanoma.org.au/about-melanoma/melanoma-facts/. (Accessed: 30 October 2023).

Melanoma Institute Australia (2023b) *What is melanoma?* [Online]. Available at: https://melanoma.org.au/about-melanoma/what-is-melanoma/. (Accessed: 30 October 2023).

Melanoma Research Foundation (MRF) (2023) *Melanoma facts and stats*. [Online].

Available at: https://melanoma.org/melanoma-education/understand-melanoma/facts-stats/. (Accessed: 30 October 2023).

Melanoma UK (2016) *The ABCDE Rule*. [Online]. Available at: https://www.melanomauk.org.uk/the-abcde-rule. (Accessed: 30 October 2023).

Melgani, F. (2006) *Robust image binarization with ensembles of thresholding algorithms*, Elec. Imag., Vol. 15 (2), pp. 15 – 15 – 11.

Melli, R., Grana, C. and Cucchiara, R. (2006) Comparison of color clustering algorithms for segmentation of dermatological images, *SPIE Medical Imag.*, pp. 61443S–61443S.

Mendonca, T., Ferreira, P. M., Marques, J. S., Marcal, A. R. and Rozeira, J. (2015) PH2 – A public database for the analysis of dermoscopic images, *Dermoscopy Image Anal.*, CRC Press, pp. 419–439.

Menegola, A., Fornaciali, M., Pires, R., Bittencourt, F. V., Avila, S. and Valle, E. (2017) Knowledge transfer for melanoma screening with deep learning, *IEEE Int. Symp. Biomed. Imag.*, pp. 297–300.

Menegola, A., Tavares, J., Fornaciali, M., Li, L. T., Avila, S. and Valle, E. (2017) *RECOD Titans at ISIC Challenge 2017*, CoRR, Vol. arxiv.org/abs/1703.04819.

Mete, M., Kockara, S. and Aydin, K. (2011) *Fast density-based lesion detection in dermoscopy images*, Comp. Med. Imag. Graph., Vol. 35 (2), pp. 128–136.

Mishra, N. K. and Celebi, M. E. (2016) *An overview of melanoma detection in dermoscopy images using image processing and machine learning*, ArXiv, Vol. abs/1601.07843.

Moradi Fard, M., Thonet, T. and Gaussier, E. (2020) *Deep k-Means: Jointly clustering with k-means and learning representations*, Patt. Recog. Lett., Vol. 138, pp. 185–192.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R. and Yu, B. (2019) *Definitions, methods, and applications in interpretable machine learning*, Proc. Nat. Academy Sci., Vol. 116 (44), pp. 22071–22080.

Naing, A. and Hajjar, J. (2017) *Immunotherapy*, Springer.

National Cancer Institute (2023) *Melanoma treatment PDQ–patient version*. [Online]. Available at: https://www.cancer.gov/types/skin/patient/melanoma-treatment-pdq. (Accessed: 30 October 2023).

Ng, V. T., Fung, B. Y. and Lee, T. K. (2005) *Determining the asymmetry of skin lesion with fuzzy borders*, Comp. Biol. Med., Vol. 35 (2), pp. 103–120.

Nielsen, F. (2016) *Hierarchical Clustering*, Springer, pp. 195–211.

Nobuyuki, O. (1979) *A threshold selection method from gray-level histograms*, IEEE Trans. Syst. Man. Cybern. Syst., Vol. 9 (1), pp. 62–66.

Ogorzalek, M., Nowak, L., Surowka, G. and Alekseenk, A. (2011) Modern techniques for computer-aided melanoma diagnosis, *Melanoma in the Clinic - Diagnosis, Management and Complications of Malignancy*, INTECH.

Oktar, Y. and Turkan, M. (2022) *Preserving spatio-temporal information in machine learning: A shift-invariant k-means perspective*, J. Sign. Process. Syst., Vol. 94, pp. 1471–1483.

Okur, E. and Turkan, M. (2018) *A survey on automated melanoma detection*, Eng. Appl. Artif. Intell., Vol. 73, pp. 50–67.

Okur, E. and Turkan, M. (2022) Patch enhancement for melanoma detection with bag of visual words, *Med. Tech. Cong. (TIPTEKNO)*, pp. 1–4.

Okur, E. and Turkan, M. (2024) *Weighted Bag of Visual Words with enhanced deep features for melanoma detection*, Expert Systems with Applications, Vol. 237, pp. 121531.

Parikh, R., Mathai, A., Parikh, S., Chandra, S. G. and Thomas, R. (2008) *Understanding and using sensitivity, specificity and predictive values*, Indian J. Ophthalmol., Vol. 56 (1), pp. 45–50.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. and Dubourg, V. (2011) *Scikit-learn: Machine learning in python*, Journal of machine learning research, Vol. 12 (Oct), pp. 2825–2830.

Peotta, L., Granai, L. and Vandergheynst, P. (2006) *Image compression using an edge adapted redundant dictionary and wavelets*, Signal Process., Vol. 86 (3), pp. 444–456.

Peyre, G. (2009) *Sparse modeling of textures*, J. Math. Imag. Vis., Vol. 34 (1), pp. 17–31.

Priddy, K. L. and Keller, P. E. (2005) *Artificial Neural Networks: An Introduction*, SPIE.

Protter, M. and Elad, M. (2009) *Image sequence denoising via sparse and redundant* *representations*, IEEE Trans. Image Process., Vol. 18 (1), pp. 27–35.

Quintana, J., Garcia, R. and Neumann, L. (2009) *A novel method for color correction in epiluminescence microscopy*, Comp Med Imag Graph, Vol. 35 (7), pp. 646–652.

Redmon, J. (2013) *Darknet: Open Source Neural Networks in C*. [Online]. Available at: http://pjreddie.com/darknet/. (Accessed: 16 November 2023).

Republic of Türkiye Ministry of Health (2022) *Türkiye kanser İstatistikleri 2018*. [Online]. Available at: https://hsgm.saglik.gov.tr/depo/birimler/kanser-db/Dokumanlar/Istatistikler/Kanser_Rapor_2018.pdf. (Accessed: 30 October 2023).

Rubegni, P., Cevenini, G., Burroni, M., Perotti, R., Dell'Eva, G., Sbano, P. and Miracco, C. (2002) *Automated diagnosis of pigment skin lesions*, Int. J. Cancer, Vol. 101 (6), pp. 576–580.

Rui, X. and Wunsch, D. (2005) *Survey of clustering algorithms*, IEEE Transactions on Neural Networks, Vol. 16 (3), pp. 645–678.

Saeed, F., Hussain, M. and Aboalsamh, H. A. (2018) Classification of live scanned fingerprints using dense sift based ridge orientation features, *2018 1st International Conference on Computer Applications Information Security (ICCAIS)*, pp. 1–4.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L. (2018) Mobilenetv2: Inverted residuals and linear bottlenecks, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4510–4520.

Santy, A. and Joseph, R. (2015) Segmentation methods for computer aided melanoma detection, *Global Conf. Communication Tech.*, pp. 490–493.

Schmid, C. (2001) Constructing models for content-based image retrieval, *IEEE Comp. Soc. Conf. Comp. Vis. Pattern Recog.*, Vol. 2, pp. 39–45.

Schmid, P. (1999) *Segmentation of digitized dermatoscopic images by two-dimensional color clustering*, IEEE Trans. Med. Imag., Vol. 18 (2), pp. 164–171.

Senel, E. (2011) *Dermatoscopy of non-melanocytic skin tumors*, Indian J. Dermatol. Venereol. Leprol., Vol. 77 (1), pp. 16–22.

Sezgin, M. and Sankur, B. (2004) *Survey over image thresholding techniques and quantitative performance evaluation*, J. Elec. Imag., Vol. 13 (1), pp. 146–165.

Sharma, A. K., Tiwari, S., Aggarwal, G., Goenka, N., Kumar, A., Chakrabarti, P., Chakrabarti, T., Gono, R., Leonowicz, Z. and Jasinski, M. (2022) *Dermatologist-level classification of skin cancer using cascaded ensembling of convolutional neural network and handcrafted features based deep neural network*, IEEE Access, Vol. 10, pp. 17920–17932.

Shekhar, R. and Jawahar, C. V. (2012) Word image retrieval using bag of visual words, *2012 10th IAPR International Workshop on Document Analysis Systems*, pp. 297–301.

Siegel, R. L., Miller, K. D. and Jemal, A. (2018) *Cancer statistics, 2018*, CA: A Cancer J. Clinicians, Vol. 68 (1), pp. 7–30.

Simonyan, K. and Zisserman, A. (2014) *Very deep convolutional networks for large-scale image recognition*, arXiv e-prints, Vol. .

Singh, S., Srivastava, D. and Agarwal, S. (2017) GLCM and its application in pattern recognition, *Int. Symp. Comput. Business Intell.*, pp. 20–25.

Situ, N., Yuan, X., Chen, J. and Zouridakis, G. (2008) Malignant melanoma detection by bag-of-features classification, *IEEE Int. Conf. Eng. Med. Biol. Soc.*, pp. 3110–3113.

Sousa, R. T. and de Moraes, L. V. (2017) *Araguaia medical vision lab at isic 2017 skin lesion classification challenge*, ArXiv, Vol. abs/1703.00856.

Steinwart, I. and Christmann, A. (2008) *Support Vector Machines*, Springer.

Stewart, C. (2023) *Skin cancer in europe - statistics and facts*. [Online]. Available at: https://www.statista.com/topics/11101/skin-cancer-in-europe/#topicOverview. (Accessed: 30 October 2023).

Sundararajan, M. and Najmi, A. (2019) The many shapley values for model explanation, *Int. Conf. Mach. Lear.*

Szegedy, C., Ioffe, S., Vanhoucke, V. and Alemi, A. (2017) Inception-v4, inception-resnet and the impact of residual connections on learning, *Proc. Thirty-First AAAI Conf. Art. Int.*, AAAI Press, pp. 4278—-4284.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z. (2016) Rethinking the inception architecture for computer vision, *IEEE Conf. Comp. Vis. Pat. Rec.*, pp. 2818–2826.

Szegedy, C., Wei, L., Yangqing, J., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A. (2015) Going deeper with convolutions, *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9.

Szegedy, C., Ioffe, S. and Vanhoucke, V. (2016) *Inception-v4, inception-resnet and the impact of residual connections on learning*, CoRR, Vol. arxiv.org/abs/1602.07261.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z. (2015) *Rethinking the inception architecture for computer vision*. [Online]. Available at: https://arxiv.org/pdf/1512.00567.pdf. (Accessed: 05 November 2023).

Tan, M. and Le, Q. (2019) EfficientNet: Rethinking model scaling for convolutional neural networks, *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97, pp. 6105–6114.

Tenenhaus, A., Nkengne, A., Horn, J., Serruys, C., Giron, A. and Fertil, B. (2010) *Detection of melanoma from dermoscopic images of naevi acquired under uncontrolled conditions*, Skin Res. Technol., Vol. 16 (1), pp. 85–97.

The MathWorks Inc. (2022) *Matlab version: 9.13.0 (r2022b)*. [Online]. Available at: https://www.mathworks.com. (Accessed: 13 November 2023).

The University of Edinburgh (2013) *Dermofit image library*. [Online]. Available at: https://licensing.edinburgh-innovations.ed.ac.uk/product/dermofit-image-library. (Accessed: 30 October 2023).

Tkalcic, M. and Tasic, J. F. (2003) Colour spaces: perceptual, historical and applicational background, *IEEE EUROCON Computer as a Tool*, Vol. 1, pp. 304–308.

Tschandl, P., Rosendahl, C. and Kittler, H. (2018) *The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions*, Scientific Data, Vol. 5.

Turkiye Kanserle Savas Vakfi (2021) *Melanom cilt kanseri*. [Online]. Available at: https://www.kanservakfi.com/kanser-turleri/melanom-cilt-kanseri/. (Accessed: 30 October 2023).

van de Sande, K. E. A., Gevers, T. and Snoek, C. G. M. (2010) *Evaluating color descriptors for object and scene recognition*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 32 (9), pp. 1582–1596.

Vasconcelos, C. N. and Vasconcelos, B. N. (2017) *Convolutional neural network committees for melanoma classification with classical and expert knowledge based image transforms data augmentation*, arXiv: Computer Vision and Pattern Recognition, Vol. .

Wang, J., Chen, Q. and Chen, Y. (2004) RBF kernel based support vector machine with universal approximation and its application, *Advances in Neural Networks*, Springer Berlin Heidelberg, pp. 512–517.

Washington, C. M. and Leaver, D. T. (2016) *Principles and Practice of Radiation Therapy*, Elsevier.

WebMD (2023) *Moles and skin cancer screening*. [Online]. Available at: https://www.webmd.com/melanoma-skin-cancer/screening-moles-cancer#1. (Accessed: 30 October 2023).

Weitz, E. (2016) *SIFT - Scale-Invariant Feature Transform*. [Online]. Available at: http://weitz.de/sift/. (Accessed: 13 November 2023).

WHO (2017) *Radiation: Ultraviolet (uv) radiation and skin cancer*. [Online]. Available at: https://www.who.int/news-room/questions-and-answers/item/radiation-ultraviolet-(uv)-radiation-and-skin-cancer. (Accessed: 30 October 2023).

Wight, P., Lee, T. K., Lui, H. and Atkins, M. S. (2011) *Chromatic aberration correction: an enhancement to the calibration of low-cost digital dermoscopes*, Skin Res. Technol., Vol. 17 (3), pp. 339–347.

Wong, A. (2011) *Automatic skin lesion segmentation via iterative stochastic region merging*, IEEE Trans. Info. Tech. Biomed., Vol. 15 (6), pp. 929–936.

Wu, Y., Xie, F., Jiang, Z. and Meng, R. (2013) Automatic skin lesion segmentation based on supervised learning, *Int. Conf. Imag. Graph.*, pp. 164–169.

Yan, L., Rosen, N. and Arteaga, C. (2011) *Targeted cancer therapies*, Chin. J. Cancer, Vol. 30 (1), pp. 1–4.

Yang, J., Jiang, Y., Hauptmann, A. G. and Ngo, C. (2007) Evaluating bag-of-visual-words representations in scene classification, *Proc. Int. Work. Mult. Inf. Ret.*, pp. 197—-206.

Yang, X., Zeng, Z., Yeo, S. Y., Tan, C., Tey, H. L. and Su, Y. (2017) *A novel multi-task deep learning model for skin lesion segmentation and classification*, ArXiv, Vol. abs/1703.01025.

Yuan, X., Situ, N. and Zouridakis, G. (2009) *A narrow band graph partitioning method for skin lesion segmentation*, Pattern Recog., Vol. 42 (6), pp. 1017–1028.

Yuksel, M. E. and Borlu, M. (2009) *Accurate segmentation of dermoscopic images* *by image thresholding based on type-2 fuzzy logic*, IEEE Trans. Fuzzy. Syst., Vol. 17 (4), pp. 976–982.

Zhang, Y., Jin, R. and Zhou, Z.-H. (2010) *Understanding bag-of-words model: A statistical framework*, International Journal of Machine Learning and Cybernetics, Vol. 1, pp. 43–52.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. and Torralba, A. (2015) *Learning deep features for discriminative localization*. [Online]. Available at: https://arxiv.org/abs/1512.04150. (Accessed: 05 November 2023).

Zhou, Y. and Song, Z. (2013) Binary decision trees for melanoma diagnosis, *Int. W. Multiple Classifier Syst.*, pp. 374–385.

Zhou, Y. and Song, Z. (2014) Melanoma diagnosis with multiple decision trees, *Comp. Vis. Tech. Diag. Skin Cancer*, pp. 267–282.