
Towards PACE - CAD Systems

*Pragmatic, Accurate, Confident & Explainable Computer-Aided
Diagnosis*

Thesis approved by
the Department of Computer Science
Technische Universität Kaiserslautern
for the award of the Doctoral Degree
Doctor of Engineering (Dr.-Ing)

to

Muhammad Naseer Bajwa

Date of Defense : August 03, 2022
Dean : Prof. Dr. Jens Schmitt
Reviewers : Prof. Dr. Prof. h.c. Andreas Dengel
: Prof. Dr. Seiichi Uchida



Executive Summary

Despite phenomenal advancements in the availability of medical image datasets and the development of modern classification algorithms, Computer-Aided Diagnosis (CAD) has had limited practical exposure in the real-world clinical workflow. This is primarily because of the inherently demanding and sensitive nature of medical diagnosis that can have far-reaching and serious repercussions in case of misdiagnosis. In this work, a paradigm called PACE (Pragmatic, Accurate, Confident, & Explainable) is presented as a set of some of must-have features for any CAD. Diagnosis of glaucoma using Retinal Fundus Images (RFIs) is taken as the primary use case for development of various methods that may enrich an ordinary CAD system with PACE. However, depending on specific requirements for different methods, other application areas in ophthalmology and dermatology have also been explored.

Pragmatic CAD systems refer to a solution that can perform reliably in day-to-day clinical setup. In this research two, of possibly many, aspects of a pragmatic CAD are addressed. Firstly, observing that the existing medical image datasets are small and not representative of images taken in the real-world, a large RFI dataset for glaucoma detection is curated and published. Secondly, realising that a salient attribute of a reliable and pragmatic CAD is its ability to perform in a range of clinically relevant scenarios, classification of 622 unique cutaneous diseases in one of the largest publicly available datasets of skin lesions is successfully performed.

Accuracy is one of the most essential metrics of any CAD system’s performance. Domain knowledge relevant to three types of diseases, namely glaucoma, Diabetic Retinopathy (DR), and skin lesions, is industriously utilised in an attempt to improve the accuracy. For glaucoma, a two-stage framework for automatic Optic Disc (OD) localisation and glaucoma detection is developed, which marked new state-of-the-art for glaucoma detection and OD localisation. To identify DR, a model is proposed that combines coarse-grained classifiers with fine-grained classifiers and grades the disease in four stages with respect to severity. Lastly, different methods of modelling and incorporating metadata are also examined and their effect on a model’s classification performance is studied.

Confidence in diagnosing a disease is equally important as the diagnosis itself. One of the biggest reasons hampering the successful deployment of CAD in the real-world is that medical diagnosis cannot be readily decided based on an algorithm’s output. Therefore, a hybrid CNN architecture is proposed with the convolutional feature extractor trained using point estimates and a dense classifier trained using Bayesian estimates. Evaluation on 13 publicly available datasets shows the superiority of this method in terms of classification accuracy and also provides an estimate of uncertainty for every prediction.

Explainability of AI-driven algorithms has become a legal requirement after Europe’s General Data Protection Regulations came into effect. This research presents a framework for easy-to-understand textual explanations of skin lesion diagnosis. The framework is called ExAID (Explainable AI for Dermatology) and relies upon two fundamental modules. The first module uses any deep skin lesion classifier and performs detailed analysis on its latent space to map human-understandable disease-related concepts to the latent representation learnt by the deep model. The second module proposes Concept Localisation Maps, which extend Concept Activation Vectors by locating significant regions corresponding to a learned concept in the latent space of a trained image classifier.

This thesis probes many viable solutions to equip a CAD system with PACE. However, it is noted that some of these methods require specific attributes in datasets and, therefore, not all methods may be applied on a single dataset. Regardless, this work anticipates that consolidating PACE into a CAD system can not only increase the confidence of medical practitioners in such tools but also serve as a stepping stone for the further development of AI-driven technologies in healthcare.

Dedication

To my Maa G

Is it paradise or the feet of my mother?
I always confuse one with the other

Acknowledgements

All extolment be to ALLAH. We have no knowledge except that which He has taught us. In truth, it is He, and only He, who is perfect in knowledge and wisdom. Peace and blessings of ALLAH be upon the Holy Prophet Muhammad, the final of the prophets, and upon his blessed offspring and his exalted companions.

This work is the product of inspiration, guidance, and support from many people and it is absolutely ineluctable to extend my profound gratitude to everyone.

First and foremost, I am ever indebted to Prof. Dr. Prof. h.c. Andreas Dengel for entrusting me with a doctoral position in his SDS research group in the esteemed German Research Centre for Artificial Intelligence (DFKI). He provided a highly conducive environment for conducting research by affording me every resource required for my work. It has been an absolute privilege to work under his patronage. I am greatly thankful to other members of my Ph.D. review committee, including Prof. Dr. Seiichi Uchida from Kyushu University, Japan, and Prof. Dr. Christoph Garth from the Technical University of Kaiserslautern, Germany, for reviewing and evaluating the quality of my work and providing insightful feedback for further improvement of my future research.

Special gratitude is due to my mentors Dr. Sheraz Ahmed from DFKI and Dr. Muhammad Imran Malik from the National University of Science and Technology (NUSY), Pakistan. Without their patience, constant advice, and at times brutal yet constructive criticism, it would not have been possible to accomplish any quality work. They were always there for me right from the conceptualisation of projects to the design of experiments, analysis of the results, and, above all, painstaking revision of my rudimentary drafts. This work is every bit their achievement as it is mine.

The whole DFKI, in general, and the SDS group, in particular, were also very supportive during my tenure there as a doctoral researcher. My friends and colleagues in the group provided honest feedback during colloquiums and valuable suggestions through thought-provoking discussions in the coffee area. They made me feel at home thousands of miles away from home. I made some of my dearest friends here, and I will always cherish the time spent and memories made with them.

I would also like to acknowledge the Government of Pakistan and NUST for funding the better part of my Ph.D. Indeed, without their assistance, I could not have done it.

Finally, I am forever grateful to my family for their understanding and cooperation throughout my Ph.D. My wife, my kids, my siblings, and most of all my mother sacrificed their rights, that were due on me, to enable me to pursue this research away from home, for the most part. I can never repay them. And I can never adequately thank them.

Author's declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it is not submitted for any other academic award. This is my own work except where indicated by specific references in the text. Any work done in collaboration with, or with the assistance of, others is indicated as such. Any views expressed in this dissertation are those of the author.

MUHAMMAD NASEER BAJWA

AUGUST 05, 2022

Table of Contents

	Page
List of Tables	xiii
List of Figures	xv
List of Abbreviations	xix
1 Introduction	1
1.1 Motivation for CAD	3
1.2 Rationale for PACE	5
1.2.1 Pragmatic	5
1.2.2 Accurate	6
1.2.3 Confident	6
1.2.4 Explainable	6
1.3 Research Goal and Objectives	7
1.4 Contributions	8
1.5 Thesis Organisation	9
2 Pragmatic CAD Solutions	11
2.1 Impractical Limitations of Existing CAD Systems	12
2.1.1 Limitations in Medical Image Datasets	12
2.1.2 Limitations in Image Classification Tasks	12
2.2 Related Work	13
2.2.1 Existing RFI Datasets	13
2.2.2 Optic Disc and Optic Cup Segmentation	14
2.2.3 Skin Lesion Classification	16
2.3 Curation of G1020 Dataset	17
2.3.1 Description of G1020	18
2.3.2 Benchmark Results	19

TABLE OF CONTENTS

2.4	Extending CAD to Clinically Relevant Skin Disease Detection	25
2.4.1	Datasets Used for Skin Lesion Classification	25
2.4.2	Experiments and Results	25
2.5	Discussion	31
3	Accuracy of CAD Systems	35
3.1	Domain Knowledge as a Means to Improve Accuracy	36
3.1.1	Understanding Glaucoma	36
3.1.2	Understanding Diabetic Retinopathy	37
3.2	Related Work	39
3.2.1	Optic Disc Localisation	39
3.2.2	Glaucoma Classification	41
3.2.3	Diabetic Retinopathy Grading	43
3.3	Two-Stage Framework for Glaucoma Classification	45
3.3.1	Datasets for Disc Localisation and Glaucoma Detection	46
3.3.2	Localisation of Optic Disc	48
3.3.3	Classification of Glaucoma	54
3.3.4	Verification of Clinical Criteria for Glaucoma Detection	58
3.4	Combined Coarse-and Fine-Grained Classifier for Diabetic Retinopathy Detection	59
3.4.1	Datasets for Diabetic Retinopathy Detection	61
3.4.2	Methodology	62
3.4.3	Experiments and Results	64
3.5	Capitalising Non-Visual Metadata to Improve Classification Accuracy . . .	68
3.6	Discussions	72
4	Uncertainty Estimation in CAD	75
4.1	Problem Definition	76
4.2	Related Work	77
4.3	Hybrid Between Deterministic and Probabilistic CNNs	79
4.3.1	Uncertainty Estimation Algorithm	80
4.3.2	Datasets for Evaluating Hybrid CNN	83
4.3.3	Preprocessing	84
4.3.4	Experimental Setup and Hyperparameter Selection	84
4.3.5	Results and Analysis	85
4.4	End-to-End Training of Hybrid CNN	90

4.4.1	Experiments and Results	91
4.4.2	Analysis	92
4.5	Discussion	93
5	Explainability of CAD	95
5.1	Problem Definition	96
5.2	Achievements and Challenges in Explainable CAD	97
5.2.1	Overview of Common XAI Methods	97
5.2.2	Achievements of xAI in Medicine	101
5.2.3	Challenges for XAI Applications in Medicine	104
5.3	Explaining Network Decision using Concept Activation Vectors	112
5.3.1	Concept Activation Vectors	113
5.3.2	Dermoscopic Concepts used for Analysis	114
5.3.3	The RECOD Model	116
5.3.4	Datasets for Concept Classification and Evaluation	117
5.3.5	Experiments and Results	118
5.4	Mapping Concepts from Latent Space to Image Space	122
5.4.1	Datasets for CLM Generation	124
5.4.2	Concept Localisation Maps	126
5.4.3	Experiments and Results	127
5.5	The ExAID Framework: Providing Multi-modal Explanations	131
5.5.1	Datasets for Skin and Concept Classification	132
5.5.2	Components of the Framework	134
5.5.3	Operation Modes	136
5.5.4	Experiments and Results	139
5.5.5	Limitations	143
5.6	Understanding Glaucoma Diagnosis using GradCAM	144
5.7	Discussion	145
6	Conclusion	149
6.1	Discussion	149
6.2	Future Outlook	154
	Bibliography	157
	Curriculum Vitae	199

List of Tables

Table	Page
2.1 Segmentation performance of Mask R-CNN on G1020 dataset	20
2.2 Mean Absolute Percentage Error (MAPE) of various parameters for correctly detected optic disc and optic cup.	21
2.3 Performance metrics for glaucoma detection on G1020 and ORIGA.	23
2.4 Overview of DermNet dataset and the distribution of classes	26
2.5 Overview of ISIC Archive dataset and the distribution of classes	26
2.6 Performance metrics for 23-Class classification of DermNet using ensemble . .	28
2.7 Performance metrics of ISIC Archive-2018 using ensemble	30
3.1 Overview of datasets used for the evaluation of the heuristic method	48
3.2 Intersection Over Union (IOU) of heuristic predictions and the ground truth .	50
3.3 Accuracy of automated disc localisation compared with heuristic method . . .	52
3.4 Generalisation performance of faster R-CNN on unseen datasets	53
3.5 Precision, Recall and F1 score of classification with random train and test split	55
3.6 Comparison of obtained Area Under the Curve (AUC) with random train and test split	56
3.7 Precision, Recall, and F1-score of classification with cross validation	56
3.8 Comparison of obtained Area Under the Curve (AUC) with cross validation .	57
3.9 Performance comparison of our method with the later approaches using ORIGA dataset for glaucoma detection	58
3.10 Performance comparison of Inception V3 trained for glaucoma detection using different variations of RFIs	60
3.11 Overview of EyePACS dataset	61
3.12 Overview of Messidor dataset showing grading criteria and class distribution .	62
3.13 Class distribution for Normal vs Abnormal classification	65
3.14 Class distribution for Referable vs Non-Referable classification	65
3.15 Detailed performance metrics for various classification settings	66

LIST OF TABLES

3.16	Class distribution for 4-Class classification	67
3.17	Class distribution for 3-Class classification	67
3.18	Classification performance of various models with and without metadata	72
3.19	Performance of best performing individual models versus two types of ensemble predictions	73
4.1	Time and space requirement of fully deterministic, fully Bayesian and hybrid models for some datasets	82
4.2	Distribution of datasets used to evaluate the proposed architecture	83
4.3	Comparison of fully deterministic, fully Bayesian, and the proposed hybrid models on different datasets without using uncertainty estimation	86
4.4	Comparison of fully Bayesian and the proposed hybrid models on different datasets with uncertainty estimation	88
4.5	Comparison of Accuracy (%) of deterministic, Bayesian, and proposed hybrid models on different datasets without using uncertainty estimation	91
4.6	Comparison of Bayesian and Hybrid models on different datasets before and after uncertainty estimation.	92
5.1	Distribution of image samples into different concept classes in PH ² and derm7pt datasets	117
5.2	The distribution of data in training, validation and test splits for disease-level classification	133
5.3	The distribution of data in training, validation and test splits for concept-level classification with D7PH2 dataset	133
5.4	Performance evaluation of lesion classifier on various datasets	139
5.5	Performance evaluation of concept classifiers on various datasets	140

List of Figures

Figure	Page
1.1 Computer-Aided Diagnosis sits at the intersection of medicine and computer science	2
2.1 Density map of optic disc in G1020 and ORIGA	18
2.2 Sample images with optic cup, optic disc and bounding box annotations . . .	19
2.3 Example images with incorrect OD and OC detection	21
2.4 Receiver operating characteristic (ROC) and AUC for 6-fold cross-validation on G1020 and 5-fold cross-validation on ORIGA datasets using Inception V3	24
2.5 Visualisation of image embeddings learnt by DL model from G1020 and ORIGA datasets plotted on 2D plane after dimensionality reduction using PCA	24
2.6 Accumulated confusion matrix of 23-ary classification of DermNet dataset . .	29
2.7 Confusion Matrix showing number of correctly classified and misclassified images per class in ISIC Archive-2018	30
2.8 Examples of correctly and incorrectly classified skin diseases from ISIC Archive dataset	31
3.1 Stages of glaucoma in retinal fundus images taken from Rim-One dataset . . .	37
3.2 Progression of diabetic retinopathy from healthy to proliferative stage is subtle and gradual	38
3.3 Complete framework of disc localization and classification	46
3.4 Workflow of semi-automatic ground truth generation mechanism	49
3.5 Binary images showing misleading bright spots	49
3.6 Results of Heuristic Localisation of OD. Subfigure 5(d) shows the only example where heuristic failed.	51
3.7 Internal components of faster RCNN	52
3.8 Results of automated localization on different datasets	53

LIST OF FIGURES

3.9	Convolutional neural network used for glaucoma classification	54
3.10	Confusion matrix showing the distribution of True Positives, False Positives, and False Negatives	55
3.11	Results of Glaucoma Classification using DCNN	57
3.12	An original RFI with two variations to obscure optic disc	59
3.13	System Overview of combining coarse-grained and fine-grained classifiers . . .	63
3.14	Effects of preprocessing steps on retinal fundus images	63
3.15	Conversion of five retinopathy grades in EyePACS to quaternary, ternary and binary classification	64
3.16	Confusion matrices for EyePACS and Messidor for multi-class classification tasks	68
3.17	ROC Curves for all classification tasks	69
3.18	Incorporating metadata in image classifier by direct concatenation with visual embeddings.	71
3.19	Converting metadata into feature vector by processing them with MLP before concatenating with visual embeddings.	71
4.1	The proposed hybrid model	79
4.2	An analysis of confidence comparison for all three approaches on various samples of CIFAR10 and ORIGA datasets	87
4.3	Comparison of output probabilities for fully Bayesian and hybrid training approaches on ORIGA dataset	89
4.4	Trade-off between number of uncertain samples and the accuracy on remain- ing predictions	89
5.1	Topology of the XAI process with optional model and data correction as well as taxonomy of common and relevant explainable AI methods in medical image analysis	98
5.2	Comparison between Grad-CAM heatmaps generated before and after cor- rection of the trained network using model correlation methods	102
5.3	Exemplary cases of skin lesion concepts from derm7pt dataset	114
5.4	Overview of training concept classifiers and calculating CAV and TCAV scores	119
5.5	Validation accuracies of all concept classifiers trained and tested individually on derm7pt and PH ² datasets	120
5.6	The TCAV scores of each concept for derm7pt with respect to each target class on <i>mixed_6h</i> layer of RECOD model	121

5.7	The TCAV scores of each concept for PH ² with respect to each target class on <i>mixed_6h</i> layer of RECOD model	122
5.8	The sorting of the test images with respect to the presence of Typical Pigment Network (<i>PN_T</i>)	123
5.9	The sorting of the test images with respect to the presence of Irregular Streaks (<i>ST_IR</i>)	123
5.10	The sorting of the test images with respect to the presence of Regression Structure (<i>RS</i>)	124
5.11	Training samples from SCDB dataset	125
5.12	Concept Localisation Maps (CLMs) for SCDB images	128
5.13	Average IOU, precision and recall over all 10 concepts for predicted CLMs applied to three network architectures	129
5.14	Examples for CLMs generated from <i>SE-ResNeXt-50</i> trained on binary classification of gender with CelebA dataset	130
5.15	The ExAID Framework architecture	134
5.16	Diagnostic mode of ExAID can be used as decision support system in routine clinical workflows	137
5.17	Educational mode of ExAID can help in training of resident dermatologists by allowing them to explore many of its interactive features.	138
5.18	Positive and negative examples of visual explanations provided by ExAID along with the corresponding samples and ground truth concept masks	141
5.19	Positive and negative examples of textual explanations are provided by ExAID along with the corresponding skin lesion samples	142
5.20	Examples of correct classification by Inception V3 albeit by looking at different regions in the RFI than analysed by ophthalmologists	145

List of Abbreviations

AAPM	American Association of Physicists in Medicine
ABCD rule	Asymmetry, Border, Colour, Diameter
ACD	Agglomerative Contextual Decomposition
AD	Automated Diagnosis
AI	Artificial Intelligence
AMD	Age-related Macular Degeneration
AOPC	Area Over the MoRF Perturbation Curve
AUC	Area Under the Curve
BCC	Basal Cell Carcinoma
BNN	Bayesian Neural Network
CAD	Computer-Aided Diagnosis
CADSC	Computer-Aided Detection in Diagnostic Imaging Subcommittee
CAM	Class Activation Map
CAV	Concept Activation Vector
CDR	Cup-to-Disc Ratio
CNN	Convolutional Neural Network
CPN	Cup Proposal Network
CT Scan	Computed Tomography Scan
DL	Deep Learning

LIST OF FIGURES

DNN	Deep Neural Network
DPN	Disc Proposal Network
DSS	Decision Support System
EG	Expressive Gradient
EHR	Electronic Health Record
ELBO	Evidence Lower BOund
FCNN	Fully Convolutional Neural Network
FPR	False Positive Rate
GAN	Generative Adversarial Network
Glow	Generative flow
GMP	Generalised Motion Pattern
GON	Glaucomatous Optic Neuropathy
GT	Ground Truth
HCI	Human-Computer Interface
HCNN	Hybrid Convolutional Neural Network
HT	Hough Transform
IAF	Inverse Autoregressive Flows
ICDR	International Clinical Diabetic Retinopathy
ILR	Initial Learning Rate
INN	Invertible Neural Network
IOP	IntraOcular Pressure
IoT	Internet of Things
IOU	Intersection Over Union
IrMA	Intraretinal Microvascular Abnormalities

ISBI	International Symposium on Biomedical Imaging
ISD	International Society of Dermoscopy
ISNT	Inferior, Superior, Nasal, Temporal
LBP	Local Binary Pattern
LRE	Local Reparametrization Estimator
LRP	Layer-wise Relevance Propagation
LSTM	Long Short-Term Memory
MAF	Masked Autoregressive Flows
MAPE	Mean Absolute Percentage Error
mAP	mean Average Precision
MCMC	Markov Chain Monte Carlo
MCQ	Multiple Choice Question
MEL	Melanoma
MIA	Medical Image Analysis
MIT	Massachusetts Institute of Technology
MLE	Maximum Likelihood Estimation
ML	Machine Learning
MLP	MultiLayer Perceptron
MRI	Magnetic Resonance Imaging
NAF	Neural Autoregressive Flows
NLP	Natural Language Processing
NMS	Non-Maximum Suppression
NV	Naevus
real NVP	real-valued Non-Volume Preserving

LIST OF FIGURES

OD	Optic Cup
OCT	Optical Coherence Tomography
OD	Optic Disc
ONH	Optic Nerve Head
OOD	Out-Of-Distribution
PACE	Pragmatic, Accurate, Confident & Explainable
PCA	Principal Component Analysis
PN_A	Atypical Pigment Network
PN	Pigment Network
PN_T	Typical Pigment Network
PPA	PeriPapillary Atrophy
RCV	Regression Concept Vector
RECOD	REasoning for COMplex Data
RFI	Retinal Fundus Image
RGB	Red, Green, Blue
RNN	Recurrent Neural Network
ROAR	RemOve And Retrain
ROC	Receiver Operating Characteristic
ROI	Region of Interest
RPN	Region Proposal Network
SBF	Sliding Band Filter
SBN	Sigmoid Belief Network
SDGs	Sustainable Development Goals
SHAP	SHapley Additive exPlanations

SK	Seborrheic Keratosis
SSIM	Structural Similarity
STD	Standard Deviation
SVM	Support Vector Machine
TPR	True Positive Rate
UN	United Nations
USD	United States Dollar
VFL	Visual Field Loss
VI	Variational Inference
VQA	Visual Question Answering
VRH	Visual Relevance Heatmap
VFT	Visual Field Test
WHO	World Health Organisation
XAI	eXplainable Artificial Intelligence

Introduction

With the advent of computers, many laborious and time-consuming tasks were delegated to these newly invented machines. Among other applications, computers were used to process and analyse medical images [1, 2], since it was one of those tasks in which computers were thought to perform better than humans. This early research on medical image processing produced promising results encouraging the researchers to dream big - a dream where a medical diagnosis was altogether delegated to machines [3, 4]. Thus the idea of Automated Diagnosis (AD) was born. But surely, the researchers hastened to place their trust in those rudimentary computers with limited computational capabilities and non-availability of advanced image processing techniques. And when the available computational power and algorithms of that time were unable to deliver on those high expectations, there was a sense of despair among the researchers. Ralph Engle notes in his review on computers as diagnostic aids in medical decision making [5],

”Thus, we do not see much promise in the development of computer programs to simulate the decision-making of a physician.

In the future, computers will certainly be used in medicine in many ways. However, after many years we have concluded that we should stop trying to make computers act like diagnosticians.”

This was perhaps the result of expecting from the computers too much too early. Luckily, a necessary course correction was made and the job description of computers

was redefined as "a second pair of eyes". This new approach shifted the research focus from automated diagnosis to Computer-Aided Diagnosis (CAD), in which the role of computers is to assist clinicians by providing a second opinion instead of replacing them in clinical workflows. Contrary to AD systems, which were expected to perform at par or better than human counterparts since they were solely responsible for final diagnosis, the performance of CAD systems was only needed to be complementary to that of diagnosticians [6]. Today, CAD has evolved into a major research area in medical diagnosis, and the rise of modern hardware accompanied by sophisticated Artificial Intelligence (AI) based algorithms has provided much-needed support.

Computer-Aided Diagnosis is a multidisciplinary research area primarily involving Medicine and Computer Science as shown in Fig. 1.1. The branch of computer science that deals with the development of expert systems, like CAD, is Deep Learning (DL) [7], which is a part AI in which a computer algorithm analyses raw data and automatically learns discriminatory features needed for recognising hidden patterns in them. Over the last decade, this field has witnessed striking advances in the ability to analyse various types of data, especially images [8] and natural language [9]. The most common DL models are trained using supervised learning, in which datasets are composed of inputs, for example, radiography images of lungs, and corresponding target output labels, for instance, any pulmonary pathology. Healthcare and medicine have greatly benefited from recent advances in image classification and object detection [10], particularly those medical disciplines in which diagnoses are primarily based on the detection of morphologic changes such as pathology, radiology, ophthalmology, and dermatology, etc. In such medical domains, digital images are captured and provided to DL algorithms for CAD. These advanced algorithms have made their mark on automated detection of many dis-

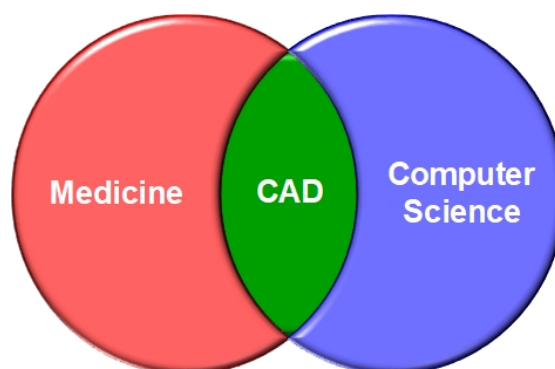


Figure 1.1: Computer-Aided Diagnosis sits at the intersection of medicine and computer science

eases like tuberculosis [11], breast malignancy [12], glaucoma [13], diabetic retinopathy [14] and serious brain findings such as stroke, haemorrhage, and mass effects [15].

1.1 Motivation for CAD

United Nations (UN) recognised healthcare and well-being as one of the 17 Sustainable Development Goals (SDGs) to create a better future for everyone by 2030 [16]. However, achieving this goal requires concerted and sustained efforts in utilising all available resources to improve healthcare since many people are needlessly suffering from preventable diseases. With a rapid increase in population, and consequently rising demand for healthcare services, World Health Organisation (WHO) estimates a global shortage of 18 million healthcare workers by 2030 [17] and nine million more nurses and midwives are required by the same time to achieve this SDG [18]. In face of this global shortage and maldistribution of healthcare resources, the potential of AI to help achieve this SDG and solve other greatest global challenges was identified by AI for Good [19], which is a UN initiative to provide a global platform for researchers.

There are some obvious benefits of using CAD over manual diagnosis by medical professionals in routine clinical scenarios. Some of them are briefly mentioned below.

- **Financial and Time Economy:** It takes hundreds of thousands of euros to train a medical doctor, whether the cost is paid by the student, the state, or shared by both. This huge financial cost comes on top of years of rigorous training. Therefore, as mentioned above, the number of medical practitioners is estimated to be far fewer than actually required and their uniform distribution poses another challenge in the way of the equitable dispensation of medical services. Furthermore, there is no way of mass-producing medical practitioners to amortise training time and cost. On the other hand, the primary cost involved in software-based CAD solutions is related to research and development. Once a system is developed, tested, and approved for use in clinical routine, it can be cheaply and efficiently mass-produced and deployed across many healthcare establishments.
- **Equitability:** Like everyone else, healthcare professionals too are subject to cognitive biases which can greatly hinder their ability to make a correct and fair assessment of the case at hand. These cognitive biases are ubiquitous across clinical practice and are increasingly recognised as the prime source of error in the diagnosis and prognosis of a patient [20, 21]. CAD systems, which are essentially complex mathematical models, do not suffer from such cognitive biases. However,

there are peculiar biases in CAD systems that are mostly related to dataset curation [22] like spectrum biases, which occur when the proportion of selected diseases in the dataset is not representative of the target population and verification biases that creep in when a dataset is limited to only those patients that have definitive verification of a given disease, or lack thereof.

- **Efficiency:** In a hospital’s outpatient department, it can take up to 24 hours to furnish a report for a crucial diagnostic modality like CT scan and MRI – in emergency situations, however, this time could be as short as an hour. This long time accounts for the availability of medical specialists and actual time spent on reading and interpreting diagnostic modalities. However, CAD systems can be ideally available round the clock and with powerful computing machines, the time to process a case is much shorter than that for humans. In addition, when used properly in an assistive role, these CAD systems can also cut the time short for human graders to review and interpret a case [23, 24].
- **Accessibility:** Availability of medical services is a major barrier, particularly in low-income countries and populations living in remote areas. Many people might suffer unnecessarily from advanced stages of diseases, which, if diagnosed early, can be easily treated and prevented from affecting a patient’s quality of life. For example, retinal disorders like Glaucoma, Diabetic Retinopathy and Age-Related Macular Degeneration (AMD) progress slowly and take decades to develop noticeable symptoms, at which stage the damage already caused is sometimes irreversible. Large-scale screening programmes for such diseases can help mitigate the needless burden on the healthcare system by nipping the evil in the bud. However, such systematic large-scale programmes are expensive and require specialised medical experts making it difficult to scale up and expand rapidly and widely enough to cater to the needs of a given population [25]. Screening programs using CAD systems have the potential to fill this niche by providing the first line of defence against disease progression [14]. They can allow a timely referral to specialist doctors and thus help improve prognosis.
- **Objectivity:** Manual diagnosis may be affected by physicians’ level of experience and different diagnostic algorithms in which they are formally trained. This can lead to multiple experts disagreeing on their diagnosis for a certain condition [26, 27]. Additionally, due to physicians’ subjective judgements, manual diagnosis is hardly reproducible [28]. On the other hand, CAD can provide standardised

and objective diagnosis of various diseases which can be reproduced. Fatigue and tiredness of doctors can also interfere with their ability to perform effectively in their clinical diagnosis whereas CAD systems are completely free from any such shortcomings and can provide consistent performance throughout their lifespan.

1.2 Rationale for PACE

CAD has evolved significantly from a rudimentary image processing tool to a potential digital aid in routine clinical workflows. This evolution was partly due to technological advancements in the past few decades and partly because of changing expectations of the end-users. There are a few examples of limited use of CAD systems in some medical domains [29] like mammography and radiology. However, this promising application of AI is not as prevalent as it should be in current times. This thesis identifies potential reasons for this scarce use of CAD in today's clinical setups and proposes a paradigm abbreviated as PACE (Pragmatic, Accurate, Confident, & Explainable), which any modern CAD system must comply with in order to be accepted and deployed in real-world scenarios. The rationale for each aspect of PACE is briefly described below.

1.2.1 Pragmatic

Early prototypes of every new invention start with a simple yet working example serving as a proof-of-concept. Such rudimentary examples need to mature into sophisticated solutions that can be used in real-world applications. In CAD using Medical Image Analysis (MIA), most publicly available medical image datasets are curated by using the over-simplistic image capturing conditions, which are not representative of medical imaging in hectic clinical routines. Such datasets, though provide encouraging results in lab settings, are unable to train robust image classifiers that can work reliably in a clinical environment. Similarly, the task of medical image classification can also be made overly easy even when there is room to venture into realistic and pragmatic image diagnosis tasks, for example choosing to classify a few diseases when images pertaining to multiple diseases are available in the dataset. Therefore, a pragmatic CAD system needs to be trained to identify a wide range of clinically relevant diseases using datasets that are representative of medical imaging in real life.

1.2.2 Accurate

Accuracy is surely one of the most salient features of any CAD system. Although in the early days CAD systems were not expected to perform at par with human counterparts since they were supposed to only provide an opinion to human diagnosticians [6] yet with the advancement of computing resources and modern DL based algorithms, these CAD systems are now anticipated to work competitively with human experts. And modern CAD systems have proved to be able to do just that [14, 30, 31]. Achieving high accuracy, however, is not only a matter of using cutting-edge hardware, modern software suits, and a humongous amount of data. It requires an in-depth understanding of the diagnostic task at hand and possibly mimicking experts' behaviour into AI algorithms to combine automated feature learning of AI and established clinical criteria for providing correct diagnoses.

1.2.3 Confident

An AI-based CAD system usually only provides numerical values corresponding to its predictions for given data samples. Sometimes these predictions are incorrect. Other times these predictions are correct but could be the result of a lucky guess. However, in critical application areas like medical diagnosis, there is justifiable reluctance by clinicians and patients to trust an algorithmic prediction without any consideration on the possibility of a fluke. Therefore, despite the extensive implementation of Deep Neural Networks (DNNs) in CAD [32–34], there has been growing advocacy for the need for uncertainty estimation in such decision support systems [35] in order to successfully deploy these solutions in the detection and diagnosis of diseases. It is, therefore, imperative for a trustworthy CAD system that it only provides an opinion in cases when it is sufficiently certain about its prediction. This can be achieved by estimating the uncertainty of a trained DNN and supplementing it with the network's prediction.

1.2.4 Explainable

Numerous remarkable studies have been conducted recently successfully applying deep learning for disease classification using various medical image modalities [10]. However, the acceptance of such CAD solutions with doctors and patients remains uncertain partly due to the fact that the exact decision-making process of these complex DNNs is not unambiguous. This lack of transparency in the whole decision-making process cannot be overlooked in various high-stake application areas including medical diagnosis. With

increasing legislation across the world conferring *Right to an Explanation* [36] to any subject of algorithmic decisions, it has become paramount for modern AI-based algorithms to supplement their decision with justifiable explanations.

This should be emphasised here that this PACE paradigm is not exhaustive in that there may be other important features, which, in the fullness of time, may be added to the final industry standard CAD solutions.

1.3 Research Goal and Objectives

The ultimate goal of this thesis is to identify potential limitations in the successful implementation of CAD systems in the clinical workflow of the near future and propose possible solutions so that CAD can have a tangible social impact on the healthcare system. In achieving this goal, this thesis presents the PACE paradigm which can act as a blueprint to develop CAD systems that are not confined to lab settings and present potential methods for realisation of each of the four dimensions of PACE. The following objectives correlate with some of the implementation strategies that may be employed to enrich a CAD system with PACE.

1. To curate and make a publicly available, high-quality, and large dataset that characterises realistic imaging conditions in routine ophthalmology for the training of robust glaucoma classification models.
2. To study the feasibility of emulating experts' decision-making process and integrating it into DL-based medical image classifiers for improving diagnostic performance.
3. To explore economical and effective ways of estimating uncertainty associated with DL-based CAD systems without compromising on the accuracy and computational cost.
4. To investigate various methods of explainable AI, their viability in image-based CAD systems, and development of an explainable CAD which provides easy-to-understand multi-modal explanations for medical professionals and patients.

1.4 Contributions

In this thesis various methods have been proposed, which may help a CAD system pick some pace. However, some of the methods are not viable in every application or dataset and therefore it is difficult to apply all of them on a certain dataset. Therefore, glaucoma classification using RFI is taken as primary use case and where required other application areas like cutaneous disease classification or diabetic retinopathy detection have also been used as secondary use cases to showcase that this PACE paradigm can be adopted in a variety of application areas. Important contributions of this thesis are as follows.

1. Since non-availability of high-quality, large, and publicly available medical datasets that have the characteristics of actual images captured in clinical practice is one of the bottlenecks of developing a reliable CAD solution, a dataset of Retinal Fundus Images (RFIs) called G1020 is gathered, curated, and published as part of this thesis. This dataset does not impose strict inclusion criteria making it sufficiently challenging for glaucoma classification and segmentation of optic discs and optic cups.
2. Many existing works on image-based disease classification tend to tread cautiously on unchallenging paths of classifying only a few diseases. However, a pragmatic CAD must be able to identify a broad range of clinically relevant diseases. Therefore, in this work, one of the largest publicly available skin lesion datasets called DermNet is used for successful classification of more than 600 distinct skin lesions.
3. To improve the classification performance of existing image classifiers, this thesis proposes to augment DL models with knowledge garnered from ophthalmologists in detecting retinal disorders like Glaucoma and Diabetic Retinopathy. For glaucoma detection, for example, doctors pay attention to the optic disc region in RFI. Therefore, a two-stage model is developed which can automatically localise optic disc from whole RFI and then analyse it closely to identify biomarkers of this disease. Similarly, diabetic retinopathy is diagnosed by first briefly glancing over the whole RFI, identifying potential Regions Of Interest (ROIs), which are usually scattered all over the image, and then closely analysing those ROIs to look for signs of diabetic retinopathy. To mimic this, it is shown that an ensemble of fine-grained and coarse-grained image classifiers can provide competitive results on a range of classification tasks using publicly available datasets like EyePACS and Messidor.

4. Understanding the importance of uncertainty estimates with AI predictions in sensitive application areas like medical diagnosis, a classifier is developed that is not compulsive in its predictions, meaning that the model has the option to withhold its prediction if it is not entirely certain about it. This model is a hybrid between deterministic and probabilistic Convolutional Neural Networks (CNN). Deterministic CNNs have been shown to provide better classification performance while being sufficiently economical with respect to computations. However, they are innately unable to provide uncertainty estimates. Bayesian CNNs, a type of probabilistic neural networks, can provide posterior distribution which can be used to estimate the network's uncertainty. However, they are computationally expensive and not as high-performing as their deterministic counterparts. A hybrid between them, therefore, combines the merits of both training paradigms.
5. Substantial research is conducted in this thesis on ways of elucidating the decision-making process of DL-based image classifiers. Using Concept Activation Vectors it is verified that DL algorithms are able to encode and utilise the same concepts as defined and employed by dermatologists. Part of this thesis presents methods to localise the region on input space which was most influential in learning those concepts. It was found that located regions conform to the spatial positions of concepts. This verification of concept learning and their localisation is integrated into a unified framework called ExAID, which can spit out easy-to-understand textual explanations justifying the prediction of the classifier.

1.5 Thesis Organisation

The rest of the thesis is organised as follows. Chapter 2 opens the core body of research by addressing the first dimension of the PACE framework. This chapter proposes two disjoint yet complementary ways of making a CAD system more pragmatic. The first way deals with curation and publication of a large publicly available dataset of RFIs to ensure training of robust glaucoma classifier and segmentation algorithms. The second approach advocates to use the full potential of DL-based image classifiers and shows that modern classifiers are fairly capable of identifying hundreds of skin lesions. Chapter 3 explores various ways of improving the classification performance of image-based CAD systems. It takes retinal disorders as example use cases and shows that DL models can be configured to follow the diagnostic approach used by ophthalmologists. This chapter also studies the possibility of modelling and incorporating non-visual clinical data into a deep model for

performance improvement. Chapter 4 presents a method to estimate the uncertainty of DL-based classifiers. The method given in this chapter employs a novel hybrid between deterministic and probabilistic CNNs to allow the estimation of uncertainty with the classifier's predictions. The last dimension of PACE, i.e. Explainability, is highlighted and addressed in Chapter 5. This chapter presents various approaches that can be used to get a sneak peek into the decision-making process of image-based CAD systems. Finally, Chapter 6 concludes the thesis with a comprehensive discussion and summarises important findings.

Pragmatic CAD Solutions

Computer-Aided Diagnosis of various diseases is receiving a lot of attention from the research community due to its far-reaching benefits of providing swift and accurate large-scale screening as well as reducing physicians' workload in routine clinical setups [37]. However, one of the biggest hurdles of CAD not being widely used in real-world healthcare environments is its incompatibility with clinical workflows. As with any new invention, CAD systems are developed in laboratory setups under relatively strict, and sometimes over-simplistic, conditions. These lab-born CAD systems must morph significantly if they are to be successfully deployed outside the controlled laboratory environment. In this chapter, two such limitations of CAD systems are identified and potential solutions are provided. First, realising that the non-availability of a publicly available medical image dataset that represents images captured in hectic clinical routine might hamper a trained DL-based image classifier's performance when put to test in the field, a large publicly available RFI dataset is curated and published for glaucoma classification and many image processing tasks. Second, instead of performing binary or tertiary classification tasks, the potential of DL-based image recognition models is stretched to classify hundreds of clinically relevant skin lesions to make it more useful in clinical scenarios. There may be many other aspect which contribute towards making a CAD system more pragmatic and practically usable like safety [38] and integration into clinical workflows [39], for instance. However, notwithstanding the significance of these other aspects, the focus in this research is maintained on the issues mentioned above.

2.1 Impractical Limitations of Existing CAD Systems

2.1.1 Limitations in Medical Image Datasets

Deep Learning based techniques have been used to automatically detect various ocular diseases like glaucoma [40], diabetic retinopathy [14], AMD [41] and many other retinal disorders [42]. Recently, it has been shown that RFIs can be used to detect non-ocular diseases as well like Type-II diabetics [43], anaemia [44], and cardiovascular risks [45]. For automated glaucoma detection, different image modalities and clinical tests are used, for instance, RFIs [46], Optical Coherence Tomography (OCT) [47], and Visual Field Tests (VFTs) [48]. However, fundus imaging is the most common and inexpensive imaging technique [49] for large-scale screening of various retinal diseases.

The scarcity of large publicly available medical image datasets for automated detection of various diseases has been the bottleneck for the successful application of AI towards practical CAD systems. A few small datasets that are available for the research community usually suffer from impractical image capturing conditions [50] and stringent inclusion criteria, for example for RFI datasets centralising Optic Disc (OD) [51] or macula and removing images containing certain artefacts [52]. These shortcomings in the already limited choice of existing datasets make it challenging to mature a CAD system so that it can perform well in a real-world environment. Since the most important application of automated glaucoma detection is cost-effective and large-scale screening [53] of the general population, these automated solutions should be able to perform well in the field with fundus images taken in day-to-day practice without many constraints [54]. Removing images that do not conform to strict inclusion criteria for example, from the available datasets might result in a CAD that works exceptionally well in a controlled *laboratory* environment but is most likely to fail in routine screening or in a clinical setting.

2.1.2 Limitations in Image Classification Tasks

Most publicly available datasets for clinical or dermoscopic images like Interactive Atlas of Dermoscopy [55], Dermofit Image Library [50], Global Skin Atlas, MED-NODE [56] and PH2 [57] etc. contain only a few hundred to a couple of thousand images. Ali et al. [58] reported that around 78% of the studies they surveyed used datasets smaller than 1000 images and the study using the largest dataset had 2430 images. Therefore, most existing works on CAD of skin diseases use either private or very small publicly available datasets. Additionally, these studies usually render overwhelming focus on only

binary or ternary classification of skin diseases, and not much attention is paid to multi-class classification to explore the full potential of DL. Therefore, such studies, though produce glamorous publishable results with performance metrics well above 90% in some cases, act merely as a proof-of-concept for the efficacy of AI in dermatology. There is a pressing need to extend previous works by showing that DL models are fairly capable of recognising hundreds of skin lesions, and therefore should be capitalised to their full extent.

2.2 Related Work

This section provides an overview of some of the research works relevant to the above-mentioned two limitations of CAD systems, namely non-availability of large public datasets for glaucoma detection and few-class classification of skin lesions even when a large number of classes are available in dermoscopic datasets.

2.2.1 Existing RFI Datasets

ORIGA Online Retinal fundus Image database for Glaucoma Analysis and research (ORIGA) [51] is one of the largest and most commonly used datasets for glaucoma detection made public since 2010. This dataset consists of 650 images (168 glaucomatous, 482 healthy) collected by Singapore Eye Research Institute between 2004 and 2007. The dataset provides class labels for healthy and glaucoma, OD and Optic Cup (OC) contours, and Cup-to-Disk Ratio (CDR) values for each image.

RIM-ONE This small dataset [59] consists of 169 high-resolution RFIs collected at three Spanish hospitals. Each image is classified as healthy, early glaucoma, moderate glaucoma, deep glaucoma, or ocular hypertension. Additionally, it provides OD segmentation annotations to evaluate disc detection algorithms.

RIGA Retinal fundus Images for Glaucoma Analysis (REGA) [60] consists of 750 images taken from Messidor dataset [61] and two clinics in Saudi Arabia. This dataset provides OD and OC boundary annotations; however, it does not provide any diagnosis with regards to glaucoma detection.

REFUGE REtinal FUndus Glaucoma Chalenge (REFUGE) [62] is one of the largest and RFI datasets publicly available for glaucoma detection. It was made public in 2018 as a grand challenge and consists of 1200 fundus images with ground truth segmentation

of OD and OC and clinical glaucoma labels. Despite the large size of this dataset, it is highly unbalanced towards the healthy class as it contains only 120 glaucoma images.

ACRIMA This new dataset [52] consists of a total of 705 fundus images with 396 glaucoma images and 309 normal images taken with a centred optic disc. The dataset does not provide any annotations for OD and OC segmentation. A relatively balanced proportion of normal and glaucomatous images in this dataset makes it particularly suitable for training DL-based classifiers.

ODIR Ocular Disease Intelligent Recognition [63] is a dataset published by Peking University China for International Competition on recognition of eight ocular conditions including glaucoma. This structured ophthalmic dataset consists of left and right fundus images of varying resolution from 5000 patients who visited various hospitals and medical centres in China. The images are taken under realistic settings with different cameras. Seven thousand images from around 3500 cases are provided for training and 1000 images are reserved for off-site testing. The number of glaucoma positive images in the training set are only 207 which accounts for 0.03% of the total images.

DRISHTI-GS1 This small dataset [64] of 101 high-resolution colour RFIs captured after pupil dilation with 30-degree FOV. The dataset is divided into 50 training and 51 testing images. The exclusion criteria include poor contrast and positioning of OD other than in the centre of the image. This dataset consists of 70 glaucoma-positive and 31 glaucoma-negative images. This is the only dataset encountered in our research where the proportion of glaucomatous images is higher than normal images.

LAG Large-scale Attention based Glaucoma dataset [65] consists of 5824 colour RFIs with 2392 glaucomatous images and 3432 healthy images collected from Beijing Tongren Hospital. In addition to binary labels, this dataset is unique in providing attention maps of grading ophthalmologists captured using a simulated eye-tracking experiment. Diagnosis is based on considerations of morphologic and functional analysis of the images like Intra-Ocular Pressure (IOP), Visual Field Loss (VFL), and manual OD assessment.

2.2.2 Optic Disc and Optic Cup Segmentation

Almazroa et al. [66] devised an image processing based heuristic algorithm for optic disc segmentation using RIGA dataset, which was later made public [60]. Their algorithm achieved an accuracy of 83.9% for marking the OD area and centroid. Al-Bander et

al. [67] used a U-Net [68] like Fully Convolutional Neural Network (FCNN) for OD and OC segmentation and evaluated their method on 1129 RFIs from five public datasets. Their method was shown to be invariant to population demography, camera models, and other ocular diseases. They outperformed the state-of-the-art on two datasets and gave competitive results on two datasets without training on these four datasets. Fu et al. [69] attempted to jointly segment OD and OC. They modified faster R-CNN [70] by replacing its Region Proposal Network (RPN) with two networks named Disc Proposal Network (DPN) and Cup Proposal Network (CPN). The proposed network is tested on a publicly available ORIGA dataset and 1676 images of a private dataset called SCES [71] and outperformed state-of-the-art methods for joint segmentation of OD and OC.

Park et al. [72] compared YOLO V3 [73], ResNet [74], and DenseNet [75] architectures for automatic Optic Nerve Head (ONH) localisation and CDR calculation. Using 2163 RFIs captured at Pusan National University Hospital, South Korea, they found that DenseNet performed best and YOLO-V3 performed worst in terms of mean Average Precision (mAP) and Intersection Over Union (IOU) for low-resolution images of 224×224 and 416×416 . However, when the image resolution increased to 832×832 , YOLO-V3 took the lead from ResNet and fared at par with DenseNet. In terms of mean detection time, ResNet took the least amount of time and YOLO-V3 took the most amount of time while running on CPU. However, YOLO-V3 appeared to capitalise better on GPUs than competing architectures. Zhou et al. [76] employed Support Vector Machine (SVM) to classify the brightest OD region in RFI based on structural and intensity features. Image processing techniques like convex hull are applied on detected candidate regions to locate centre of OD. They achieved 96.7%, 97.8%, and 100% localisation accuracy on a total of 259 test images taken from DIARETDB0 [77], DIARETDB1 [78], and DRIVE [79] datasets, respectively. For training SVM, 81 images from STARE [80] dataset are used. In [81], Sreng et al. performed OD segmentation using a combination of DeepLabV3+ [82] and MobileNet [83]. The encoder part of DeepLabV3+ is replaced with various CNN architectures. Evaluation is performed on 2787 images from five datasets with an accuracy of 99.7% and dice coefficient of 91.73% on the combined test set. Joshi et al. [84] proposed two methods for OD segmentation. The first method uses Interference Maps [85], which are obtained from Generalised Motion Pattern (GMP) [86] of the images. The second method makes use of Grab Cut algorithm [87]. Their methods are evaluated on DRISHTI-GS1 dataset and they achieved 97% precision, 90% recall, and accuracy of 88%.

Veena et al. [88] performed OD and OC segmentation, and using these segmentation results they calculated CDR. They first employ various image processing techniques

like adaptive histogram equalisation, Sobel edge detection algorithm, and Watershed algorithm to enhance and capture salient image features. Afterwards, two U-Net based models are used for separately segmenting OD and OC. They achieved 98% accuracy for OD segmentation and 97% for OC on DRISHTI-GS dataset. Similarly, using DRISHTI-GS, DRIONSDB, and RIM-ONE v3 datasets, Magnipudi et al. [89] achieved 96.62%, 96.15% and 98.42% IOU respectively for OD segmentation via U-Net based model.

2.2.3 Skin Lesion Classification

Towards automated skin disease classification, Jibhakate et al. [90] detected seven skin cancers from HAM10000 [91] dataset. They used 10% of this dataset for validation of pre-trained models and 20% for CNNs trained from scratch. However, their train/validation split was selected 'after considerable permutations of various train-test splits'. Although they achieved up to 99% accuracy with some pretrained models, in absence of the exact train/validation split, these results cannot be reproduced. Salian et al. [92] classified six skin lesions using HAM10000 and PH² [57] datasets. Data augmentation was used to balance the under-represented classes. They report interesting results where three DL-based classifiers achieved better F-1 scores without data augmentation as compared to using augmentation.

Kawahara et al. [93] employed CNNs to extract features and trained a linear classifier on them using 1300 images of Dermofit Image Library to perform 10-ary classification. A similar approach was used by Ge et al. [94] on MoleMap dataset to do 15-ary classification. Esteva et al. [31] used a pre-trained Inception v3 on around 130,000 images. Although their results for two binary-classification tasks are merely "on par with all tested experts", yet this work was the first credible proof of concept based on a large dataset that DL can make a practical contribution in real-world diagnosis. Following their steps, Haenssle et al. [30] pitched their fine-tuned Inception v4 model against 58 dermatologists after evaluating binary classification performance of their model on two test sets of size 100 and 300 only. The sensitivity and specificity of their DNN model were certainly higher than that of dermatologists' mean performance on two private test sets, however, their performance on publicly available International Symposium on Biomedical Imaging (ISBI) 2016 Challenge [95] test data is below the performance of the first two winning entries in that challenge.

The non-availability of high-quality, large, and publicly available skin lesion datasets makes realisation of reliable CAD systems significantly challenging. Deep learning methods trained on small datasets can produce very good results but are generally unpre-

dictable when tested on large datasets [96]. Many researchers make up for this paucity of large datasets by using private data. However, it makes it more difficult to reproduce the results and draw a fair comparison among various CAD systems. Ntoutsis et al. [97] note that in addition to the small size of publicly available datasets, another issue regarding their quality is that they do not represent the demographic distribution of various races in the dataset. This non-representation or under-representation of some demographic populations can induce racial bias in the dataset which can affect DL-based models to provide optimal predictions for such populations. To address the scarcity of available data for tracking and detecting skin diseases, Li et al. [98] developed a domain-specific data augmentation technique by merging individual lesions with full-body images to generate a large volume of synthetic data. Li and Shen [99] also used DNN to segment lesions, extract their dermoscopic features and classify them.

Cullell-Dalmau et al. [100] presented a pedagogical framework for implementing skin lesion classification model using CNNs. They provide a hands-on educational activity to understand and develop a DL-based classification model using commonly used APIs.

2.3 Curation of G1020 Dataset

We curated and published a new publicly available RFI dataset called G1020¹ for segmentation of OD and Optic Cup (OC) and detection of glaucoma. This dataset is curated by conforming to standard practices in routine ophthalmology and contains images taken under realistic conditions without many imaging constraints and, as a result, is fairly representative of real-world fundus imaging practices. We provided ground truth annotations for glaucoma diagnosis, OD and OC segmentation, bounding box coordinates for OD localisation, vertical Cup-to-Disc Ratio (CDR), and size of the neuroretinal rim in Inferior, Superior, Nasal and Temporal quadrants to see if ISNT rule is followed. We also provide a gold standard clinical diagnosis for glaucoma and many other ocular disorders. We believe that this challenging dataset can be used as a benchmark dataset to train robust algorithms for glaucoma detection capable of performing in the field or in clinics. We also report baseline results by conducting extensive experiments for automated glaucoma diagnosis and segmentation of optic disc and optic cup.

¹Available at: <https://www.dfki.uni-kl.de/g1020>

2.3.1 Description of G1020

G1020 database consists of 1020 high-resolution colour fundus images. The images are collected at a private clinical practice in Kaiserslautern, Germany between the years 2005 and 2017 with a 45-degree field of view after using dilation drops. The records were subsequently anonymised and random unique patient identifiers were assigned to each case. Because the images are collected retrospectively and are fully anonymised the informed consent of the patients was not required. To achieve a dataset that reflects routine clinical practice at busy healthcare facilities, no specific imaging constraints, like centring of OD or macula, were imposed. Figure. 2.1 shows the density map of OD in all images of G1020 as compared to the corresponding density map of OD in ORIGA. It can be seen that the images in G1020 have OD at a wider spatial area making post-processing of any segmentation algorithm significantly challenging. The images are stored in .JPG format. In the final dataset released, the black background is truncated and only the fundus region is preserved resulting in images of size between 1944×2108 and 2426×3007 pixels.

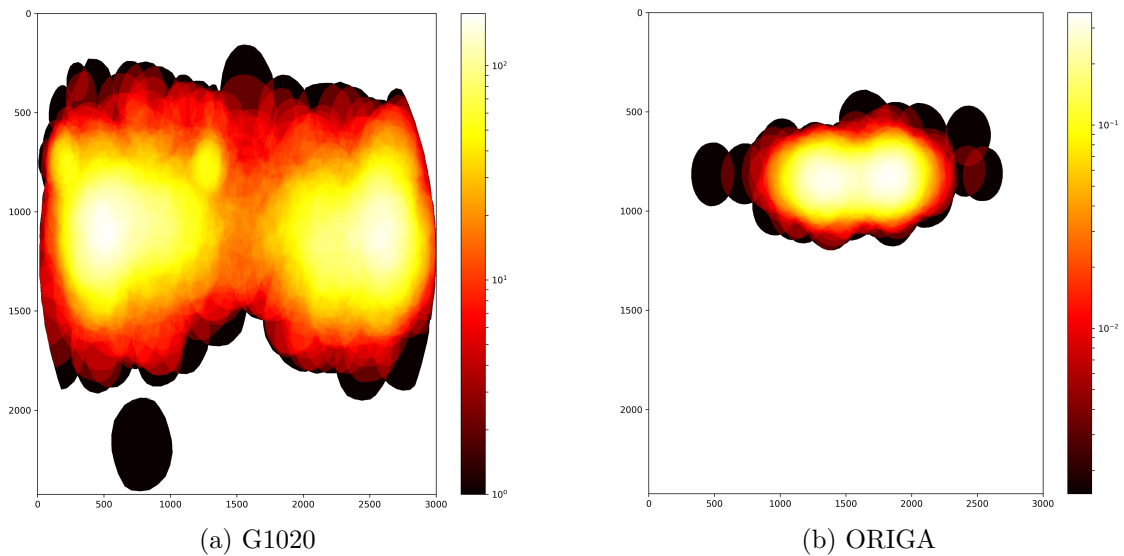


Figure 2.1: Density map of optic disc in G1020 and ORIGA. The optic discs in G1020 are not centralised, making post-processing of segmentation algorithms more challenging

The images of G1020 are taken from 432 patients. Each patient has a minimum of 1 image and a maximum of 12 images. Out of 1020 images, 296 images from 110 patients were found to have glaucoma and 724 images from 322 patients were healthy. There was no patient with images belonging to both healthy and glaucomatous classes.

Clinical diagnosis is provided for each patient with regards to the presence or absence of glaucoma and any other ocular disorder observed. To provide segmentation Ground Truth (GT), an expert marked OD and OC boundaries as well as bounding box annotations using *labelme* [101], which is an open-source annotation tool developed by MIT. These manual annotations are verified and corrected (if required) by a veteran ophthalmologist with more than 25 years of clinical experience. The annotations are saved in JSON files corresponding to each image. Based on the ground truth annotations for OD and OC, vertical CDR is calculated and the size of the neuroretinal rim in four quadrants is measured to see if ISNT rule is followed. In 60 glaucomatous images, OC was not visible whereas 170 healthy images also do not show any visible OC. In the absence of visible OC, the diagnosis was made using other clinical assessment and testing modalities. Fig. 2.2 shows sample images with OD, OC, and bounding box annotations.



(a) Sample image with all three annotations

(b) Sample image without optic cup

Figure 2.2: Sample images with optic cup (black polygon), optic disc (white polygon) and bounding box (red rectangle) annotations

2.3.2 Benchmark Results

2.3.2.1 Segmentation of OD and OC

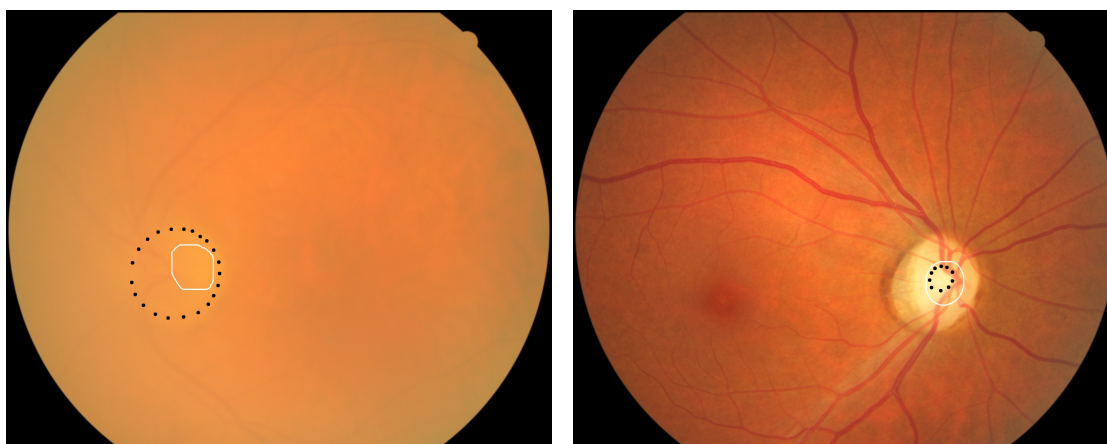
The state-of-the-art segmentation algorithms and image classification networks are evaluated on the G1020 dataset. Mask R-CNN [102] pre-trained on ImageNet [103] is used for automated segmentation of OD and OC with ResNet-50 [74] as convolutional backbone. Separate models are trained for the segmentation of OD and OC. First, Mask-RCNN is

trained using 80% random images from G1020 and tested on the remaining 20% images. The names of images in both training and testing splits are given with the dataset. Secondly, the model is trained using all images of ORIGA and evaluated on all images of G1020. Table 2.1 summarises segmentation results. Multiple criteria are employed to consider a detected OD and OC as correct or incorrect. Table 2.1 shows results for three such criteria, namely when Intersection Over Union (IOU) between predicted object and ground truth object is > 40 , 50 , or 60 .

Table 2.1: Segmentation performance of Mask R-CNN on G1020 dataset

Train/Test Splits	Object	Criterion	Average IOU	Precision	Recall	F1-Score
Train: G1020 (random 80%)	Optic Disc	IOU>0.4	0.8852	0.9951	0.9951	0.9951
		IOU>0.5	0.8852	0.9951	0.9951	0.9951
		IOU>0.6	0.8852	0.9951	0.9951	0.9951
Test: G1020 (random 20%)	Optic Cup	IOU>0.4	0.7276	0.9810	0.9810	0.9810
		IOU>0.5	0.7364	0.9494	0.9494	0.9494
		IOU>0.6	0.7645	0.8228	0.8228	0.8228
Train: ORIGA (all images)	Optic Disc	IOU>0.4	0.8641	0.9920	0.9774	0.9847
		IOU>0.5	0.8665	0.9861	0.9716	0.9786
		IOU>0.6	0.8719	0.9692	0.9549	0.9620
Test: G1020 (all images)	Optic Cup	IOU>0.4	0.6496	0.9071	0.9014	0.9042
		IOU>0.5	0.6809	0.7812	0.7762	0.7787
		IOU>0.6	0.7256	0.5489	0.5752	0.5770

To refine these segmentation results, Non-Maximum Suppression (NMS) is employed and all but one contour with the highest probability score is preserved. If the overlap (IOU) between a predicted object (OD or OC) and its ground truth is less than the criterion (IOU > 0.4 , for example), it's considered as both a False Negative (FN) since the actual object is not detected, and a False Positive (FP) since an object other than the actual object is predicted. For training and testing on G1020, the network was able to predict OC and OD for each image. In this experiment, there was only one image with IOU = 0.2689 below three criteria given in Table 2.1. The second minimum IOU was found to be 0.6429. Therefore, precision, recall, and F-1 score for all three criteria are the same. Furthermore, since the only misclassified image resulted in one FP and one FN, therefore, the values of precision and recall are also the same. For the experiment with training using ORIGA and testing on G1020, the network was able to detect 786 cups out of 791 actual cups and 1005 discs out of 1020 discs. Therefore, precision and recall are different in that experiment for each criterion. Figure 2.3 shows sample images with incorrectly detected OD and OC.



(a) Image with the smallest IOU ($= 0.2689$) between prediction and GT of OD

(b) Image with the smallest IOU ($= 0.308$) between prediction and GT of OC

Figure 2.3: Example images with incorrect OD and OC detection. Dotted annotations correspond to GT, whereas solid annotations represent prediction

Using correctly predicted OD and OC, CDR and size of the neuroretinal rim in inferior, superior, nasal, and temporal quadrants are predicted. Mean Absolute Percentage Error (MAPE) between various predicted values and ground truth values is given in Table 2.2. All the values in this table are calculated using $\text{IOU} > 0.5$.

Table 2.2: Mean Absolute Percentage Error (MAPE) of various parameters for correctly detected optic disc and optic cup. STD stands for Standard Deviation

Train/Test Split	Parameters	Mean	STD	
Train: G1020 (random 80%) Test: G1020 (random 20%)	Cup Diameter	0.2242	0.1933	
	Disc Diameter	0.0502	0.0664	
	CDR	0.2304	0.1852	
	Neuroretinal Rim	Inferior	0.1226	0.1002
		Superior	0.0206	0.0314
		Nasal	0.0880	0.0881
		Temporal	0.0669	0.0688
Train: ORIGA (all images) Test: G1020 (all images)	Cup Diameter	0.1396	0.1031	
	Disc Diameter	0.0593	0.0692	
	CDR	0.1674	0.1181	
	Neuroretinal Rim	Inferior	0.2102	0.2170
		Superior	0.2066	0.1278
		Nasal	0.2177	0.1933
		Temporal	0.2150	0.1483

2.3.2.2 Classification of Glaucoma

After localising and extracting ODs from the whole fundus images, these extracted discs are used to train Inception V3 [104] for classification of healthy and glaucomatous images. Cross validation with $k = 6$ is used with respect to patients to ensure that all images belonging to one patient are in either training set or validation set. The k-fold cross validation is a statistical method to ensure that the classifier’s performance is less biased towards a randomly taken train/test split. It is implemented by dividing the whole dataset into k , possibly equal, portions or folds. During a training iteration, one of these folds is kept aside for validation and the rest of $k - 1$ folds are used for training the model. In next training iteration a different fold is kept aside for validation and remaining $k - 1$ are used for training. This way, the train and test sets in each iteration are completely mutually exclusive. This process is repeated k times such that each of the k-folds is used for validation exactly once. This cross-validation approach provides a more realistic generalisation approximation. The inception model with the same experimentation setup was also used to classify ORIGA dataset using 5-fold cross validation. Performance of another custom CNN presented later in section 3.3 that gave state-of-the-art results on ORIGA was also evaluated for detection of glaucoma in G1020 dataset. Table 2.3 shows performance metrics for both classifiers on both datasets. It is evident from the table that both network were able to classify images from ORIGA with high precision and recall. However, those networks struggled hard against G1020. The difference between the performance of both networks on these two datasets could be correlated with the way these datasets are collected. ORIGA, and most other publicly available RFI datasets impose many constraints on imaging techniques and selection of images into final dataset so that the resulting image set is no longer representative of realistic image capturing practices. A DL model trained on such carefully curated datasets may have the ability to perform well in laboratory conditions but is likely to be unsuccessful in the field.

The Receiver Operator Characteristic (ROC) curve is a popular performance metric used to evaluate the discriminative ability of a binary classifier. It uses a varying threshold, on the confidence of an instance being positive, to measure the performance of the classifier by plotting *sensitivity* against *specificity*. Sensitivity or True Positive Rate (TPR) is defined as,

$$Sensitivity = \frac{TruePositives}{TruePositives + FalseNegatives} \quad (2.1)$$

Table 2.3: Performance metrics for glaucoma detection on G1020 and ORIGA.

Method	Dataset	Class	Precision	Recall	F1-Score
inception v3	ORIGA	Healthy	0.8578±0.0383	0.9170±0.0208	0.8861±0.0252
		Glaucoma	0.6947±0.0869	0.5581±0.1408	0.6157±0.1165
		Total	0.8157±0.0486	0.8246±0.0419	0.8164±0.0476
	G1020	Healthy	0.7150±0.1053	0.8183±0.0289	0.7587±0.0619
		Glaucoma	0.2894±0.0834	0.1920±0.0637	0.2219±0.0513
		Total	0.6055±0.0940	0.6344±0.0722	0.6080±0.0988
Bajwa et al. (2019) [105]	ORIGA	Healthy	0.8231±0.0288	0.9186±0.0229	0.8681±0.2460
		Glaucoma	0.6552±0.0665	0.4366±0.0495	0.5237±0.5340
		Total	0.7797±0.0378	0.7938±0.0342	0.7788±0.0366
	G1020	Healthy	0.4735±0.3348	0.6667±0.4714	0.5537±0.3916
		Glaucoma	0.0970±0.1373	0.3333±0.4714	0.1503±0.2126
		Total	0.3646±0.1979	0.5706±0.1976	0.4371±0.2162

Similarly, Specificity or True Negative Rate (TNR) is defined as,

$$Specificity = \frac{TrueNegatives}{TrueNegatives + FalsePositives} \quad (2.2)$$

The AUC of this ROC gives a quantitative measure to compare the performance of different classifiers. Fig. 2.4 shows Area Under ROC Curve (AUC) for each fold and their mean for both datasets. The network was able to achieve competitive AUC compared to state-of-the-art AUC results on ORIGA classification by Bajwa et al.[105] (AUC = 0.874) and Fu et al. [69] (AUC = 0.851), but suffered from serious performance degradation on G1020.

To provide deeper insight into the complexity of G1020 dataset and compare it with ORIGA, image embeddings of both datasets from the final convolutional layer of the inception model are deeply analysed. To do so, Principal Component Analysis (PCA) is applied on image embeddings to obtain two of the most significant principal components and the same are visualised on a 2D plane. Fig. 2.5 illustrates the results of PCA. It can be seen that glaucoma images (blue dots) and healthy images (red dots) are fairly separable in the ORIGA dataset. However, both classes have a huge overlap in the latent representation of the classifier trained on G1020 images.

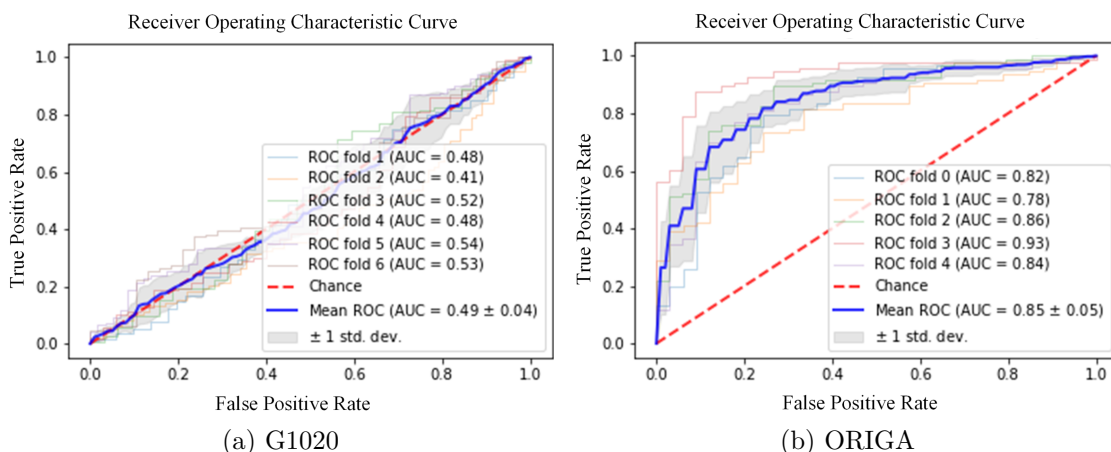


Figure 2.4: Receiver operating characteristic (ROC) and AUC for 6-fold cross-validation on G1020 and 5-fold cross-validation on ORIGA datasets using Inception V3

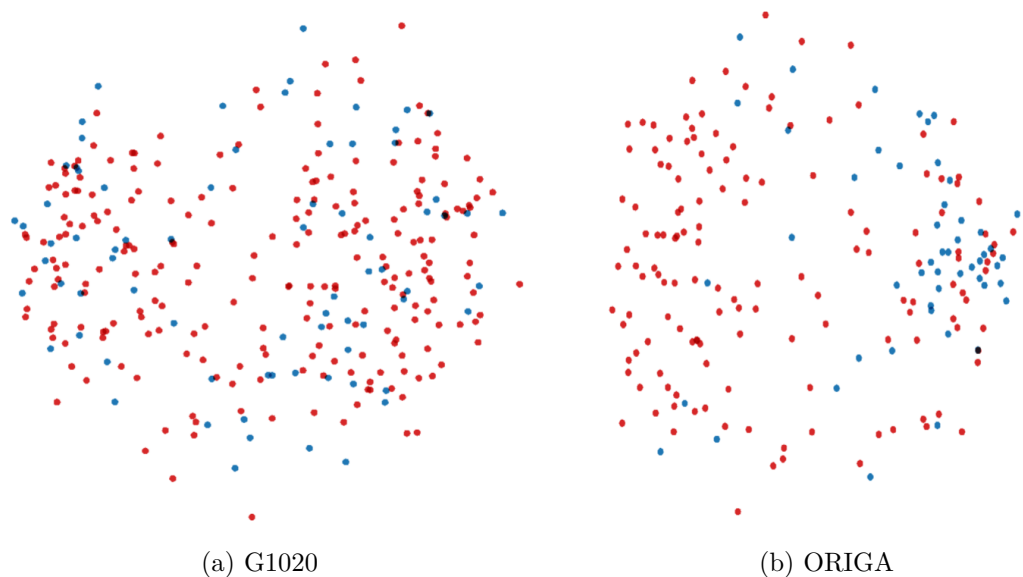


Figure 2.5: Visualisation of image embeddings learnt by DL model from G1020 and ORIGA datasets plotted on 2D plane after dimensionality reduction using PCA. Blue dots represent embeddings corresponding to glaucomatous images whereas red dots stand for embeddings of healthy images

2.4 Extending CAD to Clinically Relevant Skin Disease Detection

Large-scale manual screening for diseases is exhaustively laborious, extremely protracted, and severely susceptible to human predisposition and fatigue. CAD can provide a swift, reliable, and standardised diagnosis of various diseases with consistency and accuracy. CAD can also afford the opportunity of efficient and cost-effective screening and prevention of advanced tumour diseases to people living in rural or remote areas where expert dermatologists are not readily available. To make CAD systems practically more usable, this section extends previous works on CAD for dermatology by exploring the potential of DL to classify hundreds of skin diseases and improve classification performance. Various DNNs are trained on two of the largest publicly available skin image datasets, namely DermNet and ISIC Archive and it is shown that modern DL algorithms are capable of undertaking daunting tasks of recognising hundreds of, sometimes, seemingly similar skin lesions.

2.4.1 Datasets Used for Skin Lesion Classification

DermNet is a freely available dataset of around 23000 images gathered and labelled by Dermnet Skin Disease Atlas. Only 22501 images could be downloaded, however, and the links for the rest of them appeared to be inactive. This dataset provides diagnosis for 23 super-classes of diseases which are taxonomically divided into 642 sub-classes. However, there were some duplicate, empty and irrelevant sub-classes in the data. After pruning, 21844 images in distinct 622 sub-classes remained. The distribution of DermNet dataset used in this work is given in Table 2.4.

The second dataset is an online archive of around 24000 images divided into seven classes. The dataset is maintained by The International Skin Imaging Collaboration (ISIC). Their growing archive of high-quality clinical and dermoscopic images is manually labelled. The distribution of images in ISIC Archive-2018 dataset can be found in Table 2.5.

2.4.2 Experiments and Results

Many state-of-the-art DNN architectures that are used in this project are developed in recent years like residual networks, inception networks, densely connected networks, and frameworks facilitating architecture search. To cope up with the never-ending appetite of deep CNNs for data, the models used were pre-trained on ImageNet, which is a large

Table 2.4: Overview of DermNet dataset and the distribution of classes

Class Label	Abbreviation	Super-Class Name	Np. of Images	No. of Sub-Classes
0	ACROS	Acne and Rosacea	912	21
1	AKBCC	Actinic Keratosis, Basal Cell Carcinoma, and other Malignant Lesions	1437	60
2	ATO	Atopic Dermatitis	807	11
3	BUL	Bullous Diseases	561	12
4	CEL	Cellulitis, Impetigo, and other Bacterial Infections	361	25
5	ECZ	Eczema Photos	1950	47
6	WXA	Exanthems and Drug Eruptions	497	18
7	ALO	Alopecia and other Hair Diseases	195	23
8	HER	Herpes, Genetal Warts and other STIs	554	15
9	PIG	Pigmentation Disorder	711	32
10	LUPUS	Lupus and other Connective Tissue diseases	517	20
11	MEL	Melanoma and Melanocytic Nevi	635	15
12	NAIL	Nail Fungus and other Nail Disease	1541	48
13	POI	Poison Ivy and other Contact Dermatitis	373	12
14	PSO	Psoriasis Lichen Planus and related diseases	2112	39
15	SCA	Scabies Lyme Disease and other Infestations and Bites	611	25
16	SEB	Seborrheic Keratoses and other Benign Tumors	2397	50
17	SYS	Systemic Disease	816	43
18	TIN	Tinea Candidiasis and other Fungal Infections	1871	36
19	URT	Urticaria	265	9
20	VASCT	Vascular Tumors	603	18
21	VASCP	Vasculitis	569	17
22	WARTS	Common Warts, Mollusca Contagiosa and other	1549	26
Total			21844	622

dataset of around 1.5 million natural scene images divided into 1000 classes. These models were fine-tuned on dermatology datasets to leverage the benefits of transfer learning. From various CNN architectures explored for this task, eventually a few were selected including ResNet-152 [74], DenseNet-161 [75], SE-ResNeXt-101 [106], and NASNet [107] for their better performance. To report the final results, the potential of all of these biologically inspired neural networks is combined by taking an ensemble of their individual predictions. For performing ensemble the average of individual predictions of four best performing CNNs is used to output the final prediction.

It is important to note here that comparing researches that use different datasets,

Table 2.5: Overview of ISIC Archive dataset and the distribution of classes

Class Label	Abbreviation	Class	Np. of Images
0	AKIEC	Bowen Disease	334
1	BCC	Basal Cell Carcinoma	583
2	BKL	Benign Keratosis-like Lesions	1674
3	DF	Dermatofibroma	122
4	MEL	Melanoma	2177
5	NV	Melanocytic Nevi	18618
6	VASC	Vascular Lesions	157
Total			23665

different subsets or train/test splits of the same dataset is not scientifically correct. Since neither of the two datasets used in this work provided instructions on dividing the data into train and test sets, we used stratified k-fold cross-validation ($k = 5$ in this work) so that any future research can be compared with our work at least. For training, we randomly cropped the images with scale probability ranging between 0.7 and 1.0 while maintaining the aspect ratio. These cropped images are then resized to 224×224 pixels (for NASNet the input is resized to 331×331) before feeding them to the network. The images are also randomly flipped horizontally with a flip probability of 0.5. During testing, an image is cropped from four corners (top left, top right, bottom left, and bottom right) and one central crop of the required size. These cropped images are given to the classifier for inference and an ensemble of five predictions is taken to provide the final output. The initial learning rate is set to 10^{-4} and is halved every five epochs. The networks are trained for 20 epochs and 10 epochs for DermNet and ISIC Archive, respectively. The number of training epochs for each dataset and initial learning rate were determined empirically. To handle class imbalance, the weighted loss was used where the weight for a certain class equals the reciprocal of that class's ratio in the dataset.

2.4.2.1 Results on DermNet

As DermNet provides the opportunity to leverage taxonomical relationships among various diseases, therefore, for 23-ary classification the experiments were conducted in two ways. In the first experiment (Exp-1), the networks were trained on 23 classes and inferred on 23 classes. This is the most prevalent approach. With this experiment, Top-1 accuracy of $77.53 \pm 0.64\%$ and Top-5 accuracy of $93.87 \pm 0.37\%$ was achieved with $97.60 \pm 0.15\%$ AUC using an ensemble of four best models. In the second experiment (Exp-2), additionally given ontology in the dataset was utilised. The networks were trained on 622 classes but inferred on 23 classes only. The use of disease ontology information translates into the incorporation of expert knowledge into the network. This was implemented by summing the predictions of all sub-classes to calculate the prediction of respective super-class. This approach gave a noticeable boost in classifiers' performance. Top-1 accuracy of $79.94 \pm 0.45\%$ and Top-2 accuracy of $95.02 \pm 0.15\%$ was obtained with $98.07 \pm 0.07\%$ AUC using ensemble.

Top-N accuracy indicates the capability of a classifier to predict the correct class in the first N attempts. This metric gives a deeper insight into the classifier's learning and discriminating ability. The obtained results, of Exp-2 for example, show that the

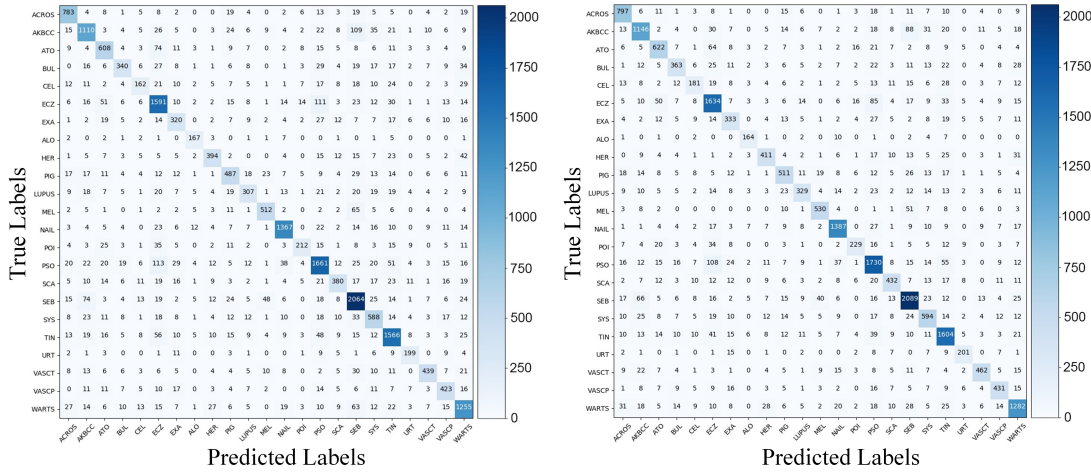
Table 2.6: Performance metrics for 23-Class classification of DermNet using ensemble. Exp-1¹ refers to training on 23 classes and testing on 23 classes without using disease ontology. Exp-2² refers to training on 622 classes and testing on 23 classes using disease ontology. Refer to Table 2.4 for full-form of class abbreviations

Class	Precision (%)		Sensitivity (%)		Specificity (%)		F-1 Score (%)	
	Exp-1 ¹	Exp-2 ²	Exp-1	Exp-2	Exp-1	Exp-2	Exp-1	Exp-2
ACROS	81.39	81.66	85.86	87.39	98.90	98.94	83.56	84.43
AKBCC	79.17	81.45	77.24	79.75	98.19	98.43	78.20	80.59
ATO	71.95	75.76	75.34	77.08	98.57	98.83	73.61	76.41
BUL	75.72	74.08	60.61	64.71	99.35	99.26	67.33	69.08
CEL	61.60	64.18	44.88	50.14	99.40	99.42	51.92	56.30
ECZ	75.19	78.41	81.59	83.79	96.69	97.24	78.26	81.01
WXA	62.99	65.17	64.39	67.00	98.88	98.97	63.68	66.07
ALO	76.96	81.19	85.64	84.10	99.70	99.78	81.07	82.62
HER	77.87	77.99	71.12	74.19	99.33	99.32	74.34	76.04
PIG	69.57	73.31	68.50	71.87	98.72	98.91	69.03	72.59
LUPUS	69.61	74.60	59.38	63.64	99.20	99.35	64.09	68.68
MEL	82.85	83.46	80.63	83.46	99.36	99.38	81.72	83.43
NAIL	89.64	89.08	88.71	90.01	99.00	98.95	89.17	89.53
POI	76.81	75.33	56.84	61.39	99.62	99.57	65.33	67.65
PSO	78.39	79.61	78.65	81.91	97.09	97.26	78.52	80.75
SCA	74.51	77.42	62.19	70.70	99.22	99.27	67.80	73.91
SEB	79.14	85.16	86.10	87.15	96.47	97.69	82.48	86.14
SYS	68.61	72.35	72.06	72.79	98.38	98.67	70.29	72.57
TIN	80.97	80.97	83.70	85.73	97.66	97.68	82.31	83.28
URT	75.67	78.21	75.09	75.85	99.62	99.68	78.38	77.01
VASCT	83.30	84.77	72.80	76.62	99.47	99.51	77.70	80.49
VASCP	72.43	77.24	74.34	75.75	99.03	99.26	73.37	76.49
WARTS	77.76	81.97	81.02	82.76	97.76	98.29	79.36	82.36
Weighted Average	71.81	79.82	77.53	79.94	98.14	98.40	77.34	79.80
Standard Deviation	06.46	05.89	11.20	09.83	00.95	00.75	08.42	07.72

model was able to predict the correct diagnosis out of 23 possible diseases in the first attempt with almost 80% accuracy. However, when allowed to make the 5 most probable predictions about a given image, the classifier achieved more than 95% accuracy. This means that even when the first prediction of the classifier is wrong, the actual correct prediction was high on the list of the next four predictions. Table 2.6 shows detailed performance metrics of 23-ary classification in both experiments.

Figure 2.6 shows that many reciprocatory misclassifications in Exp-1, like between Eczema (Abbreviated as ECZ in Figure 2.6) and Psoriasis Lichen Planus (PSO) and between Actinic Keratosis BCC (AKBCC) and Seborrheic Keratosis (SEB), are corrected to a large extent in Exp-2 by utilising taxonomical relationship among diseases.

2.4. EXTENDING CAD TO CLINICALLY RELEVANT SKIN DISEASE DETECTION



(a) Confusion Matrix for Exp-1 on DermNet (b) Confusion Matrix for Exp-2 on DermNet

Figure 2.6: Accumulated confusion matrix of 23-ary classification of DermNet dataset

Previous works on DermNet have generally opted for a subset of 23 super-classes for classification. Classification of 23-super classes in DermNet is performed by a few other research works as well. Haofu Liao [108] chose to classify all 23 classes and reported the best Top-1 accuracy of 73.1% and Top-5 accuracy of 91% on 1000 randomly chosen test images. Cícero et al. [109] reported Top-1 accuracy of 60% on 24 classes (they split “Melanoma and Melanocytic Nevi” into malignant and benign classes). They chose only 100 examples of each class for their test set.

Detailed literature survey on the classification of skin lesions using DermNet revealed that previously the classification task with the highest number of classes using DermNet has been performed by Prabhu et al. [110]. They performed 200-ary classification and obtained the highest Mean Class Accuracy (MCA) around 51%. However, this work took a step forward and tried to classify all 622 unique sub-classes to study the potential of DNNs in distinguishing among these skin lesions. Using similar experimental setup and DNN models, Top-1 accuracy of $66.74 \pm 0.64\%$ and Top-5 accuracy of $86.26 \pm 0.54\%$ was achieved with $98.34 \pm 0.09\%$ AUC. Small values of standard deviation in all of these results signify the stability and consistency of DNN classifiers’ performance.

2.4.2.2 Results on ISIC Archive-2018

ISIC Archive consists of high-resolution clinical and dermoscopic images. It does not provide any ontology information about the diseases. Therefore, the approach used in Exp-2 for DermNet cannot be applied here. Aforementioned experimental setup and

CNN architectures yielded Top-1 accuracy of $93.06\% \pm 0.31\%$ and Top-2 accuracy of $98.18\% \pm 0.06\%$ with $99.23\% \pm 0.02\%$ AUC using ensemble approach. Since this dataset has only seven classes, only Top-2 accuracy was calculated instead of Top-5 as was the case with 23 and 622 classes of DermNet. Table 2.7 shows that the ensemble of four classifiers was able to achieve high precision of over 80% for all classes except Vascular Lesions that can be justified by the small number of images (157 only) in this class. Confusion matrix showing the number of correctly classified and misclassified images per class in this dataset is shown in Figure 2.7.

Table 2.7: Performance metrics of ISIC Archive-2018 using ensemble

Class	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)
Bowen Disease (AKIEC)	80.43	78.74	99.71	79.58
Basal Cell Carcinoma (BCC)	91.85	86.96	99.79	89.34
Benign Keratosis-like Lesions (BKL)	85.55	77.48	98.95	81.32
Dermatofibroma (DF)	91.67	81.15	99.96	86.09
Melanoma (MEL)	84.64	66.05	98.75	74.20
Melanocytic Nevi (NV)	94.90	98.30	79.09	96.57
Vascular Lesions (VASC)	66.10	74.52	99.73	70.06
Weighted Average	85.02	80.46	96.57	82.45
Standard Deviation	09.10	09.38	07.15	08.38

The ISIC Challenges of 2016 [95] and 2017 [111] have focused on binary classification of skin lesions whereas ISIC Challenge 2018 [112] included seven classes. However, as shown in these experiments DL has an enormous capacity to discern far many diseases

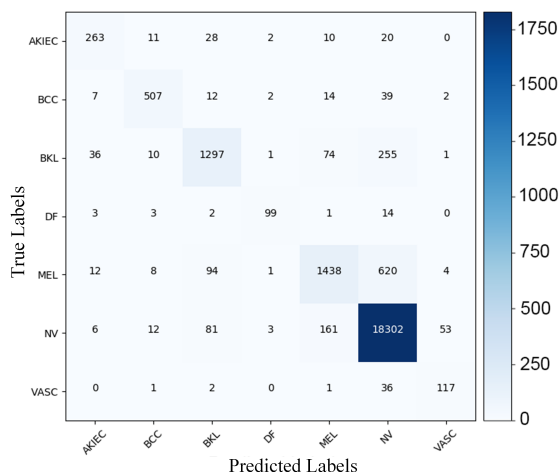


Figure 2.7: Confusion Matrix showing number of correctly classified and misclassified images per class in ISIC Archive-2018

with high sensitivity and specificity if given enough data. While reliable and accurate detection of melanoma is of utmost importance, because of its lethality, it might also be of interest for dermatologists to use CAD to detect other non-lethal skin diseases.

Figure 2.8 shows some examples of correct and misclassified images. It can be observed that some of these misclassified images had a very high correlation with other classes. For example, there is significantly small inter-class variance between Figure 2.8a and Figure 2.8e and between Figure 2.8d and Figure 2.8h. Therefore, CAD had a really hard time classifying those classes.

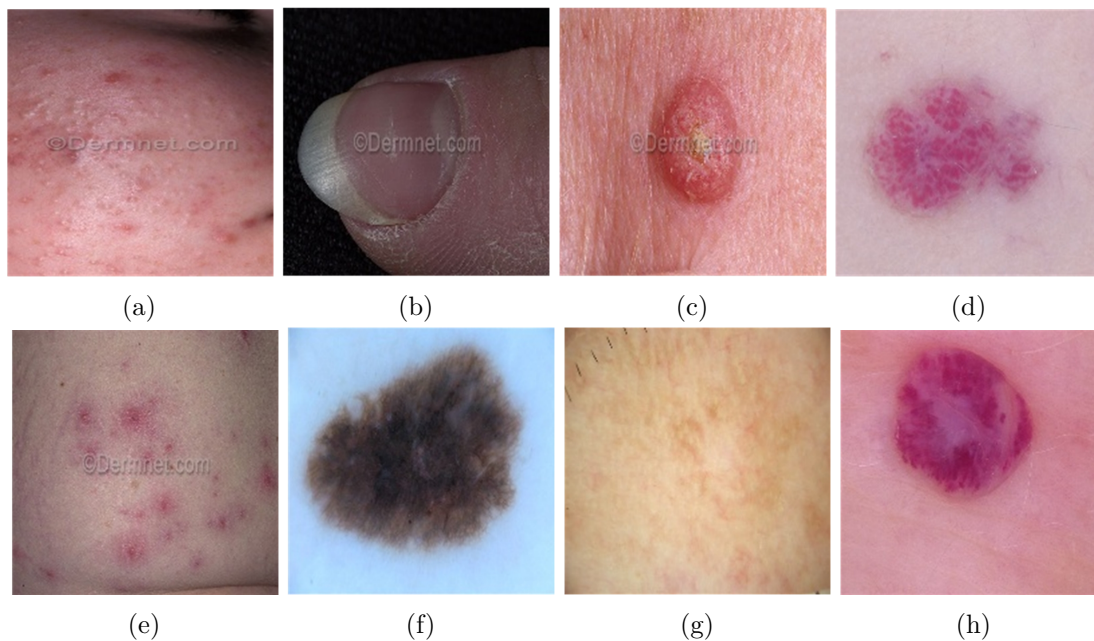


Figure 2.8: Examples of correctly and incorrectly classified skin diseases from ISIC Archive dataset. (a) Correctly classified ACROS in DermNet (b) Correctly Classified NAIL in DermNet (c) Correctly Classified SEB in DermNet (d) Correctly Classified VASC in ISIC (e) CEL Misclassified as ACROS in DermNet (f) Correctly Classified AKIEC in ISIC (g) BKL Misclassified as MEL in ISIC (h) NV Misclassified as VASC in ISIC. All images are resized to fit in square windows

2.5 Discussion

To make a CAD solution practically relevant in large-scale screening or routine clinical practices, it is inevitable that the CAD is trained using data that are representative of real-world image capturing practices. Most of the existing RFI datasets for glaucoma

detection are very small in size (a few hundred images) and almost all of them are collected in a controlled environment. These datasets do not consider practical limitations in imaging and usually exclude images that have other retinal artefacts [52]. It has been reported in the literature that the presence of multiple eye diseases degrades the performance of DL algorithms trained on such datasets [67]. Due to these reasons, most of the publicly available datasets for glaucoma detection are unable to train a robust CAD system that can perform equally well in a real clinical environment. In this chapter, a new large publicly available dataset of RFIs is presented that closely represents fundus imaging in practical clinical routine and does not enforce strict inclusion criteria on the captured images. The initial evaluation of various DL methods for OD and OC segmentation and glaucoma classification highlights challenges that need to be addressed to develop a practical CAD system for swift and reliable glaucoma screening. Obtained results set a baseline for comparison by future works in this domain. The research community is invited to utilise this dataset and evaluate its segmentation and classification algorithms.

Glaucoma is a multiplex disorder and sometimes requires more than one diagnostic modality, like fundoscopy, OCT, and VFT, to reliably and accurately identify its early stages. Fundus imaging is the preferred approach for quick screening for various ocular diseases including glaucoma due to its low cost and portability [113], however, relying on this single test does not provide dependable diagnosis either in field testing or in clinics. Therefore, a multi-modal classification model could be of interest that can process various testing modalities and makes a fairly informed decision regarding the presence or absence of glaucoma or other ocular disorders.

Another important factor to ease integration of CAD systems in the healthcare system as a Decision Support System (DSS) is to enable it to assist clinicians in identifying a wide range of highly prevalent diseases. Despite a lot of research focusing on classifying skin diseases using AI, most of these researches confine themselves to only binary or ternary classification [114–119] even when a large number of classes are available [120]. The importance of early detection of melanoma is understandable given the growing risk it poses to the patient’s survival with every passing day. However, there are thousands of other skin diseases [31] that might not be as fatal as melanoma but have an enormous impact on a patient’s quality of life. DL is extremely competent to take on hundreds of classes simultaneously, as evident by results recorded in this chapter. It is believed that this is the right time to harvest the potential of DL to its full extent and accelerate conducting impactful research that can translate into an industry-standard solution for automated skin disease diagnosis on a larger scale. These solutions can

have a far-reaching social impact by not only helping dermatologists with their diagnosis in a clinical setup but also providing an economical and efficient initial screening for underprivileged people in both developed and developing countries.

An important consideration in terms of the application of DL in medical image analysis is that many researchers either use private datasets or public datasets with their own choice of train/test splits (although randomly taken) and the number of classes. For this reason, there is little common ground, and often no ground at all, to compare various classification methods – as also noted by Brinker et al. [121]. This issue of non-comparability can be resolved by collecting and maintaining a standardised publicly available large dataset with explicitly specified train/test splits and standard performance metrics for benchmarking. Notwithstanding that some public datasets, like ISIC Challenges datasets, do provide this beforehand train/test split but their size is normally small and the task is usually restricted to binary or ternary classification. Any research on such small datasets cannot be reliably generalised and although the results are publishable, they cannot be used as a foundation stone for practical applications of AI in real-world diagnosis. On the other hand, large public datasets normally have a lot of noise, images with disgracefully low resolution, or are watermarked. Significant useful information required for fine-grained classification of seemingly similar diseases is lost in such low resolution or watermarked images. Additionally, non-visual metadata, like medical history, is not usually available with most medical image datasets. However, this additional information could be pivotal for a confident and accurate diagnosis. This project was able to utilise disease taxonomy for DermNet dataset and improved the results by 2.5% (refer to Table 2.6). If multi-modal datasets are curated and provided publicly, AI can surely leverage additional information to improve its classification performance.

While understanding and interpreting the results of any AI-based classifier it is important to realise that accuracy, or even sensitivity and specificity, might not portray the complete picture of a model’s performance. That is why AUC is also reported along with other performance metrics. From AI point of view, one might argue that achieving around 80% average sensitivity with 1.6% average false positive rate (Table 2.6, Exp-2) for 23-ary classification task using highly unbalanced datasets of low-resolution and watermarked images is a reasonable achievement. Nevertheless, the actual performance of any AI-based classifier can be significantly different in practical clinical setup as noted by Navarrete-Dechent et al. [122]. They found that the classifier developed by Han et al. [123] did not generalise well when presented with data from an archive of different demography than the one which was used to train the classifier. For a medical

practitioner, it is certainly a cause of concern. However, Han et al. advocated in their response [124] that a classifier should not be judged merely on the basis of sensitivity and specificity. The ROC curves indicate the true ability of a classifier to perform under a wide range of operating points or thresholds while making a diagnosis prediction for a given image. Varying this threshold from 0 to 1 on the model's output can change the trade-off between sensitivity and specificity and yield different accuracy. Therefore, higher AUC values ensure that the model has the ability to correctly predict a certain disease, for examples melanoma, with a minimum chance of classifying any other disease as that particular disorder.

Accuracy of CAD Systems

With the advancement of powerful image processing and machine learning techniques, CAD has become ever more prevalent in all fields of medicine. These computing methods have helped CAD evolve into a reliable DSS that can provide accurate and standardised large-scale screening of various image modalities to assist clinicians in identifying diseases. Today's CAD systems are expected to be at least at par with human counterparts in terms of accuracy. However, continuous efforts are still being exerted on finding new ways of making AI-based disease classifiers even more accurate and hence reliable. In addition to curating bigger medical image datasets and developing deeper DNNs, AI developers must work in close liaison with medical practitioners to understand their thought process and possibly follow that in their smart solutions.

This chapter focuses on improving the classification performance of DL-based medical image classifiers by taking advantage of domain knowledge from ophthalmology. Two retinal disorders that can be identified using RFIs are taken as example use cases and DL-based classification models are tailored around the way human experts analyse these images for detection of glaucoma and diabetic retinopathy.

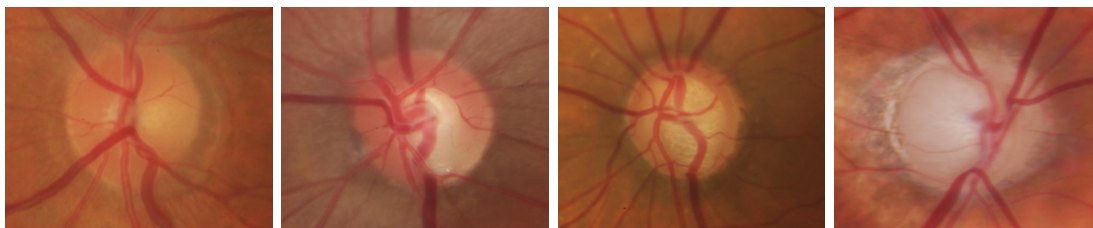
3.1 Domain Knowledge as a Means to Improve Accuracy

3.1.1 Understanding Glaucoma

Glaucoma is a syndrome of eye disease that leads to subtle, gradual, and eventually total loss of vision if left untreated. The disordered physiological processes associated with this disease are multifactorial. However, the causes of glaucoma are usually associated with the build-up of IOP in the eye that results from blockage of intraocular fluid drainage [125]. Although the exact cause of this blockage is unknown, it tends to be inherited and can be linked to old age, ethnicity, a steroid medication, and other diseases like diabetes and hypertension [126]. The increased IOP damages the optic nerve that carries visual information of photoreceptors from eye to brain. Generally, glaucoma does not show any signs or symptoms until it has progressed to an advanced stage at which point the damage becomes irreversible. It has been reported that the damage to optic nerve fibres becomes noticeable and a reduction in the visual field is detected when about 40% of axons are already lost [125]. However, it is possible to slow down the impairment caused by glaucoma if it is diagnosed sufficiently early. World Health Organisation recognised glaucoma as the third biggest cause of blindness after un-operated cataract and uncorrected refractive errors [127] and the leading cause of irreversible vision loss.

Glaucoma is normally diagnosed by obtaining the medical history of patients, measuring IOP, performing VFL test, and conducting a manual assessment of OD using ophthalmoscopy to examine the shape and colour of optic nerve [46, 128]. The optic disc is the cross-sectional view of the optic nerve connecting to the retina of each eye. It looks like a bright round spot in RFIs. In the case of glaucoma, the IOP damages the nerve fibres constituting the optic nerve. As a result, OD begins to form a cavity and develops a crater-like depression, at the front of the nerve head, called the optic cup. The boundary of the disc also dilates and the colour changes from healthy pink to pale. The CDR is one of the major structural image cues considered for glaucoma detection [129]. Figure 3.1 shows a healthy optic disc and its condition during various stages of glaucoma.

In retinal images, some of the important structural indications of glaucoma are disc size, CDR, the width of the neuroretinal rim in inferior, superior, nasal, and temporal quadrants (ISNT rule), and peripapillary atrophy (PPA) [131] etc. These indications are usually concentrated in and around OD. Therefore, segmentation of this ROI, that is detecting the contour of OD, is not only useful for a more focused clinical assessment by the ophthalmologists but also helpful in training a DL-based automated method for classification. However, automated glaucoma detection techniques based upon segmented



(a) Healthy optic disc image (b) Early glaucoma image (c) Moderate glaucoma image (d) Advanced glaucoma image

Figure 3.1: Stages of glaucoma in retinal fundus images taken from Rim-One dataset [130]

discs are very sensitive to the accuracy of segmentation and even a small error in the delineation of OD may affect the diagnosis significantly [132]. Localisation, on the other hand, gives the exact location of OD in the whole image with some surrounding context. Automatic methods for glaucoma detection based upon this approach of ROI extraction are more resilient to localization errors.

From an automated classification point of view, the disease pattern in retinal fundus images is inconspicuous and complex. Detecting ROI from natural scene images is comparatively easy because it has an obvious visual appearance, for example, colour, shape, and texture, etc. In contrast, the significant features of the disease in medical images are hidden and not readily discernible except by highly trained and qualified field experts. Since OD is the most important part of retinal fundus image for glaucoma detection, it is prudent to first detect and localise it before a thorough analysis is performed for the classification of healthy or glaucomatous images.

3.1.2 Understanding Diabetic Retinopathy

Diabetic patients are at constant risk of developing diabetic retinopathy that may eventually lead to permanent vision loss if left unnoticed or untreated. In such patients, increased blood sugar, blood pressure, and cholesterol can cause small blood vessels in the retina to protrude and, in due course, haemorrhage blood into retinal layers and/or vitreous humour [133]. In severe conditions, scar tissues and newly proliferated fragile blood vessels blanket the retina and obstruct incoming light from falling on it. As a result, the retina is unable to translate light into neural signals which results in blindness. Diabetic retinopathy advances slowly and gradually and may take years to reach the proliferative stage. However, almost every diabetic patient is potentially susceptible to this complication.

Timely diagnosis is the key to an appropriate prognosis. Ophthalmologists usually detect diabetic retinopathy by examining retinal fundus and looking for any signs of microaneurysms (bulging of blood vessels), blood leakage, and/or neovascularization [134]. While the indications of advanced stages of diabetic retinopathy are rather prominent, these symptoms remain largely discrete in the early stages. Figure 3.2 shows progress of diabetic retinopathy from healthy to proliferative stage in RFIs taken from EyePACS dataset. It can be observed from the figure that the difference between healthy and early stages of diabetic retinopathy is very subtle and not readily discernible. Manual analysis of these images requires highly qualified and specialised ophthalmologists who may not be easily accessible in developing countries or remote areas of developed countries.

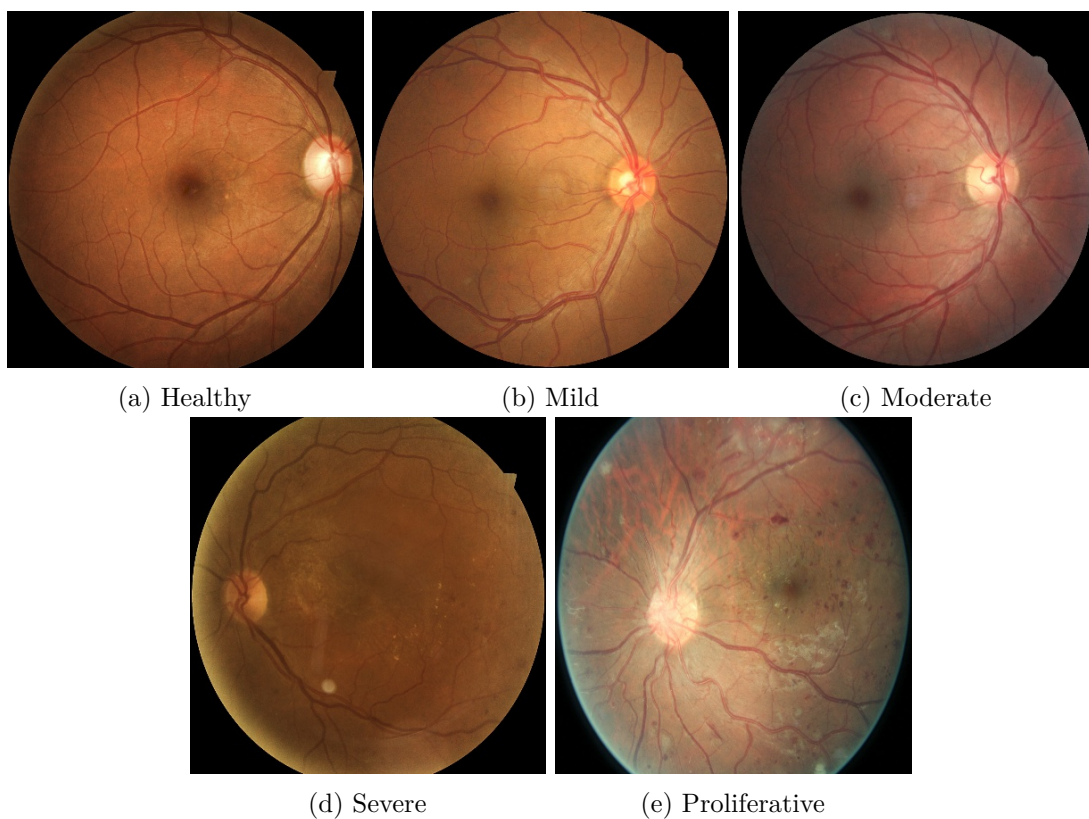


Figure 3.2: Progression of diabetic retinopathy from healthy to proliferative stage is subtle and gradual. Images are taken from EyePACS train set

Visual artefacts of early diabetic retinopathy in RFIs are usually small in size, inconspicuous, and scattered all over the retina. Detecting diabetic retinopathy requires physicians to visually scan the whole image and fixate on some specific regions to identify potential biomarkers of the disease. Therefore, getting inspiration from ophthalmolo-

gists, a DL-based model can be designed that combines coarse-grained classifiers, which detect discriminating features from the whole images, and a recent breed of fine-grained classifiers that discovers and pays special attention to pathologically significant regions.

3.2 Related Work

This section discusses various image processing, Machine Learning (ML), and DL approaches making use of various diagnostic criteria for optic disc localization, glaucoma identification, and diabetic retinopathy grading.

3.2.1 Optic Disc Localisation

Although OD can be spotted manually as a round bright spot in a retinal image, yet performing large-scale manual screening can prove to be tiresome, time-consuming, and prone to human fatigue and predisposition. Usually, the disc is the brightest region in the image. However, if ambient light finds its way into the image while capturing the photo it can look brighter than OD. Furthermore, occasionally some shiny reflective areas appear in the fundus image during image capturing. These shiny reflections can also look very bright and mislead a heuristic algorithm in considering them as candidate ROIs. There are many approaches laid out by researchers for OD localization exploiting different image characteristics. Some of these approaches are briefly covered below.

Intensity variations in the image can help locate OD in fundus images. To make use of this variation the image contrast is first improved using some locally adaptive transforms. The appearance of OD is then identified by noticing rapid variation in intensity as the disc has dark blood vessels alongside bright nerve fibres. The image is normalised and average intensity variance is calculated within a window of size roughly equal to the expected disc size. The disc centre is marked at the point where the highest intensity is found. Eswaran et al. [135] used such intensity variation based approach. They applied a 25×35 averaging filter with equal weights of 1 on the image to smooth it and get rid of low-intensity variations and preserve ROI. Chrástek et al. [136] used 31×31 averaging filter and the ROI is assumed to be 130×130 pixels. They used Canny Edge Detector [137] to plot the edges in the image. To localise the OD region they used only the green channel of RGB image. Abramoff et al. [138] proposed that the OD can be selected by taking only the top 5% brightest pixels and hue values in the yellow range. The surrounding pixels are then clustered to constitute a candidate region. The clusters which are below a certain threshold are discarded. Liu et al. [139] used

a similar approach. They first divided the image into 8×8 pixels grid and selected the block with a maximum number of top 5% brightest pixels as the centre of the disc. Nyúl [140] employed adaptive thresholding with a window whose size is determined to approximately match the size of the vessel thickness. A mean filter with the large kernel is then used with threshold probing for rough localization.

Another extensively used approach is threshold-based localization. A quick look at the retinal image tells that the OD is mostly the brightest region in the image. This observation is made and exploited by many including Siddalingaswamy and Prabhu [141]. It is also noticed that the green channel of RGB has the greatest contrast compared to red and blue channels [142–144], however, the red channel has also been used [145] since it has fewer blood vessels that can confuse the rule-based localization algorithm. The Optimal threshold is chosen based upon the approximation of the image histogram. The histogram of the image is gradually scanned from a high-intensity value I_1 , slowly decreasing the intensity until it reaches a lower value I_2 that produces at least 1000 pixels with the same intensity. It results in a subset of the histogram. The optimal threshold is taken as the mean of the two intensities I_1 and I_2 . Applying this threshold produces several connected candidate regions. The region with the highest number of pixels is taken as the OD. Dashtbozorg et al. [146] used Sliding Band Filter (SBF) [147] on downsampled versions of high resolution images since SBF is computationally very expensive. They apply this SBF first to a larger region of interest on downsampled images to get a rough localization. The position of this roughly estimated ROI is then used to establish a smaller ROI on the original sized image for a second application of SBF. The maximum filter response results in k -candidates pointing to potential OD regions. They then use a regression algorithm to smooth the disc boundary. Zhang et al. [148] proposed a fast method to detect the OD. Three vessel distribution features are used to calculate possible horizontal coordinates of the disc. These features are local vessel density, compactness of the vessels, and their uniformity. The vertical coordinates of the disc are calculated using Hough Transform according to the global vessel direction characteristics.

Hough Transform (HT) has also been widely utilised to detect OD [148–150] due to the disc's inherent circular shape and bright intensity. The technique is applied to binary images after they have undergone morphological operations to remove noise or reflection of light from the ocular fundus that may interfere with the calculation of Hough Circles. The HT maps any point (x, y) in the image to a circle in a parameter space that is characterised by centre (a, b) and radius r , and passes through the point (x, y) by following the equation of circle. Consequently, the set of all feature points in

the binary image are associated with circles that may almost be concentric around a circular shape in the image for some given value of radius r . This value of r should be known a priori by experience or experiments. Akyol et al. [151] presented an automatic method to localise OD from retinal images. They employ keypoint detectors to extract discriminative information about the image and the Structural Similarity (SSIM) index for textual analysis. They then used a visual dictionary and random forest classifier [152] to detect the disc location.

3.2.2 Glaucoma Classification

Automatic detection and classification of glaucoma have been widely studied by researchers since long. A brief overview of some of the current works is presented below. For a thorough coverage of glaucoma detection techniques using AI, [113, 153, 154] may be consulted.

Maheshwari et al. [155] used pre-trained AlexNet [103] on RIM-ONE dataset of RFIs to classify glaucoma. They split RGB images into their constituting Red, Green, and Blue channels and compute Local Binary Pattern (LBP) [156] on each of the three channels. These LBP images are used for training the classifier. During test time, individual channels of test images are fed directly to the classifier without LBP-augmentation, and the classifier’s predictions for each of R, G, and B channels are fused to get the final decision. Raghavendra et al. [157] used 1426 private RFIs to train and test an 18-layer DNN and achieved 95.6% accuracy, 95.5% sensitivity, and 95.7% specificity for glaucoma classification. In a large and comprehensive study using around 40,000 RFIs, Li et al. [53] evaluated the performance of inception v3 for detecting referable Glaucomatous Optic Neuropathy (GON). They defined GON as vertical CDR greater than 0.7. They achieved 92.9% accuracy and 98.6% AUC with 95.6% sensitivity and 92.0% specificity. They found that the leading reason for false-positive results was the presence of other eye conditions in the fundus images. Al-Bander et al. [158] used 455 images of RIM-ONE v2 dataset and extracted discriminating features using DNN before classifying them using SVM. They obtained 88.2% accuracy, 85% sensitivity, and 90.8% specificity.

R. Shinde [159] used a combination of image processing, ML, and DL methods to recognise glaucoma images. She used LeNet[160] for validating input images and bright spot algorithm for detecting ROIs. The OD and OC are segmented using UNet and finally, classification is performed using SVM, NN, and Adaboost classifiers. The CAD system is trained and evaluated on six small datasets, five of which are publicly available. The total number of images in the training and validation sets is merely 666. Fuente-

Arriaga et al. [161] proposed measuring blood vessels displacement within the disc for glaucoma detection. The authors first segment the vascular bundle in OD to set a reference point in the temporal side of the cup. Centroid positions of inferior, superior, and nasal vascular bundles are then determined which are used to calculate $L1$ distance between the centroid and the normal position of vascular bundles. They applied their method on a set of 67 images carefully selected for clarity and quality of the retina from a private dataset and report 91.34% overall accuracy. Ahmad et al. [162] and Khan et al. [163] have used almost similar techniques to detect glaucoma. They calculate CDR and ISNT quadrants and classify an image as glaucomatous if the CDR is greater than 0.5 and it violates the ISNT rule. Ahmad et al. applied the method on 80 images taken from DMED dataset, FAU data library, and Messidor dataset and reported 97.5% accuracy whereas Khan et al. used 50 images taken from the above-mentioned datasets and reported 94% accuracy. Though the accuracies reported by the aforementioned researchers are well above 90%, their test images are handpicked and so fewer in number that the results are not statistically significant and cannot be reliably generalised to large public datasets.

Xu et al. [164] formulated a reconstruction-based method for localising and classifying optic discs. They generate a codebook by random sampling from manually labelled images. This codebook is then used to calculate OD parameters based on their similarity to the input and their contribution towards the reconstruction of the input image. They report AUC for glaucoma diagnosis at 0.823. Noting that classification-based approaches perform better than segmentation-based approaches for glaucoma detection, Li et al. [165] proposed to integrate local features with holistic features to improve glaucoma classification. They ran various CNNs like AlexNet, VGG-16 and VGG-19 [166] and found that combining holistic and local features with AlexNet as the classifier gives the highest AUC at 0.8384 using 10-fold cross-validation, while the manual classification gives AUC equal to 0.839 on ORIGA dataset. Chen et al. [128] also used DNN based approach for glaucoma classification on the ORIGA dataset. Their method inserts micro neural networks within more complex models so that the receptive field has a more abstract representation of data. They also make use of a contextualisation network to get the hierarchical and discriminative representation of images. Their achieved AUC is 0.838 with 99 randomly selected train images and the rest for testing. In another of their publications, Chen et al. [46] used a six-layer CNN to detect glaucoma from ORIGA images. They used the same strategy of taking 99 random images for training and the rest for testing and obtained 0.931 AUC.

Franco et al. [167] designed and evaluated an automated glaucoma classifier based on

ResNet-50 using more than ten thousand RFIs from seven datasets. They achieved 95% accuracy and 91% AUC based on 50 cross-validation sets comprising of a total of 3551 images. A recent study by Wang et al. [168] performed multi-label classification of RFIs using EfficientNet [169]. Their model consists of EfficientNet-based feature extractor followed by an NN-based custom classifier for multi-label prediction. The models are trained using ODIR-2019 dataset and achieved 0.73 AUC and 0.88 F1-score on the test set. Al-Bander et. al [67] used deep learning approach to segment OC and OD from fundus images. Their segmentation model has a U-Shape architecture inspired from U-Net [68] with densely connected convolutional blocks, inspired from DenseNet [75]. They outperformed state-of-the-art segmentation results on various fundus datasets including ORIGA. For glaucoma diagnosis, however, in spite of combining commonly used vertical CDR with horizontal CDR, they were able to achieve AUC at 0.778 only. Similarly, Fu et. al [69] also proposed a U-Net like architecture for joint segmentation of OC and OD and named it M-Net. They added a multi-scale input layer that gets the input image at various scales and gives receptive fields of respective sizes. The main U-shaped convolutional network learns hierarchical representation. The so-called side-output layers generate prediction maps for early layers. These side-output layers not only relieve the vanishing gradient problem by back-propagating side-output loss directly to the early layers but also help achieve better output by supervising the output maps of each scale. For glaucoma screening on ORIGA data set, they trained their model on 325 images and tested on the rest of 325 images. Using vertical CDR of their segmented discs and cups they achieved 0.851 AUC.

3.2.3 Diabetic Retinopathy Grading

Recently, a large-scale study on detecting four stages of diabetic retinopathy, excluding healthy class, has been conducted by Dai et al. [170] using a private dataset of 666,383 images from 173,346 patients. Their proposed CAD system for diabetic retinopathy called DeepDR is evaluated on 200,136 private images and 9186 images from publicly available datasets and gives AUC for four classes in the range of 94% to 97% for private images and 91% to 97% for external images. In addition to diabetic retinopathy detection, they also detected microaneurysms, cotton-wool spots, hard exudates, and haemorrhages and obtained promising results. Mushtaq et al. [171] classified five stages of diabetic retinopathy using the Diabetic Retinopathy Detection 2015 dataset and APTOS-2019 datasets with the help of DenseNet-169. Data augmentation was used to obtain 7000 images per class and remove class imbalance. They have reported 90% validation accuracy and 80%

Cohen’s kappa score. However, they have not mentioned their train and test split or whether they trained the model on individual datasets or merged them. Their results also lack the confusion matrix necessary to perform in-depth error analysis and evaluate their model.

Welikala et al. [172] detected proliferative DR by identifying neovascularization. They used an ensemble of two networks trained separately on 100 different patches for each network. The patches are taken from a selected set of 60 images collected from Messidor [61] and a private dataset. Since the dataset had only 60 images they performed leave-one-out cross-validation and achieved 0.9505 AUC and sensitivity of 1 with the specificity of 0.95 at the optimal operating point. Wang et al. [173] identified suspicious regions in RFIs and classified diabetic retinopathy into normal (nDR) vs abnormal (aDR) and referable (rDR) vs non-referable (nrDR). They developed a CNN-based model called Zoom-in-Network to identify important regions. To classify an image the network uses the overview of the whole image and pays particular attention to important regions. They took 182 images from the EyePACS dataset and had a trained ophthalmologist draw bounding boxes around 306 lesions. On the Messidor dataset, they achieved 0.921 AUC, 0.905 accuracy, and 0.960 sensitivity at 0.50 specificity for nDR vs aDR.

Gulshan et al. [14] conducted a comprehensive study to distinguish rDR from nrDR grades. They trained a deep CNN on 128175 fundus images from a private dataset and tested on 9963 images from EyePACS-I and 1748 images of Messidor-2. They achieved an AUC of 0.991 on EyePACS-I and 0.990 on Messidor-2. Guan et al. [174] proposed that modelling each classifier after individual human grader instead of training a single classifier using average grading of all human experts improves classification performance. They trained 31 classifiers using a dataset of a total of 126522 images collected from EyePACS and three other clinics. The method is tested on 3547 images from EyePACS-I and Messidor-2 and achieved 0.9728 AUC, 0.9025 accuracy, and 0.8181 specificity at 0.97 sensitivity. However, it would have been more interesting if they had provided a comparison of their suggested approach with a simple ensemble of 31 networks modelled after average grading. Costa et al. [175] used adversarial learning to synthesise colour retinal images. However, the performance of their classifier trained on synthetic images was less than the classifier trained on real images. Aujih et al. [176] found that blood vessels play important role in disease classification and fundus images without blood vessels resulted in poor performance by the classifier.

The role of multiple filter sizes in learning fine-grained features was studied by Vo et al. [177]. They used VGG network with extra kernels and combined kernels with multiple loss networks. They achieved 0.891 AUC for rDR vs nrDR and 0.870 AUC for

normal vs abnormal on Messidor dataset using 10-fold cross-validation. Somkuwar et al. [178] performed classification of hard exudates by exploiting intensity features using 90 images from the Messidor dataset and achieved 100% accuracy on normal and 90% accuracy on abnormal images. Seoud et al. [179] focused on red lesions in RFIs, like haemorrhages and microaneurysms, and detected these biomarkers using dynamic shape features to classify DR. They achieved 0.899 AUC and 0.916 AUC for nDR vs aDR and rDR vs nrDR, respectively on Messidor. Rakhlin et al. [180] used around 82000 images taken from EyePACS for training and around 7000 EyePACS images and 1748 images from Messidor-2 for testing their deep learning-based classifier. They achieved 0.967 AUC on Messidor and 0.923 AUC on EyePACS for binary classification. Ramachandran et al. [181] used 485 private images and 1200 Messidor images to test a third-party DL-based classification platform, which was trained on more than 100000 images. Their validation gave them 0.980 AUC on Messidor dataset for rDR vs nrDR classification. Quellec et al. [182] capitalised a huge private dataset of around 110000 images and around 89000 EyePACS images to train and test a classifier for rDR vs nrDR grades and achieved 0.995 AUC on EyePACS. A comprehensive review on Diabetic retinopathy detection through deep learning techniques can be found in [183]

3.3 Two-Stage Framework for Glaucoma Classification

Providing whole RFI to an image classifier for glaucoma detection does not enable the model to concentrate on clinically significant ROI. Therefore, a two-stage framework is developed as shown in Fig. 3.3. The first stage is based on Regions with Convolutional Neural Network (R-CNN) and is responsible for localising and extracting OD from an RFI while the second stage uses DNN to classify the extracted disc into healthy or glaucomatous. Unfortunately, none of the publicly available retinal fundus image datasets provides any bounding box ground truth required for disc localization. Therefore, in addition to the proposed two-stage solution, a rule-based semi-automatic ground truth generation method is also developed that provides necessary annotations for training R-CNN based model for automated disc localisation.

For the automatic localisation stage of this framework, no fully automated disc localization method could be found at that time that could give robust and accurate results independent of the datasets. Also, many existing heuristic methods, for example, [184–186], set the bar for correct localization as low as accepting a predicted disc location correct if IOU between actual and predicted locations is greater than zero. To address these issues a dataset-independent fully automated disc localization method is proposed

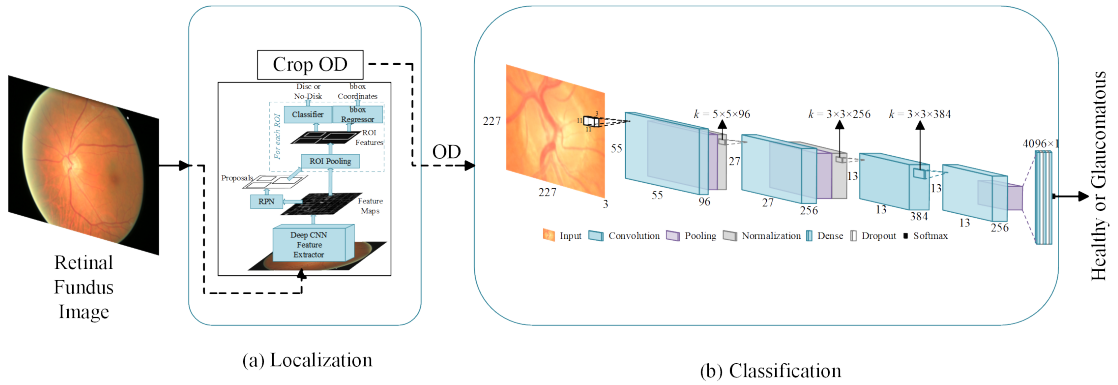


Figure 3.3: Complete framework of disc localization and classification. Detailed diagrams and description of both sub modules are given in their respective sections

based on faster R-CNN [70] as shown in Figure 3.3(a). This approach sets new state-of-the-art on six out of seven datasets for localization while setting the bar for correct localization at $\text{IOU} > 50$.

3.3.1 Datasets for Disc Localisation and Glaucoma Detection

It can be noticed from the brief introduction of publicly available RFI datasets used in this work given below that none of these datasets provide any bounding box ground truth for disc localisation, thus prompting for development of a new ground truth generation mechanism.

ORIGA(-light)

ORIGA [51] dataset already introduced in 2.2 aims to provide clinical ground truth to benchmark segmentation and classification algorithms. It uses a custom-developed tool to generate manual segmentation for OD and OC. It also provides CDR and labels for each image as glaucomatous or healthy. This dataset has been used as a standard dataset in some of the recent state-of-the-art researches for glaucoma classification.

HRF Image Database

High-Resolution Fundus [187] (HRF) Image database is provided by the Department of Ophthalmology, Friedrich-Alexander University Erlangen-Nuremberg, Germany. It consists of 15 healthy images, 15 glaucomatous images, and 15 images with diabetic

retinopathy. For each image, binary gold standard vessel segmentation is provided by a group of experts and clinicians.

OCT & CFI

This dataset [188] contains OCT and Colour RFIs of both eyes of 50 healthy persons collected at the Ophthalmology Department of Feiz Hospital, Isfahan, Iran. As the images were taken as part of a study on the comparison of macular OCTs in right and left eyes of normal people, it doesn't provide any ground truth for segmentation of OD or blood vessels, or OD localization.

DIARETDB1

Standard DIAbetic RETinopathy DataBase calibration level 1 (DIARETDB1) [78] is a publicly available dataset consisting of 89 colour RFIs taken at Kuopio University Hospital, Finland. The prime objective of this dataset is to benchmark the performance of automated methods for diabetic retinopathy detection. Four independent medical experts are employed to annotate the dataset and provide the markings for microaneurysms, haemorrhages, and hard and soft exudates. Based upon the markings provided, 84 of the images were found to have at least mild non-proliferative diabetic retinopathy while the rest of the five images were found to be healthy. The dataset does not provide retinopathy grades following International Clinical Diabetic Retinopathy (ICDR) severity grade or ground truth for OD localization.

DRIVE

Digital Retinal Images for Vessel Extraction (DRIVE) [79] consists of 40 images taken in the Netherlands as part of a diabetic retinopathy screening programme. The dataset is divided into train and test splits. Train set contains 20 images with manual segmentation masks for blood vessels. The test set also contains 20 images with two manual segmentation masks. This dataset also does not provide any annotation for OD localization.

DRIONS-DB

Digital Retinal Images for Optic Nerve Segmentation DataBase [189] commonly known as DRIONS-DB is a dataset for benchmarking ONH segmentation from retinal images. The data were collected at Ophthalmology Service at Miguel Servet Hospital, Saragossa,

Spain and contains 110 images. It provides ONH contours traced by two independent experts using a software tool for image annotation.

Messidor

Methods to evaluate segmentation and indexing techniques in the field of retinal ophthalmology (Messidor) [61] is a large publicly available dataset of 1200 high-resolution colour fundus images. The dataset contains 400 images collected from three ophthalmology departments each, under a project funded by the French Ministry of Research and Defence. It provides diabetic retinopathy grade for each image from 0 (healthy) to 3 (severe) as well as the risk of macular oedema at a scale of 0 (no risk) to 2 (high risk).

3.3.2 Localisation of Optic Disc

A heuristic method is developed to approximate the location of OD in retinal images. Results generated by this heuristic method are manually checked and necessary corrections are made where needed. Figure 3.4 depicts the workflow of this mechanism. It consists of a heuristic algorithm that gives a proposal for OD location which is then manually verified by an expert. This way localization ground truth for all seven datasets discussed in the previous section was generated.

Three publicly available datasets of high-resolution colour retinal fundus images were chosen to evaluate the performance of the heuristic localization algorithm. Table 3.1 gives an overview of the datasets used. Out of 780 images, 525 were randomly selected for training, 48 images were taken for validation and the rest of 207 images were kept aside for testing. The validation set was used to find various empirical parameters like retinal rim crop margin and maximum size of valid disc radius etc. The mixture of three different datasets introduces enough inter-dataset variations in the images to thoroughly and rigorously validate the accuracy and robustness of the heuristic method.

Table 3.1: Overview of datasets used for the evaluation of the heuristic method

Dataset	Total Size	Healthy	Glaucoma	Split		
				Train	Validate	Test
ORIGA	650	482	168	441	36	173
HRF	30	15	15	12	04	14
OCT&CFI	100	100	Nil	72	20	08
Total	780	597	183	525	48	207

3.3.2.1 Heuristic Algorithm for OD Localisation

This section details a heuristic algorithm to find approximate OD location from retinal images. The basic flow of the method is shown in Figure 3.4(a). It can be observed from RFIs that OD is usually the brightest region in the RFI. However, there could be other bright spots in the image, due to some disease or imperfect image capturing conditions, that can affect the performance of any empirical or heuristic method. Figure 3.5 shows two examples of such misleading bright spots.

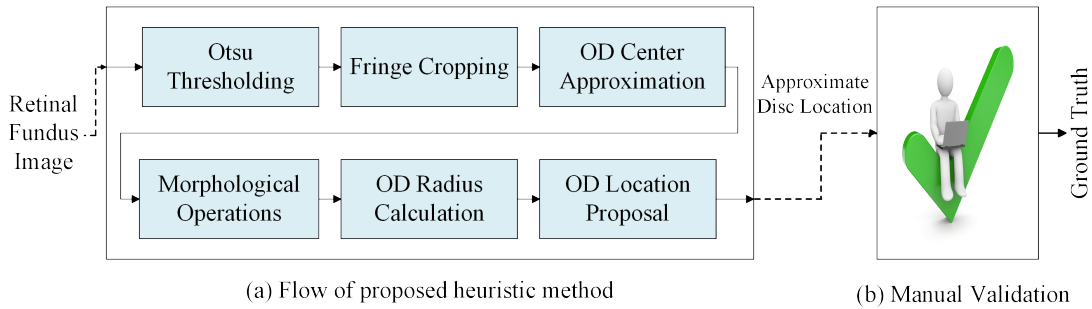
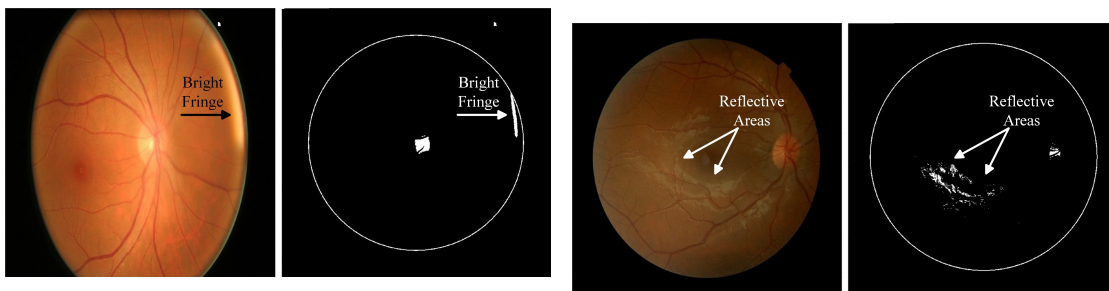


Figure 3.4: Workflow of semi-automatic ground truth generation mechanism

The first column of each subfigure shows colour RFI and the second column shows the binary image corresponding to the respective colour image. The bright fringe at the retinal rim, as shown in Figure 3.5(a), occurs when a patient does not place his/her eye correctly on the image capturing equipment and the ambient light gets through the corners of the eye. Figure 3.5(b) shows an example of shiny cloud-like spots around the macular region caused by the reflection of light from the ocular fundus which is a common phenomenon in younger patients.



(a) Binarisation of image with bright fringe at retinal rim

(b) Binarisation of image with reflection spots

Figure 3.5: Binary images showing misleading bright spots. RGB image in (a) is rescaled to fit in square window

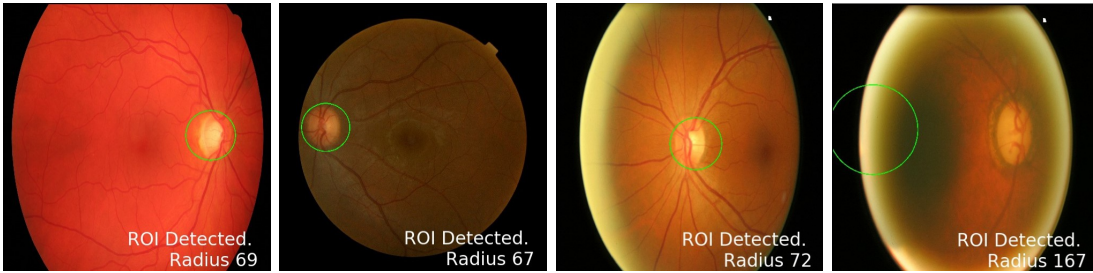
In this heuristic method, the fringe is removed by first finding the diameter of the retinal rim inside the image. This is done by applying Otsu thresholding [190] on the image. Otsu binarisation method assumes that the image consists of only two classes of pixels (foreground and background) following a bi-model histogram. It adaptively calculates the most appropriate threshold value that can categorise all the pixels into two classes. As a result, it turns the whole retina into a white disc and keeps the background black. This output is used to calculate the centre and radius of the retina. A circular mask with a radius less than the retinal radius is created and applied to the original image to crop and possibly get rid of the fringe.

A custom adaptive binarisation is then applied on the resultant image with a threshold for each image calculated as the mean of the top 1% brightest pixels. This technique locates the approximate core of OD. Before finding the centre of this approximate OD, morphological erosion operation is applied to remove small reflective areas and random impulse noise. This is followed by a dilation operation to connect disjoint white spots into fewer and bigger connected blobs. The result of these operations is a better approximation of OD. The radius and centre of this approximate disc location are then calculated and a circle with a radius greater than the calculated radius is drawn on the image to identify and localise OD. Lastly, these proposed locations are manually verified by an expert and necessary corrections are made where necessary.

Visual inspection of the output of train and test datasets showed that the method failed on only 3 out of 573 (test+validate) images and on only 1 of 207 test images from three different datasets which is shown in Figure 3.6. To quantify the accuracy of this approach IOU between bounding boxes given by the proposed method and manual ground truth is calculated. Table 3.2 shows the accuracy of this method in terms of overlap between predicted disc and actual disc. The results show that more than 96% of ODs are localised with more than 50% of actual disc present in the prediction. Also, about 52% of the predicted discs contain more than 70% of the actual disc. The average overlap between predicted disc area and ground truth for the test images is around 70%. It is also worth mentioning here that the minimum IOU of a correctly localised disc in this method is more than 20% whereas some researchers [184–186] have opted to consider their localization correct if the distance between predicted disc centre and actual disc

Table 3.2: Intersection Over Union (IOU) of heuristic predictions and the ground truth

IOU (%)	20	50	60	70	80
Test Accuracy	99.52	96.14	75.96	51.97	09.18



(a) Correct localisation of an HRF image (b) Correct localisation of an OCT&CFI image (c) Correct localisation of an ORIGA image (d) Incorrect localisation of ORIGA image

Figure 3.6: Results of Heuristic Localisation of OD. Subfigure 5(d) shows the only example where heuristic failed.

centre is less than expected disc diameter – in other words if $\text{IOU} > 0$.

3.3.2.2 Automated Disc Localisation

Although the results of the heuristic-based approach are very promising, yet they are dataset specific and might not work well in real-world scenarios on a diverse spectrum of fundus images. Therefore, a fully automated approach of precise disc localization without using any empirical knowledge about the dataset is also explored. Necessary corrections were made in the annotations given by the heuristic approach and these semi-automated annotations were provided to the automated localization method as the ground truth.

As shown in Fig. 3.7, the model consists of three major modules: Region Proposal Network (RPN), CNN classifier, and Bounding Box Regression. Given an image for object detection, RPN generates a number of random rectangular object proposals with associated objectness scores. These proposals are fed to the CNN that classifies whether a given object is present in the proposal. Then bounding box regression is performed to fit the rectangle closely to the object and provide the precise location of the object in the image.

For automated localisation of OD, the model was trained for 100,000 iterations using VGG16 as classifier pre-trained on Pascal VOC2007 [191]. The GT generated by our semi-automated method is used along with 573 images, previously employed for training and validation of the heuristic method, to train the network. The disc localisation outcome of faster R-CNN on three datasets, shown in Table 3.1, is given in Table 3.3. As can be seen in the Table, faster R-CNN gives 100% correct localization for 60% IOU and an average overlap of 97.11% on these three datasets combined.

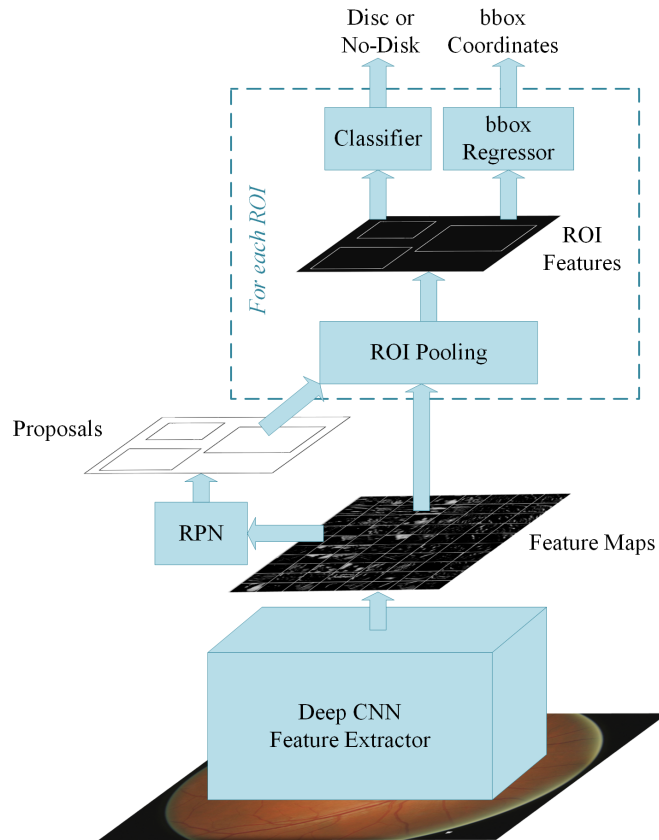


Figure 3.7: Internal components of faster R-CNN (Figure 3.3(a))

Once trained and evaluated on ORIGA, HRF, and OCT & CFI datasets, the model was also tested on other publicly available databases and the results are compared with some state-of-the-art methods developed specifically for those datasets. The results highlight the comparative performance of a fully automated method with state-of-the-art heuristic algorithms. The accuracies of our method are taken for 50% IOU. The results reported by [184–186] are for IOU > 0 whereas rest also have considered a localization correct if 50% overlap is achieved.

Table 3.3: Accuracy of automated disc localisation compared with heuristic method

Method	IOU (%)				
	20	50	60	70	80
Heuristic	99.52	96.14	75.96	51.97	09.18
Automated	100.0	100.0	100.0	99.52	94.69

Table 3.4: Generalisation performance of faster R-CNN on unseen datasets

Papers	Criterion (IOU >)	DIARETDB1 N = 89	DRIVE N = 40	DRIONS-DB N = 110	MESSIDOR N = 1200
Our Method (RCNN-based)	50	100.0	97.50	99.09	99.17
Giachetti et al. [184]	0	N/A	N/A	N/A	99.83
Yu et al. [185]	0	N/A	N/A	N/A	99.08
Aquino et al. [186]	0	N/A	N/A	N/A	98.83
Akyol et al. [151]	50	94.38	95.00	N/A	N/A
Qureshi et al. [192]	50	94.02	100.0	N/A	N/A
Godse et al. [193]	50	96.62	100.0	N/A	N/A
Lu et al. [194]	50	96.91	N/A	N/A	N/A

As can be seen from Table 3.4, the automated method performed significantly better than existing heuristics methods, which means that it was able to learn the discriminative representation of OD. It should be noted here that heuristics methods are normally designed with a particular dataset in focus. Figure 3.6 and Figure 3.8 show that there exists substantial variations in the colour, brightness, contrast, and resolution etc. among images of different datasets. The proposed fully automated method was not trained on any of the four datasets listed in table 3.4 and yet it performed superior to those methods tailored specifically for those individual datasets. The average overlap of predicted and actual OD bounding boxes is 84.65% for DIARETDB1, 84.13% for DRIVE, 80.46% for DRIONS-DB, and 84.82% for MESSIDOR.



(a) Sample image from DRIVE dataset (b) Sample image from DIARETDB1 dataset (c) Sample image from DRIONS-DB dataset (d) sample Image from Messidor dataset

Figure 3.8: Results of automated localization on different datasets. Notice the illumination and contrast variations amongst the datasets

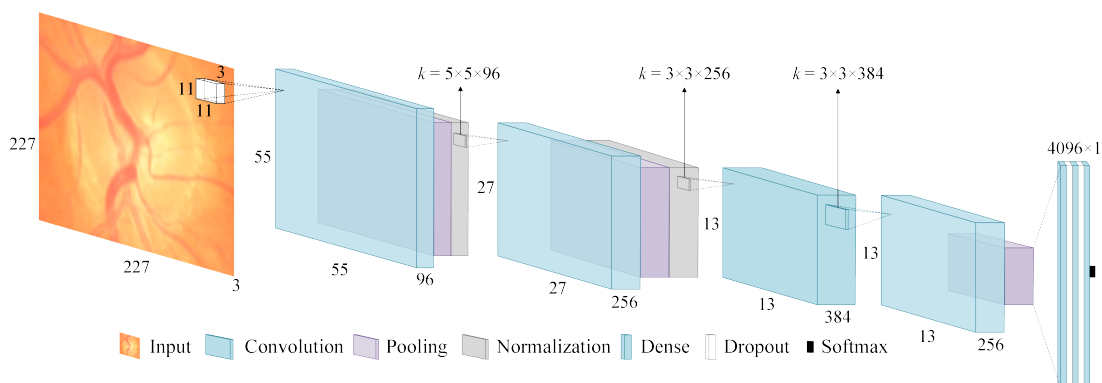


Figure 3.9: Convolutional neural network used for glaucoma classification (Figure 3.3b)

3.3.3 Classification of Glaucoma

In the first stage, the OD is extracted because most of the glaucoma-related information is contained in this region [46, 128]. Extracting this ROI not only produces a smaller initial image that is computationally efficient but also allows a DNN to focus on the most important part of the image. Figure 3.9 depicts the architecture of the CNN used in this work.

The network consists of four convolutional layers followed by three fully connected layers. Max pooling with overlapping strides and local response normalisation is used after the first two convolutional layers. Max Pooling also follows the fourth convolutional layer. The first two fully connected layers are followed by dropout layers with a dropout probability of 0.5. The output of the last dense layer is fed to the softmax function that gives prediction probabilities for each class.

3.3.3.1 Results of Classification

Due to class imbalance in the ORIGA dataset, as shown in Table 3.1, a stratified sampling technique is implemented where it is made sure that each batch for training contains some of the glaucoma images. This technique is used to prevent any bias towards the healthy class. Furthermore, a constant learning rate of 0.0001 along with Adam optimiser and Cross Entropy loss was used during training.

Results with Random Training

As no standard split of train and test set is available for this dataset, therefore, to compare the proposed model with other recently reported works the same training setup

Table 3.5: Precision, Recall and F1 score of classification with random train and test split

Class	Precision (%)	Recall (%)	F1 Score	No. of Samples
Healthy	81.12	94.9	0.8747	412
Glaucoma	69.57	34.53	0.4615	139
Total	78.21	79.67	0.7705	551

used by most of them [46, 128, 164] is used first. The model is trained repeatedly every time randomly taking 99 images for training and the rest for testing. From more than 1500 training runs the best combination of train and test split resulted in overall classification accuracy of 79.67%. Class-based average precision, recall, and F1 scores are tabulated in the Table 3.5.

Figure 3.10 shows the confusion matrix. It can be observed from the figure that out of 412 images without glaucoma, 391 are correctly classified and 21 such images are misclassified as having glaucoma. On the other hand, only 48 of the total 139 glaucomatous images are correctly classified and 91 images with glaucoma are incorrectly classified as healthy.

Table 3.6 shows the superiority of our model over other comparative studies in terms of AUC. Most of the works cited in Table 3.6 reported only AUC as a performance

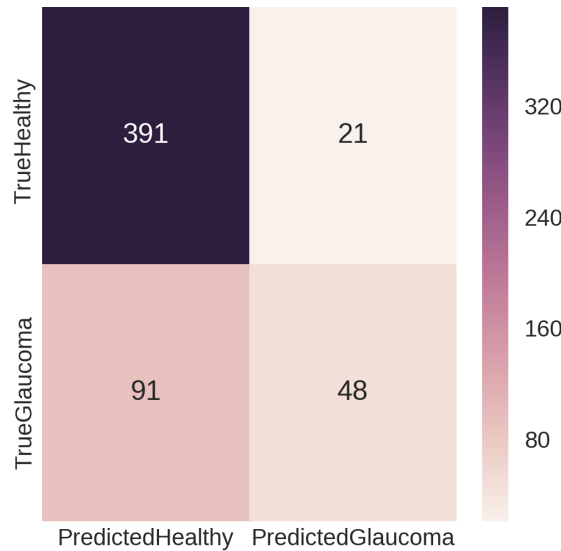


Figure 3.10: Confusion matrix showing the distribution of True Positives, False Positives, and False Negatives

metric for their classifiers. However, it was found during the experimentation that for some combinations of 99 train and 551 test images the model was able to achieve higher AUC, 84.87%, than four results in [46, 128, 164, 165] while predicting only healthy class for every test image leading to the healthy class recall of 1 and glaucoma class recall of 0. It means that the trade-off between sensitivity and specificity of the models can result in higher AUC without learning anything. Therefore, in the absence of clearly defined train and test splits and without knowing the proportion of healthy and glaucomatous images in both sets, AUC only may not depict the complete and true picture of a classifier. Other performance measures like precision, recall, and F1-scores should also be reported for a fair analysis and comprehensive comparison with other models. In the case of a well-defined train and test split, however, AUC alone might be enough to quantify the classification ability of a model.

Table 3.6: Comparison of obtained Area Under the Curve (AUC) with random train and test split

Performance Metric	[46]	[128]	[165]	[164]	[69]	Our Method
AUC	0.831	0.838	0.838	0.823	0.851	0.868

Results with Cross Validation

Realising this pitfall in performance evaluation of classifiers, and to facilitate future researchers in thorough comparison of their models, 10-fold cross-validation is performed on the dataset. The whole dataset was randomly divided into 10 equal portions. In one training session, for example, the first part is reserved for testing and the other nine are used for training. In the next session, the second part, for example, is kept aside for testing, and the rest of the nine are used for training. Average is taken over 10 training sessions and the accumulative test accuracy is found to be $79.39\% \pm 3.42\%$. Class-based precision, recall, and F1-score are tabulated in Table 3.7.

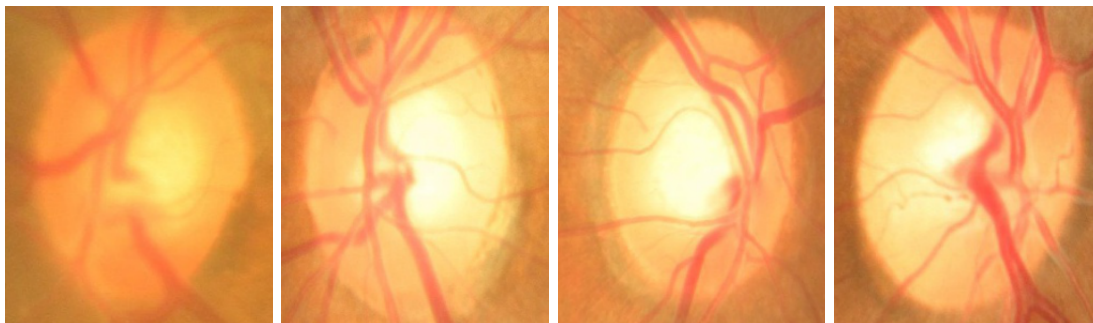
Table 3.7: Precision, Recall, and F1-score of classification with cross validation

Class	Precision (%)	Recall (%)	F1 Score
Healthy	82.31 \pm 2.88	91.86 \pm 2.29	0.8681 \pm 0.246
Glaucoma	65.52 \pm 6.65	43.66 \pm 4.95	0.5231 \pm 0.534
Total	77.97 \pm 3.78	79.38 \pm 3.42	0.7788 \pm 0.366

Table 3.8: Comparison of obtained Area Under the Curve (AUC) with cross validation. The sensitivity is calculated at observed specificity of 85%

Performance Metric	Chen et al.		Cheng et al.	Xu et al.	Fu et al.	Proposed Model	
	[46]	[128]	[165]	[164]	[69]	Random Training	Cross Validation
AUC	0.831	0.838	0.838	0.823	0.851	0.868	0.874
Sensitivity (%)	N/A	N/A	N/A	58	N/A	71	71.17

The comparison of AUC obtained using cross-validation with other works is summarised in Table 3.8 which clearly shows that the proposed network outperforms state-of-the-art results for glaucoma classification on ORIGA dataset. Fig. 3.11 shows sample images of correctly and incorrectly classified glaucoma and healthy images.



(a) Glaucoma Image Correctly Classified (b) Glaucoma Image Incorrectly Classified (c) Healthy Image Correctly Classified (d) Healthy Image Incorrectly Classified

Figure 3.11: Results of Glaucoma Classification using DCNN

Data augmentation was also performed to study its effects on the accuracy of classification. It was implemented by horizontal and vertical flips and cropping $227 \times 227 \times 3$ patches from four corners and centre of $256 \times 256 \times 3$ extracted images of OD. However, the experiments performed with and without data augmentation showed no significant difference between the performances of both approaches. The effect of network complexity on classification accuracy was also explored. For this purpose, Alexnet was used as the reference model and the impact of the number of layers on the network's performance was assessed while all the other conditions were the same. It was observed that increasing network complexity actually deteriorated the accuracy of the classifier. The reason for this performance degradation can be the small size of the dataset. Deeper networks have a habit of overfitting during training when not enough training samples

are provided. The networks working better than others had four convolutional layers. The best working model among all the different versions tried is used for classification and is the one shown in Figure 3.9.

Comparison with Later Researches

We compared the performance of our method with some of the research works that emerged after the publication of our results. It is obvious from Table 3.9 that over the years, this topic has attracted a lot of attention from the research community who have employed newer techniques on this task and have achieved superior results. However, this comparison should be analysed with extreme caution. Since changing the samples in train and test splits can have noticeable effect on the performance of any trained classifier, no two methods can be compared head-to-head unless they follow exactly the same train/test split. Whereas, Table 3.9 clearly shows that some researchers report their results without any mention of train/test split. The authors, who do mention this split, sometime, make the split as per their wishes instead of following any previous research or using conventional k-fold cross-validation. Even different instances of, for instance 10-fold, cross validation cannot be directly compared because of distinct possibility of folds composition.

Table 3.9: Performance comparison of our method with the later approaches using ORIGA dataset for glaucoma detection. Direct comparison between these methods is not fair due to different train/test splits.

Paper	Year	Train/Test Split	AUC	Sensitivity	Specificity
Our Method	Jun 2019	99/551 (random)	0.868	0.710	0.850
Our Method	Jun 2019	10-fold CV	0.874	0.712	0.850
Liao et al. [195]	Oct 2019	10-fold CV	0.880	N/A	N/A
Nazir et al. [196]	2020	unknown	0.940	0.941	0.945
Nazir et al. [197]	2021	455/195 (random)	0.970	0.963	N/A
Nawaz et al. [198]	2022	unknown	0.979	0.970	N/A
Doperlioglu et al. [199]	2022	unknown	0.951	0.977	0.926

3.3.4 Verification of Clinical Criteria for Glaucoma Detection

As mentioned above in the section 3.1, the majority of visual artefacts significant for glaucoma detection from RFI are centred around the OD. While examining a funduscopy image, ophthalmologists focus exclusively on this region to find out if glaucoma-related morphological changes can be spotted. This diagnostic criterion has been developed since

long and has become a standard protocol for identifying glaucoma using ophthalmoscopy. In this project, a small study is conducted to validate whether OD is indeed the integral part of RFI for the reliable identification of glaucoma.

To endorse the veracity of this clinically established criterion, the vital information in OD was systematically obscured and these manipulated images were used for training glaucoma classifiers. The obscuring of the disc achieved in two ways: masking the whole optic disc by replacing disc pixels with its mean value, and inpainting the disc from the outside to its centre as shown in Fig. 3.12.

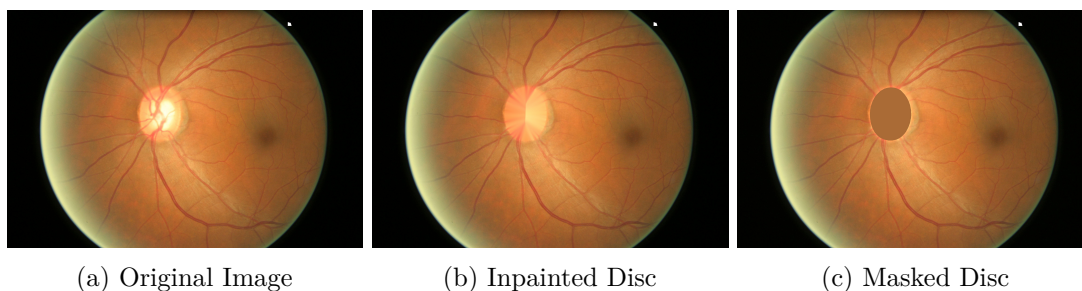


Figure 3.12: An original RFI with two variations to obscure optic disc

Pretrained Inception V3 was used to fine-tune the model on these doctored images since it was found to perform better in classification experiments discussed in section 2.3. For ORIGA dataset, 5-fold cross-validation was used for training and evaluation. Separate classifiers were trained for each type of manipulated images and the results are compared with classifiers trained with the whole RFI and with disc only. Table 3.10 clearly shows that the classification performance of the model noticeably drops when the most important region in the whole RFI for glaucoma detection is either masked or inpainted. The classifier trained using the whole unadulterated images gives relatively better performance. However, when the optic disc is extracted and a deep model is trained with only this part of the image, the performance of the model improves significantly.

3.4 Combined Coarse-and Fine-Grained Classifier for Diabetic Retinopathy Detection

Automated image recognition can be divided into coarse-grained classification and fine-grained classification. In the former case, images are classified into high-level categories like humans, animals, vehicles, and other objects in a natural scene, for example. In the

Table 3.10: Performance comparison of Inception V3 trained for glaucoma detection using different variations of RFIs

Input Image	Precision	Recall	F1-Score	AUC
Inpainted Disc	0.7014±0.03	0.6877±0.08	0.6538±0.05	65±0.07
Masked Disc	0.7018±0.06	0.7292±0.03	0.694±0.04	66±0.04
Whole Image	0.6474±0.09	0.7462±0.02	0.6701±0.06	67±0.04
Only Disc	0.7834±0.02	0.7969±0.02	0.7774±0.03	82±0.02

latter case, classification is focused on low-level categories like species of dogs or models of cars, etc. Fine-grained classification is particularly challenging owing to high intra-class variations and low inter-class variations. Since the difference between two consecutive grades of diabetic retinopathy is not always very obvious, as shown in Fig. 3.2, it could be treated as a fine-grained classification task, although it has normally been addressed using simple coarse-grained classification algorithms.

This section presents a combination of coarse-grained and fine-grained deep CNNs to analyse RFIs and predict automated diagnosis for diabetic retinopathy. The models used include two of the most popular conventional image classification architectures i.e. Residual Networks [74] and Densely Connected Networks [75], a network search framework called NASNet [107] and two methods for fine-grained classification namely NTS-Net [200] and SBS Layer [201]. This amalgamation of various types of models explores to draw on the combined potential of both fine-grained and coarse-grained approaches by training them separately and taking their ensemble during inference. Two commonly used RFI datasets for diabetic retinopathy grading named EyePACS and Messidor are used for evaluation. Since previous researches have used vastly disparate experimental setups, as evident from section 3.2, it is not possible to directly compare obtained results with most of them. However, a broad range of experiments are performed, following the most common problem settings in the literature like normal vs abnormal, referable vs non-referable, ternary and quaternary classification to set benchmarks, which can afford future works with an opportunity of fair comparison. In the following, details on the datasets used in this work and the ensemble methodology employed to perform classification are presented.

3.4.1 Datasets for Diabetic Retinopathy Detection

The EyePACS dataset is published publicly by Kaggle for a competition ¹ on diabetic retinopathy detection. Table 3.11 gives overview of EyePACS dataset. Although this dataset is very large in size, only about 75% of its images are of reasonable quality that they can be graded by human experts [180]. EyePACS is graded on a scale of 0 to 4 following International Clinical Diabetic Retinopathy (ICDR) guidelines [202]. However, the low gradeability of this dataset raises suspicions about the fidelity of labels provided with each image. The train set was pruned to get rid of 657 completely uninterpretable images. For evaluation on EyePACS, 33423 images were taken from the test set.

Table 3.11: Overview of EyePACS dataset. IrMA stands for IntraRetinal Microvascular Abnormalities

Severity Grade	Criterion	Train Set		Test Set	
		Images	Percentage	Images	Percentage
0	No Abnormalities	25810	73.48	39533	73.79
1	Microaneurysms Only	2443	6.95	3762	7.02
2	More than just microaneurysms but less than Grade 3	5292	15.07	7861	14.67
3	More than 20 intraretinal haemorrhages in each of 4 quadrants OR Definite venous beading in 2+ quadrants OR Prominent IrMA in 1+ quadrant AND no signs of proliferative retinopathy	873	2.48	1214	2.27
4	Neovascularization OR Vitreous/preretinal haemorrhage	708	2.02	1206	2.25
Total		35126	100	53576	100

As can be observed from Table 3.11, the data is highly unbalanced. About three-quarters of the images in the training set belong to the healthy category, which leaves only around 26% of the images for the classifiers to learn the minute details significant to discern among four grades of this disease.

Messidor dataset [61], publicly available since 2008, consists of 1200 high-resolution colour images of the posterior pole collected at three different ophthalmology departments in France. Each participating site contributed 400 images. This dataset is graded

¹Available at <https://www.kaggle.com/c/diabetic-retinopathy-detection>

for diabetic retinopathy on a scale of 0 to 3 following the criteria given in Table 3.12 and for macular oedema on a scale of 0 to 2. Two-thirds of the images were taken with pupil dilation and the remaining one-third were captured without pupil dilation. The Messidor dataset is carefully validated by experts and is, therefore, of higher quality than EyePACS in terms of both image resolution and labels.

Table 3.12: Overview of Messidor dataset showing grading criteria and class distribution

Severity Grade	Criterion	Images	Percentage
0	No microaneurysms AND No haemorrhages	546	45.50
1	Microaneurysms ≤ 5 AND No haemorrhages	153	12.75
2	$5 < \text{Microaneurysms} < 15$ AND $0 < \text{Haemorrhages} < 5$ AND No Neovascularization	247	20.58
3	Microaneurysms ≥ 15 OR Haemorrhages ≥ 5 OR Neovascularization	254	21.17
Total		1200	100

3.4.2 Methodology

Figure 3.13 illustrates the complete pipeline of the system combining coarse-grained and fine-grained classifiers. Before feeding an image to the network, it is preprocessed as shown in Fig. 3.14. First, Otsu Thresholding is applied to extract and crop retinal rim from RFI and get rid of the superfluous black background, as shown in Fig. 3.14a. Since the images in both datasets are taken with different cameras and under different clinical settings, they suffer from large brightness and colour variations. To compensate for that, adaptive histogram equalisation is used to normalise brightness and enhance the contrast of visual artefacts which are critical for diabetic retinopathy detection. Adaptive histogram equalisation can be applied on single-channel images, whereas the images provided in the datasets are in RGB colour space. Therefore, the images are first translated into YCbCr colour space to distribute all luminosity information in the Y channel and colour information in Cb and Cr channels. Then, adaptive histogram equalisation is applied on the Y channel only. This equalised Y channel is merged with Cb and Cr channels and the resultant image is converted back to RGB colour space. The effect of this contrast enhancement can be seen in Fig. 3.14b. The images are further normalised by subtracting local average colour from each pixel to highlight the

3.4. COMBINED COARSE-AND FINE-GRAINED CLASSIFIER FOR DIABETIC RETINOPATHY DETECTION

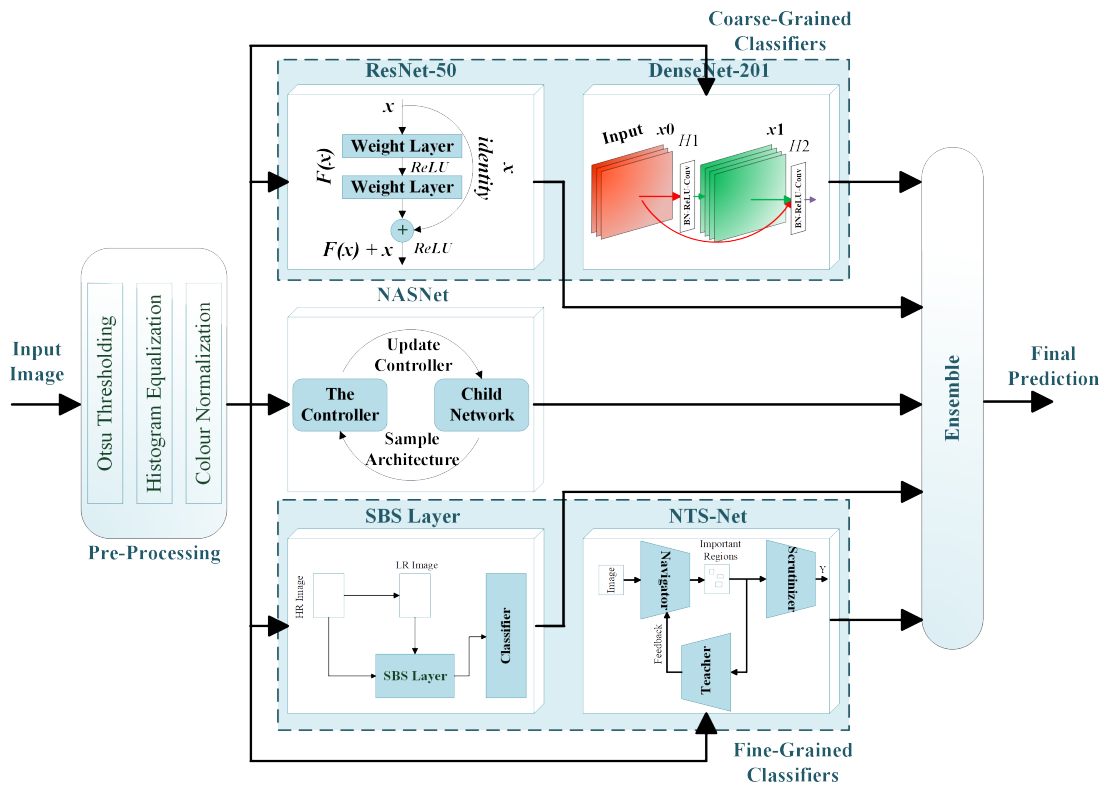
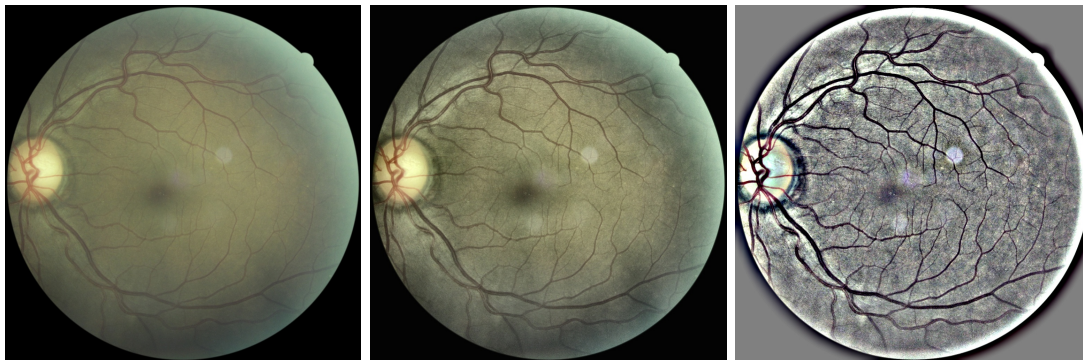


Figure 3.13: System Overview of combining coarse-grained and fine-grained classifiers

foreground and help our network detect small features, as shown in Fig. 3.14c. These pre-processed images are then used to train all five networks individually. During inference,



(a) Original Image before Pre-processing (b) After Adaptive Histogram Equalisation (c) After Local Average Colour Subtraction

Figure 3.14: Effects of preprocessing steps on retinal fundus images

each network gives its independent diagnoses which are ensembled to calculate the final prediction.

3.4.3 Experiments and Results

From the EyePACS train set, 30000 images are randomly selected for training, and the rest of the 4469 images are used for validation. Images from the test set of EyePACS are used for reporting evaluation results on this dataset. From Messidor, 800 images are used for training and 400 images from Lariboisière Hospital for testing (as done by Lam et al. [203]). A broad range of hyperparameters was explored during training. All networks are initialised with pre-trained weights and fine-tuned on ophthalmology datasets. To evaluate these models on EyePACS and Messidor datasets under similar problem settings, diabetic retinopathy grades of both datasets are first parallelised using criteria given in Figure 3.15.

EyePACS	Grade-0	Grade-1	Grade-2	Grade-3	Grade-4
Messidor	Grade-0	Grade-1	Grade-2	Grade-3	
Ternary	Grade-0	Grade-1	Grade-2		
Binary	Non-Referable		Referable		
Binary	Normal	Abnormal			

Figure 3.15: Conversion of five retinopathy grades in EyePACS to quaternary, ternary and binary classification

From section 3.2, it can be observed that previous works on EyePACS and Messidor have used the disparate train and test splits and different classification tasks, for example, Quaternary, Ternary, and Binary (rDR vs nrDR and nDR vs aDR). Furthermore, different researchers use different performance metrics to evaluate their methods. Therefore, in such scenarios comparison of any two works is not directly possible [204]. However, in this work, extensive experiments are conducted to perform all four classification tasks mentioned above and report comprehensive results to allow a rough comparison with some of the published state-of-the-art results on these datasets.

3.4.3.1 Results of Binary Classification

As discussed above, many previous works focus primarily on binary classification as nDR vs aDR or rDR vs nrDR grading. This binary classification is useful for large-scale screening programmes where the only objective is to screen for potential diabetic retinopathy patients, who can be directed to promptly consult a specialist for a thorough assessment of disease and appropriate treatment regimen. The criteria to convert 4 or 5 grades into binary grades is given in Fig. 3.15. For binary classification, the number of images used for training, validation, and testing from EyePACS and Messidor is given in Table 3.13 and Table 3.14. It can be seen from the tables that there is an extensive class imbalance between both classes.

Table 3.13: Class distribution for Normal vs Abnormal classification

Grade	EyePACS			Messidor		
	Train	Validate	Test	Train	Validate	Test
Normal	22668	2744	24741	346	49	151
Abnormal	7332	1725	8682	354	51	249
Total	30000	4469	33423	700	100	400

Table 3.15 provides detailed performance metrics for all classification tasks including nDR vs aDR and nDR vs aDR classification. These results show that for normal vs abnormal classification using Messidor, the proposed approach outperformed all three methods from the literature except for accuracy in which Wang et al. [173] performed slightly better. It should be noted here that Wang et al. performed 10-fold cross-validation and although their sensitivity of 96 is higher than 89.75 obtained with the proposed approach, it is calculated at 50% specificity while in these experiments it is calculated at 90% specificity. Therefore, it can be argued that getting a 90% true positive rate with less than 10% false positive is rate is better than having a 96% true positive rate with a 50% false positive rate.

Table 3.14: Class distribution for Referable vs Non-Referable classification

Grade	EyePACS			Messidor		
	Train	Validate	Test	Train	Validate	Test
Non-Referable	28825	4177	31937	453	65	181
Referable	1175	292	1486	247	35	219
Total	30000	4469	33423	700	100	400

Results of rDR vs nrDR classification can also be found in Table 3.15. All networks performed significantly better for this task than for normal vs abnormal classification on

Table 3.15: Detailed performance metrics for various classification settings

Results of Binary (Normal vs Abnormal) Classification								
Model	Accuracy (%)		AUC (%)		Sensitivity (%)		Specificity (%)	
	EyePACS	Messidor	EyePACS	Messidor	EyePACS	Messidor	EyePACS	Messidor
NTS-Net	88.19	88.00	92.72	95.20	88	88.00	72	87.51
SBS Layer	80.11	89.50	86.20	95.17	80	89.50	54	92.07
ResNet-50	82.86	87.75	89.46	95.06	83	87.75	75	90.49
DenseNet-201	82.66	87.75	89.69	95.89	83	87.75	77	88.14
NASNet	82.19	87.25	88.49	95.04	82	87.25	73	89.14
Ensemble	87.74	89.75	93.44	96.50	88	89.75	75	91.44
Vo et. al	N/A	87.10	N/A	87.00	N/A	88.2	N/A	85.7
Wang et. al	N/A	90.50	N/A	92.10	N/A	96	N/A	50
Soud et. al	N/A	N/A	N/A	89.90	N/A	N/A	N/A	N/A
Results of Binary (Referable vs Non-Referable) Classification								
NTS-Net	94.93	93.25	99.10	96.56	95	93	75	94
SBS Layer	95.89	88.75	99.44	94.90	96	89	67	90
ResNet-50	95.08	86.75	98.97	94.95	95	87	81	89
DenseNet-201	94.70	89.25	99.05	95.33	95	89	82	91
NASNet	91.98	87.50	97.45	95.16	92	88	85	89
Ensemble	95.34	89.25	99.23	96.45	95	89	81	91
Lam et. al	N/A	74.5	N/A	N/A	N/A	N/A	N/A	N/A
Vo et. al	N/A	89.70	N/A	89.10	N/A	89.3	N/A	90
Wang et. al	N/A	91.10	N/A	95.70	N/A	97.8	N/A	50
Seoud et. al	N/A	74.5	N/A	91.60	N/A	N/A	N/A	N/A
Results of Ternary Classification								
NTS-Net	84.43	84.50	94.89	94.61	84	85	72	94
SBS Layer	76.93	84.50	90.95	94.12	77	85	50	91
ResNet-50	81.23	80.50	93.51	93.79	81	81	74	92
DenseNet-201	79.20	80.25	92.87	94.25	79	80	77	93
NASNet	78.95	81.75	91.93	94.00	79	82	71	89
Ensemble	84.94	85.25	95.28	95.40	85	85	73	92
Lam et. al	N/A	68.8	N/A	N/A	N/A	N/A	N/A	N/A
Results of Quaternary Classification								
NTS-Net	82.53	74.50	95.72	91.84	83	75	76	92
SBS Layer	82.00	65.00	95.69	88.43	82	65	67	88
ResNet-50	81.82	70.25	95.53	91.31	82	70	71	89
DenseNet-201	79.38	74.00	95.04	92.26	79	74	75	91
NASNet	73.73	71.75	92.06	90.84	74	72	74	86
Emsemble	83.42	76.25	96.31	92.99	83	76	73	91
Lam et. al	N/A	57.2	N/A	N/A	N/A	N/A	N/A	N/A

EyePACS dataset reaching maximum accuracy around 96% with 99.44% AUC using SBS layer architecture. For Messidor dataset, both fine-grained classifiers i.e. NTS-Net and SBS Layer stand out from the coarse-grained classifiers. NTS-Net outperforms all other methods in all metrics, whereas ensemble of all methods gives sub-optimal performance

than individual fine-grained methods. This can happen when the majority of classifiers used for ensembling have a skewed performance towards the downside and only a few give standout results.

3.4.3.2 Results of Multi-Class Classification

In addition to serving as a screening tool at primary healthcare facilities, CAD can also be used in secondary and tertiary healthcare establishments to provide a more precise diagnosis. Therefore, the complexity of the classification task was gradually increased from binary to ternary and eventually quaternary classification. Table 3.16 and Table 3.17 show the class distribution in train, validation and test splits for this multi-class setting for both datasets. For ternary classification, conversion criterion used by [203, 205] is used in these experiments, as shown in Fig. 3.15.

Table 3.16: Class distribution for 4-Class classification

Grade	EyePACS			Messidor		
	Train	Validate	Test	Train	Validate	Test
0	22668	2744	24741	346	49	151
1	6157	1433	7196	107	16	30
2	685	166	753	155	22	70
3	490	126	733	92	13	149
Total	30000	4469	33423	700	100	400

Performance of individual networks and their ensemble for ternary and quaternary classification is also summarised in Table 3.15. Ensemble of all models gave a better performance in this case. It can also be observed that the performance of NTS-Net is higher than all other individual networks. The accuracy values for both ternary and quaternary classification are superior to the values reported by Lam et al. [203].

Table 3.17: Class distribution for 3-Class classification

Grade	EyePACS			Messidor		
	Train	Validate	Test	Train	Validate	Test
0	22668	2744	24741	346	49	151
1	6157	1433	7196	107	16	30
2	1175	292	1486	247	35	219
Total	30000	4469	33423	700	100	400

Figure 3.16 shows confusion matrices providing a detailed overview of classification performance of ensemble in multi-class scenarios. These confusion matrices are consis-

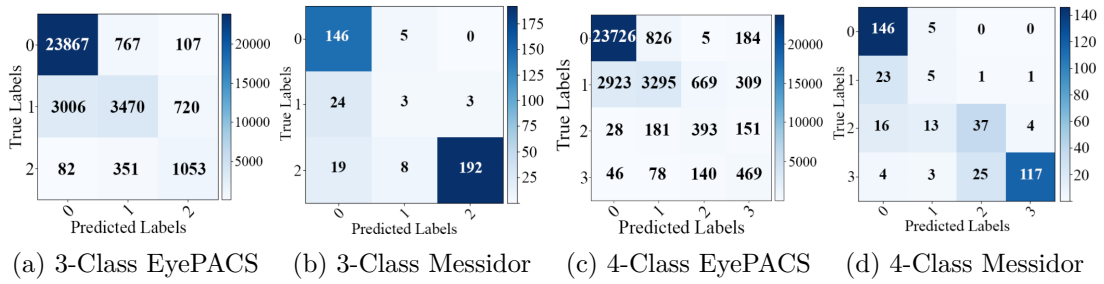


Figure 3.16: Confusion matrices for EyePACS and Messidor for multi-class classification tasks

tent with challenges of identifying grade 1 and 2 diabetic retinopathy reported in the literature [203]. For example, in ternary classification the sensitivity for healthy, mild, and severe diabetic retinopathy is 96.5%, 48.2%, and 70.9% respectively for EyePACS (Fig. 3.16a) and 96.7%, 10%, and 86.7% respectively for Messidor dataset (Fig. 3.16b). These statistics conform to established challenges in precisely grading diabetic retinopathy, where grading of the stages on the opposite spectrum of this progressive disease has higher sensitivity than intermediate stages. Furthermore, mild diabetic retinopathy appears to be misclassified with healthy images more frequently (80.7% and 88.9% for EyePACS and Messidor respectively) compared to misclassified with severe disease. This trend can be explained by the progressive nature of such disease, where visual artefacts start very slowly and unremarkably and manifest gradually into noticeable features.

Similarly, for quaternary classification the recall for all four grades starting from healthy class are 95.9%, 45.8%, 52.2%, and 64% for EyePACS dataset (Fig. 3.16c) and 96.7%, 16.7%, 52.8%, 78.5% for Messidor (Fig. 3.16d). Once again it can be seen that the disease grades at the opposite sides of the spectrum have higher sensitivities compared to intermediate stages. Low recall of mild class in Messidor data in both ternary and quaternary classification tasks can be attributed to the infinitesimally small number of samples in those classes. Figure 3.17 shows ROC curves for all four classification tasks for both datasets and all individual classifiers as well as their ensemble.

3.5 Capitalising Non-Visual Metadata to Improve Classification Accuracy

Many DL-based classification and diagnosis models base their decisions solely on images and ignore non-visual metadata such as sex, age, and ethnicity of the patients and anatomical location of the skin lesions, for example. Although the availability of datasets

3.5. CAPITALISING NON-VISUAL METADATA TO IMPROVE CLASSIFICATION ACCURACY

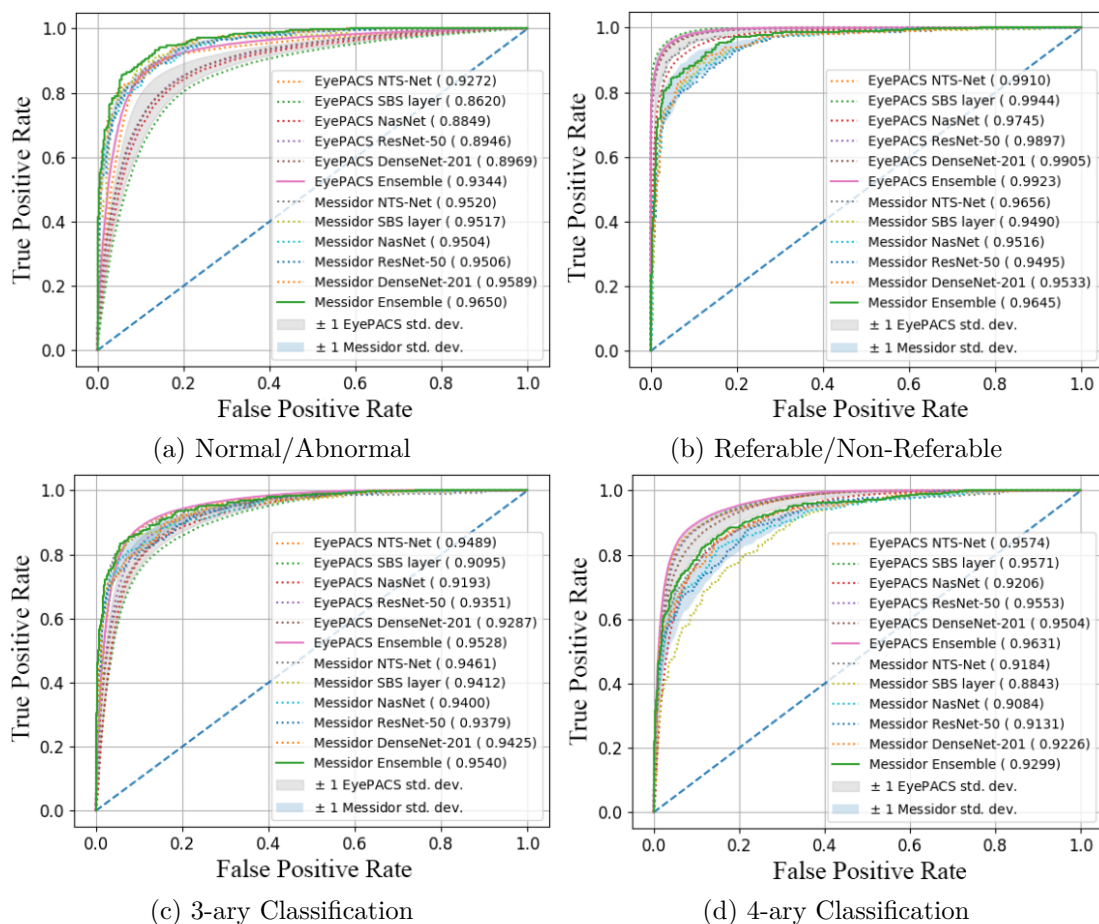


Figure 3.17: ROC Curves for all classification tasks

that provide this metadata is still a challenge, recent works on this issue show that incorporating metadata, when available with medical images, in DL-based diagnosis systems may result in improvement of classification and prediction accuracy of DNNs [206]. For instance, Kharazmi et al. [207] demonstrate that data-driven feature learning by combining dermoscopic images with patient’s profile, which consists of patient’s age, sex, location of the lesion and its size and elevation, improves sensitivity, specificity, and accuracy of detection of Basal Cell Carcinoma (BCC). However, the prime question regarding incorporating metadata into deep learning models is about the way these metadata should be processed and modelled [208]. Pachecho et al. [209] unsuccessfully tried to apply Naive Bayes and Decision Tree on metadata but could not achieve any performance improvement for their skin lesion classification task. Mitani et al. [44] also exploited metadata along with 114205 RFIs from a private dataset to detect anaemia

but could not achieve any significant and noticeable boost in their model’s performance.

To explore ways of incorporating non-visual clinical data with medical images, ISIC-2020 dataset ² is selected. This dataset consists of around 33000 training images and around 11000 test images. The task is to classify between melanoma and benign skin lesions. The ratio of benign lesions to melanoma in the training set is almost 98.25:1.75, so the data is heavily unbalanced. In addition to image-level disease labels, three types of clinical metadata are available, namely the age and sex of the patient and the anatomical location of the lesion. For many data points, one or more of these metadata are missing, which are replaced by the ‘unknown’ category during metadata preprocessing. To set a baseline for comparison with metadata incorporation, experiments were performed using the given training set and many CNN architectures. Various techniques were employed to handle exceptionally high-class imbalance like weighted loss, stratified batch sampling, and gradient accumulation. Once the experimental setup was established and hyperparameters tuned, metadata was introduced into the classification pipeline using different ways.

Firstly, the metadata was directly concatenated with visual embeddings obtained by processing a dermoscopic image with convolutional layers of DNNs, as shown in Fig. 3.18. Each piece of metadata, e.g. sex, anatomical location, and age, are individually one-hot encoded with sex represented with a vector of size 3 (male, female and unknown), anatomical location with a vector of size 7 (six defined locations plus unknown), and age with a vector of size 18 (unknown age plus seventeen defined age levels with increment of 5 years). These encodings are then concatenated to form a feature vector of size 28, which is then concatenated with the visual embeddings extracted by the relevant image classifier. The results of this type of metadata incorporation are tabulated in Table 3.18 and represented by ‘concat’ under the Metadata field.

Secondly, the one-hot encoded metadata was processed through a two-layered Multi-Layer Perceptron (MLP) network simultaneously as the corresponding image is processed by CNN. The resulting feature vector is then concatenated with visual embeddings before they are passed on to the softmax classifier. Figure 3.19 shows the schematic of this approach. The results of this approach can also be found in Table 3.18 and are represented by ‘MLP’ under Metadata filed.

In these experiments, the inclusion of metadata appears to have mixed results, sometimes improving the accuracy slightly and other times having a negative impact compared to classification with images only. This shows that contextual information is easier

²<https://challenge2020.isic-archive.com/>

3.5. CAPITALISING NON-VISUAL METADATA TO IMPROVE CLASSIFICATION ACCURACY

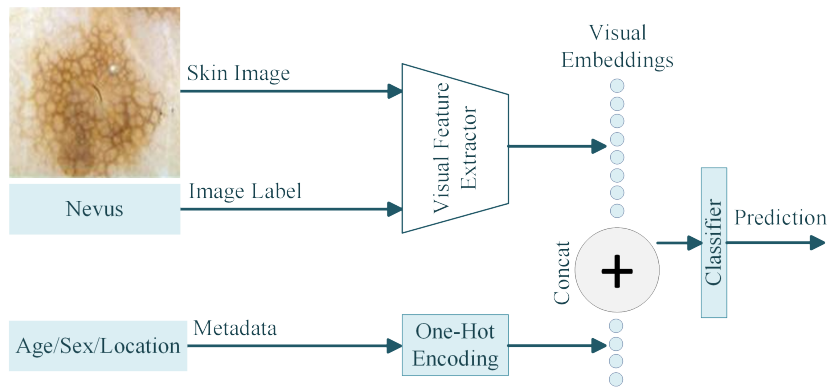


Figure 3.18: Incorporating metadata in image classifier by direct concatenation with visual embeddings.

to understand for human medical practitioners, yet tricky to model and incorporate in DL algorithms. Geesert et al. [206] who won ISIC 2019 challenge with 8 skin lesion classes also report a similar trend with the incorporation of metadata.

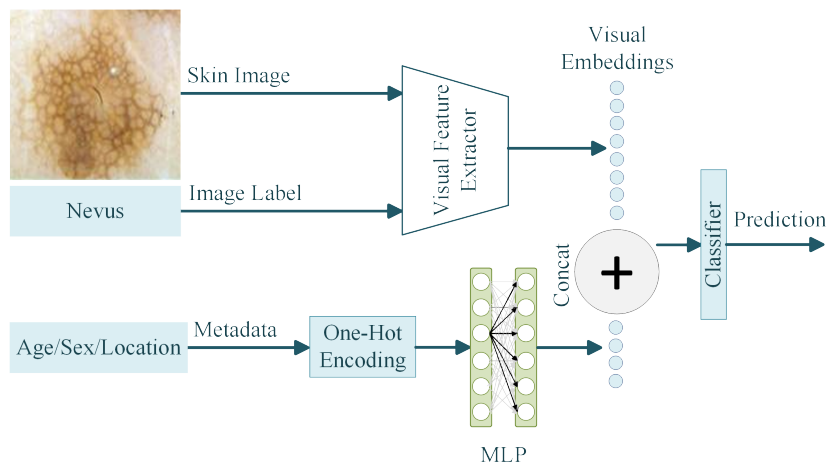


Figure 3.19: Converting metadata into feature vector by processing them with MLP before concatenating with visual embeddings.

An ensemble was also taken for the individual predictions of all classifiers listed in Table 3.18 using all three experimental setups namely, without metadata, direct concatenation of metadata, and passing metadata through a neural network. The ensemble was performed in two ways: averaging individual predictions of the classifiers and taking maximum voting. Table 3.19 summarises ensemble results of three experiments. No significant improvement in the classification performance could be achieved.

Table 3.18: Classification performance of various models with and without metadata

Model	Metadata	Accuracy	F-1 Score	AUC	Test AUC
Inception-V3	None	0.78	0.86	0.8657	0.761
	Concat	0.76	0.85	0.8520	0.7638
	MLP	0.74	0.83	0.809	0.7452
ResNet-18	None	0.71	0.81	0.7996	0.732
	Concat	0.68	0.79	0.7934	0.7355
	MLP	0.69	0.80	0.7796	0.7355
ResNet-152	None	0.69	0.80	0.7652	0.7255
	Concat	0.68	0.80	0.7441	0.7332
	MLP	0.78	0.86	0.8645	0.7811
ResNext-50	None	0.77	0.85	0.8625	0.75
	Concat	0.78	0.86	0.8636	0.7588
	MLP	0.76	0.85	0.867	0.749
ResNext-101	None	0.79	0.86	0.8765	0.7833
	Concat	0.81	0.88	0.885	0.7991
	MLP	0.80	0.87	0.8937	0.7946
DenseNet-121	None	0.78	0.86	0.8615	0.7946
	Concat	0.78	0.86	0.8704	0.7652
	MLP	0.77	0.85	0.8512	0.7664
DenseNet-201	None	0.80	0.87	0.8868	0.7938
	Concat	0.73	0.83	0.8073	0.7394
	MLP	0.78	0.86	0.8696	0.7768
EfficientNet-B7	None	0.69	0.80	0.7799	0.7401
	Concat	0.70	0.80	0.7785	0.7392
	MLP	0.69	0.80	0.775	0.7362

3.6 Discussions

Improving the accuracy of CAD systems is not restricted merely to using advanced CNN architectures and larger high-quality image datasets. It also requires a deeper understanding of the task and smart improvisation. For glaucoma detection using RFIs, realising that OD is instrumental in examining the eye for glaucoma, this chapter presented a fully automated disc localization method based on faster R-CNN. This method eliminates the need for the development of dataset-specific empirical or heuristic localisation methods by providing robust and accurate localisation across several datasets. The performance of these fully automated systems sets new state-of-the-art results in six out of seven publicly available datasets.

The classification of images into diseased and healthy using CNN has also been

Table 3.19: Performance of best performing individual models versus two types of ensemble predictions

Metadata	Ensemble Method	Leaderboard Score
None	DenseNet-201	0.7938
	Average	0.7795
	Voting	0.7755
Concat	ResNext-101	0.7991
	Average	0.7678
	Voting	0.7651
MLP	ResNext-101	0.7946
	Average	0.7797
	Voting	0.7788

investigated. Although some researchers have reported around 95% accuracy on private datasets or carefully selected a small set of images from public datasets, the classification accuracy and AUC for publicly available ORIGA dataset has been challenging to improve. Even though the experiments submitted in this chapter were able to achieve significantly higher AUC on ORIGA with both random training and k-fold cross-validation, the detailed performance measures of the classifier on this dataset revealed that the network has difficulty in learning discriminative features to classify glaucomatous images in this public dataset. It appears that the fine-grained discriminative details in the images of this dataset are lost with the increase in the hierarchy of the network. Therefore, more effort is required to tailor classifiers capable of identifying glaucomatous images with reliability. The empirical evaluation of glaucoma classification on ORIGA also shows that reporting only AUC, for datasets with the class imbalance and without pre-defined train and test splits, does not portray the true picture of the classifier’s performance and calls for additional performance metrics to substantiate the results.

Diabetic retinopathy detection using retinal fundus images is a fine-grained classification task. The biomarkers of this disease on retinal images are usually very small in size, especially for early stages, and are scattered all across the image. The ratio of the pathologically important region to the whole input volume is therefore minuscule. Due to this reason traditional deep CNNs usually struggle to identify regions of interest and do not learn discriminatory features well. This problem of small and distributed visual artefacts coupled with the unavailability of a large publicly available high-quality dataset with reasonable class imbalance makes diabetic retinopathy detection particularly challenging for DNN models. However, fine-grained classification networks have a high potential to pro-

vide standardised and large-scale initial screening of diabetic retinopathy and help in the prevention and better management of this disease. These networks are equipped with specialised algorithms to discover the important region from the image and pay heed to learning characterising features from those regions. The results recorded in this chapter exhibit superior performance for diabetic retinopathy detection on binary, ternary and quaternary classification tasks than many previously reported results. However, due to hugely different experimental setups and the choice of performance metrics, it is unfair to draw a direct comparison with any of the cited research. Nevertheless, a wide spectrum of performance metrics and detailed experimental setup are provided for comparison by any future work.

In addition to understanding the classification task, for example, visual biomarkers and their spatial distribution for glaucoma and diabetic retinopathy, utilising non-visual metadata can also be useful for improving the accuracy of CAD systems. However, availability of such medical image datasets and exploring effective modelling and incorporation methods in CNN architectures are significant challenges in effective use of such metadata [44]. ISIC 2019 Skin Lesion Classification Challenge ³ invited researchers to utilise available metadata with skin images. However, all top-ranking entries in the competitions were unable to propose an incorporation strategy that would yield substantial improvement in the performance of skin lesion classifiers. Close collaboration with medical practitioners to understand the way these metadata are capitalised by doctors could help AI researchers and can also aid in paving the way for efficient assimilation of this vital modality.

³<https://challenge2019.isic-archive.com/>

Uncertainty Estimation in CAD

Despite remarkable performance of AI in many classification tasks [107, 210], including sensitive and critical automated decision-making scenarios like autonomous driving [211], financial systems [212], and medical image analysis for disease prediction [32–34], there is justifiable reluctance by the users of these models to trust an algorithmic prediction without any supplementing estimate of algorithm’s uncertainty. In medical diagnosis, human diagnosticians may often refrain from providing any concrete diagnosis if they are not sufficiently confident about a given case. They may require additional information about the case, run some more tests or seek consultation from their fellow doctors. In CAD, no such facilities are usually at the disposal of AI algorithms. Image classifiers trained using the supervised learning paradigm are provided with a limited number of distinct classes and are expected to produce a prediction for each and every test sample. Even when these CAD systems are provided with an ambiguous or completely unknown case, for which the classifier was not trained at all, they lack the liberty to say, “Well! I don’t know”. In such cases, these algorithms will categorise an unknown sample to the ‘nearest’ class known to them. Such compulsive behaviour of traditional medical image classifiers may have deep and unwanted repercussions on diagnosis and prognosis. Therefore, there has been growing advocacy for the need for uncertainty estimation in such DSS [213], in order to successfully deploy these solutions in the detection and diagnosis of diseases.

4.1 Problem Definition

Convolutional Neural Networks gained significant attention due to their parameter efficiency, in contrast to other deep learning models like densely-connected MLPs, resulting in comparatively better generalisation performance. They are particularly powerful in analysing visual modalities like images and videos [7] but have also proved their worth in time-series analysis where they have been used for classification [214, 215] and anomaly detection [216, 217].

The fundamental principle behind conventional CNNs is to learn the optimal combination of network parameters (weights and biases) that can capture an encoded representation of the training data. These conventional CNNs use point-estimates to represent network parameters and although they work astonishingly well in most image recognition tasks, they have a large insatiable appetite for data [218]. Additionally, the *softmax* function tips the odds in favour of one class by squashing classification probabilities for others. Therefore, often it results in overly confident predictions even when the network is completely wrong. This compulsive behaviour of traditional point-based neural networks to always be relentlessly assertive in their prediction raises serious concerns in many crucial application areas like medical image analysis, security, autonomous driving, financial transactions, and IoT (Internet of Things) based human health monitoring. Also, the very nature of these point-based classifiers prohibits them to associate uncertainty with their predictions, which is a highly desired characteristic of any AI-based classifier.

Bayesian estimation introduces a probabilistic perspective to the neural networks and addresses many shortcomings of traditional point-based deterministic neural networks. It represents each parameter with a probability distribution instead of a single point-estimate. As a result, Bayesian Neural Networks (BNNs) are able to learn effectively from a relatively small amount of data and are fairly robust to overfitting [219]. They can provide an inherent regularisation effect [220] by constraining the network parameters within a distribution instead of allowing them to grow out of bounds. Most importantly, Bayesian inference can permit the estimation of the network’s uncertainty about any prediction. However, a full Bayesian estimation over all network parameters is computationally expensive, and finding true posterior probability is intractable [35]. These limitations are normally addressed by employing various tricks like Markov Chain Monte Carlo (MCMC) sampling [221] and Variational Inference (VI) [222], or a combination of the two [223] to approximate the true posterior with a manageable distribution. A CNN trained using Bayesian estimates for network parameters is shown to lag its

counterpart trained using point-estimates in terms of classification accuracy [219, 224].

In this chapter, the need for uncertainty estimation with CAD is recognised and a potential solution is proposed by acknowledging specific merits of each training approach discussed above and combining them into a hybrid training paradigm. This hybrid approach integrates deterministic CNNs, where each parameter assumes only one numerical value, with probability-driven Bayesian CNNs, where each parameter may take any value drawn from a probability distribution characterised by a mean and a standard deviation. The probability distribution is learnt for each parameter during training. This proposed training method provides an estimate of uncertainty, using a Bayesian classifier, without compromising on classification accuracy owing to a deterministic feature extractor. It also captures maximum weight configurations from small datasets while still remaining computationally manageable. The approach is tested on medical image datasets from ophthalmology and dermatology. To show that the performance of this method is not limited to MIA or image analysis in general, it is evaluated on different classification datasets including benchmark image datasets and time-series datasets. The proposed hybrid method is shown to be superior to both fully deterministic and fully Bayesian CNN approaches in terms of classification accuracy.

4.2 Related Work

Although applications of the Bayesian method into neural networks have been investigated for many decades [225–227], it was only after Blundell et al. [228] proposed Bayes by Backprop that training of deep neural networks was made possible using Bayesian estimation. This method of variational inference allowed backpropagation of so-called Expected Lower Bound (ELBO) loss and regularising weight distributions. A CNN trained using the Bayesian method was proposed by Shridhar et. al [224] as a fundamental construct for other network architectures. They used Bayes by Backprop for training CNNs and reported comparable results on many benchmark datasets.

Acknowledging the excessive computational cost of Bayesian models, Gal and Ghahramani [229] proposed a Monte Carlo dropout method to approximate Bayesian inference in deep Gaussian processes. The method is equivalent to performing multiple forward passes through the network and taking the average of results to model the uncertainty of the network. Combining deterministic and probabilistic models in various fashions has also been studied before. Tang and Salakhutdinov [230] pointed out that the conditional distribution $p(Y|X)$ does not need to be unimodal, as normally assumed by MLPs, but can also be represented as a multimodel output distribution for many structured

prediction problems. They proposed a hybrid Sigmoid Belief Network (SBN) with some stochastic hidden variables and some deterministic hidden variables and achieved superior performance on synthetic and facial expression datasets. Similarly, other neural networks with partially Bayesian parameters have been proposed for regression tasks as an alternative to Gaussian Processes [220, 231], which do not scale well with the number of training samples. Hybrid optimisation of MLPs [232, 233] has also been studied in depth. Furthermore, Bi-level CNNs have been employed to prove the competitive advantages for the point-based and probabilistic interval prediction [234].

Kwon et al. [235] recognised the importance of uncertainty quantification especially in the medical domain and proposed to calculate it by splitting the uncertainty into aleatoric uncertainty, which corresponds to model’s uncertainty; and epistemic uncertainty, which represents inherent noise in the data. Kendall and Gal [236] examined the advantages of modelling epistemic uncertainty as compared to aleatoric uncertainty in deep Bayesian models. The problem of estimating uncertainty has been addressed in a variety of ways, for example, Out-Of-Distribution (OOD) sample detection [237, 238] and density estimation using flow-based models. Normalising flows and autoregressive models have been successfully combined to produce state-of-the-art results in density estimation via Masked Autoregressive Flows (MAF) [239], and to accelerate WaveNet-based speech synthesis to 20x faster than real-time [240] via Inverse Autoregressive Flows (IAF) [241]. Huang et al. [242] presented Neural Autoregressive Flows (NAFs) and demonstrated that these models are universal approximators for continuous probability distributions, and their greater expressivity allows them to better capture multimodal target distributions. Adding on to their work, Cao et al. [243] proposed Block Neural Autoregressive Flow which is a much more compact universal approximator of density functions, where a bijection is directly modelled using a single feedforward network. Dinh et al. [244] introduced a set of transformations called real-valued Non-Volume Preserving (real NVP) as a tractable and expressive way to modelling high-dimensional data. Ardizzone et al. [245] extended real NVP architecture and argued that their proposed Invertible Neural Networks (INNs) are well suited for determining full posterior parameter distribution conditioned on training data. They noted that alternating backward and forward training passes and accumulating gradients from both sides before updating parameters allow efficient training. Kingma et al. [246] advanced flow-based generative models [247], which are useful for calculating exact log-likelihood, by performing exact latent-variable inference and parallelising training and synthesis pipelines. Their Generative flow (Glow) model uses an invertible 1×1 convolution and is shown to be capable of efficient and accurate synthesis of large images.

Laves et al. [248] compared Bayesian ResNet and Variational ResNet trained for detecting various retinal disorders using Optical Coherence Tomography (OCT). They evaluated these models for integrating prediction uncertainty into medical image classifiers. They found that the models showed up to 8 times higher uncertainty for misclassification as compared to correct classification. QuickNAT [249] is a method based on Fully Convolutional Neural Networks (FCNN) for quick segmentation of neuroanatomy using MRI scans. This method samples multiple segmentations to estimate segmentation uncertainty as a means to ensure quality control. In another study, Laves et al. [250] used four DL-based architectures to segment seven regions in the human larynx. They used stochastic inference to obtain an approximate distribution of softmax probabilities. The variance of this distribution is then used to model uncertainties of individual architectures.

4.3 Hybrid Between Deterministic and Probabilistic CNNs

A CNN primarily consists of two main modules: a convolutional feature extractor and a dense classifier. The proposed network consists of a set of convolutional layers trained with point estimates followed by fully-connected layers trained using Bayesian estimates. It provides a trade-off between the high accuracy of deterministic models and the uncertainty estimation of Bayesian models. It also restricts the parameter space of the network as compared to fully Bayesian models because the probability distribution is

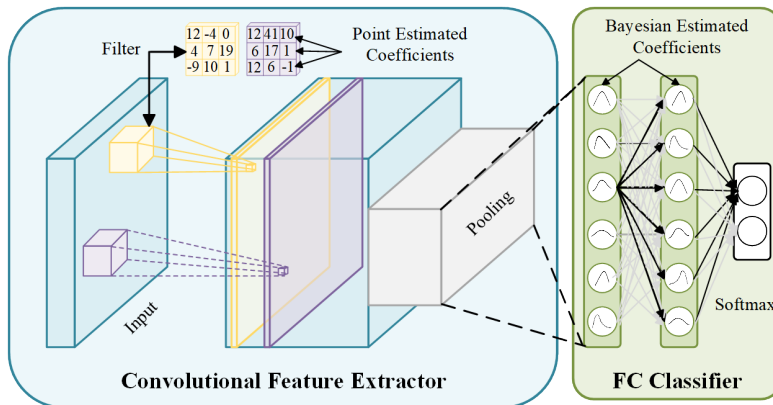


Figure 4.1: The proposed hybrid model. Convolutional feature extractor is trained separately using point estimates. The parameters of the convolutional layers are then frozen and Bayesian classifier is retrained

placed only on the parameters of the classifier part of the network. Figure 4.1 shows a schematic diagram of the proposed hybrid model. The network initially trains to optimise parameters for both convolutional feature extractor and dense classifier as given below in equation 4.1.

$$\mathcal{W}_C^*, \mathcal{W}_D^* = \operatorname{argmin}_{\mathcal{W}_C, \mathcal{W}_D} \frac{1}{|\mathcal{X}|} \sum_{(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}} \mathcal{L}(\psi(\Phi(\mathbf{x}; \mathcal{W}_C); \mathcal{W}_D), y), \quad (4.1)$$

where \mathcal{L} denotes the loss function, Φ represents the convolutional part of the network parameterised by \mathcal{W}_C and ψ represents the dense layers parameterised by \mathcal{W}_D .

Once the network is trained using traditional point-estimates, fully connected layers are reinitialised with random variables following normal distribution and retrained using Bayesian estimation. The parameters of the convolutional feature extractor are frozen throughout this retraining. This whole training paradigm allows to capitalise on economically learned features by deterministic convolutional block and use expensive Bayesian inference only to approximate posterior distribution, which can then be used for uncertainty estimation. Mathematically, the learning of dense classifier of the hybrid model is given by equation 4.2

$$\theta_D^* = \operatorname{argmin}_{\theta_D} \frac{1}{|\mathcal{X}|} \sum_{(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}} \mathcal{L}(\Psi(\Phi(\mathbf{x}; \mathcal{W}_C^*); \theta_D), y), \quad (4.2)$$

where Ψ represents the Bayesian layers learned through Bayes by Backprop and θ_D denotes the distribution of weights. Since the weights are described by a distribution instead of point-wise estimates, \mathcal{L} , in this case, denotes the ELBO loss. Convolutional feature extractor trained with point-estimates learns crisp features of the input data while probabilistic classifier allows to sample from the posterior distribution and offers an insight into network's confidence.

4.3.1 Uncertainty Estimation Algorithm

After this retraining is finished, the inference is performed by passing test samples a number of times from the network. Since the parameters of the last fully-connected layers of the network are sampled from a probability distribution, each pass of the same test sample gives a different prediction. These output predictions are used to draw a posterior distribution and help estimate the network's uncertainty. The complete algorithm used for this task is given in Algorithm 1.

For uncertainty analysis in Bayesian and hybrid architectures during inference, the algorithm works by sampling 10 classifier models from Bayesian weights distribution for

Algorithm 1 Uncertainty Estimation

Inputs *modelOutput*: Array containing softmax probabilities of all images for all models
allPredictions: Array containing class predictions for all images and for all models
allTargets: Array containing actual targets for all images and for all models
percentile: A scalar parameter to ascertain uncertain images to ignore
consensus: A scalar parameter representing minimum number of confident models to reach certain prediction
Outputs *certainAccuracy*: Accuracy when model is certain
uncertainImages: A percentage of uncertain images filtered out

```

1: procedure ESTIMATEUNCERTAINTY
2:   for each model i in allModels do
3:     for each image j in allImages do
4:       differences = differences of top two classes' probabilities in
         modelOutput[i][j]
5:     end for
6:   end for
7:   threshold = calculate for each model by filtering percentile number of images
         from differences of each model and average them.
8:   for each image j in allImages do
9:     Let confPred = 0, uncertain = 0, confModels = 0 be new variables
10:    for each model i in allModels do
11:      if differences[i][j] > threshold then
12:        if allPredictions[i][j] == allTargets[i][j] then
13:          increment confModels
14:        end if
15:      end if
16:    end for
17:    if confModels >= consensus then
18:      increment confPred
19:    else
20:      increment uncertain
21:    end if
22:  end for
23:  return confPred/(len(allImages) - uncertain), uncertain/len(allImages)
24: end procedure

```

every test sample and taking their output predictions. This way, instead of a single prediction, a set of predictions are obtained representing a probability distribution on the network's output. This set of predictions are normalised in the [0-1] range using min-

max normalisation for direct comparison. Predictions for the top two classes are taken and the difference in their values is recorded. After having the normalised differences, a distribution of all these differences is built and a percentile value (40% in this case) is used to automatically select a threshold for the measure of uncertainty. The percentile value of 40% is determined heuristically. This parameter can be considered as a knob to control how confident predictions are desired in any given application area. In circumstances where 'no prediction' is deemed better than a 'wrong prediction' – medical diagnosis, for example – this value can be raised to ensure that only the most confident predictions are given by the network. For other, relatively less critical, scenarios this knob can be adjusted accordingly. The underlying assumption for this uncertainty estimation is that if the network is able to recognise a given test sample then the difference in softmax probabilities of the top two classes should be greater than the threshold and the model is regarded as certain about prediction; otherwise, it is considered uncertain. If a test sample is regarded as certain by more than half models – represented by *consensus* parameter – using simple majority voting then it is output as a fairly certain prediction.

4.3.1.1 Time and space complexity analysis

The hybrid model uses fewer parameters than its Bayesian counterpart as is evident from Table 4.1. This table shows the number of trainable parameters in each method and training time per epoch for some of the datasets. The hybrid model does not incur any additional cost in terms of network parameters or training time for combining the benefits of both deterministic and Bayesian methods.

Table 4.1: Time and space requirement of fully deterministic, fully Bayesian and hybrid models for some datasets

Dataset	Network Parameters (Millions)			Execution Time per epoch (s)		
	Deterministic	Bayesian [224]	Hybrid [Ours]	Deterministic	Bayesian [224]	Hybrid [Ours]
MNIST	2.457	4.914	2.459	15	70	27
CIFAR-10	5.851	11.703	9.528	25	129	49
ISIC-Subset	58.294	116.587	112.840	338	832	602
ORIGA	58.29	116.579	112.831	5	16	6
Electric Devices	0.655	3.277	0.577	2	16	3
Mallat	3.801	33.423	3.486	2	10	3
Thorax-1	2.726	24.589	2.569	2	10	5

The time complexity of the Algorithm 1 is $O(2M \times S)$, where M represents the number of models sampled and S denotes the number of test samples. Also, the algorithm

computes in constant space since, regardless of the number of total models and test samples, only one model and one test sample are loaded at any given time.

4.3.2 Datasets for Evaluating Hybrid CNN

A total of 13 datasets of disparate modalities and from diverse areas of application are used to ascertain the viability of this proposed hybrid CNN architecture. A brief description of all the datasets used and the overall experimental setup is given below.

Table 4.2 gives an overview of all the datasets used in this work. Standard benchmark image datasets, as well as challenging fine-grained medical image classification datasets and many time-series datasets, are selected so that the validity of the approach on a broad range of datasets may be extensively investigated.

Table 4.2: Distribution of datasets used to evaluate the proposed architecture

Datasets	Modality	No. of Classes	No. of Samples		
			Train	Test	Total
Image Datasets					
MNIST	Grey Images	10	60k	10k	70k
CIFAR-10	Color Images	10	50k	10k	60k
Medical Image Datasets					
ORIGA	Color Retinal Fundus Images	2	520	130	650
ISIC-Subset	Color Clinical Skin Images	3	5201	600	5801
Time Series Datasets					
Fish	Image-derived data	7	175	175	350
ShapesAll	Image-derived data	60	600	600	1200
Plane	Sensor data	7	105	105	210
TwoPattern	Simulation data	4	1000	4000	5000
ECG5000	ECG data	5	500	4500	5000
MedicalImages	Image-derived data	10	381	760	1141
ElectricalDevices	Device data	7	8926	7711	16637
Mallat	Simulation data	8	55	2345	2400
ECG Thorax1	ECG data	42	1800	1965	3765

4.3.2.1 Image Datasets

The detail of the ORIGA dataset is given in chapter 2.2.1. It provides clinical ground truth to benchmark segmentation of optic disc and classification of healthy and glaucomatous images. Since this dataset is very small and no predefined train and test splits are given, 5-fold cross-validation is used. The second dataset of medical images was taken from ISIC Archive 2018 version ¹. It consists of around 24000 clinical and dermoscopic images of skin lesions categorised into 7 classes. Some of the classes in this dataset have as few as 122 images, therefore, a subset with the three largest classes namely Benign Keratosis-like Lesions (BKL), Melanoma (MEL), and melanocytic Nevi (NV) is taken and randomly divided into training and test sets.

Two of the most common benchmark datasets i.e. MNIST [160] and CIFAR-10 [251] are also used. For these datasets, standard pre-defined train and test splits are used.

4.3.2.2 Timeseries Datasets

Nine datasets from UCR archive [252] are also selected for evaluation of this approach. The time-series datasets were generated based on different modalities including device usage, sensors data, ECG, motion sensor, and simulation, etc. Each time-series contains a different number of classes; the number of observations also varies in each dataset. All datasets are already divided into train and test sets by the publisher.

4.3.3 Preprocessing

To preprocess medical image datasets, histogram equalisation is applied to enhance contrast and normalise brightness. Different data augmentation techniques like rotations, flipping, and random crops are also utilised to increase the dataset size. In addition to preprocessed images, original images are also kept in the dataset. Data augmentation was done keeping in mind the class ratio, such that the minor class can have more augmentations and more copies generated. On benchmark image datasets (MNIST and CIFAR-10), random crop and normalisation by mean subtraction are applied. Time-series datasets are used without any preprocessing.

4.3.4 Experimental Setup and Hyperparameter Selection

All of the image datasets were trained and compared with a similar experimental setup. A 5-layer convolutional block is used as baseline CNN. However, experiments with varying

¹<https://challenge2018.isic-archive.com/>

depths and breadths of CNN showed that the approach is fairly scalable to more advanced CNN architectures. This CNN is trained using Maximum Likelihood Estimation (MLE) for 60 epochs with a learning rate of 10^{-3} , weight decay of 5×10^{-4} , and batch size of 32. For probabilistic models, the same setup is used as described above but instead of using point estimates the convolutional and fully connected layers are trained with distribution-based weights using Bayes by Backprop for 60 epochs.

In the hybrid approach, a fully-connected classifier is employed with a frozen convolutional feature extractor, pre-trained using MLE, and is fine-tuned using Bayesian estimation for 60 epochs with similar parameters. Two hyperparameters used in Algorithm 1, i.e. *percentile* and *consensus* can be selected according to the use case requirements. In critical application areas, for example, medical image diagnosis or stock market prediction, where there is little room for incorrect classification, higher values of these parameters can be selected to ensure only the most certain predictions are given by the network. In other applications, a relaxed criterion for uncertainty estimation might be acceptable. In these experiments, *percentile* = 40% and *consensus* of more than half models (i.e. 6 models) is used. These values were selected empirically and they worked well in all 13 datasets of different kinds. It should be emphasised here that, for a given dataset, the same underlying architecture (number, width, and depth of convolutional layers and size of dense layers) is used in all three training paradigms, i.e. fully deterministic, fully Bayesian, and Hybrid, to ensure fair comparison among three approaches.

For time-series modality, a CNN with two convolutional layers is used, each followed by a max-pooling layer for deterministic model analysis. On top of that, two fully connected layers were added as the classifier. For probabilistic and hybrid approaches, the same settings are kept as explained before.

4.3.5 Results and Analysis

Table 4.3 summarises classification accuracies obtained by traditional fully deterministic CNN, Bayesian CNN [224] and the proposed hybrid CNN (HCNN). The table shows that the HCNN outperforms not only purely Bayesian CNNs but also their deterministic counterparts in 9 out of 13 datasets while giving comparable results on the rest of them. Even when the hybrid approach lagged other methods in classification accuracies, the difference was very small and came at no additional cost in terms of time or number of parameters as shown in Table 4.1. The results in *Bayesian Accuracy* field in Table 4.3 are generated by running experiments using the implementation of Shridhar et al. [224]

Table 4.3: Comparison of fully deterministic, fully Bayesian, and the proposed hybrid models on different datasets without using uncertainty estimation

Datasets	Deterministic Accuracy (%)	Bayesian [224] Accuracy (%)	Hybrid [Proposed] Accuracy (%)
Benchmark Datasets			
MNIST	99.0	99.01	99.3
CIFAR-10	88	72.0	88.7
Medical Image Datasets			
ORIGA	76	74.4	80.3
ISIC-Subset	74	65.5	75.7
Time Series Datasets			
Fish	85.1	80.7	84.7
ShapesAll	67.0	70.9	72.3
Plane	97.0	96.7	95.1
TwoPattern	89.0	81.0	89.4
ECG5000	92.0	93.2	91.9
MedicalImages	69.0	62.4	64.7
ElectricalDevices	55.0	54.0	56.6
Mallat	88.0	82.5	89.3
ECG Thorax1	90.0	89.1	91.3

for Bayesian CNNs.

Figure 4.2 shows output probabilities of deterministic, Bayesian, and hybrid models for various correctly classified and misclassified images from CIFAR-10 and ORIGA. It can be observed from Fig. 4.2 that when the hybrid model was unable to make a correct prediction (subfigures (b), (d), (e), and (h)), it associated relatively smaller probability scores with its misclassification than its competing models who also misclassified but did so with overconfidence. For example, consider Fig4.2 (h) where the original label of the image is healthy. Although the hybrid model failed to correctly classify this image it predicted glaucoma with only a 74.14 probability score. In contrast, deterministic and Bayesian models also misclassified this image but predicted glaucoma with 95.97 and 95.37 probability scores, respectively. Additionally, in cases where both deterministic and Bayesian models failed to correctly classify an image and hybrid network succeeded (subfigures (c), (f), and (g)), it predicted very cautiously with reasonable probability scores. The probability scores of the hybrid model were at par with the other two methods for relatively easy examples as shown in subfigure (a).

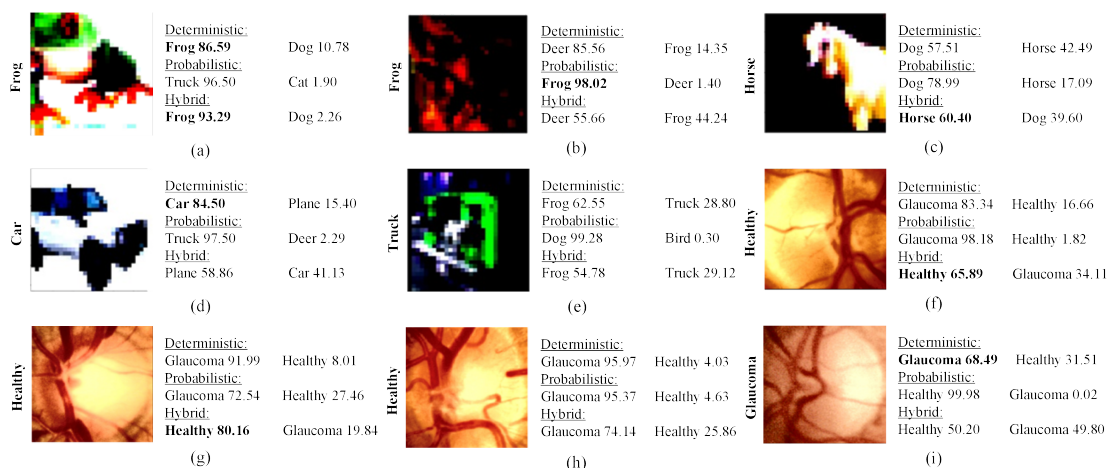


Figure 4.2: An analysis of confidence comparison for all three approaches on various samples of CIFAR10 and ORIGA datasets. The actual class is mentioned on the left side of each image in bold vertical text

4.3.5.1 Uncertainty Estimation

Since the deterministic model does not have the intrinsic ability to estimate uncertainty – although some works like [229, 253] have used deterministic models and applied some post-processing to get confidence estimates – this section focuses on Bayesian and Hybrid models only and compares their performance. As the classifier part of both Bayesian and Hybrid methods are trained using Bayesian estimates, both networks provide posterior distribution which is used to estimate uncertainty using Algorithm-I. Table 4.4 compares the accuracies of both training methods before and after using Algorithm 1. In this table, *Overall Accuracy* refers to the accuracy of the model before applying Algorithm 1, whereas *Certain Accuracy* refers to the accuracy on the predictions for which the network was certain according to Algorithm 1. When the algorithm is not sure about the prediction it tags the test sample as uncertain.

It can be observed from Table 4.4 that the accuracies for both fully Bayesian and hybrid approaches improved after the uncertainty estimation algorithm was applied. The accuracy of the hybrid approach is higher than the fully Bayesian model especially when it was fairly certain about the predictions. However, in the case of medical image datasets, HCNN outperformed the Bayesian approach even without uncertainty analysis. Removing uncertain predictions from all predictions, accuracies for both Bayesian and hybrid models improved with HCNN outperforming by up to 15% for fairly confident predictions. Figure 4.3 shows some examples of images that were considered certain

Table 4.4: Comparison of fully Bayesian and the proposed hybrid models on different datasets with uncertainty estimation

Datasets	Bayesian Model [224]			Hybrid Model		
	Overall Accuracy (%)	Certain Accuracy (%)	Uncertain Samples (%)	Overall Accuracy (%)	Certain Accuracy (%)	Uncertain Samples (%)
Image Datasets						
MNIST	99.01	99.17	20.5	99.26	99.28	9.6
CIFAR-10	65.41	72	66.9	88.70	91.11	46.2
Medical Image Datasets						
ORIGA	74.42	77.10	35.65	80.31	77.21	38.7
ISIC-Subset	58.15	65.48	34.3	75.67	81.5	53.8
Time Series Datasets						
Fish	80.7	92.4	9.1	84.7	100.0	6.8
ShapesAll	70.9	71.8	1.0	72.3	72.9	1.3
Plane	96.7	98.9	0.95	95.1	97.1	0.0
TwoPattern	81.0	84.4	25.0	89.4	91.3	24.9
ECG5000	93.2	93.8	36.2	91.9	93.9	36.8
MedicalImages	62.4	62.9	0.13	64.7	66.5	0.13
ElectricalDevices	54.0	55.8	14.6	56.6	57.9	14.8
Mallat	82.5	84.2	35.6	89.3	92.1	37.7
ECG Thorax1	89.1	90.9	14.9	91.3	91.6	14.8

or uncertain by both the Bayesian model (top row) and hybrid model (bottom row). It is very interesting to observe that the algorithm enabled both models to confidently categorised those images that had clearly defined optic disc border (black dotted elliptical boundary drawn on images to highlight disc boundary). In both training approaches the images where the boundary of the disc dwindled, for example, because of papilledema (Fig. 4.3d and Fig. 4.3h) or optic atrophy (Fig. 4.3b and Fig. 4.3f), were filtered out and the models did not predict on these images because of high uncertainty.

Figure 4.4 depicts the trade-off between the number of uncertain samples and classification accuracy for both Bayesian and Hybrid models. It can be seen from this figure that the accuracy of the networks increases with the increase in the percentage of uncertain samples. Though one can argue from these curves that since *difficult* samples have been skipped by the classifier and prediction is given for *easy* samples only, that is why there is a positive trend in the accuracy with a growing number of uncertain samples. However, in many crucial application areas, it is better to abstain from giving any half-cooked prediction than making a potentially costly mistake. In medical image analysis, for instance, such non-compulsive classifiers can reduce the workload of human

4.3. HYBRID BETWEEN DETERMINISTIC AND PROBABILISTIC CNNs

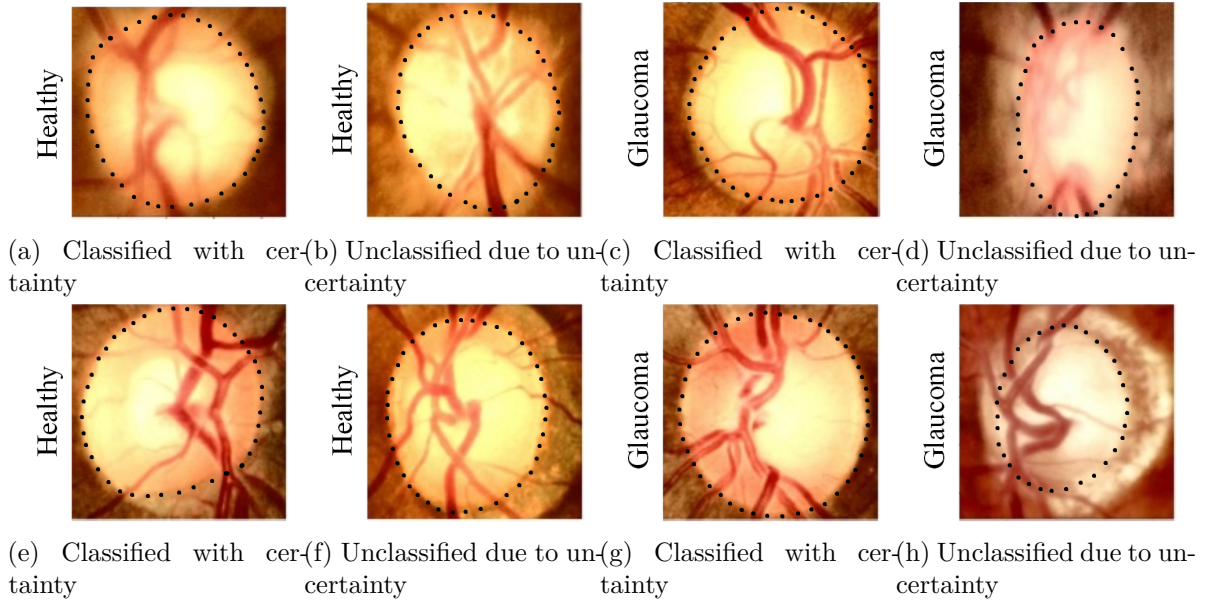


Figure 4.3: Comparison of output probabilities for fully Bayesian and hybrid training approaches on ORIGA dataset

experts by screening relatively easy disease patterns and allowing the physicians to focus their time and energy only on the most challenging of the cases.

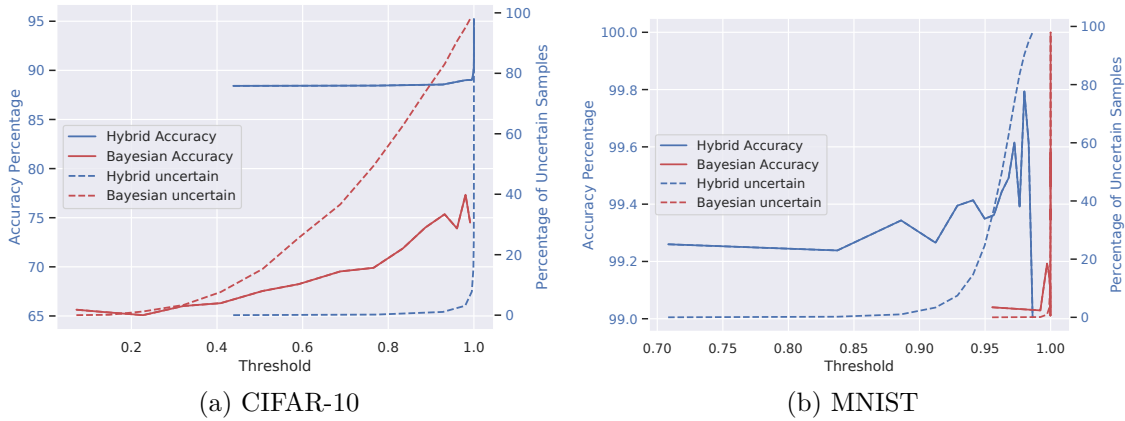


Figure 4.4: Trade-off between number of uncertain samples and the accuracy on remaining predictions. The threshold on x-axis is calculated using *percentile* parameter as shown in Algorithm 1

4.4 End-to-End Training of Hybrid CNN

The method described above in section 4.3 has a practical limitation regarding training the architecture. The model needs to be trained in two stages: first, training the whole architecture using cross-entropy loss, and second re-training probabilistic classifier using ELBO loss. To remove this limitation and enable the proposed hybrid CNN train smoothly in a single pass, the hybrid architecture shown in Fig. 4.1 is re-implemented. The traditional fully-connected layer is replaced with an analogue that effectuates Bayesian variational inference by assuming that the convolutional kernels or bias, or both, are drawn from probabilistic distributions. This layer implements Local Reparametrisation Estimator (LRE) proposed by Kingma et al. [254] which reduces the variance of stochastic gradients for variational Bayesian inference of a posterior over layer’s parameters. This parametrisation technique uses a Monte Carlo approximation of the distribution on kernel’s and bias’s hidden units. However, instead of multiplicative parametrisation as employed in [254], this layer uses additive noise reparameterisation [255] which is shown to achieve faster convergence and reduce the variance of stochastic gradients.

Using Local Reparametrization Estimator in the dense layers of an otherwise traditional CNN allowed training the parameters of convolutional layers with point-estimates and the parameters of the dense classifier with probabilistic kernels in a single training run. The loss function used to train the whole network in the backward pass is categorical cross-entropy,

$$CE_{Loss} = - \sum_{i=1}^C t_i \log(p_i) \quad (4.3)$$

where t_i and p_i are ground-truth and network predictions, respectively, for each class i in C . In order to update the parameters in the probabilistic dense layer forward KL divergence is used,

$$D_{KL}(P||Q) = \sum_{x \in X} \log\left(\frac{P(x)}{Q(x)}\right) \quad (4.4)$$

where for all random variables $x \in X$, $P(X)$ is the true distribution, and $Q(X)$ is its approximated distribution. KL Divergence calculates the weighted average on the difference between $P(X)$ and $Q(X)$ probability distributions at x .

The datasets given in Table 4.2 are used along with the G1020 dataset described in section 2.3 to perform another set of experiments with this new HCNN. The experiments for fully deterministic and fully Bayesian counterparts are also repeated with slight

modifications for hyperparameters. For image datasets, the CNN architecture consists of six convolutional layers and two dense layers, while for time-series datasets the network has two convolutional layers and two dense layers as used in section 4.3 above. For uncertainty estimation, Algorithm 1 is used.

4.4.1 Experiments and Results

Table 4.5 shows the comparison of the accuracy using deterministic CNN, BCNN, and HCNN. The hybrid network outperforms the other two on medical image datasets and does far better than the Bayesian network on benchmark and time-series datasets. Although the accuracies of deterministic CNN are better than the Hybrid approach on most datasets, yet in many sensitive application areas higher prediction accuracies without any measure of uncertainty are of little practical value.

Table 4.5: Comparison of Accuracy (%) of deterministic, Bayesian, and proposed hybrid models on different datasets without using uncertainty estimation

Dataset	Deterministic	Bayesian	Hybrid
Medical Image Datasets			
ORIGA	76.92	73.84	80.76
ISIC-Subset	78.79	42.37	82.73
G1020	81.86	62.74	76.00
Benchmark Image Datasets			
MNIST	99.66	98.54	99.42
CIFAR-10	85.26	58.04	80.14
Time Series Datasets			
Electrical Devices	64.93	59.11	61.08
Mallat	93.98	81.96	84.35
ECG5000	93.75	93.60	94.24
Medical Images	69.08	62.23	66.05
Fish	88.00	56.57	84.00
Shapes All	75.00	54.17	70.33
Plane	96.19	94.28	96.20
Two Pattern	88.12	84.20	85.30

The performance of the new HCNN shines brightly when compared to the other network that allows uncertainty estimates, namely BCNN. Here, the hybrid approach does not only give higher classification accuracies compared to BCNN but also more confident predictions as shown in Table 4.6. Using the hybrid approach resulted in a

smaller percentage of samples that the model was not confident about and was classified as uncertain. Additionally, the certain accuracy, which is the accuracy of the confident predictions, is higher for the hybrid approach than its Bayesian counterpart for all datasets. This implies that the hybrid model gives fewer confident predictions that are misclassified.

Table 4.6: Comparison of Bayesian and Hybrid models on different datasets before and after uncertainty estimation.

Dataset	Bayesian Model			Hybrid Model		
	Uncertain Accuracy	Certain Accuracy	Uncertain Samples	Uncertain Accuracy	Certain Accuracy	Uncertain Samples
Medical Image Datasets						
ORIGA	73.84	72.88	9.23	80.76	83.00	23.08
ISIC-Subset	42.37	53.73	18.78	82.73	87.59	21.82
G1020	62.74	67.36	6.87	75.98	76.92	17.16
Benchmark Image Datasets						
MNIST	98.54	99.74	5.64	99.42	99.87	2.2
CIFAR-10	58.04	68.25	28.30	80.14	91.76	27.4
Time Series Datasets						
Electrical Devices	59.11	62.43	9.76	61.08	66.63	16.31
Mallat	81.96	83.22	3.41	84.35	86.01	4.01
ECG5000	93.60	93.90	0.6	94.24	94.15	0.49
Medical Images	62.23	66.33	8.55	66.05	69.78	8.54
Fish	56.57	61.68	4.57	84.81	84.00	9.71
Shapes All	54.17	73.32	33.2	70.33	82.56	24.5
Plane	94.28	94.12	2.86	96.20	96.20	0.00
Two Pattern	84.20	83.79	0.65	85.89	85.30	0.6

4.4.2 Analysis

On all medical image datasets, it was observed that HCNN by far outperformed its Bayesian counterpart. However, the more crucial observation is that HCNN gave not just higher accuracy but a higher certain accuracy justifying that the hybrid implementation can be used for confident CAD. On the ORIGA dataset, HCNN gave better accuracy along with higher precision and recall. The HCNN had a precision of 0.69 and a recall of 0.61, higher than both the deterministic and probabilistic models, which had precision and recall values of 0.64 and 0.57, and 0.50 and 0.50 respectively. What is of greater significance is the finding that the precision and recall values were fairly high for the

malignant images in ISIC-subset. The HCNN had a precision of 0.84 and recall of 0.94 on the malignant images, showing that the model did fairly well to correctly classify malignant tumours.

The HCNN also gave high certain accuracy. A deeper analysis of the results showed that for confident results that were incorrectly classified, the model misclassified normal images as glaucoma more often than it misclassified glaucoma images as normal. This observation was consistent across both the Bayesian and the hybrid architectures. However, the ratio of confidently misclassified normal images to confidently misclassified glaucoma images is 3:1 on BCNN whereas the same is 16:1 on HCNN. This suggests that even when HCNN gave wrong confident predictions, it wrongly predicted normal images as glaucoma rather than the opposite. A false positive at an early screening of a disease can be corrected by advanced testing before prognosis. However, a false negative is more likely to result in negligence of a serious condition.

On the ISIC dataset, BCNN gave a low certain accuracy suggesting that the model could do better in correctly and confidently making predictions. However, HCNN showed more promising results, giving high certain accuracy. Moreover, while BCNN had a precision of 53.2 and a recall of 52.4, the hybrid model far exceeded this performance standard with a precision of 82.2 and recall of 81.1. Higher precision and recall values of the hybrid model corroborate that the hybrid model is a suitable candidate for realising non-compulsive confident CAD systems.

4.5 Discussion

Practical applications of DL-based medical image classification models require high accuracy, better generalisation, computational efficiency, and an estimate of the uncertainty in the model's predictions. All these characteristics are not readily available with either traditional deterministic CNNs or Bayesian CNNs. Deterministic models, though provide better accuracies, do not facilitate uncertainty estimation on their own. Bayesian method, on the other hand, allows explication of posterior distribution but has a significantly larger number of parameters that require more memory and time for tuning. Therefore, in this chapter, a hybrid CNN is conceptualised and implemented, which is capable of combining some of the merits of deterministic and Bayesian methods. The proposed method is validated on a number of different datasets and shows promising results. The experimentation with different architectures with a varying number of convolutional and dense layers showed that the hybrid training approach performed consistently better than its deterministic and Bayesian equivalents. This work might serve as a stepping

stone for further exploration of such hybrid CNNs since it has the potential of performing noticeably better while at the same time facilitating estimation of the network's certainty for every prediction. Improved HCNN with end-to-end training in a single run is efficient to train and does not cost extra in terms of training time or memory requirement. A thorough architecture search and hyper-parameter tuning might be required to increase baseline accuracies for each dataset. However, the experimentation with various data modalities and application areas has shown great promise to prompt further comprehensive investigation into this training paradigm. One logical next step in this research could be to incorporate this hybrid approach with dataset-specific architectures obtained through, for instance, NAS-Net [107] and ENAS [256] algorithms.

Explainability of CAD

In 2016, Ribeiro *et al.* [257] reported an image classifier that was able to inadvertently classify correctly but for wrong reasons. They found out that their wolf versus dog classifier learnt an undesirable correlation between the wolf and the background snow and, therefore, would classify a given image as a wolf if there was snow in the background. If it were not due to the authors' vigilance in finding explanations to the model's predictions, it would have been difficult to properly evaluate the trustworthiness of this image classifier. The inherently inquisitive human nature prompts us to unfold and understand the rationale behind decisions taken by DNN based algorithms. This curiosity has led to the rise of eXplainable Artificial Intelligence (XAI), which deals with making AI-based models considerably transparent and building trust in their predictions. Over the past few years, AI researchers are increasingly turning their attention to this rapidly developing area of research not only because it is driven by human nature but also because legislations across the world are mandating the explainability of AI-based solutions [258, 259].

Although the case of correct classification for incorrect reasons as reported in [257] was an inconsequential example of spurious correlations learnt from a large amount of data, medical diagnosis resulting from such misunderstandings can potentially have a grave impact on human lives. One of the biggest advantages of consulting a doctor is the opportunity to discuss one's medical conditions, ask questions about the differential diagnosis and talk over the likely course of action. Similarly, when a group of doctors deliberates over a case, they provide an explanation on their viewpoint and justify

their opinion through arguments. Such discussions that offer justifiable explanations of medical diagnosis and prognosis allow patients to confer their trust on their healthcare providers and help medical practitioners avoid any pitfall in their decision-making process. Moreover, many health insurance companies require that any medical procedure, test, or course of treatment must be justified to be medically necessary before a claim to cover the cost of such services is settled. Therefore, in a routine clinical environment, simply naming a medical condition might not be enough. Commonly developed medical image-based disease classifiers only provide a numerical value corresponding to the class label without giving a quick peep into their decision-making process. This lack of transparency could be one of the mightiest hurdles in the successful integration of CAD systems in real-world healthcare systems. The requirement for a CAD to be explainable arose with early applications of AI in healthcare [260] and became more relevant with recent ethical and legal standards [261, 262]. The consequent increase in research activity in the domain of XAI also reflects the growing interest of the community to provide explanations for CAD systems [261]. In addition to evaluating the reasons behind a model's predictions, explanation methods can also help in revealing new diagnostic criteria [263] previously unknown to medical practitioners.

This chapter addresses the need for explainable AI, especially in medical image diagnosis. It provides a comprehensive overview of existing achievements and open challenges in explainable CAD systems and presents methods to help explain disease prediction of DL-based classifiers. These methods are then unified into a framework for generating easy-to-understand textual explanations for medical diagnosis.

Skin cancer is the most common type of cancer in the U.S [264]. According to a recent study, [265], skin cancer related death rate forecast for the U.S in 2019 amounted to 11,650 people. These rising rates of skin cancer incidences can not only cost precious lives but also incur a huge burden on healthcare systems. It is estimated that approximately 3 million people are treated annually for skin cancer in the U.S and it costs around 8.1 billion USD [266]. Therefore, in this thesis, the classification of malignant melanoma from benign naevi is chosen as a use case to study explainable CAD systems.

5.1 Problem Definition

The human-centric explainability of AI-based DSS using visual input modalities is directly related to the reliability and practicality of such algorithms. An otherwise accurate and robust DSS might not enjoy the trust of domain experts in mission-critical application areas if it is not able to provide reasonable justifications for its predictions. It

is, therefore, the need of the hour to elucidate the working principle of deep learning based classifiers so that practical applications of AI in medical diagnosis can be realised expeditiously.

Compared to other fields of applications of DNNs, the MIA often presents unique challenges due to the inherent complexity of this task. Manual classification of complex diseases involves recognising subtle features and high-level concepts that are challenging to grasp without expert knowledge. Even with expert knowledge, doctors' subjective understanding of disease biomarkers leads to low inter-expert agreement [267, 268]. Therefore, common explanation methods like visualisation of saliency maps, which strongly rely on spatial divisibility of concepts, work well on common object detection tasks [269–271] that have well-distinguishable features but fail on more complex medical image analysis tasks.

5.2 Achievements and Challenges in Explainable CAD

This section provides a comprehensive analysis of AI approaches successfully employed in explainable CAD systems and some of the most prominent open challenges requiring further attention.

5.2.1 Overview of Common XAI Methods

Methods explaining the decision-making process of DNNs exist in a variety of forms. Not only the derivation of the explanations differs but also the way it is communicated to the user. There are a number of taxonomies available in the literature to differentiate these methods. An important distinction for AI users, for example, is made between post-hoc and ante-hoc methods. Methods that can explain the decision of a so-called 'black box' model after it is developed and trained are called post-hoc (literally meaning, after-this event) methods. Ante-hoc (literally meaning, before-this event) methods, on the other hand, are already interpretable – to some extent at least – due to their architecture. Since these ante-hoc explanations are usually achieved by architectural or conceptual restrictions in the learning process that limits modelling capacity, such inherently interpretable models are often thought to be under-performing than their unrestricted conspecifics in terms of final model performance. However, this effect can sometimes be mitigated by pre-processing raw data with noisy features into meaningfully structured representations [272]. Another distinction among these explanation methods can be made with respect to a classifier's ability to explain their decision-making process

on a global scale or locally on a single data point at a time. Although local explanations might be initially sufficient for clinical applications as assistive diagnosis systems, global explanations are crucial for understanding a model’s behaviour as a whole. This is specifically important for identifying decision biases and hence for the development of autonomous decision systems. As mentioned before, there exist various taxonomies for XAI methods in the literature. A few types of explainable methods are discussed below that are specifically relevant to medical imaging. A visual overview of the grouping is provided in Fig. 5.1.

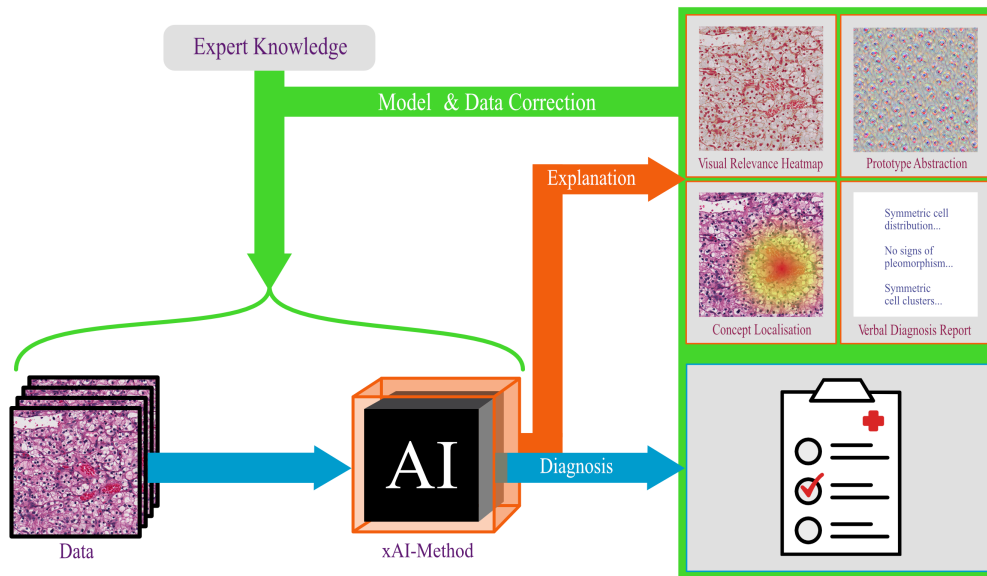


Figure 5.1: Topology of the XAI process with optional model and data correction as well as taxonomy of common and relevant explainable AI methods in medical image analysis

5.2.1.1 Visual Relevance Heatmaps

Probably the most popular group of methods for explaining and interpreting image-based classification methods is the generation of visual heatmaps representing the influence of individual pixels on the result of the classification. Existing methods differ significantly in the computation of relevance values. The most obvious approach is the visualisation of the internal activations of a model [273]. Therefore, single or combinations of intermediate, two-dimensional activation values are scaled to input size and are visualised. Other common methods rely on the attribution of the classification results to the individual pixels. In practice, this is done using, for instance, weighted activations in Class Activation Maps (CAMs) [274], gradient-based methods like Saliency [275],

Gradient*Input [276], Grad-CAM [271], Integrated Gradient [277], DeepLift [278] or methods based on mathematical decomposition like Layerwise-Relevance Propagation (LRP) [279], Agglomerative Contextual Decomposition (ACD) [280] and SHapley Additive exPlanations (SHAP) [281]. All these methods require access to the model parameters and thus an understanding of the model architecture.

Perturbation-based methods, on the other hand, are completely model-agnostic and can therefore be used for model-independent explanation without knowledge of their internal constructs. In order to explain a given sample, it is modified several times and evaluated by the model repeatedly in order to systematically record the changes caused by the perturbations. Methods like Occlusion [273], RISE [282], and Extremal Perturbation [283] differ in the occlusion strategy (procedure and perturbation). LIME [257] goes one step further and trains local approximation models based on the results of the randomly modified images.

In addition to the post-hoc methods mentioned so far, there is also a possibility to generate relevance heatmaps in an ante-hoc process. Here, model architectures can be extended by attention mechanisms that force the model to focus its attention explicitly on certain parts of the input and to hide the remaining part. This distribution of attention can often be visualised in a heatmap [284], using pointers [285] or by explicitly cropping the input to the intended region [286] to gain insight into the network’s decision-making process.

5.2.1.2 Class- and Prototype Abstraction

Visual Relevance Heatmaps (VRHs) usually help to explain the decision on individual samples. Another approach that aims towards both global and local explanations of DL models is the generalised representation of prototypes of individual classes or neurons as learned by the model. This includes, for instance, methods maximising the activation of particular outputs [275] or intermediate neurons [287] by optimising over an input image to determine their ”prototypical” activation patterns. Many variations of this approach have already yielded interesting results and insights [288] for general image recognition tasks. However, only a few works can be found applying abstraction methods to medical problems [289–291]. This might be attributed to the complexity and entanglement of disease criteria and consequently complications in interpreting the prototypical results.

5.2.1.3 Conceptual Explainability and Biomarker Identification

The aim of concept-based explanation methods is to map human-understandable semantic concepts to the concepts learned by DL models after training in order to make their decision-making processes more comprehensible. Such concepts can be very simple characteristics such as colours, shapes, or textures. However, complex concepts can also be defined, consisting of combinations of simpler concepts. The Testing with Concept Activation Vector (TCAV) method developed by Kim et al. [292] requires a small number of sample images per concept to compute global concept influence scores. Further exploitation of this method allows explicit localisation of the concepts recognised by the network in the input domain, extending its application to regression tasks [289, 293] and introduce improved metrics [294]. Other concept-based approaches include Network Dissection [295] the quantifies how interpretable the latent representation of a CNN is by evaluating the alignment between individual hidden units in the network and a set of semantic concepts, and Interpretable Basis Decomposition [296] that provides visual explanations for image-based classification models by decomposition intermediate activations pertaining to an input image into semantically interpretable components that are pretrained from a concept dataset.

Especially in the application of DL in the medical domain, the detection and localisation of biomarkers by the model is popular in addition to the diagnosis of diseases. This approach allows intermediate steps of the models to be validated by experts. As has been shown in recent works [289, 293], even post-hoc concept-based methods can be used to detect such biomarkers. However, more common approaches in the literature are ante-hoc methods based on multi-task learning [297], where the models are trained for the combined classification or localisation of biomarkers [298–300]. Segmentation networks are often used for localization also as in [299], however, such explicit approaches presuppose that correspondingly annotated data are available. An alternative approach by Zhang et al. [301] combines the optimisation of a CNN and a Generative Adversarial Network (GAN) in a single end-to-end architecture for the localisation of biomarkers without the presence of explicit biomarker annotations. Generative DNNs can be trained to learn the underlying data generating process of a given training dataset, which can be used to interpolate among samples and synthesise new images.

5.2.1.4 Textual Explainability

There are different methods for generating verbal explanations of DL model decisions. These methods can be categorised into those that use a template-based approach [302–

304], rule-based methods [305–307], and those that utilise Natural Language Processing (NLP) models to generate an explanatory text [308]. An early use case of NLP-based, textual explanation generation in the medical domain is MDNet framework developed by Zhang et al. [309]. This framework allows the generation of a textual diagnostic report based on a medical image. In addition, a heatmap is generated for each word of the diagnostic report, which shows users the model’s attention at that step.

5.2.2 Achievements of xAI in Medicine

The number of research papers on interpretability and explainability of AI has mushroomed in the last few years [261] and thereby the application and adaption of XAI methods to specific medical domains have also increased. In the following, some influential research works are presented with the most practical significance towards clinical DSS.

5.2.2.1 Interventional Methods

The explanation of high-performing AI algorithms that utilise spurious indicators for classification allows revealing biases. To make practical use of these explanations, methods that facilitate intervention and correction of working of algorithms are required. Common methods for penalisation and correction of explanations in DL models work by imposing a loss on explanation heatmaps, for example from VRH method, or conceptual predictions, like TCAV, against ground truth explanations provided by the human experts [310, 311]. This area is strongly related to the field of explicit expert knowledge incorporation. Examples of successful application of such methods in the medical domain are disease grading in diabetic retinopathy [312], lymph node histopathology [313] and dermoscopic skin lesion classification [314, 315]. Rieger et al. [315], for instance, were able to correct a classifier trained on the ISIC 2019 dataset, which is heavily biased towards benign predictions when coloured patches appear beside the lesion. A comparison between Grad-CAM maps generated before and after correction of the network can be seen in Fig. 5.2. Inspired by the concept-based explanation method of TCAV, Graziani et al. [313] fine-tuned a deep classifier for histopathologic lymph node tumour detection. By penalising undesired control targets (concepts), they managed to increase AUC by 2%.

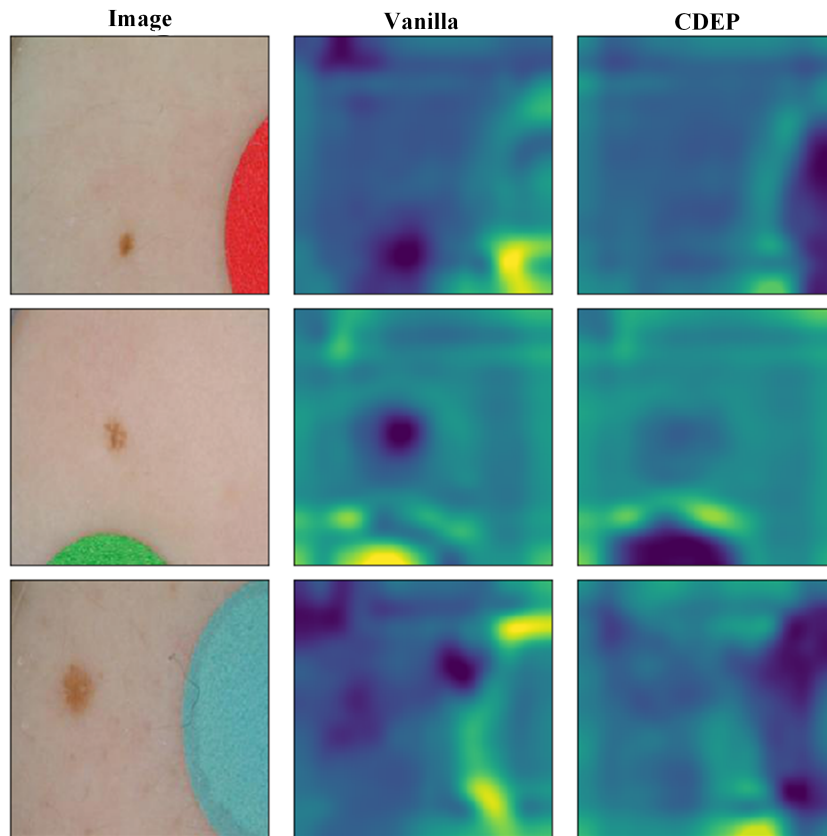


Figure 5.2: Comparison between Grad-CAM heatmaps generated before and after correction of the trained network using model correlation methods. The left column shows the original image samples from the dataset. The middle and right columns show grad-CAM heatmaps before and after correction

5.2.2.2 Revealing New Criteria

Explainability methods are often employed in specific, sometimes medical application, areas by expert computer scientists to prove their effectiveness. Lack of domain knowledge on part of computer scientists often hampers the proper interpretation of presented results, rendering the provided explanations less useful for assessing the correctness of the network. However, an increasing trend of collaborations between medical professionals and computer scientists is apparent in the application and tuning of DL models. The need to obtain domain knowledge for both computer scientists and domain experts in order to understand and explain models has been reflected in a growing number of publications on XAI.

A team of computer scientists and neurosurgeons succeeded in training a CNN for

the localization of diagnostic features in confocal laser endomicroscopy images for glioma detection using only image-level annotations. Izadyyazdanabadi et al. [263] sequentially applied a visual relevance localization method to a multi-head network, merging the resulting maps by collateral integration as well as biologically inspired lateral inhibition principle. Their diagnostic localization maps correctly identified familiar diagnostic features and also revealed new diagnostic regions that were previously unknown to the neurosurgeons. Using a complex model architecture consisting of two autoencoders and further processing steps, an interdisciplinary team of pathologists and computer scientists successfully predicted the recurrence of prostate cancer from digitised slides of histological sections in [316]. A custom-made method for calculating an impact score, which provides information about the direction of influence of an image section for diagnosis, offers further insights. It has been confirmed that the model independently learned the concept of the Gleason Score, an established prognostic value for prostate cancer among experts worldwide, and identified the occurrence of stroma, which is an intermediate tissue running through the parenchymatous organs, as a prognostic factor for prostate cancer in areas of the incision-free of cancer cells.

No clear physiological characteristics of insomnia are known yet. Researchers from Charité Berlin, HTW Berlin, and University Medicine Göttingen have used machine learning models in [317] to detect insomnia in polysomnographic data with the aim of revealing such physiological features through AI. By applying DeepLift [278] method, some factors such as increased and less synchronous eye movement were highlighted as relevant for the prediction of insomnia. However, the authors themselves stress that the results should be interpreted with caution, as neither the bias of the results due to laboratory conditions can be excluded nor can the validity of the factors be definitively confirmed.

A team of computer scientists and biologists used samples of microbiomes of human female skin to determine phenotypes such as age, skin moisture, menopause status, and smoking status in [318]. The SHAP method was used to assess the relevance of each bacterial genus in the microbiome. As this method generates local explanations, SHAP values for all bacterial genera were averaged over the subset of samples with correct and good results for classification and regression. The most relevant bacterial genera and their influence on the respective task were reported. For the determination of all phenotypes, a number of relevant bacteria genera were identified. In the case of skin moisture determination, for instance, the genera identified by the model as particularly important were already associated with skin moisture in previous studies. Essemli et al. [319] were able to determine whether patients suffer from mild cognitive impairment

or even Alzheimer’s dementia using their two-dimensional connectivity matrix of brain regions. They used a specially adapted CNN architecture for this purpose. To explain the disease prognosis, the gradients of all images of the respective classes were averaged to obtain a global explanation. These averaged heatmaps of different classes were subtracted to emphasize the crucial differences between the two conditions. The results confirmed that the connectivity of the entorhinal cortex is crucial for the separation between healthy and Alzheimer’s disease subjects and the hippocampus for the separation between healthy subjects and those with mild cognitive impairment. Their results have been discussed with an expert neuroanatomist.

A project at the German Research Center for Artificial Intelligence (DFKI) is specifically focused on the development of a CAD system for the detection of skin diseases [320, 321]. The system developed in the Skincare project is capable of analysing images of skin diseases taken with a smartphone, generating a differential diagnosis, and segmenting the skin lesion and individual biomarkers. The explainability of the system is ensured through the calculation of expert scores and VRHs. A demo of the system can be tested on the project webpage¹.

5.2.3 Challenges for XAI Applications in Medicine

Since the initial applications of modern DL-based systems in medical domains, there have been remarkable strides in the explanation of systems that in some cases already led to correction and verification of AI as well as disclosure of new potential diagnostic criteria. However, there are still a number of challenges pertinent to medical image diagnosis, which should be addressed by concerted efforts from AI researchers, medical practitioners, and regulatory authorities.

5.2.3.1 Evaluation of Explanation Methods

Before XAI methods can be practically deployed, it must be ensured that their explanations are reliable, trustworthy, and useful. This evaluation of explanations must take into account the truthfulness and usefulness of the explanations and their interpretation by the users.

¹<http://www.dfki.de/skincare/classify.html>

Evaluation of Truthfulness

One of the key challenges in explainable AI is difficulty in evaluating if the explanation of a model's behaviour is reliable. This is primarily because there is no gold standard ground truth available for such evaluations [322]. Truthfulness or fidelity of an explanation refers to whether it is reliable and reflects the actual decision process of the AI. In order to practically install AI in clinical environments such that it increases the efficiency and accuracy of human doctors, it is of paramount importance to ensure the fidelity of XAI methods. However, due to the lack of explanation ground truth, evaluation of such methods is largely subjective.

There have been attempts to quantify and measure the quality of explanations. Samek et al. [323] introduced a metric called Area Over the MoRF Perturbation Curve (AOPC) to quantitatively compare VRHs. The measure gradually perturbs input images starting from the regions that are marked as the most relevant according to a given explanation method. High AOPC values indicate that a model is sensitive to perturbations in those regions, thus confirming the validity. The RemOve And Retrain (ROAR) framework [322] is an advancement of AOPC approach. As image perturbations lead to a change in image distribution, they retrain the network on the perturbed images to avoid distribution gaps and evaluate the achieved accuracy. However, the evaluation of an altered model cannot give reliable insights into the sensitivity of the original model. In [324] a synthetic dataset with ground truth explanations has been generated for easier XAI method evaluation. Adebayo et al. [325] introduced randomisation tests in which model weights and data labels were systematically randomised to reveal if explanation methods were really model and data-dependent. Although this method has not been used to quantify fidelity, its results are certainly meaningful for evaluation.

Truthfulness is the basis for robust and useful XAI. Results from works like [325] showed that some methods produce convincing explanations that are worth no more than simple edge detectors. Eitel et al. [326] performed a quantitative comparison of visual relevance methods for MRI-based Alzheimer's disease classification. They found that guided backpropagation attribution maps [327] averaged over all true positives for multiple training runs highlighted different regions in brain MRI. However, despite the variance, which makes it harder to compare and replicate outcomes of individual experiments, some regions like the hippocampus, cerebellum, and edges of the brain were commonly identified as salient regions. Other visual relevance methods like Gradient*Input, Occlusion Sensitivity, and LRP also showed similar behaviour, which raises serious questions on the robustness and coherence of these explanation methods. How-

ever, this could also indicate an abundance of biomarkers in the data that allows DNN's to perform the same task in a variety of ways.

Evaluation of Usefulness

Besides evaluating the fidelity and completeness of explanation methods, it is also crucial to quantify and qualify the usefulness of generated explanations. Doshi-Velez and Kim [328] proposed the distinction between application-grounded, human-grounded, and functionally-grounded evaluation of explanations. In [329] the first functionally-grounded metrics were introduced, allowing to objectively judge the quality of an explanation. This quantification has the advantage of being independent of human subjectivity. On the other hand, human-grounded evaluation makes use of non-specialist human evaluators to subjectively compare or rate explanations. The evaluation approach that is found to be the most important for XAI in medicine is the application-grounded evaluation. Depending on the domain or problem, medical practitioners have a very specific way of thinking about a problem, communicating or explaining a diagnosis. Hence, the application-grounded evaluation is necessary to find and optimise the right explanation methods for a medical use case.

Evaluation with respect to Evaluators

An equally decisive factor in the use of XAI methods is their interpretation by the end-user. One explanation can be interpreted differently by different individuals. A wrong or too naive interpretation of decision processes by developers or users can lead to serious consequences in the practical use of AI. The approach to the interpretation of explanations differs significantly for AI researchers and medical practitioners, but also overlaps to some extent.

For AI developers, explainability methods can help them design better models by understanding the interactions between the model and the data. However, AI developers and data scientists can sometimes over-trust or misuse these interpretability tools as noted by [330]. They conducted a small-scale study to learn how data scientists utilise publicly available interpretation tools and found that visual explanations are usually taken at their face values and used for rationalisation of suspicious observations instead of understanding how AI models worked. Experienced data scientists, on the other hand, were able to capitalise on these interpretability tools and effectively understand issues with models and data.

For medical practitioners, such tools can provide reasoning for model predictions and, therefore, develop trust and ease their acceptance into routine clinical workflow. Sayres et al. [331] evaluated the impact of DL-based diabetic retinopathy detection algorithm on the performance of human graders in the computer-assisted setting. They found that the accuracy of human graders improved when assisted by the algorithm that provided only disease prediction without any explanation. However, when the graders were provided prediction plus visual explanation by the algorithm, their detection accuracy improved only for patients who had diabetic retinopathy (resulting in high sensitivity) and decreased for patients without the disease (resulting in low specificity). Although the qualitative feedback of human graders on the explanations provided by the algorithm was generally positive, the participants were not able to harness this additional information to notably improve their performance. This could partly be because the pathologic features of diabetic retinopathy are very tiny in size, inconspicuous, and occupy only a fraction of the whole image space.

To meet the challenges in the evaluation of XAI, special focus should be placed on the evaluation of the realistic applicability of methods in a clinical environment. This includes truthfulness, robustness, quality, and the actual usefulness of the methods. Through such detailed analyses, the agreement between medical expert knowledge and the knowledge gained from the model and data can be evaluated and validated and, possibly, new knowledge can be gained. A further dimension that should not be neglected when evaluating xAI applications in healthcare is the ethical assessment of the impact on individuals and society. There is an increasing commercial interest in explaining AI decisions. This requires the development of regulatory measures that take into account different needs of different individuals and user groups and are adaptable to the constantly evolving AI technology [332]. However, this also requires clearly defined evaluation and certification processes to assess the ethical conformity of the use of AI in a specific context. z-Inspection [333] is one of the first ethical evaluation and certification processes that integrates theoretical principles for the ethical evaluation of AI into a practically applicable framework.

5.2.3.2 Deployment in Clinical Workflow

Proof of concept studies and prototype methods are required to be tested rigorously to analyse their contextual fit in a real-world clinical environment. However, many obstacles have been discovered and highlighted by researchers in implementing laboratory research in clinical settings. These challenges include lack of utility to clinicians' logistical hurdles

that hamper clinical deployment and trials [334]. Ineffective use, or misuse, of these assistive systems can even lead to performance degradation of human graders [335–337]. Cai et al. [337] developed interactive user-centric techniques for pathologists to improve diagnostic utility and trust in algorithmic predictions in laboratory settings. Previously, such Human-Computer Interface (HCI) techniques have been used only to improve the algorithm. However, these interactive tools have the potential to enable users to test, understand, and grapple with AI algorithms, leading to new ways for improving their explainability. Instead of waiting for algorithms to generate human-understandable explanations [257, 338], interactive techniques can allow users to play an active role in the interpretation of algorithm predictions and hypothesis-test their intuitions. In a study [339] designed for the field assessment of a DSS for cardiologists, it was found that the clinicians were more likely to embrace and use such systems if it was seamlessly and unobtrusively integrated into their existing workflow. However, the misuse of these systems can sometimes let the clinicians develop their own tolerance and workarounds in order to trust the algorithm results [340].

There are a few examples of such translation of AI into commercial applications, for instance, in the detection of diabetic retinopathy [341], cancer, and analysis of radiology images [342]. Deployment of CAD solutions in clinical settings can also help focus on the effects of a workflow when new diagnostic and information systems are introduced into clinical environments. Arbabshirani et al. [343] integrated their AI-based model for identification of Intracranial Haemorrhage (ICH) using head CT scans into a clinical workflow for three months. During the trials, the model was able to reduce the median time to diagnosis for routine studies from more than eight hours to only 19 minutes, while at the same time discovering some probable ICH cases which were overlooked by radiologists.

5.2.3.3 Diverse and Complete Explanations

Most applications of XAI in research focus on utilising single approaches and modalities for the explanation of AI models in given use cases. This can be seen in the analysis of achievements of XAI in section 5.2.2 as well as many reviews on this topic [324, 344–346]. However, the integration of XAI in the clinical workflow can benefit more from a combination of multiple explanatory views to draw explanations that are diverse and as complete as possible. This is inspired by medical practitioners in routine healthcare environments using textual descriptions alongside visualisations and temporal coherence to communicate decisions effectively and reliably. On one hand, this should motivate

AI researchers to think of new and creative paths for XAI methods for complementing existing methods and on the other to not only evaluate the effectiveness of approaches in isolation but also in combination with diverse methods and leverage synergies. Early efforts towards diverse explanations have been recently made in the Visual Question Answering community in works like [347] and [348]. Huk Park et al. [347] show the positive complementary effect of visual relevance and textual explanations which is backed up by human evaluation. Completeness of explanations can be considered from the point of view of the model and the user. Completeness from a model’s point of view is directly related to fidelity. Yeh et al. [349] introduced a measure that quantifies the completeness of a given concept-based explanation for a model’s prediction. Completeness from a user’s point of view is subjective but equally relevant to usefulness.

5.2.3.4 Human-Centric Explanations

High-performing DNNs often utilise unintelligible notions of concepts to reach a prediction. Integration of AI assistants in clinical workflows requires a human-centric explanation of a decision that is able to not only describe a decision with high fidelity but also conforms to human-understandable thought models. Compared to simpler use-cases like visual object classification or part segmentation, complex medical concepts used for diagnosis particularly necessitate making explanations as human-understandable as possible.

Human-Understandable Concepts

One way to explain the decisions of AI-based CAD systems in a human-centric way is to investigate the role of human-understandable concepts, learned by DL-based algorithms. It is very important to analyse the learned features of an algorithm that makes the right decisions but is based on wrong reasons. It is a major issue that can affect performance when the system is deployed in the real world. Explaining the role of a model’s concepts can reduce reliability concerns of medical practitioners and help develop their trust in CAD.

Application of concept-based XAI methods in MIA has been challenging partly because these methods require concept datasets [296] or image patches corresponding to those concepts that are human-understandable [292], which are not always available. An unsupervised approach, extending the Concept Activation Vectors (CAVs) method, is developed by Ghorbani et al. [350] to cluster object datasets by performing segmentation of single objects and clustering their relevant activations into semantically meaningful

groups. This approach cannot be directly applied to, for example, skin lesion classification where there is a substantial overlap between various concepts that can not be segmented into distinct spatial patches. Also, this method does not guarantee the discovery of human-understandable concepts and requires thorough human evaluation effort.

Sometimes general explanation methods cannot be readily used for certain medical image tasks due to technical requirements or inappropriateness to the domain. Besides the continuous development of advanced XAI methods, it is important that developers pay attention to the domain-specific needs of particular medical applications and their users. There have been many studies extending existing methods to better suit the challenges of MIA. For example, Yang et al. [351] proposed Expressive Gradients (EG), an extension of commonly used Integrated Gradients [277] to cover the retinal lesions better while [293] extended CAVs from [292] for continuous concepts like eccentricity and contrast. A part of this thesis extended the method for localising and highlighting image regions significant for network’s concept recognition in a medical inspired dataset. This could allow doctors to verify the network’s concept learning and suggest precise image regions for concepts. Such studies lead to the advancement of the XAI domain and provide specialisation to application domains without designing new methods from the scratch.

Challenges in Textual Explanations

Most disease classification algorithms using medical images attempt to answer Multiple Choice Questions (MCQs) in which the algorithm is expected to select one disease from a list of all possible diseases. In this type of experimental setting, there is a fair chance that a correct prediction given by AI-based CAD is nothing more than a fluke – though the probability of fluke decreases with the increase in the total number of classes. Therefore, such classification algorithms require explicit interpretations of network predictions to validate their results.

In many medical domains like radiology and histopathology, doctors routinely write textual reports clearly noting salient findings before providing their impression (diagnosis). The nature of this type of detailed diagnosis substantiated by textual descriptions of the image is self-explanatory – at least for the domain experts. AI-based CAD can be enabled to process this multi-modal data (image and text) and generate textual reports to mimic the behaviour of radiologists and histopathologists. Such systems embed explanations of their decisions within their predictions. These natural language explanations, using domain-specific terminology and mimicking the structure of communication

provide an intuitive and effective way of explaining decision processes to practitioners. However, providing textual explanations in the form of clinically accurate medical reports for medical images has some differences compared to other application areas where NLP is used to describe an image.

Generating long coherent reports (more than a few dozens of words) is one of the major challenges in textual XAI. Language generation models usually start with a few coherent sentences and after that their performance tapers off generating completely random words that have no association with the previously generated words or phrases. This happens generally due to very long temporal dependency among words which Long Short-Term Memory (LSTM) [352] models have difficulty handling. One way to address this problem is to use transformer networks [353] as a language model decoder. These models are able to capture the relationship between words in a longer sentence better than Recurrent Neural Network (RNN) based models. Input text reports are tokenised and passed to the transformer network that consists of a decoder layer and generates a query vector for another transformer model that generates reports by combining this query with information obtained from the image processing model. The size of the generated reports and vocabulary can also be limited to ensure that the text is coherent and clinically meaningful.

Most of the reports written by doctors are free text reports, which means that they do not always follow any defined template. Reports written by two radiologists, for example, for a given X-ray image can be vastly different. There can be superfluous information that does not contribute directly to the final diagnosis. This makes it very difficult to compare AI-generated reports with human-generated reports especially when some of the reports depend on the previous examination of the patients and provide a continuous diagnosis. This problem can be addressed by removing those parts of the input reports which bear no influence on the diagnosis such as at what time the doctor saw the patient or who was the doctor on call.

Incorporation of Context

Traditional AI algorithms overwhelmingly rely on one input modality, for example, images in medical image analysis. However, medical practitioners routinely incorporate context, in the form of, for instance, a patient's clinical history, age, and sex, etc., in their decision-making process. Compared to raw image pixels, this contextual information is much easier to understand for practitioners. However, incorporation of this metadata into AI algorithms is tricky since context is difficult to represent in a form that

is appropriate for processing by AI algorithms [354]. Not leveraging this useful context in DNNs can not only restrict their performance but also make explanations challenging. Therefore, another direction of research to make DNNs more transparent and explainable is to use multi-modal data like medical images and patients' records together in the decision-making process and attribute the model decisions to each of them [345]. This approach simulates the diagnostic workflow of a clinician where both images and physical parameters of a patient are used to make the decision. It can not only improve the diagnostic performance of these algorithms but also explain the phenomena more comprehensively.

5.3 Explaining Network Decision using Concept Activation Vectors

The applications of XAI are at least as widespread as AI itself including in medical image analysis for disease predictions, text analytics [355], industrial manufacturing [356], autonomous driving [357], and insurance sector [358]. Many of these application areas utilise visual inputs in the form of images or videos. Humans recognise these images and videos by identifying and localising various concepts that are associated with objects – for example, concepts of shape (bananas are long and apples are round) and colour (bananas are generally yellow and apples are mostly red or green). XAI methods dealing with images also employ a similar approach of identifying and localising regions in the input space of a given image that corresponds strongly with the presence or absence of a certain object, or concept associated with the object.

One way of elucidating a deep learning based CAD could be to verify that the model learns and utilises similar disease-related concepts as defined and employed by human diagnosticians. The objective of this study is to scrutinise if the concepts learnt by deep image-based classifiers in complex skin lesion classification tasks are similar to those used by dermatologists. To do so, human-understandable concepts are mapped to the RECOD image classification model, which is a well-trained and high-performing DNN developed by REasoning for COMplex Data (RECOD) Lab for the classification of skin tumours, with the help of Concept Activation Vectors (CAVs). The RECOD model is trained for the classification of three skin tumours, i.e. Melanocytic Naevi, Melanoma, and Seborrheic Keratosis. A detailed analysis is performed on the latent space of DNNs to comprehend what they learn and what they rely on for their predictions in medical diagnosis. Two well-established and publicly available skin disease datasets, PH² and

derm7pt, are used for experimentation. These datasets are selected because they provide concept annotations in addition to image-level diagnosis labels.

A thorough survey on the use of concept-based explanation methods for skin lesion classification showed that these methods have not previously been explored for this application area. Due to the nature of this problem, not all of the previously described methods can be directly applied to this task. Unsupervised clustering as used in [359], for example, is not suitable in skin lesions as there is a huge spatial concept overlap and thus no possibility for distinct part segmentation. Regression Concept Vectors (RCVs) are also not applicable as skin lesion concepts are hardly quantifiable. The method in [296] requires a concept corpus that is not readily available for this specific task. Any type of textual explanation generation is also not applicable, as no diagnostic reports or descriptions of diagnosis are provided with any public dermoscopic skin lesion dataset. The computation of CAVs as given in [292] requires patches corresponding to general human-understandable concepts. In this work, the TCAV method is adopted to the problem of skin lesion classification. Instead of providing general, OOD concept patches, concept classifiers are trained using samples from identically distributed datasets to map human-understandable concepts to the network’s latent space.

The CAVs and the method of calculating TCAV scores are briefly described below as used in this work to quantify the contribution of a concept to DNN’s prediction. Moreover, dermoscopic concepts explaining the classifier’s decisions are also introduced.

5.3.1 Concept Activation Vectors

To achieve human-centred interpretability of DNNs, Kim et al. [292] introduced *Concept Activation Vectors*. A CAV, denoted by \vec{v}_c , is a vector in the embedding space of a neural network pointing into the direction that encodes the concept c . The CAVs can be calculated by training a binary concept classifier that distinguishes samples containing a given concept from samples where the concept is absent. The CAV is then defined as the normal to the hyperplane separating the two classes.

TCAV Score The metric introduced in [292] to estimate the influence of a CAV on a class of input images is the TCAV score. It makes use of directional derivatives $S_{C,k,l}(x)$ to measure the contextual sensitivity of a concept towards an entire input class, therefore providing global explanations. The TCAV score is given by:

$$TCAV_{Q_{c,k,l}} = \frac{|\{x \in X_k : S_{C,k,l}(x) > 0\}|}{|X_k|}, \quad (5.1)$$

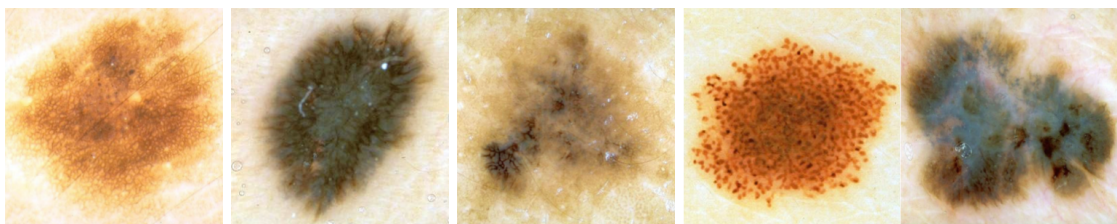
where X_k denotes all inputs, k represents the class labels and $S_{C,k,l}(x)$ is the directional derivative of a sample's activation x from layer l with respect to class k and concept C . The TCAV score effectively measures the ratio of class k 's inputs, that are positively affected by concept C without taking any magnitude into the account. As compared to saliency maps or other per-feature metrics, the TCAV score allows for quantitative evaluation of concepts on whole input classes.

5.3.2 Dermoscopic Concepts used for Analysis

The concepts used in this work to understand the decision-making of a deep classifier are briefly defined below in accordance with standardised terminology agreed upon by expert dermatologists in the 3rd Consensus Conference of the International Society of Dermoscopy (ISD) [360]. Figure 5.3 depicts examples of some concepts mentions below.

5.3.2.1 Pigment Networks

Pigment Networks consist of interconnected pigmented lines forming a grid-like pattern. Depending on the subtype of Pigment Networks, it can either have minimal variability in colour, thickness, and spacing of the lines, forming a symmetric grid (Typical Pigment Network), or have greater variability in colour, thickness, and spacing of the lines, forming an asymmetric grid (Atypical Pigment Network). Apart from those two general types, more subtypes are also defined in the literature. Atypical Pigment Networks can be a clue for Melanoma (although many dysplastic naevi also have atypical networks) whereas typical Pigment Networks normally indicate benign melanocytic lesions (Naevi).



(a) Typical Pigment Network (b) Regular Streaks (c) Regression Structure (d) Regular Dots & Globules (e) Blue Whitish Veil

Figure 5.3: Exemplary cases of skin lesion concepts from derm7pt dataset

5.3.2.2 Streaks

Streaks describe an abnormality of the lesion that can either have the form of straight radial extensions, radial extensions with bulbous and often kinked projections on their ends, or a widening of broken lines with incomplete connections. Streaks are referred to as irregular if they are irregularly distributed along the edge of the lesion and are brown-black in colour [361]. Regular Streaks indicate benign lesions and Irregular Streaks are clues for malignant Melanoma.

5.3.2.3 Regression Structures

Regression Structures are characterised by the appearance of either area of fine, grey-blue dots, or areas of skin whiter than the surrounding normal-looking skin without blood vessels or shiny-white structures. Its presence is highly indicative of melanoma [361].

5.3.2.4 Dots and Globules

Dots are small structures of pigmented areas clustered in any distribution in or around the lesion. Dots clustered in the centre or on the network lines are referred to as regular, otherwise, they are called irregular. Globules are round, oval, or polygonal structures larger than dots that can have high variability in colour, size, and shape along with asymmetric distribution for Irregular Globules, or minimal variability along with symmetric distribution for Regular Globules. Regular Dots and Globules are indicators for benign melanocytic lesions and irregular Dots and Globules indicate melanoma [361].

5.3.2.5 Blue-Whitish Veils

Blue-Whitish Veils describe an irregularly shaped, structureless blotch on the lesion area that is characterised by a blue hue with an overlying whitish ground-glass haze. In [362] it is rated as the most useful single diagnostic indicator for melanoma.

5.3.2.6 Asymmetry

Asymmetry is the most important factor in malignant melanoma identification using ABCD rule [363]. In this work, asymmetry refers to an asymmetrical lesion contour as well as asymmetrical distributions of structures and colours within a lesion [57]. The asymmetry concept is further divided into symmetric or asymmetric in one or two axes.

5.3.2.7 Colour

This concept refers to the colours present within the lesion area. As the appearance of a single colour is not yet indicative of any diagnosis, a combined concept of *three or more colours* is used in the analysis. The presence of three or more colours increases the probability of melanoma drastically [361].

The intricate explanations of concepts given above along with the concepts' innate variability offer much room for interpretation, implying the complexity of the problem itself. This is evident by the fact that even expert dermatologists tend to have notable disagreements when it comes to diagnosis, localization, or identification of concept [267, 268].

5.3.3 The RECOD Model

The model used in this work as the basis for the exploration and experimentation is developed by the University of Campinas in Brazil. Their RECOD Lab made their submission [364] to the IEEE International Symposium on Biomedical Imaging (ISBI) 2017 challenge and is publicly available on github². By applying a transfer learning approach combined with extensive ensembling using an SVM meta-layer on top of seven base models trained on different data subsets, they achieved the best AUC for Melanoma (MEL) classification (87.4%), 3rd best AUC for Seborrheic Keratosis (SK) classification (94.3%), and 3rd best combined/mean AUC (90.8%) in part 3 of 2017 challenge. In this part of the thesis, this RECOD model is used in lieu of training another skin lesion classification model as the primary objective is the explainability of deep models instead of their classification performance. In the later part of this chapter (section 5.5) it is shown that this method is equally effective with any other DNN trained for any classification problem, as long as relevant concept annotations are available. Thus, for these experimentations, attention is only focused on a single module from RECOD's well-trained architecture. The base models³ with Inception v4 [104] architecture is used, which is subsequently referred to as *the model* or *the network*. This base model was trained on RECOD's "deploy" set of 9,640 images using per-image normalisation.

²<https://github.com/learningtitans/isbi2017-part3>

³*checkpoint.rc25 of RECOD model*

5.3.4 Datasets for Concept Classification and Evaluation

The datasets used for concept training are PH² dataset [57] and Seven-Point Checklist Dermatology dataset abbreviated as derm7pt [298].

The PH² dataset consists of only 200 dermoscopic images of melanocytic lesions, including 80 common naevi, 80 atypical naevi, and 40 melanomas. Along with the images, colour and lesion segmentation masks are provided along with extensive well-curated annotations. The derm7pt dataset consists of 1,011 clinical and dermoscopic images. Each sample is assigned to either a miscellaneous class or one of 4 diagnosis classes. Two of these diagnosis classes i.e. Melanoma and Naevi (NV) are further divided into 13 sub-classes. From this dataset, only MEL and NV samples have been considered, resulting in 823 images. SK samples have been discounted due to their low count of only 45 samples. Table 5.1 provides an overview of some samples for each concept class.

Table 5.1: Distribution of image samples into different concept classes in PH² and derm7pt datasets. Note that PH² dataset does not distinguish between regular and irregular streaks

Concepts	Presentation	Abbreviation	PH ² [57]	derm7pt [298]
Pigment Network		PN	N/A	551
	Typical	PN_T	84	335
	Atypical	PN_AT	116	216
Streaks		ST	30	333
	Regular	ST_R	N/A	96
	Irregular	ST_IR	N/A	237
Regression Structures		RS	25	233
Dots & Globules		DG	113	690
	Regular	DG_R	54	300
	Irregular	DR_IR	59	390
Blue-Whitish Veils		BWV	36	182
Asymmetry		Sym	117	N/A
	1-Axis	Asym_1	31	N/A
	2-Axis	Asym_2	52	N/A
Colours	3 or more	C_3	39	N/A
Total Samples			200	823

For evaluation purposes, the original ISBI 2017 challenge dataset [111] is used. The train set of the ISBI 2017 challenge contains 1372 samples of NV, 374 samples of MEL, and 254 samples of SK whereas the test set contains 393 images of NV, 117 images of MEL, and 90 images of SK.

To verify the statistical significance of the results, CAVs for random concepts are calculated to compare against the CAVs for real dermoscopic concepts. For this purpose, random concept labels are assigned to a subset of the ISIC archive⁴ images, excluding MEL and NV classes, resulting in 2870 samples. The idea behind leaving out those two classes is that the remaining samples hardly contain concepts similar to the ones used for concept training.

5.3.5 Experiments and Results

As previously described, all experiments have been conducted on one of the Inception v4 base models from [364]. For each concept, binary classifiers are trained on the network’s activations to find the concepts’ directions in the embedding space. The training and evaluation scheme is depicted in Fig. 5.4. First, the activations are extracted from *mixed_6h* layer of the model using PH² and derm7pt datasets. A clustering-based under-sampling technique along with stratified splitting is applied to ensure evenly balanced train and validation splits for each binary concept training. These dataset splits are balanced with respect to not only concept labels, but also target class labels. Train and validation data are split with a ratio of 2:1. Second, the TCAV score is used to evaluate a concept’s importance to a specific target class. To account for differences in pre-processing and classifier initialisation, each classifier training is repeated 20 times on a randomly sampled dataset split, resulting in different CAVs and different TCAV scores.

Additional 50 random CAVs per layer are trained to ensure the statistical significance of the learnt concepts. The random datasets are produced by repeatedly sampling 1,000 random images from the ISIC archive subset and assigning them random binary labels. The distribution of random concept TCAV scores and real concept TCAV scores is then compared by conducting a two-sided *t*-test with $\alpha = 0.05$ to assure significance of the calculated CAVs. In the results section, statistical insignificance is represented by red asterisks on top of the plotted bars. The lack of quantifiability in most of the relevant explanation methods does not allow for proper comparison with previous approaches.

⁴<https://isic-archive.com/>

Hence, the focus is placed on the quantitative evaluation of the concept classifier’s accuracies and TCAV scores as well as a qualitative analysis of the resulting CAVs.

Figure 5.5a shows all mean validation accuracies achieved by individual binary concept classifiers, and their standard deviation, trained on *derm7pt* embeddings from *mixed_6h* layer. The mean baseline results from training on 50 random concept subsets are depicted by horizontal red line along with light red shaded area marking standard deviation. It is evident from the figure that all concept classifiers achieved significantly higher validation accuracies than random baseline. At first look, the overall accuracies achieved might not seem very high. However, it has to be mentioned here that computation of CAVs requires the use of linear classifiers to calculate normal vector to decision hyperplane. The results are clear evidence that the network’s latent space is structured in a way that allows activation’s separation with respect to similar concepts.

Figure 5.5b shows the classifiers’ validation accuracy trained on PH^2 dataset embeddings from *mixed_6h* layer. It is notable that many concepts achieved relatively mediocre accuracies near the random baseline. This can be explained by a very small number of positive concept samples available in PH^2 dataset.

The TCAV score quantifies the positive or negative influence of a given concept towards a specific target class. Values above 0.5 indicate a positive influence of the concept on the prediction and lower values indicate negative influence. Figure 5.6 shows the TCAV scores achieved by evaluating 20 CAVs per concept on the *mixed_6h* layer

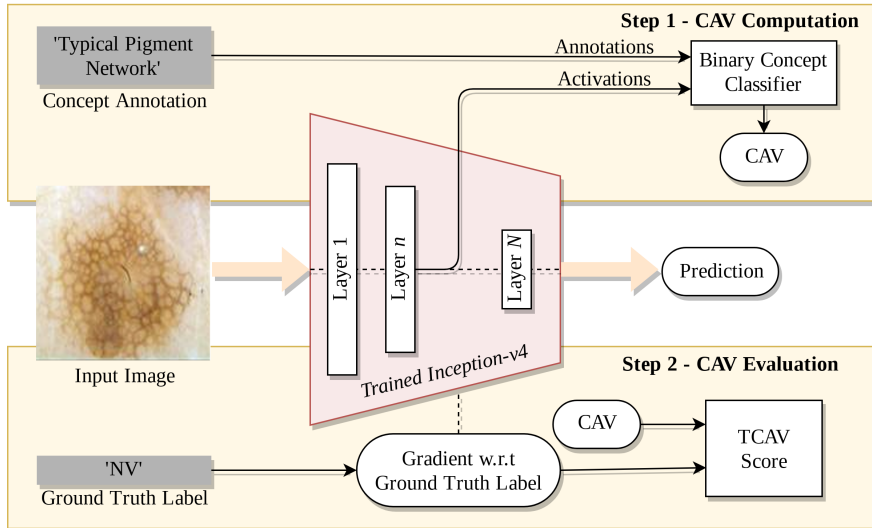


Figure 5.4: Overview of training concept classifiers and calculating CAV and TCAV scores

trained on derm7pt dataset. Average baseline scores of all 50 random concepts are again depicted by red horizontal lines along with their standard deviation in light red. Statistically insignificant results are marked by red asterisks.

The results for NV and MEL classes for concepts trained using derm7pt look very much as expected. Although the score for *PN* turned out to be insignificant in one experiment, features indicating benign melanocytic lesions like *PN_T*, *ST_R* and *DG_R* all contributed positively towards NV class. On the other hand, strong signs for malignant melanoma like *PN_AT*, *ST_IR*, *RS*, *DG_IR* and *BWV* show strong negative influence. Also, it is notable that the presence of Streaks in general (*ST*) has a stronger negative influence as compared to the presence of regular Streaks (*ST_R*). Results for MEL class show the exact opposite behaviour, which is perfectly aligned with the descriptions in the medical literature. It is again noticeable that the presence of Dots and Globules (*DG*) and the presence of Streaks (*ST*) show a higher positive impact on MEL class as compared to their regular kind, for example, regular Streaks (*ST_R*). The results for the SK class show similar concept influence as for MEL, except for (*PN*) exhibiting negative influence. In [365] the appearance of network-like structures in Seborrheic Keratosis has

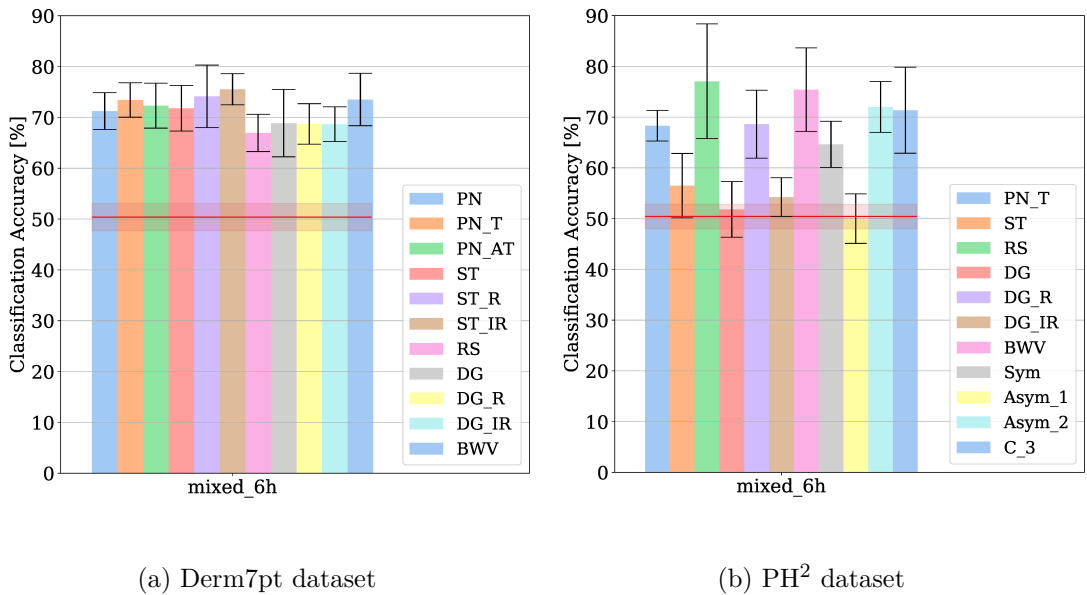


Figure 5.5: Validation accuracies of all concept classifiers trained and tested individually on derm7pt and PH² datasets. Random baseline is denoted by horizontal red line along with light red area marking standard deviation. Insignificant classifiers are marked with a red asterisk

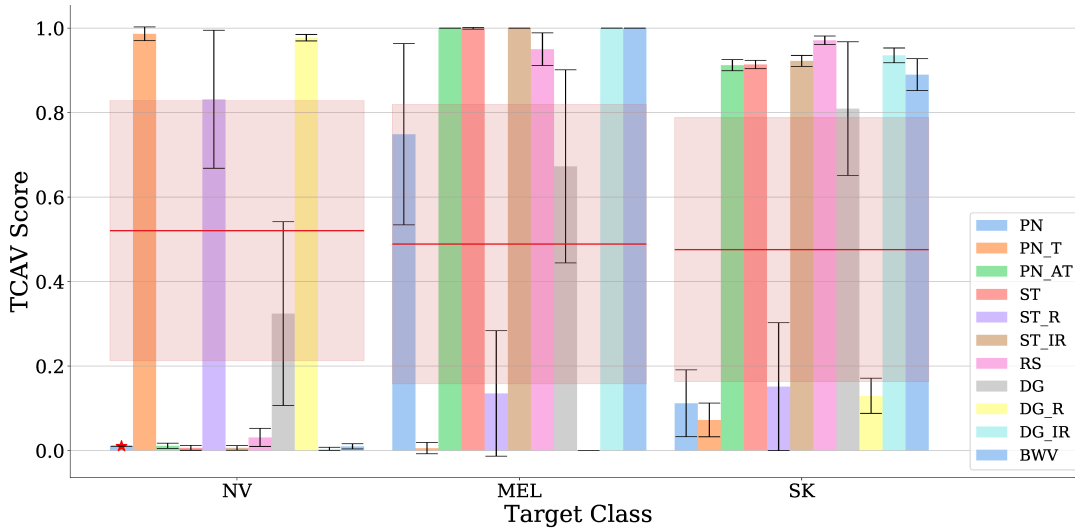


Figure 5.6: The TCAV scores of each concept for derm7pt with respect to each target class on *mixed_6h* layer of RECOD model

been confirmed. The model might have encoded those structures in the (PN_AT) concept, as their appearance slightly differs from the classical pigment networks definition. In the same study, evidence for Dots and blue-gray areas in SK lesions have been found as well.

Figure 5.7 shows resulting TCAV scores for CAVs trained on PH² dataset. All concepts achieving less than 55% validation accuracy have not been considered. Again, TCAV scores for NV and MEL show expected behaviour. Only typical Pigment Networks (PN_T), regular Dots and Globules (DG_R) and Symmetry (Sym) contribute positively towards Naevi class. For melanoma, the exact opposite holds again which can be confirmed by the concept descriptions in Section 5.3.2. Additionally, from the results, it appears that asymmetric lesions ($Asym_2$) and lesions containing more than three colours (C_3) tend to be classified as melanoma. For SK we can again observe the low influence of typical Pigment Networks (PN_T) as well as high influence for all other concepts including asymmetry ($Asym_2$) and colour diversity (C_3).

To further validate that the model has comprehensively learnt these disease-related concepts instead of learning something randomly, the model was made to sort all the test images with respect to the degree of visibility of a certain concept in each image. The model ordered all 300 test images starting from those that presented the very obvious existence of a concept and ending with those which had the least evidence of that concept. This ordering is performed based on Euclidean distance in a CAV’s direction.

Figure 5.8 through Figure 5.10 show the first five and the last five images from the sorted test set with respect to different concepts. The first row of each figure shows positive examples, where the concept is most clearly visible, and the second row shows negative examples, where the concept is virtually absent. It is evident from these figures that the proposed method for explaining skin disease classifiers does not only provide justification of classifier’s decision on global dataset scale but also sensibly identifies reasons for per-image predictions.

5.4 Mapping Concepts from Latent Space to Image Space

This section builds upon CAVs and extends it by introducing visual Concept Localisation Maps (CLMs), which are generated to locate human-understandable concepts that are learnt and encoded by a classifier in its latent space, in the input image. These CLMs validate that DNNs learn to focus on pertinent regions in the image while understanding relevant concepts. Furthermore, a new synthetic dataset called Simple Concept DataBase (SCDB) is developed, which consists of geometric shapes with annotations for 10 concepts and their segmentation maps. This dataset mimics complex relationships between concepts and classes in real-world skin lesion analysis tasks and can assist researchers in the classification and localization of complex concepts. These CLMs are

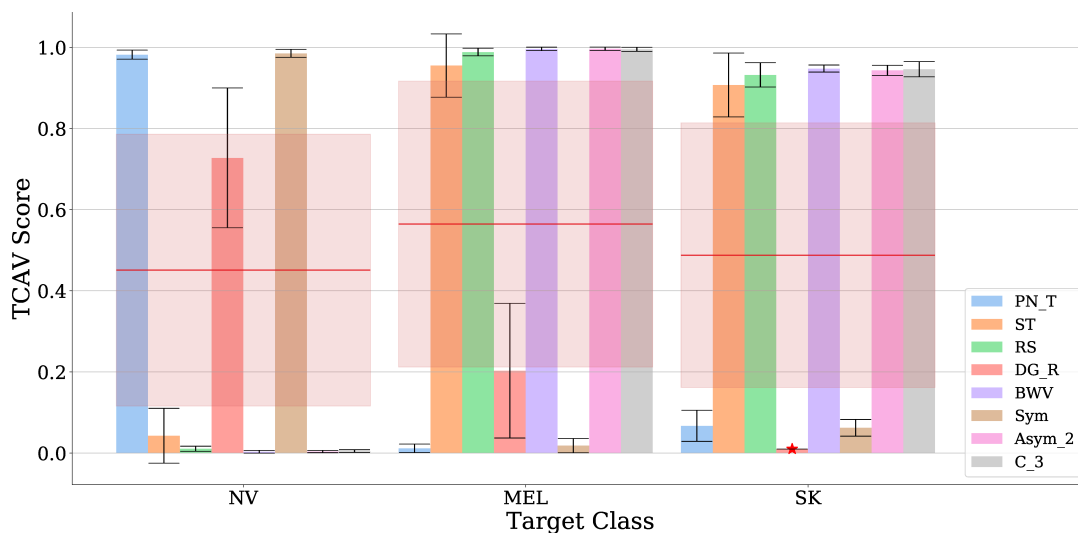


Figure 5.7: The TCAV scores of each concept for PH^2 with respect to each target class on *mixed_6h* layer of RECOD model

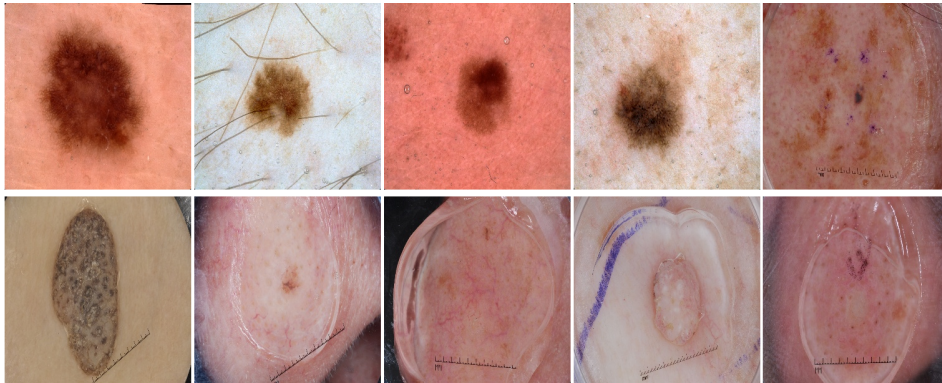


Figure 5.8: The sorting of the test images with respect to the presence of Typical Pigment Network (PN_T). The first row depicts the test images that show the strongest presence of Typical Pigment Network (PN_T). The second row shows the images with the weakest presence of this concept

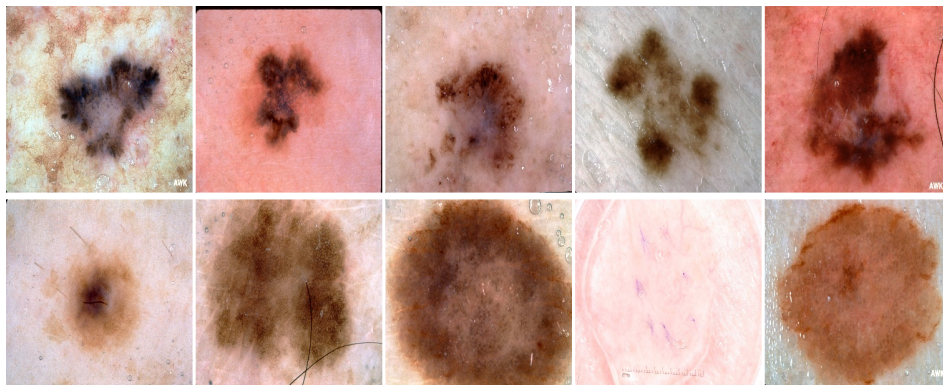


Figure 5.9: The sorting of the test images with respect to the presence of Irregular Streaks (ST_{IR}). The first row depicts the test images that show the strongest presence of Irregular Streaks (ST_{IR}). The second row shows the images with the weakest presence of this concept

qualitatively and quantitatively evaluated using three different model architectures i.e., VGG16, ResNet50, and SE-ResNeXt-50 trained on SCDB dataset to show that the proposed method works across different network architectures. The practicality of this method in real-world applications is also demonstrated by applying it on SE-ResNeXt-50 trained on CelebA dataset.

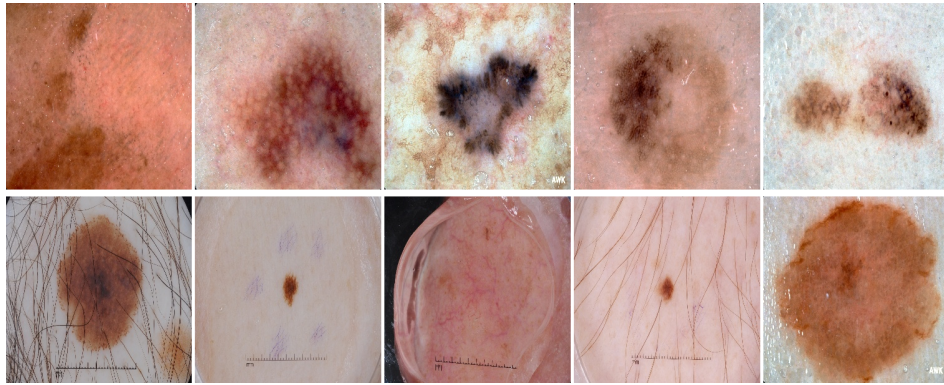


Figure 5.10: The sorting of the test images with respect to the presence of Regression Structure (RS). The first row depicts the test images that show the strongest presence of Regression Structure (RS). The second row shows the images with the weakest presence of this concept

5.4.1 Datasets for CLM Generation

5.4.1.1 SCDB: Simple Concept DataBase

Attribution methods proved to work well in simpler detection tasks where entities are spatially easy to separate [270, 271] but often fail to provide meaningful explanations in more complex and convoluted domains like dermatology, where concepts indicative of the predicted classes are spatially overlapping. Therefore, SCDB is developed and released ⁵, which is a new synthetic dataset of complex composition inspired by the challenges in skin lesion classification using dermoscopic images. In SCDB, skin lesions are modelled as randomly placed large geometric shapes (called base shapes) on black background. These base shapes are randomly rotated and have varying sizes and colours. The *disease biomarkers* indicative of the ground truth labels are given as combinations of smaller geometric shapes within a larger base shape. These *biomarkers* can appear in a variety of colours, shapes, orientations, and at different locations. Semi-transparent fill colour allows *biomarkers* to spatially overlap. The dataset has two defined classes, C1 and C2, indicated by different combinations of *biomarkers*. Class C1 is represented by joint presence of concepts $hexagon \wedge star$ or $ellipse \wedge star$ or $triangle \wedge ellipse \wedge starmarker$. Class C2 is characterised by joint presence of concepts $pentagon \wedge tripod$ or $star \wedge tripod$ or $rectangle \wedge star \wedge starmarker$. In addition to these combinations, additional *biomarkers* are randomly generated within the base shape without violating the classification rules. Two more *biomarkers* (i.e. *cross* and *line*) are randomly generated on the base shape

⁵<https://github.com/adriano-lucieri/SCDB>

without any relation to target classes. Finally, random shapes are generated outside of the base shape as noise.

The dataset consists of 7500 samples for binary image classification and is divided into train, validation, and test splits of 4800, 1200, and 1500 samples, respectively. Another 6000 images are provided separately for concept training. Along with each image, binary segmentation maps are generated and made available for every concept present in the image in order to evaluate concept localization performance. Segmentation maps are provided as the smallest circular area enclosing the *biomarker*. Figure 5.11 shows examples of SCDB dataset samples.

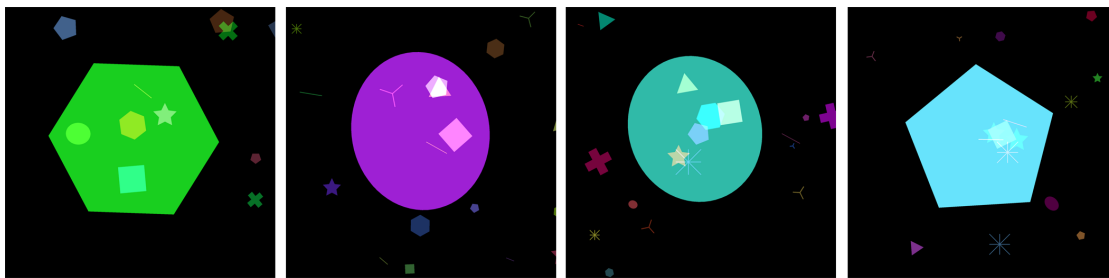


Figure 5.11: Training samples from SCDB dataset. Large hexagons, ellipses, and pentagons are examples of base shapes, akin to skin lesions; small squares, stars, and triangles etc. represent disease-related concepts.

5.4.1.2 CelebA

CelebA [366] is a dataset containing 202,599 face images each annotated with regards to 40 binary attributes. The dataset is split into train, validation, and test splits of 129,664, 32,415, and 40,520 samples, respectively. The images in each split are evenly divided with respect to gender labels. This dataset is chosen for qualitative evaluation because the gender annotation allows for solving a non-trivial, high-level classification task that relies on some of the fine-grained face-attributes like *baldness*, *moustache*, and *makeup*. An important aspect to consider while selecting datasets was to find a dataset that not only contains annotations of fine-grained concepts but also high-level concepts that can be reflected by solving an interim task of fine-grained concept detection. Gender annotations in CelebA dataset allow solving a non-trivial classification task that relies on some of the remaining annotated concepts like *baldness*, *moustache*, *lipstick* and *makeup*, which statistically suggests the gender in the given data distribution.

5.4.1.3 Dermatology Datasets

Two skin lesion datasets with rich dermoscopic criteria annotations, namely PH^2 dataset and `derm7pt` are also used. Both datasets are merged for evaluation, resulting in 1023 images of melanoma and nevus.

5.4.2 Concept Localisation Maps

The CLM method obtains a localisation map m_{Cl} for a concept C learnt on DNN’s layer l , that locates the relevant region essential for the prediction of a concept classifier $g_C(f_l(x; \theta))$ given an input image $x \in X$. The linear concept classifier g_C generates a concept score for concept C given a latent vector of trained DNN $f_l(x; \theta)$ with optimal weights θ at layer l . The resulting map m_{Cl} corresponds to the region in the latent space of DNN that encodes the concept C .

5.4.2.1 g-CLM

To apply gradient-based attribution methods for concept localization a binary mask m_{binC} is required that filters out latent dimensions that contribute the least to the classification of concept C . For each concept, those dimensions are determined by thresholding the concept classifier’s weight vector ν_C , also known as CAV. High absolute weight values imply a stronger influence of the latent feature dimension on the concept prediction and shall thus be retained. Therefore, a threshold value T_C is computed automatically based on the 90th percentile of weight values in ν_C .

Gradient-based attribution methods are applied once the latent feature dimension is masked and the concept-relevant latent subset $f_{lC}(x, \theta)$ is obtained. The methods evaluated in this work apply SmoothGrad² [367] and VarGrad [325] as ensembling approaches using plain input gradients as base-estimator. The noise vector $g_j \sim \mathcal{N}(\mathbf{0}, \sigma^2)$ is drawn from a normal distribution and sampling is repeated $N = 15$ times. SmoothGrad² and VarGrad were proven to be superior to the classical SmoothGrad [322] in terms of trustworthiness and spatial density of attribution maps. Henceforth, all experiments referring to gradient-based CLM will be denoted by g-CLM.

5.4.2.2 p-CLM

The application of perturbation-based attribution methods requires local manipulation of the input image to observe changes in prediction output. In the case of CLM, the output is the predicted score of the concept classifier instead of the image classifier.

The systematic occlusion method from [273] is used in all experiments with a patch-size of 30 and stride of 10 since it provides a good trade-off between the smoothness of obtained maps and localization performance. Occluded areas are replaced by black patches. Experiments referring to the perturbation-based CLM method are denoted as p-CLM.

5.4.3 Experiments and Results

Three DNN types, namely VGG16, ResNet50, and SE-ResNeXt-50 are examined using CLM to study the influence of architectural complexity on concept representation and localisation. All models were initialised with weights pre-trained on ImageNet. Hyperparameter tuning on optimiser and Initial Learning Rate (ILR) provided best results for optimisation using RMSprop [368] with ILR of 10^{-4} . Experiments were conducted for a maximum of 100 epochs using learning rate decay with factor 0.5 and tolerance of 5 epochs, and early stopping if no improvement in the validation loss is achieved after 10 epochs.

The VGG16, ResNet50, and SE-ResNeXt-50 achieved 97.5%, 93.5%, and 95.6% image classification accuracy and 85.7%, 81.1%, and 72.8% concept classification accuracy, respectively. Surprisingly, the simplest and shallowest architecture achieved the highest test accuracy. However, the average concept classification accuracies on the architectures' last pooling layers (*pool5*) indicate that complex architectures possess more informed representations of concepts.

Figure 5.12 shows some examples of SCDB along with generated CLMs. Rows two and three correspond to g-CLM (SG-SQ) and p-CLM, respectively. The examples presented in this figure reveal that g-CLMs can be used to localise concepts in many cases. However, it appears that the method often highlights additional *biomarkers* that do not correspond to the investigated concept. For some concepts, localisation was not successful for almost all examples. Furthermore, the generated maps appear to be sparse and distributed, which is typical for methods based on input gradients. The heatmaps obtained from p-CLM are extremely meaningful and descriptive, as shown in the last row of Fig. 5.12. The granularity of these heatmaps is restricted by the computational cost (through chosen patch size and stride) as well as the average concept size on the image. The method can separate the contributions of specific image regions to the prediction of a certain concept. This even holds if shapes are overlapping.

5.4.3.1 Quantitative Evaluation:

To quantify CLMs performance, average IoU, precision, and recall are computed between predicted CLMs and their respective ground truth masks for all images in the validation set of SCDB dataset. The predicted CLMs are binarised using a per-map threshold from the 98th percentile. The metrics are computed for all images with a positive concept ground truth which means that images with incorrect concept prediction are included as well. Average results for all 10 concepts for all networks and variants are presented in Fig. 5.13. Concept localisation performance of all methods increased with the model complexity. This suggests that concept representations are most accurate in SE-ResNeXt-50. The results also clearly show that both variants of g-CLM are outperformed by p-CLM over all networks, achieving the best average localization recall of 68% for all 10 concepts, followed by g-CLM (SG-SQ) with 38% and g-CLM (VarGrad) with 36%. Most concepts relevant to the classification achieved recalls over 80% with p-CLM. The best IoU of 26% is also scored by p-CLM. It needs to be noted that IoU is an

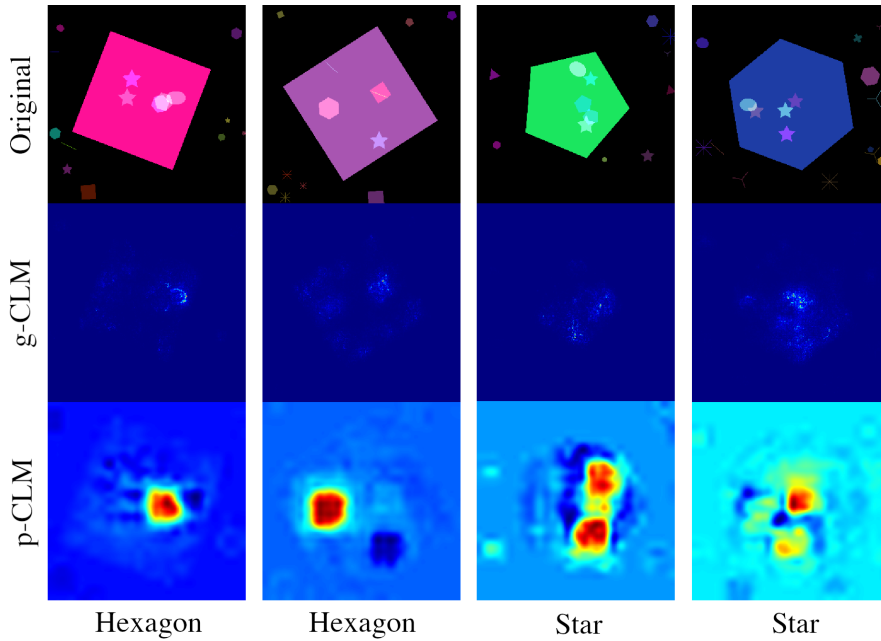


Figure 5.12: Concept Localisation Maps (CLMs) for SCDB images. Input images are shown in the first row along with corresponding concept localisation maps from SE-ResNeXt-50 on layer *pool5*. The middle and the bottom rows show corresponding g-CLM (SG-SQ) and p-CLM, respectively. The respective concept names for the CLM computation is given below each column

imperfect measure considering the sparsity of gradient-based CLMs and the granularity of p-CLMs.

IoU scores are mostly low with *starmarker* achieving the best value of only 0.38. This can have several reasons. First, the patch size for occlusion was chosen to be 30×30 pixels to avoid excessive computation time resulting in relatively coarse CLMs. Secondly, the IoU may be sensitive to the binarisation threshold. However, a better indicator for the viability of the method is the recall, as it describes the portion of concept pixels that are correctly localised. For many concepts, recalls over 80% were achieved. It is interesting that SE-ResNeXt-50 and ResNet50 show better overall concept localisation. This includes concepts like *cross* and *line* that are uninformative for the target task. Furthermore, SE-ResNeXt-50 is the only architecture that shows constant localisation performance in the last three layers.

Both qualitative and quantitative analyses suggest that the performance of CLM and thus the representation of concepts is improved with the complexity of the model architecture. This finding is contrary to the recent claims by Hu et al. [369]. They concluded that simpler architectures allow for easier disentanglement and are therefore more interpretable, comparing VGG16 to ResNet and DenseNet.

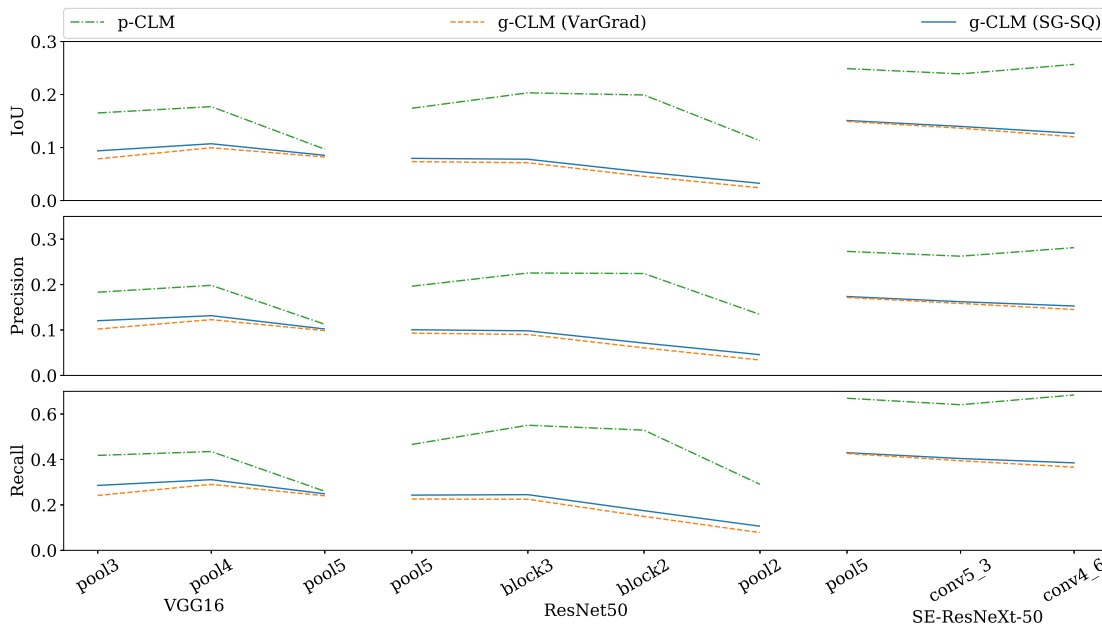


Figure 5.13: Average IOU, precision and recall over all 10 concepts for predicted CLMs applied to three network architectures

5.4.3.2 Evaluation on CelebA Dataset

Learning from the experiments on SCDB, the *SE-ResNeXt-50* model was trained on the binary gender classification task using CelebA. The resulting network achieved 98.6% accuracy on the test split. Concepts that achieved highest accuracies are often strongly related to single classes like facial hair (e.g. *goatee*, *moustache*, *beard* and *sideburns*) or makeup (e.g. *heavy makeup*, *rosy cheeks* and *lipstick*). Figure 5.14 shows images with their corresponding CLMs generated with the proposed method. Due to the absence of ground truth segmentation masks in this dataset, results are only discussed qualitatively.

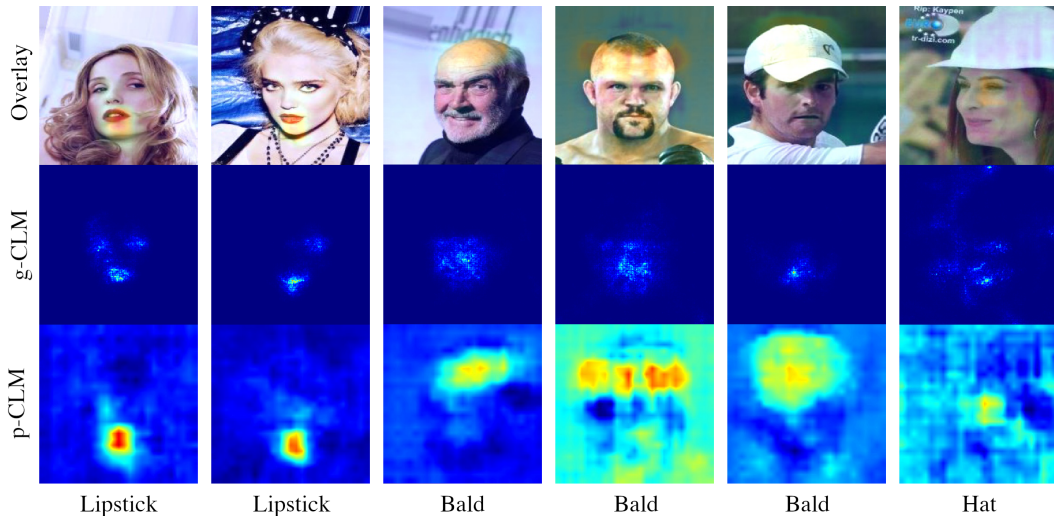


Figure 5.14: Examples for CLMs generated from *SE-ResNeXt-50* trained on binary classification of gender with CelebA dataset. The first row shows the original images with the heatmap overlay of p-CLM, the second row shows g-CLM (SG-SQ) and the last row shows p-CLM. The respective concept names for the CLM computation is given below each column

The first two columns in Fig. 5.14 show examples of CLMs for the *lipstick* concept. Although it is quite likely that the network learnt a more abstract notion of female and male lips for the classification, the robust localisation indicates that the network indeed encodes a lip-related concept in the learnt CAV direction. It is striking that g-CLM often fails and highlights the cheeks as well.

All concepts related to facial hair achieved concept accuracies exceeding 80%. However, inspecting the generated CLMs reveals that the CAVs do not properly correspond with the nuances in concept definitions. The localisation maps reveal that the concept *sideburns* never actually locates sideburns but beards in general. For the *goatee* and *moustache* it can be observed that a distinction between both is rarely made. It is thus

very likely that the network learned a general representation of facial hair instead of different styles, as it would not aid in solving the target task of classifying males versus females.

The *bald* concept produces almost perfect p-CLMs focusing on the forehead and bald areas, as can be seen in columns 3 and 4 of Fig. 5.14. It perfectly demonstrates how the network learnt an intermediate-level feature from raw input that is strongly correlated to a target class. However, this also shows a tendency to classify the male class when too much forehead is detected. An intriguing finding is that hats are often confused with baldness, as shown in the CLM for *baldness* in the second last column. However, g-CLM consistently failed to locate this concept. In addition to being sometimes mistaken for baldness, the CLM for *hat* in the last column shows that the network struggled to learn the correct representation of a hat.

5.5 The ExAID Framework: Providing Multi-modal Explanations

Explanation methods for AI come in a variety of forms and provide explanations using a range of modalities such as visualisations [273], text [302], or quantitative relevance measures for abstract concepts [292]. They differ not only in the way they are presented to the users but also in their derivation, resulting in varying levels of insight provided regarding the decision-making of the AI. However, most model explanations given by a single XAI method are usually not sufficient to provide plausible and easy-to-understand decision justification to the end-users.

This section presents a framework called Explainable Artificial Intelligence for Dermatology (ExAID) which is able to provide easy-to-understand textual, visual, and conceptual explanations for automated analysis of dermoscopic images of malignant and benign skin lesions. This framework is built upon the works explained in section 5.3, which verifies that deep learning models are able to learn and utilise similar disease-related concepts as described by dermatologists and employed by them during the manual analysis of dermoscopic images; and section 5.4, which localises these concepts, learnt and embedded in the latent space of the model, on the original image space. The ExAID extends these explanation modalities by introducing concept-based textual explanations and integrates all modalities in a unified framework to further enhance the intelligibility of AI's decision-making in a diagnostic setting, providing in-depth analysis tools for medical researchers and students. The framework offers two distinct interfaces for clin-

ical diagnosis and research purposes, laying the foundation for the understandable and transparent integration of AI in medical workflows.

5.5.1 Datasets for Skin and Concept Classification

The ExAID contains two types of classifiers: Disease-level classifiers for lesion diagnosis and concept-level classifiers for detection of dermatological concepts in a given image. To train these two classifiers, it requires datasets with two types of labels, namely disease labels, like *Melanoma* and *Nevus*, and concept annotations, for example, presence or absence of dermoscopic concepts.

5.5.1.1 Datasets for Disease-level Classification

The training set for disease-level classification consists of Melanoma and Nevi images taken from ISIC 2019, PH², and derm7pt datasets. These datasets are already described in the previous sections of this chapter. A brief account of their usage and distribution in this section is given below.

ISIC 2019 dataset is a public collection of 25,331 images of different provenance divided into eight different classes. This dataset is a coalition of three datasets, HAM10000 [91], BCN20000 [370], and MSK [111]. Since the common denominator of ISIC2019, PH², and derm7pt datasets are Melanoma and Nevi classes, a subset of the three datasets was assembled consisting of images from these two classes only. The subset was manually cleansed for duplicates and samples with low quality, e.g. systematic artefacts, resulting in a total of 6475 images. As PH² and derm7pt are used for training the concept-level classifiers, a custom dataset split is gathered from a combination of all three stratified datasets to avoid covariate shifts between disease-level and concept-level training stages. The distribution of images in training, validation, and test sets for disease-level classification is given in Table 5.2. Additionally, the generalisability of the model is evaluated on a range of other datasets including 2016 and 2017 ISIC challenge datasets and SKINL2 [371] dataset.

5.5.1.2 Datasets for Concept-level Classification

Training of concept classifiers requires annotations regarding the presence or absence of specific dermoscopic concepts. These annotations are not usually available with dermoscopic image datasets, which limits the selection of training and evaluation datasets primarily to PH² and derm7pt. In PH² dataset, colour, and lesion segmentation masks

Table 5.2: The distribution of data in training, validation and test splits for disease-level classification

Split	Dataset	Lesions		
		Melanoma	Nevi	Total
Train	ISIC2019	1250	2894	4144
	Derm7pt	158	368	526
	PH2	26	102	128
Validate	ISIC2019	313	723	1036
	Derm7pt	40	92	132
	PH2	6	26	32
Test	ISIC2019	391	904	1295
	Derm7pt	50	115	165
	PH2	8	32	40
Total	ISIC2019	1954	4521	6475
	Derm7pt	248	575	823
	PH2	40	160	200

and extensive, well-curated annotations with respect to the presence or absence of various concepts are given for each image. From derm7pt dataset, 823 images belonging to Melanoma and Naevi classes are selected. The combination of derm7pt and PH² used for concept classification is subsequently referred to as D7PH2. Table 5.3 shows the distribution of images used in the concept-level classification task. The ISIC 2016 and 2017 challenge datasets are also used for the evaluation of the concept classifier’s generalisability. However, both datasets only include annotations of two dermoscopic concepts each, namely Pigment Networks and Streaks as well as Dots & Globules and Streaks, respectively.

Table 5.3: The distribution of data in training, validation and test splits for concept-level classification with D7PH2 dataset

Split	Dataset	Lesions		
		Melanoma	Nevi	Total
Train	Derm7pt	158	368	526
	PH2	26	102	128
Validate	Derm7pt	40	92	132
	PH2	6	26	32
Test	Derm7pt	50	115	165
	PH2	8	32	40
Total	Derm7pt	248	575	823
	PH2	40	160	200

5.5.2 Components of the Framework

At its core, ExAID is a generic toolbox for human-centred post-hoc explanations capable of explaining arbitrary DL-based models even beyond applications in dermatology. In addition to the DL model to be explained, its computational foundation consists of three basic components, namely Concept Identification, Concept Localisation, and Decision Explanation modules as depicted in Fig. 5.15.

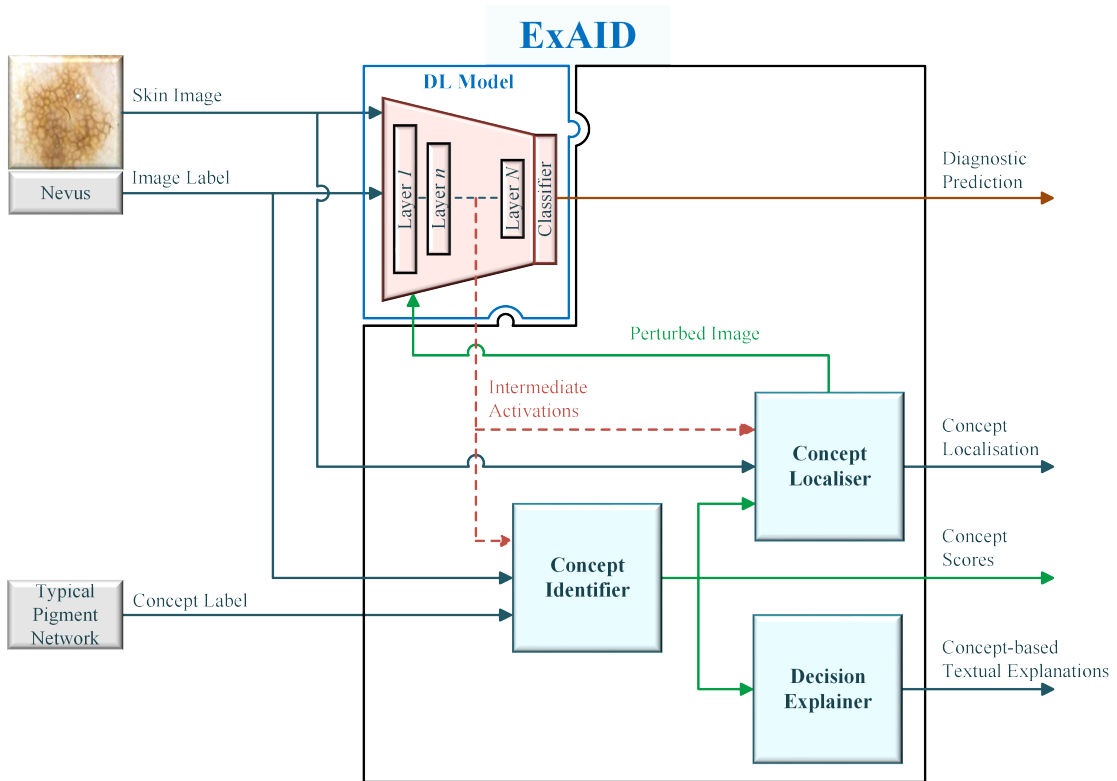


Figure 5.15: The ExAID Framework architecture. The schematic drawing shows the input, output and the flow of information through ExAID as well as the relationship between its components

5.5.2.1 Concept Identifier

The Concept Identifier module is based on the work discussed in detail in section 5.3. It maps disease-related dermatological concepts to their corresponding representations learnt by the DL-based model in its latent space using Concept Activation Vectors [292] (CAVs). For each pre-defined concept, a linear binary classifier is trained on the detection of the said concept from the model's activation space, resulting in a CAV that

represents the main concept direction in this latent space. The CAV training can be executed on arbitrary model layers, automatically selecting each concept’s best performing activations, for inference. Once trained, the concept classifiers allow predicting the presence or absence of individual concepts on unseen images, based on the model’s latent activation patterns. The CAVs additionally allow for computation of the global TCAV metric, estimating a concept’s overall contribution to the prediction of a certain target class.

5.5.2.2 Concept Localiser

Concept Localisation Maps (CLMs), as explained in section 5.4, extend CAVs by localising regions pertinent to a learned concept in the latent space of a trained image classifier. They provide qualitative and quantitative assurance of the model’s ability to learn the right interpretation of a concept by indicating the exact spatial location that contributed to a concept prediction and moreover enable the visualisation of other, potentially abstract, concepts.

5.5.2.3 Decision Explainer

The Decision Explainer receives all concept prediction scores for a given image from the Concept Identifier. A rule-base is derived from a calibration dataset and applied to the translation of single concept scores into a textual decision explanation grounded in human-understandable conceptual evidence. The explanations derived from concept detection are composed of coherent and easy-to-understand explanation texts. An explanation sentence is constructed based on concept scores and directional derivatives computed during concept detection under discrimination between absence, moderate evidence, and strong evidence of concepts to reflect the fuzzy nature of concepts’ appearance. Manifestation of a concept is decided by means of thresholds derived from the concept training data. This is achieved by first scaling the unbound concept prediction using a two-sided normalisation scheme to obtain a centred probability of concept presence. Thresholds are then derived by maximising False Positive Rate (FPR) and True Positive Rate (TPR) among all positive predictions on the training dataset for moderate and strong evidence thresholds, respectively. The directional derivatives of the predicted class along with the individual CAV are used to indicate a positive or negative influence of concept on the prediction. Conceptual evidence is mentioned after the keyword "despite" in the case of negative class influence to signalise contraindication.

5.5.3 Operation Modes

The ExAID offers two complementary operation modes that are meant for different use cases. The diagnostic mode provides functionality meant to support dermatologists during clinical examination of patient’s skin lesions. For research and education purposes, ExAID offers an educational mode including a collection of tools for holistic analysis of the deep model’s behaviour as well as the collected case data.

5.5.3.1 Diagnostic Mode

The majority of a dermatologist’s clinical routine consists of a visual examination of patients’ skin lesions to reach a decision regarding the further investigation of a potentially malignant lesion. Provided enough evidence for malignancy is available, the suspicious tissue may be excised under local anaesthesia. Physicians with considerable experience in dermoscopy develop an intuition that enables them to promptly reach a conclusion while novices initially need to pay greater attention to the assessment of a particular skin lesion. This is among other things owing to the disarray of dermoscopic terms and concepts and their usage in different schools of thought. Having developed a routine and diagnostic intuition not only bears the risk of subjective bias in a decision, it might also lead to negligence in the identification of important diagnostic details, which is furthermore aggravated by emotional stress and time constraints.

The diagnostic mode of ExAID aims at mediating subjectivity by offering a supplement to the experienced physician’s first impressions, serving as a second opinion that stimulates the physician’s thought and breaks the routine. Through this additional information, it is made sure that cues vital for the successful identification of malignant conditions are not overseen during manual examination. The user interface of the diagnostic mode is presented in Fig 5.16. While allowing dermatologists to examine the dermoscopic image manually, an initial diagnosis suggestion is provided supported by concept-based textual, quantitative, and visual explanations. Through its neutral design, the interface assures that users are not biased towards the proposed diagnosis but are free to reconstruct the AI’s decision process by considering and validating biomarker scores along with their optional localisations provided in the form of CLMs.

5.5.3.2 Educational Mode

Explanations of classifiers’ decisions in ExAID have further utility beyond mere information and guidance of the algorithm’s users. It is of central importance for the validation of individual automated decisions and verification of plausibility of a model’s global

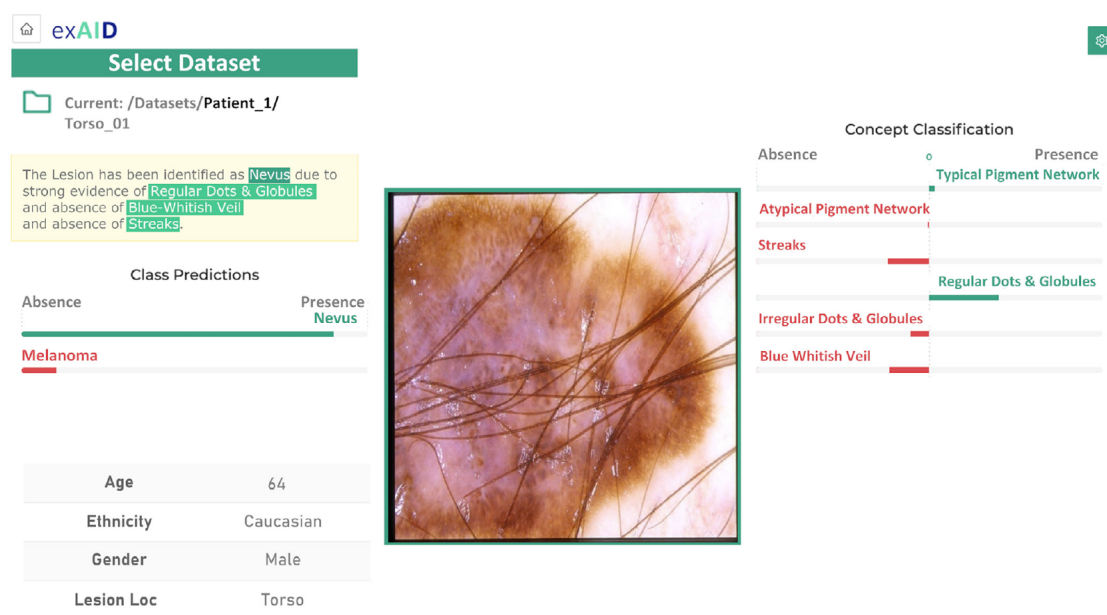


Figure 5.16: Diagnostic mode of ExAID can be used as decision support system in routine clinical workflows

generalisation behaviour and can additionally aid in the decryption of unintelligible decision-relevant concepts learned by the AI. With its educational mode, as presented in Fig. 5.17, ExAID offers an extensive toolbox for the investigation of model behaviour and data distribution. Dataset-level model behaviour analysis is enabled through a combination of class-wise performance evaluation metrics and concept-wise global explanation metrics in combination with tools for facilitated overview of individual decision outcomes and explanations.

5.5.3.3 Interactive Features of ExAID

Some of the most salient interactive features of the ExAID framework are introduced below.

Filtering The filtering option allows filtering arbitrary subsets of samples by metadata such as age, concept presence, concept prediction, or correct prediction. An adaptive data distribution plot helps to quickly identify important statistical characteristics related to biomarker presence as well as certain failure modes of the model.

Highlighting A highlighting feature allows spotlighting certain useful properties of samples to further facilitate the review of model behaviour and data. This feature allows the highlighting of not only binary attributes such as the correct target class prediction but also more complex relationships such as the presence of classes or concepts in the annotations as well as the class and concept prediction by the model. Complex highlighting is always supported by visual cues indicating the accordance of attribute prediction with expert annotations.

Localisation In addition to individual localisation of concepts in data samples, ExAID allows to visualise concept localisation simultaneously for all samples of a dataset. This allows for quick examination of a model’s concept localisation behaviour, aiding the validation of system behaviour and identification of potential systematic errors in the dataset or model by revealing patterns in the localisation process.

Latent Inspection Examination of the model’s latent space structure gives further insight into the disentanglement of data representations and potential biases captured by the model parameters. A latent view functionality based on Tensorboard’s projector⁶ allows to intuitively examine the latent distribution of data samples by means of

⁶<https://projector.tensorflow.org/>

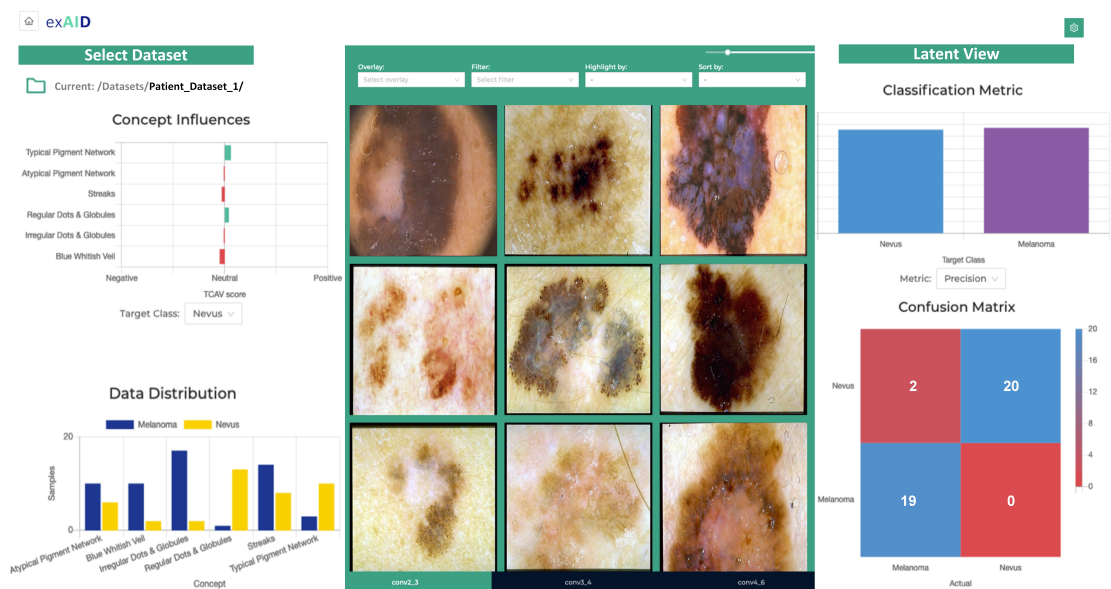


Figure 5.17: Educational mode of ExAID can help in training of resident dermatologists by allowing them to explore many of its interactive features.

dimensionality reduction techniques.

5.5.4 Experiments and Results

5.5.4.1 Classifier Training & Evaluation

To demonstrate the utility of the proposed framework, a deep network for binary classification of Melanoma and Naevi from dermoscopic skin lesion images is trained. Experimentation included VGG16, ResNet, DenseNet, NASNet, SEResNeXt architectures with Adam, SGD, and RMSprop optimisers using learning rates ranging from 1e-3 to 1e-4. Among various architecture, learning rate, and optimiser combinations, the best results are achieved using SEResNeXt architecture with RMSprop optimiser and a learning rate of 1e-4 trained for 100 epochs. Training images were augmented by random horizontal and vertical flip as well as random cropping to 85% of the image size, resulting in input images of size 224×224.

Evaluation on a variety of datasets is presented in Table 5.4. Results clearly show the strong generalization capability of the model, even on unseen datasets like SKINL2, consisting of high-quality images of 20 Melanomas and 35 Naevi. Poor performance on the ISIC2017 test dataset can be explained by the large fraction of artefacts present in the images, which have been intentionally left out of the training procedure to restrict the use case to a realistic, adapted environment focusing on the image acquisition specifically for AI processing.

Table 5.4: Performance evaluation of lesion classifier on various datasets

Datasets	Size	Accuracy (%)	Precision (%)	Recall (%)	AUC
Derm7pt (Test)	165	83.6	81.7	78.0	0.85
PH2 (Test)	40	100.0	100.0	100.0	1.00
ISIC2019 (Test)	1295	88.9	88.2	84.9	0.91
ISIC2017 (Test)	510	78.4	68.5	62.3	0.70
ISIC2016 (Test)	379	89.7	83.7	84.0	0.92
SkinL2	55	90.9	89.9	90.7	0.99

5.5.4.2 Explanation Training and Evaluation

For the explanation of the final DL-based classifier’s decisions, the procedure outlined in section 5.3 is followed. To this end, concept annotated samples from D7PH2 have been utilised to assure generalisation while learning CAVs. In each run, the data are internally

split into folds for concept training and validation under stratification of both concept and disease labels. Linear concept classifiers are trained for 200 runs using stochastic gradient descent with early stopping for each concept.

5.5.4.3 Concept Detection

The final CAV for a concept is chosen based on the average concept direction for all runs. Due to concept annotation requirements, concept detection performance is evaluated only on ISIC 2016 and ISIC 2017 datasets as well as D7PH2 test set. Table 5.5 presents Macro Average F1-Scores for concept detection.

Table 5.5: Performance evaluation of concept classifiers on various datasets. The results are given as macro average F1-Scores (%) to account for class imbalance.

Datasets	Streaks	Pigment Networks	Dots & Globules	Regression Structures	Blue-Whitish Veil
derm7pt (test)	70.91	78.74	63.14	59.41	71.66
ISIC-2017	51.75	50.37	-	-	-
ISIC-2016	56.53	-	53.03	-	-

It can be seen that concept generalisation to unseen datasets such as ISIC 2017 and ISIC 2016 is poor. This is most likely a consequence of diverging annotation standards between the derm7pt and PH² datasets used for CAV training and furthermore influenced by artefacts present in the challenge test sets. Moreover, results show the superiority of coarse-grained biomarkers such as Streaks, Pigment Networks, and Blue-Whitish Veils over more fine-grained ones such as Dots & Globules.

5.5.4.4 Concept Localisation

Fair quantitative evaluation of a network’s CLMs for skin lesions poses a number of difficulties including the selection of a suitable binarisation scheme, subjectivity of concept annotations as well as lack of representative metrics for fuzzy localisation tasks. Proper binarisation is especially difficult as it depends on the size of a particular ROI, its significance to the prediction score as well as further noise stemming from the saliency method used. Furthermore, evaluation is limited by the availability of annotated concept segmentation maps. The ISIC 2016 and ISIC 2017 challenge datasets each provide concept segmentation maps for two concepts which are used to provide a qualitative assessment of the trained model’s concept localisation ability. The CLMs were binarised

using variable percentiles, which are manually chosen based on the size of the respective ROI in a specific image.

Figure 5.18 shows examples of the model’s concept localisation ability for concept classes Streaks and Pigment Network using an adaptation of the method proposed in 5.4. Whereas in some cases, CLM localisation aligned very well with the concept annotations, most of the time CLMs highlighted slightly different regions. However, these highlights often depict areas that could plausibly count as concept regions, as can be seen in the second row of Fig. 5.18. The qualitative evaluation confirmed quantitative results and showed that the network performed better in localising concepts Streaks and Pigment Networks as compared to the more fine-grained Dots & Globules concept. Scattered spots in CLMs outside the lesion regions highlight noise problems inherent in perturbation-based CLM computation and the dependence on a proper binarisation scheme, especially when quantitatively evaluating the maps.

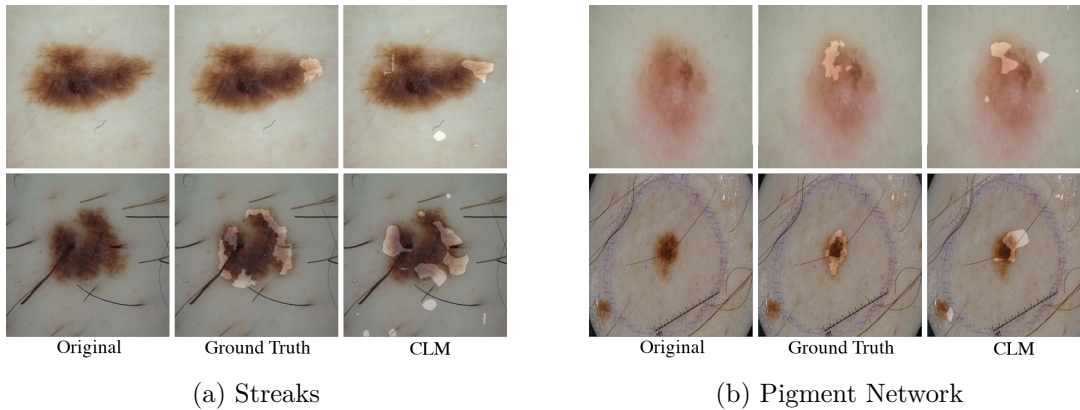


Figure 5.18: Positive and negative examples of visual explanations provided by ExAID along with the corresponding samples and ground truth concept masks

5.5.4.5 Textual Explanation

Quantitative evaluation of textual explanation results is performed based on the performance evaluation for concept detection presented in Table 5.5. A qualitative evaluation of these textual explanations is provided below.

Figure 5.19 shows examples of images along with correct and incorrect textual explanations provided by ExAID. The examples in Fig. 5.19a showcase the simplicity and intelligibility of the generated explanations. These explanations reflect the most important criteria necessary for experts to understand the network decision. Interestingly, it appeared that although correct concept predictions were given, the network sometimes

misclassified the underlying disease as seen in the third row of Fig. 5.19a. This could be a result of wrong ground truth annotation or the presentation of an ambiguous borderline case. However, the explanation exposes Streaks, Irregular Dots & Globules, and Blue-Whitish Veil as contraindications for the prediction of Naevus. In a clinical setting, such contraindication would raise the suspicion of a user, possibly initiating a more thorough review of the case. This emphasises the utility of such a system since a correct explanation will allow physicians to scrutinise a given prediction instead of solely relying on an automated opaque categorical output value.

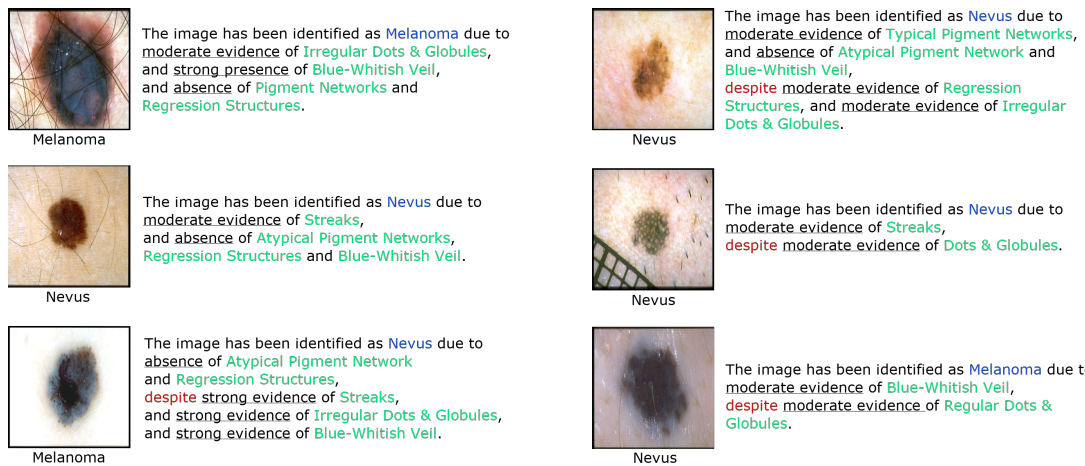


Figure 5.19: Positive and negative examples of textual explanations are provided by ExAID along with the corresponding skin lesion samples. The ground truth class of the sample is given below the image

Figure 5.19b on the contrary shows failure cases where the network confused different visual cues for concepts. Whereas Irregular Dots & Globules have been correctly detected in the top-right image, the middle-right image contains white blobs which might have been confused as Dots & Globules by the model. The bottom right case shows a Blue Naevus which has been confused by the network as a Melanoma showing signs of Blue-Whitish Veil although it is actually containing Regular Dots & Globules. It is also evident that samples with incorrect concept predictions already expose a certain uncertainty by exhibiting moderate concept detections more frequently as compared to the samples from Fig. 5.19a as well as their prevalence of contraindications. This shows that irrespective of the model used for the prediction, ExAID can provide well-founded justifications which help to express model uncertainty, encouraging closer examination of rare and borderline cases.

5.5.5 Limitations

This study on multi-modal easy-to-understand explainable CAD system presents a unified framework that primarily focuses on the comprehensible user interface, conveying textual, visual, and conceptual explanations for reliable DSS in dermatology. Although concept classification, localisation, and textual explanation abilities of ExAID are remarkable given the fact that the DL model has not explicitly been trained on those tasks, some challenges must be resolved before an application in real clinical settings becomes feasible.

Current public datasets often suffer from low sample quality that can be attributed to a lack of process standardisation – different camera setups, operators, and techniques like polarised and non-polarised dermoscopy resulting in varying image quality, lighting, alignment, and artefacts – that can lead to uncertain diagnosis and subjective annotation. Together with the low number of overall available images, particularly those with detailed concept annotation, this results in a significant shift of data distributions between datasets, which could have been a major reason for the sub-optimal generalisation of the proposed concept classifiers to other datasets.

The concept localisation ability of ExAID currently suffers from limitations stemming from the perturbation-based nature of saliency map generation which results in noisy heatmaps and high sensitivity to hyperparameters, especially in case of varying biomarker size. Future work applying optimisation-based perturbation methods for concept localisation can mitigate these issues and result in more flexible and robust heatmaps. Textual explanations are generated based on concept predictions as well as directional derivatives as used in TCAV scores. Lacking a meaningful scaling of gradients, only the direction without the magnitude of a concept’s influence is currently used to improve the explanation text. Incorporation of more robust concept influence measures could add another level of details to the rule-base, making the explanations more differentiated and rendering the system even more useful in practice.

Quantitative evaluation of concept detection or localisation is still limited due to the lack of similarly and sufficiently annotated data from other sources. To solve this issue, an agreed-upon definition and consensual annotation of a large number of representative images are required, which will reflect higher-quality explanations. Moreover, evaluation of CLMs is aggravated by noise artefacts emerging during binarisation and a lack of definite measures for fuzzy localisation tasks. A qualitative evaluation in the real-world setting by medical experts is of extreme value for the evaluation of the explanations’ utility to the diagnostic workflow.

The influence of subjectivity is not only reflected in the data annotations, but also the general uncertainty surrounding the field of dermoscopy. Despite first attempts towards standardisation of dermoscopic terminologies and concepts [360], no general consensus has yet been broadly established among physicians. Thus, a variety of diagnostic schools exist and interpretation of terms and concepts is still largely depending on the education, preference, and experience of the individual dermatologists. This work focused on the 7-point checklist criteria [372] as well as further dermoscopic concepts from [57], due to the public availability of annotated data.

The commitment to a specific set of concepts before the decision of a standard consensus might hamper the acceptance of the framework by physicians accustomed to various methods and the mixture of different schools and interpretations of concepts bears the risk of contrasting labelling. Productive deployment of such a system requires diligent assessment by medical practitioners in real-world environments and providing their valuable feedback to evaluate and improve such a system. Prior to performing clinical trials, the system should be fed with carefully selected data properly representing a set of meaningful and unambiguously defined dermoscopic concepts as agreed by a large body of dermatology experts.

5.6 Understanding Glaucoma Diagnosis using GradCAM

The ExAID framework described in the previous section can be used on datasets which clearly define clinical concepts used for disease classification and provide annotations for such concepts. However, other medical domains, like glaucoma classification, may not have such elaborately defined concepts attributed with the disease and therefore no such concept annotations are available with such datasets. To explain the predictions of DL models in such use cases, one can resort to simpler explanation methods as described in section 5.2.1.

GradCAM method was used to visualise the input regions, which were deemed the most significant by Inception V3 trained on G1020 and ORIGA datasets (refer to section 2.3 for detailed experimental setup). Figure 5.20 shows some sample images along with their GradCAM results for correctly classified healthy and glaucoma images from both datasets. It is interesting to note here that, when supplied with the whole RFI instead of cropped optic disc, the classifier may focus on input regions other than optic disc, which is considered most important by ophthalmologists. These GradCAM results on glaucoma prediction by DL model should be analysed with great caution. While it cannot be ruled out that the classifier might have made a mistake in understanding the

discriminatory features of the disease, as Ribeiro’s classifier did with the classification of wolf [257], and these correct classifications are merely a fluke – after all the accuracy of the classifier dropped noticeably when whole RFIs were provided instead of ODs as seen in Table 3.10 – these maps may also hint at some hidden patterns in the image, tucked away from normally focused OD, which might have escaped human eye until now. To know for sure, ophthalmologists should have a comprehensive analysis on these class relevance maps. This prospect of AI finding new criteria for disease classification has a lot of potential. To make the scope of this thesis more manageable, further exploration into this topic is left for future work.

5.7 Discussion

One principal impediment in the successful deployment of AI-based CAD systems in the everyday clinical workflow is their lack of transparent decision-making. Although commonly used explanation methods provide some insight into these largely opaque algorithms, yet such explanations are usually convoluted and not readily comprehensible

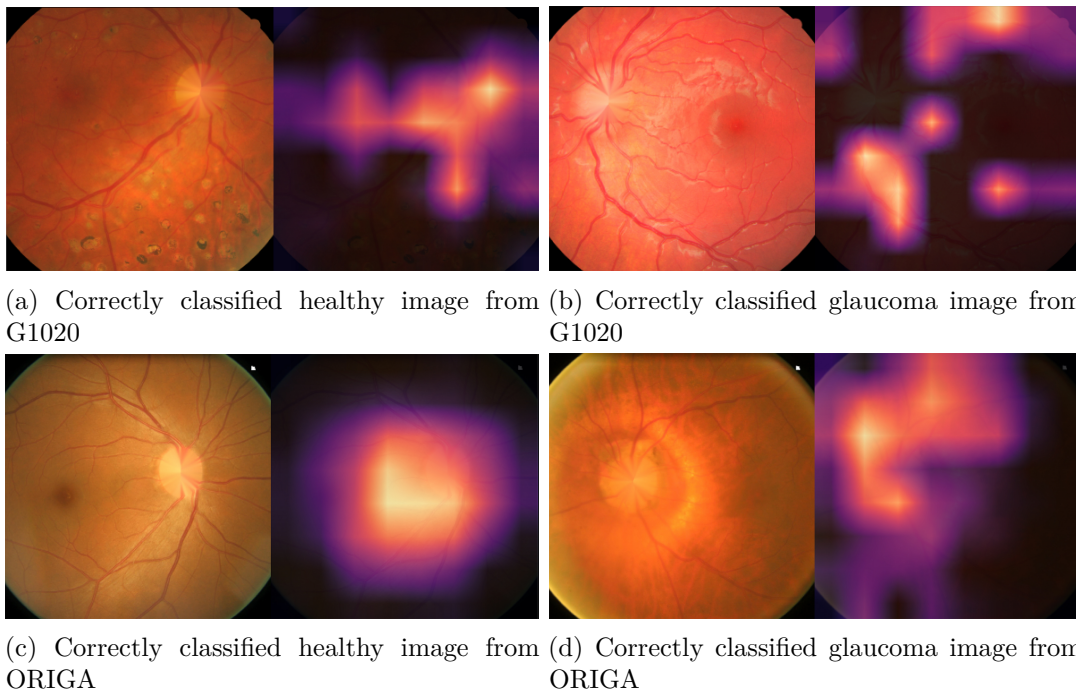


Figure 5.20: Examples of correct classification by Inception V3 albeit by looking at different regions in the RFI than analysed by ophthalmologists

except by highly trained AI experts. That is why beyond academic research and proof of concept studies, there has been a healthy scepticism about to what extent, if at all, AI should make or support medical decisions in real clinical workflows [373]. Although the sequential computation of DL-based models is traceable, they often lack explicit declarative representation of knowledge. Justifying the decisions taken by AI using explanations can help bring such academic research one step closer to practical deployment in the healthcare sector. Concept-based methods for network explanation offer great potential especially for complex classification tasks in sensitive application areas like MIA. In this chapter leveraging of these methods is explored to verify the ability of DNNs to learn and utilise human-understandable concepts for skin lesion classification. It is shown that a strong correlation exists between DNN's learnt representation of various concepts and those routinely used by dermatologists. These findings corroborate that deep learning based CAD systems are able to learn and capitalise on similar disease-related concepts for prediction as used by dermatologists. It has also been shown that Testing with CAVs (TCAV) is applicable using complete identically distributed images instead of general concept patches. However, this approach can further be improved by using more granular labelling of disease indicative concepts to get a deeper insight into the model's classification processes as well as further validation of its decisions.

Due to the complexity of the problem, possibly subjective annotations of a training set by various experts, and a small number of concept training samples, not all human-defined disease-related concepts were thoroughly analysed. Since supervised concept learning is highly dependent on high-quality and precisely annotated human concept examples, more focus should be placed on generating clean datasets of high-quality concept annotations that can be used for explaining models in medical imaging applications. Standardising the annotation according to one *school of thought* in the dermatology community, for example, following [360], can decrease inter-observer disagreement but it would require an enormous amount of time and effort by dermatology experts. To allow for a more comprehensive interpretation of the TCAV scores for this specific task, it would be desirable to curate a high-quality dataset with reliable fine-grained labels of concepts that are known to be highly indicative of specific diagnoses. As supervised concept classification from network activations has already been proven to be effective, an extension of unsupervised concept discovery should be considered. This alternative, or perhaps supplementary approach, could not only allow improvement towards simplifying the interpretability of networks by eliminating the necessity for laborious expert annotations but it could also provide insights into a network's own concepts, potentially revealing new knowledge for domain experts or unexpected biases in the network.

A novel direction of concept localisation for the explanation of AI-based DSS is also introduced and a robust perturbation-based concept localization method (p-CLM) is proposed that has been evaluated on a synthetically generated dataset as well as on a publicly available dataset of natural face images. The p-CLM considerably outperformed two gradient-based variants (g-CLM) in qualitative and quantitative evaluation. The initial results are promising and encourage further refinement of this approach. The computational efficiency and quality of heatmaps can be greatly improved by utilising optimisation-based perturbation methods like [283] and [374]. Not only will they reduce the number of network propagations by optimising the prediction score, but also the flexible shape of masks would be beneficial for the quality of CLMs. Perturbation-based methods always introduce some distribution shift that might distort predicted outcomes. However, more sophisticated methods like image in-painting could minimise distribution shifts through perturbation.

The ExAID framework consolidated and built upon previous works on detection of human-defined concepts for skin lesions diagnosis in the DL model’s latent space and their localisation on the input image to provide an intelligible textual explanation of the model’s predictions. In spite of severe limitations in terms of data and annotation availability, the system provides useful insights into the DL classifier’s decision-making, even in case of wrong predictions. However, when the current limitations of ExAID framework are thoroughly addressed, it will not only play a useful assistive role in reliable, efficient, and objective screening of melanoma, which is one of the most serious skin cancers but also help train new dermatologists efficiently and effectively. It is anticipated that physicians would be able to confer higher confidence to such CAD systems that are able to justify their prediction by listing the concepts that influenced positively or negatively towards a certain output.

Collaboration of AI developers and medical professionals already led to interesting advances in medical AI, including practical AI evaluation and discovery of new potential diagnostic criteria. However, despite success of concept-based and heatmap-based explanation methods, sensible and comprehensible explanations for medical image analysis tasks are still one of the greatest challenges related to medical image diagnosis, which should be addressed by concerted efforts from AI researchers, medical practitioners and regulatory authorities. Qualitatively evaluating an explanation with regards to its interpretability and completeness can be substantially subjective. Recently, there have been many efforts to quantify and qualify XAI methods and their explanations in objective and subjective ways. However, there are no agreed-upon and standardised evaluation procedures for explanation methods that can guarantee fidelity and rate quality. De-

velopment of standardised and objective evaluation criteria can greatly help benchmark upcoming explainable CAD systems and is thus an extremely important requirement for application in routine clinical environments. Moreover, appropriate regulatory measures should provide an ethical framework for the application of AI in healthcare, which can ensure safety and transparency through standardised evaluation and certification procedures.

Conclusion

Present status of CAD is the culmination of years of efforts expended by medical professionals and computer scientists. Yet, owing to the substantially sensitive nature of this application, these solutions fall short of very high standards of performance and reliability expectations from patients, healthcare providers, and regulators. The most straightforward approach to improve existing CAD systems is to enable them to emulate the time-tested diagnostic decision-making processes designed by medical practitioners. This, however, requires a comprehensive understanding of such diagnostic criteria on part of AI developers and their ability to mathematically model this decision-making process so that it can be integrated into DNNs. Therefore, it is advocated that well-coordinated collaboration must be established among researchers and professionals from all stakeholders.

6.1 Discussion

This thesis identified many open challenges requiring smart engineering and innovation. In any CAD development process, the first step is to gather relevant data and assess its quality. There are many issues worth addressing with regard to data curation. Since DL-based models are data-driven, they suffer from limitations and biases inherent in the data [97]. Although AI algorithms are not capable of showing any bias, yet they can inadvertently behave with partiality due to any bias that crept in during data collection. This contamination of data with bias can seriously hamper the ability of CAD systems

to perform effectively and objectively. For example, some data might suggest that a cohort who took a certain drug recovered quickly compared to those who did not. Deep models can detect this correlation easily. However, if the causality between the drug and the recovery is missing from the data, the models can overlook this causality and found their decision purely on association. The CAD systems trained on such datasets cannot propose an acceptable explanation of their decision either [375]. Therefore, CAD solutions will hugely benefit from a carefully curated dataset that incorporates the context and does not leave out any confounding variables. Such dataset curation can be achieved by concerted and close collaboration between medical practitioners and AI developers right from the onset.

Class imbalance in existing medical image datasets is another huge problem. Since some diseases have a very low prevalence, for example, melanoma, therefore it is really difficult to collect a large number of samples with such rarely occurring diseases. Subjectivity in annotating datasets is also a bottleneck. The performance of a CAD system is compared to the ground truth generated by human graders. When there is a significant disagreement of human graders on the diagnosis of a sample in the dataset, it can only deteriorate the performance of AI-based models. Therefore, while AI developers should be able to devise methods to identify erroneous data and handle missing values or outliers, profound efforts must be exerted by medical professionals in curating and publishing high-quality datasets.

Another challenge is capitalising on expert knowledge. The utilisation of expert knowledge in AI-based solutions does not only improve prediction accuracy of these systems as shown in section 2.4, section 3.3, and section 3.4, but also help explain these predictions as given in section 5.3.1. However, sometimes this expert knowledge and unstructured or semi-structured clinical data upon which this expert knowledge could be based happens to be phenomenally complex, massive and challenging to process effectively by AI algorithms [376] as seen in section 3.5. This could be partly because of a lack of guidelines and standardised Electronic Health Records (EHR) protocols. These EHRs play a vital role in routine manual diagnosis and can also be helpful in improving CAD performance. However, EHRs are not sometimes treated with the same level of diligence as bestowed upon other research and diagnostic data [377]. Although the quantity of EHRs is increasing since more and more healthcare establishments are using them in one form or the other, their availability with public medical image datasets remains scarce. Even when these additional data are available, there may be frequent missing or incorrect records as with ISIC-2019 and ISIC-2020 datasets. Therefore, improving the quality assessment of EHRs by emphasising accuracy, completeness, and

credibility can have a direct positive impact on the ability of CAD systems to make the most out of these data. Moreover, the advancement of interventional methods for the correction of algorithms and incorporation of explicit expert knowledge could account for retrospective adjustments during trial phases of CAD systems.

An interesting example in which clinical metadata and context of the case mattered and lack of their inclusion in AI models resulted in a technically valid yet misleading ML prognostic model, was the use of mortality risk prediction to make decisions about whether to provide treatment on an inpatient or outpatient basis for more than 14000 patients with pneumonia [378]. In this study, the algorithm counter-intuitively suggested that patients with pneumonia and asthma were at a lower risk of death compared to patients with only pneumonia, an indication that surprised the researchers who eventually ruled it out. A closer analysis of the data revealed that, at the hospital hosting this study, patients with a history of asthma who presented with pneumonia were usually admitted directly to intensive care units to prevent complications. This led to a pattern in the data that reflected better outcomes for such patients compared to patients with pneumonia and without a history of asthma with approximately 50% less mortality rate. This example not only emphasises the importance of representative training data for such algorithms but also that a contextually complete description of the data is of crucial importance.

One of the major use cases of CAD systems is in large-scale screening for early detection of possibly asymptomatic diseases. In such scenarios, these systems are expected to be highly accurate and sensitive to the early stages of the disease for which the screening is conducted. However, this usage has the risk of over-diagnosis [379]. Over-diagnosis means that a patient indeed had a disease and it was correctly identified by CAD but this diagnosis has little to no benefit for the patient. Instead, it can even be harmful to the patients [380]. In cancer screening, for example, it has been observed that repeated testing can result in increased detection of findings that are consistent with cancer, yet the mortality rates corresponding to such findings do not decrease accordingly. This peculiar and unexpected discovery raises questions about the actual benefits of such early screening of diseases. Over-diagnosis happens with manual diagnosis as well and some researchers argue for its benefits [381, 382]. However, with CAD it can quickly become overwhelming when coupled with the prediction of false positives. Therefore, in addition to becoming highly sensitive in their diagnosis, the CAD system may be able to reconcile this problem.

There is, unfortunately, no standardised method of evaluating CAD systems [22, 383]. This creates a hurdle for regulators to approve such solutions for practical use [384]. As

evident by section 2.2 and section 3.2, for example, researchers have used different performance metrics for the same datasets and the same classification tasks. While one performance metric may deal with certain aspects of the CAD system's performance, it may miss other crucial information. For example, classification accuracy is the most easily understandable metric. However, it does not always portray a true and complete picture of a classifier's performance, especially with highly imbalanced datasets. Other methods like sensitivity, specificity, and AUC are very useful in evaluating CAD systems. For multiclass classification, confusion matrices, in addition to the above-mentioned metrics, are also important to perform detailed error analysis. Therefore, to thoroughly and fairly evaluate the diagnostic performance of a system, a comprehensive and standardised assessment mechanism needs to be developed and followed. Additionally, an objective performance evaluation of a CAD system should also carefully consider unit misclassification costs for false positive and false negative errors [383]. The American Association of Physicists in Medicine (AAPM) constituted the Computer-Aided Detection in Diagnostic Imaging Subcommittee (CADSC) to develop standardised evaluation methods for medical image-based CAD systems [22] and raise awareness in medical professionals about various aspects of these systems like appreciating their effectiveness and cautious usage given the open challenges they still face. There is a pressing need to duplicate such regulating bodies to streamline the scientific evaluation of these systems and expedite their introduction in healthcare workflows.

Although there have been numerous studies on CAD systems that report encouraging results in lab settings, there are very limited instances where such algorithms are actually validated in clinical practice [385]. Effective and constant feedback from clinicians, who test such systems in their day-to-day routine, to AI researchers, who design these systems, can greatly hasten the development of practically usable CAD systems [386]. This back and forth feedback can also mitigate another major barrier in the successful deployment of CAD systems in a clinical environment, which is the lack of training of medical practitioners in terms of correct use and interpretation of the outcome of a CAD system [387]. Therefore, in addition to refining CAD systems, adequate training and education of their users are also equally vital since it has been observed that the communication gap between AI developers and their users can lead to misuse of technology [335, 336]) and eventual performance degradation [331].

There have been some studies with controversial findings of existing CAD systems, for example, with regards to generalisability of CAD [122] and their usefulness [336, 388]. Prima facie such studies might discourage other researchers to continue research or medical insurance companies to not covering the cost of using these systems [383].

Therefore, it is ever more important now to address all dubious aspects of CAD and win the confidence of physicians, patients, and regulators. In doing so, it is also necessary to signify that, as with every other medical procedure, drug, and device, the use of CAD systems has specific risks and will probably continue to have some imperfections. However, as long as the benefits of CAD usage are overwhelmingly higher than their potential, albeit rare, risks and the prerogative of taking final decision remains with human experts, CAD should be allowed a fair opportunity to show their mettle in the field. This can only be achieved with close and concerted efforts by medical and AI researchers.

In the medical domain, it is imperative to explain the output of algorithms in a human-understandable language as to support and not distract experts. Holzinger et al. [389] believe that the only way forward towards explainable CAD is to combine knowledge-driven and data-driven approaches, which could harness interpretability of the former method and high accuracy of the latter. This thesis advocates that the transition towards multimodal, diverse, and complete explanations that combine human-understandable modalities such as text, human-understandable concepts, and context will substantially support the way of XAI in clinical assistive settings. In medical diagnosis, explanations can be different for different users. For instance, a doctor might use different language, modality, or depth of explanation depending upon whether he/she is explaining to a patient, a regulator, or a fellow doctor. Similarly, explainable AI for healthcare serves a different purpose for medical practitioners and AI developers. It is inevitable that AI engineers design solutions that provide diverse explanations fitting the need of specific use-cases.

Finally, CAD should be used by medical researchers as it was intended by its developers since misuse of such systems can do more harm than good as discussed in section 5.2.3. These systems are developed as an aid, not as a primary decision-maker. In spite of all the technological panoply of AI, most CAD systems only achieve high sensitivity comparable to human graders at the expense of low specificity. However, the types of mistakes that CAD makes are different from those made by human experts, and, therefore, the complementary use of CAD by medical practitioners has the potential to improve overall performance. For example, a CAD system for screening mammography images to detect breast cancer was approved by the FDA only as a second reader [34]. Therefore, radiologists are expected to analyse a given case as observantly as they would in the absence of this helping hand, and only use CAD as a spell-checker to verify their diagnosis. Even with the second opinion given by the CAD, an expert diagnostician must not readily dismiss their own findings and should be able to discern between a

true negative and a false positive or vice versa. Appropriate and efficient use of CAD by clinicians can surely improve their performance and allow effective dispensation of medical services to the patients.

6.2 Future Outlook

From a healthcare perspective, CAD can have a promising future in general medicine encompassing a range of applications in the healthcare process like risk assessment, prognosis prediction, and monitoring of disease recurrence, in addition to using CAD systems for detection of certain diseases. In any role, however, the interface of these systems should be smooth and intuitive so that it can be seamlessly integrated into modern clinical practices without compromising on workflow efficiency.

In addition to the challenges identified and addressed in this thesis, this thesis also found some other interesting research directions which can be useful in realising a diverse CAD system and may help improve their accuracy and explainability. In radiology and histopathology, for instance, doctors do not just label an image with a certain disease or condition. They interpret the image by noting their findings and giving an impression based on those findings. These findings and impressions are usually free-hand text and do not follow any structure. Enabling a CAD system to analyse a visual modality and produce a coherent textual report can have an inherent explainability advantage since the impression (diagnosis) is substantiated by findings (justification of diagnosis). Although there is a reasonable body of research on this topic, a break-through in producing clinically meaningful detailed textual reports is yet awaited.

Many diseases, like diabetic retinopathy, are graded with respect to their severity. These grades have an ordinal relationship between two successive stages. In deep learning algorithms, various disease or their stages are usually one-hot encoded before they can be fed to the model. This encoding is very easy for DL models to process, however, it loses vital information regarding how close or distant two stages are from each other. For example, if Cheetah, Leopard, and Dog are three classes for an image classifier, it is evident from the classes that taxonomically Cheetah and Leopard are closer to each other and either feline is very distant from a dog. However, when these three classes are one-hot encoded, the encoded representation of each class is equally dissimilar from the other two, therefore, losing vital information which might be helpful for DL methods to exploit. There is a need to investigate alternative encoding schemes for such scenarios and take advantage of inherent relationships in class labels.

There is a growing interest in commercialising CAD solutions and developing use-

case-specific frameworks. However, caution must be exercised to carefully gauge those achievements and continue investing efforts in standardised evaluation and investigation of CAD methods in close cooperation with domain experts. Modern AI technologies have the potential to revolutionise healthcare in innumerable ways and plays a crucial role in creating a solid foundation of understanding and improving CAD functionality. Current advancements show that close collaboration of medical domain experts and computer scientists paired with persistent efforts of AI experts to advance and develop new methods will eventually lead to many practical applications which are just an anticipation of what will be possible in the future.

Bibliography

- [1] P. H. Meyers, C. M. Nice Jr, H. C. Becker, W. J. Nettleton Jr, J. W. Sweeney, and G. R. Meckstroth, “Automated Computer Analysis of Radiographic Images,” *Radiology*, vol. 83, no. 6, pp. 1029–1034, 1964.
- [2] J.-I. Toriwaki, Y. Suenaga, T. Negoro, and T. Fukumura, “Pattern recognition of chest X-ray images,” *Computer Graphics and Image Processing*, vol. 2, no. 3-4, pp. 252–271, 1973.
- [3] R. P. Kruger, J. R. Townes, D. L. Hall, S. J. Dwyer, and G. S. Lodwick, “Automated Radiographic Diagnosis via Feature Extraction and Classification of Cardiac Size and Shape Descriptors,” *IEEE Transactions on Biomedical Engineering*, no. 3, pp. 174–186, 1972.
- [4] R. P. Kruger, W. B. Thompson, and A. F. Turner, “Computer Diagnosis of Pneumoconiosis,” *IEEE Transactions on Systems, Man, and Cybernetics*, no. 1, pp. 40–49, 1974.
- [5] R. L. Engle Jr, “Attempts to Use Computers as Diagnostic Aids in Medical Decision Making: A Thirty-Year Experience,” *Perspectives in biology and medicine*, vol. 35, no. 2, pp. 207–219, 1992.
- [6] K. Doi, “Computer-aided diagnosis in medical imaging: Historical review, current status and future potential,” *Computerized medical imaging and graphics*, vol. 31, no. 4-5, pp. 198–211, 2007.
- [7] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [8] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, “ImageNet Large Scale Visual Recognition Challenge,” *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

- [9] J. Hirschberg and C. D. Manning, “Advances in natural language processing,” *Science*, vol. 349, no. 6245, pp. 261–266, 2015.
- [10] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, “A survey on deep learning in medical image analysis,” *Medical image analysis*, vol. 42, pp. 60–88, 2017.
- [11] L.-T. Wen, A. Tanaka, and M. Nonoyama, “Identification of Marek’s disease virus nuclear antigen in latently infected lymphoblastoid cells,” *Journal of virology*, vol. 62, no. 10, pp. 3764–3771, 1988.
- [12] P. Teare, M. Fishman, O. Benzaquen, E. Toledano, and E. Elnekave, “Malignancy Detection on Mammography Using Dual Deep Convolutional Neural Networks and Genetically Discovered False Color Input Enhancement,” *Journal of digital imaging*, vol. 30, no. 4, pp. 499–505, 2017.
- [13] D. M. Barros, J. C. Moura, C. R. Freire, A. C. Taleb, R. A. Valentim, and P. S. Morais, “Machine learning applied to retinal image processing for glaucoma detection: review and perspective,” *Biomedical engineering online*, vol. 19, no. 1, pp. 1–21, 2020.
- [14] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, *et al.*, “Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs,” *JAMA*, vol. 316, no. 22, pp. 2402–2410, 2016.
- [15] L. M. Prevedello, B. S. Erdal, J. L. Ryu, K. J. Little, M. Demirer, S. Qian, and R. D. White, “Automated Critical Test Findings Identification and Online Notification System Using Artificial Intelligence in Imaging,” *Radiology*, vol. 285, no. 3, pp. 923–931, 2017.
- [16] United Nations General Assembly, “Transforming our World: The 2030 Agenda for Sustainable Development.” available at <https://sustainabledevelopment.un.org/post2015/>, October 2015.
- [17] World Health Organization *et al.*, “Health workforce policy and management in the context of the COVID-19 pandemic response: Interim guidance,” tech. rep., World Health Organization, 2020.

-
- [18] World Health Organization *et al.*, “State of the World’s Nursing 2020: Investing in Education, Jobs and Leadership,” tech. rep., World Health Organization, 2020.
- [19] H. Zhao and M. Shingles, “AI for Good: Global Summit 2018,” tech. rep., The International Telecommunication Union (ITU), June 2017.
- [20] E. D O’Sullivan and S. Schofield, “Cognitive Bias in Clinical Medicine,” *Journal of the Royal College of Physicians of Edinburgh*, vol. 48, no. 3, pp. 225–231, 2018.
- [21] T. S. Doherty and A. E. Carroll, “Believing in Overcoming Cognitive Biases,” *AMA Journal of Ethics*, vol. 22, no. 9, pp. 773–778, 2020.
- [22] N. Petrick, B. Sahiner, S. G. Armato III, A. Bert, L. Correale, S. Delsanto, M. T. Freedman, D. Fryd, D. Gur, L. Hadjiiski, *et al.*, “Evaluation of computer-aided detection and diagnosis systems,” *Medical physics*, vol. 40, no. 8, p. 087001, 2013.
- [23] M. Brown, P. Browning, M. W. Wahi-Anwar, M. Murphy, J. Delgado, H. Greenspan, F. Abtin, S. Ghahremani, N. Yaghmai, I. da Costa, *et al.*, “Integration of Chest CT CAD into the Clinical Workflow and Impact on Radiologist Efficiency,” *Academic radiology*, vol. 26, no. 5, pp. 626–631, 2019.
- [24] T. Kozuka, Y. Matsukubo, T. Kadoba, T. Oda, A. Suzuki, T. Hyodo, S. Im, H. Kaida, Y. Yagyū, M. Tsurusaki, *et al.*, “Efficiency of a computer-aided diagnosis (CAD) system with deep learning in detection of pulmonary nodules on 1-mm-thick images of computed tomography,” *Japanese Journal of Radiology*, vol. 38, no. 11, pp. 1052–1061, 2020.
- [25] C. Y. Cheung, F. Tang, D. S. W. Ting, G. S. W. Tan, and T. Y. Wong, “Artificial Intelligence in Diabetic Eye Disease Screening,” *The Asia-Pacific Journal of Ophthalmology*, vol. 8, no. 2, pp. 158–164, 2019.
- [26] R. M. Nishikawa, “Computer-aided Detection and Diagnosis,” in *Digital Mammography*, pp. 85–106, Springer, 2010.
- [27] C. Carrera, M. A. Marchetti, S. W. Dusza, G. Argenziano, R. P. Braun, A. C. Halpern, N. Jaimes, H. J. Kittler, J. Malvehy, S. W. Menzies, *et al.*, “Validity and Reliability of Dermoscopic Criteria Used to Differentiate Nevi From Melanoma. A Web-Based International Dermoscopy Society Study,” *JAMA dermatology*, vol. 152, no. 7, pp. 798–806, 2016.

- [28] A. Masood and A. Ali Al-Jumaily, “Computer Aided Diagnostic Support System for Skin Cancer: A Review of Techniques and Algorithms,” *International journal of biomedical imaging*, vol. 2013, 2013.
- [29] S. Benjamens, P. Dhunnoo, and B. Meskó, “The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database,” *NPJ digital medicine*, vol. 3, no. 1, pp. 1–8, 2020.
- [30] H. A. Haenssle, C. Fink, R. Schneiderbauer, F. Toberer, T. Buhl, A. Blum, A. Kalloo, A. B. H. Hassen, L. Thomas, A. Enk, *et al.*, “Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists,” *Annals of Oncology*, vol. 29, no. 8, pp. 1836–1842, 2018.
- [31] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.
- [32] A. Ebigbo, R. Mendel, A. Probst, J. Manzeneder, L. A. de Souza Jr, J. P. Papa, C. Palm, and H. Messmann, “Computer-aided diagnosis using deep learning in the evaluation of early oesophageal adenocarcinoma,” *Gut*, vol. 68, no. 7, pp. 1143–1145, 2019.
- [33] M. K. Santos, J. R. Ferreira Júnior, D. T. Wada, A. P. M. Tenório, M. H. N. Barbosa, and P. M. d. A. Marques, “Artificial intelligence, machine learning, computer-aided diagnosis, and radiomics: advances in imaging towards to precision medicine,” *Radiologia brasileira*, vol. 52, no. 6, pp. 387–396, 2019.
- [34] H.-P. Chan, L. M. Hadjiiski, and R. K. Samala, “Computer-aided diagnosis in the era of deep learning,” *Medical physics*, vol. 47, no. 5, pp. e218–e227, 2020.
- [35] M.-H. Laves, S. Ihler, T. Ortmaier, and L. A. Kahrs, “Quantifying the uncertainty of deep learning-based computer-aided diagnosis for patient safety,” *Current Directions in Biomedical Engineering*, vol. 5, no. 1, pp. 223–226, 2019.
- [36] L. Edwards and M. Veale, “Enslaving the Algorithm: From a ”Right to an Explanation” to a ”Right to Better Decisions”?,” *IEEE Security & Privacy*, vol. 16, no. 3, pp. 46–54, 2018.

- [37] Y. Hagiwara, J. E. W. Koh, J. H. Tan, S. V. Bhandary, A. Laude, E. J. Ciaccio, L. Tong, and U. R. Acharya, “Computer-aided diagnosis of glaucoma using fundus images: A review,” *Computer methods and programs in biomedicine*, vol. 165, pp. 1–12, 2018.
- [38] G. Falco, B. Shneiderman, J. Badger, R. Carrier, A. Dahbura, D. Danks, M. Eling, A. Goodloe, J. Gupta, C. Hart, *et al.*, “Governing ai safety through independent audits,” *Nature Machine Intelligence*, vol. 3, no. 7, pp. 566–571, 2021.
- [39] L. Barinov, A. Jairaj, M. Becker, S. Seymour, E. Lee, A. Schram, E. Lane, A. Goldszal, D. Quigley, and L. Paster, “Impact of data presentation on physician performance utilizing artificial intelligence-based computer-aided diagnosis and decision support systems,” *Journal of Digital Imaging*, vol. 32, no. 3, pp. 408–416, 2019.
- [40] T. W. Rogers, N. Jaccard, F. Carbonaro, H. G. Lemij, K. A. Vermeer, N. J. Reus, and S. Trikha, “Evaluation of an AI system for the automated detection of glaucoma from stereoscopic optic disc photographs: the European Optic Disc Assessment Study,” *Eye*, vol. 33, no. 11, pp. 1791–1797, 2019.
- [41] E. Pead, R. Megaw, J. Cameron, A. Fleming, B. Dhillon, E. Trucco, and T. MacGillivray, “Automated detection of age-related macular degeneration in color fundus photography: a systematic review,” *survey of ophthalmology*, vol. 64, no. 4, pp. 498–511, 2019.
- [42] J. Son, J. Y. Shin, H. D. Kim, K.-H. Jung, K. H. Park, and S. J. Park, “Development and Validation of Deep Learning Models for Screening Multiple Abnormal Findings in Retinal Fundus Images,” *Ophthalmology*, vol. 127, no. 1, pp. 85–94, 2020.
- [43] F. G. Heslinga, J. P. W. Pluim, A. Houben, M. T. Schram, R. M. A. Henry, C. D. A. Stehouwer, M. J. van Greevenbroek, T. T. Berendschot, and M. Veta, “Direct classification of type 2 diabetes from retinal fundus images in a population-based sample from the Maastricht study,” in *Medical Imaging 2020: Computer-Aided Diagnosis* (H. K. Hahn and M. A. Mazurowski, eds.), vol. 11314, pp. 383 – 388, International Society for Optics and Photonics, SPIE, 2020.
- [44] A. Mitani, A. Huang, S. Venugopalan, G. S. Corrado, L. Peng, D. R. Webster, N. Hammel, Y. Liu, and A. V. Varadarajan, “Detection of anaemia from retinal

- fundus images via deep learning,” *Nature Biomedical Engineering*, vol. 4, no. 1, pp. 18–27, 2020.
- [45] R. Poplin, A. V. Varadarajan, K. Blumer, Y. Liu, M. V. McConnell, G. S. Corrado, L. Peng, and D. R. Webster, “Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning,” *Nature Biomedical Engineering*, vol. 2, no. 3, p. 158, 2018.
- [46] X. Chen, Y. Xu, D. W. K. Wong, T. Y. Wong, and J. Liu, “Glaucoma detection based on deep convolutional neural network,” in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, pp. 715–718, IEEE, 2015.
- [47] G. An, K. Omodaka, K. Hashimoto, S. Tsuda, Y. Shiga, N. Takada, T. Kikawa, H. Yokota, M. Akiba, and T. Nakazawa, “Glaucoma Diagnosis with Machine Learning Based on Optical Coherence Tomography and Color Fundus Images,” *Journal of healthcare engineering*, vol. 2019, 2019.
- [48] Ş. S. Kucur, G. Hollo, and R. Sznitman, “A deep learning approach to automatic detection of early glaucoma from visual fields,” *PloS one*, vol. 13, no. 11, 2018.
- [49] M. Abramoff and C. N. Kay, “Chapter 6 - Image Processing,” in *Retina (Fifth Edition)* (S. J. Ryan, S. R. Sadda, D. R. Hinton, A. P. Schachar, S. R. Sadda, C. Wilkinson, P. Wiedemann, and A. P. Schachar, eds.), pp. 151 – 176, London: W.B. Saunders, fifth edition ed., 2013.
- [50] L. Ballerini, R. B. Fisher, B. Aldridge, and J. Rees, “A Color and Texture Based Hierarchical K-NN Approach to the Classification of Non-melanoma Skin Lesions,” in *Color Medical Image Analysis*, pp. 63–86, Springer, 2013.
- [51] Z. Zhang, F. S. Yin, J. Liu, W. K. Wong, N. M. Tan, B. H. Lee, J. Cheng, and T. Y. Wong, “ORIGA-light : An Online Retinal Fundus Image Database for Glaucoma Analysis and Research ,” in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pp. 3065–3068, IEEE, 2010.
- [52] A. Diaz-Pinto, S. Morales, V. Naranjo, T. Köhler, J. M. Mossi, and A. Navea, “CNNs for automatic glaucoma assessment using fundus images: an extensive validation,” *Biomedical engineering online*, vol. 18, no. 1, p. 29, 2019.

-
- [53] Z. Li, Y. He, S. Keel, W. Meng, R. T. Chang, and M. He, “Efficacy of a Deep Learning System for Detecting Glaucomatous Optic Neuropathy Based on Color Fundus Photographs,” *Ophthalmology*, vol. 125, no. 8, pp. 1199–1206, 2018.
- [54] D. C. Hood and C. G. De Moraes, “Efficacy of a Deep Learning System for Detecting Glaucomatous Optic Neuropathy Based on Color Fundus Photographs,” *Ophthalmology*, vol. 125, no. 8, pp. 1207–1208, 2018.
- [55] G. Argenziano, H. P. Soyer, D. Giorgi, D. Piccolo, P. Carli, and M. Delfino, *Interactive Atlas of Dermoscopy*. EDRA Medical Publishing & New Media, 2000.
- [56] I. Giotis, N. Molders, S. Land, M. Biehl, M. F. Jonkman, and N. Petkov, “MED-NODE: a computer-assisted melanoma diagnosis system using non-dermoscopic images,” *Expert systems with applications*, vol. 42, no. 19, pp. 6578–6585, 2015.
- [57] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. Marcal, and J. Rozeira, “PH 2-A dermoscopic image database for research and benchmarking,” in *2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pp. 5437–5440, IEEE, 2013.
- [58] A.-R. A. Ali and T. M. Deserno, “A Systematic Review of Automated Melanoma Detection in Dermatoscopic Images and its Ground Truth Data,” in *Medical Imaging 2012: Image Perception, Observer Performance, and Technology Assessment*, vol. 8318, p. 83181I, International Society for Optics and Photonics, 2012.
- [59] F. Fumero, S. Alayón, J. L. Sanchez, J. Sigut, and M. Gonzalez-Hernandez, “Rim-one: An open retinal image database for optic nerve evaluation,” in *2011 24th international symposium on computer-based medical systems (CBMS)*, pp. 1–6, IEEE, 2011.
- [60] A. Almazroa, S. Alodhayb, E. Osman, E. Ramadan, M. Hummadi, M. Dlaim, M. Alkatee, K. Raahemifar, and V. Lakshminarayanan, “Retinal fundus images for glaucoma analysis: the RIGA dataset,” in *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications*, vol. 10579, p. 105790B, International Society for Optics and Photonics, 2018.
- [61] E. Decencière, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay, B. Charton, and J.-C. Klein, “Feedback

- on a Publicly Distributed Image Database: The Messidor Database,” *Image Analysis & Stereology*, vol. 33, no. 3, pp. 231–234, 2014.
- [62] J. I. Orlando, H. Fu, J. B. Breda, K. van Keer, D. R. Bathula, A. Diaz-Pinto, R. Fang, P.-A. Heng, J. Kim, J. Lee, *et al.*, “REFUGE Challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs,” *Medical image analysis*, vol. 59, p. 101570, 2020.
- [63] Peking University China, “Ocular Disease Intelligent Recognition ODIR-5K.” <https://odir2019.grand-challenge.org/>, 2019.
Accessed: September 23, 2021.
- [64] J. Sivaswamy, S. Krishnadas, A. Chakravarty, G. Joshi, A. S. Tabish, *et al.*, “A Comprehensive Retinal Image Dataset for the Assessment of Glaucoma from the Optic Nerve Head Analysis,” *JSM Biomedical Imaging Data Papers*, vol. 2, no. 1, p. 1004, 2015.
- [65] L. Li, M. Xu, X. Wang, L. Jiang, and H. Liu, “Attention Based Glaucoma Detection: A Large-scale Database and CNN Model,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10571–10580, 2019.
- [66] A. Almazroa, W. Sun, S. Alodhayb, K. Raahemifar, and V. Lakshminarayanan, “Optic disc segmentation for glaucoma screening system using fundus images,” *Clinical ophthalmology (Auckland, NZ)*, vol. 11, 2017.
- [67] B. Al-Bander, B. M. Williams, W. Al-Nuaimy, M. A. Al-Tae, H. Pratt, and Y. Zheng, “Dense Fully Convolutional Segmentation of the Optic Disc and Cup in Colour Fundus for Glaucoma Diagnosis,” *Symmetry*, vol. 10, no. 4, p. 87, 2018.
- [68] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.
- [69] H. Fu, J. Cheng, Y. Xu, D. W. K. Wong, J. Liu, and X. Cao, “Joint Optic Disc and Cup Segmentation Based on Multi-Label Deep Network and Polar Transformation,” *IEEE transactions on medical imaging*, vol. 37, no. 7, pp. 1597–1605, 2018.

-
- [70] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” in *Advances in neural information processing systems*, pp. 91–99, 2015.
- [71] M. Baskaran, R. C. Foo, C.-Y. Cheng, A. K. Narayanaswamy, Y.-F. Zheng, R. Wu, S.-M. Saw, P. J. Foster, T.-Y. Wong, and T. Aung, “The Prevalence and Types of Glaucoma in an Urban Chinese Population: The Singapore Chinese Eye Study,” *JAMA ophthalmology*, vol. 133, no. 8, pp. 874–880, 2015.
- [72] K. Park, J. Kim, and J. Lee, “Automatic optic nerve head localization and cup-to-disc ratio detection using state-of-the-art deep-learning architectures,” *Scientific reports*, vol. 10, no. 1, pp. 1–10, 2020.
- [73] A. Farhadi and J. Redmon, “Yolov3: An incremental improvement,” in *Computer Vision and Pattern Recognition*, pp. 1804–02, Springer Berlin/Heidelberg, Germany, 2018.
- [74] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- [75] 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*, pp. 4700–4708, 2017.
- [76] Z. Zhou, J. Yao, and L. Wang, “A Robust Optic Disc Localization Algorithm in Retinal Images Based on Support Vector Machine,” in *Proceedings of the 2020 5th International Conference on Biomedical Signal and Image Processing*, pp. 25–31, 2020.
- [77] T. Kauppi, V. Kalesnykiene, J.-K. Kamarainen, L. Lensu, I. Sorri, H. Uusitalo, H. Kälviäinen, and J. Pietilä, “DIARETDB0: Evaluation Database and Methodology for Diabetic Retinopathy Algorithms,” *Machine Vision and Pattern Recognition Research Group, Lappeenranta University of Technology, Finland*, vol. 73, pp. 1–17, 2006.
- [78] T. Kauppi, V. Kalesnykiene, J.-K. Kamarainen, L. Lensu, I. Sorri, A. Raninen, R. Voutilainen, H. Kalviainen, and J. Pietila, “DIARETDB1 diabetic retinopathy database and evaluation protocol,” in *Medical Image Understanding and Analysis, 2007., Proceedings of 11th Conference on*, 2007.

- [79] J. Staal, M. Abramoff, M. Niemeijer, M. Viergever, and B. van Ginneken, "Ridge-Based Vessel Segmentation in Color Images of the Retina," *IEEE Transactions on Medical Imaging*, vol. 23, no. 4, pp. 501–509, 2004.
- [80] A. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating Blood Vessels in Retinal Images by Piecewise Threshold Probing of a Matched Filter Response," *IEEE Transactions on Medical imaging*, vol. 19, no. 3, pp. 203–210, 2000.
- [81] S. Sreng, N. Maneerat, K. Hamamoto, and K. Y. Win, "Deep Learning for Optic Disc Segmentation and Glaucoma Diagnosis on Retinal Images," *Applied Sciences*, vol. 10, no. 14, p. 4916, 2020.
- [82] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018.
- [83] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *arXiv preprint arXiv:1704.04861*, 2017.
- [84] P. Joshi, R. Raj KS, and J. Alike, "Optic disc localization using interference map and localized segmentation using grab cut," *Automatika*, vol. 62, no. 2, pp. 187–196, 2021.
- [85] A. Chakravarty, J. Sivaswamy, *et al.*, "An assistive annotation system for retinal images," in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pp. 1506–1509, IEEE, 2015.
- [86] K. S. Deepak, N. K. Medathati, and J. Sivaswamy, "Detection and discrimination of disease-related abnormalities based on learning normal cases," *Pattern Recognition*, vol. 45, no. 10, pp. 3707–3716, 2012.
- [87] C. Rother, V. Kolmogorov, and A. Blake, "'GrabCut' - Interactive Foreground Extraction using Iterated Graph Cuts," *ACM transactions on graphics (TOG)*, vol. 23, no. 3, pp. 309–314, 2004.
- [88] H. Veena, A. Muruganandham, and T. S. Kumaran, "A novel optic disc and optic cup segmentation technique to diagnose glaucoma using deep learning convolutional neural network over retinal fundus images," *Journal of King Saud University-Computer and Information Sciences*, 2021.

-
- [89] P. S. Mangipudi, H. M. Pandey, and A. Choudhary, “Improved optic disc and cup segmentation in Glaucomatic images using deep learning architecture,” *Multimedia Tools and Applications*, pp. 1–21, 2021.
- [90] A. Jibhakate, P. Parnerkar, S. Mondal, V. Bharambe, and S. Mantri, “Skin Lesion Classification using Deep Learning and Image Processing,” in *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, pp. 333–340, IEEE, 2020.
- [91] 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.
- [92] A. C. Salian, S. Vaze, P. Singh, G. N. Shaikh, S. Chapaneri, and D. Jayaswal, “Skin Lesion Classification using Deep Learning Architectures,” in *2020 3rd International conference on communication system, computing and IT applications (CSCITA)*, pp. 168–173, IEEE, 2020.
- [93] J. Kawahara, A. BenTaieb, and G. Hamarneh, “Deep features to classify skin lesions,” in *2016 IEEE 13th international symposium on biomedical imaging (ISBI)*, pp. 1397–1400, IEEE, 2016.
- [94] Z. Ge, S. Demyanov, B. Bozorgtabar, M. Abedini, R. Chakravorty, A. Bowling, and R. Garnavi, “Exploiting local and generic features for accurate skin lesions classification using clinical and dermoscopy imaging,” in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pp. 986–990, IEEE, 2017.
- [95] D. Gutman, N. C. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, and A. Halpern, “Skin Lesion Analysis toward Melanoma Detection: A Challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC),” *arXiv preprint arXiv:1605.01397*, 2016.
- [96] 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, p. 1390, 2021.
- [97] E. Ntoutsis, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdil, M.-E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, *et al.*, “Bias in data-driven artifi-

- cial intelligence systems — An introductory survey,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 3, p. e1356, 2020.
- [98] Y. Li, A. Esteva, B. Kuprel, R. Novoa, J. Ko, and S. Thrun, “Skin Cancer Detection and Tracking using Data Synthesis and Deep Learning,” in *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- [99] Y. Li and L. Shen, “Skin Lesion Analysis towards Melanoma Detection Using Deep Learning Network,” *Sensors*, vol. 18, no. 2, p. 556, 2018.
- [100] M. Cullèl-Dalmau, S. Noé, M. Otero-Viñas, I. Meić, and C. Manzo, “Convolutional Neural Network for Skin Lesion Classification: Understanding the Fundamentals Through Hands-On Learning,” *Frontiers in Medicine*, vol. 8, p. 213, 2021.
- [101] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, “LabelMe: A Database and Web-Based Tool for Image Annotation,” *International journal of computer vision*, vol. 77, no. 1-3, pp. 157–173, 2008.
- [102] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.
- [103] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.
- [104] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, “Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning,” in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- [105] M. N. Bajwa, M. I. Malik, S. A. Siddiqui, A. Dengel, F. Shafait, W. Neumeier, and S. Ahmed, “Two-stage framework for optic disc localization and glaucoma classification in retinal fundus images using deep learning,” *BMC medical informatics and decision making*, vol. 19, no. 1, p. 136, 2019.
- [106] J. Hu, L. Shen, and G. Sun, “Squeeze-and-Excitation Networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

-
- [107] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, “Learning Transferable Architectures for Scalable Image Recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8697–8710, 2018.
- [108] H. Liao, “A deep learning approach to universal skin disease classification,” *University of Rochester Department of Computer Science, CSC*, 2016.
- [109] F. M. Cicero, A. H. M. Oliveira, G. M. Botelho, and C. d. C. da Computação, “Deep Learning and Convolutional Neural Networks in the Aid of the Classification of Melanoma,” in *Conference on Graphics, Patterns and Images, SIBGRAPI*, 2016.
- [110] V. Prabhu, A. Kannan, M. Ravuri, M. Chablani, D. Sontag, and X. Amatriain, “Prototypical Clustering Networks for Dermatological Disease Diagnosis,” *arXiv preprint arXiv:1811.03066*, 2018.
- [111] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, *et al.*, “Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC),” in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 168–172, IEEE, 2018.
- [112] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, *et al.*, “Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC),” *arXiv preprint arXiv:1902.03368*, 2019.
- [113] A. Shabbir, A. Rasheed, H. Shehraz, A. Saleem, B. Zafar, M. Sajid, N. Ali, S. H. Dar, and T. Shehryar, “Detection of glaucoma using retinal fundus images: A comprehensive review,” *Mathematical Biosciences and Engineering*, vol. 18, no. 3, pp. 2033–2076, 2021.
- [114] A. Menegola, M. Fornaciali, R. Pires, F. V. Bittencourt, S. Avila, and E. Valle, “Knowledge transfer for melanoma screening with deep learning,” in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pp. 297–300, IEEE, 2017.
- [115] E. Nasr-Esfahani, S. Samavi, N. Karimi, S. M. R. Soroushmehr, M. H. Jafari, K. Ward, and K. Najarian, “Melanoma detection by analysis of clinical images

- using convolutional neural network,” in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1373–1376, IEEE, 2016.
- [116] L. Yu, H. Chen, Q. Dou, J. Qin, and P.-A. Heng, “Automated Melanoma Recognition in Dermoscopy Images via Very Deep Residual Networks,” *IEEE transactions on medical imaging*, vol. 36, no. 4, pp. 994–1004, 2016.
- [117] N. Codella, J. Cai, M. Abedini, R. Garnavi, A. Halpern, and J. R. Smith, “Deep Learning, Sparse Coding, and SVM for Melanoma Recognition in Dermoscopy Images,” in *International workshop on machine learning in medical imaging*, pp. 118–126, Springer, 2015.
- [118] Z. Ma and J. M. R. Tavares, “Effective features to classify skin lesions in dermoscopic images,” *Expert Systems with Applications*, vol. 84, pp. 92–101, 2017.
- [119] A. R. Lopez, X. Giro-i Nieto, J. Burdick, and O. Marques, “Skin lesion classification from dermoscopic images using deep learning techniques,” in *2017 13th IASTED international conference on biomedical engineering (BioMed)*, pp. 49–54, IEEE, 2017.
- [120] D. A. Shoieb, S. M. Youssef, and W. M. Aly, “Computer-Aided Model for Skin Diagnosis Using Deep Learning,” *Journal of Image and Graphics*, vol. 4, no. 2, pp. 122–129, 2016.
- [121] T. J. Brinker, A. Hekler, J. S. Utikal, N. Grabe, D. Schadendorf, J. Klode, C. Berking, T. Steeb, A. H. Enk, and C. von Kalle, “Skin cancer classification using convolutional neural networks: Systematic review,” *Journal of medical Internet research*, vol. 20, no. 10, p. e11936, 2018.
- [122] C. Navarrete-Dechent, S. W. Dusza, K. Liopyris, A. A. Marghoob, A. C. Halpern, and M. A. Marchetti, “Automated dermatological diagnosis: hype or reality?,” *The Journal of investigative dermatology*, vol. 138, no. 10, p. 2277, 2018.
- [123] S. S. Han, M. S. Kim, W. Lim, G. H. Park, I. Park, and S. E. Chang, “Classification of the Clinical Images for Benign and Malignant Cutaneous Tumors Using a Deep Learning Algorithm,” *Journal of Investigative Dermatology*, vol. 138, no. 7, pp. 1529–1538, 2018.

- [124] S. S. Han, W. Lim, M. S. Kim, I. Park, G. H. Park, and S. E. Chang, "Interpretation of the Outputs of a Deep Learning Model Trained with a Skin Cancer Dataset," *The Journal of investigative dermatology*, vol. 138, no. 10, p. 2275, 2018.
- [125] E. Dervisevic, S. Pavljasevic, A. Dervisevic, and S. S. Kasumovic, "Challenges in Early Glaucoma Detection," *Medical Archives*, vol. 70, no. 3, p. 203, 2016.
- [126] A. Jackson and S. Radhakrishnan, "Understanding and living with glaucoma." Glaucoma Research Foundation, 2016.
- [127] R. R. Bourne, S. R. Flaxman, T. Braithwaite, M. V. Cicinelli, A. Das, J. B. Jonas, J. Keeffe, J. H. Kempen, J. Leasher, H. Limburg, *et al.*, "Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis," *The Lancet Global Health*, vol. 5, no. 9, pp. e888–e897, 2017.
- [128] X. Chen, Y. Xu, S. Yan, D. W. K. Wong, T. Y. Wong, and J. Liu, "Automatic Feature Learning for Glaucoma Detection Based on Deep Learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 669–677, Springer, 2015.
- [129] Y. Xu, J. Liu, S. Lin, D. Xu, C. Y. Cheung, T. Aung, and T. Y. Wong, "Efficient Optic Cup Detection from Intra-image Learning with Retinal Structure Priors," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 58–65, Springer, 2012.
- [130] F. Fumero, S. Alayón, J. Sanchez, J. Sigut, and M. Gonzalez-Hernandez, "RIM-ONE: An Open Retinal Image Database for Optic Nerve Evaluation," in *Computer-Based Medical Systems (CBMS), 2011 24th International Symposium on*, pp. 1–6, IEEE, 2011.
- [131] Q. Abbas, "Glaucoma-Deep: Detection of Glaucoma Eye Disease on Retinal Fundus Images using Deep Learning," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 6, pp. 41–45, 2017.
- [132] M. R. K. Mookiah, U. R. Acharya, C. M. Lim, A. Petznick, and J. S. Suri, "Data mining technique for automated diagnosis of glaucoma using higher order spectra and wavelet energy features," *Knowledge-Based Systems*, vol. 33, pp. 73–82, 2012.

- [133] J. Amin, M. Sharif, M. Yasmin, H. Ali, and S. L. Fernandes, “A method for the detection and classification of diabetic retinopathy using structural predictors of bright lesions,” *Journal of Computational Science*, vol. 19, pp. 153–164, 2017.
- [134] M. U. Akram, S. Khalid, A. Tariq, S. A. Khan, and F. Azam, “Detection and classification of retinal lesions for grading of diabetic retinopathy,” *Computers in biology and medicine*, vol. 45, pp. 161–171, 2014.
- [135] C. Eswaran, A. W. Reza, and S. Hati, “Extraction of the Contours of Optic Disc and Exudates Based on Marker-Controlled Watershed Segmentation,” in *Computer Science and Information Technology, 2008. ICCSIT’08. International Conference on*, pp. 719–723, IEEE, 2008.
- [136] R. Chrástek, M. Wolf, K. Donath, G. Michelson, and H. Niemann, “Optic Disc Segmentation in Retinal Images,” in *Bildverarbeitung für die Medizin 2002*, pp. 263–266, Springer, 2002.
- [137] J. Canny, “A Computational Approach to Edge Detection,” in *Readings in Computer Vision*, pp. 184–203, Elsevier, 1987.
- [138] M. D. Abràmoff, M. K. Garvin, and M. Sonka, “Retinal Imaging and Image Analysis,” *IEEE reviews in biomedical engineering*, vol. 3, pp. 169–208, 2010.
- [139] J. Liu, D. Wong, J. Lim, X. Jia, F. Yin, H. Li, W. Xiong, and T. Wong, “Optic cup and disk extraction from retinal fundus images for determination of cup-to-disc ratio,” in *Industrial Electronics and Applications, 2008. ICIEA 2008. 3rd IEEE Conference on*, pp. 1828–1832, IEEE, 2008.
- [140] L. G. Nyúl, “Retinal image analysis for automated glaucoma risk evaluation,” in *MIPPR 2009: Medical Imaging, Parallel Processing of Images, and Optimization Techniques* (F. Zhang and F. Zhang, eds.), vol. 7497, pp. 332 – 340, International Society for Optics and Photonics, SPIE, 2009.
- [141] P. Siddalingaswamy and K. G. Prabhu, “Automatic Localization and Boundary Detection of Optic Disc Using Implicit Active Contours,” *International Journal of Computer Applications*, vol. 1, no. 6, pp. 1–5, 2010.
- [142] B. Harangi, R. J. Qureshi, A. Csutak, T. Peto, and A. Hajdu, “Automatic detection of the optic disc using majority voting in a collection of optic disc detectors,” in *Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on*, pp. 1329–1332, IEEE, 2010.

- [143] L. Kovacs, R. J. Qureshi, B. Nagy, B. Harangi, and A. Hajdu, "Graph based detection of optic disc and fovea in retinal images," in *Soft Computing Applications (SOFA), 2010 4th International Workshop on*, pp. 143–148, IEEE, 2010.
- [144] K. A. Goatman, A. D. Fleming, S. Philip, G. J. Williams, J. A. Olson, and P. F. Sharp, "Detection of New Vessels on the Optic Disc Using Retinal Photographs," *IEEE transactions on medical imaging*, vol. 30, no. 4, pp. 972–979, 2011.
- [145] S.-C. Cheng and Y.-M. Huang, "A novel approach to diagnose diabetes based on the fractal characteristics of retinal images," *IEEE Transactions on Information Technology in Biomedicine*, vol. 7, no. 3, pp. 163–170, 2003.
- [146] B. Dashtbozorg, A. M. Mendonça, and A. Campilho, "Optic disc segmentation using the sliding band filter," *Computers in biology and medicine*, vol. 56, pp. 1–12, 2015.
- [147] H. Kobatake, "A convergence index filter for vector fields and its application to medical image processing," *Electronics and communications in Japan (Part III: fundamental electronic science)*, vol. 89, no. 6, pp. 34–46, 2006.
- [148] D. Zhang and Y. Zhao, "Novel Accurate and Fast Optic Disc Detection in Retinal Images With Vessel Distribution and Directional Characteristics," *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 1, pp. 333–342, 2014.
- [149] M. C. V. S. Mary, E. B. Rajsingh, J. K. K. Jacob, D. Anandhi, U. Amato, and S. E. Selvan, "An empirical study on optic disc segmentation using an active contour model," *Biomedical Signal Processing and Control*, vol. 18, pp. 19–29, 2015.
- [150] S. Samanta, S. K. Saha, and B. Chanda, "A Simple and Fast Algorithm to Detect the Fovea Region in Fundus Retinal Image," in *Emerging Applications of Information Technology (EAIT), 2011 Second International Conference on*, pp. 206–209, IEEE, 2011.
- [151] K. Akyol, B. Şen, and Ş. Bayır, "Automatic Detection of Optic Disc in Retinal Image by Using Keypoint Detection, Texture Analysis, and Visual Dictionary Techniques," *Computational and mathematical methods in medicine*, vol. 2016, 2016.

- [152] T. K. Ho, “Random decision forests,” in *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, vol. 1, pp. 278–282, IEEE, 1995.
- [153] M. W. Khan, M. Sharif, M. Yasmin, and T. Saba, “CDR based glaucoma detection using fundus images: a review,” *International Journal of Applied Pattern Recognition*, vol. 4, no. 3, pp. 261–306, 2017.
- [154] S. Sengupta, A. Singh, H. A. Leopold, T. Gulati, and V. Lakshminarayanan, “Application of Deep Learning in Fundus Image Processing for Ophthalmic Diagnosis - A Review,” *Artificial Intelligence in Medicine*, vol. 102, p. 101758, 2020.
- [155] S. Maheshwari, V. Kanhangad, and R. B. Pachori, “CNN-based approach for glaucoma diagnosis using transfer learning and LBP-based data augmentation,” *arXiv preprint arXiv:2002.08013*, 2020.
- [156] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.
- [157] U. Raghavendra, H. Fujita, S. V. Bhandary, A. Gudigar, J. H. Tan, and U. R. Acharya, “Deep convolution neural network for accurate diagnosis of glaucoma using digital fundus images,” *Information Sciences*, vol. 441, pp. 41–49, 2018.
- [158] B. Al-Bander, W. Al-Nuaimy, M. A. Al-Tae, and Y. Zheng, “Automated glaucoma diagnosis using deep learning approach,” in *2017 14th International Multi-Conference on Systems, Signals & Devices (SSD)*, pp. 207–210, IEEE, 2017.
- [159] R. Shinde, “Glaucoma detection in retinal fundus images using U-Net and supervised machine learning algorithms,” *Intelligence-Based Medicine*, vol. 5, p. 100038, 2021.
- [160] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, *et al.*, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [161] J. A. De La Fuente-Arriaga, E. M. Felipe-Riverón, and E. Garduño-Calderón, “Application of vascular bundle displacement in the optic disc for glaucoma

- detection using fundus images,” *Computers in biology and medicine*, vol. 47, pp. 27–35, 2014.
- [162] H. Ahmad, A. Yamin, A. Shakeel, S. O. Gillani, and U. Ansari, “Detection of glaucoma using retinal fundus images,” in *Robotics and Emerging Allied Technologies in Engineering (iCREATE), 2014 International Conference on*, pp. 321–324, IEEE, 2014.
- [163] F. Khan, S. A. Khan, U. U. Yasin, I. ul Haq, and U. Qamar, “Detection of glaucoma using retinal fundus images,” in *Biomedical Engineering International Conference (BMEiCON), 2013 6th*, pp. 1–5, IEEE, 2013.
- [164] Y. Xu, S. Lin, D. W. K. Wong, J. Liu, and D. Xu, “Efficient Reconstruction-Based Optic Cup Localization for Glaucoma Screening,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 445–452, Springer, 2013.
- [165] A. Li, J. Cheng, D. W. K. Wong, and J. Liu, “Integrating holistic and local deep features for glaucoma classification,” in *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*, pp. 1328–1331, IEEE, 2016.
- [166] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *CoRR*, vol. abs/1409.1556, 2014.
- [167] P. F. Abad, D. Coronado-Gutierrez, C. Lopez, and X. P. Burgos-Artizzu, “Glaucoma patient screening from online retinal fundus images via Artificial Intelligence,” *medRxiv*, 2021.
- [168] J. Wang, L. Yang, Z. Huo, W. He, and J. Luo, “Multi-Label Classification of Fundus Images With EfficientNet,” *IEEE Access*, vol. 8, pp. 212499–212508, 2020.
- [169] M. Tan and Q. V. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” *CoRR*, vol. abs/1905.11946, 2019.
- [170] L. Dai, L. Wu, H. Li, C. Cai, Q. Wu, H. Kong, R. Liu, X. Wang, X. Hou, Y. Liu, *et al.*, “A deep learning system for detecting diabetic retinopathy across the disease spectrum,” *Nature communications*, vol. 12, no. 1, pp. 1–11, 2021.

- [171] G. Mushtaq and F. Siddiqui, "Detection of diabetic retinopathy using deep learning methodology," in *IOP Conference Series: Materials Science and Engineering*, vol. 1070, p. 012049, IOP Publishing, 2021.
- [172] R. Welikala, J. Dehmeshki, A. Hoppe, V. Tah, S. Mann, T. H. Williamson, and S. Barman, "Automated detection of proliferative diabetic retinopathy using a modified line operator and dual classification," *Computer methods and programs in biomedicine*, vol. 114, no. 3, pp. 247–261, 2014.
- [173] Z. Wang, Y. Yin, J. Shi, W. Fang, H. Li, and X. Wang, "Zoom-in-Net: Deep Mining Lesions for Diabetic Retinopathy Detection," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 267–275, Springer, 2017.
- [174] M. Y. Guan, V. Gulshan, A. M. Dai, and G. E. Hinton, "Who Said What: Modeling Individual Labelers Improves Classification," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [175] P. Costa, A. Galdran, M. I. Meyer, M. Niemeijer, M. Abràmoff, A. M. Mendonça, and A. Campilho, "End-to-End Adversarial Retinal Image Synthesis," *IEEE transactions on medical imaging*, vol. 37, no. 3, pp. 781–791, 2018.
- [176] A. Aujih, L. Izhar, F. Mériaudeau, and M. I. Shapiai, "Analysis of Retinal Vessel Segmentation with Deep Learning and its Effect on Diabetic Retinopathy Classification," in *2018 International conference on intelligent and advanced system (ICIAS)*, pp. 1–6, IEEE, 2018.
- [177] H. H. Vo and A. Verma, "New Deep Neural Nets for Fine-Grained Diabetic Retinopathy Recognition on Hybrid Color Space," in *Multimedia (ISM), 2016 IEEE International Symposium on*, pp. 209–215, IEEE, 2016.
- [178] A. C. Somkuwar, T. G. Patil, S. S. Patankar, and J. V. Kulkarni, "Intensity features based classification of hard exudates in retinal images," in *India Conference (INDICON), 2015 Annual IEEE*, pp. 1–5, IEEE, 2015.
- [179] L. Seoud, T. Hurtut, J. Chelbi, F. Cheriet, and J. P. Langlois, "Red Lesion Detection Using Dynamic Shape Features for Diabetic Retinopathy Screening," *IEEE transactions on medical imaging*, vol. 35, no. 4, pp. 1116–1126, 2016.
- [180] A. Rakhlin, "Diabetic Retinopathy detection through integration of Deep Learning classification framework," *bioRxiv*, p. 225508, 2018.

-
- [181] N. Ramachandran, S. C. Hong, M. J. Sime, and G. A. Wilson, “Diabetic retinopathy screening using deep neural network,” *Clinical & experimental ophthalmology*, vol. 46, no. 4, pp. 412–416, 2018.
- [182] G. Quellec, K. Charrière, Y. Boudi, B. Cochener, and M. Lamard, “Deep image mining for diabetic retinopathy screening,” *Medical image analysis*, vol. 39, pp. 178–193, 2017.
- [183] W. L. Alyoubi, W. M. Shalash, and M. F. Abulkhair, “Diabetic retinopathy detection through deep learning techniques: A review,” *Informatics in Medicine Unlocked*, vol. 20, p. 100377, 2020.
- [184] A. Giachetti, L. Ballerini, and E. Trucco, “Accurate and reliable segmentation of the optic disc in digital fundus images,” *Journal of Medical Imaging*, vol. 1, no. 2, pp. 024001–024001, 2014.
- [185] H. Yu, E. S. Barriga, C. Agurto, S. Echegaray, M. S. Pattichis, W. Bauman, and P. Soliz, “Fast Localization and Segmentation of Optic Disk in Retinal Images Using Directional Matched Filtering and Level Sets,” *IEEE Transactions on information technology in biomedicine*, vol. 16, no. 4, pp. 644–657, 2012.
- [186] A. Aquino, M. E. Gegúndez-Arias, and D. Marín, “Detecting the Optic Disc Boundary in Digital Fundus Images Using Morphological, Edge Detection, and Feature Extraction Techniques,” *IEEE transactions on medical imaging*, vol. 29, no. 11, pp. 1860–1869, 2010.
- [187] A. Budai, R. Bock, A. Maier, J. Hornegger, and G. Michelson, “Robust Vessel Segmentation in Fundus Images,” *International journal of biomedical imaging*, vol. 2013, 2013.
- [188] T. Mahmudi, R. Kafieh, H. Rabbani, M. Akhlagi, *et al.*, “Comparison of macular OCTs in right and left eyes of normal people,” in *Medical Imaging 2014: Biomedical Applications in Molecular, Structural, and Functional Imaging*, vol. 9038, p. 90381W, International Society for Optics and Photonics, 2014.
- [189] E. J. Carmona, M. Rincón, J. García-Feijoó, and J. M. Martínez-de-la Casa, “Identification of the optic nerve head with genetic algorithms,” *Artificial Intelligence in Medicine*, vol. 43, no. 3, pp. 243–259, 2008.
- [190] N. Otsu, “A Threshold Selection Method from Gray-Level Histograms,” *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

- [191] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The PASCAL Visual Object Classes (VOC) Challenge,” *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.
- [192] R. J. Qureshi, L. Kovacs, B. Harangi, B. Nagy, T. Peto, and A. Hajdu, “Combining algorithms for automatic detection of optic disc and macula in fundus images,” *Computer Vision and Image Understanding*, vol. 116, no. 1, pp. 138–145, 2012.
- [193] D. A. Godse and D. S. Bormane, “Automated Localization of Optic Discin Retinal Images,” *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 2, pp. 65–71, 2013.
- [194] S. Lu and J. H. Lim, “Automatic optic disc detection through background estimation,” in *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pp. 833–836, IEEE, 2010.
- [195] W. Liao, B. Zou, R. Zhao, Y. Chen, Z. He, and M. Zhou, “Clinical interpretable deep learning model for glaucoma diagnosis,” *IEEE journal of biomedical and health informatics*, vol. 24, no. 5, pp. 1405–1412, 2019.
- [196] T. Nazir, A. Irtaza, A. Javed, H. Malik, D. Hussain, and R. A. Naqvi, “Retinal image analysis for diabetes-based eye disease detection using deep learning,” *Applied Sciences*, vol. 10, no. 18, p. 6185, 2020.
- [197] T. Nazir, A. Irtaza, and V. Starovoitov, “Optic disc and optic cup segmentation for glaucoma detection from blur retinal images using improved mask-rcnn,” *International Journal of Optics*, vol. 2021, 2021.
- [198] M. Nawaz, T. Nazir, A. Javed, U. Tariq, H.-S. Yong, M. A. Khan, and J. Cha, “An efficient deep learning approach to automatic glaucoma detection using optic disc and optic cup localization,” *Sensors*, vol. 22, no. 2, p. 434, 2022.
- [199] O. Deperlioglu, U. Kose, D. Gupta, A. Khanna, F. Giampaolo, and G. Fortino, “Explainable framework for glaucoma diagnosis by image processing and convolutional neural network synergy: Analysis with doctor evaluation,” *Future Generation Computer Systems*, vol. 129, pp. 152–169, 2022.
- [200] Z. Yang, T. Luo, D. Wang, Z. Hu, J. Gao, and L. Wang, “Learning to Navigate for Fine-grained Classification,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 420–435, 2018.

-
- [201] A. Recasens, P. Kellnhofer, S. Stent, W. Matusik, and A. Torralba, “Learning to Zoom: a Saliency-Based Sampling Layer for Neural Networks,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 51–66, 2018.
- [202] American Academy of Ophthalmology, “International clinical diabetic retinopathy disease severity scale,” *International Council of Ophthalmology*, 2002.
- [203] D. Y. Carson Lam, M. Guo, and T. Lindsey, “Automated detection of diabetic retinopathy using deep learning,” *AMIA Summits on Translational Science Proceedings*, vol. 2017, p. 147, 2018.
- [204] M. D. Abràmoff, Y. Lou, A. Erginay, W. Clarida, R. Amelon, J. C. Folk, and M. Niemeijer, “Improved Automated Detection of Diabetic Retinopathy on a Publicly Available Dataset Through Integration of Deep Learning,” *Investigative ophthalmology & visual science*, vol. 57, no. 13, pp. 5200–5206, 2016.
- [205] M. Shaban, Z. Ogur, A. Mahmoud, A. Switala, A. Shalaby, H. Abu Khalifeh, M. Ghazal, L. Fraiwan, G. Giridharan, H. Sandhu, *et al.*, “A convolutional neural network for the screening and staging of diabetic retinopathy,” *Plos one*, vol. 15, no. 6, p. e0233514, 2020.
- [206] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer, “Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data,” *MethodsX*, vol. 7, p. 100864, 2020.
- [207] P. Kharazmi, *Automated analysis of vascular structures of skin lesions: segmentation, pattern recognition and computer-aided diagnosis*.
PhD thesis, University of British Columbia, 2018.
- [208] J. S. Ellen, C. A. Graff, and M. D. Ohman, “Improving plankton image classification using context metadata,” *Limnology and Oceanography: Methods*, vol. 17, no. 8, pp. 439–461, 2019.
- [209] A. G. Pacheco, A.-R. Ali, and T. Trappenberg, “Skin cancer detection based on deep learning and entropy to detect outlier samples,” *arXiv preprint arXiv:1909.04525*, 2019.
- [210] Y. Huang, Y. Cheng, D. Chen, H. Lee, J. Ngiam, Q. V. Le, and Z. Chen, “GPipe: Efficient Training of Giant Neural Networks using Pipeline Parallelism,” *arXiv preprint arXiv:1811.06965*, 2018.

- [211] X. Yang, Y. Zhang, Y. Yang, and W. Lv, “Deterministic and Probabilistic Wind Power Forecasting Based on Bi-Level Convolutional Neural Network and Particle Swarm Optimization ,” *Applied Sciences*, vol. 9, no. 9, p. 1794, 2019.
- [212] J. M.-T. Wu, Z. Li, N. Herencsar, B. Vo, and J. C.-W. Lin, “A graph-based CNN-LSTM stock price prediction algorithm with leading indicators,” *Multimedia Systems*, pp. 1–20, 2021.
- [213] D. Khaledyan, A. Tajally, R. Sarkhosh, A. Shamsi, H. Asgharnezhad, A. Khosravi, and S. Nahavandi, “Confidence Aware Neural Networks for Skin Cancer Detection,” *arXiv preprint arXiv:2107.09118*, 2021.
- [214] T. Li, Y. Zhang, and T. Wang, “SRPM–CNN: a combined model based on slide relative position matrix and CNN for time series classification,” *Complex & Intelligent Systems*, vol. 7, no. 3, pp. 1619–1631, 2021.
- [215] B. K. Iwana and S. Uchida, “An empirical survey of data augmentation for time series classification with neural networks,” *Plos one*, vol. 16, no. 7, p. e0254841, 2021.
- [216] M. Munir, S. A. Siddiqui, A. Dengel, and S. Ahmed, “DeepAnT: A Deep Learning Approach for Unsupervised Anomaly Detection in Time Series,” *IEEE Access*, vol. 7, pp. 1991–2005, 2019.
- [217] C.-Y. Hsu and W.-C. Liu, “Multiple time-series convolutional neural network for fault detection and diagnosis and empirical study in semiconductor manufacturing,” *Journal of Intelligent Manufacturing*, vol. 32, pp. 823–836, 2021.
- [218] D. Ho, E. Liang, I. Stoica, P. Abbeel, and X. Chen, “Population Based Augmentation: Efficient Learning of Augmentation Policy Schedules,” *arXiv preprint arXiv:1905.05393*, 2019.
- [219] K. Shridhar, *A comprehensive guide to Bayesian CNN with variational inference : with implementation in PyTorch*. LAP Lambert Academic Publishing, 2019.
- [220] J. Snoek, O. Rippel, K. Swersky, R. Kiros, N. Satish, N. Sundaram, M. Patwary, M. Prabhat, and R. Adams, “Scalable Bayesian Optimization Using Deep Neural Networks,” in *International conference on machine learning*, pp. 2171–2180, 2015.

-
- [221] R. M. Neal, *Probabilistic inference using Markov chain Monte Carlo methods*. Department of Computer Science, University of Toronto Toronto, Ontario, Canada, 1993.
- [222] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, “An Introduction to Variational Methods for Graphical Models,” *Machine learning*, vol. 37, no. 2, pp. 183–233, 1999.
- [223] T. Salimans, D. P. Kingma, and M. Welling, “Markov Chain Monte Carlo and Variational Inference: Bridging the Gap,” *arXiv preprint arXiv:1410.6460*, 2014.
- [224] K. Shridhar, F. Laumann, A. Llopart Maurin, and M. Liwicki, “Bayesian Convolutional Neural Network,” *arXiv preprint arXiv:1806.05978*, 2018.
- [225] I. Kononenko, “Bayesian neural networks,” *Biological Cybernetics*, vol. 61, no. 5, pp. 361–370, 1989.
- [226] A. Bate, M. Lindquist, I. R. Edwards, S. Olsson, R. Orre, A. Lansner, and R. M. De Freitas, “A Bayesian neural network method for adverse drug reaction signal generation,” *European journal of clinical pharmacology*, vol. 54, no. 4, pp. 315–321, 1998.
- [227] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, “Variational Inference: A Review for Statisticians,” *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.
- [228] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, “Weight Uncertainty in Neural Network,” *arXiv preprint arXiv:1505.05424*, 2015.
- [229] Y. Gal and Z. Ghahramani, “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning,” in *international conference on machine learning*, pp. 1050–1059, 2016.
- [230] Y. Tang and R. R. Salakhutdinov, “Learning Stochastic Feedforward Neural Networks,” in *Advances in Neural Information Processing Systems*, pp. 530–538, 2013.
- [231] M. Lázaro-Gredilla and A. R. Figueiras-Vidal, “Marginalized Neural Network Mixtures for Large-Scale Regression,” *IEEE transactions on neural networks*, vol. 21, no. 8, pp. 1345–1351, 2010.

- [232] J. Kwon, C. Shin, E. Lee, Y. Han, and S. Hong, “Hybrid HMM-MLP classifier for prosthetic arm control purpose,” in *Proceedings of Digital Processing Applications (TENCON’96)*, vol. 1, pp. 21–24, IEEE, 1996.
- [233] A. Lins and T. B. Ludermir, “Hybrid optimization algorithm for the definition of MLP neural network architectures and weights,” in *Fifth International Conference on Hybrid Intelligent Systems (HIS’05)*, pp. 6–pp, IEEE, 2005.
- [234] M. Yang, S. Wang, J. Bakita, T. Vu, F. D. Smith, J. H. Anderson, and J.-M. Frahm, “Re-Thinking CNN Frameworks for Time-Sensitive Autonomous-Driving Applications: Addressing an Industrial Challenge,” in *2019 IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, pp. 305–317, IEEE, 2019.
- [235] Y. Kwon, J.-H. Won, B. J. Kim, and M. C. Paik, “Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation,” *Computational Statistics & Data Analysis*, vol. 142, p. 106816, 2020.
- [236] A. Kendall and Y. Gal, “What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?,” in *Advances in neural information processing systems*, pp. 5574–5584, 2017.
- [237] D. Hendrycks and K. Gimpel, “A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks,” *arXiv preprint arXiv:1610.02136*, 2016.
- [238] K. Lee, H. Lee, K. Lee, and J. Shin, “Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples,” *arXiv preprint arXiv:1711.09325*, 2017.
- [239] G. Papamakarios, T. Pavlakou, and I. Murray, “Masked Autoregressive Flow for Density Estimation,” in *Advances in Neural Information Processing Systems*, pp. 2338–2347, 2017.
- [240] A. v. d. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. v. d. Driessche, E. Lockhart, L. C. Cobo, F. Stimberg, *et al.*, “Parallel WaveNet: Fast High-Fidelity Speech Synthesis,” *arXiv preprint arXiv:1711.10433*, 2017.

-
- [241] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling, “Improved Variational Inference with Inverse Autoregressive Flow,” in *Advances in neural information processing systems*, pp. 4743–4751, 2016.
- [242] C.-W. Huang, D. Krueger, A. Lacoste, and A. Courville, “Neural Autoregressive Flows,” *arXiv preprint arXiv:1804.00779*, 2018.
- [243] N. De Cao, I. Titov, and W. Aziz, “Block Neural Autoregressive Flow,” *arXiv preprint arXiv:1904.04676*, 2019.
- [244] L. Dinh, J. Sohl-Dickstein, and S. Bengio, “Density estimation using Real NVP,” *arXiv preprint arXiv:1605.08803*, 2016.
- [245] L. Ardizzone, J. Kruse, S. Wirkert, D. Rahner, E. W. Pellegrini, R. S. Klessen, L. Maier-Hein, C. Rother, and U. Köthe, “Analyzing Inverse Problems with Invertible Neural Networks,” *arXiv preprint arXiv:1808.04730*, 2018.
- [246] D. P. Kingma and P. Dhariwal, “Glow: Generative Flow with Invertible 1x1 Convolutions,” in *Advances in Neural Information Processing Systems*, pp. 10215–10224, 2018.
- [247] L. Dinh, D. Krueger, and Y. Bengio, “NICE: Non-linear Independent Components Estimation,” *arXiv preprint arXiv:1410.8516*, 2014.
- [248] M.-H. Laves, S. Ihler, and T. Ortmaier, “Uncertainty Quantification in Computer-Aided Diagnosis: Make Your Model say “I don’t know” for Ambiguous Cases,” *arXiv preprint arXiv:1908.00792*, 2019.
- [249] A. G. Roy, S. Conjeti, N. Navab, C. Wachinger, A. D. N. Initiative, *et al.*, “Quick-NAT: A fully convolutional network for quick and accurate segmentation of neuroanatomy,” *NeuroImage*, vol. 186, pp. 713–727, 2019.
- [250] M.-H. Laves, J. Bicker, L. A. Kahrs, and T. Ortmaier, “A dataset of laryngeal endoscopic images with comparative study on convolution neural network-based semantic segmentation,” *International journal of computer assisted radiology and surgery*, vol. 14, no. 3, pp. 483–492, 2019.
- [251] A. Torralba, R. Fergus, and W. T. Freeman, “80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 11, pp. 1958–1970, 2008.

- [252] H. A. Dau, E. Keogh, K. Kamgar, C.-C. M. Yeh, Y. Zhu, S. Gharghabi, C. A. Ratanamahatana, Yanping, B. Hu, N. Begum, A. Bagnall, A. Mueen, and G. Batista, “The UCR Time Series Classification Archive,” October 2018.
https://www.cs.ucr.edu/~eamonn/time_series_data_2018/.
- [253] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On Calibration of Modern Neural Networks,” in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1321–1330, JMLR. org, 2017.
- [254] D. P. Kingma, T. Salimans, and M. Welling, “Variational dropout and the local reparameterization trick,” *Advances in neural information processing systems*, vol. 28, pp. 2575–2583, 2015.
- [255] D. Molchanov, A. Ashukha, and D. Vetrov, “Variational Dropout Sparsifies Deep Neural Networks,” in *International Conference on Machine Learning*, pp. 2498–2507, PMLR, 2017.
- [256] H. Pham, M. Y. Guan, B. Zoph, Q. V. Le, and J. Dean, “Efficient Neural Architecture Search via Parameters Sharing,” *arXiv preprint arXiv:1802.03268*, 2018.
- [257] M. T. Ribeiro, S. Singh, and C. Guestrin, “”Why Should I Trust You?”: Explaining the Predictions of Any Classifier,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.
- [258] W. Guo, “Explainable Artificial Intelligence (XAI) for 6G: Improving Trust between Human and Machine,” *IEEE Communications Magazine*, vol. 58, no. 6, pp. 39–45, 2020.
- [259] B. Walzl and R. Vogl, “Explainable artificial intelligence - the new frontier in legal informatics,” *Jusletter IT*, vol. 4, pp. 1–10, 2018.
- [260] R. L. Teach and E. H. Shortliffe, “An analysis of physician attitudes regarding computer-based clinical consultation systems,” *Computers and Biomedical Research*, vol. 14, no. 6, pp. 542–558, 1981.
- [261] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, *et al.*, “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” *Information Fusion*, vol. 58, pp. 82–115, 2020.

- [262] Council of the European Union, “Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).” available at <http://data.europa.eu/eli/reg/2016/679/2016-05-04>, April 2016.
- [263] M. Izadyazdanabadi, E. Belykh, C. Cavallo, X. Zhao, S. Gandhi, L. B. Moreira, J. Eschbacher, P. Nakaji, M. C. Preul, and Y. Yang, “Weakly-Supervised Learning-Based Feature Localization for Confocal Laser Endomicroscopy Glioma Images,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 300–308, Springer, 2018.
- [264] American Cancer Society, “Cancer Facts & Figures 2017.” Available at: <https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2017/cancer-facts-and-figures-2017.pdf>, 2017.
- [265] R. L. Siegel, K. D. Miller, and A. Jemal, “Cancer statistics, 2019,” *CA: A Cancer Journal for Clinicians*, vol. 69, no. 1, pp. 7–34, 2019.
- [266] E. S. Ruiz, F. C. Morgan, C. M. Zigler, R. J. Besaw, and C. D. Schmults, “Analysis of national skin cancer expenditures in the United States Medicare population, 2013,” *Journal of the American Academy of Dermatology*, vol. 80, no. 1, pp. 275–278, 2019.
- [267] A. B. Fortina, E. Peserico, A. Silletti, and E. Zattra, “Where’s the naevus? Inter-operator variability in the localization of melanocytic lesion border,” *Skin Research and Technology*, vol. 18, no. 3, pp. 311–315, 2012.
- [268] R. Corona, A. Mele, M. Amini, G. De Rosa, G. Coppola, P. Piccardi, M. Fucci, P. Pasquini, and T. Faraggiana, “Interobserver variability on the histopathologic diagnosis of cutaneous melanoma and other pigmented skin lesions,” *Journal of clinical Oncology*, vol. 14, no. 4, pp. 1218–1223, 1996.
- [269] A. Ghorbani, D. Ouyang, A. Abid, B. He, J. H. Chen, R. A. Harrington, D. H. Liang, E. A. Ashley, and J. Y. Zou, “Deep learning interpretation of echocardiograms,” *bioRxiv*, p. 681676, 2019.

- [270] S. Jolly, B. K. Iwana, R. Kuroki, and S. Uchida, “How do Convolutional Neural Networks Learn Design?,” in *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 1085–1090, IEEE, 2018.
- [271] 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*, pp. 618–626, 2017.
- [272] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.
- [273] M. D. Zeiler and R. Fergus, “Visualizing and Understanding Convolutional Networks,” in *European conference on computer vision*, pp. 818–833, Springer, 2014.
- [274] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning Deep Features for Discriminative Localization,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.
- [275] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” *arXiv preprint arXiv:1312.6034*, 2013.
- [276] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, “Not Just a Black Box: Learning Important Features Through Propagating Activation Differences,” *arXiv preprint arXiv:1605.01713*, 2016.
- [277] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic Attribution for Deep Networks,” in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3319–3328, JMLR.org, 2017.
- [278] A. Shrikumar, P. Greenside, and A. Kundaje, “Learning Important Features Through Propagating Activation Differences,” in *International Conference on Machine Learning*, pp. 3145–3153, PMLR, 2017.
- [279] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, “On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation,” *PloS one*, vol. 10, no. 7, 2015.

-
- [280] C. Singh, W. J. Murdoch, and B. Yu, “Hierarchical interpretations for neural network predictions,” *arXiv preprint arXiv:1806.05337*, 2018.
- [281] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in *Advances in neural information processing systems*, pp. 4765–4774, 2017.
- [282] V. Petsiuk, A. Das, and K. Saenko, “RISE: Randomized Input Sampling for Explanation of Black-box Models,” in *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.
- [283] R. Fong, M. Patrick, and A. Vedaldi, “Understanding Deep Networks via Extremal Perturbations and Smooth Masks,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2950–2958, 2019.
- [284] J. Ba, V. Mnih, and K. Kavukcuoglu, “Multiple Object Recognition with Visual Attention,” *arXiv preprint arXiv:1412.7755*, 2014.
- [285] S. Jetley, N. A. Lord, N. Lee, and P. H. Torr, “Learn To Pay Attention,” *arXiv preprint arXiv:1804.02391*, 2018.
- [286] M. Jaderberg, K. Simonyan, A. Zisserman, *et al.*, “Spatial Transformer Networks,” in *Advances in neural information processing systems*, pp. 2017–2025, 2015.
- [287] A. Mahendran and A. Vedaldi, “Visualizing Deep Convolutional Neural Networks Using Natural Pre-images,” *International Journal of Computer Vision*, vol. 120, no. 3, pp. 233–255, 2016.
- [288] C. Olah, A. Satyanarayan, I. Johnson, S. Carter, L. Schubert, K. Ye, and A. Mordvintsev, “The Building Blocks of Interpretability,” *Distill*, vol. 3, no. 3, p. e10, 2018.
- [289] M. Graziani, J. M. Brown, V. Andrearczyk, V. Yildiz, J. P. Campbell, D. Ergogmus, S. Ioannidis, M. F. Chiang, J. Kalpathy-Cramer, and H. Müller, “Improved interpretability for computer-aided severity assessment of retinopathy of prematurity,” in *Medical Imaging 2019: Computer-Aided Diagnosis*, vol. 10950, p. 109501R, International Society for Optics and Photonics, 2019.
- [290] S. Toba, Y. Mitani, H. Ohashi, H. Sawada, N. Yodoya, H. Hayakawa, M. Hirayama, A. Futsuki, N. Yamamoto, H. Ito, *et al.*, “Quantitative Analysis of Chest X-Ray Using Deep Learning to Predict Pulmonary to Systemic Flow Ratio in

- Patients With Congenital Heart Disease,” *Circulation*, vol. 140, no. Suppl_1, pp. A14250–A14250, 2019.
- [291] V. Couteaux, O. Nempont, G. Pizaine, and I. Bloch, “Towards Interpretability of Segmentation Networks by Analyzing DeepDreams,” in *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, pp. 56–63, Springer, 2019.
- [292] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, *et al.*, “Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV),” in *International conference on machine learning*, pp. 2668–2677, PMLR, 2018.
- [293] M. Graziani, V. Andrearczyk, and H. Müller, “Regression Concept Vectors for Bidirectional Explanations in Histopathology,” in *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, pp. 124–132, Springer, 2018.
- [294] H. Yeche, J. Harrison, and T. Berthier, “UBS: A Dimension-Agnostic Metric for Concept Vector Interpretability Applied to Radiomics,” in *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, pp. 12–20, Springer, 2019.
- [295] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, “Network Dissection: Quantifying Interpretability of Deep Visual Representations,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6541–6549, 2017.
- [296] B. Zhou, Y. Sun, D. Bau, and A. Torralba, “Interpretable Basis Decomposition for Visual Explanation,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 119–134, 2018.
- [297] R. Caruana, “Multitask learning,” *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.
- [298] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh, “Seven-Point Checklist and Skin Lesion Classification Using Multitask Multimodal Neural Nets,” *IEEE Journal of Biomedical and Health Informatics*, vol. 23, pp. 538–546, mar 2019.

-
- [299] J. Kawahara and G. Hamarneh, “Fully Convolutional Neural Networks to Detect Clinical Dermoscopic Features,” *IEEE journal of biomedical and health informatics*, vol. 23, no. 2, pp. 578–585, 2018.
- [300] D. Coppola, H. Kuan Lee, and C. Guan, “Interpreting Mechanisms of Prediction for Skin Cancer Diagnosis Using Multi-Task Learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 734–735, 2020.
- [301] R. Zhang, S. Tan, R. Wang, S. Manivannan, J. Chen, H. Lin, and W.-S. Zheng, “Biomarker Localization by Combining CNN Classifier and Generative Adversarial Network,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 209–217, Springer, 2019.
- [302] L. A. Hendricks, R. Hu, T. Darrell, and Z. Akata, “Grounding visual explanations,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 264–279, 2018.
- [303] P. Guo, C. Anderson, K. Pearson, and R. Farrell, “Neural Network Interpretation via Fine Grained Textual Summarization,” *arXiv preprint arXiv:1805.08969*, 2018.
- [304] M. Munir, S. A. Siddiqui, F. Küsters, D. Mercier, A. Dengel, and S. Ahmed, “TSXplain: Demystification of DNN Decisions for Time-Series using Natural Language and Statistical Features,” in *International Conference on Artificial Neural Networks*, pp. 426–439, Springer, 2019.
- [305] B. Hancock, M. Bringmann, P. Varma, P. Liang, S. Wang, and C. Ré, “Training Classifiers with Natural Language Explanations,” in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2018, p. 1884, NIH Public Access, 2018.
- [306] H. Lakkaraju, E. Kamar, R. Caruana, and J. Leskovec, “Faithful and Customizable Explanations of Black Box Models,” in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 131–138, 2019.
- [307] J. Rabold, H. Deininger, M. Siebers, and U. Schmid, “Enriching Visual with Verbal Explanations for Relational Concepts - Combining LIME with Aleph,” in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 180–192, Springer, 2019.

- [308] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell, “Generating Visual Explanations,” in *European Conference on Computer Vision*, pp. 3–19, Springer, 2016.
- [309] Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang, “MDNet: A Semantically and Visually Interpretable Medical Image Diagnosis Network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6428–6436, 2017.
- [310] A. S. Ross, M. C. Hughes, and F. Doshi-Velez, “Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations,” *arXiv preprint arXiv:1703.03717*, 2017.
- [311] G. Erion, J. D. Janizek, P. Sturmfels, S. Lundberg, and S.-I. Lee, “Learning Explainable Models Using Attribution Priors,” *arXiv preprint arXiv:1906.10670*, 2019.
- [312] M. Mitsuhashi, H. Fukui, Y. Sakashita, T. Ogata, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, “Embedding Human Knowledge in Deep Neural Network via Attention Map,” *arXiv preprint arXiv:1905.03540*, vol. 5, 2019.
- [313] M. Graziani, S. Otálora, H. Muller, and V. Andrearczyk, “Guiding CNNs towards Relevant Concepts by Multi-task and Adversarial Learning,” *arXiv preprint arXiv:2008.01478*, 2020.
- [314] Y. Yan, J. Kawahara, and G. Hamarneh, “Melanoma Recognition via Visual Attention,” in *International Conference on Information Processing in Medical Imaging*, pp. 793–804, Springer, 2019.
- [315] L. Rieger, C. Singh, W. Murdoch, and B. Yu, “Interpretations are Useful: Penalizing Explanations to Align Neural Networks with Prior Knowledge,” in *International Conference on Machine Learning*, pp. 8116–8126, PMLR, 2020.
- [316] Y. Yamamoto, T. Tsuzuki, J. Akatsuka, M. Ueki, H. Morikawa, Y. Numata, T. Takahara, T. Tsuyuki, K. Tsutsumi, R. Nakazawa, *et al.*, “Automated acquisition of explainable knowledge from unannotated histopathology images,” *Nature communications*, vol. 10, no. 1, pp. 1–9, 2019.
- [317] C. Jansen, T. Penzel, S. Hodel, S. Breuer, M. Spott, and D. Krefting, “Network physiology in insomnia patients: Assessment of relevant changes in network

- topology with interpretable machine learning models,” *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 29, no. 12, p. 123129, 2019.
- [318] A. P. Carrieri, N. Haiminen, S. Maudsley-Barton, L.-J. Gardiner, B. Murphy, A. Mayes, S. Paterson, S. Grimshaw, M. Winn, C. Shand, *et al.*, “Explainable AI reveals key changes in skin microbiome associated with menopause, smoking, aging and skin hydration,” *bioRxiv*, 2020.
- [319] A. Essemlali, E. St-Onge, M. Descoteaux, and P.-M. Jodoin, “Understanding Alzheimer disease’s structural connectivity through explainable AI,” in *Medical Imaging with Deep Learning*, pp. 217–229, PMLR, 2020.
- [320] F. Nunnari and D. Sonntag, “A CNN toolbox for skin cancer classification,” *arXiv preprint arXiv:1908.08187*, 2019.
- [321] D. Sonntag, F. Nunnari, and H.-J. Profitlich, “The Skincare project, an interactive deep learning system for differential diagnosis of malignant skin lesions. Technical Report,” *arXiv preprint arXiv:2005.09448*, 2020.
- [322] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, “A Benchmark for Interpretability Methods in Deep Neural Networks,” in *Advances in Neural Information Processing Systems*, pp. 9734–9745, 2019.
- [323] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, “Evaluating the Visualization of What a Deep Neural Network Has Learned,” *IEEE transactions on neural networks and learning systems*, vol. 28, no. 11, pp. 2660–2673, 2016.
- [324] E. Tjoa and C. Guan, “Quantifying Explainability of Saliency Methods in Deep Neural Networks,” *arXiv preprint arXiv:2009.02899*, 2020.
- [325] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, “Sanity Checks for Saliency Maps,” in *Advances in Neural Information Processing Systems*, pp. 9505–9515, 2018.
- [326] F. Eitel, K. Ritter, A. D. N. I. (ADNI, *et al.*), “Testing the Robustness of Attribution Methods for Convolutional Neural Networks in MRI-Based Alzheimer’s Disease Classification,” in *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, pp. 3–11, Springer, 2019.

- [327] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, “Striving for Simplicity: The All Convolutional Net,” *arXiv preprint arXiv:1412.6806*, 2014.
- [328] F. Doshi-Velez and B. Kim, “Towards A Rigorous Science of Interpretable Machine Learning,” *arXiv preprint arXiv:1702.08608*, 2017.
- [329] A.-p. Nguyen and M. R. Martínez, “On quantitative aspects of model interpretability,” *arXiv preprint arXiv:2007.07584*, 2020.
- [330] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan, “Interpreting Interpretability: Understanding Data Scientists’ Use of Interpretability Tools for Machine Learning,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2020.
- [331] R. Sayres, A. Taly, E. Rahimy, K. Blumer, D. Coz, N. Hammel, J. Krause, A. Narayanaswamy, Z. Rastegar, D. Wu, *et al.*, “Using a Deep Learning Algorithm and Integrated Gradients Explanation to Assist Grading for Diabetic Retinopathy,” *Ophthalmology*, vol. 126, no. 4, pp. 552–564, 2019.
- [332] E. Tjoa and C. Guan, “A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, p. 4793–4813, Nov 2021.
- [333] R. V. Zicari, J. Brodersen, J. Brusseau, B. Düdler, T. Eichhorn, T. Ivanov, G. Kararigas, P. Kringen, M. McCullough, F. Möslein, *et al.*, “Z-inspection@: A process to assess trustworthy ai,” *IEEE Transactions on Technology and Society*, 2021.
- [334] G. Elwyn, I. Scholl, C. Tietbohl, M. Mann, A. G. Edwards, C. Clay, F. Légaré, T. van der Weijden, C. L. Lewis, R. M. Wexler, *et al.*, “”Many miles to go ...”: A systematic review of the implementation of patient decision support interventions into routine clinical practice,” *BMC medical informatics and decision making*, vol. 13, no. 2, pp. 1–10, 2013.
- [335] E. B. Cole, Z. Zhang, H. S. Marques, R. Edward Hendrick, M. J. Yaffe, and E. D. Pisano, “Impact of Computer-Aided Detection Systems on Radiologist Accuracy With Digital Mammography,” *American Journal of Roentgenology*, vol. 203, no. 4, pp. 909–916, 2014.
- [336] A. Kohli and S. Jha, “Why CAD Failed in Mammography,” *Journal of the American College of Radiology*, vol. 15, no. 3, pp. 535–537, 2018.

-
- [337] C. J. Cai, E. Reif, N. Hegde, J. Hipp, B. Kim, D. Smilkov, M. Wattenberg, F. Viégas, G. S. Corrado, M. C. Stumpe, *et al.*, “Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2019.
- [338] C. J. Cai, J. Jongejan, and J. Holbrook, “The effects of example-based explanations in a machine learning interface,” in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pp. 258–262, 2019.
- [339] Q. Yang, A. Steinfeld, and J. Zimmerman, “Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–11, 2019.
- [340] M. Jirotko, R. Procter, M. Hartswood, R. Slack, A. Simpson, C. Coopmans, C. Hinds, and A. Voss, “Collaboration and trust in healthcare innovation: The eDiaMoND case study,” *Computer Supported Cooperative Work (CSCW)*, vol. 14, no. 4, pp. 369–398, 2005.
- [341] E. Beede, E. Baylor, F. Hersch, A. Iurchenko, L. Wilcox, P. Ruamviboonsuk, and L. M. Vardoulakis, “A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2020.
- [342] A. L. Beam and I. S. Kohane, “Translating Artificial Intelligence Into Clinical Care,” *JAMA*, vol. 316, no. 22, pp. 2368–2369, 2016.
- [343] M. R. Arbabshirani, B. K. Fornwalt, G. J. Mongelluzzo, J. D. Suever, B. D. Geise, A. A. Patel, and G. J. Moore, “Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration,” *NPJ digital medicine*, vol. 1, no. 1, pp. 1–7, 2018.
- [344] G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Verbert, and L. Cilar, “Interpretability of machine learning based prediction models in healthcare,” *arXiv preprint arXiv:2002.08596*, 2020.

- [345] A. Singh, S. Sengupta, and V. Lakshminarayanan, “Explainable Deep Learning Models in Medical Image Analysis,” *Journal of Imaging*, vol. 6, no. 6, p. 52, 2020.
- [346] G. Vilone and L. Longo, “Explainable Artificial Intelligence: a Systematic Review,” *arXiv preprint arXiv:2006.00093*, 2020.
- [347] D. Huk Park, L. Anne Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, and M. Rohrbach, “Multimodal Explanations: Justifying Decisions and Pointing to the Evidence,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8779–8788, 2018.
- [348] K. Alipour, J. P. Schulze, Y. Yao, A. Ziskind, and G. Burachas, “A Study on Multimodal and Interactive Explanations for Visual Question Answering,” *arXiv preprint arXiv:2003.00431*, 2020.
- [349] C.-K. Yeh, B. Kim, S. Arik, C.-L. Li, T. Pfister, and P. Ravikumar, “On Completeness-aware Concept-Based Explanations in Deep Neural Networks,” *Advances in Neural Information Processing Systems*, vol. 33, 2020.
- [350] A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim, “Towards Automatic Concept-based Explanations,” in *Advances in Neural Information Processing Systems*, pp. 9277–9286, 2019.
- [351] H.-L. Yang, J. J. Kim, J. H. Kim, Y. K. Kang, D. H. Park, H. S. Park, H. K. Kim, and M.-S. Kim, “Weakly supervised lesion localization for age-related macular degeneration detection using optical coherence tomography images,” *PloS one*, vol. 14, no. 4, p. e0215076, 2019.
- [352] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [353] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention Is All You Need,” in *Advances in neural information processing systems*, pp. 5998–6008, 2017.
- [354] F. Cabitza, R. Rasoini, and G. F. Gensini, “Unintended Consequences of Machine Learning in Medicine,” *JAMA*, vol. 318, no. 6, pp. 517–518, 2017.



- [355] M. A. Qureshi and D. Greene, “EVE: explainable vector based embedding technique using Wikipedia,” *Journal of Intelligent Information Systems*, vol. 53, no. 1, pp. 137–165, 2019.
- [356] J.-R. Rehse, N. Mehdiyev, and P. Fettke, “Towards Explainable Process Predictions for Industry 4.0 in the DFKI-Smart-Lego-Factory,” *KI-Künstliche Intelligenz*, vol. 33, no. 2, pp. 181–187, 2019.
- [357] J. A. Glomsrud, A. Ødegårdstuen, A. L. S. Clair, and Ø. Smogeli, “Trustworthy versus Explainable AI in Autonomous Vessels,” in *Proceedings of the International Seminar on Safety and Security of Autonomous Vessels (ISSAV) and European STAMP Workshop and Conference (ESWC) 2019*, pp. 37–47, Sci-endo, 2020.
- [358] J. Kahn, “Artificial Intelligence Has Some Explaining to Do.” <https://www.bloomberg.com/news/articles/2018-12-12/artificial-intelligence-has-some-explaining-to-do>, 2018.
Accessed: April 26, 2020.
- [359] A. Ghorbani, J. Wexler, and B. Kim, “Automating Interpretability: Discovering and Testing Visual Concepts Learned by Neural Networks,” *arXiv preprint arXiv:1902.03129*, 2019.
- [360] H. Kittler, A. A. Marghoob, G. Argenziano, C. Carrera, C. Curiel-Lewandrowski, R. Hofmann-Wellenhof, J. Malvehy, S. Menzies, S. Puig, H. Rabinovitz, *et al.*, “Standardization of terminology in dermoscopy/dermatoscopy: Results of the third consensus conference of the International Society of Dermoscopy,” *Journal of the American Academy of Dermatology*, vol. 74, no. 6, pp. 1093–1106, 2016.
- [361] G. Argenziano, H. P. Soyer, S. Chimenti, R. Talamini, R. Corona, F. Sera, M. Binder, L. Cerroni, G. De Rosa, G. Ferrara, *et al.*, “Dermoscopy of pigmented skin lesions: Results of a consensus meeting via the Internet,” *Journal of the American Academy of Dermatology*, vol. 48, no. 5, pp. 679–693, 2003.
- [362] S. Menzies, C. Ingvar, and W. McCarthy, “A sensitivity and specificity analysis of the surface microscopy features of invasive melanoma,” *Melanoma research*, vol. 6, no. 1, pp. 55–62, 1996.

- [363] R. J. Friedman, D. S. Rigel, and A. W. Kopf, “Early detection of malignant melanoma: The role of physician examination and self-examination of the skin,” *CA: A Cancer Journal for Clinicians*, vol. 35, no. 3, pp. 130–151, 1985.
- [364] A. Menegola, J. Tavares, M. Fornaciali, L. T. Li, S. Avila, and E. Valle, “RECOD Titans at ISIC challenge 2017,” *arXiv preprint arXiv:1703.04819*, 2017.
- [365] R. P. Braun, H. S. Rabinovitz, J. Krischer, J. Kreusch, M. Oliviero, L. Naldi, A. W. Kopf, and J. H. Saurat, “Dermoscopy of Pigmented Seborrheic KeratosisA Morphological Study,” *Archives of dermatology*, vol. 138, no. 12, pp. 1556–1560, 2002.
- [366] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep Learning Face Attributes in the Wild,” in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.
- [367] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, “SmoothGrad: removing noise by adding noise,” *arXiv preprint arXiv:1706.03825*, 2017.
- [368] T. Tieleman and G. Hinton, “Lecture 6.5 - RMSProp: Divide the gradient by a running average of its recent magnitude.” COURSERA: Neural Networks for Machine Learning, 2012.
- [369] J. Hu, L. Cao, T. Tong, Q. Ye, S. Zhang, K. Li, F. Huang, L. Shao, and R. Ji, “Architecture Disentanglement for Deep Neural Networks,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 672–681, 2021.
- [370] M. Combalia, N. C. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, C. Carrera, A. Barreiro, A. C. Halpern, S. Puig, *et al.*, “BCN20000: Dermoscopic Lesions in the Wild,” *arXiv preprint arXiv:1908.02288*, 2019.
- [371] S. M. de Faria, J. N. Filipe, P. M. Pereira, L. M. Tavora, P. A. Assuncao, M. O. Santos, R. Fonseca-Pinto, F. Santiago, V. Dominguez, and M. Henrique, “Light Field Image Dataset of Skin Lesions,” in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 3905–3908, IEEE, 2019.
- [372] G. Argenziano, G. Fabbrocini, P. Carli, V. De Giorgi, E. Sammarco, and M. Delfino, “Epiluminescence Microscopy for the Diagnosis of Doubtful

- Melanocytic Skin Lesions Comparison of the ABCD Rule of Dermatoscopy and a New 7-Point Checklist Based on Pattern Analysis,” *Archives of dermatology*, vol. 134, no. 12, pp. 1563–1570, 1998.
- [373] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, “Causability and explainability of artificial intelligence in medicine,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 4, p. e1312, 2019.
- [374] R. C. Fong and A. Vedaldi, “Interpretable Explanations of Black Boxes by Meaningful Perturbation,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3429–3437, 2017.
- [375] J. Pearl and D. Mackenzie, *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018.
- [376] J. Yanase and E. Triantaphyllou, “A systematic survey of computer-aided diagnosis in medicine: Past and present developments,” *Expert Systems with Applications*, vol. 138, p. 112821, 2019.
- [377] J. S. Adelman, M. A. Berger, A. Rai, W. L. Galanter, B. L. Lambert, G. D. Schiff, D. K. Vawdrey, R. A. Green, H. Salmasian, R. Koppel, *et al.*, “A national survey assessing the number of records allowed open in electronic health records at hospitals and ambulatory sites,” *Journal of the American Medical Informatics Association*, vol. 24, no. 5, pp. 992–995, 2017.
- [378] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, “Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission,” in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1721–1730, 2015.
- [379] H. G. Welch and W. C. Black, “Overdiagnosis in Cancer,” *Journal of the National Cancer Institute*, vol. 102, no. 9, pp. 605–613, 2010.
- [380] R. Moynihan, *Preventing Overdiagnosis: How to Stop Harming the Healthy*. PhD thesis, Faculty of Health Sciences and Medicine, Bond University, Australia, 2016.
- [381] R. A. Smith, O. W. Brawley, and R. C. Wender, “Screening and Early Detection,” *The American Cancer Society’s Principles of Oncology: Prevention to Survivorship*, pp. 110–35, 2018.

- [382] M. Wallis, “How do we manage overdiagnosis/overtreatment in breast screening?,” *Clinical radiology*, vol. 73, no. 4, pp. 372–380, 2018.
- [383] J. Yanase and E. Triantaphyllou, “The seven key challenges for the future of computer-aided diagnosis in medicine,” *International journal of medical informatics*, vol. 129, pp. 413–422, 2019.
- [384] B. Evans and P. Ossorio, “The Challenge of Regulating Clinical Decision Support Software After 21st Century Cures,” *American journal of law & medicine*, vol. 44, no. 2-3, pp. 237–251, 2018.
- [385] E. E. Bron, M. Smits, W. M. Van Der Flier, H. Vrenken, F. Barkhof, P. Scheltens, J. M. Papma, R. M. Steketee, C. M. Orellana, R. Meijboom, *et al.*, “Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge,” *NeuroImage*, vol. 111, pp. 562–579, 2015.
- [386] A. Cahan and J. J. Cimino, “A Learning Health Care System Using Computer-Aided Diagnosis,” *Journal of medical Internet research*, vol. 19, no. 3, p. e54, 2017.
- [387] D. Regge and S. Halligan, “CAD: How it works, how to use it, performance,” *European Journal of Radiology*, vol. 82, no. 8, pp. 1171–1176, 2013.
- [388] C. D. Lehman, R. D. Wellman, D. S. Buist, K. Kerlikowske, A. N. Tosteson, D. L. Miglioretti, B. C. S. Consortium, *et al.*, “Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection,” *JAMA internal medicine*, vol. 175, no. 11, pp. 1828–1837, 2015.
- [389] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, “What do we need to build explainable AI systems for the medical domain?,” *arXiv preprint arXiv:1712.09923*, 2017.

Muhammad Naseer Bajwa

naseer.bajwa@dfki.de 
linkedin.com/in/naseeralibajwa 

WORK EXPERIENCE

RESEARCH ASSISTANT, DFKI GmbH, Germany

Dec 2017 - Aug 2022



My primary responsibilities in this position were to conduct applied research towards development of practically usable Computer-Aided Diagnosis systems. Additionally, I helped in supervising masters and undergraduate theses and writing research proposals.

LAB ENGINEER, COMSATS University Islamabad, Pakistan

Aug 2009 - Aug 2012



As lab engineer I designed and conducted lab experiments in various core Electrical and Computer Engineering courses. I also contributed in MultiRate Communication Networks (MRCN) research group on adaptive impulsive noise cancellation in Power Line Communications.

EDUCATION

DOCTOR OF ENGINEERING IN DEEP LEARNING

Apr 2017 - Aug 2022



Technische Universität Kaiserslautern, Germany

My PhD research was towards development of a Pragmatic, Accuracy, Confident and Explainable Computer-Aided Diagnosis (CAD). The thesis identified various limitations in successful deployment of existing CAD systems and proposed multiple solutions to overcome those limitations.

magna cum laude (0.6)

MS COMPUTER ENGINEERING

Sep 2012 - Jan 2015



King Fahd University of Petroleum and Minerals, Saudi Arabia

In my masters thesis, I presented empirical proof of concept that Bluetooth and Industrial WiFi can serve as suitable wireless solutions in Reconfigurable Production Lines using QoS-aware DDS-based middleware.

CGPA: 3.29/4

BS COMPUTER ENGINEERING

Feb 2005 - Jan 2009



COMSATS University Islamabad, Pakistan

My bachelors thesis was about effectively detecting, analyzing, and neutralizing acoustic noise from environment in the real-time. The system was modelled in MATLAB Simulink and implemented on C6713 DSP Starter Kit.

CGPA: 3.39/4

PUBLICATIONS

BOOK CHAPTER



Muhammad Naseer Bajwa*, Adriano Lucieri*, Andreas Dengel and Sheraz Ahmed. Erklärbare KI in der medizinischen Diagnose – Erfolge und Herausforderungen. in *Künstliche Intelligenz im Gesundheitswesen - Springer Nature 2022*

PEER REVIEWED JOURNAL ARTICLES



Adriano Lucieri, **Muhammad Naseer Bajwa**, Stephan Alexander Braun, Muhammad Imran Malik, Andreas Dengel, Sheraz Ahmed. "ExAID: A Multimodal Explanation Framework for Computer-Aided Diagnosis of Skin Lesions", *Computer Methods and Programs in Biomedicine 2022*



Muhammad Naseer Bajwa, Suleman Khurram, Mohsin Munir, Shoaib Ahmed Siddiqui, Muhammad Imran Malik, Andreas Dengel, Sheraz Ahmed. "Confident Classification using a Hybrid between Deterministic and Probabilistic Convolutional Neural Networks", *IEEE Access* 2020



Muhammad Naseer Bajwa, Kaoru Muta, Muhammad Imran Malik, Shoaib Ahmed Siddiqui, Stephan Alexander Braun, Bernhard Homey, Andreas Dengel, Sheraz Ahmed. "Computer-Aided Diagnosis of Skin Diseases using Deep Neural Networks", *MDPI Applied Sciences* 2020



Muhammad Naseer Bajwa, Muhammad Imran Malik, Shoaib Ahmed Siddiqui, Andreas Dengel, Faisal Shafait, Wolfgang Neumeier, Sheraz Ahmed. "Two-Stage framework for optic disc localization and glaucoma classification in retinal fundus images using deep learning", *BMC Medical Informatics and Decision Making* 2019



Basem Almadani, **Muhammad Naseer Bajwa**, Shuang-Hua Yang, Abdul Wahid Al-Saif. "Performance evaluation of DDS-based middleware over wireless channel for reconfigurable manufacturing systems", *International Journal of Distributed Sensor Networks* 2015



Basem Almadani, Shehryar Khan, **Muhammad Naseer Bajwa**, Tarek R. Sheltami, Elhadi Shakshuki (2015). "AVL and Monitoring for Massive Traffic Control System over DDS", *Mobile Information Systems* 2015

CONFERENCE PROCEEDINGS



Muhammad Naseer Bajwa, Gur Amrit Pal Singh, Wolfgang Neumeier, Muhammad Imran Malik, Andreas Dengel and Sheraz Ahmed. "G1020: A Benchmark Retinal Fundus Image Dataset for Computer-Aided Glaucoma Detection", *International Joint Conference on Neural Networks, (IJCNN-2020)*



Adriano Lucieri, **Muhammad Naseer Bajwa**, Stephan Alexander Braun, Muhammad Imran Malik, Andreas Dengel and Sheraz Ahmed. "On Interpretability of Deep Learning based Skin Lesion Classifiers using Concept Activation Vectors", *International Joint Conference on Neural Networks, (IJCNN-2020)*



Adriano Lucieri, **Muhammad Naseer Bajwa**, Andreas Dengel and Sheraz Ahmed. "Explaining AI-based Decision Support Systems using Concept Localization Maps", *27th International Conference on Neural Information Processing, (ICONIP-2020)*



Muhammad Naseer Bajwa, Yoshibonum Taniguchi, Muhammad Imran Malik, Wolfgang Neumeier, Andreas Dengel, and Sheraz Ahmed. "Combining Fine- and Coarse-Grained Classifiers for Diabetic Retinopathy Detection", *23rd International Conference on Medical Image Understanding and Analysis, (MIUA-2019)*

ACHIEVEMENTS AND AWARDS



PHD SCHOLARSHIP

National University of Science and Technology (NUST), Pakistan

2017 - 2021



MS SCHOLARSHIP

King Fahd University of Petroleum and Minerals (KFUPM), Saudi Arabia

2012 - 2014



BS SCHOLARSHIP

COMSATS Institute of Information Technology (CIIT), Pakistan

2005 - 2009



SILVER MEDAL

COMSATS Institute of Information Technology (CIIT), Pakistan

March 2009

SKILLS & INTERESTS

IT SKILLS



Python 2.7/3.6



Tensorflow and Keras



Pytorch



C/C++



MATLAB and Simulink



LaTeX



Ubuntu



Windows

LANGUAGES



Punjabi Native Speaker



Urdu Native Speaker



English Fluent Speaker

INTERESTS



Literature and History



Music and Movies



Chess and Cricket