Novel Deep Learning Frameworks for Image Classification-based Computer Vision Tasks

Thesis submitted in partial fulfillment of the requirements for the degree of

Master of Technology
in
Computer Technology

Submitted by

Soumyajit Gayen
Regn. No. - 160083 of 2021 - 2022
Exam Roll No. - M6TCT24009
Class Roll No. - 002110504012
Session - 2021 - 2024

Under the supervision of

Prof. Ram Sarkar

Department of Computer Science and Engineering
Jadavpur University
188, Raja S.C. Mallick Rd,
Kolkata - 700032, West Bengal, India

Department of Computer Science and Engineering
Faculty of Engineering and Technology
Jadavpur University, Kolkata - 700032

CERTIFICATE OF RECOMMENDATION

This is to certify that the work embodied in this thesis entitled **"Novel Deep Learning Frameworks for Image Classification-based Computer Vision Tasks"** has been satisfactorily completed by **Soumyajit Gayen** (Registration Number 160083 of 2021 - 2022; Class Roll No. 002110504012; Examination Roll No. M6TCT24009). It is a bonafide piece of work carried out under my supervision and guidance at Jadavpur University, Kolkata, for partial fulfillment of the requirements for the awarding of the **Master of Technology** in **Computer Technology** degree of the Department of Computer Science and Engineering, Faculty of Engineering and Technology, Jadavpur University, during the academic year 2021 - 2024.

Prof. Ram Sarkar,
Department of Computer Science and Engineering,
Jadavpur University.
(Supervisor)

Forwarded By:

Prof. Nandini Mukhopadhyay,
Head, Department of Computer Science and Engineering,
Jadavpur University.

Prof. Dipak Laha,
DEAN, Faculty of Engineering and Technology,
Jadavpur University.

Department of Computer Science and Engineering
Faculty of Engineering and Technology
Jadavpur University, Kolkata, 700032

CERTIFICATE OF APPROVAL

This is to certify that the thesis entitled **"Novel Deep Learning Frameworks for Image Classification-based Computer Vision Tasks"** is a bonafide record of work carried out by **Soumyajit Gayen** (Registration Number 160083 of 2021 - 2022; Class Roll No. 002110504012; Examination Roll No. M6TCT24009) in partial fulfillment of the requirements for the award of the degree of **Master of Technology** in **Computer Technology** in the Department of Computer Science and Engineering, during the period of September 2021 to May 2024. It is understood that by this approval, the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein, but approve the thesis only for the purpose of which it has been submitted.

Signature of Examiner 1
Date:

Signature of Examiner 2
Date:

Department of Computer Science and Engineering
Faculty of Engineering and Technology
Jadavpur University, Kolkata, 700032

DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC ETHICS

I hereby declare that the thesis entitled **"Novel Deep Learning Frameworks for Image Classification-based Computer Vision Tasks"** contains a literature survey and original research work by the undersigned candidate, as a part of his degree of **Master of Technology** in **Computer Technology, Jadavpur University**. All the information has been obtained and presented in accordance with academic rules and ethical conduct.
I also declare that, as required by these rules and conduct, I have fully cited and referenced all materials and results that are not original to this work.

Name: Soumyajit Gayen

Examination Roll No.: M6TCT24009

Registration No.: 160083 of 2022 - 2023

Thesis Title: Novel Deep Learning Frameworks for Image Classification-based Computer Vision Tasks

Signature of the Candidate:

. . . . . .

Dedicated to my unwavering pillars of support,
my parents
whose love and sacrifice have fueled every chapter of my journey.

. . . . . .

Acknowledgements

I tender my deepest gratitude and high sense of indebtedness to my guide, **Dr. Ram Sarkar**, Professor, Department of Computer Science and Engineering, Jadavpur University, Kolkata, India, for his constant support and supervision in carrying out this thesis, which has been instrumental in the materialization of this work. His expertise and guidance were crucial in shaping both the direction and the quality of this thesis, and I am profoundly thankful for his steadfast dedication.

I express my sincere gratitude to the Head of the Department, **Dr. Nandini Mukhopadhyay**, Professor, Department of Computer Science and Engineering, Jadavpur University, Kolkata, India, for being kind enough to permit me to carry out my research at this prestigious institution.

I must express my sincere gratitude to **Prof. Mahantapas Kundu**, Coordinator of **Centre for Microprocessor Applications for Training, Education, and Research (CMATER) Lab**, Department of Computer Science and Engineering, Jadavpur University, Kolkata, India, for permitting me to conduct this study in the renowned CMATER lab and rendering valuable suggestions.

My special thanks to my teachers, **Prof. Debotosh Bhattacharjee**, **Prof. Subhadip Basu**, and **Dr. Mita Nasipuri (former Professor)**, Department of Computer Science and Engineering, Jadavpur University, Kolkata, India, for imparting their wealth of knowledge to me.

I thank my seniors, Mr. Sujan Sarkar, Mr. Sourajit Maity, Dr. Neelotpal Chakraborty, Mr. Debam Saha, and Mrs. Rakhi Tewari, as well as my labmates, Mr. Amartya Ray, Mr. Akash Halder, Mr. Diptarka Mandal, and Mr. Ankit Das, for their unwavering support, validation, and understanding throughout this journey. Their companionship and camaraderie have been invaluable to me.

I must also extend a special word of thanks to my parents and family members for the encouragement and support they have rendered in preparing this thesis.

Soumyajit Gayen

Examination Roll No.: M6TCT24009

Registration No.: 160083 of 2021 - 2022

Department of Computer Science and Engineering,

Jadavpur University.

Abstract

In recent years, there has been a notable surge in interest from the research community regarding the improvement and refinement of image classification techniques. The thesis at hand specifically delves into two pivotal areas within these domains: vehicle make and model recognition, and the classification of skin cancer. These areas have been selected due to their critical importance and the potential impact that advancements in image classification methodologies can have within these fields.

Vehicle make and model recognition is crucial for automatic vehicle recognition systems. It has received significant attention from the computer vision and artificial intelligence communities in recent decades. Previous research has focused on improving recognition accuracy through attention mechanisms, which, while effective, often increase complexity and computational burden unnecessarily. A novel deep neural network model, named the Simple Sequential Attention Network (SimSANet), is proposed to address this challenge. SimSANet integrates a sequential multi-kernel approach to balance complexity and performance, effectively reducing computational load while efficiently capturing essential information from feature maps of varying scales. The proposed approach exhibits superior performance, as evidenced by experimental results on publicly accessible VMMR datasets such as Stanford Cars, CompCars, CompCarsSV, and the Vehicle Images dataset.

For medical contexts like skin cancer classification, the accurate identification of skin abnormalities is imperative for effective diagnosis and treatment. Traditional convolutional neural network based ensemble approaches, while effective, often incur high computational costs due to the combination of multiple models. In contrast, snapshot-based approaches offer a streamlined alternative by utilizing a single baseline model. To mitigate computational burden without sacrificing performance, SEFFNet, a novel snapshot ensemble-based feature fusion network coupled with a machine learning-based classifier, is proposed. SEFFNet captures diverse feature representations at different training stages, eliminating the need for multiple models and refraining from data augmentation and pre-processing techniques. The effectiveness of SEFFNet is demonstrated through experimental assessments conducted on publicly accessible datasets such as HAM10000 and ISIC 2017.

Keywords: Vehicle Model and Make Recognition, Skin Cancer Classification, Machine Learning, Deep Learning, Snapshot Ensemble, Feature Fusion, Attention Mechanism.

Contents

Abstract i

Table of Contents iii

List of Figures iv

List of Tables viii

1 Introduction 1
- 1.1 Vehicle Make and Model Recognition . . . 2
- 1.2 Skin Cancer Classification . . . 3
- 1.3 Scope of the Thesis . . . 5
- 1.4 Organization of the Thesis . . . 6

2 Literature Survey 9
- 2.1 Vehicle Make and Model Recognition . . . 9
- 2.2 Skin Cancer Classification . . . 11

3 SimSANet: A Method for VMMR Classification 13
- 3.1 Proposed Architecture . . . 13
  - 3.1.1 Fundamental Feature Extraction Module . . . 13
  - 3.1.2 Multi-headed SA Module . . . 14
    - 3.1.2.1 SimAM: A Simple Parameter-free Attention Module 14
    - 3.1.2.2 SA Module . . . 15
  - 3.1.3 Classification Module . . . 18
- 3.2 Materials and Implementation Details . . . 19
  - 3.2.1 Datasets . . . 19
  - 3.2.2 Data Pre-processing and Base Model Selection . . . 19
  - 3.2.3 Experimental Setup . . . 20
  - 3.2.4 Evaluation Metrics . . . 21
- 3.3 Results and Discussion . . . 21
  - 3.3.1 Performance of SimSANet on various VMMR datasets . . . 21
  - 3.3.2 Quantitative Analysis on the Stanford Cars Dataset . . . 22
  - 3.3.3 Quantitative Analysis on the CompCars Dataset . . . 24
  - 3.3.4 Quantitative Analysis on the CompCarsSV Dataset . . . 24
  - 3.3.5 Quantitative Analysis on the Vehicle Images Dataset . . . 25
  - 3.3.6 Analysis of Accuracy-Loss curves on various VMMR datasets 25

3.3.7 Analysis of Grad-CAM based feature extraction on various VMMR datasets . . . 26
3.3.8 Analysis of Misclassified Images on various VMMR datasets 28
3.3.9 Ablation Study . . . 28
3.3.9.1 Ablation Study on Model's Architecture . . . 29
3.3.9.2 Ablation Studies on Varying Hyperparameters . . . 30
3.3.10 Statistical Analysis . . . 31

4 SEFFNet: A Method for Skin Cancer Classification 33
4.1 Proposed Architecture . . . 33
4.1.1 Deep Feature Extraction Module . . . 34
4.1.2 Feature Fusion Module . . . 37
4.1.3 Classification Module . . . 38
4.1.4 Hybrid loss Function . . . 38
4.2 Materials and Implementation Details . . . 40
4.2.1 Datasets . . . 40
4.2.2 Implementation Details . . . 41
4.2.2.1 Experimental Setup . . . 41
4.2.2.2 Evaluation Metrics . . . 41
4.3 Results . . . 41
4.3.1 Performance of different baseline models . . . 41
4.3.2 Quantitative Analysis of Results . . . 42
4.3.3 Comparative Analysis on the HAM10000 Dataset . . . 45
4.3.4 Comparative Analysis on the ISIC 2017 Dataset . . . 46
4.3.5 Impact of the proposed Hybrid loss function . . . 47
4.3.6 Model Interpretability using Grad-CAM . . . 49
4.3.7 Results of the Ablation Study . . . 50
4.3.8 Statistical Analysis . . . 51

5 Conclusion 53
5.1 Limitations and Open challenges . . . 54
5.2 Future Research Scopes . . . 55

List of Publications 57

Bibliography 59

List of Figures

3.1 Overall pipeline of the proposed SimSANet model used for VMMR. 14
3.2 Architectural overview of the proposed SA module. The basis of this generated module is the information extraction at various levels. Each block's layer dimension and associated parameters have been omitted from this module for better visualization. 16
3.3 Architecture of the generic SMDC block used in the present work. . 17
3.4 Diagram showing the row-wise and column-wise SMDC blocks. . . . 17
3.5 Few sample images from various VMMR datasets 19
3.6 Accuracy and loss curves generated by the proposed model on four standard VMMR datasets. 26
3.7 Visualization of feature maps on four standard VMMR datasets using Grad-CAM analysis. 27
3.8 Some misclassified images with their true class and predicted class. 29

4.1 The overview of proposed SEFFNet. This architecture includes three main modules: Deep Feature Extraction, Feature Fusion, and Classification Modules. Utilizing the Cosine Annealing Learning Rate scheduling, the deep feature extraction module extracts features (i.e., snapshots) at different training stages, which are then fused and employed for classification via a Random Forest classifier. 34
4.2 The overview of the CBAM block. 35
4.3 The overview of the TA block. 36
4.4 Few examples of Skin Cancer Samples from Diverse Datasets. . . . 40
4.5 Accuracy and loss curves generated by the proposed method on HAM10000 and the ISIC 2017 datasets. 44
4.6 Confusion matrices obtained for HAM10000 and ISIC 2017 datasets. 44
4.7 Results using different loss functions on the HAM10000 and ISIC 2017 datasets. 48
4.8 Grad-CAM results on various samples images. 49

List of Tables

3.1 Statistics of the VMMR datasets used in the present experimentation. 20
3.2 Comparison of classification results and number of parameters on the Stanford Cars dataset using various baseline models 20
3.3 Comparison of the classification results on the Stanford Cars dataset. 23
3.4 Comparison of the classification results on the CompCars dataset. 24
3.5 Comparison of the classification results on the CompCarsSV dataset. 25
3.6 Comparison of the classification results on the Vehicle Images. . . 25
3.7 Experimental results of different layout methods of proposed SimSANet on Stanford Cars dataset 30
3.8 Experimental results on the Stanford Cars dataset with varying hyperparameters 31
3.9 Statistical comparison of the proposed model with other models on four VMMR datasets based on p-values of the Wilcoxon rank-sum test. 32

4.1 Statistics of the datasets used in the present experimentation. . . . 40
4.2 The performance of different baseline models on the HAM10000 and ISIC 2017 datasets. All scores are in %. Bold values represent the best results. 42
4.3 Results after the basic feature extraction module on the HAM10000 and ISIC 2017 datasets. All scores are in %. Bold values represent the best results. 43
4.4 Final classification results after employing the SelectKBest approach and ML classifiers on the HAM10000 and ISIC 2017 datasets. All scores are in %. Bold values represent the best results. 43
4.5 Comparison of the classification results on the HAM10000 dataset. All scores are in %. Bold values represent the best results. 45
4.6 Comparison of the classification results on the ISIC 2017 dataset. All scores are in %. Bold values represent the best results. 47
4.7 The performance of different loss functions on the HAM10000 and ISIC 2017 datasets. All scores are in %. Bold values represent the best results. 48

4.8 Experimental results of different layouts of the proposed SEFFNet on the HAM10000 dataset. All scores are in %. Bold values represent the best results. 51
4.9 Experimental results of different layouts of the proposed SEFFNet on the ISIC 2017 dataset. All scores are in %. Bold values represent the best results. 51
4.10 Statistical comparison of SEFFNet layout variants on two skin cancer datasets using Wilcoxon Rank-Sum test in terms of p-values. . . 52

CHAPTER 1

Introduction

Image classification, a fundamental task in computer vision, involves assigning labels to images based on their content. Over the past decade, the advent of deep learning has revolutionized this field, enabling significant advancements across various domains. Deep learning approaches, particularly convolutional neural networks (CNNs), have demonstrated exceptional performance in image classification tasks due to their ability to automatically extract and learn hierarchical features from raw image data. However, image classification using deep learning has found widespread applications across diverse domains. In retail and e-commerce, image classification algorithms are employed for product recognition and recommendation systems. By analyzing product images, these systems can automatically categorize items, recommend similar products to users, and personalize shopping experiences. Similarly, in agriculture, image classification aids in crop monitoring, disease detection, and yield prediction. By analyzing aerial or satellite images of farmland, farmers can identify areas requiring intervention, optimize resource allocation, and maximize crop yields. Furthermore, image classification plays a crucial role in the field of security and surveillance. Surveillance cameras equipped with image classification algorithms can detect and recognize objects of interest, such as intruders, suspicious behavior, or unauthorized vehicles. These systems enable real-time monitoring and alerting, enhancing security measures in public spaces, airports, and critical infrastructure facilities. Moreover, image classification has made significant strides in environmental monitoring and conservation efforts. Remote sensing technologies coupled with deep learning algorithms allow researchers to analyze satellite imagery to track changes in land cover, monitor deforestation, and assess habitat health. This information is invaluable for conservation initiatives, biodiversity assessments, and environmental policy-making.

In this thesis, the focus is on two major domains where image classification using deep learning techniques has shown considerable promise: vehicle make and model recognition (VMMR) and medical skin cancer image classification. These domains represent critical areas where accurate and efficient image analysis can have a profound impact on industries and society as a whole. By leveraging deep learning methodologies in these domains, the advancement of image classification techniques is aimed at, addressing real-world challenges in the VMMR and skin cancer classification problems.

1.1 Vehicle Make and Model Recognition

First, the VMMR domain will be introduced, focusing on the importance of accurate VMMR applications such as automatic vehicle recognition systems. Enhancements in the VMMR domain can lead to improved traffic management, enhanced security measures, and better customer experiences in transportation services. Next, the medical domain will be explored, emphasizing the critical role of skin cancer image classification in early diagnosis and effective treatment. Advanced image classification in this field can lead to improved patient outcomes and more efficient healthcare delivery. By discussing these introductions separately, a comprehensive overview of the specific challenges and opportunities in each domain will be provided, setting the stage for the detailed exploration of the proposed methodologies.

In the VMMR domain, accurate classification of vehicles is crucial for various applications such as surveillance, traffic monitoring, and autonomous driving. Deep learning methods offer a promising approach to address the challenges associated with vehicle classification, including variations in appearance, occlusions, and background clutter. By exploring different deep learning architectures and methodologies tailored to vehicle make and model classification, this thesis aims to improve the accuracy and robustness of vehicle recognition systems. The vehicle industry has become one of the largest industries in the world, and it has been witnessing rapid growth, especially with the implementation of intelligent transport systems. Vehicle manufacturing has served as a significant revenue source for any country and has played a vital role in the economies of many developing nations. Therefore, a large number of vehicle makers and their models can be found in the industry. For this reason, research on developing automatic VMMR systems is much needed in today's scenario. With the essential progress of numerous artificial intelligence (AI) systems and computer vision tasks, vehicle monitoring has played an important role in our daily lives. There have been multiple applications such as vehicle management, vehicle behavior analysis, vehicle detection [1] [2],

vehicle classification [3], vehicle segmentation [4], vehicle surveillance [5], vehicle make and model recognition [6], etc. With the increasing number of self-driving cars and intelligent transport systems, research on VMMR has witnessed a significant increment in the last decade. Especially, VMMR is crucial for developing vehicle surveillance systems and security reasons. Various factors are important to ensure the robustness of a VMMR model. One such factor is a realistic dataset in the model training process, covering diverse classes and images taken in different weather conditions and times, encompassing both night-time and daytime scenarios. Publicly available datasets are rarely found, and when they are, the absence of annotations renders them unsuitable for model training. Some datasets have been identified, where the number of classes is very limited, and the absence of images in various weather conditions makes them less beneficial for developing realistic models. Furthermore, present VMMR datasets do not fully depict real-life circumstances. For example, due to heavy traffic congestion on the road, numerous vehicles overlap in a single frame in the Indian traffic scene. Because of this issue, the classification of make and model, localization, detection, and segmentation of vehicles become extremely challenging. These are some challenges that are not only in India, but also in Bangladesh, Pakistan, Sri Lanka, and many other South Asian nations. Nevertheless, formulating and executing VMMR systems has posed a challenging endeavor. Despite technological advancements enhancing its functionality and effectiveness, multiple challenges must have been meticulously addressed. The primary concern is that the images utilized for VMMR have frequently been obtained in uncontrolled scenarios. The algorithms for VMMR should have effectively identified the make and model features regardless of picture quality, resolution, lighting conditions, or camera placement relative to the plate's plane. In the realm of VMMR, the second obstacle has entailed the diversity in model configurations across various manufacturers. These configurations have not only varied from one nation to another but have also exhibited distinctions within a single country, where one has come across divergent designs despite sharing the same make and model creation date. Such issues have been discussed in recent surveys by Maity et al. [3] and Gayen et al [6].

1.2 Skin Cancer Classification

Image classification is crucial in diagnosing and treating various diseases. Specifically, skin cancer classification is a challenging task that demands accurate identification of malignant and benign lesions. Deep learning techniques have shown great potential in automating the classification of skin lesions, providing dermatologists with valuable assistance in decision-making processes. Skin cancer is defined

as abnormal skin cell proliferation, which often appears on sun-exposed skin. It primarily affects the upper layer of the skin, comprising squamous cells, basal cells, and melanocytes. While non-melanocyte skin cancers such as squamous cell and basal cell types are often treatable and less likely to spread, malignant melanoma poses a greater threat if not detected early. Skin cancer, the most prevalent form of cancer globally, typically originates in the epidermis, the skin's outermost layer. Melanoma, squamous cell carcinoma, and basal cell carcinoma are the three main forms of skin cancer. Areas of skin exposed to the sun, such as the scalp, face, lips, ears, neck, chest, arms, and hands, as well as the legs, are the primary areas, where skin cancer occurs. It can also develop in places of your body that are seldom exposed to daylight, such as your palms, the space under your fingernails, or toenails. All skin tones, particularly those with darker complexions, are susceptible to skin cancer. People with dark skin tones are more likely to develop melanoma on parts of their bodies, like the palms of their hands and the soles of their feet, that are not often exposed to the sun. Melanoma, although less common than other types of skin cancer, poses a grave threat if not detected in its early stages. It can arise from an already-existing mole or grow into a new, atypical mole. It is frequently distinguished from ordinary moles by its bigger size, uneven borders, and color variety.

However, skin cancer diagnosis primarily involves three methods: self-examination, dermoscopy, and skin biopsy. Accurate self-examination of skin cancer is highly challenging due to the striking resemblance between malignant and benign skin conditions. Dermoscopy, a widely accepted imaging technique among dermatologists, involves magnifying the surface of a skin lesion for enhanced visibility during examination. However, its practice is limited to trained physicians, relying heavily on their visual skills and experience [7]. The scarcity of well-trained physicians, attributed to the high costs and efforts involved in training, renders dermoscopy impractical for large-scale skin cancer detection. Skin biopsy, another method used by doctors, entails the removal of a sample from a suspected skin lesion for examination to determine its cancerous nature [8]. Despite being a reliable diagnostic tool, this procedure is characterized by pain and time consumption. Tackling the intricacies of skin cancer classification requires navigating through numerous inherent challenges, such as the wide variability in appearance among skin lesions, imbalanced datasets, and the similarity and variability within and between different types of dermoscopic images.

In a recent comprehensive survey conducted by [9], the application of machine learning and deep learning methodologies in the domain of skin cancer classification was extensively explored. Prior attempts to classify skin cancer have predominantly relied on machine learning techniques, a fact underscored by the

methodologies outlined in Hussain et al.'s study. However, the conventional machine learning approach necessitates manual feature extraction from image data. This requirement often proves to be a bottleneck, particularly when confronted with complex pattern recognition tasks. In contrast, deep learning methods offer a more effective solution by enabling automatic feature extraction, thereby enhancing the accuracy and robustness of skin cancer classification systems. Consequently, computer-aided diagnosis (CAD) systems have emerged as an effective technique for melanoma diagnosis. Such a system can offer a user-friendly environment, which is particularly beneficial for less experienced dermatologists. The automated CAD system can aid dermatologists by reducing time, costs, and effort [10]. Additionally, it can serve as a valuable second opinion for doctors when dealing with complex melanoma cases. Over the years, deep learning approaches have demonstrated remarkable success in this domain by leveraging complex neural network architectures to extract meaningful features from dermatological images.

In the quest for accurate classification of skin cancer, deep learning models, despite their sophistication, sometimes struggle to capture all relevant features effectively. The existing methods face challenges in extracting higher-level semantic features critical for distinguishing between different types of skin lesions. Additionally, single-scale networks often fail to leverage the full spectrum of complementary information available across various scales, resulting in suboptimal performance. To overcome this, researchers have turned to ensemble approaches, recognizing their potential to enhance classification performance. Works such as [11], [12], [13], [14] and [15] have highlighted the significance of ensembles approaches not only in this field but also in related medical domains. However, the effectiveness of these models often comes at the cost of increased computational burden, especially when employing ensemble methods that rely on multiple baseline models for the training process. This limitation hampers their practical deployment, especially in resource-constrained settings. Given these challenges, there is ample room for improvement in the current ensemble methodologies for skin cancer classification.

Overall, this thesis explores the application of different deep learning approaches to tackle image classification tasks in the VMMR and skin cancer classification domains, to improve classification accuracy, efficiency, and applicability in real-world scenarios.

1.3 Scope of the Thesis

This thesis makes significant contributions to the both VMMR and skin cancer classification research domains.

- **Simple Sequential Attention Network (SimSANet):**
  - Introduction of SimSANet for VMMR, a novel multi-kernel-based sequential attention-based model that effectively extracts discriminative information by integrating global and local features.
  - Demonstrated advantages in speed, effectiveness, and efficiency with significantly fewer parameters compared to existing models.
  - Conducted extensive experiments on multiple public benchmark datasets.
- **Snapshot Ensemble-based Feature Fusion Network (SEFFNet):**
  - Presented SEFFNet for skin cancer classification, a novel snapshot ensemble-based approach for feature extraction to enhance the feature representation by combining lower-level and higher-level features, promoting model generalization.
  - Streamlined computational burden by eliminating traditional ensemble-based approach, along with traditional data pre-processing and augmentation techniques, thus accelerating training and enhancing efficiency.
  - Proposed a hybrid loss function with dynamic hyperparameter adjustment to overcome the class imbalance and overfitting problems.
  - Conducted extensive experiments on multiple public benchmark datasets.

these scopes propel the domains of VMMR and skin cancer classification research forward by introducing efficient models, innovative feature extraction methods, and tackling critical challenges like overfitting and class imbalance.

1.4 Organization of the Thesis

The remainder of this thesis is organized as follows:

- In chapter 1, a comprehensive overview is provided. This chapter examines current research efforts, detailing various methodologies, challenges, and advancements in the fields of VMMR and skin cancer classification.
- In chapter 2, a comprehensive literature survey is presented. This survey delves into existing research endeavors, exploring various methodologies, challenges, and breakthroughs in the field of VMMR and skin cancer classification.

- The chapter 3 illuminates a novel methodology alongside its results and discussions. This chapter also delineates the effectiveness of the proposed method compared to other comparable approaches in the VMMR domain.

- Subsequently, chapter 4 elucidates a novel methodology along with its corresponding results and discussion. Additionally, this chapter evaluates the efficacy of the proposed method in comparison to other relevant approaches within the Skin cancer classification domain.

- Finally, chapter 5 encapsulates essential conclusions drawn from the study and offers insights into future research directions. This concluding chapter synthesizes key findings, highlights overarching themes, and underscores the significance of the study's contributions to the field of VMMR and skin cancer classification.

Through this research, the aim is to contribute to the advancement of VMMR and skin cancer classification domains, ultimately facilitating more accurate and efficient solutions in both domains.

CHAPTER 2

Literature Survey

The literature survey serves as a comprehensive exploration of existing research in the fields of VMMR and skin cancer classification-based applications. This chapter delves into a detailed examination of relevant methodologies, techniques, and models employed in both domains. By synthesizing a wide array of literature, the chapter aims to provide a contextual backdrop for the subsequent discussion and development of novel contributions. Special attention is paid to recent advancements and state-of-the-art approaches, laying the groundwork for the proposed methodologies introduced in subsequent chapters. Through critical analysis and comparison with existing techniques, this chapter sets the stage for the evaluation and validation of the proposed methodologies in the following sections.

2.1 Vehicle Make and Model Recognition

In this section, some relevant VMMR methods found in the literature are discussed. Li et al. [16] have presented a unique framework for fine-grained recognition with a hierarchical label tree that uses a two-branch network, (a coarse branch and a fine branch), or semantic bilinear pooling. In this study, they have created a generalized cross-entropy loss to increase the distance between samples of various coarse classes and to fully utilize the semantic priors to learn the semantic information of the hierarchical levels. For this study, they have used two available datasets, namely CompCars, and Stanford Cars to test the correctness of their algorithm, and have achieved accuracy scores of 97.8% and 94.3% respectively. The attention convolution model has been greatly aided by the channel attention architecture in SENet [17]. However, according to Yang et al.'s [17] research, SENet concentrates more on certain feature channels than on the objects within the channels. They have built a channel expansion convolution network (CENet) by first performing channel sorting and group convolution on the feature map based on

the channel attention-based convolution network. FGVC-Aircraft, Stanford Cars, CIFAR-10, and CIFAR-100 are the four datasets on which they have performed experimentation. The achieved accuracy scores are 95.35%, 87.91%, 79.38%, and 89.02%, respectively.

Ajitha et al. [18] have developed a framework, that uses a shallow convolutional neural network (CNN) architecture to avoid overfitting and guarantee that the right features are learned. In addition, by using the image data generation model in Keras, an augmentation technique is applied to create synthetic images because of class imbalance. The created images are first given to a shallow CNN model to extract features, and they are subsequently classified using Softmax activation. The results of the trials show that the shallow CNN model can function effectively in practical scenarios. Stanford Cars dataset is used to evaluate the model, and they can attain 92.36% accuracy. For the fine-grained vehicle recognition challenge, Lu et al. [19] have provided an efficient way to improve and aggregate discriminative part-level features, therefore bolstering the discriminative performance of high-level deep features. In this context, they have suggested to use of a novel feature extraction method to solve this problem from two perspectives: aggregating part-level discriminative features and creating additional feature descriptors of part areas through deep network learning. With the use of a feature aggregation module and a part-level feature enhancement module, this approach aims to increase backbone networks' ability to produce finer-grained part features and explore the intrinsic features of those networks. To verify the accuracy of the model they have considered the Stanford Cars dataset, and achieved an accuracy of 94.5%.

A multi-attention CNN (MA-CNN) has been proposed by Zheng et al. [20] as a unique part-learning method, where part generation and feature learning can reinforce each other. Convolution, channel grouping, and part classification subnetworks make up the MA-CNN. Through the process of classifying an image by individual parts, the component classification network facilitates the learning of more discriminative fine-grained features. MA-CNN can be trained end-to-end and does not require bounding box/part annotation. The authors have used two publicly available datasets, FGVC-Aircraft, and Stanford-Cars, to obtain the model's performance. Furthermore, the attention transformer module is employed to ensure that the network captures discriminative features. Various tests conducted on the Stanford Cars datasets show that this approach performs well. To solve the issue of fine-grained object classification, Zhao et al. [21] have presented a diversified visual attention network (DVAN), which significantly reduces the dependency in contrast to attention-less models, on strongly-supervised information for learning to localize discriminative regions. A recurrent long short-term memory (LSTM)

unit has been used to teach the discernment and awareness of eye-catching canvases. The suggested DVAN is capable of attending the sample's granularity from coarse to fine, as well as a dynamic internal representation is progressively constructed for categorization integrating data from many scales. The accuracy of the model is 87.1% on a publicly available dataset Stanford cars. A novel progressive-attention CNN (PA-CNN) has been proposed by Zheng et al. [22] to progressively localize portions at several scales. A part proposal network (PPN) creates several local attention maps, and a part rectification network (PRN) extracts part-specific information from each proposal to provide the PPN with more precise part positions. Because of their interaction, the PPN and PRN can be tuned in a way that strengthens one another, improving the pinpointing of fine-grained components. Using the FGVC-Aircraft and Stanford Cars datasets, the authors have attained 91% and 93.3% accuracies, respectively.

2.2 Skin Cancer Classification

In this section, some relevant skin cancer-based methods found in the literature are discussed. Skin cancer classification is very tough, and despite significant progress, there are still a lot of challenges that AI researchers have yet to overcome. The individual skin lesion images are generally large, with only a limited infected portion being relevant. Also, there is a lack of reliable annotated images, and many images are almost indistinguishable from one another, with low inter-class variation. Prior to 2016, mainly machine learning-based techniques were being widely used, yielding limited success in this field. However, with the advent of deep neural networks, there has been a rapid improvement in skin cancer classification. A pre-trained deep convolutional neural network (CNN) architecture, VGG16, with the last three layers fine-tuned and five convolutional blocks, was proposed by [23]. Mendes and da Silva et al. proposed a deep CNN architecture based on pre-trained ResNet152 to classify different types of skin lesions [24]. [25] proposed a strategy to extract features from various well-established and pre-trained deep CNNs like AlexNet, ResNet18, and VGG16 for skin lesion classification. This was followed by a multi-class support vector machine (SVM) classifier that was trained on these features. Finally, the classifier results were combined to perform the classification. [26] have emphasized the need for data augmentation to tackle the dearth of annotated skin lesions. Some research work on the ISIC 2017 dataset, where they have tried to resolve some skin cancer-related issues, has been surveyed.

Deep Attention Branch Networks (DABN), along with a novel Entropy-guided Loss Weighting (ELW) technique to combat class imbalance present in skin lesion datasets, was proposed by [27]. [28] proposed Dermo-DOCTOR, where

they employed end-to-end dual encoders for segmentation before performing the classification task. They also introduced the DermoExpert model which combines pre-processing techniques with a hybrid CNN through segmentation, transfer learning, and data augmentation. [29] proposed an attention residual learning CNN (ARL-CNN) model for the classification of skin lesions. This model comprised multiple ARL blocks, a global average pooling layer, and a classification layer. Each ARL block employed both residual learning and a unique attention learning mechanism to improve its capacity for capturing discriminative representations. The attention learning mechanism, rather than introducing extra learnable layers, aimed to leverage the inherent self-attention ability of deep CNNs. Specifically, it utilized feature maps learned by a higher layer to generate the attention map for a lower layer. [30] introduced ARDT-DenseNet, a densely connected convolutional network tailored for skin lesion classification. This model featured dense blocks, transition blocks, and attention and residual modules within the ARDT block. Compared to a residual network with a similar number of convolutional layers, ARDT-DenseNet halved the parameter size while incorporating attention mechanisms and residual learning after each dense and transition block. This enhancement provided additional functionality without increasing the number of parameters.

In conclusion, the literature survey conducted in this thesis provides a comprehensive overview of the existing research landscape in the fields of VMMR and skin cancer classification. Through an in-depth analysis of various methodologies, techniques, and models proposed by previous studies, key insights have been gleaned to inform the development and implementation of novel approaches in this research. Moving forward, the findings from this literature survey will serve as a solid foundation for the development of innovative solutions and the advancement of knowledge in VMMR and skin cancer classification

CHAPTER 3

SimSANet: A Method for VMMR Classification

In this chapter, the architecture of the proposed SimSANet model is thoroughly examined, outlining its design, components, and the impact of sequential attention mechanisms on performance enhancement. An in-depth analysis of the model's efficiency is provided, comparing it with other leading approaches in the field. Additionally, extensive studies conducted on various state-of-the-art public datasets are presented, highlighting the model's strengths and potential limitations.

3.1 Proposed Architecture

A deep learning-based model, called Simple Sequential Attention Network (SimSANet), has been proposed for automatic VMMR. Figure 3.1 illustrates the general layout of the suggested framework. The primary objective of this model is to establish a mapping function capable of predicting the accurate label associated with a given input image (here, a vehicle). There are three key stages in our suggested model. In the first stage, fundamental feature extraction uses the DenseNet201-based CNN architecture. A multi-headed Sequential Attention (SA) module is introduced in the second stage, whereas the third stage comprises a classification module. Details of these three stages are given in subsequent subsections.

3.1.1 Fundamental Feature Extraction Module

In the first stage, processed images are used by the model. Next, DenseNet201 is employed as the baseline model to extract general features. An average pooling

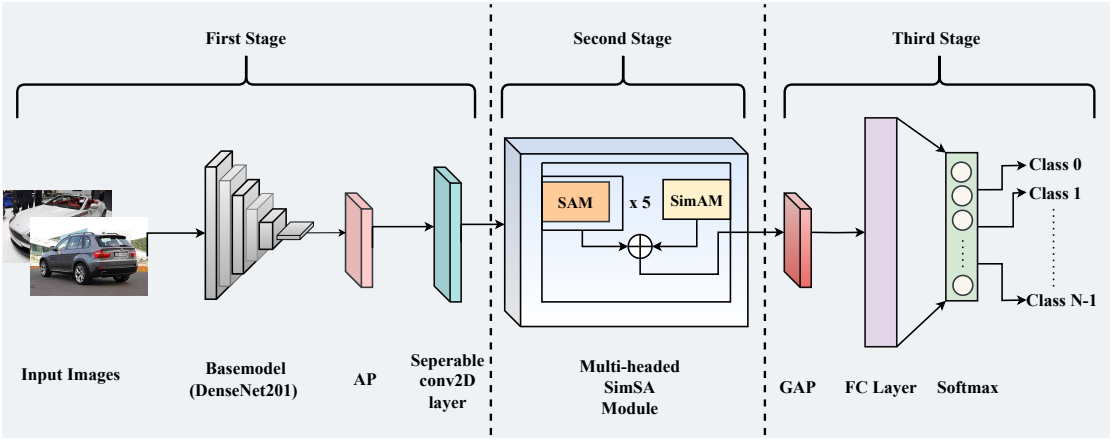


Figure 3.1: Overall pipeline of the proposed SimSANet model used for VMMR.

(AP) layer is then added to enhance the feature extraction techniques. By incorporating a separable convolutional layer with 512 filters and a specific kernel size of 5x5, the dimensions are further reduced.

3.1.2 Multi-headed SA Module

In the second stage, the proposed multi-headed SA module is employed, designed to implement a multi-head attention approach within a neural network architecture. This module is structured to make the extraction of complex features by dividing the input tensor into multiple heads, each focusing on distinct aspects of the information. Initialization entails setting parameters such as the number of heads and channels while creating individual attention heads and employing attention blocks separately to capture relevant features. The depthwise convolutional layers with the construction of the attention block involve depthwise convolutional layers with sequential kernel sizes used in the development of the attention block to capture features ranging from local to global. Here, each head processes the input independently, producing individual outputs. These outputs are then concatenated and then the concatenated output is fused with attention maps of SimAM [31]. A global average pooling layer (GAP) receives the previous layer's output. Lastly, a fully connected layer with a softmax activation function is utilized as the final layer of this proposed model. This layer turns the gathered attributes into predictions for a range of output classes.

3.1.2.1 SimAM: A Simple Parameter-free Attention Module

This SimAM [31] is a simple parameter-free attention module, which is more effective than the traditional attention mechanisms including Convolutional Block

Attention Module (CBAM) [32], the combination of spatial attention and channel attention mechanisms, alongside the SENet [33], play pivotal roles in capturing most discriminative features from the input data. Whereas channel attention highlights the strengths of certain channels, spatial attention focuses on the region's importance. On the other hand, SENet uses a squeeze-and-excitation mechanism to give inter-channel relationships priority. These attention modules improve a network's representation capabilities, but they also bring new variables and parameters. Conversely, SimAM is shown to be a lightweight, effective, and well-designed three-dimensional attention module. By taking into account spatial and channel correlations through feature mapping, it computes three-dimensional attention weights. It is noteworthy that this is achieved without adding many parameters to the original network.

3.1.2.2 SA Module

This is our proposed module, as shown in Figure 3.2, which stands for sequential attention module. In this stage, the first Generic Sequential Multi-kernel Depthwise Convolutional (SMDC) block receives the input from the preceding layer. Here, each convolution process is factorized into two parts by these layers: a depth-wise convolution and a point-wise convolution. As a result, these layers perform better than traditional convolution layers in terms of effectiveness and computational cost. This block generates 4D tensors feature maps, which are then passed to two parallel branches: a GAP branch and a Global Max Pooling (GMP) branch. The GAP branch averages the values of each feature map, and the GMP branch takes the maximum values of each feature map. The output of both the GAP and GMP branches are then concatenated together and reshaped into 4D tensors again. The next generic SMDC block receives these 4D tensors after that. The output of this block is now concatenated with the output of the first generic SMDC block.

At this point, the first row-wise SMDC, and first column-wise SMDC blocks receive the output of the first generic SMDC block in parallel. Now, the output of the first row-wise and first column-wise SMDC blocks is parallelly concatenated with the previous resultant output feature maps. Next, the two resulting feature maps are multiplied together to generate the output features. This multiplied result is now sent to the third generic SMDC block as the last set of feature maps from this initial stage. Next, the fourth Generic SMDC receives the output from the third generic SMDC block. The outputs of this generic SMDC block and the output of the first generic SMDC block are concatenated. At this point, The second row-wise SMDC and second column-wise SMDC blocks receive the first

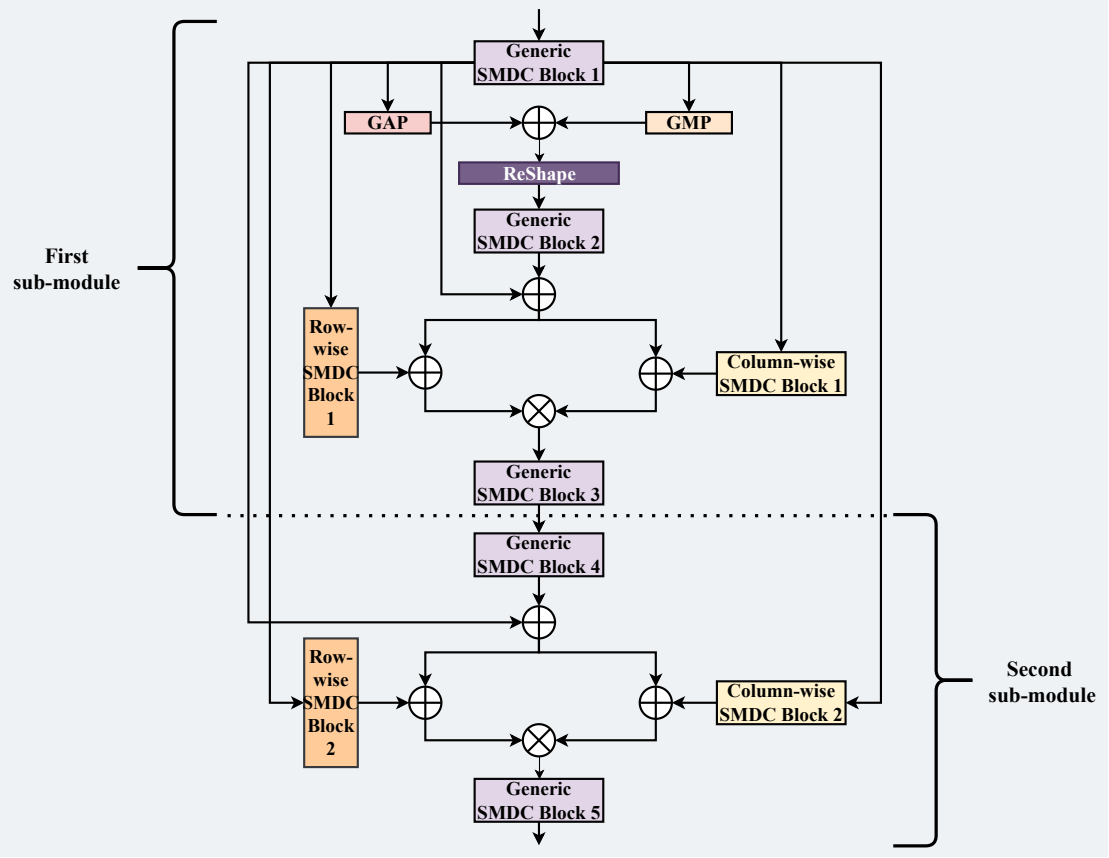


Figure 3.2: Architectural overview of the proposed SA module. The basis of this generated module is the information extraction at various levels. Each block's layer dimension and associated parameters have been omitted from this module for better visualization.

generic SMDC block's output in parallel. Now, the resulting output of these blocks is concatenated with the previous output feature maps. Next, multiplying the two resulting feature maps together yields the final output. To prevent overfitting dropouts, regularizers, and initializers are also applied to every layer of depth-wise convolution layers. This is the extracted final output of the SimSANet module.

- **Generic SMDC block:** As shown in Figure 3.3, convolution layers with fixed kernel sizes are a common aspect of traditional CNN architectures, which restrict their ability to collect features at various scales efficiently. In response, an SMDC block is introduced to capture features. The SMDC block is constructed with a series of convolution layers with increasing kernel sizes (1x1, 2x2, 3x3, 4x4, and 5x5) to progressively collect information across the different receptive fields. This methodology comprises a sequential hierarchical feature extraction procedure that aids in collecting fine-grained information, ranging from local to global, from the input data for the network.

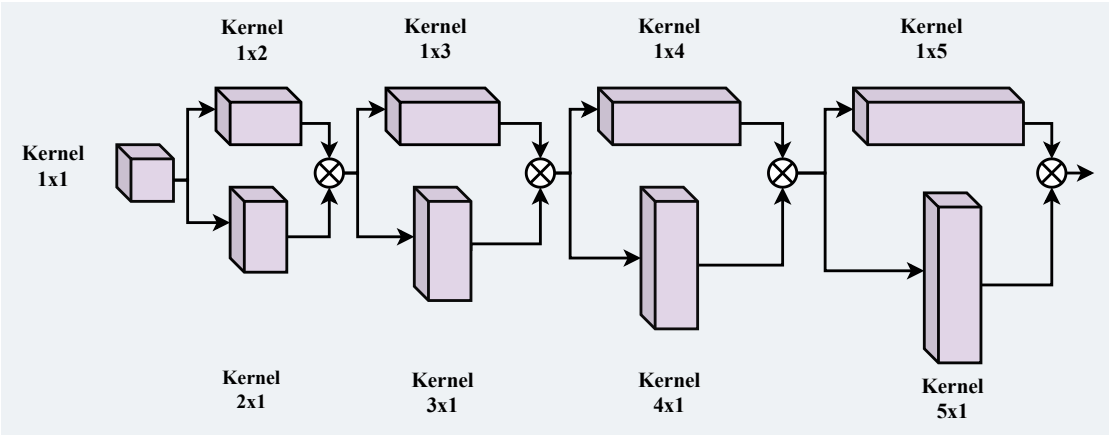


Figure 3.3: Architecture of the generic SMDC block used in the present work.

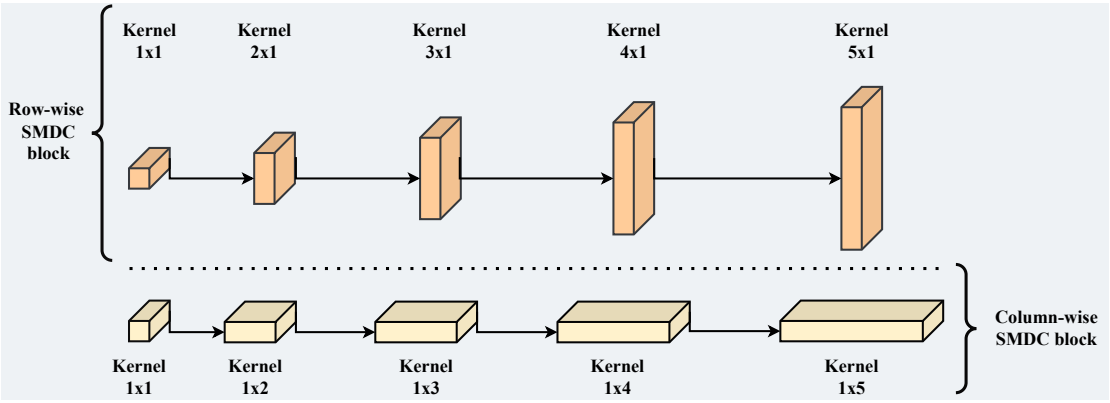


Figure 3.4: Diagram showing the row-wise and column-wise SMDC blocks.

Initially, a depthwise convolutional layer, utilizing a 1x1 kernel, processes the output from the previous layer. This layer acts as the primary feature extractor, adept at capturing local contextual information while preserving the input's spatial dimensions. Subsequently, the output undergoes two concurrent depthwise convolutional layers: a 2x1 layer operating row-wise and a 1x2 layer operating column-wise at the same time. The final result is then obtained by multiplying the outputs of these two layers. This process expands the receptive fields by allowing the network to gather more row-wise and column-wise contextual information and higher-level features.

Continuing this approach, the output of the previous sub-block is further treated simultaneously with 3x1 and 1x3 kernel depthwise convolutional layers at the same time. These are then multiplied to produce the final incorporation another step to capture the mid-level features, which works on a larger receptive field. Here, the previous sub-block output enters simultaneous 4x1 and 1x4 kernel depthwise convolutional layers at the same

time to capture more intricate patterns from the input data. Then multiplication of these outputs produces the final output of this sub-block. Moving forward, the feature map derived from the prior sub-block is processed concurrently through 5x1 and 1x5 kernel depthwise convolutional layers. Their multiplication yields the desired output, concluding the sequence. This layer completes the sequence by enabling the network to extract higher-level and global context features by incorporating a wider spatial scope. By integrating increasingly larger kernel sizes sequentially, the SMDC block can extract hierarchical features with varying receptive fields. Moreover, using different kernel sizes within a single block allows them to capture different patterns within the same input region. This approach enriches the block's capability to learn effective features within the input data.

- **Row-wise and Column-wise SMDC Blocks:** For the row-wise direction, the depthwise convolutional layers are designed with increasing row-wise kernel sizes (1x1, 2x1, 3x1, 4x1, 5x1). This implies that these convolutional layers are scanning the input tensors across rows, starting with a single-row kernel and gradually expanding to cover a larger section of each row as shown in Figure 3.4. Similarly, for the column-wise direction, the depthwise convolutional layers are designed with increasing column-wise kernel sizes (1x1, 1x2, 1x3, 1x4, 1x5).

  Here, these convolutional layers process the input tensors column by column, starting with a single-column kernel and gradually expanding to cover a larger section of each column as shown in Figure 3.4. By employing this strategy, the model can gather systematic information from the various receptive fields. By varying kernel sizes, the model can capture diverse patterns and relationships present within the rows and columns of the input tensors.

3.1.3 Classification Module

As shown in Figure 3.1, this module begins by employing a GAP layer to extract further relevant features. Lastly, a fully connected (FC) layer with n units (n is the number of classes) and a softmax activation function are utilized as the final layer of this proposed model. Next, these features are converted into predictions for a variety of output classes by this layer.

3.2 Materials and Implementation Details

3.2.1 Datasets

The experiments have been conducted on four benchmark VMMR datasets: Stanford Cars [34], CompCars [35], and CompCarsSV [35], along with a small dataset proposed by [36]. These datasets consist of images taken by security cameras or images gathered by web crawlers. These collections showcase variations in viewpoints, and encompass diverse classes, and complex backgrounds, resulting in varying levels of complexity. As a result, the classification outcomes derived from these datasets serve as valuable metrics to evaluate the module's performance precisely. A few sample images from these datasets are shown in Figure 3.5 for a visual depiction. Table 3.1 complies with the essential details about the datasets considered here for experiments.

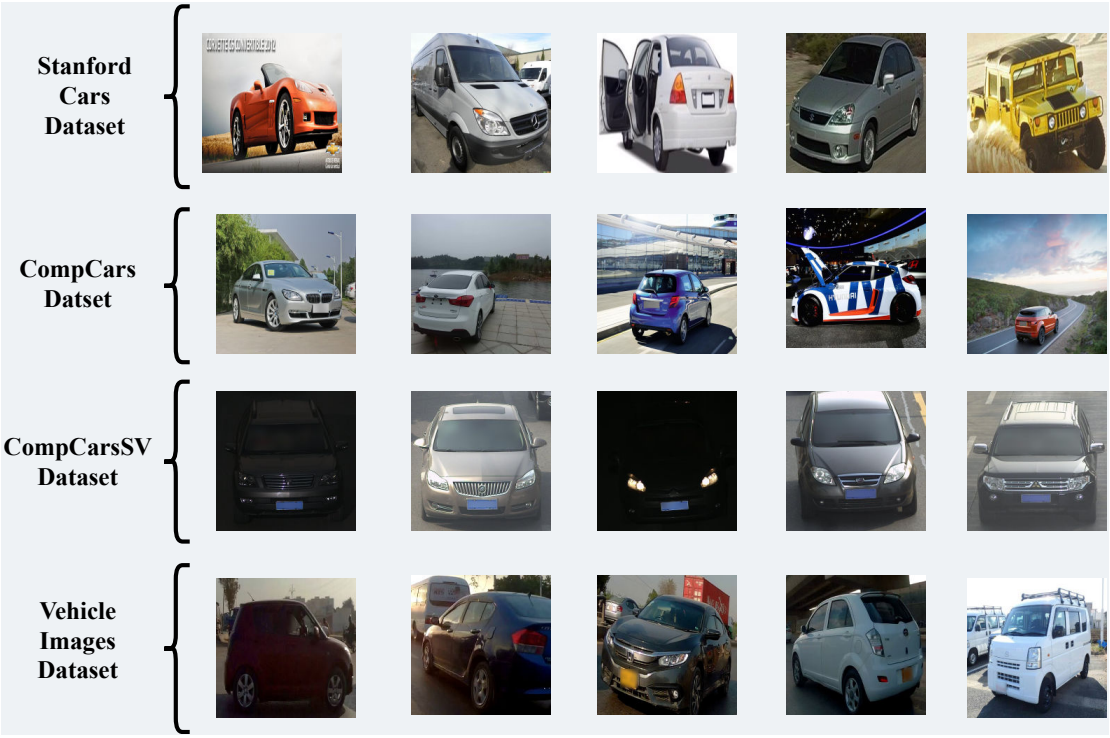


Figure 3.5: Few sample images from various VMMR datasets

3.2.2 Data Pre-processing and Base Model Selection

The images of Stanford Cars, CompCars, and CompCarsSV datasets include training and testing images with their associated bounding box locations available inside the annotation file. The image size taken as input is 400 x 400. The images are normalized to zero mean and unit variance. Additionally, data augmentation

Table 3.1: Statistics of the VMMR datasets used in the present experimentation.

Dataset	Training images	Testing images	Split ratio	Classes	Viewpoint
Stanford Cars	8144	8044	50:50	196	Multi
CompCars	36334	15749	70:30	431	Multi
CompCarsSV	31010	13471	70:30	281	Single
Vehicle Images	3096	751	80:20	48	Multi

methods have been included in the pipeline for training. These include several transformations, including horizontal flipping, fill mode parameters, shear range, zoom range, rotation, rescaling, and width and height shifts.

Our evaluation process comprises a thorough analysis of many baseline models. Finding and identifying the baseline model that best fits our experiment setup is the main goal. The findings of these assessments are summarized in Table 3.2. This comprehensive depiction provides a fine-grained analysis of each baseline model. Based on the outcomes presented in Table 3.2, our study indicates that the DenseNet201 model, pre-trained on the ImageNet dataset encompassing 1.2M images categorized into 1000 classes, stands out as the best-performing baseline model among all the baseline models that are considered here. Combining our suggested model with DenseNet201 results in significant improvements in performance measures. Moreover, it boasts 20.2M parameters. Hence, the DenseNet201 model has been chosen as the ultimate baseline model across all datasets.

Table 3.2: Comparison of classification results and number of parameters on the Stanford Cars dataset using various baseline models

Work	Baseline	Parameters(M)	Accuracy(%)
[37], 2016	InceptionV3	23.9	68.41
[37], 2016	InceptionResNetV2	55.9	68.92
[38], 2016	ResNet50V2	25.6	70.05
[39], 2017	DenseNet201	20.2	75.87

3.2.3 Experimental Setup

The proposed model is implemented with Tensorflow and runs on the Kaggle platform with 1 x Tesla P100 GPU, having 3586 CUDA cores and 16 GB GDDR6 VRAM. A pre-trained DenseNet201 baseline model has been used on the ImageNet dataset [40] and then combined with the proposed Multi-headed SimSAMNet. The model is trained using an initial learning rate set at 0.0001 on the various VMMR datasets. The learning rate is then lowered by 0.1 if the testing accuracy fails to improve for 4 epochs consecutively, and an early stopping condition is imposed if

the testing accuracy remains stagnant for 8 consecutive epochs. The performance of the model is optimized by carefully selecting certain hyperparameters. Five heads assist in collecting various parts of the data, and a batch size of 16 permits efficient processing of the incoming data. Furthermore, a l2 regularizer of 0.001 aids in managing model complexity and avoiding overfitting, while a Simam value of 0.0001 guarantees smooth convergence throughout training. Finally, a 0.05 dropout rate improves generalization.

3.2.4 Evaluation Metrics

The experiments conducted in this study are assessed through various standard metrics that include Accuracy (Acc), Precision (Pre), Recall (Rec), and F1-score (F1).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{3.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.3}$$

$$F1\text{-}score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \tag{3.4}$$

Here, *TP*, *TN*, *FP* and *FN* signify the true positives, true negatives, false positives, and false negatives, respectively.

3.3 Results and Discussion

3.3.1 Performance of SimSANet on various VMMR datasets

The SimSANet architecture features a SimSA module. The major objectives are to investigate how different baseline models affect the SimSA module and how SimSA outperforms the other attention-based methods. Detailed performance metrics for SimSANet on four datasets are illustrated in this section. The SimSANet obtains the highest accuracy of 94.47% when using DenseNet201 as the baseline model on the Stanford Cars dataset. Similarly, on the Compcars, CompCarsSv, and Vehicle Images datasets, 98.34%, 99.20%, and 97.20% accuracy are obtained using DenseNet201 as the baseline model. Analyzing the performance of the proposed model with DenseNet201 as the baseline emerges as

the top-performing variant across all datasets. Based on the evaluation indices, the combination of DenseNet201+SimSANet stands out as an excellent VMMR classification network Table 3.2.

3.3.2 Quantitative Analysis on the Stanford Cars Dataset

Table 3.3 summarizes the classification results on the Stanford Cars dataset. Without using bounding box information during the testing phase, our suggested SimSANet achieves a testing accuracy of 94.47%. The outcome is similar to the GRAD-CAM guided CBAM [41] technique, which utilizes the most discriminative parts to achieve an accuracy of 94.41%. However, it takes nearly 100 epochs which is a lot longer than ours, for this method to properly train. Comparably, the ResNet101-MRDMN [42] approach employs a discriminative mix-up strategy to attain 94.4% accuracy. However, compared to our approach, this one has 51.2M parameters, which is far more than ours.

On the other hand, the hierarchical feature extraction method used by the ResNet-ACNet [43] method achieves an accuracy of 94.60%, which turns out a marginally better score (0.13%) better than ours. Notably, to improve the accuracy, the ACNet undergoes a rigorous training procedure which, in turn, depends significantly on additional parameters. However, the model requires approximately 240 epochs for training the model. This approach is noticeably more complex than ours. Similarly, the ResNet50-DCL [44] based method attains 94.5% accuracy, which is again 0.03% better than ours. However, in comparison to our approach, this method requires nearly 180 epochs to fully train. Additionally, some other approaches yield results that are lower than ours but still effective. The Semantic Enhanced Feature (SEF) [45] approach with ResNet50 as baseline model obtains 94.2% accuracy by using semantic grouping and feature enhancement techniques, which is computationally cheaper as well as effective. Both the methods as reported in [46] and [47], achieve similar accuracy of 94.10% using multi-scale feature fusion and deep layer feature aggregation strategies, respectively. Interestingly, the authors in [48] achieve an accuracy of 94.06% utilizing a unique Center Fill with Shifted Quadrants (CFSQ) data augmentation strategy. With a similar baseline model DenseNet201 as ours, along with five homogeneous models, the authors in [49] yield 93.96% classification accuracy, which is 0.51% less than ours.

Furthermore, a novel Channel Max Pooling (CMP) based method [52] yields a 93.71% accuracy. Additionally, 98.04% accuracy is achieved using Dual CE loss [16]. Interestingly, despite using comparable baseline models, the approaches proposed in [52] and [16] achieve 0.76% and 1.43% less accuracy than ours. Similarly, Recurrent Attention Unite (RAU) based approaches are used by ResNet50-LRAU

Table 3.3: Comparison of the classification results on the Stanford Cars dataset.

Work	**Method**	**Parameters(M)**	**Epochs**	**Accuracy(%)**
[50], 2017	ResNet101 + SWP	-	90	93.10
[51], 2017	VGG16 + Part localization	-	-	93.10
[47], 2018	DLA	33.5	120	94.10
[16], 2019	DenseNet161 + Dual CE loss	-	-	93.04
[52], 2019	DenseNet161 + CMP	30	300	93.71
[53], 2019	ResNet50 + TASN	37.3	90	93.80
[44], 2019	ResNet50 + DCL	Approx 27	180	94.50
[54], 2020	ResNet101 + RAU	-	120	93.81
[45], 2020	ResNeXt50 + SEF	Approx 43.52	50	94.20
[43], 2020	ResNet + ACNet		240	**94.60**
[55], 2021	ResNet50 + LRAU	25.2	120	93.94
[49], 2021	DenseNet201 + Ensemble approach	-	-	93.96
[48], 2021	EfficientNetB3 + CFSQ augmentation	-	150	94.06
[42], 2021	ResNet101 + MRDMN	51.2	50	94.40
[41], 2021	ResNet50 + CBAM	-	100	94.41
[56], 2022	ResNet50 + CA-MSNet	44.2	-	93.52
[46], 2022	ResNet101 + Multi-scale feature fusion	-	-	94.10
Ours	**SimSANet**	**21.3**	**25**	94.47

method [55] and ResNet50-RAU method [54] to obtain 93.94% and 93.81% accuracy, respectively. Although using the Trilinear Attention Sampling Network (TASN) module, the work done in [53] reports an accuracy of 93.80%, which is %0.67 less than ours and it takes 90 epochs to fully train the model. With the help of a cross-channel attention mechanism (CAM) with the multi-scale sparse network, the CA-MSNet module [56] attains a 93.52% accuracy rate. Similar to our suggested strategy, this module aids the system in both horizontal and vertical directions. However, it achieves 0.95% less accuracy than ours and has 44.2M parameters, which is also higher than ours. Despite the part localization technique, the authors in [51] attain an accuracy of 93.10% using VGG16 as the baseline model. Similarly, the spatially weighted pooling method (SWP) [50] along with the ResNet101 model achieves an accuracy of 93.10%. The quantitative analysis conducted on the Stanford Cars dataset indicates that our proposed SimSANet model efficiently improves the model and make recognition capacity by extracting most discrimination characteristics from lower to higher levels with the advantage of having a lesser training epoch and fewer parameters.

3.3.3 Quantitative Analysis on the CompCars Dataset

Table 3.4 summarizes the classification results on the CompCars dataset. Using our proposed method, An accuracy of 98.34% is achieved, which is marginally (0.03%) better than the recently proposed VGG16-LRAU [55] based method. Notably, The VGG16-LRAU method undergoes a rigorous training procedure that requires 120 epochs for training to reach 98.31% accuracy. Using ResNet50 as the baseline model, a part-level feature aggregation approach [57] achieves an accuracy of 98.10%. By using the feature aggregation method on the same baseline model, the authors in [58] reach an accuracy of 98%.

On the other hand, the ResNet101-RAU-based method [54] achieves 97.81%, which is lower than the previous RAU-based approach. Under a similar experiment setup, the DenseNet161-CMP [52] based approach obtains 97.89% accuracy, which is superior to the ResNet101-SWP [50] based approach which attains 97.60% accuracy. By utilizing both global and local feature extraction approaches to localize deep crucial portions, the selective multi-convolutional region (SMCR) feature extraction based on ID-CNN [59] achieves 96.24% classification accuracy.

Table 3.4: Comparison of the classification results on the CompCars dataset.

Work	**Method**	**Parameters(M)**	**Epochs**	**Accuracy(%)**
[50], 2017	ResNet101 + SWP	-	90	97.60
[59], 2018	ID-CNN + SMCR feature extraction	-	-	96.24
[52], 2019	DenseNet161 + CMP	40.6	300	97.89
[54], 2020	ResNet101 + RAU	-	120	97.81
[55], 2021	VGG16 + LRAU	**19.3**	120	98.31
[58], 2023	ResNet50 + Feature aggregation	33.4	90	98.00
[57], 2023	ResNet50 + Part-level feature aggregation	-	-	98.10
Ours	**SimSANet**	22.95	**50**	**98.34**

3.3.4 Quantitative Analysis on the CompCarsSV Dataset

Table 3.5 summarizes the classification accuracies achieved by our proposed SimSANet model as compared with other standard previous models on the CompCarsSV dataset. On this dataset, our proposed method yields an accuracy of 99.20%, which is 0.19% more accurate than the recently proposed multi-agent system (MAS) based approach with ensembles of DenseNet201, ResNet101 and VGG19 models [60]. Interestingly, that accuracy is achieved in just 25 epochs, demonstrating the effectiveness and quick convergence of the model.

Table 3.5: Comparison of the classification results on the CompCarsSV dataset.

Work	Method	Parameters(M)	Epochs	Accuracy(%)
[61], 2016	CaffeNet + CF-CNN	-	-	98.29
[62], 2018	LW CNN	**1.2**	100	98.71
[63], 2018	FF-CMNet	1.6	300K	98.90
[54], 2020	ResNet101 + RAU	-	120	98.90
[60], 2022	MAS + Ensembles	-	145	99.01
[58], 2023	ResNet50 + Feature aggregation	Approx 32	90	99.00
Ours	**SimSANet**	21.88	**25**	**99.20**

On the other hand, the feature aggregation approach [58] with ResNet50 baseline model obtains approximately 99% accuracy. In addition, our proposed model has 21.88M parameters, which suggests a balanced complexity in light of the better accuracy obtained. The accuracy attained by this approach is 99.01%. A comparable good accuracy of 98.90% is also achieved using the FF-CMnet-based approach. Conversely, a lightweight CNN-based approach [62] achieves an accuracy of 98.71% using only 1.2M parameters. The ResNet50-feature aggregation-based approach outperforms the ResNet101-RAU-based approach and obtains an accuracy of 98.90% using the same baseline setting.

3.3.5 Quantitative Analysis on the Vehicle Images Dataset

Our proposed model SimSANet, achieves 97.20% classification accuracy, which is 0.3% better than that achieved by the state-of-the-art methods, as demonstrated from Table 3.6. Using ResNet50 as the baseline model, the feature aggregation strategy [58] achieves 96.90% accuracy. However, this model takes nearly 50 epochs to train properly.

Table 3.6: Comparison of the classification results on the Vehicle Images.

Work	Method	Parameters(M)	Epochs	Accuracy(%)
[58]	ResNet50 + Feature aggregation	Approx 31	50	96.90
Ours	**SimSANet**	**20.22**	**25**	**97.20**

3.3.6 Analysis of Accuracy-Loss curves on various VMMR datasets

Figure 3.6 presents comprehensive accuracy and loss curves across both training and testing images for all four VMMR datasets. The model shows increasing

accuracy for both training and test datasets as the number of training epochs increases, along with a corresponding decrease in loss. The testing accuracy for the Stanford Cars dataset is 94.47% in Figure 8(a), which shows a decrease in loss as the number of epochs increases. Similarly, Figure 8(b) displays an accuracy of 98.34% on the CompCars dataset, paired with a declining loss. Figures 8(c) and 8(d) demonstrate how the loss drops as the model's accuracy rises to astounding levels of 99.20% and 97.20%, respectively, on the CompCarsSV and Vehicle Images datasets. These observations strongly indicate that the model is effectively learning to adapt and optimize its performance in fitting the training data.

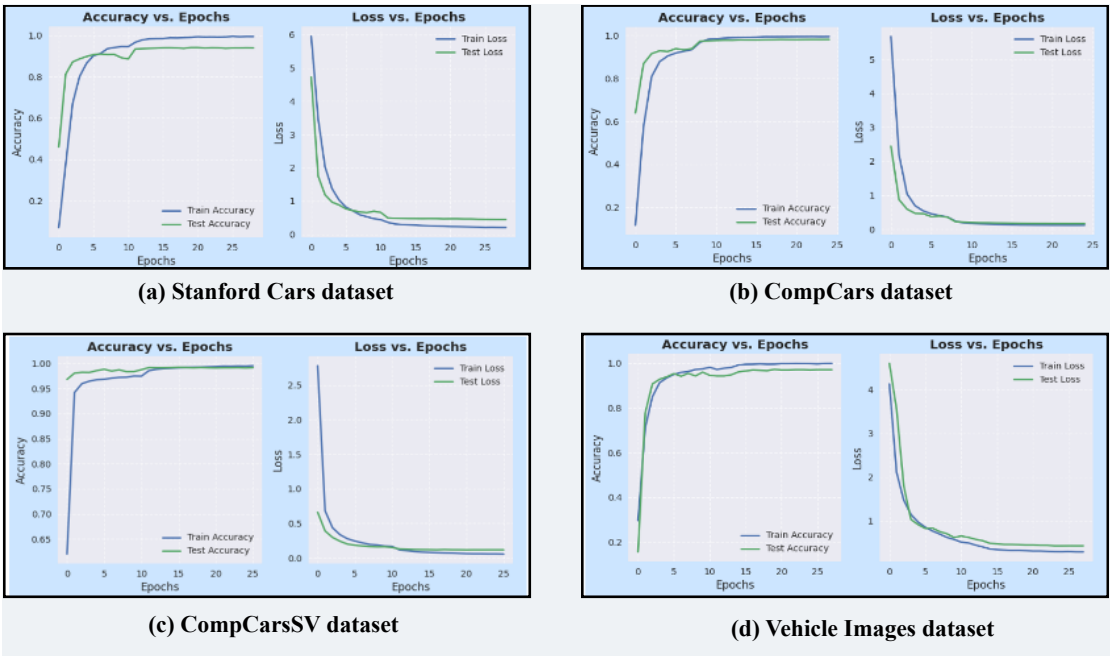


Figure 3.6: Accuracy and loss curves generated by the proposed model on four standard VMMR datasets.

3.3.7 Analysis of Grad-CAM based feature extraction on various VMMR datasets

To qualitatively verify the improvement of the proposed model, Gradient-Weighted Class Activation Mapping (Grad-CAM) [64] was utilized as the visualization approach to visually compare the baseline model and proposed SA module regarding their ability to capture unique vehicle components. Grad-CAM identifies the key areas in an image that have a major impact on the decision-making process of a model. In the final convolutional layer, it computes the gradient of the target class score concerning the feature maps. By precisely identifying the areas in which the model concentrates its efforts, this approach improves interpretability

and offers insightful information about the model's reasoning for making decisions. The feature maps produced by the last layer of the SA module are compared with those from the DenseNet201 baseline model to demonstrate the efficacy of the SA module.

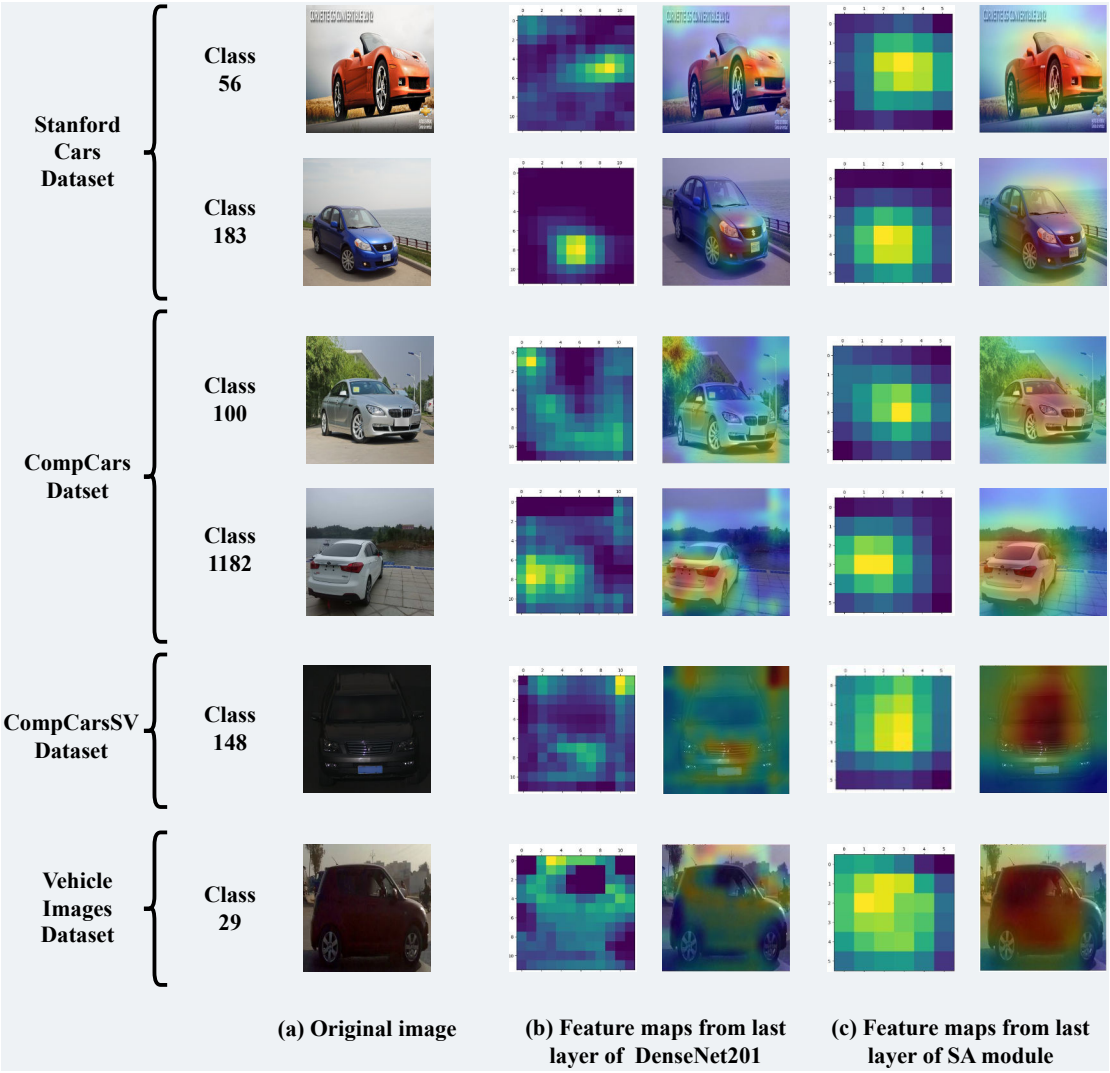


Figure 3.7: Visualization of feature maps on four standard VMMR datasets using Grad-CAM analysis.

The Grad-CAM-visualized top-level feature maps for different VMMR datasets are displayed in Figure 3.7. It is important to note that less discriminative regions are shown in blue, while the class-specific discriminative regions are emphasized in red. When the two extracted feature maps are compared, it becomes clear that the heat maps from the SA module's final layer cover a larger area. On the other hand, the heatmaps that originate from the last layer of the DenseNet201 model are mostly focused on certain components, such as bumpers, headlights, or other unique pieces. Although this concentration on discriminative traits is valuable,

the classification task becomes more complicated due to this focused attention. Conversely, the heatmap of the SA module throws a larger net on individual components found in the entire frontal or back regions. This expansive emphasis helps to extract important data that is necessary for performing VMMR tasks, thus improving the derived features' level of comprehensiveness.

Examining Figure 3.7, it can also observed that class '183' from the Stanford Cars dataset, shows a sharp contrast. As a result, the heatmap generated by the DenseNet201 model predominantly focuses only on the headlights. On the other hand, the heatmap generated by the SA module shows a more comprehensive emphasis by covering the whole frontal part. Furthermore, the SA module produces a heatmap that is nearly ideal for class '148' from the CompCarsSV dataset. However, the DeneseNet201 model is unable to produce flawless feature maps from this image captured during nighttime.

3.3.8 Analysis of Misclassified Images on various VMMR datasets

Despite the overall high classification accuracy, some images fail to capture satisfactory features adequately. This constraint becomes apparent when examining the misclassified images, which highlights the shortcomings in the feature extraction procedure. For additional study, samples of incorrectly identified images that are misclassified are provided in Figure 3.8. An analysis of these false samples confirms that the suggested approach is validated. There are several factors for the misclassifications. Primarily the presence of both intra and inter-class similarities within the datasets. It is difficult for the models to discern between closely related classes as a result of these similarities, which confuses the classification process. Significant issues can arise from the photos with blurriness, poor visibility, and general uncertain image quality. To mitigate the unfavorable impacts of intra and inter-class similarities on the overall classification accuracy, it is still necessary to increase the quality and uniqueness of the samples. These improvements will assist in improving the models' ability to identify finer features and traits, which will help to improve classification accuracy and decrease the incidence of misidentification.

3.3.9 Ablation Study

The SimSANet model architecture design and hyperparameter selection are evaluated through an ablation study on the Stanford Cars dataset. The impact

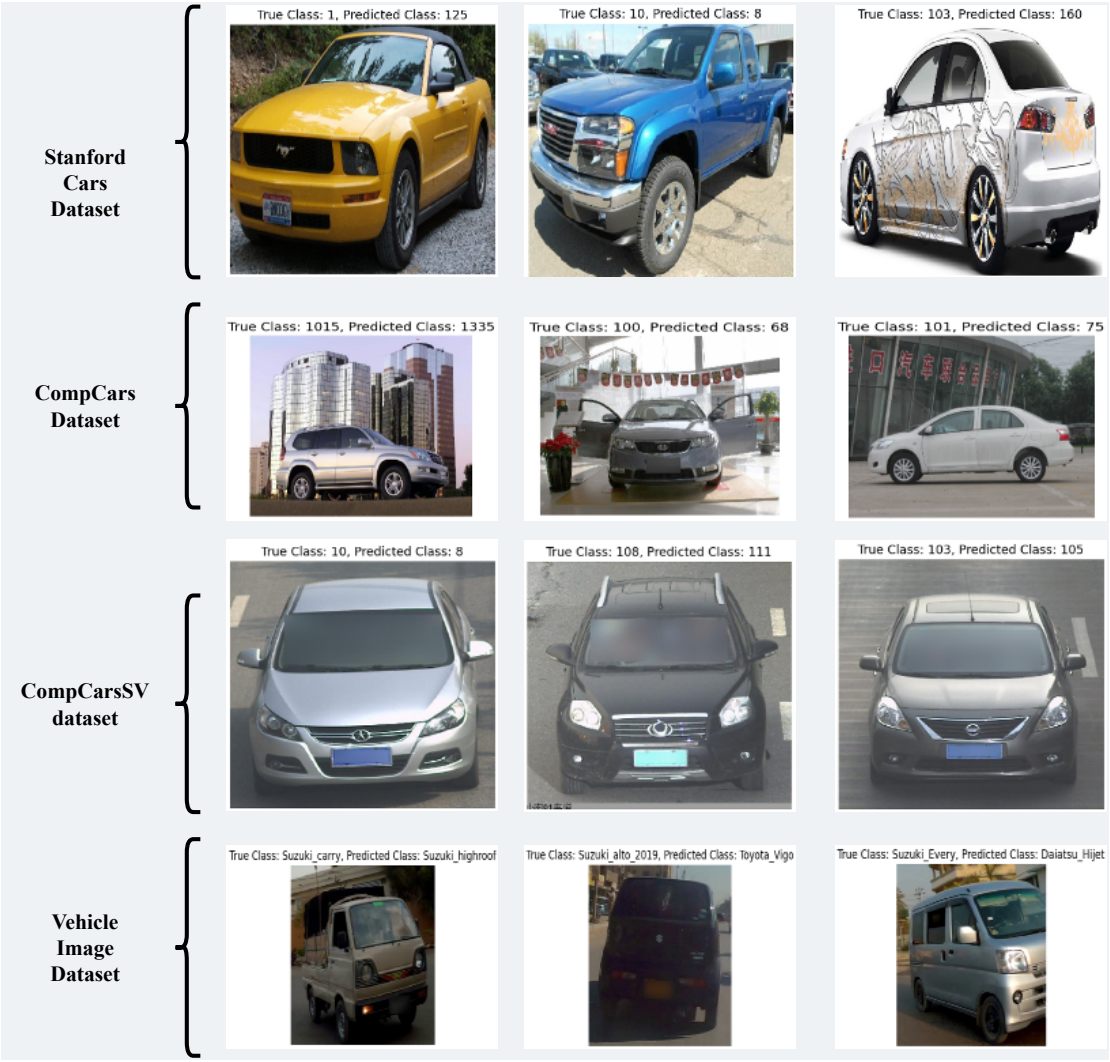


Figure 3.8: Some misclassified images with their true class and predicted class.

of each layer was evaluated by carefully examining the experimental data. The results of the ablation study on the architecture of the model are explained in Table 3.7, and the results of the ablation study on the selected hyperparameters are explained in Table 3.8.

3.3.9.1 Ablation Study on Model's Architecture

- **Effectiveness of AP layer**. It can be demonstrated that the classification accuracy decreases to 93.94% from 94.47% when the pooling layer is not included in the model's architecture. This finding emphasizes the value and efficiency of employing the pooling layer, which significantly enhances the performance of the proposed model.

- **Effectiveness of seperable convolutional layer**. The absence of the separable convolutional layer in the model's design causes the classification

Table 3.7: Experimental results of different layout methods of proposed SimSANet on Stanford Cars dataset

Baseline	AP layer	Separable Conv layer	Multi-headed SA module	Accuracy(%)
✓	×	×	×	92.93
✓	✓	✓	×	93.23
✓	✓	×	✓	93.81
✓	×	✓	✓	93.94
✓	✓	✓	✓	**94.47**

accuracy to drop from 94.47% to 93.81% or a (-0.66%) reduction. This result highlights the importance and effectiveness of using this layer, which greatly improves the suggested model's performance.

- **Effectiveness of Multi-headed SA module**. The exclusion of this module from the model design leads to a discernible decrease in classification accuracy, dropping from 94.47% to 93.23%, representing a reduction of (-1.24%). This outcome underscores the pivotal significance of incorporating this specific module, as its absence notably impacts the model's performance. The presence of this specific module significantly contributes to enhancing the performance of the suggested model.

The three ablation experiments mentioned above demonstrate how each component of the SimSANet model is important and makes a significant contribution.

3.3.9.2 Ablation Studies on Varying Hyperparameters

- **Varying dropout rates**. In this ablation study, the dropout rates are systematically manipulated across four distinct values: 0.02, 0.05, 0.1, and 0.2. The regularization process in neural networks is greatly influenced by dropout rates, which also impact the generalizability of the model. By examining these rates, the impact on the model's performance is scrutinized. Notably, our findings reveal that a dropout rate of 0.02 consistently demonstrates superior performance in terms of generalization, yielding an impressive accuracy of 94.47%. The aforementioned finding implies that a reduced incidence of dropouts considerably enhances the model's capacity for generalization.

- **Varying the number of attention heads**. The impact of varying the number of attention heads (2, 5, 8, 10) on the dataset is rigorously examined. The study's goal is to determine how changing the number of attention heads affects the model's functionality. The configuration with 5 attention heads

Table 3.8: Experimental results on the Stanford Cars dataset with varying hyperparameters

Learning Rate	**Epochs**	**Accuracy(%)**
0.001	25	60.05
0.0001	25	94.47
0.00001	25	91.69
No. of Attention Heads	**Epochs**	**Accuracy(%)**
2	25	94.05
5	25	94.47
8	25	94.11
10	25	93.93
Dropout Rate	**Epochs**	**Accuracy(%)**
0.01	25	94.11
0.05	25	94.47
0.1	25	93.93
0.2	25	93.78

yields the most promising results, achieving a notable accuracy of 94.47%. This finding suggests that there is a key optimization point, meaning that the model's capacity to extract and process pertinent information from the dataset is significantly improved by having a reasonable number of attention heads, which is 5 in this case.

- **Varying learning rates**. The impact of varying learning rates on the performance of the dataset across 25 epochs, distinct trends emerge. A thorough investigation reveals that different learning rates have a substantial impact on the model's ultimate accuracy and convergence. Remarkably, an accuracy of 60.05% is obtained with a learning rate of 0.001, however, a significant improvement to 94.47% is achieved with a rate of 0.0001. However, intriguingly, a further reduction to 0.00001 demonstrates a slight regression to 91.69% accuracy. This ablation study underscores the critical influence of learning rate on model convergence and accuracy.

These hyperparameters are important and beneficial in the SimSANet model, as demonstrated by the above ablation experiments that were mentioned.

3.3.10 Statistical Analysis

The choice of certain layers in the proposed SimSANet architecture is supported by thorough tests carried out on several VMMR datasets. Furthermore,

Table 3.9: Statistical comparison of the proposed model with other models on four VMMR datasets based on p-values of the Wilcoxon rank-sum test.

Comparison	**Stanford Cars**	**CompCars**	**CompCarsSV**	**Vehicle Images**
M1 vs M2	0.0020	0.0076	0.0020	0.0020
M1 vs M3	0.0020	0.0059	0.0020	0.0488
M1 vs M4	0.0137	0.0020	0.0020	0.0098
M1 vs M5	0.0020	0.0020	0.0020	0.0020

this section includes a comprehensive comparison of several approaches, including our suggested model and a number of its variations, together with a detailed analysis of their architectural differences and performance nuances. Here, **M1** represents the full version of our proposed SimSANet model, **M2** refers to our proposed model without utilizing the SA module, **M3** denotes our SimSANet model excluding the AP layer, **M4** represents the SimSANet model without the Separable Convolutional layer, and **M5** is a model that uses just the baseline model.

A statistical test (as mentioned in [65], [66]) was conducted to demonstrate the statistical efficacy of the suggested approach. This test is performed to find strong evidence that contradicts a null hypothesis. Within this framework, it was hypothesized that "The suggested model produces comparable outcomes in contrast to the other models". To disprove this theory, the Wilcoxon rank-sum test [67] is applied. The proposed model is tested using ten separate runs of each model on each dataset. The accuracy of each independent run is logged. The hypothesis is rejected if the p-value is less than 0.05 (5%), indicating that there is not enough data to support this conclusion. The statistical study comparing the performance of our proposed model M1, with other models M2, M3, M4, and M5, across numerous datasets is presented in Table 3.9.

The Wilcoxon rank-sum test is used to evaluate how significant performance variations within the model are. At a 5% confidence level, the null hypothesis is readily rejected. As a result, it can be confidently stated that, in the setting of these examined datasets, model M1 performs better than other models, as evidenced by the persistent pattern of substantial differences across several VMMR datasets.

CHAPTER 4

SEFFNet: A Method for Skin Cancer Classification

In this chapter, the architecture of the proposed SEFFNet model is comprehensively detailed, including its design, components, and the role of sequential attention mechanisms in enhancing performance. An in-depth efficiency analysis is conducted, comparing SEFFNet with other leading approaches in the field. Furthermore, extensive studies on various state-of-the-art public datasets are presented, highlighting the model's strengths and potential limitations.

4.1 Proposed Architecture

Our study introduces a novel approach to improve the multi-class skin cancer classification performance to overcome the complexities encountered in handling the various skin lesion datasets. By leveraging an advanced snapshot ensemble-based feature fusion approach, our method endeavors to synergistically exploit the inherent complexity of the datasets, thereby enhancing classification accuracy. This approach offers a versatile framework for integrating heterogeneous information, allowing for a more comprehensive representation of the extracted features. By fusing features from multiple snapshots, our approach seeks to mitigate the limitations of individual feature representations, such as sensitivity to noise or bias towards certain classes.

In this work, a novel SEFFNet approach for skin lesion classification is proposed. Figure 4.1 represents the overall layout of the suggested framework. Our proposed model operates through three main stages, each contributing distinct functionalities to the overall framework. The process initiates with the normalization and standardization of raw input images, transforming them into tensors to commence further processing. No pre-processing or data augmentation techniques

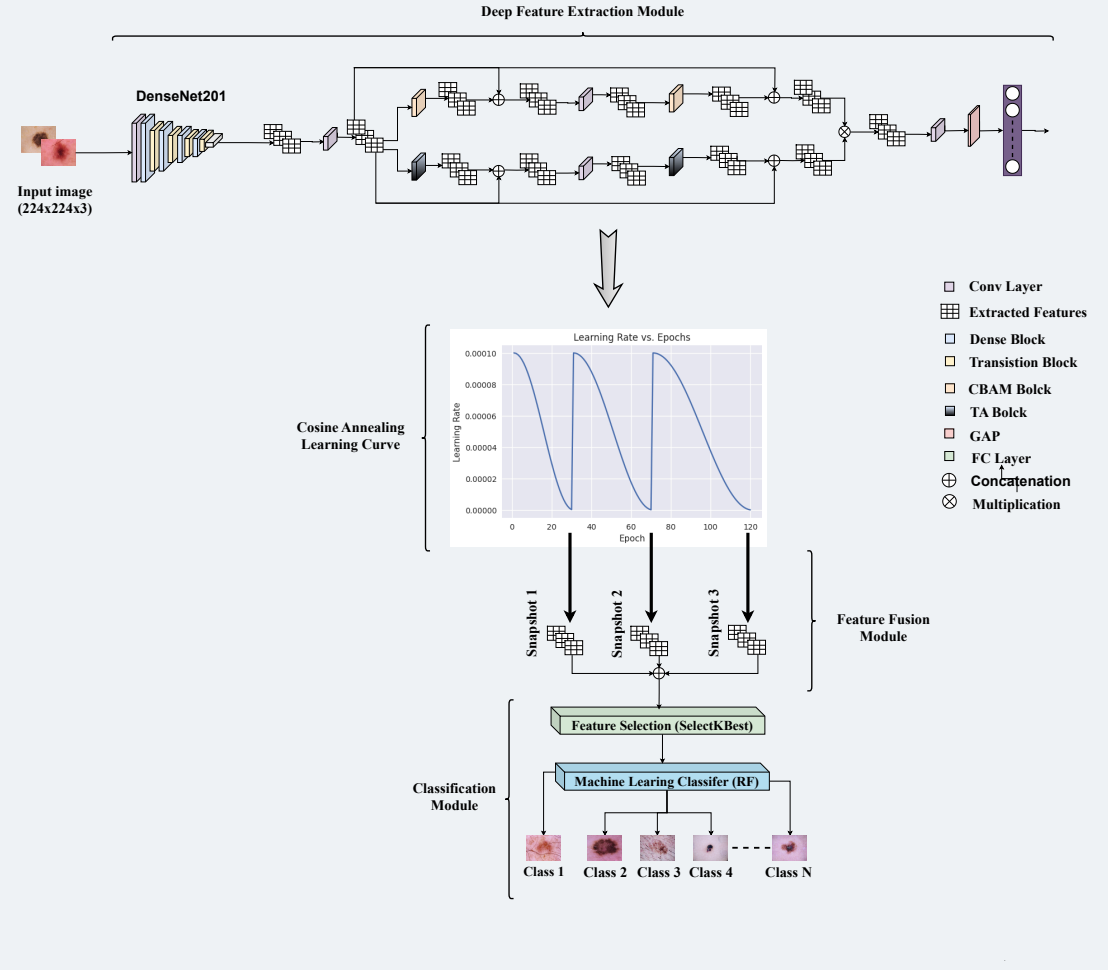


Figure 4.1: The overview of proposed SEFFNet. This architecture includes three main modules: Deep Feature Extraction, Feature Fusion, and Classification Modules. Utilizing the Cosine Annealing Learning Rate scheduling, the deep feature extraction module extracts features (i.e., snapshots) at different training stages, which are then fused and employed for classification via a Random Forest classifier.

are applied at this stage. Following this, the deep feature extraction module operates to capture diverse features from various training phases. These extracted features are then concatenated in the feature fusion module. Finally, in the third stage, the refined features are employed in the classification module, facilitating precise classification through machine learning methodologies. Through this comprehensive approach, our model achieves robust classification performance, effectively addressing the challenges posed by complex datasets.

4.1.1 Deep Feature Extraction Module

In this step, the model operates on input images. Subsequently, to capture comprehensive features, the pre-trained DenseNet201 [68] is employed as the basic

feature extractor. Within the this module, the features, $X_{\text{DenseNet201}} \in \mathrm{R}^{7\times7\times1920}$, retrieved from the DenseNet201 model traverse a sequence of processing steps. At first, these features go through a separable convolutional layer, $X_{\text{Sep Conv11}} \in \mathrm{R}^{7\times7\times512}$ and then a Convolutional Block Attention Module (CBAM) [69] block. Figure 4.2 illustrates the simplified architecture of the CBAM block, which is designed to dynamically re-calibrate the feature maps based on both channel-wise and spatial-wise attention mechanisms. This block allows the selection of relevant features, enhancing the discriminative power of the extracted features.

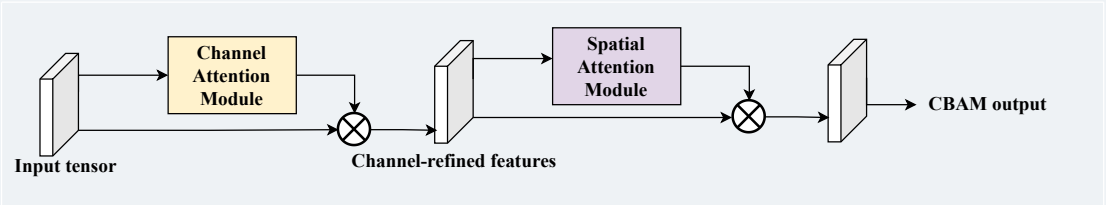


Figure 4.2: The overview of the CBAM block.

$$X_{\text{Sep Conv11}} = \text{Conv2d}(X_{\text{DenseNet201}}) \tag{4.1}$$

$$X_{\text{CBAM1}} = \text{CBAM}(X_{\text{Sep Conv11}}) \tag{4.2}$$

The resulting feature representation, $X_{\text{CBAM1}} \in \mathrm{R}^{7\times7\times512}$, is then concatenated by the output of the previous separable convolutional layer, $X_{\text{Sep Conv11}}$. The representation of the feature stream, $X_{\text{Concat11}} \in \mathrm{R}^{7\times7\times1024}$, is then refined by passing it through an additional separable convolutional layer, $X_{\text{Sep Conv12}} \in \mathrm{R}^{7\times7\times512}$. Next, the output is sent into a second CBAM block, $X_{\text{CBAM2}} \in \mathrm{R}^{7\times7\times512}$, which applies more levels of refinement to ensure a rich and nuanced feature representation. Furthermore, the resulting output X_{CBAM2} is then concatenated again with the $X_{\text{Sep Conv11}}$ and yields $X_{\text{Concat12}} \in \mathrm{R}^{7\times7\times1024}$.

$$X_{\text{Concat11}} = [X_{\text{Sep Conv11}}, X_{\text{CBAM1}}] \tag{4.3}$$

$$X_{\text{Sep Conv12}} = \text{Conv2d}(X_{\text{Concat11}}) \tag{4.4}$$

$$X_{\text{CBAM2}} = \text{CBAM}(X_{\text{Sep Conv12}}) \tag{4.5}$$

$$X_{\text{Concat12}} = [X_{\text{Sep Conv11}}, X_{\text{CBAM2}}] \tag{4.6}$$

Concurrently, within the same branch, features extracted from the DenseNet201 model again traverse through a separable convolutional layer and engage with a Triplate Attention (TA) [70] block. Figure 4.3 illustrates the simplified architecture of the TA block. Within this TA block, a sophisticated mechanism is utilized to intricately orchestrate the interaction between the spatial dimensions (height and width) and the channel dimension of the input tensor. By incorporating this block, the model gains the capability to discern and emphasize significant patterns and correlations. Now, similar to the previous processing pipeline, the resulting output, $X_{\text{TA1}} \in \text{R}^{7\times7\times512}$, is then concatenated by the output of the previous separable convolutional layer, $X_{\text{Sep Conv11}}$. Furthermore, the output of the process $X_{\text{Concat21}} \in \text{R}^{7\times7\times1024}$, passes through another separable convolution layer $X_{\text{Sep Conv22}} \in \text{R}^{7\times7\times512}$, and its output is directed to the second TA block X_{TA2}, for further distilled and refined features. Furthermore, the resulting output X_{TA2} is then concatenated again with the $X_{\text{Sep Conv11}}$, and produces $X_{\text{Concat22}} \in \text{R}^{7\times7\times1024}$.

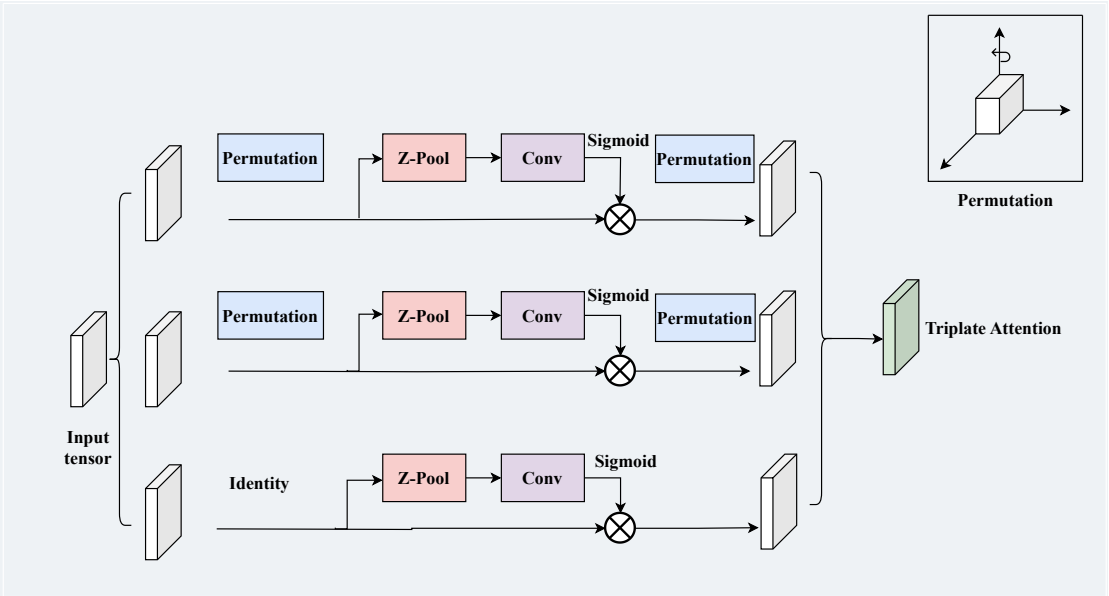


Figure 4.3: The overview of the TA block.

$$X_{\text{TA1}} = \text{TA}(X_{\text{Sep Conv11}}) \tag{4.7}$$

$$X_{\text{Concat21}} = [X_{\text{Sep Conv11}}, X_{\text{TA1}}] \tag{4.8}$$

$$X_{\text{Sep Conv22}} = \text{Conv2d}(X_{\text{Concat21}}) \tag{4.9}$$

$$X_{\text{TA2}} = \text{TA}(X_{\text{Sep Conv22}}) \tag{4.10}$$

$$X_{\text{Concat22}} = [X_{\text{Sep Conv11}}, X_{\text{TA2}}]$$

In the final stage of this module, the convergence of these parallel processing sequences, denoted as $X_{\text{Mul}} \in \text{R}^{7\times7\times1024}$, is resulting by the multiplication of extracted features from second CBAM block X_{CBAM12} and the second TA block $X_{\text{Triplate Attention22}}$.

$$X_{\text{Mul}} = X_{\text{Concat12}} \odot X_{\text{Concat22}} \tag{4.11}$$

Subsequently, X_{Mul}, again traverses through a separable convolutional layer, $X_{\text{Sep Conv}} \in \text{R}^{7\times7\times256}$ and a Global Average Pooling layer, $X_{\text{GAP}} \in \text{R}^{1\times1\times256}$ followed by a flattening layer, culminating in the final layer, X_{Final} of this module.

$$X_{\text{Sep Conv}} = \text{Conv2d}(X_{\text{Mul}}) \tag{4.12}$$

$$X_{\text{GAP}} = \text{Global Average Pooling}(X_{\text{Sep Conv}}) \tag{4.13}$$

$$X_{\text{Final}} = \text{Flatten}(X_{\text{GAP}}) \tag{4.14}$$

4.1.2 Feature Fusion Module

Within this feature fusion module, the cosine annealing learning rate scheduler, as proposed by [71] was adopted, which efficiently guides the model's performance towards diverse local minima. This integration of cosine annealing significantly enhances the efficacy and adaptability of our framework, leading to improved performance and generalization capabilities. To further leverage this approach, a multi-stage snapshot-based deep feature fusion strategy is also employed. This approach entails extracting three distinct feature representations obtained from different training stages. This is achieved by saving the model weights at specific intervals during training: at 30 epochs, 70 epochs, and 120 epochs. Subsequently, these diverse features $X_{\text{Snapshot1}} \in \text{R}^{256}$, $X_{\text{Snapshot2}} \in \text{R}^{256}$ and $X_{\text{Snapshot3}} \in \text{R}^{256}$ are combined to form an ensemble, resulting in the creation of $X_{\text{fused features}} \in \text{R}^{768}$.

$$X_{\text{fused features}} = [X_{\text{Snapshot1}}, X_{\text{Snapshot2}}, X_{\text{Snapshot3}}] \tag{4.15}$$

This fusion approach enables the model to leverage a comprehensive view of data, incorporating knowledge acquired at different stages of its learning journey. By concatenating diverse feature representations, the model gains a more holistic understanding of patterns and structures within the dataset, thereby enhancing

its ability to generalize and perform effectively on unseen data.

4.1.3 Classification Module

In the classification module, the Random Forest (RF) is employed as a machine-learning classifier. Before initiating the classification process, the SelectKBest feature selection method is used. This technique assesses the significance of each feature based on statistical measures (ANOVA and F-value). By integrating SelectKBest feature selection with the RF classifier, the classification accuracy and interpretability of the model are aimed to be improved. By combining the advantages of both methods, this approach allows for efficient categorization of data instances into predefined classes.

4.1.4 Hybrid loss Function

In this section, a Hybrid loss function is implemented that dynamically incorporates class frequencies to compute a hybrid of label-smoothed cross-entropy and focal loss. The number of target classes, initial values for hyperparameters, and weights for cross-entropy and focal-loss components are among the parameters that the class sets during the initialization. Throughout the training, it dynamically adjusts these parameters based on the current epoch relative to the total number of epochs, using a decay factor to reduce the influence of specific hyperparameters as training progresses. The class also calculates weights for each class in a batch by inversely correlating to their frequency, which helps correct imbalances by giving more weight to underrepresented classes.

During the loss computation, the algorithm converts model outputs to log probabilities using the Softmax function. It then applies label smoothing by initializing a distribution tensor that adjusts for overconfidence in the model's predictions. The smoothed cross-entropy loss, calculated as the mean of the product of this distribution tensor and the negative log probabilities, is used along with the focal loss. The focal loss focuses more on challenging or misclassified examples by utilizing a formula that incorporates alpha and an adjustment based on the probability of correct classification, enhancing focus on hard-to-classify instances.

To address the class imbalance and concentrate training on hard data points, the final loss output is the weighted sum of the smoothed cross-entropy and focused losses. Through the dynamic adjustment of hyperparameters and the integration of strategies to manage class imbalance, this Hybrid loss function provides an advanced mechanism for neural network training process optimization, which may improve model robustness and generalization performance.

Algorithm 1 Hybrid loss with Dynamic Hyperparameter Adjustment

class HybridLoss
parameters:
    num_classes: Number of classes
    initial_alpha: Initial value for focal loss alpha
    initial_gamma: Initial value for focal loss gamma
    initial_ce_weight: Initial weight for cross-entropy loss
    initial_focal_weight: Initial weight for focal loss
    initial_smoothing: Initial smoothing parameter for label smoothing

procedure ComputeLoss(output, target, current_epoch, total_epochs)
    decay_factor $\leftarrow \frac{\text{current_epoch}}{\text{total_epochs}}$
    $\alpha \leftarrow$ initial_alpha $\times$ (1 $-$ decay_factor)
    $\gamma \leftarrow$ initial_gamma $\times$ (1 $-$ decay_factor)
    focal_weight $\leftarrow$ initial_focal_weight $\times$ (1 $-$ decay_factor)
    smoothing $\leftarrow$ initial_smoothing $\times$ (1 $-$ decay_factor)
    class_counts $\leftarrow$ zero vector of size num_classes
    for $c = 0$ to num_classes-1 **do**
        class_counts[c] $\leftarrow$ sum(target $==$ c)
    end for
    class_weights $\leftarrow \frac{1.0}{\text{class_counts}+1e-6}$
    Normalize class_weights to sum to 1
    log_probs $\leftarrow$ log(softmax(output, dim $= 1$))
    Create true_dist, adjusted for smoothing
    ce_loss $\leftarrow$ mean(sum($-$true_dist $\times$ log_probs, dim $= 1$))
    pt $\leftarrow$ exp($-$ce_loss)
    focal_loss $\leftarrow \alpha \times (1 - \text{pt})^{\gamma} \times$ ce_loss
    total_loss $\leftarrow$ initial_ce_weight $\times$ ce_loss + focal_weight $\times$ focal_loss
    return total_loss
end procedure

4.2 Materials and Implementation Details

4.2.1 Datasets

Experiments have been conducted on two distinct benchmark skin lesion datasets: HAM10000 [72] and ISIC 2017 [73]. The HAM10000 dataset contains a total of 10015 images from seven categories: Actinic Keratoses and Intraepithelial Carcinoma (akiec), Basal cell carcinoma (bcc), Benign keratosis (bkl), Dermatofibroma (df), Melanoma (mel), nevus (nv), and Vascular (vasc). Of these 8912 images were utilized for training and 1103 for testing. On the other hand, the ISIC 2017 dataset includes 2750 images overall from three categories: melanoma (mel), nevus (nv), and seborrheic keratosis (sk). Of these, 2000 images were employed for training, 150 for validation, and 600 for testing. A few sample images from these datasets are shown in Figure 4.4 for a visual depiction. Table 4.1 complies with the essential details about the datasets considered here for experiments.

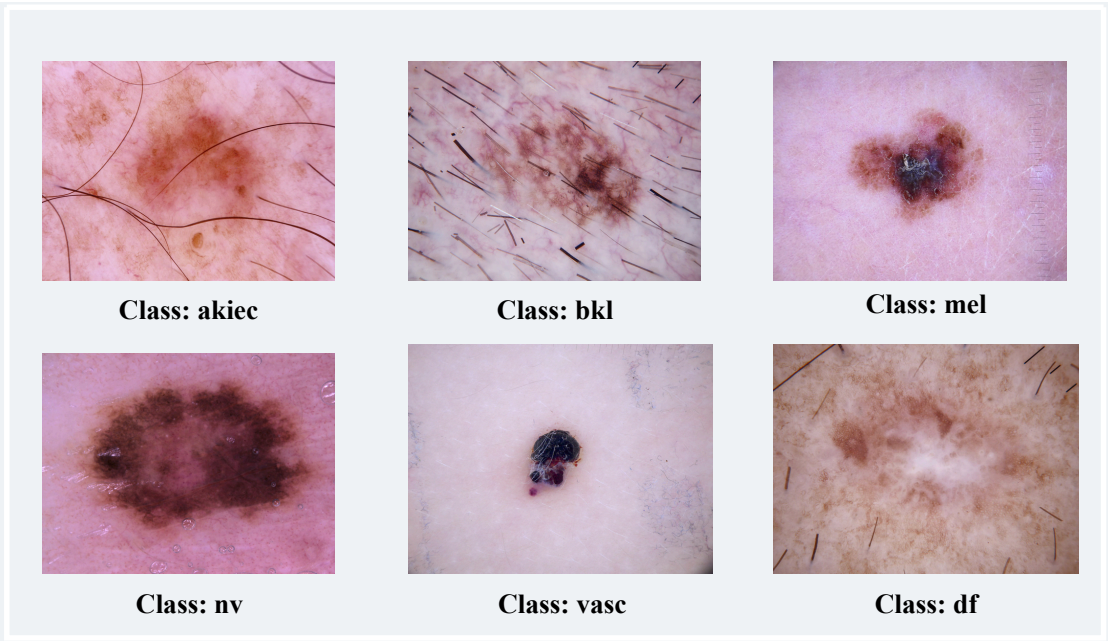


Figure 4.4: Few examples of Skin Cancer Samples from Diverse Datasets.

Table 4.1: Statistics of the datasets used in the present experimentation.

Dataset	Training images	Validation images	Testing images	Split ratio	Classes
HAM10000	8912	-	1103	80:-:20	7
ISIC 2017	2000	150	600	72:5:22	3

4.2.2 Implementation Details

4.2.2.1 Experimental Setup

In this study, Pytorch was utilized to develop the model, which was executed on the Kaggle platform. The system was equipped with NVIDIA Tesla T4 (x2) GPUs and a total of 32GB of DDR4 RAM. Our training approach involved using a cosine annealing learning rate, starting from an initial learning rate of 0.0001 and gradually decreasing to a minimum of 0.00000001. Each training snapshot was taken at specific intervals during the training process: after 30 epochs, 70 epochs, and 120 epochs, totaling three snapshots for our model. Additionally, a dropout rate of 0.02 was applied, with sigmoid used as the activation function and Adam as the optimizer throughout the experiments. Notably, a Hybrid loss function, amalgamating focal loss and multi-class cross-entropy loss, was employed to optimize model performance across experiments.

4.2.2.2 Evaluation Metrics

For a detailed discussion on the evaluation metrics used in this study, please refer to subsection 3.2.4

4.3 Results

4.3.1 Performance of different baseline models

Our evaluation process involves a comprehensive analysis of several baseline models, aiming to identify the model that best aligns with our experimental setup. The outcomes of these evaluations are succinctly summarized in Table 4.2, offering a detailed examination of each baseline model's performance. From the findings, it's evident that among the considered models, DenseNet201, pre-trained on the ImageNet dataset comprising 1.2M images across 1000 classes, emerges as the top-performing baseline model. Utilizing DenseNet201 as our baseline model leads to notable enhancements in performance metrics, notwithstanding its parameter count of 20.2M.

The architecture of the snapshot-based feature ensemble network (SEFFNet) comprises two essential modules: basic snapshot feature extraction and machine learning-based classification. Additionally, comprehensive performance metrics for SEFFNet across two skin lesion datasets are presented in this section. Remarkably, the proposed approach achieves an impressive accuracy of 92.93% on the HAM10000 and 81.17% on the ISIC 2017 datasets. The evaluation highlights that

Table 4.2: The performance of different baseline models on the HAM10000 and ISIC 2017 datasets. All scores are in %. Bold values represent the best results.

Dataset	Baseline	Snap No.	Accuracy	Precision	Recall	F1-Score
HAM10000	ResNet50	1	91.12	90.68	91.12	90.59
		2	91.84	91.72	91.84	91.74
		3	91.93	91.60	91.93	91.58
	RegNet	1	90.57	90.34	90.57	90.35
		2	91.39	90.98	91.39	90.95
		3	91.66	91.42	91.66	91.28
	EfficientNetV2S	1	90.48	90.06	90.48	90.05
		2	91.02	90.35	91.02	90.53
		3	91.57	91.30	91.57	91.37
	DenseNet201	1	92.02	91.69	92.02	91.61
		2	92.20	91.74	92.20	91.82
		3	**92.66**	**92.39**	**92.66**	**92.46**
ISIC 2017	ResNet50	1	71.17	68.76	71.17	69.13
		2	73.5	71.45	73.5	70.93
		3	76.67	75.11	76.67	74.43
	RegNet	1	70.17	66.52	70.17	64.39
		2	73.87	72.31	73.87	72.92
		3	74.5	72.87	74.5	72.21
	EfficientNetV2S	1	75.17	75.27	75.17	75.11
		2	77.33	76.62	77.33	76.81
		3	79.00	77.99	79.00	78.07
	DenseNet201	1	74.83	74.58	74.83	74.25
		2	**79.67**	**78.71**	**79.67**	**78.63**
		3	77.67	77.65	77.67	76.19

employing the SEFFNet approach consistently yields impressive results across all datasets. This underscores the effectiveness and robustness of the SEFFNet architecture in skin lesion classification tasks.

4.3.2 Quantitative Analysis of Results

The classification accuracy results on the HAM10000 and ISIC 2017 datasets are summarized in Table 4.3. Without using any augmentations and pre-processing techniques, our suggested method yields commendable testing accuracies on the HAM10000 dataset. Specifically, accuracies of 92.02%, 92.20%, and 92.66% are achieved across three distinct training snapshots. Conversely, on the ISIC 2017 dataset, our approach exhibits testing accuracies of 74.83%, 79.67%, and 77.67%

across different snapshots.

Table 4.3: Results after the basic feature extraction module on the HAM10000 and ISIC 2017 datasets. All scores are in %. Bold values represent the best results.

Dataset	Snapshot No.	Acc	Pre	Rec	F1
HAM10000	1	92.02	91.69	92.02	91.61
	2	92.20	91.74	92.20	91.82
	3	**92.66**	**92.39**	**92.66**	**92.46**
ISIC 2017	1	74.83	74.58	74.83	74.25
	2	**79.67**	**78.71**	**79.67**	**78.63**
	3	77.67	77.65	77.67	76.19

Table 4.4: Final classification results after employing the SelectKBest approach and ML classifiers on the HAM10000 and ISIC 2017 datasets. All scores are in %. Bold values represent the best results.

Dataset	Classifier	Selected Features	Acc	Pre	Rec	F1
HAM10000	RF	753	**92.93**	**92.66**	**92.93**	**92.71**
	SVM	**55**	92.84	92.41	92.84	92.52
ISIC 2017	RF	740	**81.17**	80.50	**81.17**	**80.29**
	SVM	**31**	80.67	**80.04**	80.67	79.35

Table 4.4 presents the outcomes following the application of the feature concatenation and machine learning classifiers for the final classification. Utilizing the RF as the classifier, the proposed method achieves a final accuracy of 92.93% on the HAM10000 dataset using a mere 753 features, out of the total 765 extracted features. Similarly, on the ISIC 2017 dataset, an accuracy of 81.17% was attained by employing only 740 features out of the same 765.

Employing SVM as the classifier, the proposed approach obtained an accuracy of 92.83%, slightly lower by 0.10% compared to the RF classifier's accuracy. Notably, this performance was achieved while employing only 55 selected features out of 765 extracted features. In contrast, on the ISIC 2017 dataset, utilizing a subset of 31 features, the suggested method achieved an accuracy of 80.67%. This finding shows the versatility of our approach in leveraging different classifiers. furthermore, the effectiveness of both RF and SVM as the classifiers on both datasets highlights their potential utility in accurately classifying dermatological images.

Figure 4.5 shows comprehensive accuracy and loss trends across both training and testing images. The depicted model exhibits an increasing trend in accuracy as the number of training epochs advances, coupled with a simultaneous decrease in loss. Moreover, the absence of overfitting issues is evident in the loss curves, owing to the utilization of a Hybrid loss function. This loss function treats equally the

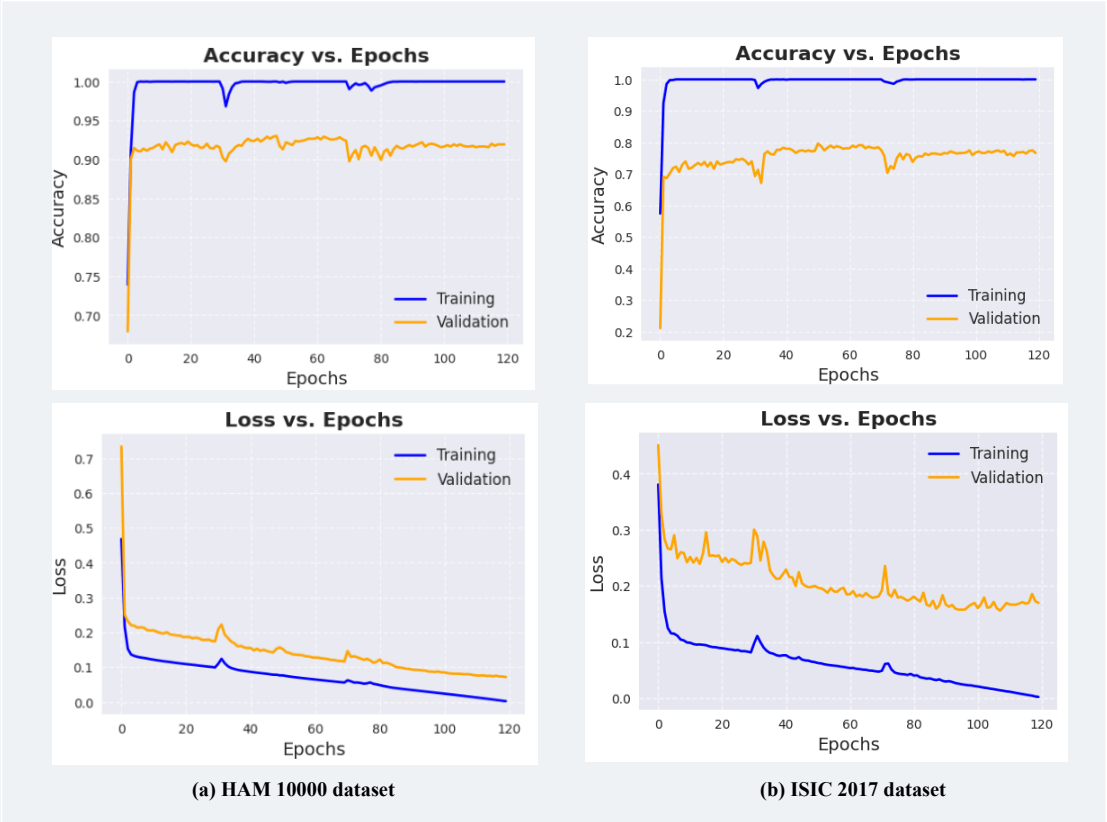


Figure 4.5: Accuracy and loss curves generated by the proposed method on HAM10000 and the ISIC 2017 datasets.

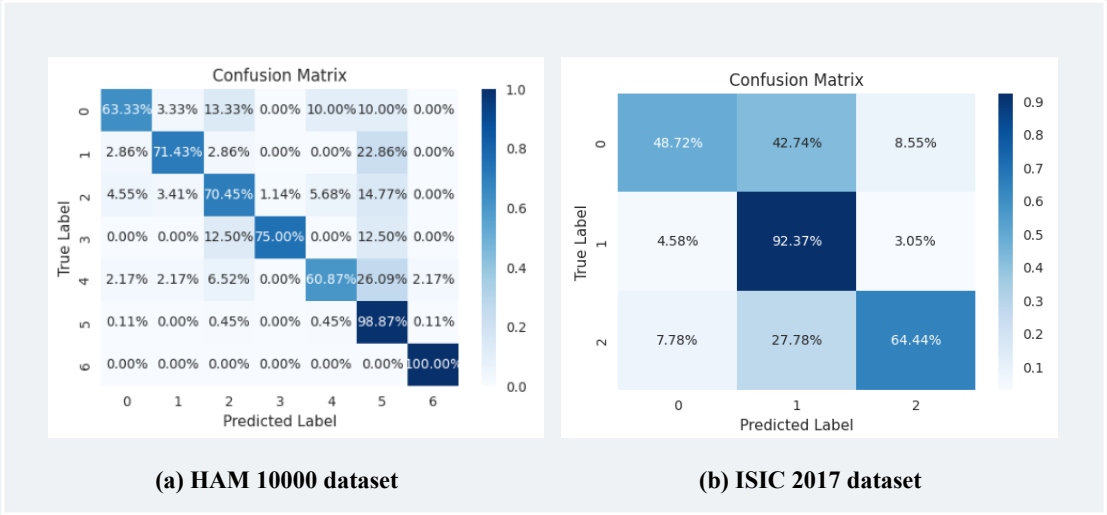


Figure 4.6: Confusion matrices obtained for HAM10000 and ISIC 2017 datasets.

less sampled classes as well as the more frequently sampled classes, ensuring robust model training and generalization across diverse class distributions. Furthermore, The confusion matrix illustrated in Figure 4.6 evaluates the performance of our classification model on the both HAM10000 and ISIC 2017 datasets. It provides

insight into specific areas where the model performs well and where it struggles. This analysis also highlights the importance of class-wise performance evaluation in understanding the strengths and limitations of the proposed model.

4.3.3 Comparative Analysis on the HAM10000 Dataset

Table 4.5 summarizes the classification results on the HAM10000 dataset. Our newly devised SEFFNet method achieves a notable accuracy of 92.93%, showcasing robust performance even without employing pre-processing or augmentation techniques. Additionally, the method demonstrates a high precision of 92.66%, recall of 92.93%, and an F1 score of 92.71%. In comparison, [74] proposed a DCNN-based approach augmented with data augmentation strategies, achieving a balanced accuracy of 85.3% on the augmented HAM10000 dataset. [75] achieved an accuracy of 87.91% using EfficientNet B4 as the baseline model. On the other hand, [76] attained an accuracy of 91.20% by employing the DenseNet169 architecture alongside pre-processing and an undersampling technique. However, it's noteworthy that their undersampling approach potentially diminishes information. [77] achieved an accuracy of 91.70% on the augmented HAM10000 dataset using a feature selection-based approach.

Table 4.5: Comparison of the classification results on the HAM10000 dataset. All scores are in %. Bold values represent the best results.

Work	Method	Acc	Pre	Rec	F1
[77]	CNN + Feature Selection	91.70	92.00	91.85	91.92
[75]	EfficientNet B4 based approach	87.91	88.00	88.00	87.00
[74]	DCNN based approach	85.30	-	-	-
[78]	DenseNet201-based approach	86.91	-	-	-
[79]	RegNetY-320 based approach	91.00	-	-	88.00
[80]	DenseNet169 based approach	92.25	-	-	-
[81]	wide-ShuffleNet	86.33	-	-	-
[76]	DenseNet169 based approach	91.20	-	-	91.70
[82]	DenseNet-121 + MLP	81.00	-	-	-
[83]	XAI-based a custom CNN model	82.00	-	-	-
[84]	SDCA-based feature fusion + ELM	87.02	-	86.89	-
[85]	FFN with ISB and IFB blocks	92.00	69.00	92.00	73.00
Ours, 2024	**SEFFNet**	**92.93**	**92.66**	**92.93**	**92.71**

Using DenseNet201 architecture as the baseline model, Lariba et al. [86] reported an accuracy of 86.91% utilizing an enhanced ResNet50 architecture on the pre-processed HAM10000 dataset, featuring fewer images than the original, sig-

nificantly lower than our model's performance. On the augmented HAM10000 dataset, [79] proposed a model using RegNetY-320 and obtained an accuracy of 91.00%. Utilizing DenseNet169 as the baseline model, [80] obtained a significance accuracy of 92.25%. [81] introduced a novel approach to skin lesion classification, combining a new segmentation methodology with wide-ShuffleNet. By employing entropy-based weighting and first-order cumulative moment calculations, lesions are effectively separated from the background. This segmentation output is then fed into a wide-ShuffleNet structure to accurately classify the type of skin lesion, which obtained an accuracy of 86.33%. Despite using the segmentation-based classification approach, this approach achieved lesser accuracy than our proposed approach. [82] attained an accuracy of 81.00% using DenseNet-121 with multilayer perceptron (MLP) as the classifier. [83] achieve an 82.00% accuracy using explainable AI (XAI), utilizing Grad-CAM and Grad-CAM++ to reveal the model's decision-making process. [84] introduced a novel approach for classification and segmentation employing multiple models. Their method incorporates two CNN-based architectures, coupled with a novel summation discriminant correlation analysis (SDCA) for feature fusion, followed by an Extreme Learning Machines (ELM) classifier. This approach obtained an accuracy of 87.02%. [85] presented a fully fused network (FFN), comprising an improved single block (ISB) and an enhanced fusion block (IFB) with a late feature fusion approach, aimed at optimizing multi-class recognition of skin diseases. This approach achieved an accuracy of 92% on the HAM10000 dataset.

4.3.4 Comparative Analysis on the ISIC 2017 Dataset

Table 4.6 summarizes the classification results on the ISIC 2017 dataset. Our proposed method, SEFFNet, achieves an accuracy of 81.17%, with a precision of 80.50%, a recall of 81.17%, and an F1-score of 80.29% on the ISIC 2017 dataset. On the other hand, [87] achieved an accuracy of only 56.00% using VGG19 architecture with a precision of only 20%, which is much less accurate than our model. [28] achieved a recall value of 78% using the DermoDoctor models. [88] presented a framework that integrates boundary segmentation and lesion classification stages for skin lesion diagnosis, leveraging different deep learning architectures like Inceptionv3, ResNet50, and InceptionResNetv2. This approach attained an accuracy of 81.57% using ResNet50 as the baseline model, slightly surpassing our proposed model's accuracy by 0.40%. However, our model exhibits a recall of 81.17%, outperforming theirs by 5.42%. [89] obtained a precision value of 72% using the selective convolutional descriptor aggregation (SCDA) architecture, which is 6% lower than our newly developed model.

Table 4.6: Comparison of the classification results on the ISIC 2017 dataset. All scores are in %. Bold values represent the best results.

Work	Method	Acc	Pre	Rec	F1
[90]	ResNet50 + MV	-	-	15.00	-
[91]	ARL-CNN	-	-	77.00	-
[89]	SCDA	-	72.00	-	-
[88]	FrCN with ResNet50	**81.57**	-	75.75	-
[28]	DermoDoctor	-	79.00	78.00	-
[87]	VGG19	56.00	20.00	41.00	26.00
[92]	Multi-feature-fusion	79.50	79.10	79.50	79.20
Ours, 2024	**SEFFNet**	81.17	**80.50**	**81.17**	**80.29**

[91] introduced an attention residual learning convolutional neural network (ARL-CNN) model, which employs attention residual learning techniques along with convolutional neural networks, featuring ARL blocks, global average pooling, and a classification layer to enhance accuracy and performance in lesion classification tasks. It obtained a recall value of 77.00%, whereas our proposed model obtained a recall value of 81.17%. [92] introduced an innovative technique for improving representation quality and discriminative power in deep learning models. By combining features from different layers of Xception and BiT-M models and incorporating the squeeze-and-excitation (SEA) attention block. Despite employing a multi-feature fusion approach, their method attained an accuracy of 79.50%, marginally trailing ours by 0.67%.

4.3.5 Impact of the proposed Hybrid loss function

Figure 4.7 and Table 4.7 present the loss curves over epochs for the proposed model trained using two different loss functions, Cross-Entropy loss (CE loss) and Hybrid loss, respectively. The x-axis denotes the number of training epochs, while the y-axis represents the loss value.

In the left graphs (a) and (c), which use CE loss, both training and testing loss decrease sharply in the initial epochs. Following this abrupt decrease, the training loss levels off and stays comparatively constant, indicating that the model's learning ability with this particular loss function has peaked. However, the test loss exhibits notable oscillations, which may point to overfitting as the model becomes overly tuned to the training set without making a good generalization to the test set that hasn't been seen yet.

Utilizing the Hybrid loss, a more progressive and consistent drop in both training and test loss is observed in the right graphs (b) and (d). This implies that a

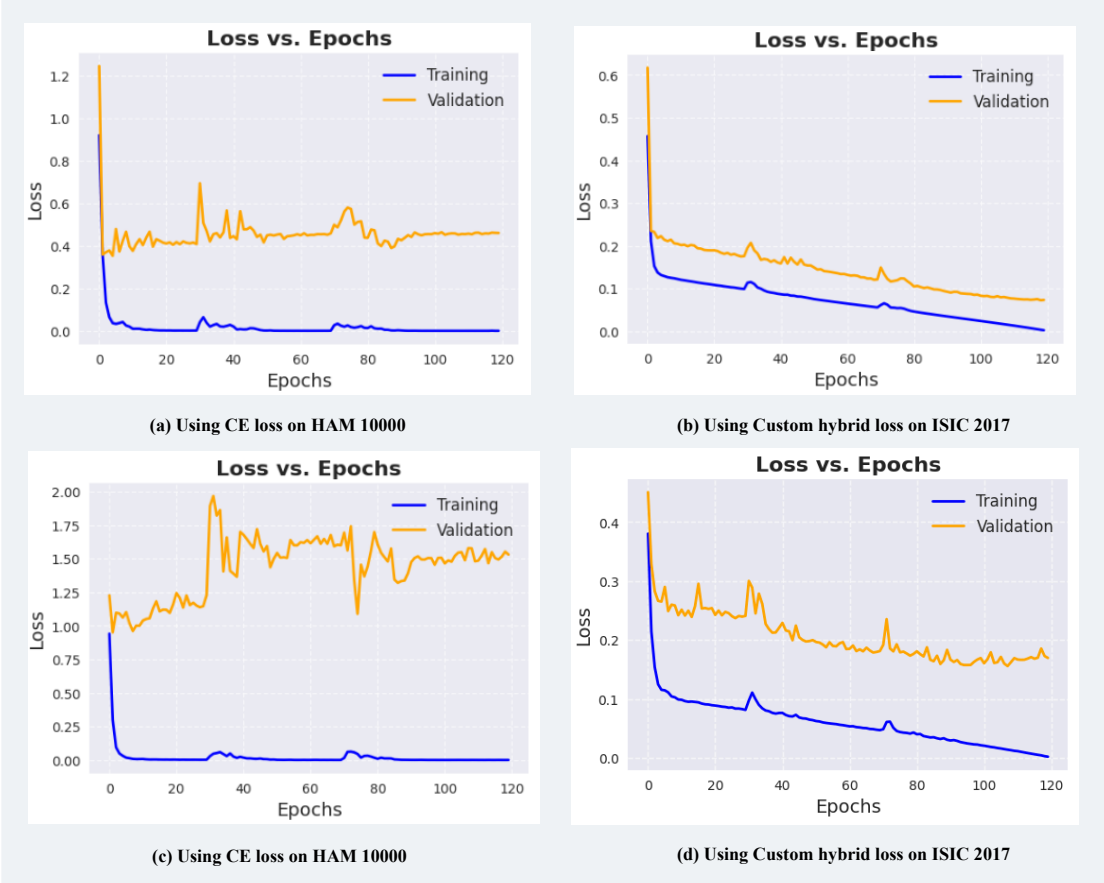


Figure 4.7: Results using different loss functions on the HAM10000 and ISIC 2017 datasets.

better-regulated learning process may be offered by hybrid loss. Although there is still some fluctuation in the validation loss, it does not change as much as the CE loss, suggesting improved generalization on the test set. This might imply that the Hybrid loss reduces overfitting, leading to a model that performs more consistently across different datasets.

Table 4.7: The performance of different loss functions on the HAM10000 and ISIC 2017 datasets. All scores are in %. Bold values represent the best results.

Dataset	Loss Function	Train Loss	Test Loss
HAM10000	CE loss	0.0001	0.4576
	Hybrid loss	0.0022	**0.0600**
ISIC 2017	CE loss	0.0009	1.1086
	Hybrid loss	0.0030	**0.1266**

These observations are crucial for understanding the impact of loss functions on model training dynamics and generalization. By incorporating a dynamic hybrid loss approach into the loss function, the proposed Hybrid loss can better capture

the nuances of the data and improve model performance in real-world scenarios. Additionally, further research is needed to explore the potential benefits of using different loss functions for specific tasks or datasets.

4.3.6 Model Interpretability using Grad-CAM

Deep learning was formerly like using a black box, meaning that it was difficult to understand the reasoning behind a model's choices, especially in important fields like healthcare. Health professionals, entrusted with decisions that impact patients' lives, face the daunting task of interpreting these decisions, often without sufficient justification. As such, there existed the possibility of drawing erroneous inferences. To enhance the interpretability of the deep learning model for skin lesion classification, Gradient-Weighted Class Activation Mapping (Grad-CAM) [93] was employed as our visualization technique.

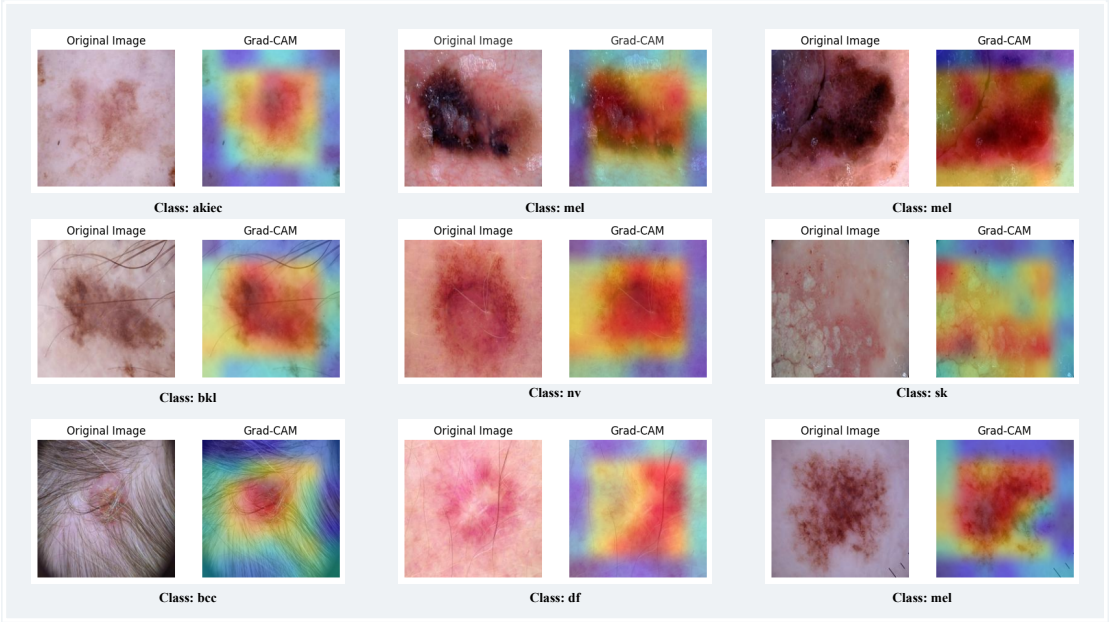


Figure 4.8: Grad-CAM results on various samples images.

The Grad-CAM visualization demonstrates the proposed model's emphasis on skin lesions, providing valuable insights into its decision-making process. The original image shows a skin lesion with uneven boundaries and diverse pigmentation, characteristics that are frequently seen in photos of skin lesions. In the meantime, the Grad-CAM generates a heatmap overlay on the original images by utilizing a color palette that spans from colder tones like blue to warmer tones like red and yellow. These colors signify the importance of different regions within the image to the model's prediction. In Figure 4.8, each image represents the original image

as well as the Grad-CAM visualization corresponding to a specific class within the HAM10000 dataset.

The red hues delineate regions of maximal activation, signifying crucial areas of focus for the model. The visual analysis of the heatmap reveals that the model predominantly focuses on the central, darker region of the lesion, which is marked in red. These regions typically correspond to features such as irregular borders, color aberrations, or unique patterns indicative of various skin ailments. Meanwhile, yellow and orange regions denote areas of heightened activation, albeit to a lesser extent than the red regions. While still influential in the model's decision-making process, they are deemed less critical compared to the prominently red regions. The regions highlighted in green signify moderate activation and denote areas of intermediate importance, While not the primary focus, these regions provide background context and supplementary information. Lastly, the blue areas denote minimal activation and represent the least significant regions in the model's decision-making process. Typically, these areas correspond to background regions or parts of the skin lacking distinctive features.

By integrating Grad-CAM into the analysis, transparency of the proposed model is improved, and clinicians are provided with an intuitive tool to verify and trust the model's predictions. This approach aligns with the broader goal of making AI-driven diagnostics more explainable and reliable in medical practice.

4.3.7 Results of the Ablation Study

An ablation study on the HAM10000 and ISIC 2017 datasets has been conducted to assess the design of the proposed SEFFNet. The impact of various components in the proposed model architecture on the performance task has been systematically evaluated. Specifically, the effects of the baseline, CBAM block, TA block, and their combinations on the model's performance have been investigated.

Table 4.8 shows the ablation results on the HAM10000 dataset. Initially, employing only the baseline configuration achieves a respectable accuracy of 92.38%, with consistent precision, recall, and F1-scores. Introducing the CBAM block along with the baseline model marginally improves the accuracy to 92.47% but shows a slight decrease in recall. Interestingly, the incorporation of the TA block along with the baseline model leads to a decrease in accuracy compared to the baseline. However, when the CBAM and TA blocks are combined, a significant boost in accuracy to 92.93% is obtained, accompanied by improvements in precision, recall, and F1 scores.

Table 4.9 shows the ablation results conducted on the ISIC 2017 dataset. Notably, an impressive accuracy of 80.00% was achieved when solely employing the

Table 4.8: Experimental results of different layouts of the proposed SEFFNet on the HAM10000 dataset. All scores are in %. Bold values represent the best results.

Baseline	CBAM	TA	CBAM + TA	Acc	Pre	Rec	F1
✓	×	×	×	92.38	92.36	92.38	92.25
✓	✓	×	×	92.47	92.02	92.47	92.17
✓	×	✓	×	92.29	91.85	92.29	91.96
✓	-	-	✓	**92.93**	**92.66**	**92.93**	**92.71**

baseline configuration. However, integrating either the CBAM block or the TA block individually did not yield significant improvements in accuracy. Intriguingly, combining both CBAM and TA blocks led to a notable enhancement in performance, with an accuracy of 81.17%. This underscores the synergistic effect of integrating both attentions, resulting in superior accuracy improvement compared to individual implementations.

Table 4.9: Experimental results of different layouts of the proposed SEFFNet on the ISIC 2017 dataset. All scores are in %. Bold values represent the best results.

Baseline	CBAM	TA	CBAM + TA	Acc	Pre	Rec	F1
✓	×	×	×	80.00	79.23	80.00	79.38
✓	✓	×	×	77.50	76.47	77.50	76.22
✓	×	✓	×	77.50	76.18	77.50	76.21
✓	-	-	✓	**81.17**	**80.50**	**81.17**	**80.29**

Overall, these findings underscore the importance of attention mechanisms along with the baseline model within our proposed architectural design for enhanced performance as well as facilitating informed decision-making for future development in this domain.

4.3.8 Statistical Analysis

The SEFFNet architecture incorporates tailored blocks, refined through rigorous evaluations across diverse skin cancer datasets. This section further conducts a comprehensive assessment, integrating our proposed model SEFFNet and its variations. Delving deeply into their architectural distinctions and nuanced performance evaluations is undertaken. Specifically, **M1** represents the baseline model alone, while **M2** integrates the CBAM block with the baseline model. Furthermore, **M3** incorporates the TA block alongside the baseline, and **M4** presents the entirety of the SEFFNet model with the combination of both the CBAM as well as the TA blocks.

Statistical testing, specifically the Wilcoxon signed-rank test, has been employed to ascertain the efficacy of the proposed approach **M4**. This test aims to provide robust evidence against a null hypothesis. Specifically, our hypothesis posited that "The suggested model, **M4** yields comparable outcomes compared to alternative models." By executing ten independent runs of each model on every dataset, the classification accuracy of each run has been measured. The hypothesis is rejected if the resulting p-value falls below 0.05 (5%), indicating insufficient evidence to support the claim. The statistical study comparing the performance of our proposed model **M4**, with other models **M1**, **M2**, and **M3**, across two skin cancer datasets is presented in Table 4.10.

Table 4.10: Statistical comparison of SEFFNet layout variants on two skin cancer datasets using Wilcoxon Rank-Sum test in terms of p-values.

Comparison	**HAM10000**	**ISIC 2017**
M1 vs M4	0.0020	0.0145
M2 vs M3	0.0020	0.0075
M2 vs M4	0.0020	0.0020
M3 vs M4	0.0020	0.0020

With a confidence level of 5%, the null hypothesis is confidently rejected. Consequently, it is evident that within the scope of the analyzed datasets, the proposed model **M4** consistently outperforms the other models, as indicated by notable differences across various datasets.

CHAPTER 5

Conclusion

In this thesis, experimentation was conducted in two distinct domains: vehicle make and model recognition (VMMR), and skin cancer classification. The research culminated in the development of two innovative models: SimSANet for VMMR tasks and SEFFNet for skin cancer classification. For VMMR, a novel multi-kernel sequential attention-based deep neural network, named SimSANet, has been developed for VMMR tasks. This model, designed to complement conventional CNN architectures, incorporates a key component known as the SA module. The SA module integrates both local and global features along with row and column-wise context information to emphasize the most discriminative characteristics in images. Extensive testing on various publicly available VMMR datasets demonstrates the superior performance of SimSANet compared to recent models, achieving accuracies of 94.47% on Stanford Cars, 98.34% on CompCars, 99.20% on CompcarsSV, and 97.20% on the Vehicle Images dataset. Through experimental analysis and feature map visualization, it is evident that the proposed model effectively focuses on target positions while optimizing accuracy and reducing processing resources, particularly in terms of model parameters. Additionally, SEFFNet, a snapshot ensemble-based feature fusion approach for skin cancer classification tasks, is introduced in this thesis. The primary objective is to capture diverse feature representations across training stages without the need for multiple models, thereby alleviating the computational burden. Remarkably, our approach achieves high accuracies of 92.93% on the HAM10000 and 81.17% on the ISIC 2017 datasets, underscoring its efficacy and computational efficiency in skin cancer classification tasks.

5.1 Limitations and Open challenges

Although the performance of the suggested SimSANet model for VMMR is deemed satisfactory, more efficient attention-based methods will be explored to produce results that are both efficient and require fewer computation resources, making them more feasible for use in real-world applications. It is important to acknowledge certain limitations and drawbacks:

- **Scarcity of Datasets:** There is a lack of large, diverse datasets that cover various vehicle types, colors, views, and conditions (lighting, weather). Most datasets are limited to typical conditions, and privacy concerns restrict public availability.
- **Intra-class Variance and Inter-class Similarities:** Significant challenges arise due to intra-class variance (differences within a single vehicle model) and inter-class similarities (similarities between different vehicle models). These variances and similarities complicate the accurate recognition and classification of vehicles, making it difficult to distinguish between models with high similarity or significant internal variability.
- **Recognition Rate and Precision:** Real-time VMMR requires high recognition speed and precision. Many studies do not address processing speed or model complexity, making it difficult to choose suitable models for resource-constrained environments.
- **Occlusion:** Occlusion remains a persistent issue, where parts of the vehicle may be obscured, complicating the recognition process. This problem is particularly prevalent in congested traffic scenarios, such as those seen in the Indian subcontinent, where over-congested traffic and violations of traffic rules are common.
- **Vehicle Make, Model, and Generation Recognition (VMMGR):** This area, which involves identifying not just the make and model but also the year range or generation of vehicles, is underexplored. Multi-stage approaches could be employed to first identify the make and model and then the generation, aiding applications in the automotive industry such as market trend analysis, insurance claims, and vehicle value assessment.

Conversely, while the proposed SEFFNet model for skin cancer classification approach demonstrates remarkable computational efficiency and classification accuracy, it is important to acknowledge certain limitations and drawbacks:

- **Limited Generalization:** The performance of the SEFFNet heavily relies on the training datasets used. Although it achieves better accuracies on both datasets when compared to other previous studies with comparable settings, its generalizability to other datasets or real-world situations is yet unknown. To evaluate its robustness, more testing on a variety of datasets is required.

- **Dependency on Snapshot-based Fusion:** The effectiveness of SEFFNet also heavily relies upon the snapshot-based feature fusion strategy, which captures diverse feature representations from different training stages. However, the best selection of snapshot intervals and fusion strategies might vary based on the datasets and tasks, which could lead to less-than-ideal performance in some situations.

- **Comparatively lower Accuracy on ISIC17 dataset:** On the ISIC17 dataset, the SEFFNet model does not perform well (achieved 81.17% accuracy), even if it achieves an impressive accuracy of 92.93% on the HAM10000 dataset. This suggests that more investigation is required to improve fusion procedures or explore different models in order to improve classification accuracy on this dataset.

5.2 Future Research Scopes

In the previous section, the open challenges and Limitations in VMMR were discussed. Here, potential solutions for real-time VMMR applications are provided:

- **Dealing with Small-sized Datasets:** One-shot or few-shot learning approaches can classify objects with limited training data. Specific loss functions like focal loss and class-weighted loss can address class imbalance by giving more weight to hard-to-classify and minority classes.

- **Limited Annotations:** Semi-supervised approaches can improve model accuracy by first training on annotated data and then predicting results for unannotated data.

- **Synthetic Datasets:** GAN-based models can generate new training data, though quality depends on the original data and GAN design. Collecting diverse data from different environments and times, using IR cameras for nighttime data, collaborating with government agencies, and crowdsourcing can enhance dataset diversity.

- **Multi-view and Multi-modal Datasets:** Using data from multiple views and modes can improve model robustness and accuracy. Techniques like image pre-processing, lightweight deep learning models, and optimized hardware systems can enhance performance and recognition speed.
- **Deployment of Models:** For real-time recognition on edge devices with limited computational resources, feature selection can reduce feature space dimensionality, making models more lightweight and efficient.

On the other hand, addressing the aforementioned issues with the SEFFNet model in skin cancer classification opens up opportunities for further research and advancements in hybrid learning, combining deep learning and machine learning approaches for medical image classification. Here, some potential solutions are provided:

- **Robustness Evaluation:** Extensive experiments on a variety of datasets, with differences in pathologies, demographics, and imaging circumstances, can provide insights into SEFFNet's generalization capabilities. Furthermore, for practical implementation, evaluation of its effectiveness in real-world clinical settings via cooperative research with healthcare professionals is crucial.
- **Exploring Fusion Strategies:** To improve the performance of SEFFNet, alternate strategies for the fusion of snapshots should be investigated. Furthermore, selecting a fusion technique with the use of domain-specific information may enhance classification performance even further.
- **Data Augmentation and Pre-processing Approaches:** As traditional augmentation and pre-processing techniques are not employed to mitigate computational overload, there is a need to explore novel approaches that effectively perform data augmentation and pre-processing while consuming fewer computational resources and training time.

Acknowledging the lack of representation of diverse complexions and demographics in existing skin cancer datasets, our goal is to bridge this gap by exploring datasets that specifically focus on a variety of complexions. By incorporating images representing different ethnicities and skin types, the aim is to create a more inclusive and robust learning environment. Furthermore, there is an opportunity to create a comprehensive dataset in the future, encompassing images from a wide range of demographics and complexions. This strategy will improve the model's capacity to generalize across various demographic groups and more accurately reflect the variety of the population.

List of Publications

Published

1. **S. Gayen**, S. Maity, P. K. Singh, Z. W. GEEM,R. Sarkar - Two decades of vehicle make and model recognition – Survey, challenges and future directions - Elsevier, Journal of KingSaud university- Computer and Information Sciences - vol. 36 - Year 2023 - Impact Factor: **6.9**

Accepted

1. T. Ghosh **S. Gayen**, S. Maity, Daria Valenkova R. Sarkar -A Feature Fusion based Custom Deep Learning Model for Vehicle Make and Model Recognition - IEEE, Mediterranean Conference on Embedded Computing

Under Review

1. **S. Gayen**, S. Maity, P.K. Singh, R. Sarkar - SimSANet: A Simple Sequential Attention aided Deep Neural Network for Vehicle Make and Model Recognition - Elsevier, Engineering Applications of Artificial Intelligence - Imapct Factor: **8** (In Review)

2. **S. Gayen**, S. Maity, S. Sarkar, R. Sarkar - SEFFNet: Snapshot Ensemble-based Feature Fusion Network for Skin Cancer Classification - Elsevier, Computers in Biology and Medicine - Imapct Factor: **7.7** (In Review)

Drafting Stage

1. A. Ray **S. Gayen**, R. Sarkar - EDB-Net: A Novel Edge-guided Dual-Branched Network for Skin Cancer Classification (Drafting Stage)

Bibliography

[1] A. Islam, S. Mallik, A. Roy, M. Agrebi, and P. K. Singh, "A filter-based feature selection framework for vehicle/non-vehicle classification," in *Measurements and Instrumentation for Machine Vision*, Taylor, 2023, pp. 677–684.

[2] S. Maity, A. Chakraborty, P. K. Singh, and R. Sarkar, "Performance comparison of various yolo models for vehicle detection: An experimental study," in *International Conference on Data Analytics & Management*, Springer, 2023, pp. 677–684.

[3] S. Maity, A. Bhattacharyya, P. K. Singh, M. Kumar, and R. Sarkar, "Last decade in vehicle detection and classification: A comprehensive survey," *Archives of Computational Methods in Engineering*, pp. 1–38, 2022.

[4] B. Chougula, A. Tigadi, P. Manage, and S. Kulkarni, "Road segmentation for autonomous vehicle: A review," in *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, IEEE, 2020, pp. 362–365.

[5] B. Tian, B. T. Morris, M. Tang, *et al.*, "Hierarchical and networked vehicle surveillance in its: A survey," *IEEE transactions on intelligent transportation systems*, vol. 16, no. 2, pp. 557–580, 2014.

[6] S. Gayen, S. Maity, P. K. Singh, Z. W. Geem, and R. Sarkar, "Two decades of vehicle make and model recognition–survey, challenges and future directions," *Journal of King Saud University-Computer and Information Sciences*, p. 101 885, 2023.

[7] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2018," *CA: a cancer journal for clinicians*, vol. 68, no. 1, pp. 7–30, 2018.

[8] R. Skaggs and B. Coldiron, "Skin biopsy and skin cancer treatment use in the medicare population, 1993 to 2016," *Journal of the American Academy of Dermatology*, vol. 84, no. 1, pp. 53–59, 2021.

[9] S. I. Hussain and E. Toscano, "An extensive investigation into the use of machine learning tools and deep neural networks for the recognition of skin cancer: Challenges, future directions, and a comprehensive review," *Symmetry*, vol. 16, no. 3, p. 366, 2024.

[10] L. Ferrante di Ruffano, Y. Takwoingi, J. Dinnes, *et al.*, "Computer-assisted diagnosis techniques (dermoscopy and spectroscopy-based) for diagnosing skin cancer in adults," *Cochrane Database of Systematic Reviews*, vol. 2018, no. 12, 1996.

[11] Z. Rahman, M. S. Hossain, M. R. Islam, M. M. Hasan, and R. A. Hridhee, "An approach for multiclass skin lesion classification based on ensemble learning," *Informatics in Medicine Unlocked*, vol. 25, p. 100 659, 2021.

[12] M. Hosni, I. Abnane, A. Idri, J. M. C. de Gea, and J. L. F. Alemán, "Reviewing ensemble classification methods in breast cancer," *Computer methods and programs in biomedicine*, vol. 177, pp. 89–112, 2019.

[13] Y. Xiao, J. Wu, Z. Lin, and X. Zhao, "A deep learning-based multi-model ensemble method for cancer prediction," *Computer methods and programs in biomedicine*, vol. 153, pp. 1–9, 2018.

[14] D. Chanda, M. S. H. Onim, H. Nyeem, T. B. Ovi, and S. S. Naba, "Dcensnet: A new deep convolutional ensemble network for skin cancer classification," *Biomedical Signal Processing and Control*, vol. 89, p. 105 757, 2024.

[15] N. Kausar, A. Hameed, M. Sattar, *et al.*, "Multiclass skin cancer classification using ensemble of fine-tuned deep learning models," *Applied Sciences*, vol. 11, no. 22, p. 10 593, 2021.

[16] X. Li, L. Yu, D. Chang, Z. Ma, and J. Cao, "Dual cross-entropy loss for small-sample fine-grained vehicle classification," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 5, pp. 4204–4212, 2019.

[17] K. Yang, X. Hu, and R. Stiefelhagen, "Is context-aware cnn ready for the surroundings? panoramic semantic segmentation in the wild," *IEEE Transactions on Image Processing*, vol. 30, pp. 1866–1881, 2021.

[18] P. Ajitha, A. Sivasangari, *et al.*, "Vehicle model classification using deep learning," in *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, IEEE, 2021, pp. 1544–1548.

[19] L. Lu, P. Wang, and Y. Cao, "A novel part-level feature extraction method for fine-grained vehicle recognition," *Pattern Recognition*, vol. 131, p. 108 869, 2022.

[20] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5209–5217.

[21] B. Zhao, X. Wu, J. Feng, Q. Peng, and S. Yan, "Diversified visual attention networks for fine-grained object classification," *IEEE Transactions on Multimedia*, vol. 19, no. 6, pp. 1245–1256, 2017.

[22] H. Zheng, J. Fu, Z.-J. Zha, J. Luo, and T. Mei, "Learning rich part hierarchies with progressive attention networks for fine-grained image recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 476–488, 2019.

[23] S. Kalouche, A. Ng, and J. Duchi, "Vision-based classification of skin cancer using deep learning," *2015, conducted on Stanfords Machine Learning course (CS 229) taught*, 2016.

[24] D. B. Mendes and N. C. da Silva, "Skin lesions classification using convolutional neural networks in clinical images," *arXiv preprint arXiv:1812.02316*, 2018.

[25] A. Mahbod, G. Schaefer, C. Wang, R. Ecker, and I. Ellinge, "Skin lesion classification using hybrid deep neural networks," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 1229–1233.

[26] E. Ayan and H. M. Ünver, "Data augmentation importance for classification of skin lesions via deep learning," in *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, IEEE, 2018, pp. 1–4.

[27] S. Ding, Z. Wu, Y. Zheng, *et al.*, "Deep attention branch networks for skin lesion classification," *Computer methods and programs in biomedicine*, vol. 212, p. 106 447, 2021.

[28] M. K. Hasan, S. Roy, C. Mondal, *et al.*, "Dermo-doctor: A framework for concurrent skin lesion detection and recognition using a deep convolutional neural network with end-to-end dual encoders," *Biomedical Signal Processing and Control*, vol. 68, p. 102 661, 2021.

[29] J. Zhang, Y. Xie, Y. Xia, and C. Shen, "Attention residual learning for skin lesion classification," *IEEE transactions on medical imaging*, vol. 38, no. 9, pp. 2092–2103, 2019.

[30] J. Wu, W. Hu, Y. Wen, W. Tu, and X. Liu, "Skin lesion classification using densely connected convolutional networks with attention residual learning," *Sensors*, vol. 20, no. 24, p. 7080, 2020.

[31] L. Yang, R.-Y. Zhang, L. Li, and X. Xie, "Simam: A simple, parameter-free attention module for convolutional neural networks," in *International conference on machine learning*, PMLR, 2021, pp. 11 863–11 874.

[32] S. Woo, J. Park, J. Lee, and I. S. Kweon, "CBAM: convolutional block attention module," *CoRR*, vol. abs/1807.06521, 2018.

[33] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *CoRR*, vol. abs/1709.01507, 2017.

[34] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *Proceedings of the IEEE international conference on computer vision workshops*, 2013, pp. 554–561.

[35] L. Yang, P. Luo, C. Change Loy, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3973–3981.

[36] M. Ali, M. A. Tahir, and M. N. Durrani, "Vehicle images dataset for make and model recognition," *Data in brief*, vol. 42, 2022.

[37] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, Springer, 2016, pp. 630–645.

[39] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[40] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[41] S. Xu, D. Chang, J. Xie, and Z. Ma, "Grad-cam guided channel-spatial attention module for fine-grained visual classification," in *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*, 2021, pp. 1–6. DOI: 10.1109/MLSP52302.2021.9596481.

[42] K. Xu, R. Lai, L. Gu, and Y. Li, "Multiresolution discriminative mixup network for fine-grained visual categorization," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[43] R. Ji, L. Wen, L. Zhang, *et al.*, "Attention convolutional binary neural tree for fine-grained visual categorization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 468–10 477.

[44] Y. Chen, Y. Bai, W. Zhang, and T. Mei, "Destruction and construction learning for fine-grained image recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5157–5166.

[45] W. Luo, H. Zhang, J. Li, and X.-S. Wei, "Learning semantically enhanced feature for fine-grained image classification," *IEEE Signal Processing Letters*, vol. 27, pp. 1545–1549, 2020.

[46] P. Wang, Y. Cao, and L. Lu, "A novel part feature integration and fusion method for fine-grained vehicle recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 1990–1994.

[47] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2403–2412.

[48] X. Xu, J. Mo, and M. Chen, "Application optimization of fine-grained vehicle classification based on backbone network," in *2021 27th International Conference on Mechatronics and Machine Vision in Practice (M2VIP)*, IEEE, 2021, pp. 54–59.

[49] A. Hassan, M. Ali, N. M. Durrani, and M. A. Tahir, "An empirical analysis of deep learning architectures for vehicle make and model recognition," *IEEE Access*, vol. 9, pp. 91 487–91 499, 2021.

[50] Q. Hu, H. Wang, T. Li, and C. Shen, "Deep cnns with spatially weighted pooling for fine-grained car recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 11, pp. 3147–3156, 2017.

[51] X. Dai, B. Southall, N. Trinh, and B. Matei, "Efficient fine-grained classification and part localization using one compact network," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 996–1004.

[52] Z. Ma, D. Chang, and X. Li, "Channel max pooling layer for fine-grained vehicle classification," *arXiv preprint arXiv:1902.11107*, 2019.

[53] H. Zheng, J. Fu, Z.-J. Zha, and J. Luo, "Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5012–5021.

[54] X. Ma and A. Boukerche, "An ai-based visual attention model for vehicle make and model recognition," in *2020 IEEE Symposium on computers and communications (ISCC)*, IEEE, 2020, pp. 1–6.

[55] A. Boukerche and X. Ma, "A novel smart lightweight visual attention model for fine-grained vehicle recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 13 846–13 862, 2021.

[56] M. Li, G. Zhou, W. Cai, *et al.*, "Multi-scale sparse network with cross-attention mechanism for image-based butterflies fine-grained classification," *Applied Soft Computing*, vol. 117, p. 108 419, 2022.

[57] L. Lu, Y. Cai, H. Huang, and P. Wang, "An efficient fine-grained vehicle recognition method based on part-level feature optimization," *Neurocomputing*, vol. 536, pp. 40–49, 2023.

[58] S. H. Tan, J. H. Chuah, C.-O. Chow, and J. Kanesan, "Coarse-to-fine context aggregation network for vehicle make and model recognition," *IEEE Access*, 2023.

[59] Y. Tian, W. Zhang, Q. Zhang, G. Lu, and X. Wu, "Selective multi-convolutional region feature extraction based iterative discrimination cnn for fine-grained vehicle model recognition," in *2018 24th International Conference on Pattern Recognition (ICPR)*, IEEE, 2018, pp. 3279–3284.

[60] A. Amirkhani and A. H. Barshooi, "Deepcar 5.0: Vehicle make and model recognition under challenging conditions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 1, pp. 541–553, 2022.

[61] J. Fang, Y. Zhou, Y. Yu, and S. Du, "Fine-grained vehicle model recognition using a coarse-to-fine convolutional neural network architecture," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 7, pp. 1782–1792, 2016.

[62] Q. Zhang, L. Zhuo, S. Zhang, J. Li, H. Zhang, and X. Li, "Fine-grained vehicle recognition using lightweight convolutional neural network with combined learning strategy," in *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*, IEEE, 2018, pp. 1–5.

[63] Y. Yu, Q. Jin, and C. W. Chen, "Ff-cmnet: A cnn-based model for fine-grained classification of car models based on feature fusion," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2018, pp. 1–6.

[64] R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization. arxiv 2016," *arXiv preprint arXiv:1610.02391*, 2022.

[65] P. K. Singh, R. Sarkar, and M. Nasipuri, "Significance of non-parametric statistical tests for comparison of classifiers over multiple datasets," *International Journal of Computing Science and Mathematics*, vol. 7, no. 5, pp. 410–442, 2016.

[66] P. K. Singh, R. Sarkar, and M. Nasipuri, "Statistical validation of multiple classifiers over multiple datasets in the field of pattern recognition," *International Journal of Applied Pattern Recognition*, vol. 2, no. 1, pp. 1–23, 2015.

[67] A. Hossain, A. R. Willan, and J. Beyene, "An improved method on wilcoxon rank sum test for gene selection from microarray experiments," *Communications in Statistics-Simulation and Computation*, vol. 42, no. 7, pp. 1563–1577, 2013.

[68] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer, "Densenet: Implementing efficient convnet descriptor pyramids," *arXiv preprint arXiv:1404.1869*, 2014.

[69] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[70] D. Misra, T. Nalamada, A. U. Arasanipalai, and Q. Hou, "Rotate to attend: Convolutional triplet attention module," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 3139–3148.

[71] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[72] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, no. 1, pp. 1–9, 2018.

[73] A. Yilmaz, M. Kalebasi, Y. Samoylenko, M. E. Guvenilir, and H. Uvet, "Benchmarking of lightweight deep learning architectures for skin cancer classification using isic 2017 dataset," *arXiv preprint arXiv:2110.12270*, 2021.

[74] S. Shen, M. Xu, F. Zhang, *et al.*, "A low-cost high-performance data augmentation for deep learning-based skin lesion classification," *BME frontiers*, 2022.

[75] K. Ali, Z. A. Shaikh, A. A. Khan, and A. A. Laghari, "Multiclass skin cancer classification using efficientnets–a first step towards preventing skin cancer," *Neuroscience Informatics*, vol. 2, no. 4, p. 100 034, 2022.

[76] H. L. Gururaj, N. Manju, A. Nagarjun, V. N. M. Aradhya, and F. Flammini, "Deepskin: A deep learning approach for skin cancer classification," *IEEE Access*, 2023.

[77] M. Arshad, M. A. Khan, U. Tariq, *et al.*, "A computer-aided diagnosis system using deep learning for multiclass skin lesion classification," *Computational intelligence and neuroscience*, vol. 2021, 2021.

[78] Y. Wu, A. C. Lariba, H. Chen, and H. Zhao, "Skin lesion classification based on deep convolutional neural network," in *2022 IEEE 4th international conference on power, intelligent computing and systems (ICPICS)*, IEEE, 2022, pp. 376–380.

[79] T. M. Alam, K. Shaukat, W. A. Khan, *et al.*, "An efficient deep learning-based skin cancer classifier for an imbalanced dataset," *Diagnostics*, vol. 12, no. 9, p. 2115, 2022.

[80] I. Kousis, I. Perikos, I. Hatzilygeroudis, and M. Virvou, "Deep learning methods for accurate skin cancer recognition and mobile application," *Electronics*, vol. 11, no. 9, p. 1294, 2022.

[81] L. Hoang, S.-H. Lee, E.-J. Lee, and K.-R. Kwon, "Multiclass skin lesion classification using a novel lightweight deep learning framework for smart healthcare," *Applied Sciences*, vol. 12, no. 5, p. 2677, 2022.

[82] H. K. Gajera, D. R. Nayak, and M. A. Zaveri, "A comprehensive analysis of dermoscopy images for melanoma detection via deep cnn features," *Biomedical Signal Processing and Control*, vol. 79, p. 104 186, 2023.

[83] K. Mridha, M. M. Uddin, J. Shin, S. Khadka, and M. F. Mridha, "An interpretable skin cancer classification using optimized convolutional neural network for a smart healthcare system," *IEEE Access*, 2023.

[84] M. A. Khan, K. Muhammad, M. Sharif, T. Akram, and S. Kadry, "Intelligent fusion-assisted skin lesion localization and classification for smart healthcare," *Neural Computing and Applications*, vol. 36, no. 1, pp. 37–52, 2024.

[85] A. K. Gairola, V. Kumar, A. K. Sahoo, M. Diwakar, P. Singh, and D. Garg, "Multi-feature fusion deep network for skin disease diagnosis," *Multimedia Tools and Applications*, pp. 1–26, 2024.

[86] C. Agyenta and M. Akanzawon, "Skin lesion classification based on convolutional neural network," *Journal of Applied Science and Technology Trends*, vol. 3, no. 01, pp. 21–26, 2022.

[87] B. Cassidy, C. Kendrick, A. Brodzicki, J. Jaworek-Korjakowska, and M. H. Yap, "Analysis of the isic image datasets: Usage, benchmarks and recommendations," *Medical image analysis*, vol. 75, p. 102 305, 2022.

[88] M. A. Al-Masni, D.-H. Kim, and T.-S. Kim, "Multiple skin lesions diagnostics via integrated deep convolutional networks for segmentation and classification," *Computer methods and programs in biomedicine*, vol. 190, p. 105 351, 2020.

[89] Z. Yu, F. Jiang, F. Zhou, *et al.*, "Convolutional descriptors aggregation via cross-net for skin lesion recognition," *Applied Soft Computing*, vol. 92, p. 106 281, 2020.

[90] S. Serte and H. Demirel, "Deep learning for diagnosis of covid-19 using 3d ct scans," *Computers in biology and medicine*, vol. 132, p. 104 306, 2021.

[91] J. Zhang, Y. Xie, Y. Xia, and C. Shen, "Attention residual learning for skin lesion classification," *IEEE Transactions on Medical Imaging*, vol. 38, no. 9, pp. 2092–2103, 2019. DOI: 10.1109/TMI.2019.2893944.

[92] D. Mandal, A. Ray, S. Sarkar, O. Ojjela, and R. Sarkar, "A robust deep feature fusion model for skin cancer classification," in *International Conference on Advanced Computing and Applications, ICACA-2024*, Springer, 2024.

[93] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.