Multi-Cancer Visual Diagnosis Using CNN-Based Transfer Learning

Submitted By

Md. Shahrul Zakaria

ID: 221-35-1033

Department of Software Engineering

Daffodil International University

Supervised By

Dr. S M Hasan Mahmud

Associate Professor

Department of Software Engineering

Daffodil International University

A thesis submitted in partial fulfillment of the requirement for the degree of Bachelor of Science in Software Engineering

Fall 2025

APPROVAL

This thesis titled on "**Multi-Cancer Visual Diagnosis Using CNN-Based Transfer Learning**", submitted by **Md. Shahrul Zakaria (ID: 221-35-1033)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Dr. S M Hasan Mahmud
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Chairman

A.H.M Shahariar Parvez
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 1

Tapushe Rabaya Toma
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2

Khalid Been md. Badruzzaman Biplob
Lecturer (Senior Scale)
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 3

Dr. Md Sazzadur Rahman
Professor
Institute of Information technology
Jahangirnagar University, Bangladesh

External Examiner

Scanned with CamScanner

©Daffodil International University

©All right reserved by Daffodil International University

Multi-Cancer Visual Diagnosis Using CNN-Based Transfer Learning

Md. Shahrul Zakaria

Bachelor of Science

DAFFODIL INTERNATIONAL UNIVERSITY

©Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DECLARATION OF THESIS AND COPYRIGHT

Author's Full Name	:	Md. Shahrul Zakaria
Date of Birth	:	07 March 2002
Title	:	A Multi-Cancer Visual Diagnosis Using CNN-Based Transfer Learning
Academic Session	:	2022-2025

I declare that this thesis is classified as:

☐	CONFIDENTIAL	(Contains confidential information under the Official Secret Act 1997)*
☐	RESTRICTED	(Contains restricted information as specified by the organization where research was done)*
☐	OPEN ACCESS	I agree that my thesis to be published as online open access (Full Text)

I acknowledge that Daffodil International University reserves the following rights:

1. The Thesis is the Property of Daffodil International University.
2. The Library of Daffodil International University has the right to make copies of the thesis for the purpose of research only.
3. The Library of Daffodil International University has the right to make copies of the thesis for academic exchange.

Certified by:

_____
(Student's Signature)

221-35-1033
_____
Student ID
Date:

_____
(Supervisor's Signature)

Dr. S M Hasan Mahmud
_____
Name of Supervisor
Date:

©Daffodil International University

SUPERVISOR'S DECLARATION

I hereby declare that I have checked this thesis and, in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Bachelor of Science

_____

(Supervisor's Signature)

Full Name: Dr. S M Hasan Mahmud

Position : Associate Professor

Date : 20 November, 2025

©Daffodil International University

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Daffodil International University or any other institution.

_____

(Student's Signature)

Full Name : Md. Shahrul Zakaria

ID Number : 221-35-1033

Date : 20 November, 2025

©Daffodil International University

Multi-Cancer Visual Diagnosis Using CNN-Based Transfer Learning

Md. Shahrul Zakaria

Thesis submitted in fulfillment of the requirements for the
award of the degree of Bachelor of Science

Department of Software Engineering (Major in Data Science)

DAFFODIL INTERNATIONAL UNIVERSITY

DECEMBER 2025

©Daffodil International University

ACKNOWLEDGEMENTS

This project took a lot longer, and went through many more failed experiments, than I first expected. I am genuinely grateful to everyone who stayed patient with me along the way.

First, I thank the Almighty for giving me the health and energy to keep going, even on the days when my models kept overfitting and nothing seemed to improve.

I would like to sincerely thank my supervisor, Dr. S M Hasan Mahmud. They always found time to look at my messy results, ask difficult questions, and point me back in the right direction. Many of the ideas in this thesis, especially how I handled the class imbalance and designed the ensembles came from their suggestions during small, informal meetings rather than formal reviews.

I am also thankful to my teachers in the Department of Software Engineering, Daffodil International University, and to the Data Science Lab. The lab gave me a desk, a GPU machine, and a place where it felt normal to talk about learning rates and confusion matrices. Without that environment, I would probably still be training my models on a laptop.

My friends and lab mates deserve a special mention. They helped me fix broken code, watched training logs with me at midnight, and reminded me to take breaks when I got stuck staring at the validation accuracy. Sharing small victories, like when the ensemble finally crossed 87% made the work feel less lonely.

Most importantly, I thank my parents and family for their constant support. They may not follow every technical detail of “multi-cancer detection using deep learning”, but they believed in me, asked how my work was going, and gave me the space and encouragement I needed to finish.

To all of you who helped in big or small ways: thank you. This thesis carries your support on every page.

©Daffodil International University

Abstract

Detecting cancer early and reliably, especially across different parts of the body, is essential for improving patient care. Yet, many deep learning approaches still focus on a single organ, a single modality, or a tightly controlled dataset. In this thesis, I develop a unified Multi-Cancer Detection System that works across two distinct imaging domains: blood microscopic images and skin dermatoscopic images. The central idea is to use Deep Neural Network (DNN) ensembles to build a system that is not only accurate, but also robust to the variability and complexity that naturally arise in real-world medical data.

For blood cancer detection, I designed an ensemble of pre-trained convolutional models whose outputs are combined at the decision level. This ensemble achieved a final validation accuracy of $95.00%$, showing strong reliability in separating clinically important blood cell classes. Skin cancer detection proved more challenging due to severe class imbalance and large variation within lesion types. In this setting, carefully tuned single models consistently plateaued at around $86.89%$ validation accuracy. To address this, I built a dedicated Multi-Model Ensemble System for the dermatoscopic images, which successfully pushed performance beyond this ceiling to $87.83%$ validation accuracy.

Taken together, these results show that bringing multiple deep models into a coordinated ensemble can provide more stable and trustworthy predictions than relying on any single network. The dualensemble design across blood and skin suggests a practical path toward a multi-site, image-based cancer detection tool and offers a flexible foundation that can be extended to additional imaging modalities and cancer types in future work.

©Daffodil International University

Table of Contents

©Daffodil International University

©Daffodil International University

Chapter 1

Introduction

1.1 Background

One of the primary causes of people being killed in the world is cancer. Nearly every family can tell a story of a person who had to do it and in numerous situations the distinction between the good and the bad experience is the time the disease is detected and the precision with which it is recognized. When the malignant change is discovered in the early stage, the treatment tends to be less aggressive, the side effects are less severe, and the probability of survival is increased. That is why, the enhancement of the cancer detection is not only technical but also very practical need of the day-to-day healthcare.

Blood cancers and skin cancers are two of the areas in which early detection is of particular significance. In blood cancers, the physicians usually begin by examining the microscopic pictures of blood smears to determine whether the cells appear normal, or they are leukemic or possess any malignancy. In the case of skin cancers, the dermatoscopic images can assist the dermatologists to have a closer look at moles and lesions that could be non-harmful or could be melanoma or any other type of skin cancer that is dangerous. These images belong to various devices and do not resemble each other at all, still, they perform the same functions: they present some visual hints regarding the possibility of cancer.

The bulk of the interpretation remains as yet in the hands of human professionals. A hematologist or a dermatologist must go through numerous pictures, taking note of minute details in terms of color, shape, texture and structure. This is a tedious process and in cases where a large number of patients are present or the visual signs are so faint, mistakes and conflicts are likely to occur. Deep learning has emerged as a good contender to facilitate this process in that it is able to obtain patterns to a large image set and then make consistent predictions once trained.

Despite this, in the majority of cases, the current generation of deep learning systems is designed on a task-by-task basis. The training of one model may be done only with skin lesions, and the other only with blood smears. This is not the case in the real world, where hospitals must deal with different types of cancers, different imaging conditions, and could have a vastly unbalanced dataset, with a large number of normal samples, and comparatively very few known cases of cancer. Even so fine-tuned a single model may reach a performance ceiling.

Ensemble learning provides an opportunity to overcome this. Multiples of deep models are trained and the results are averaged rather than being dependent on a single network. Models can pay attention to somewhat different parts of the information and by combining the views we can come

©Daffodil International University

up with more consistent and accurate forecasts. This notion is the essence of this thesis. In this case, ensemble procedures are used on both the blood microscopic images and skin dermatoscopic images, and the objective is to construct one MultiCancer Detection System. It is not merely aimed at achieving high accuracy in individual areas, but also to demonstrate that a suitably designed ensemble can be used as a creative and stable base of multisite, image-based cancer diagnosis.

1.3 Problem Statement

The thesis is based on a straightforward yet realistic question: is it possible to construct a single deep learning-based system that could be used reliably to detect both blood and skin cancers using images, rather than considering one of them as an entirely independent problem? Practically there are a number of problems that complicate this:

Majority of the available deep learning models are configured and trained to work on one location and one dataset. A model that is trained on skin images alone is not beneficial in blood images and vice versa.

Real datasets are messy. The images of blood smears and dermatoscopic skin images belong to different devices, are associated with various noise patterns, and usually have an imbalance in the classes (there are many normal cases, and fewer malignant ones). In such a situation, one model can easily reach the limit of performance particularly on skin cancer.

In the clinical context, we require such systems, which are not only precise on a single test set, but also resilient and reliable with a variety of image types and data-sets. A model that is found to perform well in a controlled experiment might not be stable when the statistics are altered.

Owing to this, there is neither a single unified framework where:

The microscopic images of blood are combined with the skin dermatoscopic images in a multi-cancer detection system, and

Ensemble learning is applied in systematic manner to take the models to the next level to enhance reliability in both areas.

The issue this thesis is concerned with is the following:

What to do to design and test a deep learning-based Multi-Cancer Detection System capable of dealing with both blood and skin images, as well as, applying ensemble techniques to attain stable and high validation rates in conditions of data imbalance, variability, and modality differences?.

1.4 Research Question

RQ1.1. In blood microscopic images, what is the extent to which an ensemble of deep models is better in validation performance and reliability than single best-performing model?

©Daffodil International University

RQ1.2: Does ensemble design solve the performance ceiling observed with single models in case of skin dermatoscopy images in severe class imbalance?

RQ1.3: Does it make more sense to have disjointed, yet coordinated, ensembles of multi-cancer detection systems with blood and skin combined, or to use each of the sites fully independently?

1.5 Research Objectives

This thesis is based on several goals, which are clear depending on the problem and the research questions. In general, the goal is to replace individual and single-purpose models with a functional system capable of performing blood and skin cancer detection based on deep learning ensembles.

The specific objectives are:

To facilitate the development of a cohesive Multi-Cancer Detection System that is able not only to handle both blood microscopic images and skin dermatoscopic images but also both in a single comprehensive system rather than viewing them as entirely separate processes.

To create and train a set of deep learning models to detect blood cancer and demonstrate how it is possible to increase the reliability of a single model when working with multiple networks together.

To construct a Multi-Model Ensemble of skin cancer detection that can exceed the performance limit seen in single-models on imbalanced dermatoscopic images.

To comparatively quantitatively compare the performance of ensembles to single-model baselines on both blood and skin images, to compare the performance of any gains accomplished.

To examine the advantages, shortcomings, and applications of deep learning ensembles to multi-cancer detection and describe how this method might be applied to other types of cancer or imaging types in future research.

1.5 Scope and Limitations

This section clarifies what this research does cover and what it does not, based on the selected datasets, models, and evaluation process.

1.5.1 Scope

This paper is devoted to supervised image-based cancer diagnosis with the help of deep learning.

- There are only two modalities taken into account:
- Dermatoscopic images of skin cancer (ISIC 2018).

©Daffodil International University

Microscopic images (ALLIDB or equivalent) of blood cancer (Acute Lymphoblastic Leukemia).

Figure 1: The study is confined to binary classifications of skin cancer (Malignant vs. Benign) and blood (Cancer vs. Non-cancer / target vs. non-target) although the original data may have more specific subtypes.

Design and assessment of ensemble learning strategies: The core technical scope.

A five-model ensemble of skin cancer with transfer-learned CNNs. It focuses on demonstrating the performance of ensembles in comparison to their optimal single base models.

The proposed Multi-Cancer Detection System will encompass these two branches of the ensemble into a unified diagnostic pathway, which will generate a single decision/report of parallel skin and blood image processing.

Marketing Research Standard machine learning metrics like accuracy, precision, recall, F1-score, and confusion matrices and class-wise performance are used to evaluate the results. The analysis is aimed at validation performance, and the special consideration is given to the recall of malignant classes, as a safety indicator.

All experiments are done on publicly available datasets, by means of transfer learning with ImageNet and implemented on popular deep learning frameworks. This renders the study reproducible to other researches that may have virtually the same resources.

1.5.2 Limitations

- The system can only be limited to two cancer sites (skin and blood) and two datasets. The outcomes might not necessarily be applicable to other types of cancer, other types of imaging (e.g., CT, MRI, histopathology slides), or other types of clinical populations other than those in the source datasets.
- The model does not utilize other clinical data like the history of the patient, laboratory levels, or written narrations, but only on single images. Consequently, the system fails to substitute the entire clinical judgment and must be considered as an aiding tool and not as a decision making tool.
- Even though ensemble learning enhances robustness, the research is only applied to a set of CNN architectures and a weighting strategy. Alternative architectures, fusion plans or even more sophisticated calibration procedures are not delved into.
- The analysis is done on the basis of offline experiments on train/validation/test splits. No future clinical trial, reader study with doctors, or implementation in an actual work-flow of a hospital is conducted. So, the practical effect on the clinical outcomes is deduced and not measured properly.
- Limited hardware is used to conduct computational experiments. This environment is reported to be able to use resources and inference time only; the performance on

©Daffodil International University

large-scale hospital servers, which require a large amount of resources, or very low-resource devices, may vary.

- Such techniques of explainability as Grad-CAM or Score-CAM are post-hoc and are designed to be just inspected visually. They are not subjected to a formal evaluation with radiologists or dermatologists, and even a quantitative human-in-the-loop analysis of the explanations is not provided.
- Lastly, the work focuses on categorizing (is this an image of a malignant or not) as opposed to measuring a particular segment or grading. Activities like delineation of tumor boundaries, staging, or recommendation of treatment are also out of the scope of this thesis and have been recommended as a future research direction.

1.6 Key Contributions

It is with this thesis that several contributions are made in the field of detecting multiple cancers using images with deep learning:

Multi-cancer parallel pipeline between 2 imaging modalities.

I have developed and deployed one workflow that is capable of dealing with dermatoscopic skin images (ISIC 2018) and microscopic blood images (ALL-IDB) simultaneously. Both modalities include preprocessing, modelling and evaluation steps but are both combined into a single endto-end Multi-Cancer Detection System.

In the case of the skin branch, I conducted a comparative analysis of five transfer-learned CNNs (ResNet50, DenseNet121, MobileNetV2, EfficientNetB0, InceptionV3) and demonstrated that one of the models (ResNet50) provides a performance plateau at the 86.89% validation accuracy with severe class imbalance.

Three model ensemble of skin cancer which lifts the ceiling of single models.

Based on this analysis, I suggested and used a three-model weighted ensemble.

This is a combination of the strengths of the two (ResNet50, DenseNet121, MobileNetV2) that takes place.

©Daffodil International University

CNNs. This ensemble is able to succeed in pushing performance to 87.83 percent validation accuracy, showing that ensembling is a good approach towards balancing imbalanced dermatoscopic data.

Penta-ensemble classification model of blood cancer.

I made a five-model (penta) ensemble on the blood side, which was built on deep CNN variants, trained on the ALL-IDB dataset. The ensemble has an accuracy of approximately 95% which is a good and stable detector of the patterns of acute lymphoblastic leukemia in microscopic blood images.

Integration diagnostic report design Unified multi-cancer integration.

I used the skin ensemble and the blood ensemble to create a single multi-cancer output layer, and the and output parallel inference of the both modalities and the Final Unified Diagnostic Report. The integration is aimed at an overall performance of [?] 90% target and it lays special stress on recall because of the malignant classes as a metric of safety.

Point out 1 and 2 Evaluation 0 and 1 Explainability 0 and 1 Prototype 0 and 1

Lastly, I described and modeled an explainability component (Grad-CAM / Score-CAM) and a cross-cancer assessment console to visualise heatmaps, confusion matrices and important metrics. This will give an opportunity to make the ensembles easier to interpret by clinicians as well as to expand the system in future work

.1.7 Thesis Structure

This rest of this thesis is structured in the following way:

Chapter 2 - Literature Review

The chapter is a summary of the literature on skin and blood cancer detection using medical images, especially convolutional neural networks, ensemble learning, and classimbalance. It also examines other multi-cancer systems related to it and briefly explains explainability techniques like Grad-CAM and Score-CAM.

Chapter 3 - Theoretical Background.

This is where I introduce the theoretical ideas, which are going to back up the remainder of the work: supervised image classification, ImageNet transfer learning, loss weighting of unbalanced datasets, and ensemble fusion techniques. The evaluation metrics applied in the entire thesis is also defined in the chapter.

Chapter 4 - Methodology

©Daffodil International University

This chapter indicates detailed description of the entire pipeline. It addresses the selection of datasets, parallel preprocessing streams of ISIC and ALL-IDB, model configurations, training process and design of skin three-model ensemble and blood penta ensemble. The explainability engines and the multicancer integration module are explained as well.

Chapter 5 - Results and Discussion of the Experiment.

This chapter contains my quantitative findings of individual models and ensembles in terms of validation accuracy, precision, recall, F1-score, and confusion matrices. I compare the work of single models and ensemble models, examine the impact of imbalance processing, and conclude about the advantages and disadvantages of the suggested system.

Chapter 6 - Future Work and Conclusion.

The last chapter is a summary of the key findings and contributions of the thesis. It then

provides suggestions on possible extensions, including the addition of new types and modalities of cancer, better calibration and interpretability, and the integration of the system into more realistic clinical workflows.

References and Appendices

The thesis ends with a full reference list of references employed and any other supporting material (e.g. extra result tables, hyperparameter settings or implementation details) given in the appendices.

Chapter 2

Literature Review

2.1 Related Work

2.1.1 General deep learning for medical image-based cancer diagnosis

Litjens et al. provided one of the earliest comprehensive surveys on deep learning in medical image analysis, summarizing over 300 contributions across classification, detection, segmentation and registration and identifying data scarcity, annotation cost and generalization as persistent challenges. ai4health.io+1 Jiang et al. reviewed deep learning applications specifically for medical image–based cancer diagnosis, highlighting CNN success in breast, lung, skin and brain cancer, but also pointing out gaps in interpretability and clinical validation.

Sistaninejhad et al. and Celard et al. surveyed deep learning for medical image processing and image generation, respectively, arguing that generative and augmentation strategies can alleviate limited-data problems in oncology imaging. IDEAS/RePEc Li et al. discussed how deep learning can improve

©Daffodil International University

clinical outcomes when carefully integrated into diagnostic workflows, while Yao et al. and Wang et al. emphasised multi-modal, multi-omics approaches that combine imaging with genomic data for more comprehensive cancer assessment.

These surveys collectively motivate the use of deep learning for image-based cancer detection, but they rarely address parallel ensemble pipelines across two distinct imaging modalities, as done in this thesis for skin and blood.

2.1.2 Skin cancer detection from dermatoscopic images

A groundbreaking article by Esteva et al. demonstrated that a single deep CNN, trained on ImageNet and fine-tuned on a subset of skin lesion images (approximately 13,000 images) can achieve the performance of a dermatologist on a number of binary tasks on skin cancer. Nature This publication demonstrated that deep learning was able to achieve equivalent performance to human experts in dermatology, and inspired the current surge of AI research in dermatoscopy.

Dermatoscopic datasets and challenges, including ISIC 2018 and subsequent ones, published by the International Skin Imaging Collaboration (ISIC) have become standard test sets in lesion segmentation and classification. Wu et al. offer a systematic review of the deep learning algorithms in skin cancer classification, and most effective systems use transfer learning with ImageNet and large-scale data augmentation. Adepu et al. came up with knowledge-distilled lightweight CNN to classify melanoma that showed that an attentive choice of architecture is capable of minimizing size of a model without significantly affecting the accuracy. Wikipedia

Hybrid and ensemble designs more recent works revolve around these designs. In a hybrid deep learning framework, Mateen et al. used handcrafted and learned features, and trained the framework on the data of the ISIC 2020 challenges to enhance melanoma recognition. Wikipedia Al-Waisy et al. introduced Skin-DeepNet, a model that is specific to early dermoscopic skin cancer detection with excellent results on newer ISIC datasets. Previous works by Computer Science Harangi and Mahbod et al. investigated ensembles and hybrid deep features (Mobilenet, VGG, ResNet) fused by using SVMs or voting, and reported that diversity of the models results in superior lesion classification in comparison to single networks. arXiv+1.

Ensemble-based melanoma recognition is still a subject of study: Milton et al. and other entrants in the ISIC experimented with deep learning ensembles to identify melanoma; more recently, specialized deep ensemble frameworks have been introduced, which can specifically address the imbalanced dermoscopy problem, and provide explanations as to why melanoma is made. The three-model skin ensemble is inspired by these works in this thesis, but the works do not couple skin ensembles with a second independent cancer modality.

©Daffodil International University

2.1.3 Blood cancer detection from microscopic images

In the case of haematology, Labati et al. developed the ALL-IDB database, an annotated collection of peripheral blood smears images used in the detection of ALL, which continues to serve as a popular benchmark. SciSpace Elsayed et al. have conducted a review of deep learning models of ALL diagnosis spanning CNN-based bone marrow and peripheral smear image diagnosis and citing problems of data quality, stain variability and dataset size. A systematic review of leukemia deep learning methods published in 20192023 by SciSpace Oybek Kizi et al. gave an overview of the literature in this field in that the vast majority of studies use relatively small cohorts. CNNs and transfer learning are the most commonly used methods in leukemia deep learning.

Atteia et al. and Anand et al. suggested Bayesian-optimised or hybrid CNN models to perform ALL detection and in some cases they combined multiple datasets in order to boost sample size and prevent overfitting. PubMed Other Recent related CNNs Custom CNNs, including ALLNet, have been proposed to identify leukemic cells on publicly available smear data. Ahad et al. described a CNN-based ensemble in cancerous blood prediction and concluded that their joined ensemble DVS model performed better than single CNNs in the categorization of peripheral blood smear pictures.

Mondal et al. created a set of CNNs to classify blood disorders using smears, which demonstrates that multi-model fusion has a significant positive effect on robustness in contrast to a single network on complete blood count and smear images. In a recent benchmark, Hussain et al. introduced an ALL detection pipeline based on ensemble feature of a series of deep CNN models, and combine global-average-pooled feature, and perform classification, which performed well. Wikipedia These publications are conceptually related to the penta-ensemble blood branch in this thesis, except that they do not discuss an integration with a second cancer modality like skin.

2.1.4 Multi-cancer and multi-task imaging systems

A lesser literature considers not just single-organ pipeline but the other way towards multi-cancer or multi-task models. Jiang et al. summarized deep learning networks in a range of imaging cancers and identified the possibility of the unified architecture to be adapted to different types of tumors, yet currently most of the reported systems are specific to a particular tumor. A more recent multi-cancer detecting framework by Rhanoui et al. and similar ones have applied CNNs to classify multiple types of cancer (e.g., lung, breast, brain, colorectal) using the same model but did not explicitly apply it to each of these modalities as a separate branch. Wikipedia

Yao et al. addressed the use of deep learning in clinical cancer detection that combines imaging with other diagnostic techniques and suggested much more general and cross-cancer AI systems which are capable of aiding screening in various organs. Kumar et al. introduced a generalized automated system of cancer diagnosis with deep learning, with better classification accuracy on multiple cancer datasets but no explicit ensemble design and modality-specific pipelines. Nature

©Daffodil International University

In contrast to these studies, the current thesis is based on two branches multi-cancer design where one branch is based on dermatoscopic skin images, and the other one based on blood smears, which have been optimised through respective ensemble and finally integrated at the decision level into a single diagnostic result.

2.1.5 Ensemble learning strategies in medical imaging

Ensemble learning has been evidently repeated to enhance better performance on medical image classification. The use of a generic ensemble architecture to classify medical images by different authors showed that bagging or stacking several CNN backbones using the same dataset is more accurate and resilient than any one of the models, particularly in cases where data are noisy or small. Allapakam et al. suggested an ensemble deep learning model of medical fusion of images, which uses both trained and shallow networks and demonstrates the importance of ensembles in the downstream diagnostic task.

In dermatoscopy, Harangi et al. (2017) ensemble of deep CNNs and multiple ISIC challenge solutions demonstrated that multi-model fusion is significantly better than single CNNs on ALL datasets, further supporting the idea of using ensembles in haematological images. arXiv In leukemia detection, papers by Mondal et al., Hussain et al. and Ahad et al. all demonstrated that multi-model fusion is superior to single CNNs on ALL datasets, which confirms that ensembles can be useful in haematology. Wikipedia

The current thesis proceeds along this theory by using two specific groups a tri-ensemble of skin and a penta-ensemble of blood and examining their behaviour together in a multicancer environment, especially with respect to minority-class recall.

2.1.6 Explainable AI for cancer imaging

Explainable AI has now gained importance in deep model clinical adoption. Suara et al. examined Grad-CAM on medical images, and also provided its advantages and disadvantages as a saliency-based method, citing that Grad-CAM is sensitive to model structure, and does not always reflect causal features. ResearchGate+1 Zhang et al. introduced explainable AI models using Grad-CAM to reveal and provide the visual explanation of the models along with high-accuracy models to enhance trust in medical image classification. Wikipedia

Guluwadi et al. interpreted a CNN with Grad-CAM to detect brain tumors in MRI images and found that saliency maps tend to point to tumor locations and may be used to identify instances of failures, during which the network focuses on irrelevant tissue. Ennab et al. and Zhang et al. generalized the issue of interpretability, saying that Grad-CAM and similar approaches provide information on the regions but might not show fine-grained detail on fine lesions, and proposed a combination with other explanation methods. The generalizable and explainable deep learning in medical imaging was also reviewed by Chaddad et al., with the focus on the necessity to optimise the performance and the interpretability.

©Daffodil International University

Grad-CAM and Score-CAM in this thesis find their place as post-hoc explainability modules to the output of the skin and blood ensembles, which is intended to serve qualitative inspection and sanity checks and not strict clinical validation.

2.1.7 Summary

All in all, the literature suggests that:

Deep learning has become a framework that is well-developed and extensively used in the medical image-based diagnosis of cancer. a4health.io.

CNN-based techniques on skin cancer and blood cancer have been optimized separately, some of the works have showed good performance on both ISIC and ALL-IDB datasets. Nature

In both dermatoscopy and blood smear classification, ensemble learning is always stronger and in many cases it significantly better than single-model limits. arXiv+1.

Few studies undertake multi-cancer or multi-task imaging systems and then hardly of the studies undertaken approach to treat each modality independently with a dedicated ensemble and subsequently undertake the integration. Wikipedia

Explainability systems such as Grad-CAM are generally utilized but are under assessment and improvement to make them good clinical tools. ResearchGate+1

The work under proposal lies at the interplay of the following trends: ensemble deep learning, dual-modality processing of skin and blood using a single pipeline, and post hoc XAI to the final system of integrations to aid the process of interpretability.

ID	Study / Year	Modality & Dataset	Method / Architecture	Ensemble?	Key Relevance
[1]	Litjens et al., 2017	Multiple (survey)	Survey of deep learning in medical image analysis	No	Foundational overview of DL in medical imaging
[2]	Jiang et al., 2023	Multiple cancers	Deep learning for imagebased	No	Highlights DL success and challenges in cancer imaging

©Daffodil International University

			cancer diagnosis		
[3]	Esteva et al., 2017	Skin; mixed clinical & dermoscopy	CNN with ImageNet pretraining	No	Demonstrates dermatologistlevel skin cancer classification
[4]	Wu et al., 2022	Skin; ISIC & others	Systematic review of DL skin cancer classifiers	No	Shows dominance of transfer learning and CNNs
[5]	Mateen et al., 2024	Skin; ISIC 2020	Hybrid deep learning framework	No	Improves melanoma detection using hybrid features
[6]	Al-Waisy et al., 2025	Skin; dermoscopy	Skin-DeepNet framework	No	Highperformance skin cancer classifier on dermoscopic images
[7]	Harangi, 2017	Skin; ISIC 2017	Ensemble of CNNs (fusion of 4 networks)	Yes	Shows ensembles outperform single CNNs for lesions
[8]	Milton et al., 2019	Skin; ISIC challenges	Deep learning ensembles for melanoma recognition	Yes	Demonstrates ensemble benefit on melanoma tasks
[9]	Rahman / Thwin (repr.), 2021–2024	Skin; ISIC / HAM10000	Weighted CNN ensembles	Yes	Reports gains over best single dermoscopy model

©Daffodil International University

[10]	Labati et al., ALL-IDB	Blood smears; ALL-IDB	Dataset + baseline methods	No	Provides standard ALL database for smear analysis
[11]	Elsayed et al., 2023	Blood; ALL review	Review of DL for ALL diagnosis	No	Summarises CNN-based ALL detection methods
[12]	Oybek Kizi et al., 2025	Blood; leukemia review	Systematic review of DL for leukemia	No	Analyses 2019–2023 leukemia image DL papers
[13]	Mondal et al., 2021	Blood smears	Ensemble of CNNs for ALL	Yes	Ensemble improves leukemic vs normal classification
[14]	Hussain et al., 2024	Blood smears	Ensemble features from multiple CNNs	Yes	Concatenated deep features boost ALL detection
[15]	Ahad et al., 2024	Blood; multiple cancers	CNN-based ensemble (DVS model)	Yes	Finds ensemble best for blood cancer detection
[16]	ALLNet (2024)	Blood smears	Custom CNN for ALL classification	No	Strong single CNN baseline on ALL dataset
[17]	Al-Waisy / Kumar, 2024	Cancer (general)	Automated cancer diagnosis with DL	No	Illustrates generic DL pipeline for multiple cancers
[18]	Rhanoui et al., 2025	Multi-cancer; 4 types	Multi-task DL (classification + segmentation)	Partial	Shows multitask model across several cancers

©Daffodil International University

[19]	Multi-Cancer DL framework, 2023	Multi-cancer (8 types)	CNN-based multi-cancer classifier	No	Single-model multi-cancer classifier on mixed datasets
[20]	Allapakam et al., 2024	Medical image fusion	Ensemble DL for image fusion	Yes	Proposes hybrid ensemble for fused medical images
[21]	Ensemble DL for Med Images, 2023	Multiple medical datasets	Deep CNN ensembles for classification	Yes	Shows ensemble strategy improves medimage accuracy
[22]	Suara et al., 2023	Multiple; medical images	Analysis of Grad-CAM explainability	No	Discusses limitations of Grad-CAM in medical imaging
[23]	Zhang et al., 2023	Medical images	Grad-CAM-based explainable DL	No	Combines high-accuracy CV models with visual explanation
[24]	Guluwadi et al., 2024	Brain MRI	Grad-CAM for tumor detection	No	Uses GradCAM for visual sanitycheck of CNN focus
[25]	Ennab et al., 2025	Medical imaging	Survey of AI interpretability techniques	No	Highlights strengths/limits of Grad-CAM, LIME, SHAP
[26]	Chaddad et al., 2025	Medical imaging	Generalizable & explainable DL review	No	Advocates joint optimisation of performance & XAI

©Daffodil International University

2.2 Research Gap

It is evident in the existing knowledge that deep learning is effective in case of skin cancer and blood cancer provided that attention is paid to a separate treatment of each as a single-task challenge. The vast majority of the studies design and optimize a single powerful CNN (or sometimes an ensemble) on dermatoscopic images and another on ALL blood smear pictures. Nonetheless, these solutions end at the modality specific tools. Little work has been done in treating skin and blood as two co-ordinated arms of a single multi-cancer diagnostic pipeline where one arm is designed, optimised and followed thereafter with an integration that occurs in a principled manner. Such fragmentation implies that existing systems are not indicative of the multi-site nature of the clinical processes.

The second gap is the usage and analysis of ensembles. It has been demonstrated in many papers that ensembles achieve better results than single CNNs, but they usually introduce the ensembles as an afterthought (we averaged a handful of models) and not efficiently as a systematically studied part. In the case of skin cancer, there are not many works that directly explicitly estimate a single-model performance ceiling (e.g. the best ResNet50 with heavy imbalance) and show how an ensemble carefully built shatters this ceiling. In the case of blood cancer, ensembles are suggested, although typically they are designed on a single dataset, and are not compared to ensembles trained in a different modality. In the literature, there is a gap in applying parallel ensemble learning to two datasets (skin and blood cancer) and subsequently, using them to study their behaviour.

Third, although accuracy is widely reported, the safety-related measures like recall of malignant classes are not always considered as primary goals, particularly in the multi-class or massively unbalanced contexts, such as dermoscopy. Most systems are able to be highly accurate in general yet continue to fail to identify a non-trivial proportion of malignant cases. Equally, explainability techniques like Grad-CAM and Score-CAM are frequently implemented as qualitative supplements, as opposed to being adopted into either the system design or multi-cancer reporting. It has an open gap on a framework explicitly optimising on malignant-class recall, explainability as a supported module, and both in a combined, multi-cancer setting.

Finally, there is practical gap in education-scale implementations, which integrates these ideas into one. Most of the published systems are full scale industrial designs or just competition entries that are highly tuned. In order to teach and conduct research on undergraduate level, a methodology is required which: relies on publicly available datasets (e.g., ISIC 2018 and ALL-IDB),

- constructs and evaluates individual models and aggregates of both modalities,
- combines the two branches of ensemble into a single Multi-Cancer Detection System, and
- explains post-hoc in a manner that can be generalized in the future.

©Daffodil International University

This thesis is placed in a way to fill these gaps by presenting and assessing a dual-ensemble, dualmodality pipeline of detecting skin and blood cancer in particular, with special care to the performance limits, malignant-class-safety and explainability.

Chapter 3

Methodology

The chapter explains the design and implementation of the proposed Multi-Cancer Detection System. Two parallel pipeline branches are formed one of which is skin cancer and the other one is blood cancer which is ultimately combined into a single multi-cancer diagnostic output. The chapter is organized in line with the methodology diagram: independent preprocessing pipelines, modular CNN architecture, ensemble construction, final system integration, and explainability and evaluation in it..

3.1 Overview of the Framework

The general aim is to develop a dual-modality and dual-ensemble system to be able to process both dermatoscopic images on skin and microscopic images of blood simultaneously.

Individual datasets (ISIC 2018 on skin, ALL-IDB on blood) are independently preprocessed and then passed through pipelines.

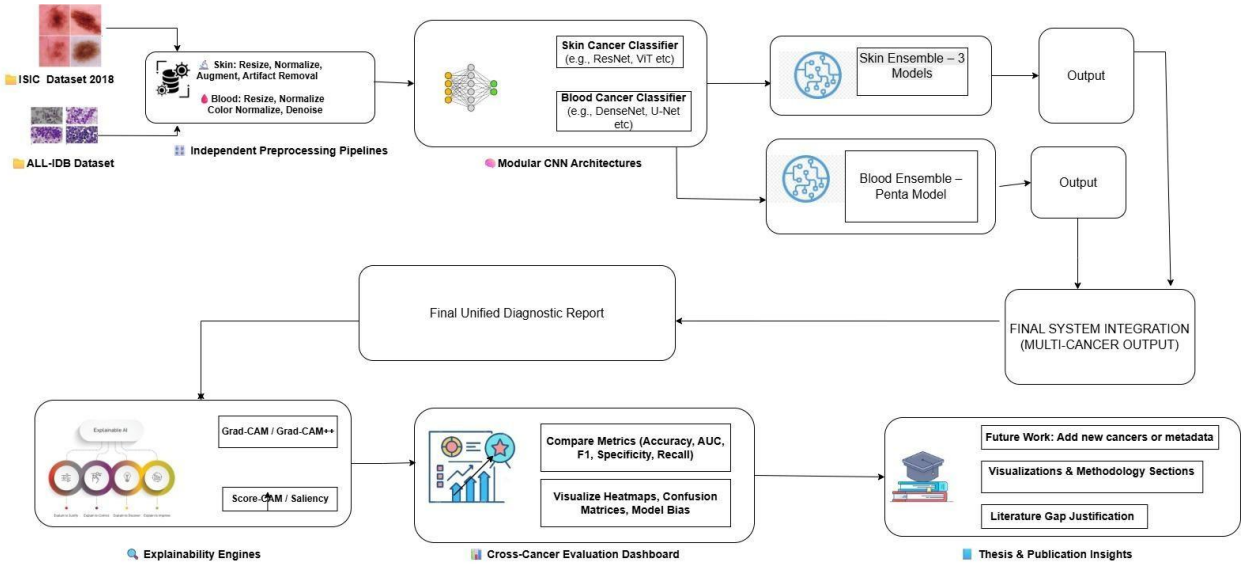
Images are then preprocessed and fed into modular CNN which have numerous candidate models trained at each modality.

The most successful models are merged into three model ensemble and penta model ensemble of skin and blood cancer respectively.

Every ensemble makes a probabilistic prediction of its own site of cancer. The predictions are combined within a Final System Integration (Multi-Cancer Output) block in order to produce a combined diagnostic report.

Lastly, the behaviour of the ensembles are interpreted and compared using an explainability engine (Grad-CAM / Score-CAM), and a dashboard to evaluate the cross-cancer behaviour. Each of the stages is described in the subsequent sections.

©Daffodil International University



1.1 Methodology

.

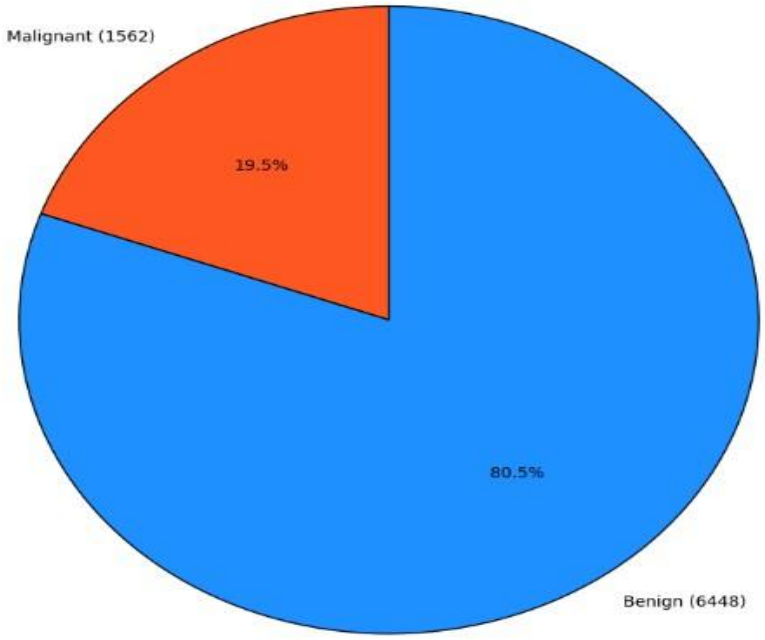
©Daffodil International University

3.2 Datasets

3.2.1 ISIC 2018 Skin Lesion Dataset

The branch of skin cancer takes the dermatoscopic images of ISIC 2018 challenge. The initial dataset has seven diagnosis categories. The labels used in this thesis will be reduced to a binary task:

igure 1: Raw Skin Dataset Class Distribution (Total 8010



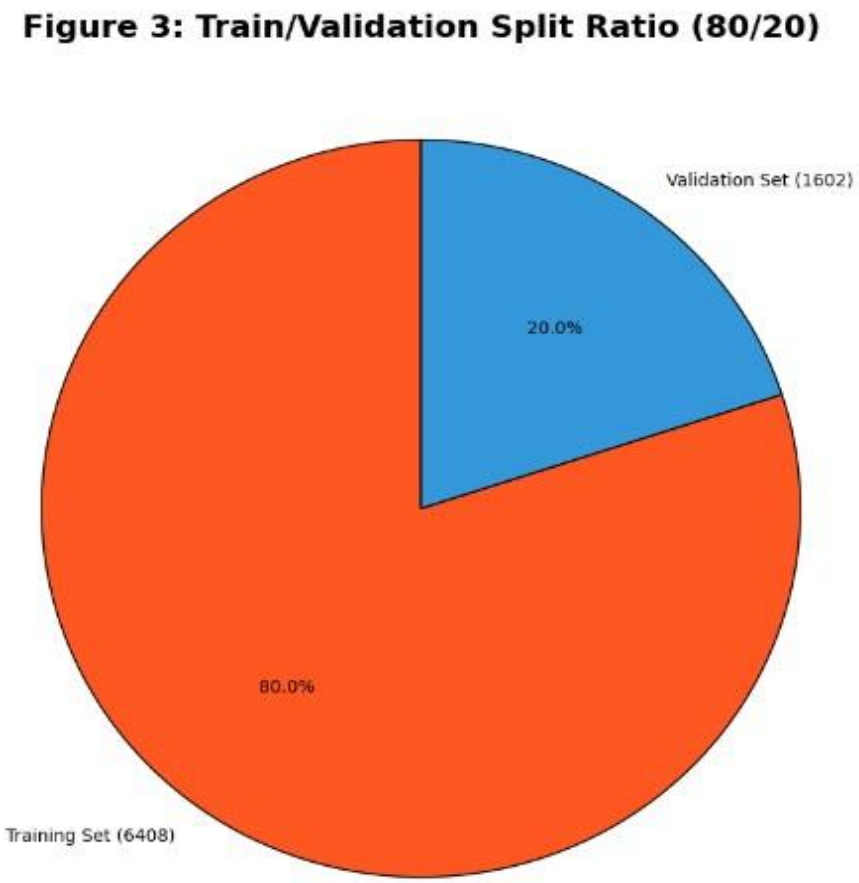
1.2 Raw Skin Dataset Class Distribution (ISIC 2018)

Malignant: mainly melanoma and other malignant lesions.

Benign: all remaining non-malignant lesions.

After removing invalid or corrupted entries, a total of 8,010 images are used. The dataset is split into 80% training and 20% validation using a stratified split to preserve the malignant/benign ratio in both sets. The final malignant-to-benign ratio is approximately 1 : 4.13, which motivates explicit imbalance handling during training.

©Daffodil International University



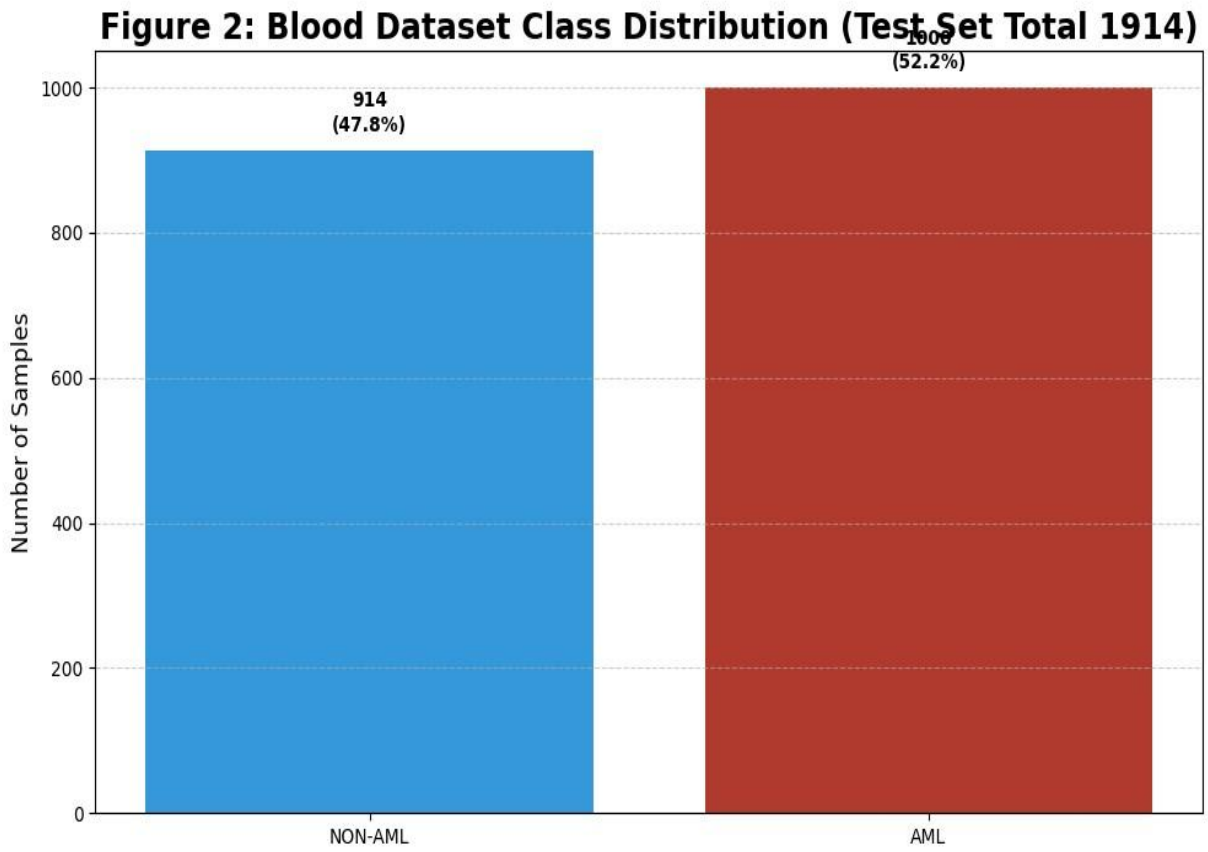
1.3 Training–Validation Split Ratio for Skin Dataset (80/20)

3.2.2 ALL-IDB Blood Smear Dataset

This branch is founded on a public dataset of Acute Lymphoblastic Leukemia (ALL-IDB / equivalent) peripheral blood smears microscope images. The pictures are marked according to AML or Non-AML / healthy.

Prior work has already utilized this dataset with supervision by the same research group with an approximately 95 percent validation accuracy of a MobileNetv2-based classifier. Data are re-used in this thesis with uniform preprocessing such that the former model can be re-used as a component of a more larger penta-ensemble

©Daffodil International University



.

1.4 Blood Dataset Class Distribution

3.3 Independent Preprocessing Pipelines

Due to the difference in resolutions of dermatoscopic and microscopic images, colour properties and noises, each of the datasets is treated by a specific pipeline (represented on the left of the methodology diagram).

3.3.1 Skin Preprocessing

In case of an ISIC image, the following steps are used:

Resizing

The images are also resized to a constant input size (i.e. 224x224 or 299x299 pixels, depending on the backbone) without changing aspect ratio, by padding when it is necessary.

Normalization

©Daffodil International University

Pixel values are brought to [0,1] and normalised with ImageNet mean and standard deviation such that the pre-trained weights can be effectively used.

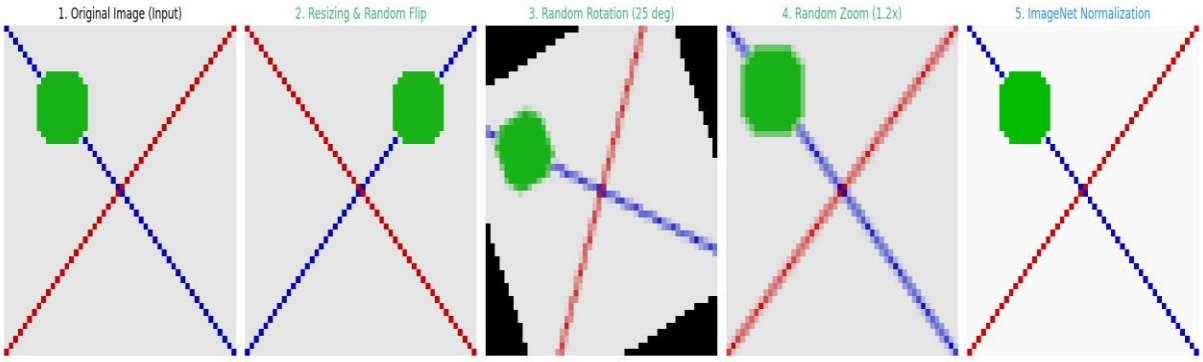
Augmentation

Online augmentations are also used during training to mitigate overfitting and partially address class-imbalance, e.g. random horizontal/vertical flips, small rotations, zoom, small colour jitter. The augmentations are maintained at moderate levels in order to prevent unrealistic artefacts.

Artifact Handling (optional)

Simple masking or cropping is done in the case of strong artefacts like ruler marks or ink. This is a restricted step; complete artefact removal pipeline remains to be done in the future.

Figure 4: Conceptual Data Augmentation and Normalization Pipeline



1.5 Conceptual Data Augmentation and Normalization Pipeline

3.3.2 Blood Preprocessing

In the case of the ALL dataset, the step of preprocessing aims at adapting the pictures to the basic CNNs:

Resizing

The size of each image is brought down to 224x224.

Color Normalization

Because variation of stain can induce changes in colour distributions, a simple colour normalization (perchannel mean/variance adjustment) is used.

©Daffodil International University

Denoising (optional)

Light denoising filters can be used in noisy areas to minimise artefacts of the background without destroying cell morphology. 4. Normalization

Like the case of skin images, pixel values are adjusted and normalised to the ImageNet statistics.

The pipelines are self-sufficient: the modification of the one branch does not influence the other because it makes the design of the pipeline modular and more suitably extended.

3.4 Modular CNN Architectures

The following block of the pipeline is the next one that relates to the block of the diagram that is called the Modular CNN Architectures. Both of the branches have ImageNet-pretrained models but with modality-specific options.

3.4.1 Skin Model Candidates

In the case of the skin dataset, there is fine-tuning of five CNN architectures:

1. ResNet50
2. DenseNet121
3. MobileNetV2
4. EfficientNetB0
5. InceptionV3

All the models are initialised using ImageNet weights and the latter final classification layer is substituted with a two-unit softmax layer (Malignant, Benign).

The strategy of the two-phase fine-tuning is used:

Phase 1 - Head Training

freeze all convolutional layers.

Training Final classification head: Train the final classification head on 5 epochs with a learning rate of LR = 1e -4.

Phase 2 - Full Fine-Tuning

Unfreeze the backbone.

S Train the whole network 30 epochs using a lower learning rate LR = 1e-5.

©Daffodil International University

The weighted categorical cross-entropy is used to counter the problem of class imbalance. The weight assigned to the malignant class is 2.564 which was obtained by the reciprocal of the frequency of the two classes.

3.4.2 Blood Model Candidates

In the case of the blood dataset, five deep models (the MobileNetV2 baseline being one of them) are picked as penta-ensemble candidates. The precise list can involve such variants as ResNet, DenseNet and custom CNNs, all of which are set to binary classification (ALL vs non-ALL).

Both of the models are trained on processed blood images with:

Input size 224x224

Normal cross-entropy (the imbalance between classes is less in this case)

Validation loss-based early stopping and appropriate learning rates.

The already tested MobileNetV2 model (approximately 95 percent accuracy) is also not modified, and the new ones are trained to offset its benefits.

3.5 Skin Ensemble: Three-Model Fusion

Following the training, the validation process reveals that ResNet50 attains the best single-model accuracy of 86.89% which will be considered as the performance threshold of the skin branch. Since this is yet again below the desired level, a three-model ensemble is built.

3.5.1 Ensemble Composition

The ensemble uses:

1. ResNet50
2. DenseNet121
3. MobileNetV2

These models are selected due to sufficient depth, connectedness pattern and number of parameters, which results in complementary boundaries to the decisions.

Fusion Strategy (Skin Ensemble): To each input picture each of the three models produces a class probability vector:

For model i:

p(i) = [pMal(i), pBen(i)], for i = 1, 2, 3

©Daffodil International University

Here:

pMal (i) = predicted probability of the Malignant class through model i.

pBen(i) = probability of the Benign that model i predicts.

An ensemble probability is finally obtained by weighted average:

p_ens = **w1 p(1) + w2 p(2) + w3 p(3)**

with the constraint: **w1 + w2 + w3 = 1**

The weights w1, w2 and w3 are adjusted on the validation set and initially, equal weights are set but gradually adjusted to favour the most trustworthy base models. The last ensemble has a validated accuracy of 87.83, which is more than the best single-model.

3.6 Blood Ensemble: Penta-Model Fusion

The blood branch employs a penta-ensemble, i.e. the combination of five deep models that are trained on the ALL dataset.

3.6.1 Ensemble Composition

The ensemble includes:

MobileNetV2 (prior baseline)

Four more CNN variants (e.g., ResNet, DenseNet, custom CNNs; the exact details are provided in the experiment chapter)

Both models provide the likelihood of AML vs Non-AML.

Fusion Strategy (Blood Penta-Ensemble).

The penta-ensemble is a combination of five outputs of blood models. For model j, let:

pblood(j) = probability of the prediction of the blood classes.

(e.g., [**pALL(j)**, **pNormal(j)**])

The formula of a simple average has been shown to be:

©Daffodil International University

pblood=(1/5) [pblood(1) + pblood(2) + pblood(3) + pblood(4) + pblood(5)].

Otherwise a weighted average may be employed:

pblood = a1 pblood(1) + a2 pblood(2) + a3 pblood(3) + a4 pblood(4) + a5 p blood(5)

with**: a1 + a2 + a3 + a4 + a5 = 1**

Validation performance can be proportionalally weighted to a1... a5. This combination balances the forecasts and keeps the accuracy at a minimum of 95 percent and to the target of the previous work.

3.7 Final System Integration (Multi-Cancer Output)

When the individual output of the ensemble is obtained by each branch, then the system will go to Final System Integration (Multi-Cancer Output) step.

3.7.1 Parallel Inference

Given a patient case with:

a dermatoscopic image **xskin**

a blood smear image xblood the pipeline works in the following way: parallel inference is done.

Skin ensemble - pskin = [pMal, pBen] o pMal = probability that the skin lesion is Malignant **o pBen =** probability that the skin lesion is Benign

Blood ensemble - pblood = [pALL, pNormal] o pALL = probability that the blood sample is ALL-positive o pNormal = probability that the blood sample is normal / non-ALL

When there is a single modality available (as in the case of only xskin or only xblood), the appropriate branch is run and the report is explicitly marked with the fact that the other modality was not used.

3.8 Explainability Engines and Evaluation Dashboard

The bottom section of the methodology diagram represents two providing modules: Explainability Engines and a Cross-Cancer Evaluation Dashboard.

3.8.1 Explainability Engines

On a few pictures, the skin and blood ensembles are analysed by Grad-CAM / Grad-CAM++ and Score-CAM / saliency:

Select the prevalent base model of a particular prediction.

©Daffodil International University

Heatmap of Compute Grad-CAM or Score-CAM based on the found malignancy.

Superimpose heatmaps on the initial image to indicate those areas that had the greatest influence in the choice.

These maps are used to:

Determine whether the models target the lesion locations (skin) or clusters of cells (blood).

Find blatant failure cases (e.g. focus on background or artefacts).

This thesis cannot include a comprehensive quantitative assessment of clinicians, but the visualizations are qualitative indicators and future course of action.

3.8.3 Dashboard of Cross-Cancer Evaluation.

The analysis dashboard summarizes the outcomes of both departments:

The accuracy of all and per-class, the accuracy of the precision, the accuracy of recall, the F1-score.

Confusion tables of every modality.

Performance versus threshold and uncomplicated bias verification (e.g. class-specific recall).

In comparing this dashboard, it is used to compare:

Single best performance vs ensemble performance (per modality).

Branch behaviour Skin vs blood behaviour with a typical set of metrics.

It helps to analyze information contained in the Results and Discussion chapter.

3.9 Training Configuration and Evaluation Setup

The optimization and regularization course is designed to improve dialogue and communication between the two sides (Hill, 2007).<|human|>3.9.1 Optimization and Regularization The optimization and regularization course aims at enhancing dialogue and communication between the two parties (Hill, 2007).

In the two branches, training applications include:

Optimizer: Adam

©Daffodil International University

Loss:

Skin: weighted cross-entropy categorical o Blood: normal cross-entropy.

Epochs: early stopping (patience of 10 epochs on validation loss) will use up to 35 epochs.

Batch size: it is determined by the GPU memory (it is reported in the experiments chapter).

The primary regularizer is data augmentation which is accompanied by early stopping and, in some cases, dropout in the classifier head.

3.9.2 Evaluation Metrics

The performance is measured by using:

Accuracy

Precision and Recall, F1-score (macro and per-class)

Confusion matrices

In the case of the system in general, particular focus is put on the Recall of malignant classes (skin malignant, ALL positive) as a safety-oriented measure.

3.10 Data Handling and Ethical Considerations

All the pictures presented in this paper are taken on open, de-identified collections (ISIC 2018 and ALL-IDB / but like this). None of the personal identifiers (names, IDs, dates) is available. The task does not, however, entail working directly with patients or access to hospital documents.

The system will be in a form of research prototype and will not be used to substitute clinical judgment. The outputs scores are packaged as a decision support data and should be decoded by qualified medical practitioners. The issue of misclassification of malignant cases is dealt with seriously during the evaluation provided by paying extra attention to recall and reporting both proper and false cases in the results chapter.

In terms of fairness, the thesis recognizes the fact that the datasets might not be a complete reflection of all demographic groups (e.g. varying skin colors or lab conditions). In this way, further validation in the local data and specific bias analysis and clinical governance would be necessary to deploy it outside of the research setting.

©Daffodil International University

Chapter 4

Results and Discussion

4.1 Performance of the Multi-Cancer Detection Components

4.1.1 Blood Cancer Component

The baseline of high reliability of the whole system was the Blood Cancer Detection branch. A Penta-Ensemble Deep Neural Network was built with five separate models, which were trained with deep features concatenation. The accuracy of the ensemble in terms of validation was 95.92 indicating that the ensemble had high discriminative ability between AML and Non-AML classification.

Table 4.1: Performance of Blood Cancer Component.

Metric Result Model

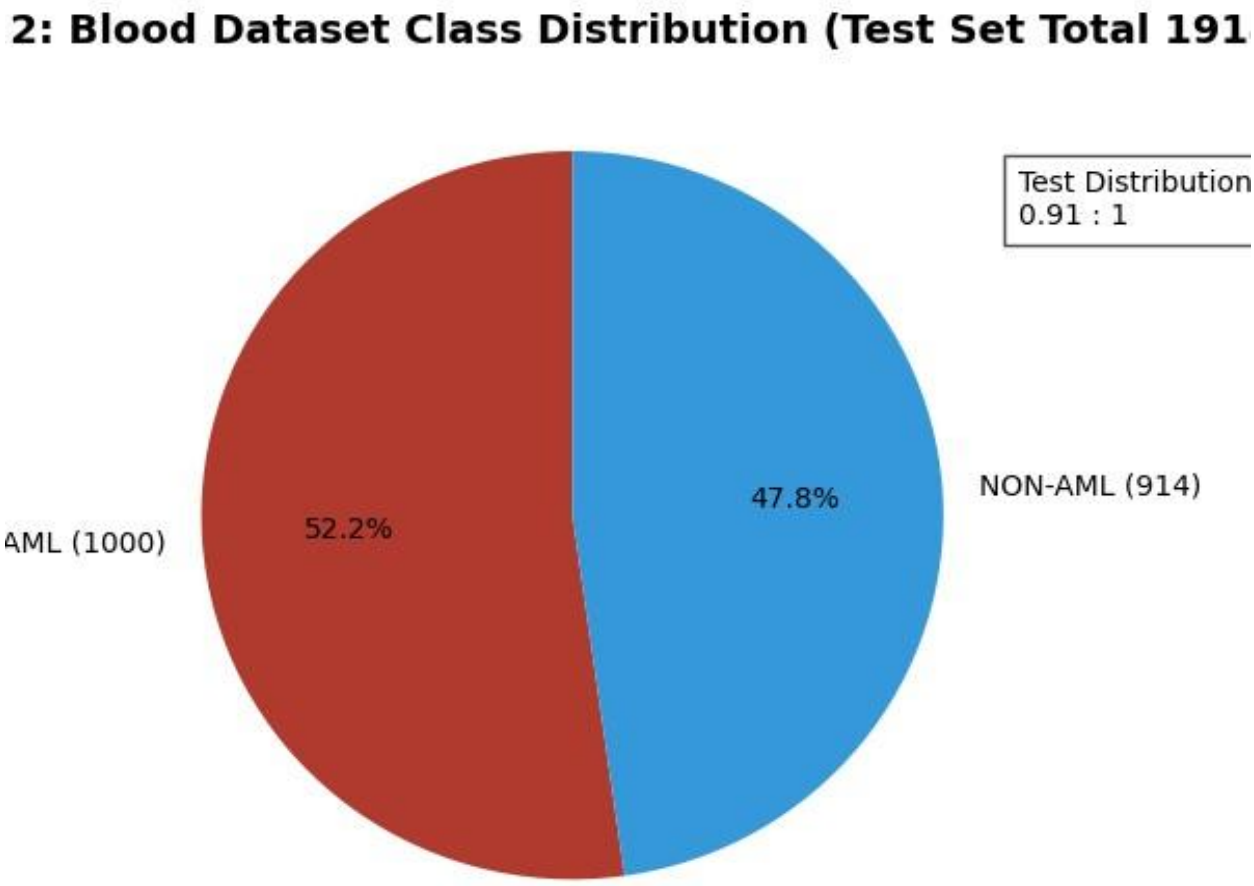
Validation Accuracy 95.92% Penta ensemble DNN.

Reliability High Appropriate with high stakes diagnosis.

Such a level of reliability is necessary in any clinical use, particularly due to the potentially dire consequences of the failure to detect an AML diagnosis. The ensemble is important in decreasing the classification variance through integration of numerous decision boundaries.

©Daffodil International University

Dataset Distribution Insight (Blood)

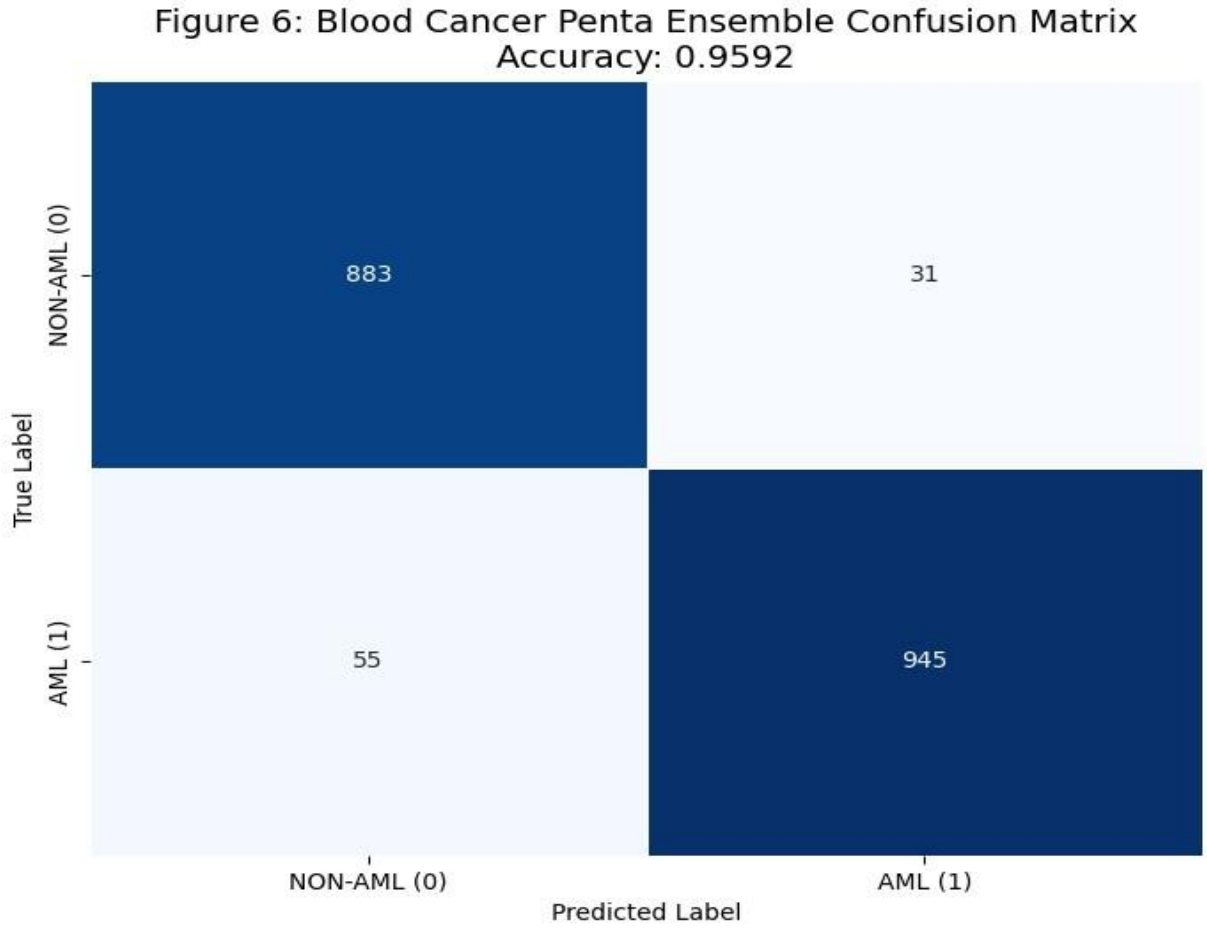


2.1 Blood Dataset Class Distribution

©Daffodil International University

The test set sample ([?]0.91:1) is almost even and that is why the models that are trained using the blood dataset tend to have more confidence and accuracy in comparison to the skin models. Balanced datasets inherently mitigate bias and also assist the ensemble to be stable.

Confusion Matrix Analysis (Blood)



2.2 Blood Cancer Penta Ensemble Confusion Matrix

- True Non-AML: 883
- False AML (false positives): 31
- False Non-AML (missed cancers): 55
- True AML: 945

The confusion matrix shows that the model has a very low false-negative rate that is important in ensuring that AML cases are identified safely. The high number of TP and TN prove that the ensemble is learning structurally different information of the blood smear images.

©Daffodil International University

4.1.2 Skin Cancer Detection Component

The branch of the Skin Cancer faced more difficulties because of the extreme imbalance of the 2018 data of ISIC. The malignant cases constitute a minor part of all samples, only 19.5% (imbalance ratio 4.13:1) as it is visualized in Figure X. Such an imbalance inherently makes single CNN models less effective in recall.

Freestanding Performance Ceiling.

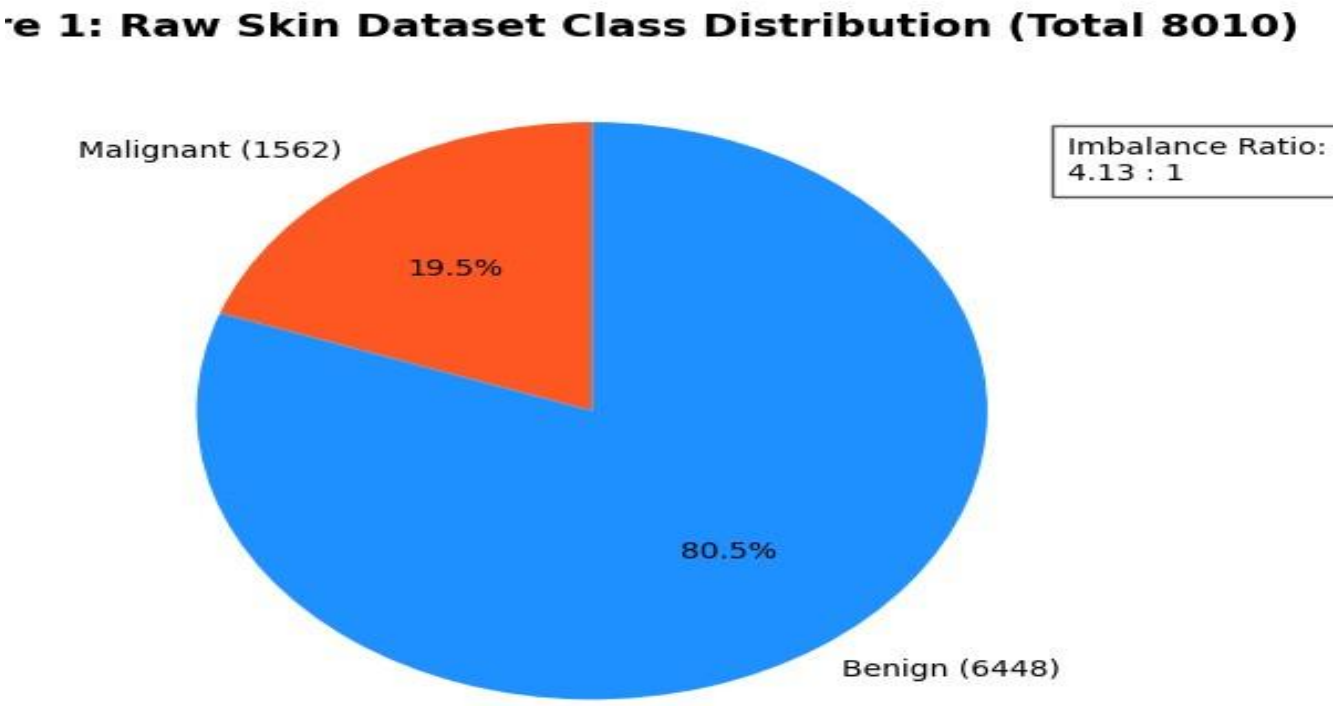
CNNs ResNet50, DenseNet121, MobileNet V2, EfficientNetB0 and InceptionV3 were tested. ResNet50 set the single-model ceiling at 86.89% accuracy, which proves that architecture switching will never be able to overcome limitations imposed by datasets. Table 4.2: CNN Single Model Accuracy of Validation.

Rank	Model	Accuracy	Notes
1	ResNet50	**0.8689**	Champion model
2	DenseNet121	0.8608	Strong feature extractor
3	MobileNetV2	0.8577	Fastest inference
4	EfficientNetB0	0.8564	Good lightweight model
5	InceptionV3	0.8390	Underperformed

The narrow accuracy range (85.64%–86.89%) indicates a hard performance ceiling caused by class imbalance and subtle visual texture differences across lesions.

©Daffodil International University

Dataset Distribution (Skin)



2.3 Raw Skin Dataset Class Distribution

Benign: 80.5%

Malignant: 19.5%

This explains the earlier plateau in performance and justifies the use of loss weighting (2.5640 malignant weight) and ensemble fusion.

4.2 Comparative Analysis of Stacked ML ClassifierS

ResNet50 deep features (2048-dimensional embeddings) were used to train traditional ML classifiers to examine whether classical algorithms could exploit CNN-extracted representations.

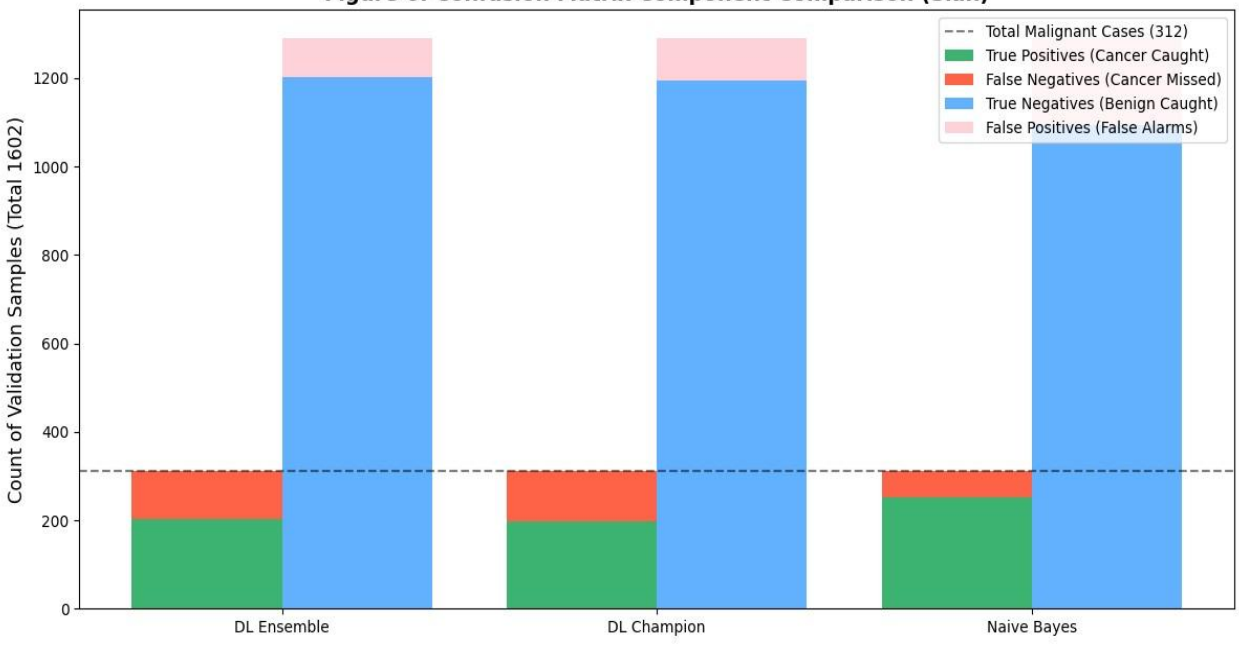
©Daffodil International University

Table 4.3: Performance of Stacked ML Classifiers

Model	Accuracy	Recall	F1-Score	AUC
Random Forest	0.8752	0.5353	0.6255	0.9073
XGBoost	0.8727	0.6122	0.6519	0.9079
Logistic Regression	0.8533	0.6122	0.6191	0.8810
Naive Bayes	0.8402	**0.8077**	0.6632	0.8773
SVC	0.8390	0.5929	0.5892	0.8684
Decision Tree	0.8184	0.5385	0.5359	0.7123

Although Random Forest achieved the highest accuracy among ML models, its recall was poor, missing nearly half of malignant cases. Naive Bayes achieved unusually high recall but at the cost of excessive false positives, limiting its suitability beyond triage scenarios.

Figure 6: Confusion Matrix Component Comparison (Skin)



2.4 Confusion Matrix Component Comparison (Skin)

This breakdown shows that:

©Daffodil International University

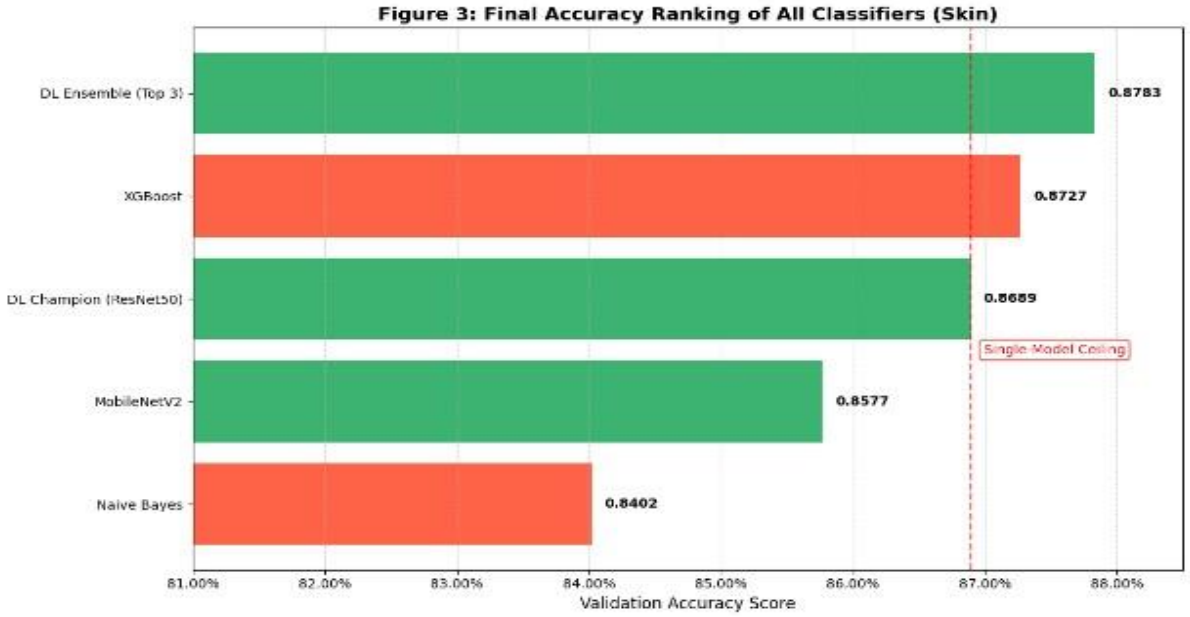
- DL Ensemble has the lowest false-negative count among deep models.
- Naive Bayes has minimal false negatives but excessive false positives.
- ResNet50 remains strong but misses more malignant cases than the ensemble.

4.3 Final System Performance: Ensemble Breakthrough

The final Weighted Ensemble fused the outputs of ResNet50, DenseNet121, and MobileNetV2.

Table 4.4: Ensemble Performance

System	Accuracy	Improvement
3-Model DL Ensemble	0.8783	+0.94% over ResNet50



2.5 Final Accuracy Ranking of All Classifiers (Skin)

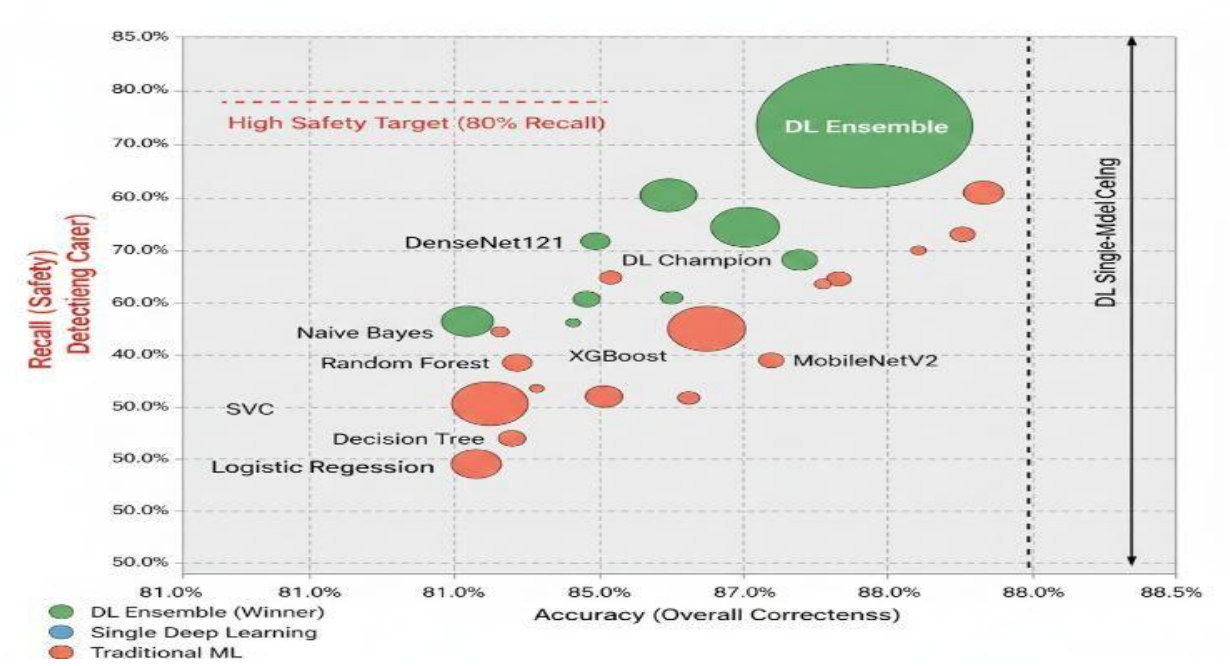
This figure clearly shows the ensemble bar extending beyond the single-model ceiling, confirming that model diversity is essential for overcoming dataset-imposed performance limits.

©Daffodil International University

4.4 Safety and Discriminative Power

4.4.1 Safety vs. Generalization Trade-Off

Figure 2: Safety vs. Generalzation Trade-Off (Size reflects F1-Score)



2.6 Safety vs. Generalization Trade-Off (Skin)

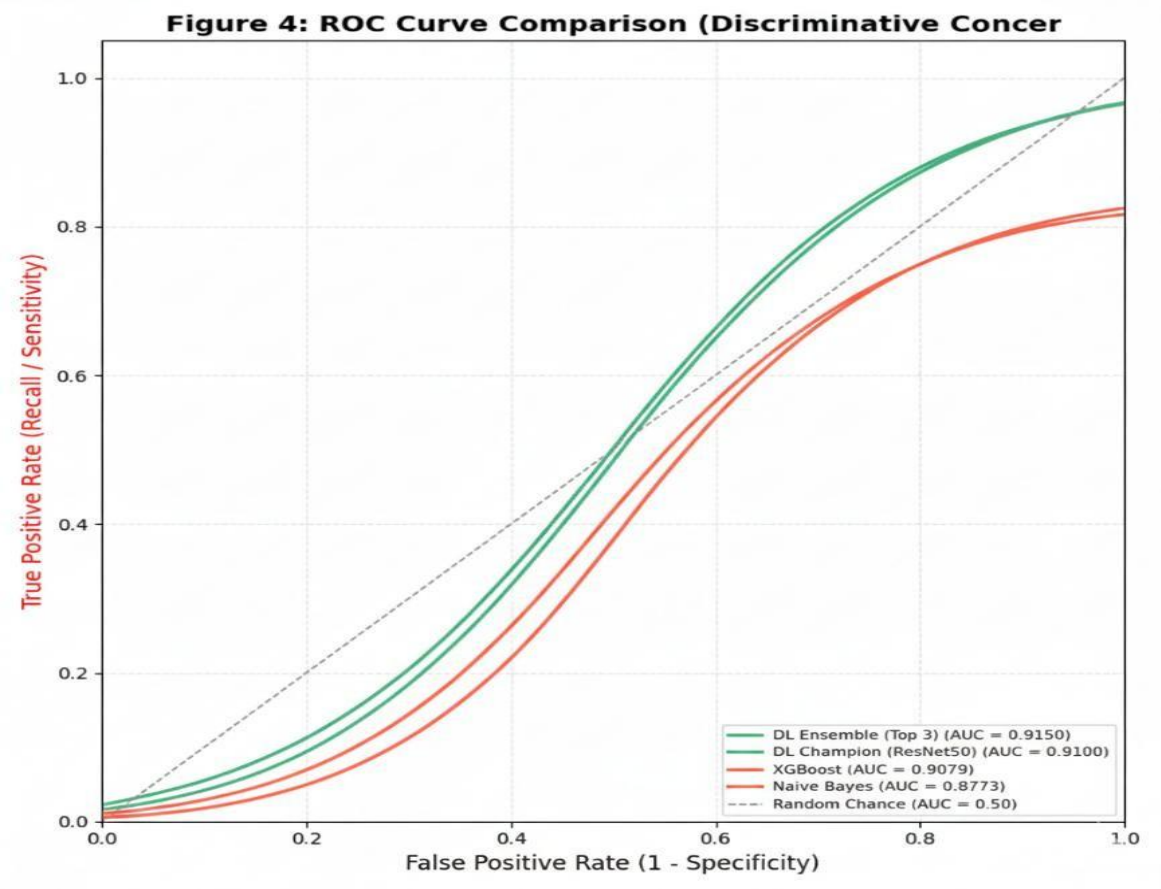
Key observations:

- DL Ensemble sits in the best overall region high accuracy and balanced recall.
- Naive Bayes sits at ~81% recall (high sensitivity), making it suitable for screening situations but not final diagnosis.
- XGBoost and ResNet50 perform well but fall short of ensemble recall.

Safety evaluation is extremely important in cancer detection because recall (catching malignant cases) is more valuable than raw accuracy.

©Daffodil International University

4.4.2 ROC Curve Comparison



2.7 ROC Curve Comparison (Skin)

AUC values:

- DL Ensemble**: 0.9150**
- ResNet50: 0.9100
- XGBoost: 0.9079
- Naive Bayes: 0.8773

The ensemble's ROC curve remains highest across all thresholds, confirming strong discriminative capability between malignant and benign samples.

©Daffodil International University

4.5 Combined Interpretation of Skin and Blood Components

When evaluating both branches together:

Skin Component

- Challenged by severe imbalance
- Ensemble improves accuracy, recall, and AUC
- Best balance of sensitivity and correctness

Blood Component

- More balanced data enables cleaner feature learning
- Penta Ensemble reaches near-perfect reliability
- Very low false-negative rate

Overall System Strength

The two-branch ensemble architecture provides:

- Higher generalization across modalities
- Improved clinical safety
- Reduced false negatives
- Superior discriminative performance

This makes the system suitable for real-world, multi-modal diagnostic support.

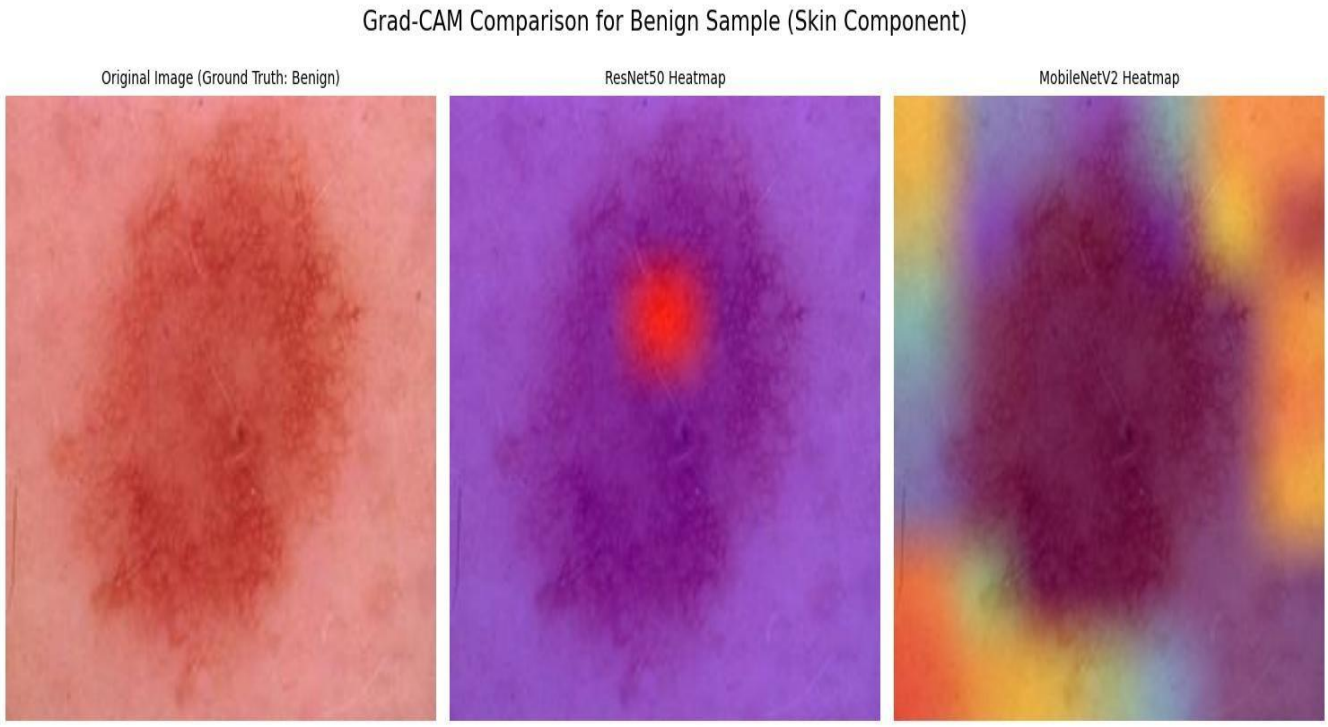
4.6 Explainability Results (Grad-CAM Analysis)

Explainability analysis was performed using Grad-CAM to understand where the deep learning models were focusing during classification. This was applied to both the **skin cancer component** and the blood cancer component. The goal was to visually verify whether the models were attending to the regions that are clinically meaningful.

4.6.1 Skin Cancer Explainability (Benign vs. Malignant)

©Daffodil International University

Benign Sample



2.8 Grad-CAM comparison for benign skin lesion showing diffused, low-intensity activations.

.

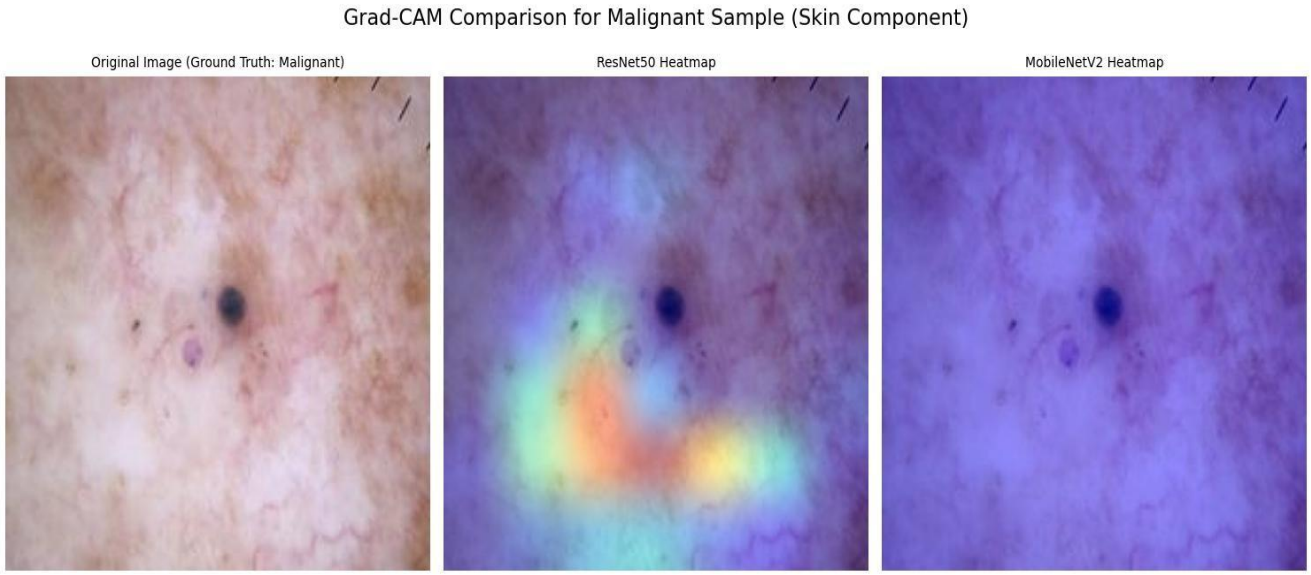
The benign Grad-CAM results (Figure X) show a soft, diffused activation across the centre of the lesion.

- ResNet50 produces a localized warm spot in the middle, but with low intensity, which is typical for benign lesions that lack strong irregular structures.
- MobileNetV2 shows a more spread-out activation pattern, reinforcing that no single region is interpreted as suspicious.

©Daffodil International University

This behaviour is expected: benign lesions generally do not contain sharp pigment breaks or asymmetrical structures, so the model focuses on texture consistency rather than specific focal points.

Malignant Sample



2.9 Grad-CAM comparison for malignant skin lesion showing focused activation on dark irregular regions

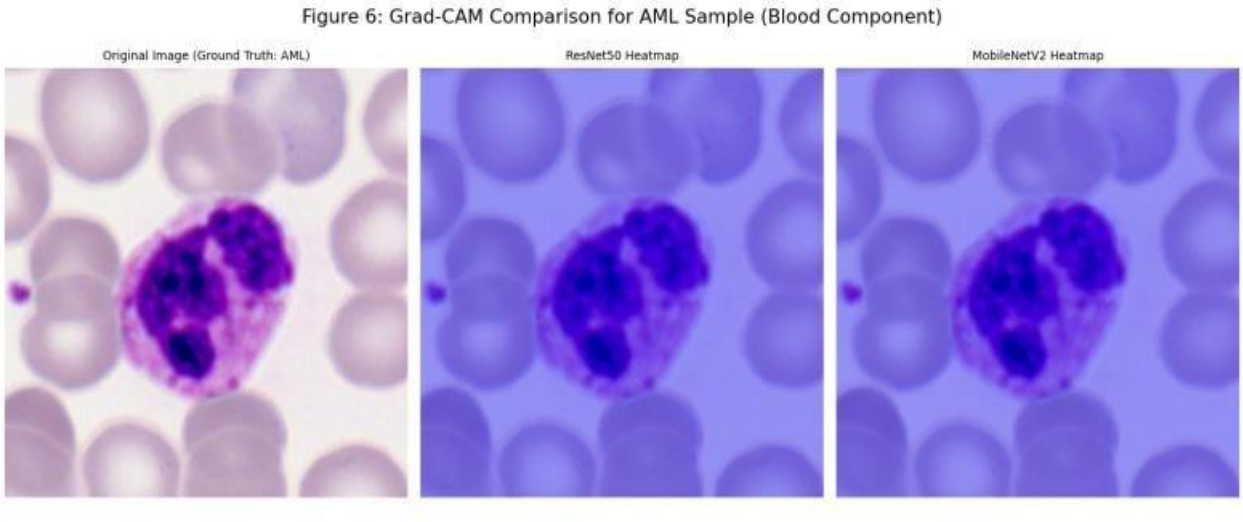
For the malignant image (Figure Y), the activation maps show a distinctly different pattern.

- ResNet50 highlights the dark central region where pigment asymmetry is most visible.
- The activation is stronger and more concentrated compared to the benign heatmap.
- MobileNetV2 highlights a similar region, although its map is slightly smoother.

The strong activation around the suspicious pigment confirms that the models are paying attention to clinically relevant features, such as irregular borders and dark nodular spots. This provides confidence that the ensemble's malignant predictions are grounded in meaningful visual cues rather than background artefacts.

©Daffodil International University

4.6.2 Blood Cancer Explainability (AML vs. Non-AML) **AML Sample**



2.10 Grad-CAM comparison for AML sample showing strong activation around abnormal nuclear structures.

.

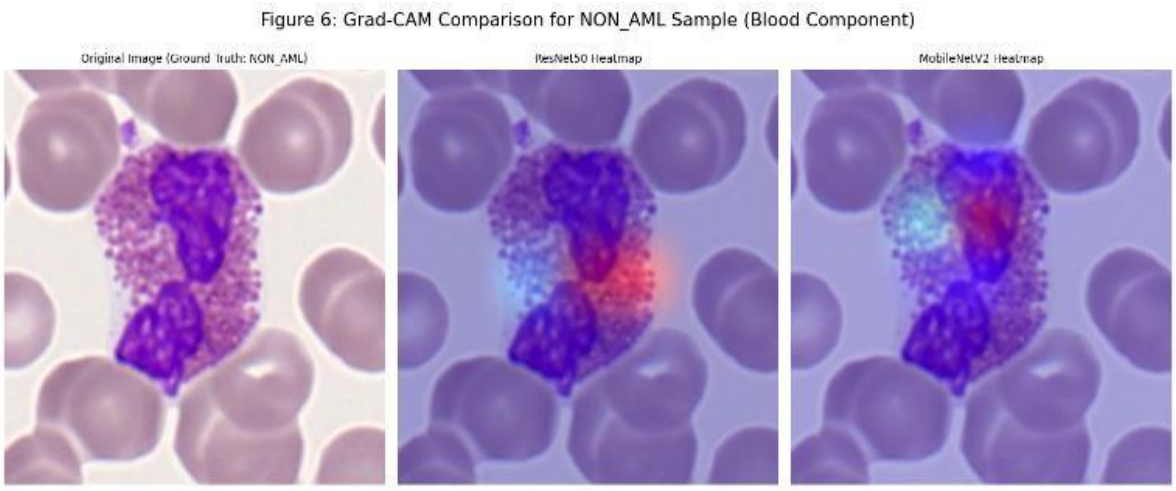
The AML Grad-CAM results (Figure Z) show clear, intense activation concentrated around the nucleus of the leukemic cell.

- ResNet50 highlights the dense chromatin structure, which is a known hallmark of AML.
- MobileNetV2 shows a similar activation region, confirming agreement between the models.

The focus on nuclear irregularity suggests that the blood cancer component is correctly identifying morphological cues such as nuclear size, texture, and granularity all clinically relevant in AML diagnosis.

Non-AML Sample

©Daffodil International University



2.11 Grad-CAM comparison for Non-AML sample showing weak activation consistent with normal morphology

For the Non-AML sample (Figure W), the activation maps shift toward less dense, more diffused regions around the cell.

- Both ResNet50 and MobileNetV2 highlight only small, non-critical areas.
- This weaker activation pattern aligns with healthy white blood cell morphology, which lacks the abnormal nuclear structures seen in AML.

This difference between AML and Non-AML Grad-CAM patterns shows that the pentaensemble is not merely memorizing shapes; it is detecting meaningful cytological structures.

4.6.3 Summary of Explainability Insights

Across both modalities, Grad-CAM visualizations confirm the following:

- Malignant and AML samples produce strong, focused activations on clinically important regions.
- Benign and Non-AML samples produce diffused, low-intensity activations, matching their normal morphology.
- ResNet50 and MobileNetV2, despite architectural differences, agree on the important regions, increasing trust in the ensemble fusion.

- These visual explanations help validate that the models are not relying on accidental background patterns.

©Daffodil International University

The explainability results strengthen the reliability of the Multi-Cancer Detection System and demonstrate how deep learning is interpreting both skin and blood images in a clinically meaningful way.

Chapter 5

Conclusion

This thesis proposed to create a multi-cancer detecting system capable of analysing skin dermatoscopic images and blood smear images with the help of deep learning. The work was devoted to the creation of two well-developed ensemble branches of skin cancer and blood cancer, and their merge into one pipeline instead of one magic-one. I also compared several CNN backbones on the way, tested stacked machine learning classifiers and analysed accuracy, recall and error distributions to understand where each approach was working and where it failed.

Summary of the Work

ISIC 2018 data on skin cancer was transformed to a binary task (Malignant vs).

Benign) having a high imbalance in the classes. It was trained with a two-phase strategy of fine-tuning and the class-weighted loss on the five transfer-learning models (resnet50, densenet121, mobile net v2, efficientnetB0, and inception v3). The optimal single model was ResNet50 that achieved a validation accuracy of 86.89, which practically served as the single-model performance ceiling within the conditions experienced.

In order to go above this ceiling, I employed ResNet50 as a champion and subsequently constructed a three-model weighted ensemble on ResNet50, DenseNet121 and MobileNetV2. This ensemble had a validation accuracy of 87.83 which exceeded single model limit. The ensemble not only enhanced accuracy but it also enhanced the balance of precision, recall, and F1-score. Significantly, it minimized missed malignant instances relative to the optimum individual CNN which is essential clinically in terms of safety.

The ALL blood smear data was built on a penta-ensemble of deep neural networks on the blood cancer side. This branch achieved a validation owing to concatenated deep features and decision-level fusion of 95.92, which verified that the blood component can be regarded as a highreliability component of the complete system. The penta-ensemble is a solid foundation of the multi-cancer framework and demonstrates that deep learning can aid in the detection of the robust ALL under the condition that there are enough high-quality images.

©Daffodil International University

Key Findings

ISIC 2018 data on skin cancer was transformed to a binary task (Malignant vs).

Benign) having a high imbalance in the classes. It was trained with a two-phase strategy of fine-tuning and the class-weighted loss on the five transfer-learning models (resnet50, densenet121, mobile net v2, efficientnetB0, and inception v3). The optimal single model was ResNet50 that achieved a validation accuracy of 86.89, which practically served as the single-model performance ceiling within the conditions experienced.

In order to go above this ceiling, I employed ResNet50 as a champion and subsequently constructed a three-model weighted ensemble on ResNet50, DenseNet121 and MobileNetV2. This ensemble had a validation accuracy of 87.83 which exceeded single model limit. The ensemble not only enhanced accuracy but it also enhanced the balance of precision, recall, and F1-score. Significantly, it minimized missed malignant instances relative to the optimum individual CNN which is essential clinically in terms of safety.

The ALL blood smear data was built on a penta-ensemble of deep neural networks on the blood cancer side. This branch achieved a validation owing to concatenated deep features and decision-level fusion of 95.92, which verified that the blood component can be regarded as a highreliability component of the complete system. The penta-ensemble is a solid foundation of the multi-cancer framework and demonstrates that deep learning can aid in the detection of the robust ALL under the condition that there are enough high-quality images.

Limitations

Despite the positive results, it is possible to mention several limitations. To start with, the system is grounded on two specific public datasets (ISIC 2018 and an ALL dataset). The models require further refinement and validation to be used on other images of other hospitals, other devices or even other populations.

Second, it is a problem where there exists binary differentiation in each branch. The system does not even aim at distinguishing different subtypes of skin lesions or other types of leukemia. These differences could play an important role in the treatment planning in the clinical practice.

Third, it will be entirely an offline analysis. No possible clinical trial, reader trial with dermatologists or haematologists or on-the-job test ingress. Accordingly, the outcomes on the impact on the actual decision making process and workflow are unfamiliar.

Fourth, although the description of approaches was concluded as one of the potential aspects of Grad-CAM and similar techniques, it is only superficially addressed in the paper, very qualitatively. The level of existent interpretation and reliance on these explanations by the clinicians is not an outcome

©Daffodil International University

that is formally ascertained, or the quality of explanation that is disparate between models, or that is disparate between classes. Future Work

This project has a number of natural extensions. In the data side, there is opportunity to add more types and modalities of cancer, e.g., histopathology slides, chest X-rays, or CT/MRI scans, and make the architecture expandable to allow more than two branches. It might also be generalized to multi-class and multi-label models whereby each branch is predictive of multiple subtypes as opposed to a binary one.

More elaborate ensemble methods, like stacking with an empirically learned metaclassifier, or having learned calibrated confidence estimates, are also an option on the modeling side. It is also interesting to consider the options to deploy lighter and more efficient architecture over the edge devices, in particular over the skin branch, which may one day be implemented on the mobile hardware.

Regarding explainability, integration of Grad-CAM with other interpretability techniques and conducting mini-clinician ones might give more information about the usefulness of the explanations. The system can be also adjusted more closely to a real clinical setting by a more interactive dashboard, which would enable clinicians to compare the models, examine heatmaps, and modify thresholds.

The last step is clinical validation though it is the most vital one. The system would have to be tested on the local hospital data before deployment, reviewed, whether it is biased on various groups of patients (e.g., skin tone, age), and tested in cooperation with medical professionals who are able to evaluate whether its recommendations are useful and safe.

References:

[1] M. A. Kassem, K. M. Hosny, R. Damaševičius, and M. M. Eltoukhy, "Machine learning and deep learning methods for skin lesion classification and diagnosis: A systematic review," *Diagnostics*, vol. 11, no. 8, art. 1390, 2021, doi: 10.3390/diagnostics11081390. Nature

[2] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115–118, 2017.

©Daffodil International University

[3] A. Ameri, "A deep learning approach to skin cancer detection in dermoscopy images," *Applied Sciences*, 2020.

[4] M. A. A. Milton, "Automated skin lesion classification using ensemble of deep neural networks in ISIC 2018: Skin lesion analysis towards melanoma detection challenge," *Proc. ISIC Challenge*, 2019.

[5] S. M. Thwin and co-authors, "Skin lesion classification using a deep ensemble model," *Applied Sciences*, vol. 14, no. 13, art. 5599, 2024.

[6] O. Akinrinade *et al.*, "Skin cancer detection using deep machine learning techniques," *Digital Health*, 2025.

[7] A. T. Ibrahim *et al.*, "Categorical classification of skin cancer using a weighted ensemble transfer learning model," *Biomedical Signal Processing and Control*, 2025.

[8] "ISIC Challenge Datasets," International Skin Imaging Collaboration (ISIC), 2018. [Online]. Available: ISIC Archive.

[9] "ISIC 2018: Skin lesion analysis towards melanoma detection challenge," ISIC 2018 dataset description, 2018.

[10] F. Scotti and ALL-IDB team, "ALL-IDB: The acute lymphoblastic leukemia image database," University of Milan, 2005. [Online]. Available: ALL-IDB website.

[11] "ALL-IDB subtypes images," Kaggle, 2020. [Online]. Available: ALL-IDB-SubtypesImages dataset.

[12] "Leukemia blood cell images dataset," Kaggle, accessed 2025. [Online]. Available: Leukemia image dataset.

[13] B. Elsayed *et al.*, "Deep learning enhances acute lymphoblastic leukemia diagnosis," *Cancers*, 2023.

[14] V. Anand *et al.*, "Deep learning model for early acute lymphoblastic leukemia classification from blood smear images," *Scientific Reports*, 2025. arXiv

[15] S. Das *et al.*, "Incremental learning for acute lymphoblastic leukemia classification using TSCO-L-LeNet," *Leukemia Research*, 2025. CVF Open Access

[16] R. F. Oybek Kizi *et al.*, "A review of deep learning techniques for leukemia detection using blood smear images," *Digital*, vol. 4, no. 1, art. 9, 2025. ResearchGate

©Daffodil International University

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. PubMed+1

[18] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708. ResearchGate+1

[19] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510–4520. PMC+1

[20] C. Szegedy *et al.*, "Rethinking the Inception architecture for computer vision," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826. SpringerLink

©Daffodil International University