

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Classification of Keratitis from Eye Corneal Photographs using Deep Learning: A Comparative Study on Amoeba, Bacteria, and Fungi

Maria Miguel Santos Beirão



Master in Bioengineering - Biomedical Engineering

Supervisor: Jaime dos Santos Cardoso, PhD

Co-supervisor: João Carlos Ramos Gonçalves Matos, MSc

Co-supervisor: Tiago Filipe Sousa Gonçalves, MSc

July 24, 2024

Classification of Keratitis from Eye Corneal Photographs using Deep Learning: A Comparative Study on Amoeba, Bacteria, and Fungi

Maria Miguel Santos Beirão

Master in Bioengineering - Biomedical Engineering

July 24, 2024

Resumo

A Ceratite é uma doença inflamatória da córnea responsável por 10% das deficiências visuais nos países de baixo e médio rendimento (PBMR); bactérias, fungos ou amebas são etiologias mais comuns. Um diagnóstico preciso e atempado do agente infeccioso é crucial não só para a opção de tratamento, mas também para garantir que a visão não é afetada. Devido ao custo elevado e à disponibilidade limitada de diagnósticos laboratoriais nos PBMR, o diagnóstico é frequentemente efectuado apenas pela observação clínica, apesar da sua menor precisão. O objetivo deste estudo foi desenvolver um modelo de deep learning que auxiliasse no diagnóstico da fonte de infeção.

Um dataset brasileiro de córnea foi utilizado, constituído por 24.692 observações, das quais 2.064 tiveram resultados positivos para infeção (56.98% bactérias, 13.42% fungos, 10.03% para ameba, 10.32% para bactérias e fungos e, por fim, 9.26% para ameba e bactérias).

Após o pré-processamento do dataset, foram implementadas várias metodologias. Foi avaliada a capacidade de previsão de características identidade com base nas imagens, a fim de compreender a quantidade de informação biométrica que o olho infetado ainda possuía. Para tal, foi implementada uma metodologia 10-fold com uma DenseNet121, pré-treinada em ImageNet responsável por uma decisão binária relativamente ao sexo. O desempenho foi avaliado usando várias métricas, matrizes de confusão e mapas de saliência. Foi alcançada uma AUROC de 0,8790-0,8994 no conjunto de teste. O mesmo método foi repetido para classificar dentro dos grupos etários, tendo sido previstas quatro classes (0-18, 18-40, 40-65 e mais de 65 anos). A AUROC foi de 0.8331-0.8624.

Relativamente à previsão de doenças, foram implementadas e comparadas três abordagens diferentes. A primeira abordagem consistia em três arquiteturas DenseNet (pré-treinadas no ImageNet), responsáveis por uma decisão binária de cada tipo de infeção. A segunda abordagem utilizou o mesmo backbone para as três tarefas, com uma camada de classificação para cada infeção, o que a torna a abordagem multitask V1. A terceira abordagem, multitask V2, consistia numa arquitetura DenseNet semelhante utilizada como backbone, com uma camada de classificação multitarefa com um neurónio para cada tarefa. A abordagem multitask produziu resultados interessantes, pelo que foram implementadas e devidamente avaliadas uma função de loss clínica e uma metodologia de threshold adaptativo, utilizando a estatística de Youden. O Multitask V2 com loss clínica foi considerado o melhor modelo, obtendo uma AUROC de 0.7413-0.7740 para Bactérias, 0.8395-0.8725 para Fungos e 0.9448-0.9616 para Amebas.

Finalmente, para avaliar a influência das características de identidade neste diagnóstico, foi efectuada uma análise estatística utilizando o teste t, para comparação de sexo, e ANOVA, para comparação de idade. Os resultados revelaram que o sexo afecta significativamente a previsão da infeção por amebas e a idade afecta significativamente a previsão da infeção por fungos e bactérias. No entanto, uma análise futura com um conjunto de dados dados equilibrado em termos de idade poderá ser necessário. Em conclusão, foi possível prever corretamente a infeção por BK, FK e AK com uma fotografia da córnea do olho. Para além disso, sexo e idade provaram ser características identificáveis a partir destas imagens, razão pela qual estratégias como o disentanglement podem ser benéficas para garantir que apenas as características médicas são utilizadas para o diagnóstico.

Abstract

Keratitis is an inflammatory corneal condition responsible for 10% of visual impairment in low- and middle-income countries (LMICs); bacteria, fungi, or amoeba are the most common infection etiology. An accurate and timely diagnosis of the infection agent is crucial not only for the treatment option but also for the patient's sight outcomes. Due to the high cost and limited availability of laboratory diagnostics in LMICs, diagnosis is often made by clinical observation alone, despite its lower accuracy. The goal of this study was to develop a deep learning framework model that could assist in the diagnosis of the source of infection.

A Brazilian cornea dataset was utilized, comprising 24,692 observations, of which 2,064 had positive results for infection (56.98% bacteria, 13.42% fungi, 10.03% for amoeba, 10.32% for bacteria and fungi and lastly, 9.26% for amoeba and bacteria).

After preprocessing the dataset and preparing the images, different frameworks were implemented. Firstly, the prediction of identity features was assessed to understand how much biometric information the infected eye still had. To do so, a 10-fold methodology with a DenseNet121, pre-trained on ImageNet responsible for a binary decision regarding sex was implemented. The performance was evaluated using several metrics, confusion matrices and saliency maps. An AUROC of 0.8790-0.8994 was achieved on the test set. The same method was then repeated to classify within age groups, with four classes being predicted (0-18 years old, 18-40 years old, 40-65 years old and more than 65). An AUROC of 0.8331-0.8624 was achieved on the test set.

Regarding disease prediction, three different approaches were implemented and compared. The first approach consisted of three DenseNet architectures (pre-trained on ImageNet), responsible for a binary decision of each infection type. The second approach used a shared backbone for the three tasks but three parallel classification layers, each responsible for a binary decision for each infection, making it the V1 multitask approach. The third approach, multitask V2, consisted of a similar DenseNet architecture used as backbone, with a multi-head classification layer for multitask learning. Since multitask approach had interesting results, improvements were made to that strategy. A Clinical Loss Function and an Adaptive Threshold Methodology (Youden's Statistic) were tested and assessed. The results were compared and Multitask V2 with clinical loss was considered the best model, achieving a 95% confidence interval on the test set for AUROC, across 10 folds, of 0.7413-0.7740 for Bacteria, 0.8395-0.8725 for Fungi and 0.9448-0.9616 for Amoeba.

Finally, to assess the influence of identity features in this diagnosis, a statistic analysis was performed using T-test, for sex comparison, and ANOVA, for age comparison. The results found out that sex significantly affects the amoeba infection prediction and age significantly affects the fungi and bacteria infection prediction; nevertheless, an analysis with a balanced dataset age-wise should be implemented to have stronger conclusions about the age influence.

In conclusion, it was possible to correctly predict for BK,FK and AK when given an eye corneal photograph. Furthermore, sex and age proved to be characteristics that are identifiable from these photographs, which is why strategies like disentanglement could be beneficial to ensure that only medical features are being used for the diagnosis.

Acknowledgements

Em primeiro lugar, agradeço ao meu orientador Professor Jaime Cardoso e co-orientadores João Matos e Tiago Gonçalves, por todos os conselhos, ideias e feedback. Em especial ao João e ao Tiago, obrigada por estarem sempre à distância de uma mensagem de WhatsApp, independentemente do fuso-horário em que se encontravam, fosse para me fazerem pensar, encontrar um bug ou apenas validar as minhas ideias com o famoso “Sounds Good”. Em segundo lugar, agradeço ao Dr Luis e à Dr Camila pela disponibilidade para responderem a todas as minhas perguntas clínicas. Lastly, I would like to extend my gratitude to Leo A. Celi and the entire group at the Laboratory for Computational Physiology at MIT for warmly welcoming me during my stay in Boston.

Às caras que foram passando na B317 enquanto eu me mantinha no mesmo lugar, obrigada por fazerem as minhas épocas de exames mais toleráveis e por terem sempre um abraço, uma fofoca, ou uma sessão de música alto para alegrar os meus dias mais negros.

À minha ShARE, que não tendo um escritório físico, foi a minha maior casa e o sítio onde mais cresci. No meio de feedbacks, reuniões em demasia e desafios que pareciam não acabar, nunca pensei que lá fosse encontrar quem encontrarei. Seja para trabalho, um bom jantar ou apenas para um abraço apertado, encontrei aqui as pessoas com o coração mais verdinho de que há memória. Tenho a certeza que daqui a 30 anos vamos continuar a encontrar tempo para comer sushi na nossa garagem de eleição. Aos meus ShARitos, obrigada por acreditarem mais em mim do que eu alguma vez acreditarei.

Aos que ao meu lado sobrevivem a este curso, obrigada. Foram muitos projetos, muitos exames e, infelizmente uma dose inesperada de máscaras. No entanto foram também muitas festas, muitas praxes, muitas frases ridículas que não deviam ter sido ditas, mas, acima de tudo, muitas memórias que levo no coração. Onde quer que a vida vos leve, terão sempre em mim uma amiga. Deixo o meu especial agradecimento à minha corretora gramatical e à minha microondas por terem sempre um conselho sábio, uma sessão de gossip ou uma mão para me estender.

À minha família, que, entre nadadores olímpicos, motores de busca de renome, motoristas privados, estilistas de topo e cozinheiros de sopa, me ensinou, ao longo dos anos, que só não é possível se não tentarmos. Um obrigada pela paciência para todas as manhãs, pelos jantares apressados antes de reuniões, pelo abraço sempre que falhei, mas acima de tudo por me ensinarem que a vida é mais bonita se formos nós próprios e nos rirmos disso. Aos cães com mais personalidade do mundo, Scooby e Doodles, obrigada por contribuírem para o caos que é esta casa.

Por último, à Maria do passado, por nunca ter dito que não a nada. Obrigada por, dos três círculos que compõem a vida de um estudante (festas, curso e dormir), teres sempre escolhido prejudicar o último.

A ti, Suzy, sempre capaz de diagnosticar com métricas excelentes o que se passava pelos meus olhos, dedico este trabalho

Maria Beirão

*"Our greatest accomplishments cannot be behind us,
because our destiny lies above us"*

Cooper, Interstellar

Contents

1	Introduction	1
1.1	Context	1
1.1.1	Keratitis	1
1.1.2	Deep Learning	1
1.2	Objectives	2
1.3	Alignment with Sustainable Development Goals (SDG)	3
1.4	Main Contributions	3
1.5	Partnerships sustained by the project	4
1.6	Structure of the Dissertation	4
2	In the Blink of an Eye: Understanding Keratitis	6
2.1	Ophthalmology	6
2.1.1	The eye	7
2.1.2	Areas and types of ophthalmology	7
2.2	Keratitis	7
2.2.1	Symptoms	9
2.2.2	Cause	9
2.2.3	Risk Factors	11
2.2.4	Diagnosis	11
2.2.5	Treatment	11
2.3	Keratitis Epidemiology Analysis	12
2.3.1	Incidence	12
2.3.2	Age	12
2.3.3	Sex	13
2.3.4	Seasonal Impact	13
2.4	Conclusions	14
2.5	Summary	14
3	Eyes on the Prize: Artificial Intelligence in Ophthalmology	15
3.1	Artificial Intelligence in Healthcare	15
3.1.1	Benefits	15
3.1.2	Challenges	16
3.1.3	Market Progression	18
3.1.4	Relevant Types of AI	18
3.2	AI in ophthalmic pathology and ophthalmology	20
3.2.1	FDA approved products for ophthalmology	21
3.3	Current Solutions for Keratitis	22
3.3.1	Bacterial Keratitis Diagnosis	22

3.3.2	Fungal Keratitis Diagnosis	23
3.3.3	Multitask Diagnosis	24
3.3.4	Datasets	27
3.3.5	Ophthalmologists vs. AI	29
3.4	Computer Vision	31
3.4.1	Models Architecture	31
3.4.2	Additional Layers	34
3.4.3	Learning Process	36
3.4.4	Performance Metrics	37
3.5	Conclusions	41
3.6	Summary	41
4	Insightful Vision: Unravelling the Dataset	43
4.1	Dataset Characterisation	43
4.2	Dataset Features	43
4.2.1	General Analysis	43
4.2.2	Feature Analysis	46
4.2.3	Extracted Features	50
4.3	Correlations	51
4.4	Conclusions	52
5	Clear Vision, Clear Methodology: Proposed Framework	55
5.1	Proposed Workflow	55
5.2	Data Preparation	56
5.3	Pre-processing	56
5.3.1	Dataset and Image Preparation	56
5.3.2	Data Augmentation of Images	57
5.4	Model preparation	60
5.4.1	K-fold strategy	60
5.4.2	Class Weights	61
5.4.3	Optimiser	61
5.4.4	Model Selection	62
5.4.5	Model Architecture	62
5.4.6	Model training and validation	63
5.4.7	Performance evaluation	65
5.5	Statistical Analysis	66
5.5.1	Confidence Intervals	66
5.5.2	P-values	67
5.5.3	Holm–Bonferroni Correction	67
5.6	Contributions	68
5.7	Summary	68
6	See to Believe: Results Exploration	69
6.1	Optimal Architecture and Parameters	69
6.2	Sex Prediction	69
6.3	Age Prediction	71
6.4	Baseline Disease Prediction	72
6.4.1	Single-Task	73
6.4.2	Multitask	78

6.5	Multitask Exploration	83
6.5.1	Multitask with Clinical Loss	83
6.5.2	Multitask with Clinical Loss and Threshold	87
6.5.3	Best Predictive Model	90
6.6	Identity Features' Impact	91
6.6.1	Sex Influence	91
6.6.2	Age Impact	91
7	Looking into the Future	93
7.1	Goals' Satisfaction	93
7.2	Future Work	93
7.3	Final Remarks	94
	References	96
A	Appendix	106
A.1	Presented Poster	106
A.2	Optimal Architecture and Parameters	107
A.2.1	Best Model Analysis	107
A.2.2	Epochs	107
A.2.3	Learning Rate	108
A.3	Loss Functions	109
A.4	Confidence Intervals for the best methodology	113
A.5	P-values for the other methodologies	116

List of Figures

1.1	Deep learning architecture from [1].	2
2.1	Anatomy of the eye from [2].	8
2.2	Healthy vs Infected eye from [3].	9
2.3	Fungi Keratitis Infections across the world from [4].	13
3.1	Progression of AI in healthcare from [5].	18
3.2	Expected growth rate of AI in healthcare in US Billions from [6].	19
3.3	Statistics on the number of AI applications in diagnosing ophthalmologic diseases, by disease (a) and by year (b) from [7].	20
3.4	FDA approved products for healthcare from [8].	21
3.5	DL model vs NCS-Oph (a) and Experts (b) from [9].	30
3.6	Confusion matrix of KeratitisNet (a) and ophthalmologists (b) in distinguishing IK from [10].	30
3.7	VGG16 Architecture diagram from [11].	32
3.8	ResNet50 Architecture diagram from [12].	33
3.9	DenseNet121 Architecture diagram from [13].	34
3.10	Average pool operation with kernel 3×3 from [14].	35
3.11	Dropout layer from [15].	35
3.12	Loss analysis from [16].	36
3.13	Backpropagation from [17].	38
3.14	Confusion Matrix scheme from [18].	39
3.15	ROC Curve from [19].	40
4.1	Bacteria Dataset Example.	45
4.2	Fungi Dataset Example.	45
4.3	Ameba Dataset Example.	45
4.4	Bacteria Dataset Example.	45
4.5	Fungi Dataset Example.	45
4.6	Ameba Dataset Example.	45
4.7	Data Cleaning.	46
4.8	Distribution of the sex parameter in the dataset.	47
4.9	Distribution of age in the dataset.	47
4.10	Distribution of the age bins in the dataset.	48
4.11	Distribution of the eye in the dataset.	48
4.12	Distribution of the ordered tests of Bacteria.	49
4.13	Distribution of the ordered tests of Fungi.	49
4.14	Distribution of the ordered tests of Acanthamoeba.	49
4.15	Distribution of surgery in the dataset.	49

4.16	Distribution of medicine in the dataset.	50
4.17	Distribution of the origin of the material in the dataset.	50
4.18	Distribution of the month of the photograph in the dataset.	51
4.19	Distribution of the quarter of the year in the dataset.	51
4.20	Distribution of the year in the dataset.	52
4.21	Normalised Correlation of the quarter and type of IK.	53
4.22	Normalised Correlation of the year and type of IK.	53
4.23	Correlations between dataset features.	54
5.1	Gaussian Filter variance with sigma from [20].	58
5.2	Examples of the pipeline of image transformations applied to the dataset.	59
5.3	Adapted 10-fold strategy.	60
5.4	Fully Connected Layers of each model.	62
5.5	Loss optimal from [21]	64
5.6	Youden's J Optimisation from [22].	65
6.1	Saliency Map for a Female Patient Correctly Diagnosed.	70
6.2	ROC Curve for one fold of the sex model.	70
6.3	Saliency Map for a Negative patient for 18-40 years old Correctly Diagnosed.	71
6.4	ROC Curve for one fold of the age model.	71
6.5	Saliency Map for a Negative Patient for Bacteria Correctly Diagnosed.	73
6.6	ROC Curve for one fold of baseline bacteria prediction.	73
6.7	Saliency Map for a Positive Patient for Fungi Correctly Diagnosed.	75
6.8	ROC Curve for one fold of baseline fungi prediction.	75
6.9	Saliency Map for a Negative patient for Amoeba Correctly Diagnosed.	76
6.10	ROC Curve for one fold of baseline amoeba prediction.	76
6.11	Saliency Map for a Negative patient for Bacteria Task Correctly Diagnosed with Multitask v1 and baseline loss.	78
6.12	Roc Curve for a fold of the Multitask v1 with baseline loss.	79
6.13	Saliency Map for a Positive patient for Fungi Task Correctly Diagnosed with Multitask v2 and baseline loss.	81
6.14	ROC Curve for one for of Multitask v2 baseline loss.	81
6.15	Saliency Map for a Positive patient for Fungi Task Correctly Diagnosed with Multitask v1 Clinical Loss.	83
6.16	ROC Curve of one fold of Multitask v1 with clinical loss and threshold.	89
6.17	ROC Curve of a random fold of Multitask v2 with clinical loss and threshold.	91
A.1	Poster Presented in IJUP and AI Cures Conference.	106
A.2	Effect of epochs on Loss Curve of the sex prediction.	108
A.3	Effect of epochs on the Multitask Loss Curve.	108
A.4	Loss evolution over epochs for sex model.	109
A.5	Loss evolution over epochs for age model.	109
A.6	Loss evolution over epochs for single task bacteria model.	110
A.7	Loss evolution over epochs for single task fungi model.	110
A.8	Loss evolution over epochs for single task ameba model.	110
A.9	Loss evolution over epochs for Multitask v1 with baseline loss.	111
A.10	Loss evolution over epochs for Multitask v2 with baseline loss.	111

A.11 Loss evolution over epochs for Multitask v1 with clinical loss (represents the loss of both results exploration cases, since the adaptive threshold method uses the same loss function).	111
A.12 Loss evolution over epochs for Multitask v2 with clinical loss (represents the loss of both results exploration cases, since the adaptive threshold method uses the same loss function).	112

List of Tables

2.1	Medicines per type of Keratitis.	12
2.2	Incidence of IK in Different Countries.	12
3.1	State of the Art of Bacterial Keratitis Diagnosis: Summary of Authors, Year of Publication, Country of the Dataset samples, Dataset Composition, Predictive Task and Performance metrics.	23
3.2	State of the Art of Fungal Keratitis Diagnosis: Summary of Authors, Year of Publication, Country of the Dataset samples, Dataset Composition, Predictive Task and Performance metrics.	24
3.3	State of the Art for multi-infection BK vs FK: Summary of Authors, Year of Publication, Country of the Dataset samples, Dataset Composition, Model Architecture, Predictive Task and Performance metrics	26
3.4	State of the Art for multi- infection: Summary of Authors, Year of Publication, Country of the Dataset samples, Dataset Composition, Predictive Task and Performance metrics.	28
4.1	Attributes and Features available in the Dataset.	44
4.2	Number of Cases and Percentages for Combinations of Infections.	46
5.1	Transformations into numerical labels.	57
5.2	The inputs of the different models.	57
6.1	Sex Prediction Confidence Intervals.	70
6.2	Average Confusion Matrix for Sex.	70
6.3	Age Prediction Confidence Intervals.	72
6.4	Average Confusion Matrix for Age.	72
6.5	Confidence Intervals for General Subset of Bacteria Baseline Model.	74
6.6	Average Confusion Matrix for Bacteria Baseline.	74
6.7	Confidence Intervals for General Subset of Fungi Baseline Model.	76
6.8	Average Confusion Matrix for Fungi Baseline.	76
6.9	Confidence Intervals for General Subset of Amoeba Baseline Model.	77
6.10	Average Confusion Matrix for Amoeba Baseline.	77
6.11	Confidence Intervals for General Subset, Tasks, and Metrics of Multitask v1 with baseline loss.	80
6.12	Average Confusion Matrices for Different Tasks with Multitask v1 baseline loss.	80
6.13	Confusion Matrix for BK-FK-AK Infections of Multitask v1 with baseline loss.	80
6.14	Confidence Intervals for General Subset, Tasks, and Metrics of Multitask v2 with baseline loss.	82
6.15	Average Confusion Matrices for Different Tasks with Multitask v2 baseline loss.	82

6.16	Confusion Matrix for BK-FK-AK Infections- Multitask v2 with baseline loss. . .	82
6.17	Confidence Intervals for General Subset, all Tasks, and Metrics on Multitask V1 with clinical loss.	84
6.18	Average Confusion Matrices for Different Tasks with Multitask v1 Clinical loss. .	84
6.19	Confusion Matrix for BK-FK-AK Infections- Multitask v1 with Clinical loss. . .	85
6.20	Confidence Intervals for General Subset, Tasks, and Metrics of Multitask v2 with Clinical Loss.	86
6.21	Average Confusion Matrices for Different Tasks with Multitask v2 Clinical loss. .	86
6.22	Confusion Matrix for BK-FK-AK Infections- Multitask v2 with Clinical loss. . .	86
6.23	Confidence Intervals for General Subset, all Tasks, and Metrics of V1 with Clinical loss and threshold.	88
6.24	Average Confusion Matrices for Different Tasks with Multitask v1 Clinical loss and Threshold.	88
6.25	Confusion Matrix for BK-FK-AK Infections - Multitask v1 with Clinical loss and threshold.	88
6.26	Confidence Intervals for Selected Subsets, Tasks, and Metrics on Multitask with Optimised Loss and Threshold v2.	90
6.27	Average Confusion Matrices for Different Tasks with Multitask v2 Clinical loss and Threshold.	90
6.28	Confusion Matrix for BK-FK-AK Infections - Multitask v2 with Clinical loss and threshold.	90
6.29	Corrected p-values for Sex.	92
6.30	Corrected p-values for Age.	92
A.1	Assessment of the best baseline models.	107
A.2	Learning Rate and weight decay impact on minimum validation loss.	109
A.3	Confidence Intervals for Sex Subsets, Tasks, and Metrics of Multitask v2 with clinical loss	113
A.4	Confidence Intervals for Age Subsets, Tasks, and Metrics of Multitask v2 with clinical loss.	114
A.5	P-values for Task Bacteria by Sex.	116
A.6	P-values for Task Bacteria by Age.	116
A.7	P-values for Task Fungi by Sex.	116
A.8	P-values for Task Fungi by Age.	116
A.9	P-values for Task Ameba by Sex.	117
A.10	P-values for Task Ameba by Age.	117
A.11	P-values for Tasks by Sex for Multitask V1.	117
A.12	P-values for Tasks by Age for Multitask V1.	117
A.13	P-values for Tasks by Sex for Multitask V2.	117
A.14	P-values for Tasks by Age for Multitask V2.	118
A.15	P-values for Tasks by Sex for Multitask V1 with clinical loss.	118
A.16	P-values for Tasks by Age for Multitask V1 with clinical loss.	118
A.17	P-values for Tasks by Sex for Multitask V1 with clinical loss and adaptive threshold.	118
A.18	P-values for Tasks by Age for Multitask V1 with clinical loss and adaptive threshold.	118
A.19	P-values for Tasks by Sex for Multitask V2 with clinical loss and adaptive threshold.	119
A.20	P-values for Tasks by Age for Multitask V2 with clinical loss and adaptive threshold.	119

Abbreviations and Symbols

AI	Artificial Intelligence
ACC	Accuracy
AK	Amoeba Keratitis
AMD	Age-related macular degeneration
ANOVA	Analysis of Variance
AUROC	Area under the receiver operating characteristic
BA	Balanced Accuracy
BK	Bacterial Keratitis
CAD	Computer-Aided Diagnosis
CAGR	Compound annual growth rate
CL	Contact Lenses
CNN	Convolutional Neural Network
DL	Deep Learning
DR	Diabetic Retinopathy
FC	Fully Connected
FDA	Food and Drug Administration
FK	Fungal Keratitis
GANs	Generative Adversarial Networks
HSK	Herpes Simplex Keratitis
HZK	Herpes Zoster Keratitis
IK	Infectious Keratitis
IOLs	Intraocular lenses
KDE	Kernel Density Estimator
LASIK	Laser-assisted in situ keratomileusis
LMICs	Low- or Middle-Income Countries
MAE	Mean Absolute Error
MIT	Massachusetts Institute of Technology
ML	Machine Learning
OCT	Optical Coherence Tomography
ROC	Receiver operating characteristic
RPA	Robotic process automation
SDG	Sustainable Development Goals
SGD	Stochastic Gradient Descent
UNIFESP	Universidade Federal de São Paulo
UV	Ultraviolet
VK	Virus Keratitis

Chapter 1

Introduction

1.1 Context

1.1.1 Keratitis

Infectious Keratitis (IK) can be caused by bacterial, fungal, amoeba, or other microorganisms and is an inflammation of the cornea [23]. As it is strongly associated with agricultural work and poor hygiene, its prevalence in low- and middle-income countries (LMICs) is very high when compared to high-income countries [24]. It is often diagnosed using a combination of clinical examination, patient history, and time-consuming laboratory testing such as corneal scraping and culture, which require specialised knowledge. In addition, these exams are very costly, which often causes the health practitioner to not test for all etiologies, making the clinician miss the real diagnoses [25]. If left untreated, it can cause serious problems such as corneal scarring, vision loss, and even blindness. Rapid and correct diagnosis is crucial for initiating appropriate therapy and avoiding lasting harm, but the similarity of symptoms between the etiologies makes the task challenging for the human eye.

1.1.2 Deep Learning

Deep learning is a subset of machine learning, which in turn is a subset of artificial intelligence (AI). It uses artificial neural networks to model and solve complex problems by learning from large amounts of data. These neural networks, inspired by the human brain's structure, consist of layers of interconnected nodes (neurons) that process input data and learn to perform tasks by adjusting the weights of the connections based on the errors in their predictions (see Figure 1.1) [26].

One of the most important characteristics of deep learning is its capacity to autonomously learn feature representations from raw data. This contrasts with typical machine learning approaches, which frequently require a feature extraction process. Deep learning (DL) models, such as convolutional neural networks (CNNs), have proven particularly effective in areas such as computer vision, natural language processing, and speech recognition.

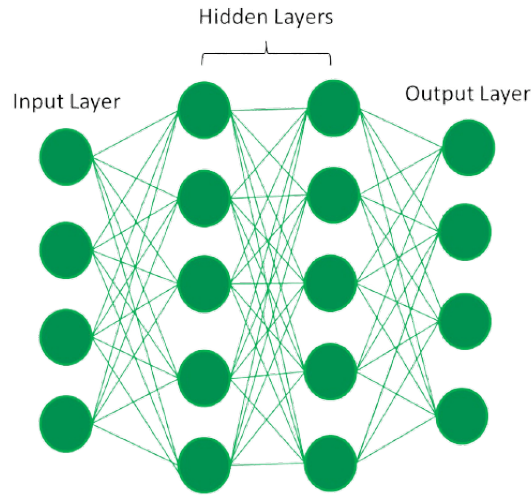


Figure 1.1: Deep learning architecture from [1].

Deep learning has driven significant advancements in technology, leading to innovations such as autonomous vehicles, virtual assistants, and advanced recommendation systems. However, it also poses challenges, including the need for large amounts of labelled data, high computational resources, and the difficulty in interpreting complex models, which are particularly challenging in the healthcare sector [1, 26, 27].

1.2 Objectives

In this dissertation, the primary aim is to develop advanced computer vision and machine learning techniques to classify the etiologies. The significance of this goal is highlighted by the impact of eye corneal diseases, with anterior segment diseases being the leading cause of monocular blindness worldwide, particularly in marginalised populations in developing countries. Corneal diseases like keratitis are responsible for 10% of visual impairment in underdeveloped regions, making accurate and timely diagnosis crucial for treatment and visual prognosis.

However, diagnosing keratitis is challenging due to reliance on clinical examination in LMICs, where access to laboratory diagnostics is limited. The growing ageing population, especially in these regions, underscores the need for efficient diagnostic methods. AI-assisted diagnosis presents a timely opportunity to address this gap, offering potential improvements in accuracy and accessibility.

The scope of this work includes three key tasks:

1. Data engineering and preparation of a new dataset, along with an in-depth analysis to evaluate its quality and relevance.
2. Implementation of computer vision and machine learning algorithms to accurately classify corneal eye photographs with respect to different types of keratitis.

3. Assessing the statistical impact of correlated features, such as sex and age, on disease presentation.

By addressing these tasks, this dissertation aims to significantly improve the diagnosis and management of corneal diseases, particularly in underserved regions.

1.3 Alignment with Sustainable Development Goals (SDG)

The core of this project directly addresses the Sustainable Development Goals of ensuring: "**SDG 3 - Good Health and Well-being**" and "**SDG 10 - Reduced Inequalities**" [28].

Keratitis, frequently linked to agricultural labour and inadequate hygiene practices, is notably prevalent in LMICs. In these settings, where access to specialised healthcare services can be limited, individuals engaged in agricultural activities are at heightened risk due to prolonged exposure to environmental factors and potential contaminants. Poor sanitation and hygiene further compound the risk, exacerbating the incidence of this condition.

By targeting the diagnosis of keratitis in such environments, the purpose of the project aligns with "SDG 3: Good Health and Well-being". Its provision of a cost-effective and accurate solution not only enables early intervention to prevent vision impairment and blindness but also addresses a specific health challenge prevalent in LMICs. Moreover, by mitigating the burden of untreated eye conditions stemming from agricultural work and poor hygiene practices, it contributes to the overall improvement of health outcomes in these regions.

Additionally, the focus on "SDG 10: Reduced Inequalities" on healthcare access is particularly pertinent in the context of keratitis. The tool's ability to provide timely and accurate diagnosis, regardless of socioeconomic status or geographic location, ensures that marginalised communities, often disproportionately affected by the condition, are not left behind. By bridging the gap in healthcare provision and promoting equitable access to diagnosis and treatment, it advances the goal of reducing disparities in health outcomes and fosters a more inclusive and sustainable healthcare system in LMICs.

1.4 Main Contributions

This dissertation's main contributions are:

- Characterisation of a new dataset for microorganism Keratitis;
- Exploration of different techniques to deal with data imbalance and small datasets;
- Comparison of Single and Multi-task Models for disease classification;
- Development of two Multitask approaches based on known deep learning architectures;
- Creation of a clinical loss function;

- Evaluation of the impact of patient features in the performance of a disease classification system;

The work developed in the scope of this project was presented in a Poster Session for "AI Cures Conference" [29], organised by Massachusetts Institute of Technology (MIT) and Mass General Brigham(MGB), and held in Boston. The poster was also presented in "Encontro de Investigação Jovem da U.Porto" (IJUP) [30], organised by the University of Porto and held in the Faculty of Economy of the University of Porto. **In IJUP, the poster was awarded a distinction as one of the best two posters in Engineering.** The developed poster can be found in Appendix A, in Figure A.1.

A paper with the contributions of this dissertation is also being prepared for the "International Conference on Bioinformatics and Biomedicine" (IEEE BIBM 2024) for the August deadline [31].

1.5 Partnerships sustained by the project

This project was initiated through a collaboration formed at MIT's Laboratory for Computational Physiology, where my co-supervisor, João Matos, met Dr. Luis Nakayama.

Dr. Nakayama, from Universidade Federal de São Paulo (UNIFESP), partnered with Dr. Camila Kase, also from UNIFESP, to collect an eye photograph dataset for Keratitis. This strategic partnership between MIT, UNIFESP and FEUP, leveraging the expertise of these distinguished researchers, has been crucial in advancing our research objectives and ensuring the project's success.

Due to the partnership established at MIT's Laboratory for Computational Physiology, I was welcomed to develop my work at the MIT lab for one month. This opportunity allowed me to engage with renowned researchers and facilitated my participation in the AI Cures Conference.

1.6 Structure of the Dissertation

This document is comprised of seven chapters, structured to build a comprehensive narrative. The current Chapter 1 serves as the introduction. Chapter 2 provides an in-depth exploration of ophthalmology, covering market dynamics, eye anatomy, the disease characterisation and its epidemiology. Chapter 3 shifts focus to the integration of artificial intelligence in ophthalmology, detailing FDA-approved products and innovative solutions addressing critical challenges in the specific scope of this disease, as well as an intensive study of the clinical performance in similar tasks. In Chapter 4, we introduce the new dataset that will underpin our research. Chapter 5 delineates the methodology employed in our deep learning-based classification framework. This is followed by Chapter 6, which presents the experimental work and a thorough analysis of the results. Finally, Chapter 7 summarizes the findings, offering conclusions and perspectives on future work. The document concludes with a set of appendices that provide an introduction to key methodologies used throughout the research and supplementary materials pertinent to the main chapters.

For the convenience of readers, a summary is included at the end of Chapters 2, 3, and 5.

Chapter 2

In the Blink of an Eye: Understanding Keratitis

2.1 Ophthalmology

Ophthalmology is a specialised field of medicine that focuses on the health of the eye and the visual system. It encompasses the diagnosis, treatment, and prevention of eye disorders and diseases, integrating both medical and surgical care [32]. The field is crucial for maintaining vision and overall eye health, which significantly impacts quality of life. Ophthalmology combines various scientific disciplines, including anatomy, physiology, and optics. The eye is, obviously, the central organ of this field of study, which functions as the body's camera, capturing and processing visual information.

Advancements in ophthalmology have led to significant improvements in diagnostic and therapeutic techniques. Innovations such as optical coherence tomography (OCT), laser-assisted in situ keratomileusis (LASIK), and intraocular lenses (IOLs) have revolutionised eye care. Moreover, the integration of artificial intelligence and telemedicine is enhancing diagnostic accuracy and expanding access to eye care, particularly in underserved regions [33]. Nevertheless, AI in this field often raises data protection problems due to all the biometric info that can be found in the iris and cornea [34].

Despite these advancements, ophthalmology faces challenges, particularly in ensuring equitable access to care. Disparities in eye health persist, especially in low- and middle-income countries (LMICs) where resources and trained professionals are limited, and contaminations are frequent. Addressing these disparities requires a multifaceted approach, including improving health-care infrastructure, increasing the availability of trained eye care professionals, and leveraging technology to reach remote populations [35].

The market for ophthalmology is experiencing substantial growth due to an ageing global population and the increasing prevalence of eye conditions such as cataracts, glaucoma, macular degeneration, and diabetic retinopathy. According to recent projections, the ophthalmology market is expected to grow at a compound annual growth rate (CAGR) of 6.4% from 2023 to 2030,

reaching approximately \$84,655.95 million by 2030 [36]. This growth is driven by advancements in technology, an increase in healthcare expenditure, and a rising awareness of the importance of eye health.

2.1.1 The eye

The eye (see Figure 2.1) sits in a socket called the orbit and has six extra-ocular muscles that are attached, responsible for the movement and rotation of the eye [37]. These muscles are attached to the sclera, a white, tough-looking tissue that covers most of the eye. The surface of the eye and inner surface of the eyelid are covered by a transparent membrane called conjunctiva [33].

The eye can be divided into the front (anterior) and the back of the eye (posterior). The anterior part starts with the cornea, a clear dome shape where the light is focused to enter the eye. Behind the cornea lies the anterior chamber, filled with a fluid called aqueous humor, which helps maintain eye pressure and drains excess fluid [38]. After, there are the iris and the pupil. The muscles in the iris contract and dilate to control the amount of light that reaches the back of the eye. Directly after, there is the lens, which is responsible for focusing the light towards the back of the eye. Its shape is mutable to help the eye focus on objects closer to the subject. Fibers called zonules or ciliary bodies are attached to the lens capsule and suspend it from the eye wall [39].

The posterior part is primarily composed of the vitreous humor, which fills the cavity between the lens and the back of the eye. The light that enters through the anterior part passes this gelatinous fluid and reaches the retina, the light-sensitive tissue in the back of the eye [38]. The retina is composed of the macula and the peripheral retina. The macula gives us detailed central vision and the peripheral retina gives us side vision. In the retina, the light is changed into energy by photoreceptors. Rods perceive black and white, and night vision, and cones perceive colour and central detailed vision. These signals are then transmitted through the optic nerve to the brain's visual cortex, enabling us to perceive and interpret visual information [39].

2.1.2 Areas and types of ophthalmology

As in other areas of medicine, there is a general ophthalmologist who is trained in all aspects of the eye, but as in other areas, there are also sub-specialists who have additional training in other areas. The subspecialties are: Cornea, Glaucoma, Retina and Vitreous, Oculoplastic Surgery, Pediatrics and Neurology [40]. Usually, the subspecialties are related to the area of the eye the clinician is focused on.

2.2 Keratitis

Corneal blindness is estimated to impact 6 million of the world population according to World Health Organization (WHO) [41]. Keratitis is the leading cause of cornea blindness in both developed and developing countries, with an estimated incidence ranging from 2.5 (developed) to 799

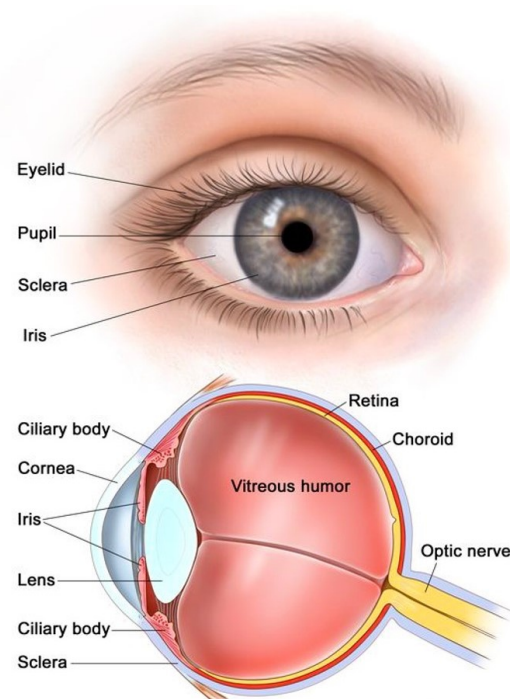


Figure 2.1: Anatomy of the eye from [2].

(developing) per 100,000 [24, 35]. It is responsible for 10% [25] of visual impairment in LMICs, with a prevalence that is thirty times higher in India compared to the USA.

Keratitis (also known as “corneal ulcer”) is an inflammation of the cornea – the clear, dome-shaped window located at the front of the eye that covers the iris and pupil [42]. Infectious keratitis (see Figure 2.2) occurs when microorganisms, including bacteria, viruses, fungi, and amoeba, invade the cornea, leading to infection and inflammation. These microorganisms can enter the cornea through various means, such as trauma, contact lens wear, or exposure to contaminated water or soil. Keratitis poses a significant public health concern, particularly in LMICs where access to healthcare and hygiene practices may be limited. The diagnosis exam associated with this disease is also extremely expensive, reaching a cost of 200\$ for the patient. That is the reason why often the exam is not asked, and the infection is treated as the most common cause, which often leads to misdiagnosis and worsening of the symptoms. The impact of keratitis extends beyond vision loss, affecting individuals’ quality of life, productivity, and economic well-being.

The risk of infectious keratitis is heightened in individuals with compromised immune systems, poor ocular hygiene, or underlying eye conditions. In addition to infectious causes, keratitis can also result from non-infectious factors such as trauma, chemical exposure, or underlying systemic conditions like autoimmune diseases. Regardless of the underlying cause, prompt diagnosis and treatment are essential to prevent complications and preserve vision [23, 43].

Effective management of keratitis requires a multidisciplinary approach involving ophthalmologists, microbiologists, and other healthcare professionals. Treatment typically involves topical or systemic antimicrobial therapy to eradicate the causative microorganism, along with supportive

measures to alleviate symptoms and promote corneal healing. In severe cases or those refractory to medical treatment, surgical intervention such as corneal transplantation may be necessary to restore vision [42].

Prevention strategies for keratitis include practising good ocular hygiene, avoiding contact with contaminated water or soil, proper contact lens care and usage, and seeking prompt medical attention for any signs of eye infection or inflammation. Public health initiatives aimed at improving sanitation, access to clean water, and healthcare infrastructure can also help reduce the burden of keratitis in underserved communities [35, 41].

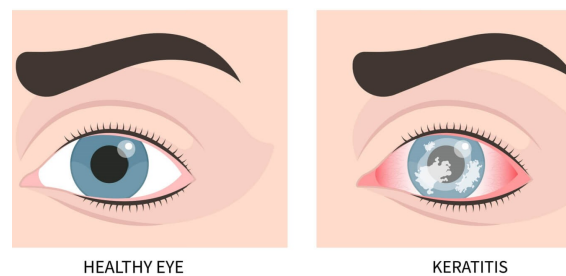


Figure 2.2: Healthy vs Infected eye from [3].

2.2.1 Symptoms

The symptoms of Keratitis (see Figure 2.2) include [23]:

- Eye redness;
- Eye pain;
- Excess tears or discharge from eye;
- Foreign body sensation in the eye;
- Difficulty in opening the eyelid;
- Blurred and decreased vision;
- Light sensitivity (photophobia);
- Watery eyes.

Naturally, these symptoms can evolve and eventually cause permanent vision loss [43].

2.2.2 Cause

Regarding the cause of Keratitis, there are two main groups, infectious or non-infectious Keratitis.

2.2.2.1 Non-Infectious Keratitis

Non-infectious Keratitis can result from several non-infectious causes, including eye injuries from surgery, accidents, or conditions where eyelashes scrape against the eye's surface. Wearing contact lenses for extended periods can also lead to keratitis. Other causes include having a foreign object in the eye, prolonged exposure to ultraviolet light, vitamin A deficiency, eyelid disorders or immune system conditions that cause dry eyes [44].

2.2.2.2 Infectious Keratitis (IK)

Although there are different types of Keratitis, it is possible that non-infectious Keratitis can evolve into Infectious Keratitis since eye injuries can attract microorganisms, mainly bacteria, to the region, causing an evolution to Infectious Keratitis. Although Bacterial Keratitis is the most common one, it can also be caused by Fungi, Amoeba, Viruses, Parasites, or a selection of multiple microorganisms [24, 45].

The most common type of Keratitis in most of the regions is Bacterial Keratitis, and it is often caused by *Pseudomonas aeruginosa* (5-24%), the Coagulase-Negative *Staphylococci* group (CoNS) (24-46%) or *Staphylococcus aureus* (5-36%) [46]. In the UK, 91% to 93% of the Keratitis cases are from Bacterial infection, 86% to 92% in North America, 79% to 88% in South America, 91.8% in Middle East and 93% to 100% in Australia. The prevalence of this type over the others often leads to misdiagnosis since in LMICs, the costs associated with the diagnosis exam can reach 200\$, which often leads the clinician to try to diagnose only based on an eye exam that will be explained later [47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59].

Fungi infection can be of two types: filamentous or yeast/yeast-like fungi. The filamentous fungus, such as *Fusarium spp* and *Aspergillus spp*, are usually associated with tropical climates and are the main causes of IK in India and China, which is not observed in European countries and North America [60]. In those areas of the globe, *Aspergillus spp* reaches 30% incidence and *Fusarium spp* reaches 24% [61]. On the other hand, yeast-like fungi infections are more common in more high-income countries because they are highly related to cornea transplant. The source of infection is often *Candida spp* [62].

Another type is the protozoa infection, which is usually caused by *Acanthamoeba*, a protozoan that is found in water, soil, air and dust. Protozoa infection represents a much smaller percentage of infections (0-5%) [46] and is often associated with contact with contaminated water and soil, and contact lenses (CL). Most of the cases of this infection are in CL wearers.

The last type of IK is virus, mainly the herpes viruses, specifically herpes simplex keratitis (HSK) and herpes zoster keratitis (HZK), but their appearance is often different than the other cases, meaning that it is not very well captured in studies [46].

Focusing now on multiple-sourced infections, they account for 2 to 15% of the total cases of IK [63], which is why clinicians should always maintain a low threshold of repeating corneal scraping if patients are not responding to the treatment.

2.2.3 Risk Factors

The prevalence of Keratitis is increased by the usage of contact lenses [64], with 2 to 4/ 10 000 of the annual cases directly caused by it. Wearing longer than recommended or improper disinfection is the most common cause of this particular type of IK. Extended usage of soft lenses is the most common cause of Contact Lens-associated keratitis.

Eye injuries, including trauma from accidents or surgery, can also lead to keratitis. Exposure to contaminated water, especially while wearing contact lenses, poses a significant risk. As with every infection, corticosteroids, reduced immunity and immune system disorders, such as diabetes or HIV/AIDS, can increase the development of an infection, as well as dry eyes. Additionally, agricultural workers are at higher risk due to potential eye exposure to plant material, soil, and chemicals. Extended exposure to ultraviolet (UV) light without adequate eye protection can also contribute to the development of keratitis. Maintaining proper eye hygiene, using protective eye-wear, and adhering to safety practices can help reduce these risks [44].

2.2.4 Diagnosis

There are four main techniques to diagnose keratitis. The first is a basic eye exam where the doctor visually inspects the eye. The second is a penlight exam, which involves examining the cornea's surface with a light. Both methods are not highly accurate, as distinguishing between bacterial, protozoal, or fungal infections is challenging. The third technique involves corneal scraping with Gram stain, microscopy, and culture analysis. This method is common in developed countries due to its high accuracy, but it is costly, approximately \$200 per patient. In underdeveloped countries, resources for such expensive diagnostic techniques are often unavailable. Consequently, a clinical examination with a slit lamp is used, providing a bright light source and magnification to observe the infection more precisely. However, distinguishing the type of infection remains difficult. Due to the higher prevalence of bacterial infections, doctors often treat keratitis as bacterial, which can worsen if the infection is of a different type [25, 65, 66].

2.2.5 Treatment

Regarding the treatment, after the test is performed and the diagnosis is made, the doctor prescribes the correct medicine according to the source of the infection. Not only the medicine is different, but also the duration of the treatment and the amount of medicine bottles used per month [67]. Table 2.1 lists the possible medicines, as well as the duration of treatment and number of flasks per month.

Table 2.1: Medicines per type of Keratitis.

Infection Type	Medicine	Flasks per month	Treatment Duration
Bacterial Infection	Moxifloxacin, Cephalothin, Gentamicin, Tobramycin, Ciprofloxacin, Ofloxacin, Gatifloxacin, Vancomycin, Amikacin, Ceftazidime	4 Flasks a month	14 days
Fungi Infection	Amphotericin B, Natamycin, Voriconazole	3 Flasks a month	1-3 months
Amoeba Infection	Biguanide, Chlorhexidine	2 Flasks a month	1-12 months

2.3 Keratitis Epidemiology Analysis

2.3.1 Incidence

Few studies have been conducted to analyse the prevalence of IK across the world, but depending on the area of the globe the incidence is very different as it was previously stated. Table 2.2 shows the prevalence of the disease per 100,000 in several countries. It is very clear to see the gap in infection cases, which is highly attributed to poorer personal hygiene and environmental hygiene, lower access to sanitation and healthcare facilities, as well as an increase in the agricultural industry prevalence.

Table 2.2: Incidence of IK in Different Countries.

Country	Incidence Range (per 100,000 population/year)
United States	2.5–27.6
United Kingdom	0.6–40.3
South India	113
Nepal	799

Looking now at a general outline of the disease across the world (see Figure 2.3), it is easy to observe the difference between developed and developing countries, with the higher incidence rates for Fungi Keratitis occurring in Nepal, India, Kenya, and Egypt, which highlights the impact of country development in the diagnosis.

In Asia and Africa, 45–71% of the patients of Keratitis were illiterate, and 62–79% lived in rural areas that were related to agriculture and had poorer access to healthcare facilities. The significant numbers can be used to conclude that IK is highly discrepant across the globe, meaning that low-income and underdeveloped countries are way more affected by the disease [46].

2.3.2 Age

Regarding age, IK mainly affects humans aged between 30 and 55, which was mainly associated with Contact Lens wear and work-related ocular trauma [48, 49, 58, 61, 63]. The infections related to agricultural trauma were found in patients aged 45–55 years old, which agrees with what was

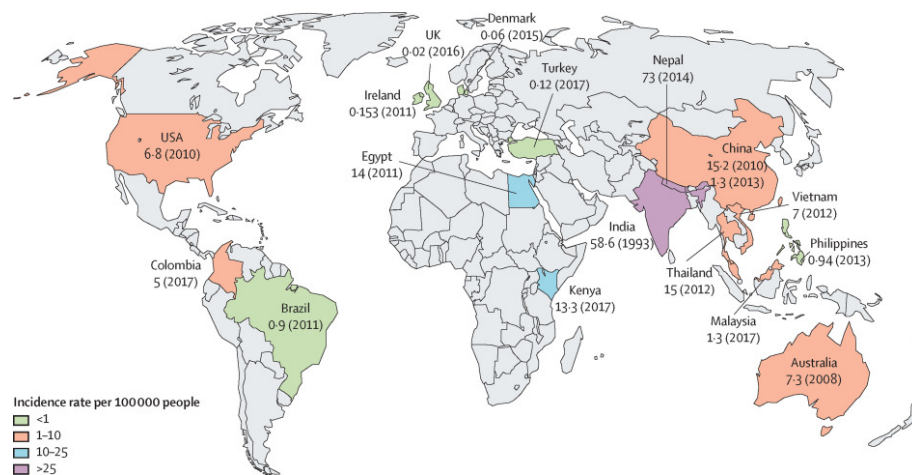


Figure 2.3: Fungi Keratitis Infections across the world from [4].

previously stated, since the workforce of some developing countries is composed of farmers and manual labourers [62, 68]. On the other hand, the CL-related IK is found between 25 and 40 years old [68]. The prevalence in extreme age groups is still very low; nevertheless, in Uganda, it was found to be the second most common cause of childhood blindness. In elders, cases of infections are usually associated with bad complications like corneal melting, perforation, and loss of eye [68].

2.3.3 Sex

There is not a lot of literature on the influence of Sex in Keratitis Infection. However, the CL-related IK was more predominant in females 57–69% and the trauma-related IK was more prominent in males 74–78% [62, 68], which is also related with their gender predominance in under-resourced regions 58–75%.

2.3.4 Seasonal Impact

Seasonal variations influence the incidence of infectious keratitis (IK) worldwide. During summer, it is often associated with Bacteria, due to increased contact lens wear and water-related activities. Conversely, in regions like India, fungal keratitis peaks during windy harvest seasons, linked to agricultural activities. *Acanthamoeba keratitis* shows conflicting seasonal trends: more common in summer in South India but in winter in the UK. These patterns highlight the complex interplay between environmental factors, human behavior, and microbial pathogens in IK epidemiology, guiding targeted preventive measures and treatment strategies [62].

2.4 Conclusions

In conclusion, keratitis remains a significant global health challenge, particularly in low-resource settings where socioeconomic disparities and limited healthcare infrastructure contribute to higher disease burden. Continued research, public health initiatives, and international collaborations are essential for addressing the epidemiological factors driving the incidence of keratitis and improving outcomes for affected individuals worldwide.

2.5 Summary

In this chapter, the epidemiology of keratitis is explored, focusing on its global distribution and factors influencing its incidence rates across different regions and populations. Keratitis, characterised by inflammation of the cornea, presents varying prevalence rates worldwide, with disparities observed between low-resource settings and developed nations. Factors such as geographical location, climate, socioeconomic status, and occupational hazards, including contact lens wear and agricultural activities, play significant roles in shaping the epidemiological landscape of keratitis. Furthermore, demographic factors such as age and sex contribute to variations in disease prevalence. By comprehensively understanding these patterns, we can develop targeted public health interventions and treatment strategies to effectively address the burden of keratitis on a global scale.

Chapter 3

Eyes on the Prize: Artificial Intelligence in Ophthalmology

3.1 Artificial Intelligence in Healthcare

Artificial Intelligence (AI) has become a common term in our daily vocabulary, yet its integration into healthcare remains a domain where considerable concerns persist, despite the widely recognised benefits. While AI technologies hold immense promise in revolutionising medical diagnostics, treatment planning, and patient care, issues such as data privacy, algorithmic bias, and ethical considerations loom large in the healthcare AI landscape [69].

3.1.1 Benefits

There are countless benefits from the use of AI in the healthcare sector:

- **Improved Diagnostics and Treatment Planning:** AI systems can quickly and accurately analyse vast amounts of medical data, resulting in more precise diagnostics. This allows for earlier detection of diseases, which is crucial for effective treatment. AI enhances diagnostic precision by reducing human error and identifying disease markers that might otherwise be overlooked. In addition, its fastness and inability to be tired, make it a strong aided diagnosis tool [5, 27, 70].

In imaging, AI algorithms can detect abnormalities in exams with high accuracy, often challenging human capabilities. BBC reported a case recently where “An AI tool tested by the NHS successfully identified tiny signs of breast cancer in 11 women which had been missed by human doctors” [71].

- **Personalised Treatment Plans:** AI enables the creation of personalised treatment plans by analysing individual patient data, including genetic information, lifestyle, and medical history. This allows for treatments tailored to the specific needs of each patient, improving

outcomes and reducing side effects. In addition, the reaction of the patient to a set of different treatments can be predicted by looking into the patient attributes and also the treatment context [27, 72].

- **Operational Efficiency:** Integrating AI in hospital and clinical settings can streamline administrative tasks, reduce the workload for healthcare professionals, and enhance overall operational efficiency. AI can automate routine tasks such as scheduling and patient record management, allowing healthcare workers to focus more on patient care. A US nurse is predicted to spend 25% of her time on regulatory and administrative work. From bureaucracies to organising doctors' notes to revenue cycle management and medical records management, AI can provide an important improvement and automation in these tasks that ultimately lead to more time for patient care [27, 73].
- **Patient Care improvement:** Even though this area is still raising concerns among users, AI can be used in chatbots for mental health and wellness and telehealth to improve the comfort of the patient. In addition, it can be used as a way to improve the experience of foreigners by providing healthcare in their language or even as a speech recognition tool for those with speech-related issues [74].
- **Cost Reduction:** By improving diagnostic accuracy, personalising treatments, and enhancing operational efficiency, AI can significantly reduce healthcare costs. AI-driven predictive analytics can help in resource allocation, ensuring that medical resources are used efficiently and reducing unnecessary expenditures [75].

3.1.2 Challenges

More cloud storage capacity, data availability and computational power have made it possible to have a lot of AI-related research focus [27]. Nevertheless, the usage of AI in healthcare is still associated with concerns. This is mainly due to concerns in the healthcare sector regarding data collection, technological development, clinical applications and ethical and societal issues [69]:

- **Data Collection Concern:** Database collection is an issue because massive datasets are necessary to accurately use Machine and Deep Learning. In this specific sector, information accessibility becomes a problem due to the confidentiality of medical data, and institutions, hospitals and clinics are reluctant to share the data in order to create proper databases. Experts state that only with a proper mentality shift from treating patients individually to improving healthcare will the creation of massive databases be possible [69, 76].
- **Data Security Concern:** Data Security concerns are also relevant since healthcare data is very valuable and vulnerable. They are often targeted by hackers during data breaches, so maintaining confidentiality and security when creating healthcare databases is essential, as well as patient consent [69, 77]. The General Computational Regulations of Europe [78]

and Health Research Regulations [79], implemented in 2018, aim to restrict personal information collection and sharing. However, these regulations may limit data accessibility for AI system training. To prevent stifling innovation, more stringent data security regulations, client-side data encryption, and federated learning are needed [80].

- **Bias Concern:** This privacy and confidentiality issue creates a big gap in multicultural databases since different countries have different laws, which makes cooperative research hard. Other concerns are raised when dealing with the quality of the existing datasets since they are often erroneous and inconsistently stored, and despite the crescent efforts to clean and organise the data, it is still impossible to control unforeseen gaps. Erroneous diagnosis on the training dataset will certainly lead to erroneous predictions, which can cause malpractice issues and a certain bias.

Another form of bias arises from the under-representation of minorities, particularly those from underdeveloped countries, leading to racial bias. While solutions like multi-ethnic training sets are being explored, their effectiveness remains to be seen [69].

AI techniques can also pick up on features that are not relevant to predict a diagnosis. For example, associate a race or gender with a certain disease and make a diagnosis based on those patient features instead of the existence of a white mass in an exam, for example. This process of the model finding unimportant associations between patients' features and outcomes is overfitting, and it is often associated with too many variables when training a model, which leads the model to have amazing results in the train set but perform very poorly on the test set [81].

- **Lack of Explanation:** Deep learning algorithms lack convincing explanations for their predictions, making it difficult for scientists to understand the data connection to their predictions. This “black box” may lead to a loss of faith in the medical system. However, the mechanism of action of commonly prescribed medications is poorly understood, and most doctors have only a basic understanding of diagnostic imaging tools. The development of AI systems that can be understood by humans is still an active field of study [27, 69, 82].
- **Responsibility:** That is one of the main concerns of artificial intelligence in the medical field since it is difficult to define who would be held accountable for the wrong predictions and erroneous diagnoses, which can affect patients' health. There is a lack of guidelines in evaluating the security and efficacy of AI systems, which are still being refined by entities like the FDA. Doubts have been raised if the doctor should be blamed since he had no part in the development of the software. Because of that, understanding how the data connects to the predictions, is also very important in the healthcare system to allow the doctor to fully understand the “thinking” behind the model decision and correct eventual outliers of the algorithm [27, 69].
- **Social Concerns:** Lastly, social concerns are raised associated with healthcare practitioners losing their jobs, which is mainly due to the general overestimation of AI capabilities [83].

Nevertheless, patient contact jobs are irreplaceable and so only healthcare professionals who deal with digital information, radiology and pathology may be affected. However, the penetration in these fields would be slow since even though there are significant improvements in these fields, their capability is still far from a reliable solution for daily use, working only as an aided diagnosis tool and not a diagnosis tool [27, 84, 85]. Making a comparison to self-driving cars, the Society of Automotive Engineers (SAE) has defined five levels of autonomy (see Figure 3.1), with Level 5 indicating full control without human backup. However, this definition is unlikely to be achieved due to unsafe conditions. Topol [5] stated that medicine will unlikely surpass Level 3, which requires human oversight for algorithmic interpretation of data. Limited human backup is unlikely for Level 4, as human health is too precious to be relegated to machines except for routine matters with minimal risk [86].

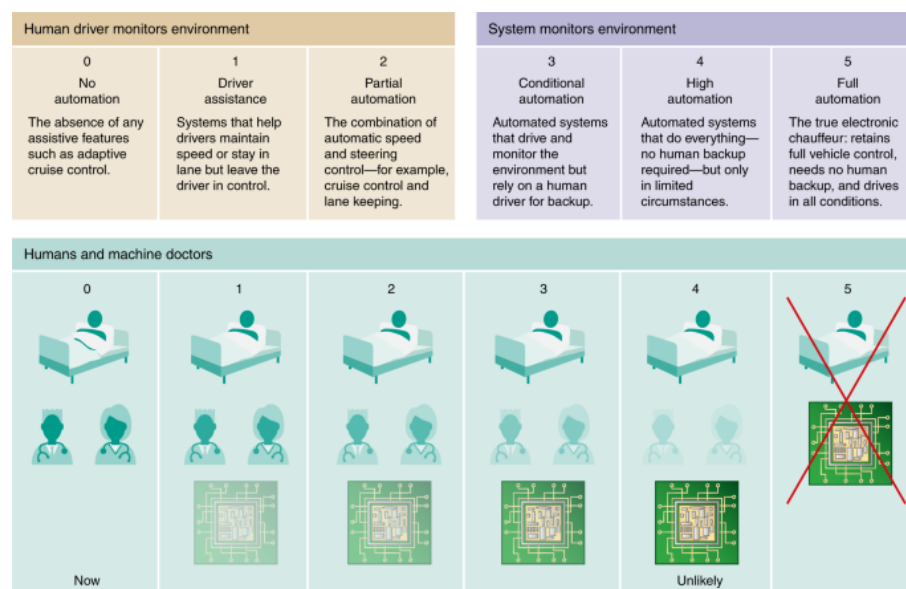


Figure 3.1: Progression of AI in healthcare from [5].

3.1.3 Market Progression

Nevertheless, AI in healthcare is becoming more and more used, and the medical sector is considered one of the most promising areas for this technique. The market of AI in healthcare is growing at an exponential pace, being projected to reach \$188 billion U.S. dollars in 2030, growing at a CAGR of 37% from 2022 to 2030. In 2021, the market size was \$11 billion U.S. dollars (see Figure 3.3) [6].

3.1.4 Relevant Types of AI

AI is a collection of technologies, some of which are of high importance to the healthcare system [27]:

Artificial intelligence (AI) in healthcare market size worldwide from 2021 to 2030

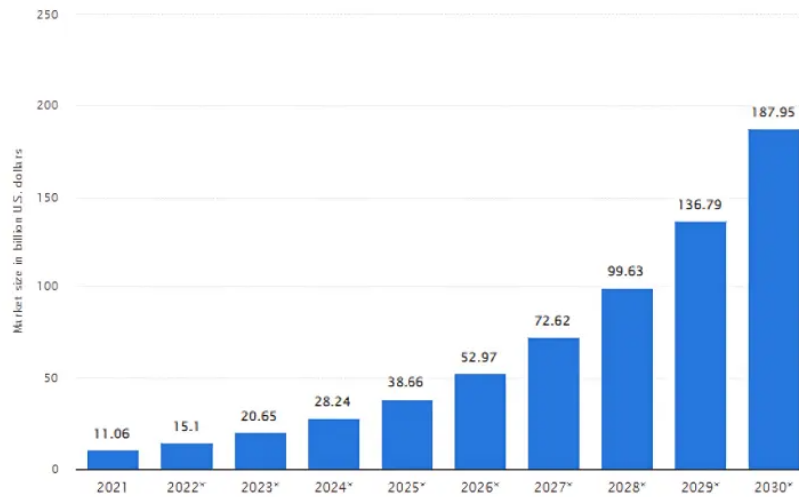


Figure 3.2: Expected growth rate of AI in healthcare in US Billions from [6].

- **Machine Learning:** Machine Learning is one of the most common types of AI, within which different techniques can be explored. Nevertheless, its usage is becoming so natural that a study conducted by Deloitte in 2018 reported that 63% of the managers of the 1,100 inquired companies were already using some source of machine learning in their businesses, highlighting the value of this tool [87]. Simple machine learning models can be used to personalise treatment plans and deep learning techniques can be used to create aided-diagnosis tools.
- **Natural Language Processing:** Focused on understanding the human language, NLP includes speech recognition and translation and is often used in the field to analyse clinical notes, create reports or even interact with a foreigner in their native language to improve comfort [27].
- **Physical Robots:** Robots have become more collaborative and trainable, with increasing intelligence through embedded AI capabilities. Surgical robots, approved in the USA in 2000, empower surgeons with enhanced abilities for precise and minimally invasive procedures. Human surgeons still make crucial decisions, particularly in gynaecologic, prostate, and head and neck surgeries [27, 88].
- **Robotic process automation:** Robotic process automation (RPA) is a technology that automates administrative tasks in information systems, mimicking human actions. Despite the name, RPAs are computer programs, not physical robots. They are cost-effective, easy to program, and transparent in operation. In healthcare, RPAs streamline tasks such as prior authorisations and billing [27].

3.2 AI in ophthalmic pathology and ophthalmology

The integration of AI into eye care has significantly transformed the landscape of diagnosis and treatment for various eye conditions. Specifically focusing on prevalent diseases such as diabetic retinopathy (DR), glaucoma, age-related macular degeneration (AMD), and cataracts, AI technologies have simplified and enhanced the process of identifying and managing these conditions. For instance, in the case of DR, AI algorithms are used to analysing images of the retina to detect early signs of the disease, facilitating timely intervention. Similarly, in glaucoma, AI plays a pivotal role in identifying important indicators like changes in the optic nerve, aiding in the early detection of the condition. Moreover, AI assists in identifying AMD-related abnormalities in images of the eye's posterior segment, enabling prompt treatment decisions. Furthermore, in the realm of cataracts, AI streamlines diagnostic procedures, enabling healthcare professionals to make well-informed decisions regarding patient care. Ultimately, the integration of AI technologies in eye care not only improves diagnostic accuracy but also ensures timely intervention, leading to better patient outcomes and overall efficiency in the management of eye diseases, which explains why constant scientific innovation is taking place in this field over the years (see Figure 3.3) [7].

Additionally, working with eye images poses unique challenges due to the presence of numerous biometric factors such as iris patterns, retinal vasculature, and ocular structures, which can introduce complexities in image analysis and interpretation due to the chance of re-identification of the patient. Moreover, ensuring the privacy and security of sensitive biometric data in ophthalmic imaging presents ethical considerations that must be carefully addressed to mitigate potential risks. To do so, one solution is to use novel learning tasks like generative adversarial networks that hide bio-identical information or work with private datasets [89].

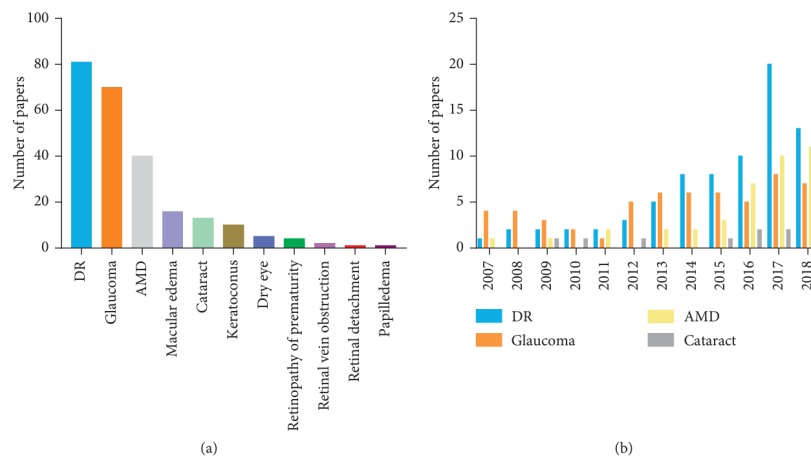


Figure 3.3: Statistics on the number of AI applications in diagnosing ophthalmologic diseases, by disease (a) and by year (b) from [7].

3.2.1 FDA approved products for ophthalmology

On May 13, 2024, the FDA updated its list of AI/ML-enabled medical devices, adding 191 new devices, bringing the total authorised to 882 [90]. From this list, only 9 are in the ophthalmic field and the majority of the approvals are, as expected, in the radiology field (see Figure 3.4). The limited number of products available for ophthalmology can be justified by what was mentioned before regarding biometric data and all the concerns that still exist with AI. The list of existing

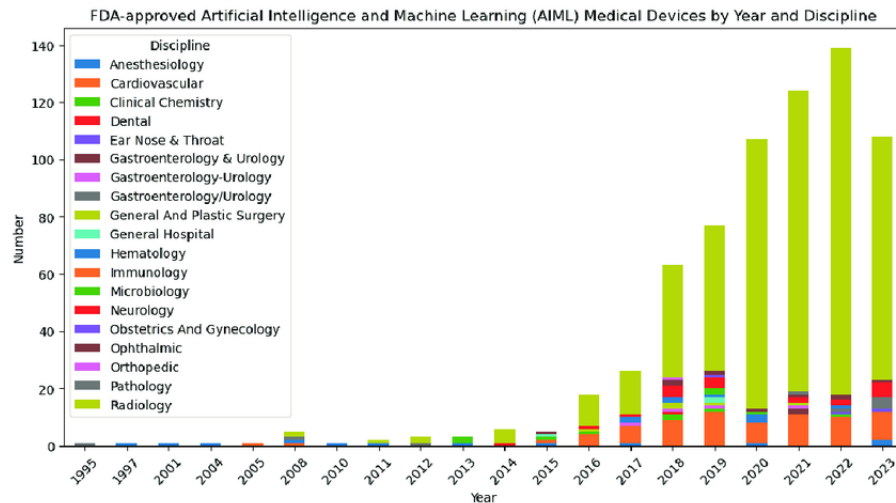


Figure 3.4: FDA approved products for healthcare from [8].

approved products contains the following:

- **IRIS Intelligent Retinal Imaging System (2015):** Intelligent Retinal Imaging Systems (IRIS) is a retinal screening software used in a telemedicine platform to screen for diabetic retinopathy, macular edema, and glaucoma. The service can screen most patients in less than 15 minutes. The digital images are enhanced and sent to a specialist. IRIS has a sensitivity rate of 97% and continuously refines its algorithms [91].
- **IDx-DR (2018, 2021, 2022):** IDx-DR is a software using AI to detect and diagnose Diabetic Retinopathy (DR) in adults over 22 years old. It uses retinal images to detect more than mild retinopathy, allowing ophthalmologists to diagnose the condition more accurately.
- **RightEye Vision System (2018):** Parkinson's disease patients often have persistent ocular tremors that hinder stability during fixation. Oculomotor testing can help diagnose the disease and detect it earlier. The RightEye Vision System, FDA-cleared in 2018, records and analyzes eye movements to identify visual tracking impairment in patients. This data can help doctors identify eye movement deficiencies or anomalies, which can indicate oculomotor and neurological issues [92].
- **CLARUS (2019):** The Clarus 700 camera by Carl Zeiss Meditec, Inc. is a state-of-the-art retinal camera designed for capturing high-quality images of the eye's interior. With its

advanced imaging technology, including ultra-wide field imaging capabilities, the Clarus 700 enables eye care professionals to diagnose and monitor various retinal conditions with exceptional precision and detail [93].

- **EyeArt (2020, 2023):** The EyeArt v2.2.0 system, developed by Eyenuk, is the first AI system FDA-cleared for multiple retinal cameras by different manufacturers, has been approved for detecting vision-threatening DR, age-related macular degeneration, and glaucomatous optic nerve damage in a single test. This follows the system's launch in August 2020, making it the first AI technology to detect these conditions. In Europe, it's the first and only AI system approved under Medical Device Regulation Class IIb to diagnose these conditions [94].
- **AEYE-DS (2022):** The AEYE Diagnostic Screening (AEYE-DS) system detects more-than-mild diabetic retinopathy (DR) in diabetes patients with no prior DR diagnosis. It captures images from both eyes using a desktop camera (Topcon NW-400) and employs AI to analyse the retinal data, delivering diagnostic results in just one minute, compared to the previous AEYE device's standard 15-20 minutes. The FDA cleared this system for AI-based DR diagnostic screening in November 2022, making it commercially available with a tabletop imaging device. The key difference from the previous device is its enhanced portability [95].

3.3 Current Solutions for Keratitis

Several deep learning algorithms have been developed to focus on this specific problem. To understand the state-of-the-art technology, the models will be separated according to the type of classification they perform: Bacterial Keratitis (BK), Fungal Keratitis (FK) or a Multi-infection prediction. It is important to highlight that no study was found for Amoeba Keratitis (AK) vs non-AK classification alone based on eye photographs, probably due to the global percentage of the etiology being so low. Nevertheless, some multi-infection classification systems were found that cover it.

The majority of the classification models focus on multi-infection cases, and because of that, only one study will be presented per infection isolated. It is important to highlight that all studies apply a deep learning algorithm to eye corneal photographs to diagnose Keratitis, and so all the studies that had microscopic images were not included.

3.3.1 Bacterial Keratitis Diagnosis

Regarding Bacterial Keratitis, a study by Kuo, MT et al. [96] in 2021 collected eye photos from patients 1512 patients in Taiwan that presented symptoms of microbial Keratitis. The goal was to diagnose between positive for BK and negative for BK, meaning the patients might have any other type of Keratitis. After normalisation, denoising and resizing, the data was submitted to 6 different models. The learning rate, number of dense blocks, growth rate and batch size were

properly tuned using the validation set. The results for 95% Wilson/Brown binomial confidence intervals regarding accuracy are shown in Table 3.1. It is possible to see that all models perform relatively well, with SE-ResNet50 retrieving the best results, with a 95% confidence interval of 67.3-73.4%, which proves that it is possible to diagnose BK based on eye corneal photographs with high accuracy.

Table 3.1: State of the Art of Bacterial Keratitis Diagnosis: Summary of Authors, Year of Publication, Country of the Dataset samples, Dataset Composition, Predictive Task and Performance metrics.

Author	Country	Dataset	Architecture	Task	Performance (Acc)
Kuo, MT et al. (2021) [96]	Taiwan	1,512 External Eye Photos	ResNet50	BK	63.9-74.5%
Kuo, MT et al. (2021) [96]	Taiwan	1,512 External Eye Photos	ResNeXt50	BK	62.8-76.8%
Kuo, MT et al. (2021) [96]	Taiwan	1,512 External Eye Photos	DenseNet121	BK	67.3-73.4%
Kuo, MT et al. (2021) [96]	Taiwan	1,512 External Eye Photos	SE-ResNet50	BK	68.3-75.0%
Kuo, MT et al. (2021) [96]	Taiwan	1,512 External Eye Photos	EfficientNet B0	BK	63.3-74.2%
Kuo, MT et al. (2021) [96]	Taiwan	1,512 External Eye Photos	EfficientNet B1	BK	65.8-74.8%
Kuo, MT et al. (2021) [96]	Taiwan	1,512 External Eye Photos	EfficientNet B2	BK	65.9-73.3%
Kuo, MT et al. (2021) [96]	Taiwan	1,512 External Eye Photos	EfficientNet B3	BK	64.6-75.9%

3.3.2 Fungal Keratitis Diagnosis

Fungal Keratitis was considered the worst type to diagnose since it is often confused with other etiologies [97]. The same happens when using deep learning solutions to diagnose, which often leads to less accuracy in comparison to other etiologies. The authors of the previous study also focused on developing a deep learning solution for the diagnosis of Fungal Keratitis [9] from other types of Keratitis. The 288 images were normalised and transformed using Gaussian blur and later, random horizontal flips and random color jitters were added with a probability of 0.5. The images were then inserted in a DenseNet based algorithm. The learning rate, number of dense blocks, growth rate and batch size were empirically tuned using the validation set. The results in the specific task were very satisfactory, specially when the difficulty of the task for the doctors is taken into account. The model was able to reach an accuracy of 63.5% to 74.5% with a Wilson/Brown binomial confidence interval of 95% (see Table 3.2).

Table 3.2: State of the Art of Fungal Keratitis Diagnosis: Summary of Authors, Year of Publication, Country of the Dataset samples, Dataset Composition, Predictive Task and Performance metrics.

Author	Country	Dataset	Architecture	Task	Performance (Acc)
Kuo, MT et al. (2020) [9]	Taiwan	288 External Eye Photos	DenseNet	FK	63.5%-74.5%

3.3.3 Multitask Diagnosis

In this section, several multi-diagnostic deep learning solutions will be explored. They were grouped by predictive task, so firstly, the differentiation between Bacterial and Fungal Keratitis will be analysed, followed by the Multi-infection models that include other etiologies. Some of these studies include a Human Comparison that will be properly discussed and analysed in the next section.

3.3.3.1 BK vs FK

Since bacterial and fungi etiologies are the most common sources of infection, they are also the ones that are most targeted when developing deep learning solutions. Throughout this comparative analysis, different methodologies from different authors and different dataset sources were used [98, 99, 100, 101] to identify Bacterial and Fungi infection on eye corneal photographs that could be taken under the slit lamp exam or not. The author, year of publication, country of dataset, dataset composition, model architecture, predictive task and performance metrics are listed in Table 3.3.

Reed et al [98] developed a solution where after image augmentation (random flipping and rotation), to reduce overfitting, the images were resized and preprocessed according to each model's requirements. They entered five pre-trained CNN architectures (MobileNetV2, DenseNet201, ResNet152v2, VGG19, Xception) that were fine-tuned using transfer learning. The final layer of each model was replaced with a global average pooling layer, dropout layer, and a single L2-regularised node, resulting in an output value ranging from 0 to 1. Optimisation was done with mini-batch gradient descent and RMSprop, in order to minimise binary cross-entropy loss. Initially, models acted as feature extractors with frozen layers, then it was fine-tuned by unfreezing deeper layers. Hyperparameters were adjusted based on validation performance, with the primary measure being the AUROC. The ensemble model was the average prediction of all models for each image. MobileNet performed the best with an AUROC of 83%. Another study by Reed et al. in 2021 [99], conducted on a different Indian dataset comprising 50% BK and 50% FK, utilised transfer learning on a pre-trained deep learning system called EyeNet, based on ResNet50. The study reported a classification accuracy of 76%.

Ghosh et al. [100] proposed a solution using a dataset acquired in Thailand, composed of 2,167 slit-lamp anterior segment images, submitted to a deep learning framework. The framework

was built on top of three pretrained CNNs: VGG19, ResNet50, and DenseNet121, all trained using the ImageNet dataset. Each pretrained architecture was expanded with two additional layers: GlobalAveragePooling2D and a dense layer with two outputs. Model enhancements such as early layer freezing, data augmentation, weight decay, and early stopping were implemented to boost performance. Grad-CAM was implemented to assess performance. The output provided the probability of BK and FK being mutually exclusive. This model was considered state-of-the-art, with an AUROC of 86% for VGG19 and 90% for the ensemble of all models.

Lastly, Hung et al. [101] also produced a deep learning system for the classification of these two etiologies. It included two parts, the segmentation and classification model. Firstly, using U^2 Net, the cornea was cropped; then, normalisation and augmentation were applied. Using a 5-fold cross-validation technique, the classification task studied several CNN backbones, including ResNet and DenseNet variants, InceptionV3, and EfficientNetB3. Age and sex were included through fully connected layers, with their output vectors concatenated with those from the global average pooling layer. The model was trained for up to 100 epochs, optimising for maximum accuracy and minimum validation loss. The best result was found for ResNet101 with 80% AUROC.

3.3.3.2 Multiple Infections

Regarding multiple infections, studies were found that included BK, FK, VK, HSK and AK. Xu et al. [102] proposed 3 different approaches that classify FK, BK and HSK based on the entire image, an anatomic-relevant annotated patch or a specific lesion. For the sake of fairness, only the full image-based classification methods are listed in Table 3.4. The image-level feature learning deep model employs transfer learning to deal with limited training data. It uses the original clinical images without annotations for CNN-based diagnostic analysis and classification. Three classic architectures for image classification were used in the experiments: VGG-16, GoogLeNet-v3, and DenseNet. The DenseNet architecture performed best on all three classifications, achieving 60.46 (BK), 56.70 (FK), 80.39 (HSK)% of accuracy.

Zhang et al. [10] developed a solution that currently achieves the best results across all etiologies. The original slit-lamp images were normalised to standardised pixel intensities. Histogram equalisation enhanced the image contrast, while data augmentation techniques like random cropping and flipping were applied to increase diversity and reduce overfitting. The pre-processed images were then fed into multiple state-of-the-art convolutional neural network architectures like ResNet, DenseNet, EfficientNet, and ResNeXt, which were initialised with pre-trained weights from ImageNet. Dynamic learning rate scheduling with a linear warm-up and cosine decay were utilised during training. To boost performance, model blending of top-performing individual models was conducted to create an ensemble that could capture diverse feature representations from the slit-lamp images for robust classification. The best performance was found for 3 types of Convolutional Neural Networks (CNN), DenseNet169, ResNet101_32X8d and ResNet10_32x16d, which have an average ACC of 92%. To try to surpass that performance, Zhange et al. created KeratitisNet (a combination of ResNet101_32x16d and DenseNet169), which ended up surpassing

Table 3.3: State of the Art for multi-infection BK vs FK: Summary of Authors, Year of Publication, Country of the Dataset samples, Dataset Composition, Model Architecture, Predictive Task and Performance metrics

Author	Country	Dataset	Architecture	Task	Performance
Redd et al.(2022) [98]	India	980 External Eye Photos	MobileNet	BK vs FK	AUROC: 83.0%
Redd et al.(2022) [98]	India	980 External Eye Photos	DenseNet	BK vs FK	AUROC: 83.0%
Redd et al.(2022) [98]	India	980 External Eye Photos	ResNet	BK vs FK	AUROC: 82.0%
Redd et al.(2022) [98]	India	980 External Eye Photos	VGG	BK vs FK	AUROC: 75.0%
Redd et al.(2022) [98]	India	980 External Eye Photos	Exception	BK vs FK	AUROC: 75.0%
Redd et al.(2022) [98]	India	980 External Eye Photos	Ensemble	BK vs FK	AUROC: 84.0%
Redd et al.(2021) [99]	India	1,010 External Eye Photos	ResNet-50	BK vs FK	ACC: 76.0%
Ghosh et al.(2022) [100]	Thailand	2167 slit-lamp anterior segment images	VGG19	BK vs FK	AUROC: 86.0%
Ghosh et al.(2022) [100]	Thailand	2167 slit-lamp anterior segment images	ResNet50	BK vs FK	AUROC: 60.0%
Ghosh et al.(2022) [100]	Thailand	2167 slit-lamp anterior segment images	DenseNet121	BK vs FK	AUROC: 73.0%
Ghosh et al.(2022) [100]	Thailand	2167 slit-lamp anterior segment images	Ensemble	BK vs FK	AUROC: 90.0%
Hung et al.(2021) [101]	Taiwan	288 slit-lamp images	DenseNet121	BK vs FK	ACC: 78.8%
Hung et al.(2021) [101]	Taiwan	1330 slit-lamp images	DenseNet161	BK vs FK	ACC: 78.6%
Hung et al.(2021) [101]	Taiwan	1330 slit-lamp images	DenseNet169	BK vs FK	ACC: 79.3%
Hung et al.(2021) [101]	Taiwan	1330 slit-lamp images	DenseNet201	BK vs FK	ACC: 78.4%
Hung et al.(2021) [101]	Taiwan	1330 slit-lamp images	EfficientNetB3	BK vs FK	ACC: 76.1%
Hung et al.(2021) [101]	Taiwan	1330 slit-lamp images	InceptionV3	BK vs FK	ACC: 78.9%
Hung et al.(2021) [101]	Taiwan	1330 slit-lamp images	ResNet101	BK vs FK	ACC: 80.0%
Hung et al.(2021) [101]	Taiwan	1330 slit-lamp images	ResNet50	BK vs FK	ACC: 77.30%

the previous models with an overall AUROC of 92.75%. The comparison with the clinicians' performance will be further analysed in section 3.3.5.

Koyama et al. [103] developed a deep learning algorithm to classify BK, FK, AK and HSK. CNNs (ResNet-50 and InceptionResNetV2) pre-trained on ImageNet were employed, utilising a dataset of 4,306 slit-lamp images from clinical cases and web sources under different illumination conditions. A hierarchical two-step CNN classifier approach was taken to handle class imbalance, with the first CNN discriminating bacterial vs. non-bacterial cases and the second categorising non-bacterial cases into acanthamoeba, fungal, or HSV. The CNN probability scores were integrated with a gradient-boosting decision tree (GBDT) model for the final classification. Data augmentation using web-scraped images and a group k-fold validation strategy were also incorporated to improve model performance and generalisation. The InceptionResnetv2 performed the best results of the CNNs tested.

Lastly, Hu et al. [104] started by augmenting the data by implementing random cropping, rotation, color jitter, horizontal flip, overall noise and strategies to compensate for data imbalance. VGG16, ResNet34, InceptionV4, DenseNet121, ViT-Base, and EfficientNetV2-M were tested as classification models. The models were trained with 150 epochs per model and a 15 batch size. The optimal model was evaluated using the Grad-CAM technique. The best overall accuracy was found for EfficientNetV2-M with 73.5% for diagnosing BK, FK and VK.

Depending on how the results are described in the publication, the overall performance (average of the performance of the metric for all tasks) or the performance for each task (ordered according to how they appear in the Task column) is shown in Table 3.4.

3.3.4 Datasets

Due to the sensitivity biometric information present in these types of images, there are no public datasets for the disease. Nevertheless, private datasets are described in the literature. For the BK vs non-BK classification, Kuo et al. [96] acquired a Taiwanese dataset, composed of 1,512 external eye photos. Of those, 929 has a BK label and the rest a non-BK label, presenting signs of infection by FK and VK. In its FK vs non-FK paper [9], another Taiwanese dataset was acquired, which comprised 288 external eye photographs. Of those, 114 had a positive label for FK and 174 a non-FK label, this group included exams that tested positive for BK and HSK.

Focusing now on the datasets used for BK vs FK, Redd et al. [98] used an Indian dataset composed by 980 external eye exams, of which 480 had BK and 500 FK. In the other published study by the author [99] a bigger dataset of 1,010 images was used and the proportions of etiologies were 50% each. Ghosh et al. [100] used a dataset acquired in Thailand that was composed by 2167 slit-lamp anterior segment images. Of those, 1388 had a positive diagnosis for BK and 779 for FK. Lastly, Hung et al. [101] utilises a Taiwanese dataset composed by 1330 slit-lamp images. Of those, 824 are positive for BK and 506 for FK.

Lastly, regarding the multi-infection studies, Xu et al [102] used a chinese dataset composed by 2284 slip lamp photos. In which, 473 had BK, 616 FK, 539 HSK and 656 other corneal diseases. The dataset of Zhang et al. [10], also chinese, has 1490 images of BK, 1670 of FK, 600

Table 3.4: State of the Art for multi- infection: Summary of Authors, Year of Publication, Country of the Dataset samples, Dataset Composition, Predictive Task and Performance metrics.

Author	Country	Dataset	Architecture	Task	Performance
Xu et al.(2021) [102]	China	2284 slit lamp photos	VGG16	BK vs FK vs HSK	ACC: 48.84, 52.57, 62.74,%
Xu et al.(2021) [102]	China	2284 slit lamp photos	GoogLeNet-v3	BK vs FK vs HSK	ACC: 53.49, 55.67, 66.67%
Xu et al.(2021) [102]	China	2284 slit lamp photos	DenseNet-v3	BK vs FK vs HSK	ACC: 60.46, 56.70, 80.39%
Zhang et al.(2022) [10]	China	4830 slit lamp images	ResNet18	BK vs FK vs AK vs HSK	AUROC: 82, 88, 95, 96%
Zhang et al.(2022) [10]	China	4830 slit lamp images	ResNet50	BK vs FK vs AK vs HSK	AUROC: 84, 89, 95, 98%
Zhang et al.(2022) [10]	China	4830 slit lamp images	DenseNet121	BK vs FK vs AK vs HSK	AUROC: 82, 87, 95, 96%
Zhang et al.(2022) [10]	China	4830 slit lamp images	DenseNet169	BK vs FK vs AK vs HSK	AUROC: 85, 90, 95, 98%
Zhang et al.(2022) [10]	China	4830 slit lamp images	ResNet101_32X8d	BK vs FK vs AK vs HSK	AUROC: 85, 89, 96, 98%
Zhang et al.(2022) [10]	China	4830 slit lamp images	ResNet101_32x16d	BK vs FK vs AK vs HSK	AUROC: 84, 90, 96, 98.0%
Zhang et al.(2022) [10]	China	4830 slit lamp images	EfficientNet-b0	BK vs FK vs AK vs HSK	AUROC: 85, 88, 95, 97%
Zhang et al.(2022) [10]	China	4830 slit lamp images	EfficientNet-b5	BK vs FK vs AK vs HSK	AUROC: 82, 87, 96, 97%
Zhang et al.(2022) [10]	China	4830 slit lamp images	EfficientNet-b7	BK vs FK vs AK vs HSK	AUROC: 83, 89, 97, 98%
Zhang et al.(2022) [10]	China	4830 slit lamp images	KeratitisNet	BK vs FK vs AK vs HSK	AUROC: 86, 91, 96, 98%
Koyama et al.(2021) [103]	Japan	4306 External Eye Photos	ResNet50	BK vs FK vs AK vs HSK	AUROC: 82, 78, 84, 73%
Koyama et al.(2021) [103]	Japan	4306 External Eye Photos	InceptionResnetV2	BK vs FK vs AK vs HSK	AUROC: 87, 86, 99, 93%
Hu et al.(2023) [104]	China	2757 slit lamp images	ResNet34	BK vs FK vs VK	ACC: 63.5%
Hu et al.(2023) [104]	China	2757 slit lamp images	Densenet121	BK vs FK vs VK	ACC: 63.7%
Hu et al.(2023) [104]	China	2757 slit lamp images	ViT-Base	BK vs FK vs VK	ACC: 69.7%
Hu et al.(2023) [104]	China	2757 slit lamp images	VGG16	BK vs FK vs VK	ACC: 70.8%
Hu et al.(2023) [104]	China	2757 slit lamp images	InceptionV4	BK vs FK vs VK	ACC: 71.6%
Hu et al.(2023) [104]	China	2757 slit lamp images	EfficientNetV2-M	BK vs FK vs VK	ACC: 73.5%

of HSK, and 1070 with AK. Koyama et al. [103] worked with a Japanese dataset of 4306 external eye photos, nevertheless the number of cases of each infection is not disclosed. Lastly, Hu et al. [104] worked with a Chinese dataset of 2757 images, of which 592 were healthy, 813 had a positive label for VK, 1125 a positive label for FK and 227 for FK.

3.3.5 Ophthalmologists vs. AI

Different studies have been conducted to evaluate the ophthalmologist's performance metrics when compared to those of deep learning solutions. Human performance is often lower than the deep learning system and much more dependent on other variables like the years of expertise and how familiar they are with the disease.

3.3.5.1 BK vs. non BK

Studies conducted on the diagnosis of BK alone have shown that the clinicians have a performance of 66.0 to 74.0%, making the performance of the system presented a valid aided-diagnosis tool since similar metrics were found [96].

3.3.5.2 FK vs. non FK

According to Dahlgren et al. [97], doctors have a sensitivity of 38% and a specificity of 45% when diagnosing FK clinically based only on eye photographs. The authors of the previously presented study for fungal Keratitis alone compared the performance of Non-Corneal Speciality Ophthalmology (NCS-Oph) and experts with the one of their model and realised that, although the DL proposed framework was capable of beating the NCS-oph, it performed lower than the Experts (Figure 3.5) [9].

3.3.5.3 Multi-infection

The diagnosis of IK using slit-lamp photography by cornea specialists has an accuracy rate of only 66.0–75.9% in distinguishing between bacterial and fungal keratitis [66]. Nevertheless, since this performance is very dependent on the dataset, several studies compared the model and the experts' performance; in its 2022 study, Redd et al. [98] compared the performance of the doctors in the same Indian dataset that was used to train the DL model. 66 international expert cornea specialists were responsible for this classification and reached an AUROC of 76% which is significantly lower ($p < 0.01$) than the one found by the DL architecture 83%. Xu et al. [102] did a similar study where doctors achieved 46.55% BK, 45.56% FK, 65.01% HSK and when the patients' history was available, the results improved to 55.55% BK, 56.28% FK, 73.25% HSK, which is still lower than the DL, metrics for image-level classification, the decrease in performance can be particularly seen for HSK. Zhang et al. [10] compared the performance (see Figure 3.6) of KeratitisNet with that of expert ophthalmologists and the results showed that the model had a correct prediction rate at

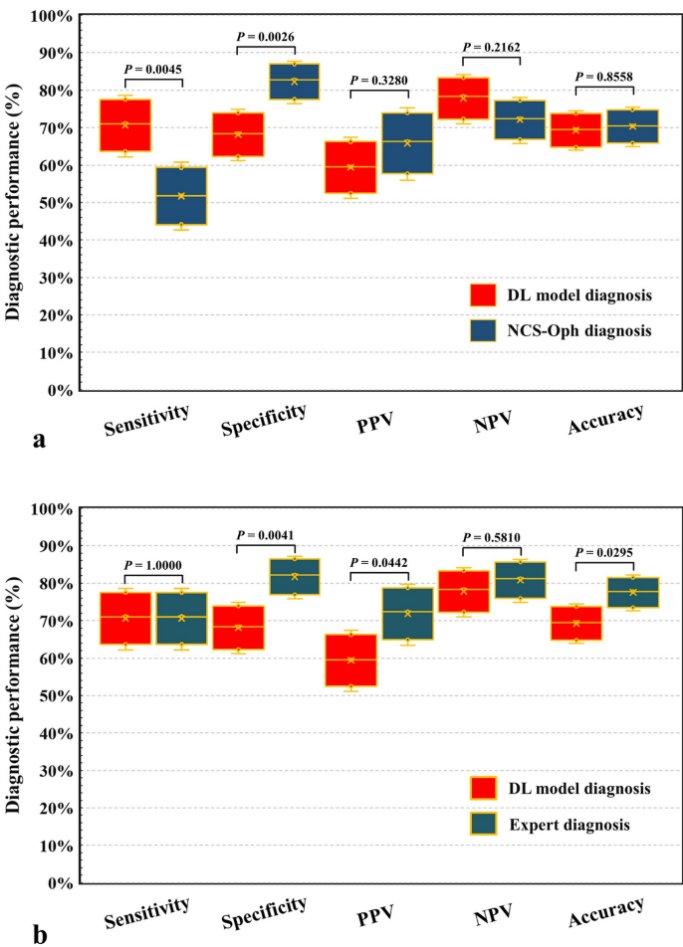


Figure 3.5: DL model vs NCS-Oph (a) and Experts (b) from [9].

least 0.10 higher than the healthcare practitioners, with the maximum difference being 0.5 for AK, proving once again how useful these developed models can be as aided diagnosis tools [105].

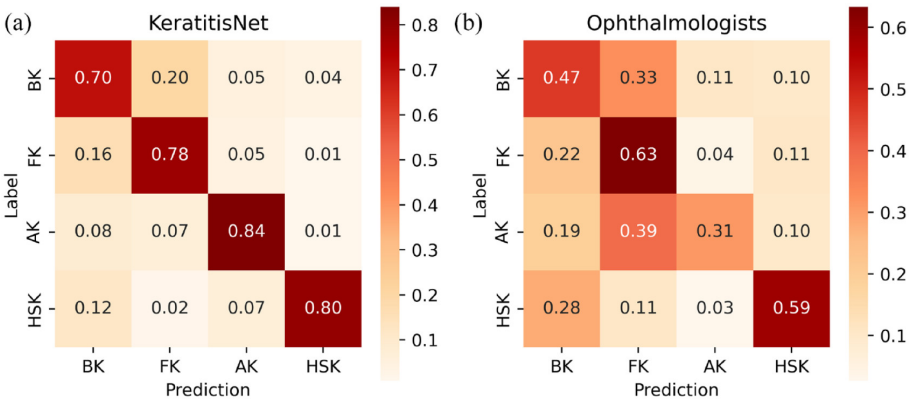


Figure 3.6: Confusion matrix of KeratitisNet (a) and ophthalmologists (b) in distinguishing IK from [10].

3.4 Computer Vision

Deep learning distinguishes itself by its ability to autonomously learn feature representations from raw data, with multiple levels of abstraction, setting it apart from traditional machine learning approaches that often require a manual feature extraction process. Models such as CNNs have demonstrated exceptional effectiveness in fields like computer vision, natural language processing, and speech recognition. Nonetheless, deep learning also presents challenges, such as the necessity for large amounts of labelled data, substantial computational resources, and the complexity of interpreting intricate models, which is particularly problematic in the healthcare sector. [1, 26, 27, 106].

3.4.1 Models Architecture

Because of the ability to automatically learn and extract meaningful features from raw images, deep learning has become particularly common in image-related tasks since CNNs have proven to be exceptionally effective in understanding and interpreting visual data. That is because of their ability to automatically learn hierarchical representations of images, starting from basic features and progressing to complex patterns and structures, mimicking the way our eyes work. This technique enables CNNs to generalise well to new, previously unseen pictures, which is critical in tasks like as object identification. Furthermore, CNNs are meant to be insensitive to spatial transformations such as translation, rotation, and scaling due to the way their weights are shared. This trait lowers the computational cost and eliminates the need to manually construct features, making CNNs more economical and scalable for large-scale picture collections [106].

Several CNN architectures have been developed specifically for image classification tasks, each offering unique design principles and performance characteristics. Examples include ResNet, DenseNet, and VGG16, which will be further explored in chronological order.

3.4.1.1 VGG16

VGG16 is a deep CNN architecture proposed by Simonyan and Zisserman [107].

VGG16 (see Figure 3.7) is recognized for its simplicity and uniform structure, featuring 16 layers, including 13 convolutional layers and 3 fully connected layers. The convolutional layers utilise 3x3 filters with a stride of 1 and padding, maintaining spatial resolution. The number of filters increases from 64 in the first layer to 512 in the deeper layers, doubling after each max-pooling layer.

Max-pooling layers, with 2x2 kernels and a stride of 2, are placed between convolutional blocks. These layers reduce spatial dimensions while preserving crucial features, aiding translation invariance and reducing computational complexity.

The fully connected layers, positioned after the convolutional layers, each contain 4096 neurons. These layers enhance high-level feature representation, followed by a softmax layer that facilitates class prediction based on the features extracted from the convolutional layers. This

combination of convolutional and fully connected layers allows VGG16 to effectively classify images by progressively extracting and interpreting intricate features, nevertheless compared to modern architectures it is considered slow and less efficient and with a very high computational cost [108].

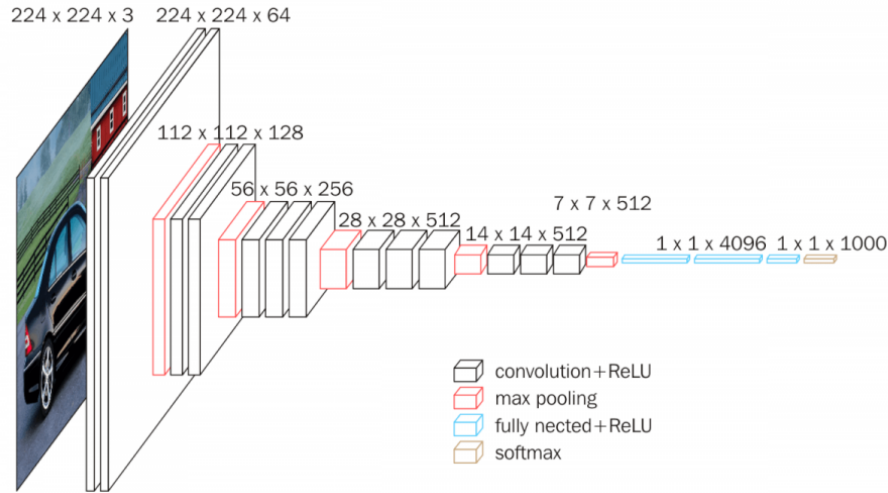


Figure 3.7: VGG16 Architecture diagram from [11].

In comparison to other similar structures, VGG16 has the following benefits:

- **Simplicity:** VGG16's uniform architecture and consistent use of 3x3 filters make it easy to understand and implement.
- **Strong Baseline Performance:** VGG16 provides a strong baseline performance for image classification tasks, achieving high accuracy on datasets like ImageNet.
- **Transfer Learning:** Pre-trained VGG16 models trained on large datasets like ImageNet can be fine-tuned for various image recognition tasks with limited data, thanks to its generalisation capabilities.

3.4.1.2 ResNet50

ResNet-50 is a convolutional neural network architecture introduced by He et al [109].

ResNet50 (see Figure 3.8) is a leading image classification model that uses residual connections to enable deep architecture learning without vanishing gradients. Its architecture has four main parts: convolutional layers, identity blocks, convolutional blocks, and fully connected layers.

The convolutional layers extract features like edges and textures from images using convolution, batch normalization, and ReLU activation, followed by max pooling to reduce spatial dimensions. Identity blocks pass the input through convolutional layers and add the input back to

the output, allowing the network to learn residual functions. Convolutional blocks are similar but include a 1x1 convolution to reduce filter numbers before a 3x3 convolution.

Finally, the fully connected layers make the final classification, with the output passed through a softmax function to produce class probabilities. This structure enables ResNet50 to efficiently process and classify images by learning complex residual functions [110].

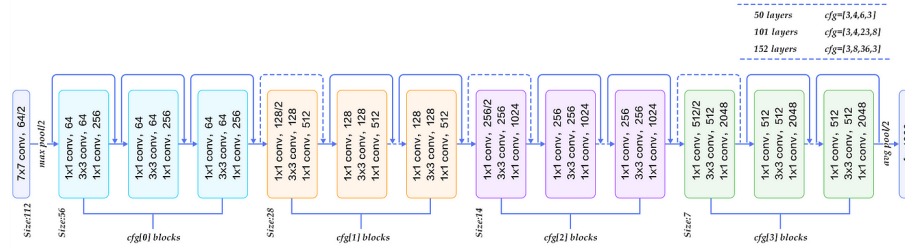


Figure 3.8: ResNet50 Architecture diagram from [12].

In comparison to other similar structures, ResNet50 has the following benefits:

- **Effective Training of Deep Networks:** ResNet50's residual learning framework enables the training of very deep networks (up to 50 layers) without suffering from the degradation problem.
- **State-of-the-Art Performance:** ResNet50 achieves state-of-the-art performance on various image recognition benchmarks, including the ImageNet dataset, surpassing previous architectures in terms of accuracy and efficiency.

3.4.1.3 DenseNet121

DenseNet-121 [111] (see Figure 3.9) is a convolutional neural network architecture that belongs to the family of Densely Connected Convolutional Networks (DenseNets), introduced by Huang et al known for its efficiency and accuracy in image classification. The number in the name is related to the number of layers with trainable weights [111, 112].

The key innovation of DenseNet is its dense connectivity pattern, where each layer receives input from all preceding layers. In traditional convolutional neural networks, feature maps are concatenated across the depth dimension, which results in a “tower-like” architecture. However, in DenseNet, each layer is connected to every subsequent layer in a feed-forward manner, meaning that each layer receives feature maps from all preceding layers and passes its feature maps to all subsequent layers. This dense connectivity promotes feature reuse and facilitates gradient flow during training.

To control the growth of feature maps and reduce computational complexity, DenseNet employs transition layers between dense blocks. These transition layers consist of convolutional and pooling operations that reduce the spatial dimensions of feature maps and the number of channels which helps maintain model compactness while preserving valuable information.

horizontally, and the padding means how many pixels are added around the maps before downsizing to retain the spatial dimensions [114].



Figure 3.10: Average pool operation with kernel 3×3 from [14].

3.4.2.2 Batch Normalisation

A Batch Normalisation layer in neural networks normalises each mini-batch's input to have a mean of zero and a standard deviation of one, hence reducing internal co-variate shift. The Batch Normalisation layer, which is applied after the linear transformation but before the activation function, has a regularisation effect that helps to reduce overfitting by adding noise via mini-batch statistics. Furthermore, it minimises sensitivity to parameter initialisation, making it an essential component in recent deep learning systems [115].

3.4.2.3 Dropout

The dropout layer randomly drops some nodes and their forward and backward connections with a certain probability rate defined. This random "dropping out" of neurons encourages the network to acquire more robust properties since it cannot rely on the existence of any specific neuron (see Figure 3.11). As a result, the dropout layer improves model generalisation by decreasing overfitting, making the network less sensitive to noise and particular patterns in training data. Usually a probability ranging from 0.3 to 0.5 is used [15].

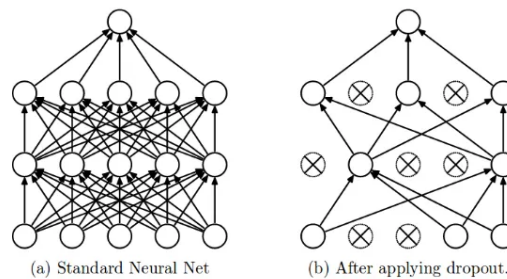


Figure 3.11: Dropout layer from [15].

3.4.3 Learning Process

3.4.3.1 Loss Function

A loss function in machine learning quantifies the difference between the predicted values produced by a model and the actual values observed in the data. It measures the error or discrepancy in predictions, providing a numerical value that indicates how well the model's predictions align with the true data. A key property of many loss functions is that they are differentiable, meaning they have well-defined gradients or derivatives at each point in their domain. This differentiability allows optimisation algorithms, such as Adam Optimiser, to compute gradients and iteratively adjust model parameters to minimise the loss. A lower loss indicates better model performance, as it signifies that the model's predictions are closer to the actual values. A common loss function is Mean Squared Error (MSE), which squares the differences between predicted and actual values, and then averages them. If predictions are perfect, MSE is zero; otherwise, it's a positive number. MSE is computed using Equation 3.1, where each variable assumes the same meaning as it did in Equation 3.8. The goal is to minimise this loss by adjusting the model's parameters.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (3.1)$$

Different tasks require different loss functions. Cross-entropy loss is used for classification tasks. It is computed by using Equation 3.2, where N is the total number of observations, $y_i \in \{0, 1\}$ represents the actual label for the prediction i and $\hat{y}_i$ is the predicted probability of the positive class for that observation [116].

$$CrossEntropy = -\frac{1}{n} \sum_{i=1}^n (y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)) \quad (3.2)$$

Choosing the right loss function is crucial (see Figure 3.12), as it affects how well the model learns from data. Additionally, monitoring the loss and the difference between train and validation loss over the epochs helps assess the model's performance and can indicate issues like overfitting or poor parameter tuning [16].

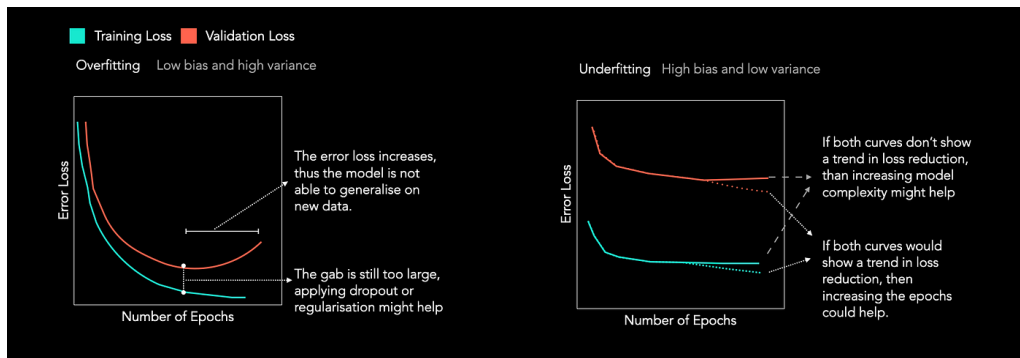


Figure 3.12: Loss analysis from [16].

3.4.3.2 Optimiser

Optimisers are an important part of deep learning algorithms since they dynamically fine-tune a model's parameters throughout the training process to minimise a predetermined loss function. They let neural networks learn by iteratively modifying weights and biases depending on data feedback. Stochastic Gradient Descent (SGD), Adam, and RMSprop are well-known deep learning optimisers, each with their own set of update rules, learning rates, and momentum strategies, all aimed at converging on optimal model parameters and thus improving overall performance. Adam (Adaptive Moment Estimation), one of the most common ones, is a powerful optimisation technique that combines the advantages of two previous stochastic gradient descent extensions: Adaptive Gradient Technique (AdaGrad) and Root Mean Square Propagation (RMSProp). It computes adaptive learning rates for each parameter by calculating the first and second moments of the gradients, enhancing convergence speed and stability. The step size between the first and second moments is defined by the learning rate. Additionally, the weight decay parameter acts as a penalty added to the loss function to reduce overfitting [117].

3.4.3.3 Backpropagation

Backpropagation plays a pivotal role in this learning process. When a neural network makes predictions, it first processes the input data through its layers in a step called forward propagation. The output of the network is then compared to the actual target values to calculate the loss.

The primary goal of backpropagation is to reduce the calculated loss by adjusting the weights in the network. It does this by propagating the error backward from the output layer to the input layer (see Figure 3.13). The process involves computing the gradient of the loss function with respect to each weight in the network. This gradient tells us how much a small change in each weight will affect the loss.

Using the chain rule of calculus, backpropagation efficiently computes these gradients. Once the gradients are calculated, the optimiser uses them to update the weights. Typically, the optimiser adjusts each weight by moving it in the direction that decreases the loss the most. This adjustment is controlled by a parameter called the learning rate, which determines the size of the steps taken during the update process.

This cycle of forward propagation, loss calculation, gradient computation, and weight update is repeated many times. Each cycle is known as an iteration, and a full pass through the entire training dataset is called an epoch. Through many epochs, the weights are iteratively adjusted to minimise the loss, leading to a trained neural network that makes accurate predictions [17].

3.4.4 Performance Metrics

To quantify the performance of a classification task, some metrics are commonly used. To understand them, it is important to define four concepts [118]:

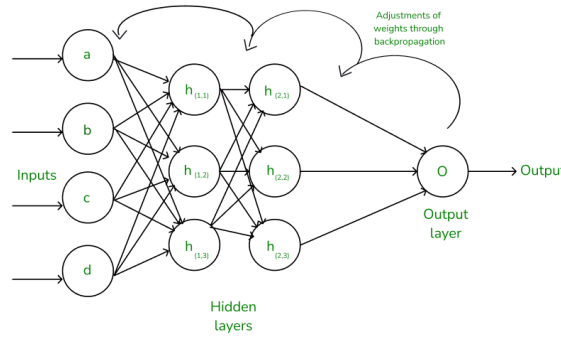


Figure 3.13: Backpropagation from [17].

- True Positive (TP): This occurs when the prediction is positive, and it correctly matches the reality.
- True Negative (TN): This happens when the prediction is negative, and it correctly aligns with reality.
- False Positive (FP): In this scenario, the prediction is positive, but it inaccurately predicts a positive outcome when the reality is negative.
- False Negative (FN): Here, the prediction is negative, but it incorrectly predicts a negative outcome when the reality is positive.

3.4.4.1 Confusion Matrices

It provides a comprehensive examination of true positive, true negative, false positive, and false negative predictions, allowing for a more in-depth understanding of a model's recall, accuracy, precision, and overall efficacy in class differentiation (see Figure 3.14). When there is an uneven class distribution in a dataset, this matrix is very useful for assessing a model's performance beyond simple accuracy measurements.

3.4.4.2 Accuracy (ACC)

Classification accuracy is probably the most straightforward statistic metric to use in model access, defined as the number of right predictions divided by the total number of predictions (see Equation 3.3). Accuracy is beneficial and usually used when the target classes are balanced. [118].

$$Accuracy = \frac{Number_of_correct_predictions}{Total_number_of_predictions} \quad (3.3)$$

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 3.14: Confusion Matrix scheme from [18].

3.4.4.3 Area Under the Curve (AUROC)

The ROC (receiver operating characteristic) curve serves as a visual representation of how well a classification model performs across different classification thresholds. It showcases the relationship between the True Positive Rate (TPR), which indicates the percentage of actual positive instances correctly identified as positive, and the False Positive Rate (FPR), which represents the percentage of actual negative instances incorrectly identified as positive.

By plotting TPR against FPR at various thresholds, the ROC curve illustrates the trade-off between these rates. As the classification threshold is lowered, more instances are classified as positive, resulting in an increase in both TPR and FPR. This dynamic is effectively captured and displayed through the ROC curve.

AUROC, short for Area Under the ROC Curve, quantifies the overall performance of a classification model by measuring the area under its ROC curve (Figure 3.15). It ranges from 0 to 1, where 0 indicates complete mispredictions and 1 indicates perfect predictions. AUROC provides a scale-invariant and threshold-invariant measure of model performance, making it valuable for assessing the quality of prediction rankings across different thresholds. However, it may not address scenarios where precisely calibrated probabilities or specific optimisation of classification errors are required [19, 118].

3.4.4.4 Precision

Precision addresses a limitation of accuracy by focusing specifically on the proportion of positive predictions that were correct. It quantifies the accuracy of positive predictions by measuring the ratio of True Positive predictions (those that are actually true) to the total positive predictions (the sum of True Positive and False Positive predictions)(see Equation 3.4). In essence, precision evaluates the model's ability to avoid false positives and ensure that positive predictions are indeed

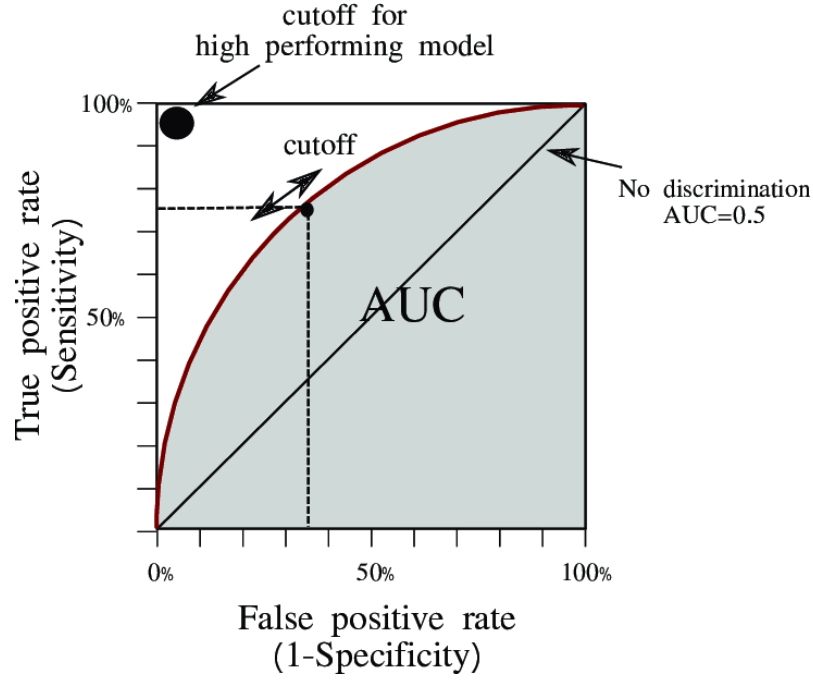


Figure 3.15: ROC Curve from [19].

accurate. If precision is maximised, the FN errors are minimised [118].

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \quad (3.4)$$

3.4.4.5 Recall

The Recall metric calculates the proportion of incorrectly identified positives, referred to as True Positive, based on predictions that are true to the total number of positives (see Equation 3.5). If precision is maximised, the TN errors are minimised.

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \quad (3.5)$$

3.4.4.6 F1-Score

The F-score, also known as the F1 Score (see Equation 3.6), is a metric used to evaluate binary classification models based on positive predictions. It represents both precision and recall and is calculated as the harmonic mean of both, assigning equal weight to each. It is useful when both are important for evaluation, but one is slightly more important than the other, such as when false negatives are more important than false positives [118].

$$F1_Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (3.6)$$

3.4.4.7 Balanced Accuracy

Balanced accuracy is a statistical metric used to evaluate classification algorithms' performance on unbalanced datasets. It represents the arithmetic mean of sensitivity (true positive rate) and specificity (true negative rate), ensuring that both minority and majority classes are given equal weight during evaluation (see Equation 3.7). 0 indicates a poor classifier and 1 a perfect one [119].

$$\text{BalancedAccuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2} = \frac{2TN \cdot TP + TP \cdot FP + TN \cdot FN}{2} \quad (3.7)$$

3.4.4.8 Mean Absolute Error

Mean Absolute Error (MAE) is commonly employed when dealing with continuous data classification tasks. It serves as a straightforward metric to assess the average magnitude of errors between predicted and actual ordinal values. Mean Absolute Error is a regressive loss metric that examines the absolute value difference between a model's predictions and ground truth, averaged over the dataset. MAE is weighted on a linear scale and hence does not place as much emphasis on outliers, it is calculated by Equation 3.8, where N is the total number of observations, y_i represents the actual value for observation i and $\hat{y}_i$ represents the predicted value for observation i . This gives a more uniform measure of performance, but it also means that both large and minor faults are weighted equally. Nevertheless, a bigger difference (which represents a higher misclassification) will impact the result.

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (3.8)$$

3.5 Conclusions

The integration of AI in ophthalmology and broader healthcare offers immense potential, if the current concerns are effectively addressed. AI has already demonstrated its capability to enhance diagnostic accuracy, streamline workflows, and reduce costs. Specifically, the state-of-the-art technology for keratitis has shown that deep learning models can outperform traditional diagnostic methods used by doctors. These models facilitate deeper, faster, and more accurate diagnoses while minimising the need for costly lab tests, making them particularly valuable in low- and middle-income countries.

3.6 Summary

AI in healthcare is revolutionising diagnostics and treatment, providing numerous benefits while also presenting several challenges. It enhances diagnostic accuracy, enables early disease detection, and supports personalised treatment by analysing extensive patient data. Additionally, AI automates routine tasks, reducing costs and saving time. However, concerns such as data privacy, potential biases from training on biased datasets, and regulatory issues pose significant challenges.

In ophthalmology, FDA-approved AI products like IDx-DR and EyeArt autonomously detect diabetic retinopathy, improving diagnostic precision and efficiency.

For the specific case of keratitis, no FDA approved product exists, but several deep-learning models have been developed to classify the etiology based on eye photographs. These models have shown significantly better results compared to ophthalmologists, proving AI's value as an aided-diagnostic tool.

Chapter 4

Insightful Vision: Unravelling the Dataset

4.1 Dataset Characterisation

Because there are no public datasets on the topic, a new private one was collected. The Dataset, acquired in São Paulo, Brazil by the clinical partners (Luis Nakayama and Camila Kase) of the Universidade Federal de São Paulo (UNIFESP), contains healthy patients for keratitis and non-healthy patients that can be infected by amoeba, bacteria or fungi. It is composed by metadata and the corresponding eye corneal photograph for each exam. There are 24,691 exams present in the dataset, belonging to 3296 patients, meaning that there are multiple exam cases for each patient, usually portraying the time progression.

4.2 Dataset Features

The dataset contains 26 attributes described in Table 4.1.

4.2.1 General Analysis

The dataset is composed by healthy and non-healthy exams for keratitis of several patients. Because of this, some cases were immediately considered healthy for Keratitis by the doctor which led to no tests and further information being asked. Those cases only have information for “exam_date”, “patient_id”, “birth_date”, “file_names”, “date”, “olho”. Of the 24,691 exam cases that compose the dataset, 6,681 have medical info although 1,914 exam cases came out negative for all laboratory tests, meaning only 4,767 have a keratitis infection.

As described in the disease characterisation, bacteria is the most common etiology. In the dataset the majority of the infections are also by bacteria, reaching 3,956 cases, which accounts for 82.99% of the dataset. Fungi keratitis represents 755 of the cases, which is 15.85%. Lastly, acanthamoeba has 736 cases that corresponds to 15.44%. That allowed the conclusion that 680 exams have multiple infections according to the laboratory diagnosis of the individual test. A

Table 4.1: Attributes and Features available in the Dataset.

Feature Name	Data Type	Description
exam_date	String	Date of the eye corneal photograph (YYYY-MM-DD)
patient_id	Int	Unique Patient Identifier
birth_date	String	Birth date of the patient (DD-MM-YYYY)
file_names	String	Name of file the eye corneal photo is stored as
date	String	Date of the sample collection (YYYY-MM-DD)
olho	String	Eye of the sample, OE (left) or OD (right)
Sexo	String	Defines the sex of the patient, can either be 'Feminino' (Female) or 'Masculino' (Male)
Idade	Float	Age of the patient (e.g. 50.0)
Data Coleta	String	Date of the collection of the sample for lab analysis (YYYY-MM-DD)
Origem Material	String	Where was the sample acquired (service wise)
material	String	How was the cornea sample collected
Histórico	String	Symptoms and patient history
medicamento_anterior	Boolean	If the patient ever took a medicine for eye related problems
Medicamentos Anteriores	String	The medicines the patient took previously for eye related problems
cirurgia_anterior	Boolean	If the patient ever had an eye surgery
Tipo Cirurgia Anterior	String	The name of the surgery
Tempo Evolução	String	Time since the first symptom
fungo	Boolean	If the cornea lab sample was submitted for fungi analysis
Cultura para FUNGOS Res	String	Result of the test for fungi, 'Positiva' for positive and 'Negativa' for negative
Fungo Filamentoso	String	Type of fungi of that category that the patient has
Fungo Leveduriforme	String	Type of fungi of that category that the patient has
acanthamoeba	Boolean	If the cornea lab sample was submitted for acanthamoeba analysis
Cultura para ACANTHAMOEBA Res	String	Result of the test for acanthamoeba, 'Positiva' for positive and 'Negativa' for negative
bacteria	Boolean	If the cornea lab sample was submitted for bacteria analysis
Cultura para BACTÉRIA Res	String	Result of the test for acanthamoeba, 'Positiva' for positive and 'Negativa' for negative
Date_Difference	Float	Time between the sample and the eye corneal photograph

deeper analysis showed the conclusion that 270 cases have bacteria and fungi infections and 410 report Bacteria and acanthamoeba infections. However, it is possible to conclude that the dataset has a significant imbalance in terms of infection type. Below, it is possible to see some examples of the images of the dataset (see Figures 4.1- 4.6), denoting not only its variety but also the different

phenotypes the disease can acquire, highlighting the difficulty of this task. When performing a deeper analysis on the dataset, it was clear that several entries of the dataset were associated with the same filename, meaning that the clinicians recorded different test acquisitions for the same image. To properly clean the dataset, and after a Clinician's validation, it was decided to summarise those different tests in one entry, meaning that if there was a positive result for that etiology in any date of the different acquisitions, the image would be considered positive for the infection (see Figure 4.7). If not, the image would be considered as a negative for that infection. This rule was applied to all the exam photos in all the infections, resulting in 2585 different exam cases of which 2024 had multiple entries that were summarised using this technique. This affected how samples were counted of each infection since, before data cleaning, multiple positives and negatives were recorded for the same exam photo. After data cleaning, it was possible to conclude that there were 521 negative samples for all infections, meaning they were healthy patients which implied they needed to be excluded from the dataset. Of the 2064 remaining exam cases, 1580 had positive results for bacteria, 490 for fungi and 298 for amoeba, which meant the multiple infections remained present. Doing an in depth analysis, there are 1176 of only bacteria, which represents 56.98% of the dataset, 277 of only fungi, accounting for 13.42% of the total samples and 207 of only amoeba, 10.03%. Regarding multiple infections, there are 213 cases of combined infection of bacteria and fungi 10.32% and 191 of combined cases of amoeba and bacteria 9.26% (Table 4.2).

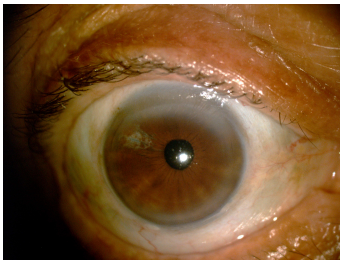


Figure 4.1: Bacteria Dataset Example.



Figure 4.2: Fungi Dataset Example.



Figure 4.3: Ameba Dataset Example.

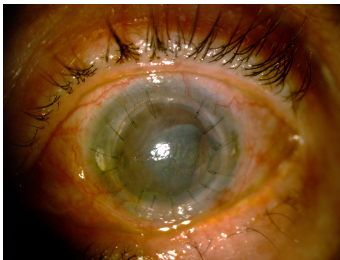


Figure 4.4: Bacteria Dataset Example.



Figure 4.5: Fungi Dataset Example.

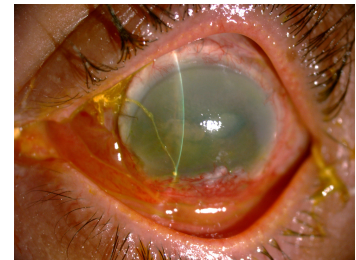


Figure 4.6: Ameba Dataset Example.

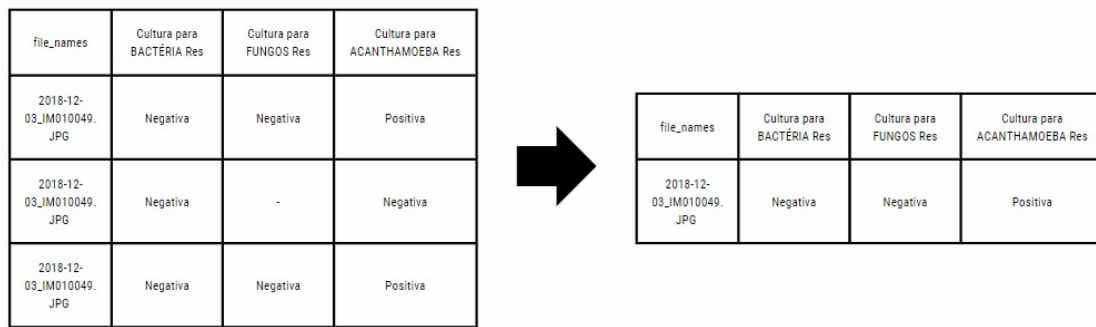


Figure 4.7: Data Cleaning.

Table 4.2: Number of Cases and Percentages for Combinations of Infections.

BACTERIA Res	ACANTHAMOEBA Res	FUNGOS Res	Cases	Percentage (%)
Positive	Negative	Negative	1176	56.98%
Negative	Negative	Positive	277	13.42%
Positive	Negative	Positive	213	10.32%
Negative	Positive	Negative	207	10.03%
Positive	Positive	Negative	191	9.26%

4.2.2 Feature Analysis

The dataset was then analysed regarding the data distribution of each individual feature. Some features were not necessarily analysed as a whole but rather processed before analysis. The features that were considered relevant, after discussion with clinicians, are the sex, the eye, the age, the combination of tests asked, the material origin, the existence of a pre-existing surgery and medicine, the time of evolution as well as the result of the tests that were already explored in the past section. The feature distribution analysis will be done on the cleaned dataset comprised of 2064 cases since that is the adequate dataset to achieve the goal of predicting the type of keratitis the patient has.

4.2.2.1 Sex

The sex distribution across the infected patients of the dataset can be seen in Figure 4.8. It is possible to see that there are 1203 cases of patients belonging to the Male sex (58,28%) and 861 belonging to the Female sex (41,72%), so we can conclude that the dataset is balanced, regarding sex.

4.2.2.2 Age

Looking at the age distribution and performing a Kernel Density Estimation of Age (KDE) (Figure 4.9) it is possible to observe a high peak of ages from 20-40 and 40-65 which, according to literature, can be related to the two most common sources of infection, bad usage of contact lenses

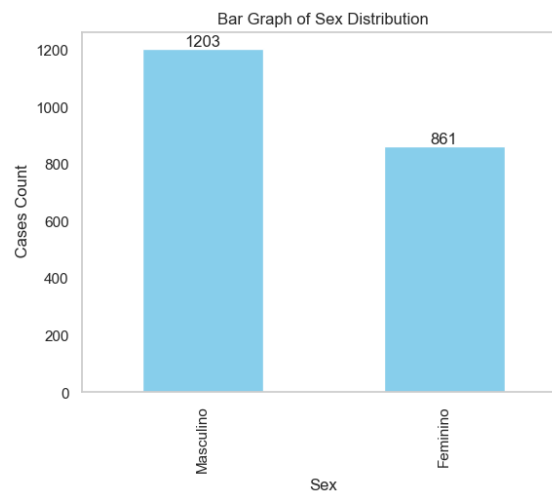


Figure 4.8: Distribution of the sex parameter in the dataset.

and agriculture work. When trying to group them by age, for a deeper analysis, according to the clinicians and the literature on the disease, 4 major bins were selected, 0-18 (label 0), 18-40 (label 1), 40-65 (label 2) and more than 65 (label 3). After that implementation, it was possible to get a better understanding of the age dispersion of the dataset (see Figure 4.10). The lack of data for patients under 18 is clear, with just 54 cases (2.62%). Also, as previously mentioned, the majority of cases exist in the 40-65 years old, with 813 (39.39%) and the 18-40 years old with 744 (36.05%), proving again that this dataset is a good representation of the infection profile.

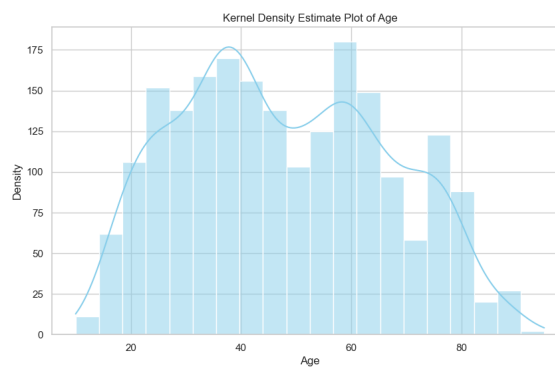


Figure 4.9: Distribution of age in the dataset.

4.2.2.3 Eye

Looking at how many exams exist of each eye, in Figure 4.11, it can be easily denoted that there are approximately 1000 images of each eye, so the dataset is well balanced in that matter.

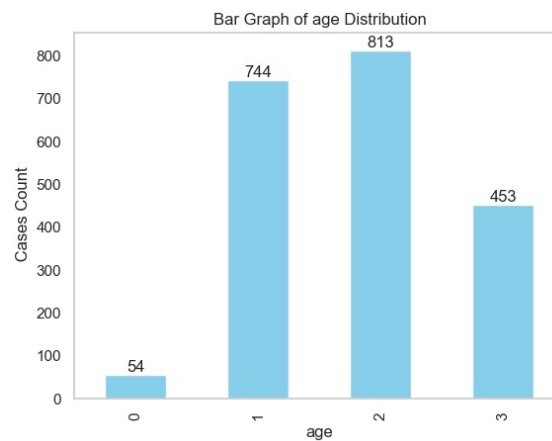


Figure 4.10: Distribution of the age bins in the dataset.

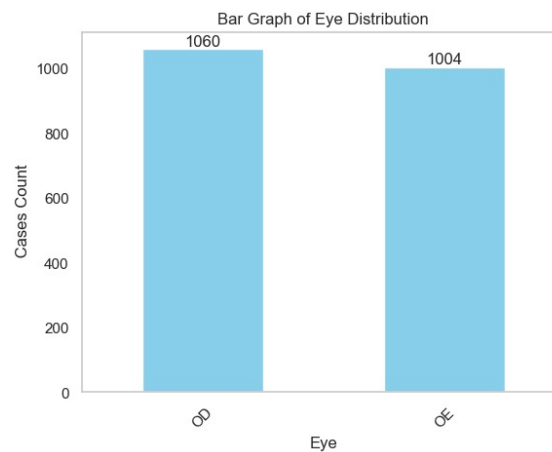


Figure 4.11: Distribution of the eye in the dataset.

4.2.2.4 Ordered Tests

When the patient reaches the clinic, in case there is a suspicion of infection, the Clinician asks for the laboratory tests. There is the possibility of asking for 1,2 or 3 tests at once, depending on the etiologies of suspicion. Because bacteria is the most common source of infection, it is also the most commonly asked test, with 1981 asked tests out of the 2064 cases where tests were asked (see Figure 4.12). Looking into the fungi infection, it is also commonly asked (1954 tests) even though the positive cases are not as frequent as the bacteria ones (see Figure 4.13). Regarding amoeba, it is the least common source of infection according to the literature, which is reflected on the amount of tests asked by the doctors. Looking into the dataset there are 1067 tests asked for acanthamoeba (see Figure 4.14).

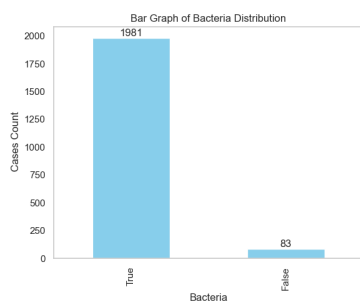


Figure 4.12: Distribution of the ordered tests of Bacteria.

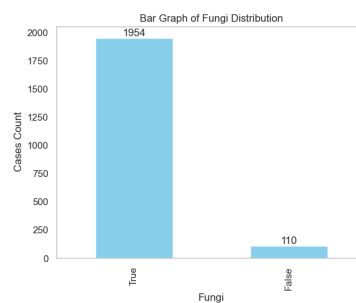


Figure 4.13: Distribution of the ordered tests of Fungi.

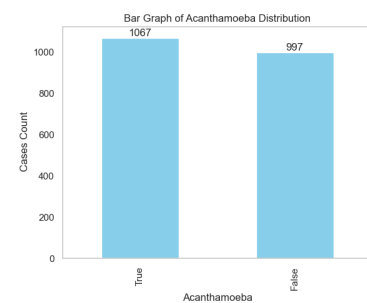


Figure 4.14: Distribution of the ordered tests of Acanthamoeba.

4.2.2.5 Surgery and Medicine

Because during eye surgery there is a high exposition of the eye to microorganisms, it is an important feature to have in consideration. Of the 2064 exam cases, 377 have a pre-existing surgery (see Figure 4.15). Looking into the surgeries that exist in the database we can clearly see that the cornea transplant is the most common surgery, with values reaching 179 reported surgeries.

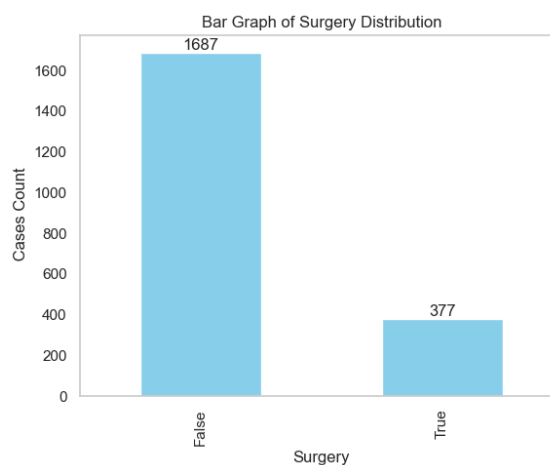


Figure 4.15: Distribution of surgery in the dataset.

In the dataset, pre-existing medicines are also recorded. They can be a great source of information regarding past infections of the patients with Keratitis. 968 have a past medicine, indicating that probably in case the result of the test is positive, it is not their first infection (see Figure 4.16).

4.2.2.6 Material Origin

Although the dataset is all collected in São Paulo, Brazil, there are different types of services where the data was collected. That discrepancy highlights the social economical aspect of the disease

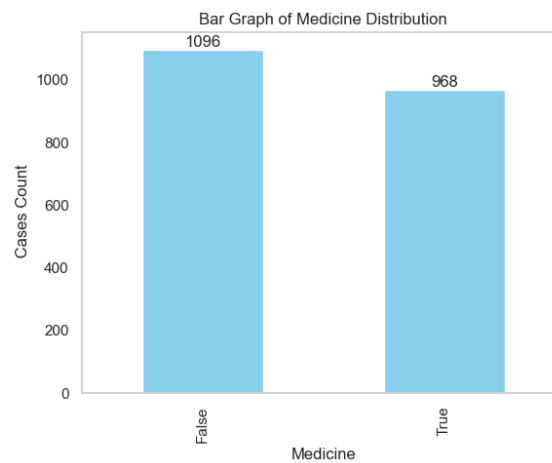


Figure 4.16: Distribution of medicine in the dataset.

with only 2 cases being registered in private clinics. The majority of the cases are registered in urgent care (1456), enhancing how IK affects the patients' life (see Figure 4.17).

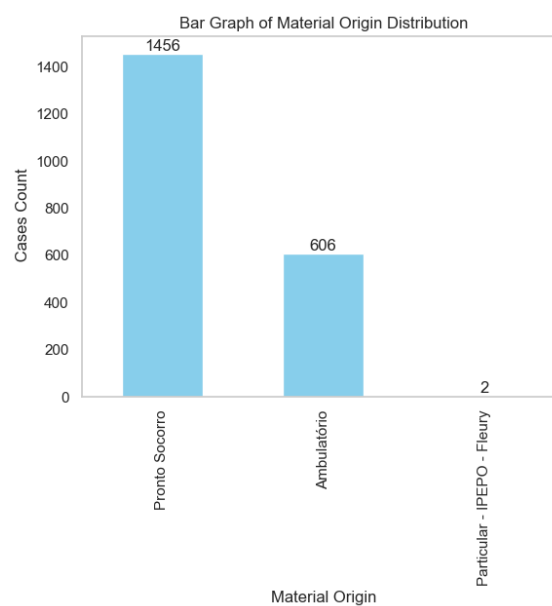


Figure 4.17: Distribution of the origin of the material in the dataset.

4.2.3 Extracted Features

To get a better understanding of the time effect in the dataset, the year, quarter, and month were extracted from the "exam_date" feature. Starting on the month it was possible to denote a peak of infections from June to October, grouping the months by quarters we could see a maximum for the third quarter with 680 cases (32,95%). Followed by the second, fourth and lastly the first. When

thinking about the logic reasoning behind it, it makes sense since those are the trimesters with the most agriculture work, and therefore where people are more prone to get infected by parasites (see Figures 4.18, 4.19). When analysing the year of infection we can see that there is data from 2018 to 2021, with the maximum existing for 2021 with 740 exams (see Figure 4.20).

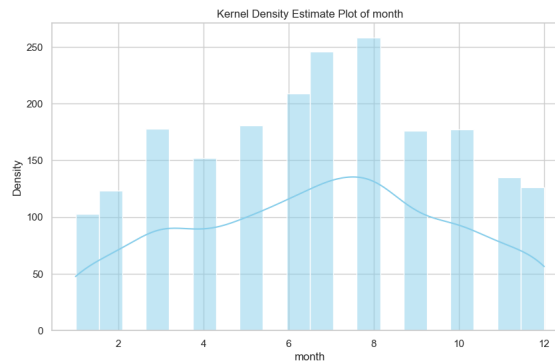


Figure 4.18: Distribution of the month of the photograph in the dataset.

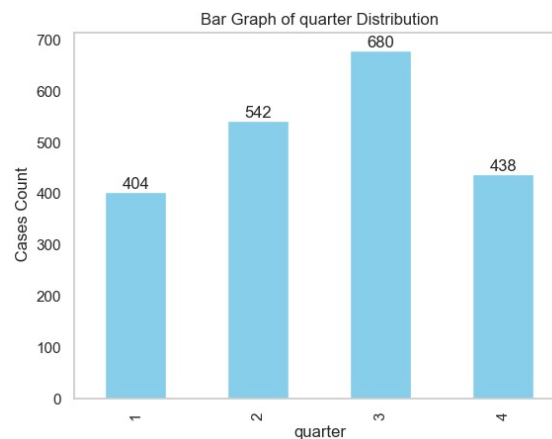


Figure 4.19: Distribution of the quarter of the year in the dataset.

4.3 Correlations

After the analysis of each relevant feature value it was important to understand potential correlations that could exist between the features and the classification of the type of Infectious Keratitis. To do so, three analysis were executed.

Firstly, to assess the influence of the period of the year on the type of infection, a correlation was performed between those features. To allow comparisons between infections and because the dataset is so imbalanced infection-wise, normalisation by the cases of each infection was implemented (see Figure 4.21). Looking at the matrix it was possible to understand a general increase of cases across all etiologies (even the combined ones) in the third quarter of the year,

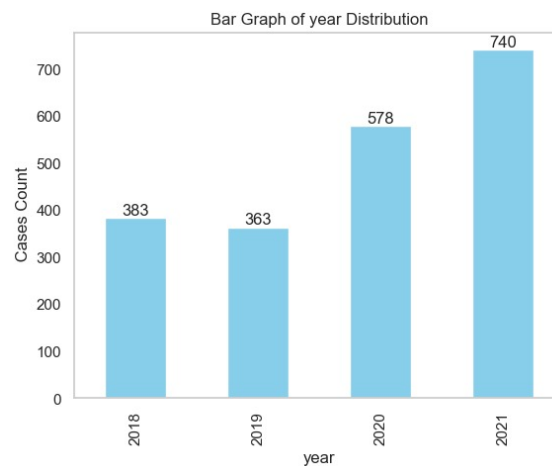


Figure 4.20: Distribution of the year in the dataset.

supporting what was seen before. It is also plausible to conclude that fungi infections seem to be frequent in the second and fourth quarters of the year.

Looking now into the influence of the year using the same process of normalisation we can see a clear peak of infections of ameba and combined infection of bacteria and fungi in 2021. We can also denote a peak in 2018 for the combined infection of bacteria and amoeba, which can represent a possible outbreak (see Figure 4.20).

Lastly, analysing the correlation between sex, age, eye, quarter, pre-existing surgery with each classification (see Figure 4.23), we can highlight apparent negative and positive correlations between the sex and both fungi (-0.22) and ameba (0.20) infection. In addition, when focusing on the age, there seems to be also a negative and a positive correlation between the age group (the 4 labels metrics) with both the fungi (0.20) and the ameba (-0.37) infection. When working with this dataset the correlations, should be taken into account in order to avoid bias. It is also important to denote that there is a high correlation (both positive and negative) between the different types of infections.

4.4 Conclusions

The extensive analysis performed on the dataset allowed the conclusion that even though the dataset is quite extensive in terms of exam cases, only 2,064 of those can be used due to the lack of medical information and image files in the rest. In addition, it is also possible to state that the dataset is highly imbalanced in terms of infections and that in order to properly use it, corrections for that will need to be implemented. Lastly, it was also inferred a possible correlation of sex and age with the type of infection.

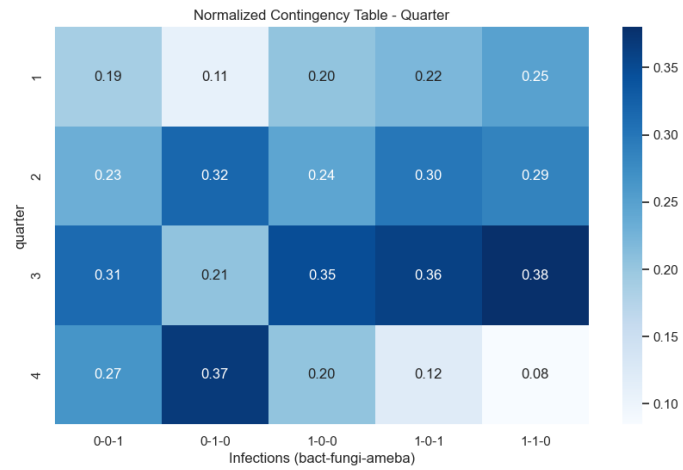


Figure 4.21: Normalised Correlation of the quarter and type of IK.

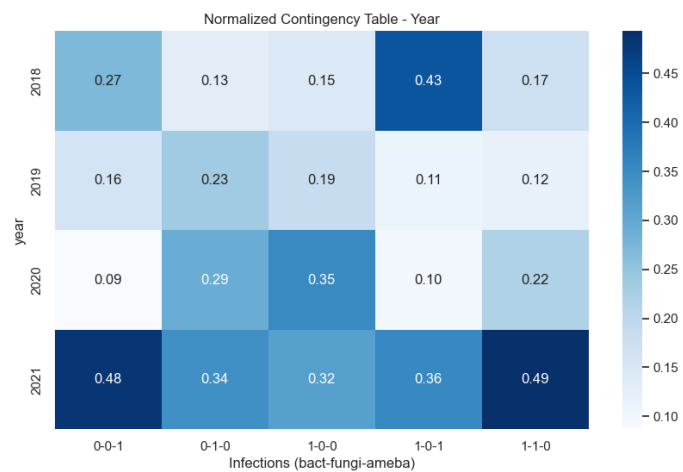


Figure 4.22: Normalised Correlation of the year and type of IK.

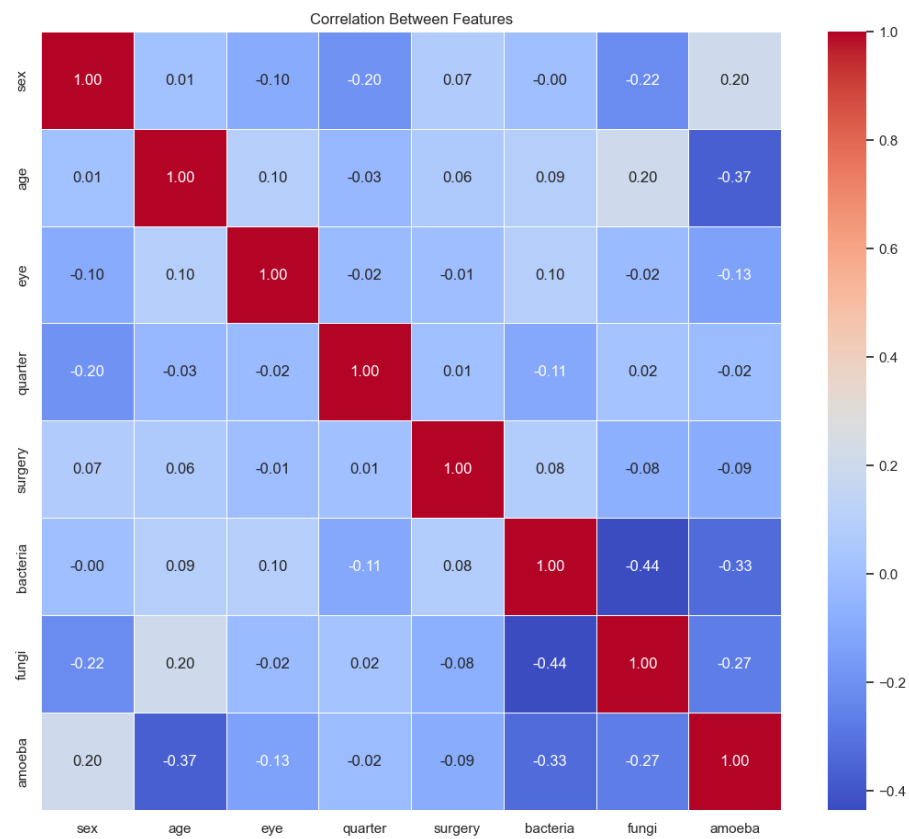


Figure 4.23: Correlations between dataset features.

Chapter 5

Clear Vision, Clear Methodology: Proposed Framework

5.1 Proposed Workflow

In order to achieve the proposed goals, different approaches were defined:

- **Baseline Approach:** A baseline binary deep learning model that predicts the infection based on the eye corneal photograph for one infection at a time. To get a diagnose for all infections, it requires three models.
- **Multitask Approach:** A deep learning solution that, with the eye corneal photograph as input, predicts the infections for all types at once, meaning that it not only reduces runtime and computational costs but also allows for multiple-infection diagnosis. Two different architectures were defined and tested.
 - **First Multitask Approach -V1:** Three different classification layers, each responsible for a binary diagnosis.
 - **Second Multitask Approach -V2:** One classification layer responsible for the three classifications.
- **Sex Prediction:** Since there is an apparent correlation between sex with the disease outcome, a deep learning system to understand if the sex could be identified from the eye photograph was implemented. The performance of this model was used to evaluate the importance the diagnosis model was giving to sex-related features of the eye, helping mitigate potential biases [120].
- **Age Prediction:** Since there is an apparent correlation of age with the disease outcome, a deep learning system to understand if the age could be identified from the eye photograph was implemented. The performance of this model was used to evaluate the importance the diagnosis model was giving to age-related features of the eye, helping mitigate potential biases.

5.2 Data Preparation

In order to save time and optimise the following steps, a data preparation system was implemented to resize and enlarge the dataset. As described before, the used dataset only included 2,064 out of the 24,691 initial exam cases. Since that significantly decreases the dataset, a strategy to flip every image along the y-axis was developed so that every eye case has a “left eye” and “right eye” photo. That not only increases the total dataset to 4,128 images but also allows the model to be equally trained to recognise both eyes, allowing it to perform equally for both cases. A new CSV file was built to account for the transformations of the version where to every image a version of its file_name with “flipped” added in the end was created.

Secondly, there was the need to standardise the input shape of the data, with 224x224 as the chosen size. The usage of 224x224 images as input in many deep learning models is a well-established practice influenced by historical precedence, computational efficiency, pre-trained model compatibility, and architectural considerations.

- **Historical Precedence:** The popular ImageNet [121] dataset, used in training many renowned models like AlexNet, VGG, ResNet and DenseNet employs 224x224 images. These models were originally designed to match the requirements of this dataset [122].
- **Computational Efficiency:** Smaller image sizes like 224x224 reduce the computational load during training and inference, requiring less memory and computational power while still retaining sufficient information [123].
- **Pretrained Models:** Many models are pre-trained on large datasets such as ImageNet using 224x224 images. Fine-tuning these models on new tasks typically uses the same input size to leverage the pre-trained knowledge effectively [124].
- **Pooling Layers:** Popular convolutional neural network (CNN) architectures use pooling layers to downsample feature maps. Starting with 224x224 images allows several pooling layers while maintaining adequate spatial information [125].

Because of all of this, the chosen image size was 224x224 and so a script was run to store every image (flipped and non-flipped) in its 224x224 format in a new folder.

5.3 Pre-processing

5.3.1 Dataset and Image Preparation

Generally, this phase is characterised by preparing the data for model prediction. After reading the CSV into a pandas DataFrame, rows with healthy patients were dropped. In addition, the dataset cleaning procedure mentioned before was implemented so that each exam photo was associated with only one csv row that summarised all the classifications, forming the 2,064 exam-cases dataset.

The values for each classification column: the result of bacteria test, the result of fungi test, the result of acanthamoeba test, age and sex; were converted into numerical labels using the rules present in Table 5.1.

Table 5.1: Transformations into numerical labels.

Column	Original Data	Transformed Data
Infection Test Result	Negative / Positive	0 (Negative), 1 (Positive)
Sex	Male / Female	0 (Male), 1 (Female)
Age	Continuous Age Values	0 (0-17), 1 (18-39), 2 (40-64), 3 (65)

Finally, the features (X) and target variables (Y) (Table 5.2) for each classification task were selected accordingly. It is important to note that for the baseline models and the sex and age models, the classification column is a single value. However, for the multitask approaches, the target variable (Y) is a vector containing the label for each task. Since sex and age factors will be relevant in future analyses, all of the infection related models receive the sex and age bin information associated with that filename. Nevertheless it is important to highlight that the sex and age are not part of the models' inputs, instead they are used to later analyse the performance of the model in a specific subset.

Following the initial data preparation, the next step involves extracting image filenames and converting the infection, age, and sex labels, making them ready for the subsequent operations that include data augmentation and model training.

This meticulous preparation ensures that the data is clean, well-labeled, and appropriately formatted, laying a solid foundation for effective model training and accurate predictions. By organizing the data in this manner, the robustness of the resulting machine learning model is significantly enhanced.

Table 5.2: The inputs of the different models.

Model Name	X	Y	Additional info
Baseline Bacteria	Filenames	Results for Bacteria test	Sex and Age
Baseline Fungi	Filenames	Results for Fungi test	Sex and Age
Baseline Amoeba	Filenames	Results for Amoeba test	Sex and Age
Multitask v1	Filenames	Results for all tests	Sex and Age
Multitask v2	Filenames	Results for all tests	Sex and Age
Sex	Filenames	Sex	-
Age	Filenames	Age	-

5.3.2 Data Augmentation of Images

During this process, it is also when transformations are randomly applied to the image data to enhance the robustness and generalisation ability of the model. The first transformation makes sure the previous resizing process was done and protects the model from an eventual error of the previous, by resizing every image to (224,224). Random rotations of up to 20 degrees were applied

to introduce variability [126]. Subsequently, random color transformations (brightness, contrast, saturation, and hue adjustments) were applied with parameters set to 0.1, as per the literature [104]. This simulates different lighting conditions and color variations, improving the model’s robustness to real-world scenarios and reducing overfitting [127]. Additionally, random vertical flips with a probability of 0.5 were included.

Lastly, Gaussian blur was applied to the image with a kernel size of 5×5 and a sigma value ranging between 0.1 and 2.0. The Gaussian filter, known for its “Mexican hat” shape, is a special case of weighted mean filters. It is particularly effective at reducing noise. As the kernel size increases, noise-induced variations in the image become less pronounced, leading to overall smoothing. This results in a higher differentiation between central and lateral values and effectively reduces noise. Conversely, increasing the sigma value reduces the differentiation between central and lateral values, leading to greater smoothing (see Figure 5.1). By adjusting these parameters, we can control the degree of smoothing applied to the image, balancing noise reduction with the preservation of important features. This can help the model become more robust to variations in image sharpness and noise levels, like the patient’s unique iris features, making the classification more invariant to the patient’s characteristics [128].

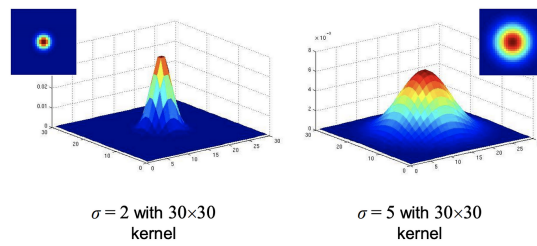


Figure 5.1: Gaussian Filter variance with sigma from [20].

Then, the image was converted to a PyTorch tensor. PyTorch tensors are the standard input format for models in PyTorch, enabling GPU acceleration and compatibility with other PyTorch functions. All the models were trained using this library [129]. Lastly, normalisation with $\text{mean}=[0.485, 0.456, 0.406]$ and $\text{std}=[0.229, 0.224, 0.225]$ was performed to adjust the values to the mean and standard deviation of the ImageNet dataset. That is because, since the models are all pre-trained on this dataset, it is important to initialize them with the same weights. This standardisation of input data ensures that pixel values are centered around zero with a standard deviation of one, which helps the model converge faster during training and can improve performance by ensuring consistent input data [130].

These transformations collectively enhance the quality and diversity of the training data, which are critical for several reasons; the impact of these transformations in the dataset can be seen on Figure 5.2. Data augmentation, through rotations, flips, filters and color adjustments, artificially increases the diversity of the training dataset, exposing the model to a wider variety of scenarios, which helps it learn more generalised features. This approach prevents overfitting, as the randomness introduced means the model is less likely to memorise the training data and more likely to

learn robust patterns that generalises well to unseen data. Furthermore, improved generalisation results in a model that becomes more resilient to variations in image appearance, such as different angles, lighting conditions, and colours, making it more applicable to real-world situations where such variations are common.

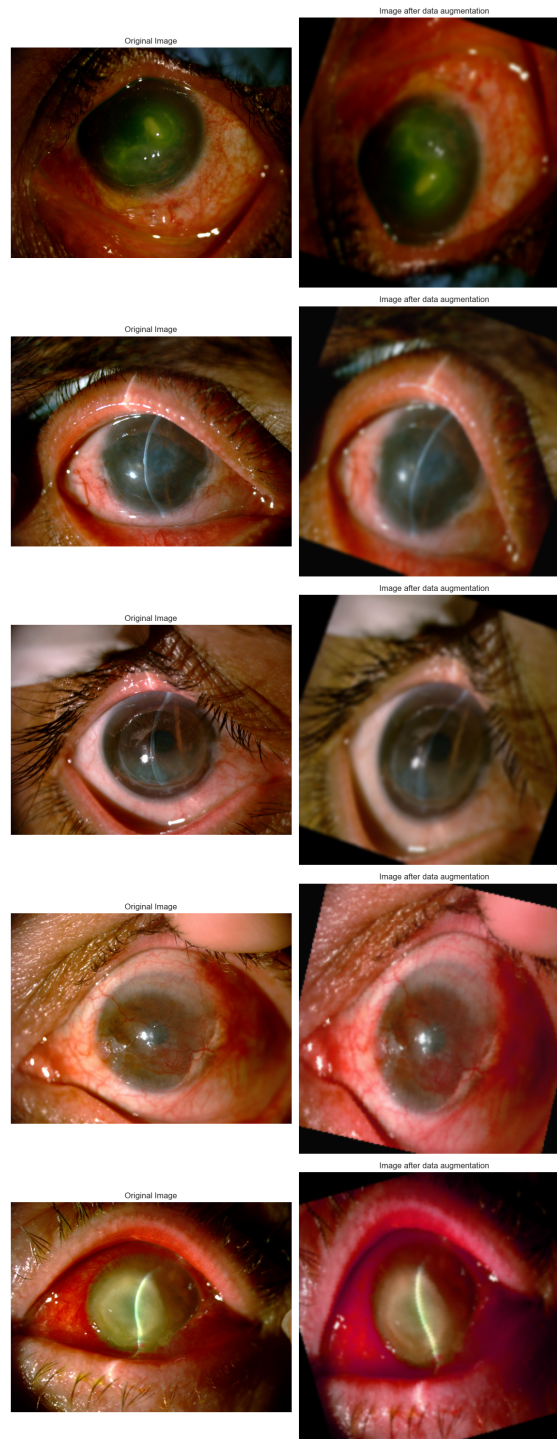


Figure 5.2: Examples of the pipeline of image transformations applied to the dataset.

5.4 Model preparation

5.4.1 K-fold strategy

To later validate the reliability of the results of the pipeline, a k-fold adaptation was implemented.

K-fold cross-validation is a statistical method used to evaluate a machine learning model's performance and generalised. In this technique, the dataset is divided into K equal-sized subsets, or folds. The model is trained and validated K times, each time using a different fold as the validation set and the remaining K-1 folds as the training set. This iterative process ensures that every data point is used for both training and validation. The performance metrics from each iteration are averaged to provide a comprehensive estimate of the model's effectiveness, helping to reduce both bias and variance [131]. An adapted version of this technique was created since it was important in our case to maintain the validation set to later select the best model.

Before starting the division in folds, the flipped version of every image was excluded from the dataset to prevent the test set from containing the flipped version of an image already seen in the train set or vice-versa, preventing data leakage. These images were later re-added after the sets were created. Because some of the models included multi-label classification and others were single-labeled, two different approaches were developed. Nevertheless, the idea behind them was the same and can be seen in Figure 5.3.

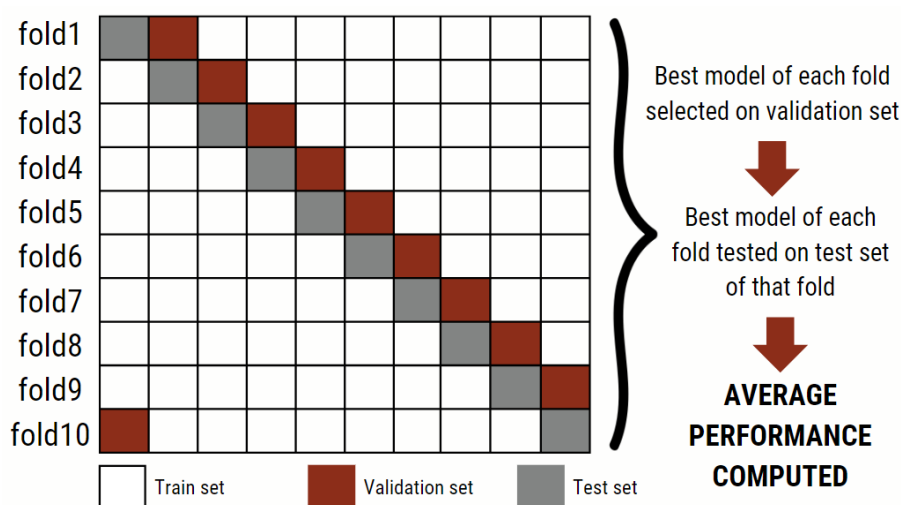


Figure 5.3: Adapted 10-fold strategy.

5.4.1.1 Age, sex and infection baseline

For the single-label option, a Stratified k-fold version was used, since it preserves the class distribution, making it ideal for problems with imbalanced datasets [132]. That allows us to correctly separate one different set as the test set for each fold, making the test set have 10% of the dataset. Because we intended to retain the validation set for later validation of the best model, the train set

group was divided using the train/test split technique with a ratio of 0.1/0.9. That way, a balanced validation set with 10% of the total dataset was created and a train set with 80% of the dataset was also formed. This approach allowed every group to be balanced and every validation and test set to be correctly excluded from the model train part, preventing data leakage.

5.4.1.2 Multitask v1 and v2

For the multi-label case, we need to ensure that every class and combination of classes was properly balanced and that took into consideration the “B-F-A” format and not only the individual classification of each case. To do so, the multilearn library [133] was explored. After careful analysis, Iterative Stratification and iterative train test split were used since they did the same as the previously explained strategy, but instead of taking into account only the class distribution it takes into consideration the combination of all infections’ distribution.

After all this was done, the names of the files corresponding to each group were accessed, and the flipped version of that image was added to that group, allowing the prevention of data leakage while ensuring these images were included in the model training and evaluation. After that, the subsets for, train, validation and test were created for each fold. To load each image, a batch size of 16 was chosen without any sampler, meaning that 16 samples would be loaded at a time, which provides a more stable gradient estimate compared to processing single samples as well as providing faster convergence during training [134].

5.4.2 Class Weights

To conquer class imbalance, the weight each positive sample of a certain label would need to have for the dataset to be considered balanced was computed. For each label, this weight was computed for the positive and negative class using the Equation 5.1, where the number of total samples of the dataset (N_s) is divided by the number of classes (N_c) and the number of samples with that label (N_{cs}). In order to get a single value for each label the positive was divided by the negative weight for that class, this value represents how much each positive sample of that class should weigh for the dataset to be considered balanced. For single-task decisions, this was only performed once, for multi-label cases this process was repeated for every infection, meaning, in the end, we had an array with the weight each positive sample of each infection would need to have for the dataset to be balanced [135].

$$Classweights = \frac{N_s}{N_c \times N_{cs}} \quad (5.1)$$

5.4.3 Optimiser

Adam optimiser has been tested with various learning rates and weight decays and the optimal value was decided by observing the minimal validation loss value that was found, as well as in which epoch it happened.

5.4.4 Model Selection

With the data loaded, the optimiser defined and the class weights computed, the model architecture needed to be built. Looking at the state of the art, it was easy to conclude that ResNet50, VGG16 and DenseNet121 were strong candidates for the model. Because deep learning models are both time- and computational-costly, we decided to create a subset of 50 images in which several epochs would be run and the duration recorded. The model with a lower runtime and better performance for the subset was the selected one.

5.4.5 Model Architecture

The backbone of the framework was inspired by the previously defined models pre-trained on ImageNet, but 3 layers were added before the fully connected layer to improve the models' performance and reduce overfitting.

Firstly, an average pool 2D layer was included with a kernel of 3×3 , stride of 2×2 and padding of 1, then a batch normalisation layer and lastly a dropout layer with a probability of 0.3.

Depending on the type of model that was being developed, different FC layer architectures were developed (see Figure 5.4).

- **Sex and Infections Baseline:** For models that predicted between two classes only, meaning that the decision was binary, a simple Fully connected layer with just one node was developed.
- **Age:** In the age case, since there were 4 age groups, the fully connected layer had 4 nodes, each predicting the probability of one age group.
- **Multitask_v1:** The multitask_v1 case is composed by 3 parallel fully connected layers, each layer has one node that makes a binary decision for that specific infection. The combination of the output of all fully connected layers represents the final classification.
- **Multitask_v2:** The multitask_v2 method uses a one fully connected layer with 3 nodes, each predicting the probability of one etiology, again the global result is obtained by the combination of all nodes.

Since the loss function that we were going to use already had a sigmoid activation function included, no activation function was used.

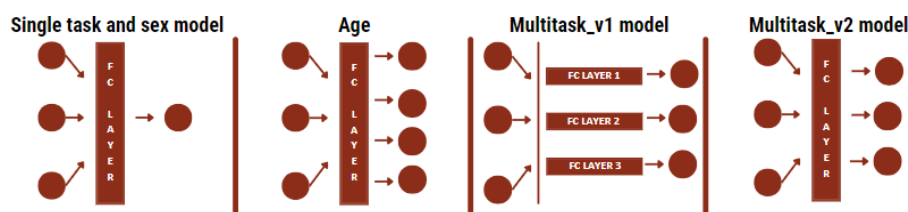


Figure 5.4: Fully Connected Layers of each model.

5.4.6 Model training and validation

Several epochs were tested for each model (100,200 and 300), and the optimal case was defined for each.

To boost the models' performance and avoid overfitting, the backbone was frozen for the first 10 epochs and later unfrozen. When we freeze specific layers in a neural network, we don't back-propagate the gradients to update their parameters. These layers usually include parameters that have been pre-trained on a big dataset. By doing so, we may take advantage of the existing information contained in these layers, which is especially useful for applications with little training data or when the task starting knowledge is critical. In contrast, unfreezing entails enabling these originally frozen layers to be updated with the remainder of the model during later training sessions. This stage allows the model to fine-tune its representations, fitting them more precisely to the requirements of the target task or dataset. This iterative process balances the benefits of pre-trained knowledge with targeted adjustments, yielding more effective and efficient models across diverse applications in deep learning [136].

Since the goal was classification, Binary Cross-Entropy was used as the loss function, to account for dataset imbalance, all losses were fed with the class weights previously calculated. It is also relevant to emphasise that the validation and train loss were tracked in order to evaluate the models' performance. Since age was not a binary task, a standard Cross-Entropy function was used, to which the weight of each class was fed. To avoid overfitting and especially since our dataset was considered small and very variant imagewise and so very prone to overfitting, the model that was implemented in the test set was always the model that had the lowest possible loss in the validation set, this was followed for all the seven prediction models. Throughout training, models are iteratively improved, but the training process can be affected by various factors such as overfitting or fluctuations in learning rates. The last model might not always be the best because it could have overfitted to the training data, meaning it performs well on seen data but poorly on unseen data. By selecting the model with the lowest validation loss, you ensure that you are choosing the model that performed best on a separate validation set, which serves as a proxy for unseen data. This approach helps in selecting a model that is more likely to generalise well, offering a better performance on real-world data and avoiding overfitting issues (see Figure 5.5). This technique is called "early-stopping". A comparison of the multitask and single task performance under these circumstances was performed and, after that, a few improvements were implemented to the multitask framework.

A combined loss based on binary cross entropy was applied where not only the class weights but also the medical cost of each infection was taken into account. This idea was suggested by the Clinicians (Dr. Kase and Dr. Nakayama), involved in the project and it meant that the cost of misidentifying a high-cost infection would be higher than that of misidentifying a low-cost infection and that our prediction should account for that. That is because since IK is common in Low and Middle countries, the patients often struggle to pay for the medicine and so, the misclassification of etiologies in which the price of the medicine was significantly high was penalised.

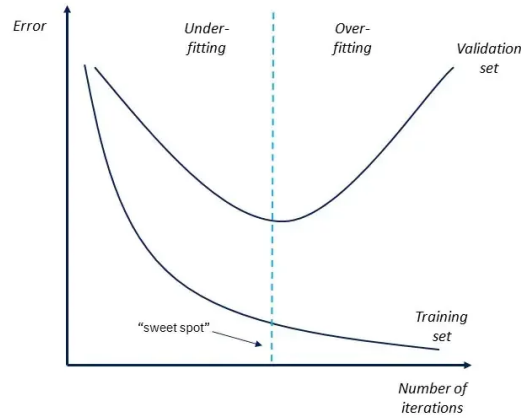


Figure 5.5: Loss optimal from [21]

To achieve this goal and based on the medicines used, duration of treatment, and flasks used per month, present in Table 2.1, the average cost of each infection was computed. To do so, the prices of the medicines used in each infection (R\$ 45,2 for BK, R\$ 203 for FK and R\$ 95.5 for AK) were multiplied by the number of flasks of each medicine used and the maximum duration of the treatment in months. The values were then divided by the cost of all infections. The values obtained were considered the hospital costs. After that, the combined loss was calculated using Equation 5.2. The performance of this strategy was then properly evaluated and the last model improvement was implemented.

$$Totalloss = 0.8 * (class_weights * BCE_loss) + 0.2 * (Hospital_weights * BCE_loss) \quad (5.2)$$

This last approach consisted of an adaptive threshold solution for considering a certain probability of a prediction. Typically, a threshold of 0.5 is used to convert model output probabilities into predictions, where probabilities above 0.5 are considered 1, and those below 0.5 are considered 0. However, in multi-label problems, this value often needs adjustment. We implemented Youden's J statistic (Equation 5.3) for each task to address this. This statistic measures a classifier's ability to make informed decisions beyond random guessing. "Informedness" refers to how well the classifier distinguishes between classes based on the available data. Higher J values indicate more reliable predictions and better-informed decision-making. To find the optimal threshold that maximized this metric, we calculated Youden's J for every threshold value, comprising the ROC curves of each task. The threshold that maximized J was then considered the optimal classification threshold for that infection (see Figure 5.6) [137].

$$J = TruePositiveRate - FalsePositiveRate \quad (5.3)$$

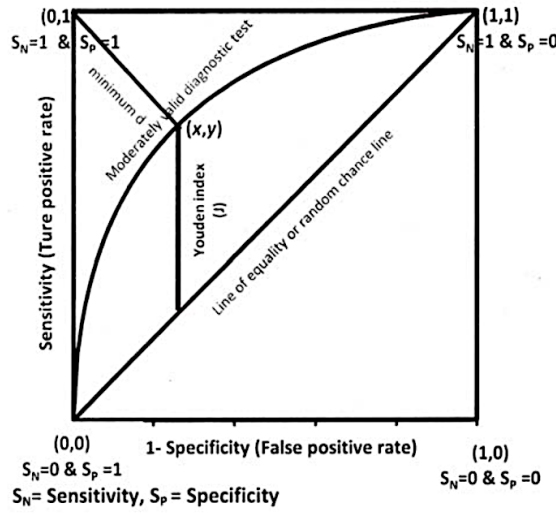


Figure 5.6: Youden's J Optimisation from [22].

5.4.7 Performance evaluation

5.4.7.1 General

To properly assess the model performance the loss evolution of both the training and validation set were evaluated, as well as the F1-Score, Recall, Precision, AUROC, Accuracy, Balanced Accuracy and Mean Absolute Error (for the age prediction) on the validation set to get an understanding of the model evolution.

Regarding the test set, the same metrics were computed, and in addition, the confusion matrices and the F1 score and Recall for each class within the same label were also computed. In the multitask models, the joint confusion matrix of all infections was also retrieved. To ensure stability, the average confusion matrix across all folds was also found, which gives us a proper sense of the models' performance. The rest of the metrics were also computed and stored across all folds but were treated with a statistical analysis that will be further analysed. Lastly, to understand the TPR and FPR, the ROC curves were plotted.

To ensure and validate the models' diagnosis, saliency maps were created for every fold and one example of each class's positive and negative case. Saliency maps are visual techniques that enable the identification of sections of an input that has the most influence on a model's prediction. They are created by calculating the gradient of the model's output concerning the input, which indicates how sensitive the output is to changes in each input pixel. Saliency maps provide several benefits, including improved interpretability by revealing which parts of the input the model is focusing on, aiding in debugging by identifying irrelevant features that the model may be overfitting to, and improving model explanations [138, 139]. All these metrics were evaluated for all the models on every fold to ensure their robustness and effectiveness; it is also relevant to add that for the multitask approaches the metrics were all computed by task with the exception of the multitask matrix to ensure a good comparison to the baseline approach.

5.4.7.2 Patients' Information subsets

To evaluate the influence of the age group and also the sex in the prediction of the etiology, the F1-Score, Recall, Precision, AUROC, Accuracy, Balanced Accuracy, and Mean Absolute Error were computed by age and group bin, for each task, meaning that a comparison of performance could be done in between sexes and ages for a specific task and metric, allowing for a deeper understanding of the influence of this factors in the classification. The goal behind it was to understand if the model was using identity features when deciding on a diagnosis. That behaviour, on a balanced dataset, would be clear if different subsets of the same category (for example both sexes) had very different performing metrics. With that in mind, will be later referenced by their names:

- General: Formed by every infected patient of the test set
- Female Sex: Formed by patients of the Female Sex (the sex that has the label 1)
- Male Sex: Formed by patients of the Male Sex (the sex with the label 0)
- 0-18 yo: Formed by patients 0 to 18 years old (the age group with label 0)
- 18-40 yo: Formed by patients 18 to 40 years old (the age group with label 1)
- 40-65 yo: Formed by patients 40 to 65 years old (the age group with label 2)
- +65yo: Formed by patients more than 65 years old (the age group with label 3)

5.5 Statistical Analysis

5.5.1 Confidence Intervals

To better summarise the performance of each metric and task across all folds, facilitating the analysis, a 95% normal confidence interval was calculated for the results of each metric and task. This involved grouping all 10 values of the metric and task across the folds to determine the interval. Using this statistic method, we could confidently state that our model in unforeseen data would achieve a performance that would be in between a certain interval, making it reliable. A normal distribution with a 95% confidence level refers to the range within which we expect 95% of the data points or observations to fall, assuming the data follows a normal (Gaussian) distribution. This means that there is a 95% probability that the true population parameter lies within this calculated interval [140]. This methodology was applied to the age and sex prediction but also to the infection prediction models. In the last ones, not only the general subset of data was analysed, but also all the other ones, meaning it was possible to get the confidence interval of F1-score for bacteria if the patient belongs to the Female sex for example.

5.5.2 P-values

After that, and only on the infection-prediction-related models, the statistical significance of the sex and age for the diagnosis was studied. The goal of this task was to understand how these parameters can impact the result and make proper conclusions on the need for disentanglement of patient features from the images before inputting them into the model based on that. To achieve that, for each task and metric, the age and sex subsets were grouped.

For the sex analysis, since only two independent groups were being compared, t-test was implemented. A t-test is a statistical method used to assess whether there is a significant difference between the means of two groups. To do so, it takes into account the mean ($\bar{X}_1, \bar{X}_2$), the sample variances (s_1^2, s_2^2) and the sample sizes (n_1, n_2) as seen in Equation 5.4. It is commonly applied in hypothesis testing, particularly when the data sets are normally distributed and the sample sizes are small [141]. In this case, the comparison was done for each metric and task combination, meaning 10 values of female sex and 10 values of male sex, one from each fold, were being compared. The p-value quantifies the probability of obtaining an effect at least as extreme as the one observed in your data, assuming that the null hypothesis (there is no difference between the groups) is true. If the p-value was less than 0.05 after the test, we can conclude that the null hypothesis is not true and there is a statistically significant difference between the sets.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad (5.4)$$

For the age analysis, Analysis of Variance (ANOVA) was used since t-test can only compare two groups of samples at a time. The strategy behind analysing the p-value was the same. ANOVA is a statistical method used to determine whether there are any statistically significant differences between the means of three or more independent groups. ANOVA calculates an F-statistic, which compares the ratio of between-group variability to within-group variability. To conduct ANOVA, we evaluate whether there is a significant difference between the means ($\bar{X}_1, \bar{X}_2, \dots, \bar{X}_k$) of k groups. This analysis considers the sample sizes ($n_1, n_2, \dots, n_k$), the overall mean ($\bar{X}$), the j -th observation in the i -th group (X_{ij}) and the total number of observations across all groups (N), as depicted in Equation 5.5 [142]. Similarly, the values were considered significantly different if the p-value was lower than 0.05.

$$F = \frac{MS_{between}}{MS_{within}} = \frac{\frac{\sum_{i=1}^k n_i (\bar{X}_i - \bar{X})^2}{k-1}}{\frac{\sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2}{N-k}} \quad (5.5)$$

5.5.3 Holm–Bonferroni Correction

Because all the analyses were done on the same dataset, usually a correction is done to the p-values to account for that. In our case, Holm–Bonferroni Correction was implemented [143]. This correction accounts for the family-wise error rate of conducting multiple tests on the same dataset. To correct for that, equation 5.6 was used where p is the previously calculated p-value, m is the number of tested hypothesis and r is the relative rank of that p-value when compared to the other

p-values. The obtained correction is the minimum between that result and 1. To classify the values as statistically different, the same threshold of below 0.05 was used.

$$p_{adjusted} = \min \left(1, \frac{m}{m - r + 1} \cdot p \right) \quad (5.6)$$

5.6 Contributions

The contributions regarding methodology include the comparison of 7 different classification frameworks for the diagnosis of Keratitis (Single task, multitaskv1, multitaskv2, multitaskv1 with loss, multitaskv2 with loss, multitaskv1 with adaptive threshold, multitaskv2 with adaptive threshold) as well as the implementation of a clinical loss function, tailored to benefit the good diagnosis of the most costly infections. In addition, the sex and age subset performance allowed interesting conclusions.

5.7 Summary

The methodology is structured to ensure robust data preparation, thorough model development, and rigorous statistical analysis. The initial phase involves data preparation, followed by pre-processing steps, including dataset and image preparation and various image transformations to enhance the quality and variability of the input data. The next stage focuses on model preparation, incorporating a k-fold strategy, the use of class weights to address class imbalances, and the selection of an appropriate optimiser. Detailed attention is given to model selection, architecture, and the processes of model training and validation to ensure optimal performance. The performance of the model is evaluated. Statistical analysis is then conducted to validate the model's results, involving the calculation of confidence intervals, p-values, and the application of the Holm–Bonferroni correction to control for multiple comparisons, ensuring the reliability and significance of the findings.

Chapter 6

See to Believe: Results Exploration

6.1 Optimal Architecture and Parameters

Starting with the model architecture and parameter definition, an intensive study, available on Appendix A.2, concluded that a DenseNet121 would be the best architecture to be implemented. The model would train for 200 epochs, and the optimal learning rate was considered $1e^{-6}$ with a weight decay of $1e^{-8}$.

The loss functions of all models, proving their stabilisation over epochs, can be found in Appendix A.3.

6.2 Sex Prediction

After including the previous parameters, the sex was evaluated. The weight of each sample associated with the minority class, in the case of the female sex, was 1.199.

To assess the reliability of the features chosen by the model to make the decision, and ensure that it was not based on a random factor, the saliency maps were analysed (see Figure 6.1) it is possible that, for the case of a Female diagnosis by the model, we can clearly see that most of the selected pixels are in the cornea area, suggesting that the decision is indeed based on the cornea, nevertheless and since this is a patient feature, it is acceptable and expectable that the decision is also based on other information. That is because the skin around, eyelids and eyelashes can contain relevant information for that classification.

Focusing now on the actual results of the sex model, the ROC curve can be seen in Figure 6.2 and reports an AUROC of 0.90, which means that the model performs relatively well with the task, as it can be seen by the analysis of the metrics across all folds in Table 6.1. A Balanced Accuracy of 0.7968 - 0.8256 was achieved, as well as an F1 Score of 0.7630 - 0.7961 and a Recall of 0.7496 - 0.7860, suggesting the model does not only perform well for the majority class (Male Sex) but also for the minority one. By looking at the F1 of each label for a specific fold, it is possible to see an F1 of 0.86 for males and 0.80 for females, proving the robustness of the model.

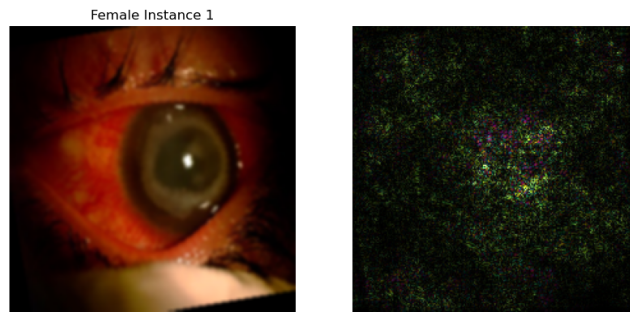


Figure 6.1: Saliency Map for a Female Patient Correctly Diagnosed.

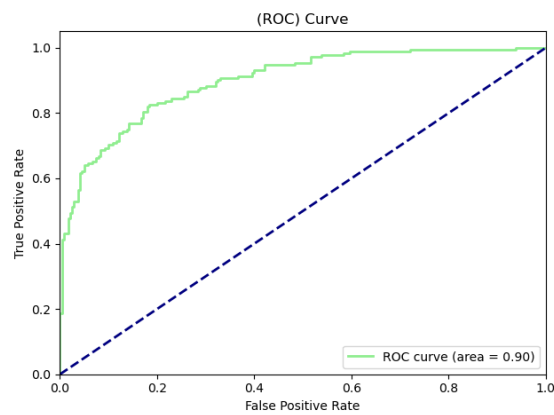


Figure 6.2: ROC Curve for one fold of the sex model.

Lastly, by analysing the average confusion matrix across all folds, it was possible to see that the model predicted 40.0 FN and 35.0 FP in 412 cases (see Table 6.2), which is still relatively high. Nevertheless, the overall prediction of the model was considered satisfactory.

Table 6.1: Sex Prediction Confidence Intervals.

Metric	Confidence Interval
F1	0.7630 - 0.7961
Recall	0.7496 - 0.7860
Precision	0.7607 - 0.8284
BA	0.7968 - 0.8256
ACC	0.8024 - 0.8343
AUROC	0.8790 - 0.8994

Table 6.2: Average Confusion Matrix for Sex.

	Predicted Male	Predicted Female
Actual Male	205.6	35.0
Actual Female	40.0	132.2

6.3 Age Prediction

Focusing now on the age prediction, the normalised class weights inputted into the Cross-Entropy function were the following: 9.5930, 0.6933, 0.6356, 1.1364.

Analysing the features used for the predictions, the saliency map for the positive and negative case of each class was analysed and they all look like the example included in Figure 6.3 for the negative classification of 18 to 40 year old bin. The map suggests that the features are mainly extracted from the cornea, with a very green area appearing there. This means that very few features are collected from the skin area, which is somewhat unexpected since visible wrinkles are clear in some images, which is often related with age.

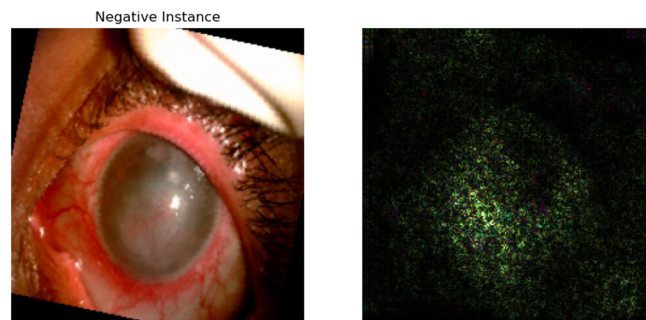


Figure 6.3: Saliency Map for a Negative patient for 18-40 years old Correctly Diagnosed.

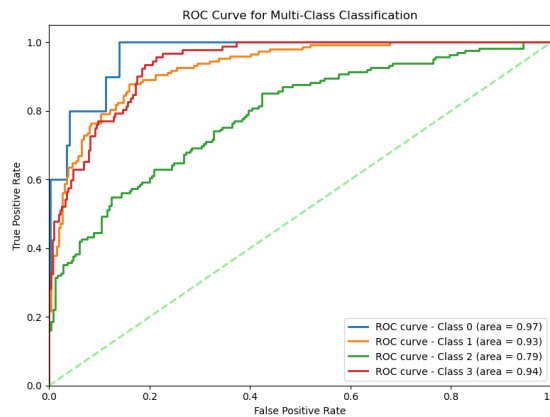


Figure 6.4: ROC Curve for one fold of the age model.

Turning the attention to the actual model performance, and analysing the ROC Curve for each task (see Figure 6.4) it is easy to denote that the AUROC is very high for almost every class (0: 0.97, 1:0.93, 2:0.79, 3:0.94) with the exception of class 2, since this class is the one with more patients and also one of the critical classes associated with the disease since it includes patients with of working age, this behaviour is somewhat expected because more severe cases of infection (and consequently less recognisable images) are probably in this subset. Another relevant take-away message from the ROC curve analysis is the box look of the class 0. That happens because

the test set cases with that label are so low (8) that the curvier look is unachievable. A deeper analysis of the metrics (see Table 6.3) allows the conclusion that the weighted AUROC is still relatively high, proving the overall effectiveness of the solution (0.8331 - 0.8624). Nevertheless, when dealing with ordinal classes like age, the metric that had the most interest is the Mean Absolute Error, looking at its confidence interval, it is possible to see that it ranges from 0.2891 to 0.3114. Even though the values are small, they are still relatively high, which means that the model is misclassifying a class as one that is further away (e.g. predict class 0 as class 3). To validate this assumption, the average confusion matrix across all folds, available in Table 6.4 was studied. Indeed, 5.7 cases of misclassification of class 1 by class 5 were found on average per fold, meaning that the approach could be more satisfactory when accounting for this. Also, if we look at the total, 91 average cases of wrong predictions were found among the 412 exam cases in an average test set, proving that there is room for improvement. Nevertheless, the current results were considered satisfactory mainly due to how infected a lot of the dataset images are, which can damage the age identity features that are present in the cornea.

Table 6.3: Age Prediction Confidence Intervals.

Metric	Confidence Interval
F1	0.7650 - 0.7940
Recall	0.7655 - 0.7936
Precision	0.7680 - 0.7973
BA	0.7329 - 0.7908
ACC	0.7655 - 0.7936
AUROC	0.8331 - 0.8624
MAE	0.2891 - 0.3114

Table 6.4: Average Confusion Matrix for Age.

	0-18	18-40	40-65	65+
0-18	7.5	2.5	0.1	0.7
18-40	2.7	121.0	19.4	5.7
40-65	0.8	21.0	120.9	19.9
65+	0.1	2.0	16.1	72.4

6.4 Baseline Disease Prediction

Starting the baseline comparison, firstly the performance of the single task model from each infection and the two baseline multitask models will be discussed.

6.4.1 Single-Task

6.4.1.1 Bacteria

Bacteria is the most common infection of the Dataset and so the normalised cost included in the loss is the lowest at 0.3078.

Looking now at the features on which the model is basing its decision, a representation of the negative diagnosis can be seen in Figure 6.5. Nevertheless, a similar looking map was seen for the positive case. Results suggest that the model is basing its decision on the cornea, especially the iris, allowing us to actually detect its shape on the feature map. This result is very positive, meaning that the model seems to find the correct targets to predict the disease. It is also important to mention that the finger in this particular image has almost no green pixels associated with it, which represents the model's robustness to noise in the image. The skin around the eye also has a low value of important features associated with it, which is very positive.

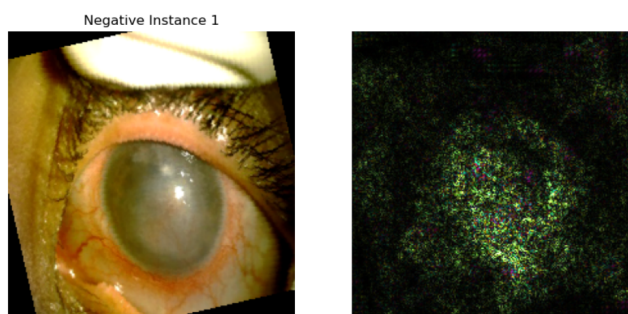


Figure 6.5: Saliency Map for a Negative Patient for Bacteria Correctly Diagnosed.

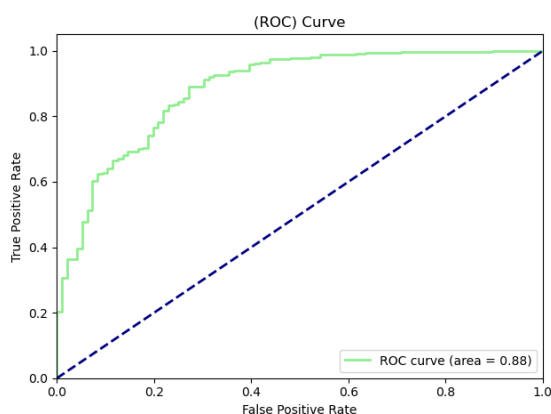


Figure 6.6: ROC Curve for one fold of baseline bacteria prediction.

Regarding performance metrics, and starting with the ROC curve (Figure 6.6), even though a high value of AUROC is achieved, 0.84, the characteristic 'elbow' looking shape is not very prominent in the actual graph, meaning that there are a lot of False positives. This behaviour was

somewhat expected from this model since the BK class has 316 samples in the test set and only 96 for non-BK. Even though this class imbalance was targeted in many ways, some discrepancies in the performance of the classification were still expected due to the lack of data. Looking into the metrics present in Table 6.5, the precision reaches values from 0.8615 to 0.8903, proving that even though there are False Positives they do not significantly affect the performance of the model. With a Balanced Accuracy ranging from 0.7215 - 0.7811 and and AUROC of 0.8306 - 0.8944 the model can be considered a baseline approach for the diagnosis of BK, with most errors found in the negative prediction. The average confusion matrix portrays exactly that, with 42.3 False Positives and only 19.1 False Negatives, proving the models' difficulty in the negative class (see Table 6.6). Lastly, analysing the individual Recall of each class for a specific fold, we can see that the class 0 has a recall of 0.54 and the 1 a recall of 0.97, proving again the problem of this model. Nevertheless, the overall performance of the model was still considered a good baseline approach.

Table 6.5: Confidence Intervals for General Subset of Bacteria Baseline Model.

Subset	Metric	Task	Confidence Interval
General	F1	Bacteria	0.8960 - 0.9167
General	Recall	Bacteria	0.9258 - 0.9533
General	Precision	Bacteria	0.8615 - 0.8903
General	BA	Bacteria	0.7215 - 0.7811
General	ACC	Bacteria	0.8344 - 0.8681
General	AUROC	Bacteria	0.8306 - 0.8944

Table 6.6: Average Confusion Matrix for Bacteria Baseline.

	Predicted Negative	Predicted BK
Actual Negative	54.5	42.3
Actual BK	19.1	296.9

6.4.1.2 Fungi

The fungi-positive infection has a normalised weight of 3.2168, which was inserted into the loss function.

Looking at the feature map of the fungi predictions, similar maps were seen when compared to the Bacteria ones, meaning it was clear that the predictions were being made based on the iris or cornea of the eye. Nevertheless, these maps seem more dispersed than the previous ones, indicating that some other features were used. As discussed before in the literature, FK is much more difficult to diagnose than BK and so this behaviour was somewhat expected since the model has more difficulty finding the right features for the diagnoses (Figure 6.7).

Focusing on the ROC (Figure 6.8), an 'elbow' shape is found, with the FPR being very low for an increased value of TPR, this behaviour indicates that, unlike the previous case, the negative class is very well predicted, which makes sense looking into the dataset, since in this

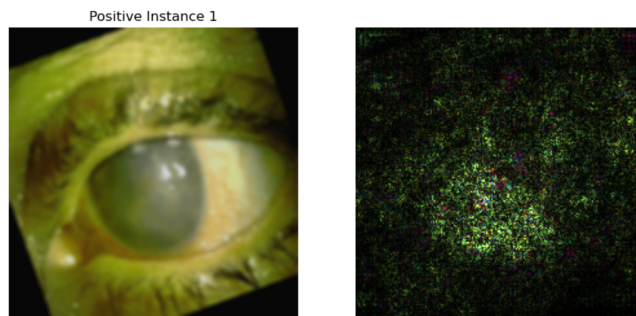


Figure 6.7: Saliency Map for a Positive Patient for Fungi Correctly Diagnosed.

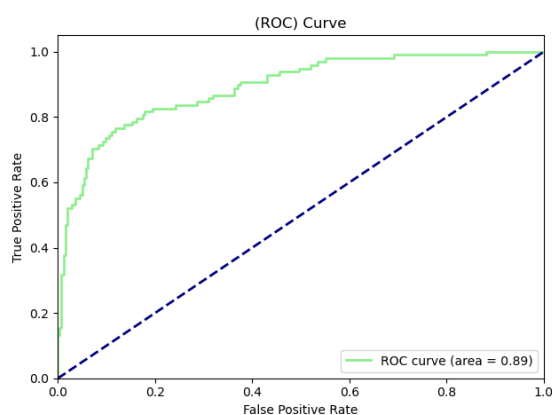


Figure 6.8: ROC Curve for one fold of baseline fungi prediction.

case the class imbalance happens in the opposite direction. Since the positive class is the one where the model could yield the worst results, the recall is the metric of interest since it directly portrays the models' ability to predict the positive class. The values range from 0.5441 - 0.6089, implying that the performance of the model in the identification of the fungi infection is not ideal, missing a lot of positive cases, which is corroborated by the analysis of the mean confusion matrix across all folds (see Table 6.8), in which an average of 41.5 False Negatives and 56.5 TP can be seen, showcasing the difficulty of the task. By observing other metrics of interest (see Table 6.7) like the BA (0.7378 - 0.7679), AUROC (0.8610 - 0.8834) and F1 (0.6138 - 0.6637) the problem is re-identified, proving the point. The F1 of each class for a random fold shows exactly that problem, with class 0 having 0.91 and class 1 0.68. The model was considered, nonetheless, a good baseline since Fungi Diagnosis was reported as a difficult task, so worse results were expected in comparison to other etiologies.

6.4.1.3 Amoeba

The weight of a positive sample of amoeba fed into the loss was 4.198, highlighting the amoeba infection as the most imbalanced etiology. The feature map (Figure 6.9) is the best of the three,

Table 6.7: Confidence Intervals for General Subset of Fungi Baseline Model.

Subset	Metric	Task	Confidence Interval
General	F1	Fungi	0.6138 - 0.6637
General	Recall	Fungi	0.5441 - 0.6089
General	Precision	Fungi	0.6859 - 0.7617
General	BA	Fungi	0.7378 - 0.7679
General	ACC	Fungi	0.8341 - 0.8568
General	AUROC	Fungi	0.8610 - 0.8834

Table 6.8: Average Confusion Matrix for Fungi Baseline.

	Predicted Negative	Predicted FK
Actual Negative	292.5	22.3
Actual FK	41.5	56.5

with a very intense green area in the centre of the iris, suggesting that the model selects good features for classification.

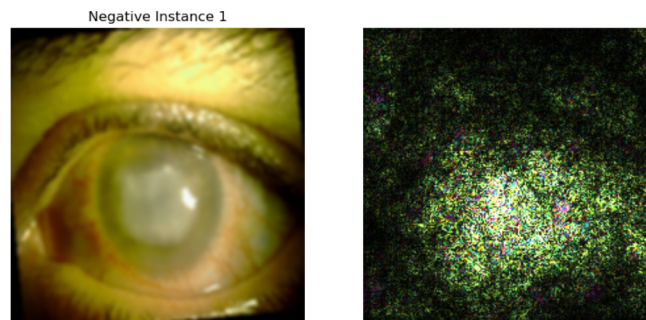


Figure 6.9: Saliency Map for a Negative patient for Amoeba Correctly Diagnosed.

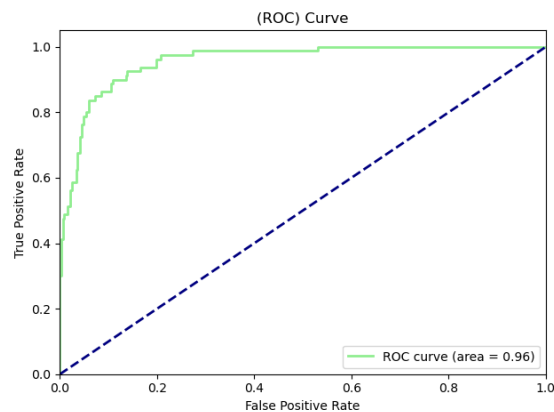


Figure 6.10: ROC Curve for one fold of baseline amoeba prediction.

The ROC curve (Figure 6.10) is the most near-ideal one of the three etiologies with an AUROC

of 0.96 for that specific fold, showcasing the model's ability to predict this task. The good AUROC is associated with good performance in the other metrics (see Table 6.9) yielding 0.8306 - 0.8769 for Balanced Accuracy, 0.7086 - 0.7908 for Recall, 0.9428 - 0.9698 for AUROC and 0.7403 - 0.8160 for the F1 Score. Looking into the worst performing metric, the recall, deeply into the model, and using a random fold, a recall of 0.96 was found for class 0, but 0.64 for class 1, which was somewhat expected since there are only 80 positive AK cases in the dataset. Nevertheless, the average confusion matrix (see Table 6.10) supports the robustness of the model with only 19.9 FN and 14.1 TP, becoming the best-performing infection. Because of that, this was considered a very good solution for the diagnosis of AK.

Table 6.9: Confidence Intervals for General Subset of Amoeba Baseline Model.

Subset	Metric	Task	Confidence Interval
General	F1	Ameba	0.7403 - 0.8160
General	Recall	Ameba	0.7086 - 0.7908
General	Precision	Ameba	0.7663 - 0.8577
General	BA	Ameba	0.8306 - 0.8769
General	ACC	Ameba	0.9037 - 0.9316
General	AUROC	Ameba	0.9428 - 0.9698

Table 6.10: Average Confusion Matrix for Amoeba Baseline.

	Predicted Negative	Predicted AK
Actual Negative	318.9	14.1
Actual AK	19.90	59.7

6.4.1.4 Baseline Downsides

Even though a minimum of classification performance was achieved for every task, the single-task approach was considered a relatively weak strategy, raising two main concerns:

- **Computational Cost:** The current approach meant that the doctor would need to submit one single eye image to three different models, and each one of them would diagnose for a specific type. This strategy represents an extremely high computational cost that, if possible, should be diminished
- **Multi-infection Cases:** The inability of the model to be trained with multiple diagnoses at once means it is also impossible for it to recognise it as an option. That implies that when the predictions are being made, a combined classification of bacteria and fungi is perceived the same way as a single bacteria infection would when training for the detection of Bacteria only. If for any chance the models predict that result, it is because of mere luck and not because the backbone of both tasks was common and the features that exam had implied that diagnosis. This limitation can put the patient's sight at risk because of an unpredicted development of the other etiology. In addition, the other case is also seen. No

simultaneous infections of Fungi and Amoeba are present in the dataset, suggesting they are mutually exclusive, a single task approach would not recognise that pattern, and so it can make mistakes by diagnosing the two types for each single task. A multitask model would be capable of learning and accounting for that.

6.4.2 Multitask

6.4.2.1 V1

Not every feature map reported as good results as the ones seen for the single task method, which makes sense since multiple infections are being predicted and so the selected features need to accommodate not only the prediction for one disease but for three of them. For example, in Figure 6.11 is possible to see the feature map of a negative prediction for Bacteria.

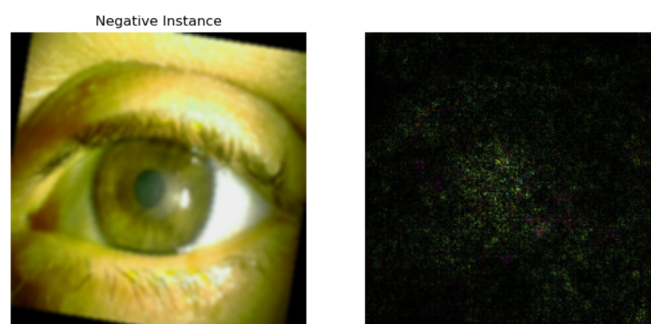


Figure 6.11: Saliency Map for a Negative patient for Bacteria Task Correctly Diagnosed with Multitask v1 and baseline loss.

The analysis of the ROC curve of one fold (see Figure 6.12) allows the conclusion that the amoeba task outperforms the fungi task, which outperforms the bacteria task. The 'elbow' shape is very evident in the Amoeba case, with an AUROC of 0.97, but it becomes less evident in Fungi, AUROC of 0.89 and almost disappears in the bacteria case, AUROC of 0.73. Again, these loss curves showcase the false positive and true positive comparison, so they tend to become critical in cases where there are a lot of FPs, which is the case of bacteria since it is the only one whose minority class is in the negative label, so this behaviour was expected (see Table 6.12). Exploring now the metrics of all folds in Table 6.11, a Balanced Accuracy of 0.5938 - 0.6561 was found for bacteria, 0.7013 - 0.7524 for fungi and 0.8223 - 0.8659 for Amoeba. These values are similar to the ones found on the single task model for fungi and amoeba, nevertheless a relevant discrepancy was encountered for the Bacteria that had in the single task model 0.7215 - 0.7811 of Balanced Accuracy. The comparative analysis of the AUROC (Bacteria: 0.7516 - 0.7918, Fungi: 0.8416 - 0.8768 and Ameba: 0.9463 - 0.9600) and its comparison with the single task model reveals that similarly good metrics were found for the Fungi and Amoeba, but the Bacteria AUROC of the single task model remained slightly better (0.8306 - 0.8944). The F1 metric retrieved a confidence interval of 0.8718 - 0.8914 for bacteria, 0.5520 - 0.6410 for fungi and 0.7373 - 0.7837 for amoeba,

which reveals the model difficulty in classifying the positive class for both ameba and fungi, as seen on the single task case, nevertheless both metrics were better in the single task version of the classification.

Focusing now on the confusion matrices, Table 6.11, they confirm what was suspected until now, with the model having a critical performance regarding FP for the Bacteria, and a similar performance than the one of single task for fungi and amoeba. The expanded confusion matrix for a specific fold was used to seek an explanation for the bad performance metrics on the positive label for fungi (since the bad performance of the bacteria was not on the disease label)(see Table 6.13). Looking at the table, it is possible to understand that the model is still struggling to separate the Fungi and Bacteria infections, probably due to the existence of the joint option. Looking at the advantage of the multilabel system, with the multiple infections, an analysis of the case of BK+FK and BK+AK, successful results were found with the model being able to identify 23 out of 44 cases of BF+FK and 16 of 34 cases of BK+AK. The misclassification is always only BK, which is in itself a good compromise since it means the model correctly identified one of the types.

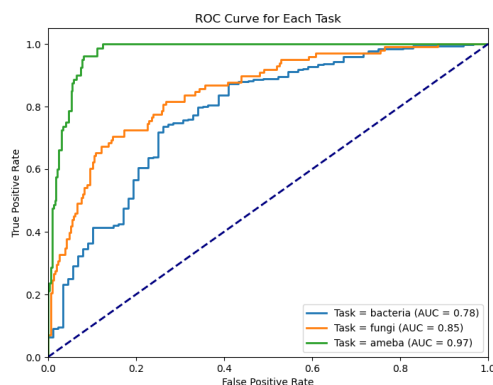


Figure 6.12: Roc Curve for a fold of the Multitask v1 with baseline loss.

6.4.2.2 V2

The feature analysis for the V2 was similar for all tasks and labels within each task. A bright green area is found on every map in the center region of the cornea, indicating the extraction of relevant features from the right area. Nevertheless, looking at this very specific example, more green dots were expected in the red areas and in the nebular-looking area of the iris instead of the top of the iris. So even though the maps look like the right area is taken into consideration, doubts remain if the right features of that area are selected (Figure 6.13).

The behaviour reported for the ROC curves (see Figure 6.14) of the multitask V1 can also be seen on this, with the 'elbow' look decreasing mainly in Bacteria. The AUROC for the ROC of the specific fold is 0.95 for Ameba, 0.85 for Fungi and 0.79 for Bacteria. Extrapolating it to an

Table 6.11: Confidence Intervals for General Subset, Tasks, and Metrics of Multitask v1 with baseline loss.

Subset	Metric	Task	Confidence Interval
General	F1	Bacteria	0.8718 - 0.8914
General	F1	Fungi	0.5520 - 0.6410
General	F1	Ameba	0.7373 - 0.7837
General	Recall	Bacteria	0.9508 - 0.9694
General	Recall	Fungi	0.4714 - 0.5755
General	Recall	Ameba	0.6833 - 0.7837
General	Precision	Bacteria	0.7984 - 0.8327
General	Precision	Fungi	0.6552 - 0.7546
General	Precision	Ameba	0.7774 - 0.8194
General	BA	Bacteria	0.5938 - 0.6561
General	BA	Fungi	0.7013 - 0.7524
General	BA	Ameba	0.8223 - 0.8659
General	ACC	Bacteria	0.7853 - 0.8193
General	ACC	Fungi	0.8191 - 0.8484
General	ACC	Ameba	0.9074 - 0.9168
General	AUROC	Bacteria	0.7516 - 0.7918
General	AUROC	Fungi	0.8416 - 0.8768
General	AUROC	Ameba	0.9463 - 0.9600

Table 6.12: Average Confusion Matrices for Different Tasks with Multitask v1 baseline loss.

	Negative	BK
Negative	27.7	69.1
BK	12.6	303.4

	Negative	FK
Negative	292.8	22.0
FK	46.7	51.3

	Negative	AK
Negative	318.1	15.1
AK	21.2	58.4

Table 6.13: Confusion Matrix for BK-FK-AK Infections of Multitask v1 with baseline loss.

	Healthy	BK	FK	AK	BK+FK	FK+AK	BK+AK	BK+FK+AK
Healthy	0	0	0	0	0	0	0	0
BK	4	199	3	0	25	0	7	0
FK	0	21	17	0	16	0	0	0
AK	2	15	0	2	1	0	26	0
BK+FK	0	20	1	0	23	0	0	0
FK+AK	0	0	0	0	0	0	0	0
BK+AK	0	14	0	4	0	0	16	0
BK+FK+AK	0	0	0	0	0	0	0	0

all-fold approach (see Table 6.14), the confidence interval of Bacteria was 0.7312 - 0.7776. When compared to the single task, the performance for AUROC is lower, which can showcase the impact of the multidagnosis system. Regarding the Fungi performance, 0.8270 - 0.8722, which is very similar to the single task performance, is only slightly decreased. The same behaviour pattern was identified in the Ameba classification, with the confidence interval being 0.9424 - 0.9619. For F1-Score, the underscoring of the multitask when compared with single task models remained,

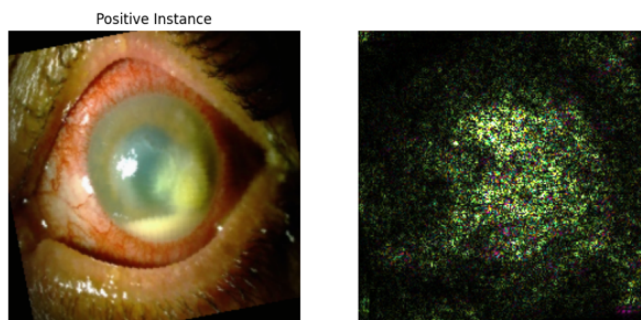


Figure 6.13: Saliency Map for a Positive patient for Fungi Task Correctly Diagnosed with Multi-task v2 and baseline loss.

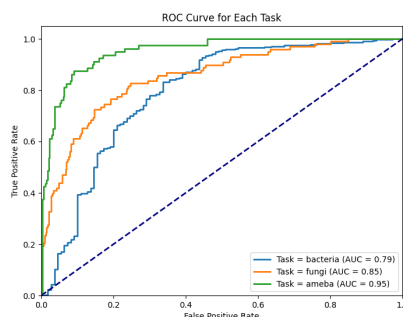


Figure 6.14: ROC Curve for one for of Multitask v2 baseline loss.

with confidence intervals of 0.8750 - 0.8881 for Bacteria and 0.5501 - 0.6388 for Fungi, the ameba interval, 0.7405 - 0.7883, is similar and slightly better than the one found in the single task approach. Regarding Balanced Accuracy, both the Fungi (0.7001 - 0.7531) and Ameba (0.8244 - 0.88637) have good metrics, similar to the single task performance. Nevertheless, for the Bacteria, the effect already reported on the V1 approach repeated itself, with a confidence interval 0.5902 - 0.6453. This decrease is due to the extremely high-class imbalance associated with the dataset. By observing the expanded confusion matrix on Table 6.16, and looking at the BK predicted column, it is possible to see that a lot of Fungi and Ameba cases (from multiple or single infections) are being classified as Bacteria, creating a lot of FPs and affecting the Balanced Accuracy of bacteria and causing the 69 FP that can be seen on Table 6.15. The multiclass v2 prediction of multilabel infections performs slightly better with an improvement in BK+AK, having only 5 misdiagnoses.

6.4.2.3 V1 vs V2

When deciding between V1 and V2, different priorities are set for each task: minimising false negatives (FN) for Fungi and Ameba, and minimising false positives (FP) for Bacteria. The focus is on recall, which directly affects FN, and precision, which relates to FP. When comparing the results of the metrics for both models, no difference is observed, with the confidence intervals being practically the same, which makes the decision of the best performing model not possible.

Table 6.14: Confidence Intervals for General Subset, Tasks, and Metrics of Multitask v2 with baseline loss.

Subset	Metric	Task	Confidence Interval
General	F1	Bacteria	0.8750 - 0.8881
General	F1	Fungi	0.5501 - 0.6388
General	F1	Ameba	0.7405 - 0.7883
General	Recall	Bacteria	0.9533 - 0.9733
General	Recall	Fungi	0.4725 - 0.5806
General	Recall	Ameba	0.6864 - 0.7737
General	Precision	Bacteria	0.7998 - 0.8264
General	Precision	Fungi	0.6534 - 0.7300
General	Precision	Ameba	0.7788 - 0.8396
General	BA	Bacteria	0.5902 - 0.6453
General	BA	Fungi	0.7001 - 0.7531
General	BA	Ameba	0.8244 - 0.8637
General	ACC	Bacteria	0.7910 - 0.8126
General	ACC	Fungi	0.8173 - 0.8459
General	ACC	Ameba	0.9065 - 0.9211
General	AUROC	Bacteria	0.7312 - 0.7776
General	AUROC	Fungi	0.8270 - 0.8722
General	AUROC	Ameba	0.9424 - 0.9619

Table 6.15: Average Confusion Matrices for Different Tasks with Multitask v2 baseline loss.

	Negative	BK
Negative	26.5	70.3
BK	11.6	304.4

	Negative	FK
Negative	291.7	23.1
FK	46.4	51.6

	Negative	AK
Negative	319.1	14.1
AK	21.5	58.1

Table 6.16: Confusion Matrix for BK-FK-AK Infections- Multitask v2 with baseline loss.

	Healthy	BK	FK	AK	BK+FK	FK+AK	BK+AK	BK+FK+AK
Healthy	0	0	0	0	0	0	0	0
BK	5	200	6	3	15	0	13	0
FK	0	19	18	0	15	0	0	0
AK	0	12	0	19	0	0	21	0
BK+FK	0	26	0	0	20	0	0	0
FK+AK	0	0	0	0	0	0	0	0
BK+AK	0	2	0	3	0	0	23	0
BK+FK+AK	0	0	0	0	0	0	0	0

Nevertheless, when focusing on multilabel infections, the V2 model outperformed the V1 model. Both of them were outperformed by the single task model, meaning that the multitask framework could be improved.

6.5 Multitask Exploration

As mentioned in the methodology, two different framework experiences were tested in both multitask models, the following sections comprise the discussion of those alterations in the results, as well as the choice for the best model

6.5.1 Multitask with Clinical Loss

The clinical loss computation resulted in a hospital weight of 0,021 for the Bacteria, 0.434 for Fungi and 0.545 for Ameba infection. Similar to what was seen until now, all the loss functions eventually stabilised and so no loss point will be shown in the discussion of this section. Because all the ROC Curves have a similar looking pattern already described in the previous section, the ROC will be further on only presented to support a discussion.

6.5.1.1 V1

Starting by analysing the saliency maps of this infection, a much noisier image is seen in comparison to the previous saliency maps. Across all folds and tasks for both the negative and positive diagnosis, this is a constant, with the maps being generally noisy and the features being less focused on the iris. This can mean that the model is learning features from the cornea not present in the iris, like the redness (see Figure 6.15). Performing a deep results exploration (see Table

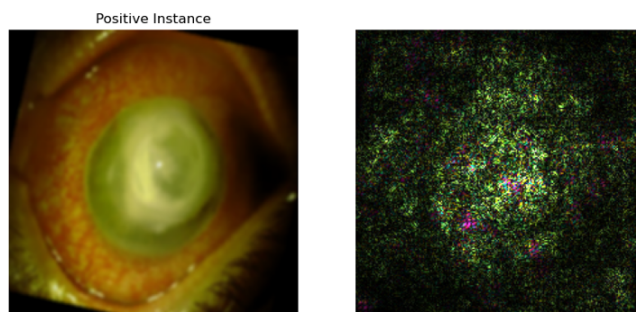


Figure 6.15: Saliency Map for a Positive patient for Fungi Task Correctly Diagnosed with Multitask v1 Clinical Loss.

6.17) on BA, AUROC and F1. Starting with the AUROC, (Bacteria: 0.7249 - 0.7717, Fungi: 0.8385 - 0.8780, Ameba: 0.9463 - 0.963), the results of the single task for each infection were matched, except bacteria, to each the result was slightly lower. Regarding Balanced Accuracy, (Bacteria: 0.6039 - 0.6374, Fungi: 0.6918 - 0.7441, Ameba: 0.8327 - 0.8731) the metrics yielded by the multitask v1 with baseline loss were slightly surpassed, nevertheless the single task model still outperforms this metrics regarding the Bacteria task, the values for the rest of the tasks were matched. Regarding the F1 with 0.8706 - 0.8872 for bacteria, 0.5322 - 0.6253 for fungi and 0.7493 - 0.7992 for ameba, the performance of the single task is improved comparing to bacteria and achieved for fungi and ameba. The confusion matrices (see Table 6.18) are very similar to the

ones encountered in multitask v1, with small improvements for ameba, which makes sense since the loss function will now attribute more weight to the diagnosis of this etiology due to its hospital costs. More importantly, observing the expanded confusion matrix (see Table 6.19), it is possible to see a decrease in the FP of bacteria, meaning that the model learned better the negative class of bacteria. When focusing on the multi-infection cases, an improvement in the correctly diagnosed cases was seen for BK+FK (27 vs 23) when comparing to the multitask baseline loss approach. For BK+AK, a even more significant improvement was seen with the identified cases shifting from 19 to 24. For both cases, all the mispredictions with the exception of one were because it was diagnosed as single infection. This analysis allows the reflection that multitask_v1 with clinical loss is capable of matching or surpassing the single task performance for every etiology except bacteria, while retaining the benefits of the multitask approach, proving itself as a strong competitor and performing better than the multitask with baseline loss.

Table 6.17: Confidence Intervals for General Subset, all Tasks, and Metrics on Multitask V1 with clinical loss.

Subset	Metric	Task	Confidence Interval
General	F1	Bacteria	0.8706 - 0.8872
General	F1	Fungi	0.5322 - 0.6253
General	F1	Ameba	0.7493 - 0.7992
General	Recall	Bacteria	0.9460 - 0.9654
General	Recall	Fungi	0.4576 - 0.5710
General	Recall	Ameba	0.7051 - 0.7948
General	Precision	Bacteria	0.8005 - 0.8275
General	Precision	Fungi	0.6282 - 0.7335
General	Precision	Ameba	0.7752 - 0.8391
General	BA	Bacteria	0.6039 - 0.6374
General	BA	Fungi	0.6918 - 0.7441
General	BA	Ameba	0.8327 - 0.8731
General	ACC	Bacteria	0.7859 - 0.8112
General	ACC	Fungi	0.8089 - 0.8406
General	ACC	Ameba	0.9086 - 0.9242
General	AUROC	Bacteria	0.7249 - 0.7717
General	AUROC	Fungi	0.8385 - 0.8780
General	AUROC	Ameba	0.9463 - 0.9631

Table 6.18: Average Confusion Matrices for Different Tasks with Multitask v1 Clinical loss.

	Negative	BK		Negative	FK		Negative	AK
Negative	27.5	69.3	Negative	289.9	24.9	Negative	318.5	14.7
BK	14	302	FK	47.6	50.4	AK	19.9	59.7

Table 6.19: Confusion Matrix for BK-FK-AK Infections- Multitask v1 with Clinical loss.

	Healthy	BK	FK	AK	BK+FK	FK+AK	BK+AK	BK+FK+AK
Healthy	0	0	0	0	0	0	0	0
BK	1	187	5	0	20	0	19	0
FK	3	17	14	1	12	0	1	0
AK	0	14	0	10	0	0	22	0
BK+FK	1	18	4	0	27	0	0	0
FK+AK	0	0	0	0	0	0	0	0
BK+AK	0	7	0	8	0	0	19	0
BK+FK+AK	0	0	0	0	0	0	0	0

6.5.1.2 V2

For the clinical adaptation of v2, similar results in the feature map were detected, with its overall noise and feature points away from iris increasing indicating that the diagnosis would be less focused in this area of the eye. The model returned (see Table 6.20) a F1 for bacteria of 0.8773 - 0.8872, for fungi, 0.5614 - 0.6420, and for ameba 0.7470 - 0.8021. An AUROC of 0.7413 - 0.7740 for Bacteria, 0.8395 - 0.8725 for fungi and 0.9448 - 0.9617 for ameba. Comparing these metrics with the previous multitask v2 approach, improvement was seen for all AUROC, and BA, mainly for the Bacteria infection, which was the worst class. When comparing to the single task models, the F1 improved in the bacteria, and remained similar for fungi and ameba. The F1, remained relatively similar. An observation of the confusion matrices on Tables 6.21 and 6.22 allows the interpretation that there is a similar behaviour to the one observed in multitask v2 baseline, with the main difference being observed in the ameba task, whose FN were reduced and FP increased. The analysis of the expanded confusion matrix origins a similar conclusion to the one taken on the v1 case with the multi-infection diagnosis being improved for this loss, BK+AK went from 23 correctly predicted cases to 30 correctly predicted cases and BK+FK went from 20 to 24. The results analysis offer the conclusion that the multitask with clinical loss outperforms the multitask with baseline loss. It was also possible to infer that again, similar performance to the one observed in the single task model is achieved, with the exception of Bacteria infection.

6.5.1.3 V1 vs V2

When deciding between V1 and V2, different priorities are set for each task: minimizing false negatives (FN) for Fungi and Ameba, and minimizing false positives (FP) for Bacteria. The focus is on recall, which directly affects FN, and precision, which relates to FP. Starting with the recall for Fungi, model v1 has a recall of 0.4576 - 0.5710, whereas model v2 has a recall of 0.4895 - 0.5758. Since higher recall means fewer false negatives, model v2 performs better in this aspect. For the recall of Ameba, model v1 has a recall of 0.7051 - 0.7948, while model v2 has a recall of 0.7115 - 0.7910. Again, model v2 performs slightly better, although the difference is minimal. Next, looking at the precision for Bacteria, model v1 has a precision of 0.8005 - 0.8275, whereas model v2 has a precision of 0.8077 - 0.8364. Higher precision means fewer false positives, so

Table 6.20: Confidence Intervals for General Subset, Tasks, and Metrics of Multitask v2 with Clinical Loss.

Subset	Metric	Task	Confidence Interval
General	F1	Bacteria	0.8773 - 0.8872
General	F1	Fungi	0.5614 - 0.6420
General	F1	Ameba	0.7470 - 0.8021
General	Recall	Bacteria	0.9401 - 0.9663
General	Recall	Fungi	0.4895 - 0.5758
General	Recall	Ameba	0.7115 - 0.7910
General	Precision	Bacteria	0.8077 - 0.8364
General	Precision	Fungi	0.6575 - 0.7315
General	Precision	Ameba	0.7814 - 0.8218
General	BA	Bacteria	0.6167 - 0.6611
General	BA	Fungi	0.7074 - 0.7529
General	BA	Ameba	0.8334 - 0.8736
General	ACC	Bacteria	0.7957 - 0.8145
General	ACC	Fungi	0.8215 - 0.8464
General	ACC	Ameba	0.9079 - 0.9247
General	AUROC	Bacteria	0.7413 - 0.7740
General	AUROC	Fungi	0.8395 - 0.8725
General	AUROC	Ameba	0.9448 - 0.9617

Table 6.21: Average Confusion Matrices for Different Tasks with Multitask v2 Clinical loss.

	Negative	BK
Negative	31.1	65.7
BK	14.8	301.2

	Negative	FK
Negative	292.0	22.8
FK	45.8	52.2

	Negative	AK
Negative	318.4	14.8
AK	19.8	59.8

Table 6.22: Confusion Matrix for BK-FK-AK Infections- Multitask v2 with Clinical loss.

	Healthy	BK	FK	AK	BK+FK	FK+AK	BK+AK	BK+FK+AK
Healthy	0	0	0	0	0	0	0	0
BK	4	193	2	2	24	0	9	0
FK	0	19	12	0	21	0	2	0
AK	4	13	1	11	0	0	11	0
BK+FK	0	20	0	0	24	0	0	0
FK+AK	0	0	0	0	0	0	0	0
BK+AK	1	5	0	2	0	0	30	0
BK+FK+AK	0	0	0	0	0	0	0	0

model v2 also performs better in this regard. In summary, model v2 is the better choice as it has higher recall for both Fungi and Ameba, which helps minimize false negatives, and higher precision for Bacteria, which helps minimize false positives. Nevertheless, their performance is very similar even when comparing the extended confusion matrixes, with v2 improving BK+AK and v1 BF+FK.

6.5.1.4 Clinical Loss evaluation

The insertion of the clinical loss in both models seems to improve the multitask performance of the model on single and multi label infections, and so, because of that, it was considered a positive addition.

6.5.2 Multitask with Clinical Loss and Threshold

After the implementation of the automatic threshold for the conversion of probability into prediction was accessed on the previous models. Since the loss is updated based on the probability and not on the prediction, the insertion of this metric does not yield any changes in the training of the model from the loss evolution to the feature maps and so only results exploration will be conducted.

6.5.2.1 V1

Comparing now the performance of the model to the previous strategy, and using the metrics present in Table 6.23. Since the threshold will only change the value at which the probability is considered a prediction, the impact of the solution should be mainly described by the Precision and Recall values and so the analysis will be focused on them, as well as the Balanced Accuracy. Looking at the Recall, for bacteria (0.6707 - 0.7856), the results are way worse than those found for the standard threshold (0.9460 - 0.9654) meaning the number of FN increased a lot, regarding Fungi, the opposite was reported, achieving 0.7162 - 0.8410. On the Ameba case, an increase of over 0.1 was observed, reaching 0.8947 - 0.9396. Focusing now on the Precision, the precision of the bacteria increased, but the ameba and fungi one significantly dropped indicating the FP increment in those two cases. To understand that fact, the ROC curve of a random fold was analysed (see Figure 6.16). The youden solution will chose the threshold that maximises the TPR-FPR value of the points used to create the ROC. The metrics show mainly a decrease of performance for bacteria, that is because, looking at the selected value by the youden's statistic, a threshold of 0.73 is chosen due to the 'flat elbow' looking ROC curve. This ROC curve shape means that there is not a clear turning-point that will improve the prediction, and so it is hard to define a threshold value where the true positive rate increases rapidly without a significant increase in the false positive rate. A threshold of 0.73 will mean that a positive result for bacteria will only be found above that value which is very high, leading to an increase of FN and decrease of FP, causing then the drop of the metrics associated with FN, the recall, and an increase in the ones associated with FP, the precision, explaining the results. For fungi and ameba, 0.22 and 0.16 are selected as the threshold values for a random fold. Since the performance of the ROC curves of this models is better, the values are more "real" and so the results of their implementation are not as severe. Nevertheless, a significant increase in FP is seen in the confusion matrices (see Table 6.24) for both cases proving that the ROC curves are not 'good' enough for this method to be implemented with good results. A deep look into the extended matrix (see Table 6.25) showcased a behaviour that was not seen in any other model, with the a lot of Bacteria infections being

misclassified. So it was easy to conclude that although in theory this implementation is beneficial, due to the imperfect performance of the model and the dataset diversity, the results suggest that the standard threshold should be used.

Table 6.23: Confidence Intervals for General Subset, all Tasks, and Metrics of V1 with Clinical loss and threshold.

Subset	Metric	Task	Confidence Interval
General	F1	Bacteria	0.7639 - 0.8315
General	F1	Fungi	0.5924 - 0.6498
General	F1	Ameba	0.7013 - 0.7793
General	Recall	Bacteria	0.6707 - 0.7856
General	Recall	Fungi	0.7162 - 0.8410
General	Recall	Ameba	0.8947 - 0.9396
General	Precision	Bacteria	0.8798 - 0.8996
General	Precision	Fungi	0.4793 - 0.5839
General	Precision	Ameba	0.5690 - 0.6849
General	BA	Bacteria	0.6951 - 0.7364
General	BA	Fungi	0.7551 - 0.7929
General	BA	Ameba	0.8732 - 0.9057
General	ACC	Bacteria	0.6843 - 0.7616
General	ACC	Fungi	0.7367 - 0.8058
General	ACC	Ameba	0.8480 - 0.8971
General	AUROC	Bacteria	0.7357 - 0.7851
General	AUROC	Fungi	0.8247 - 0.8645
General	AUROC	Ameba	0.9392 - 0.9615

Table 6.24: Average Confusion Matrices for Different Tasks with Multitask v1 Clinical loss and Threshold.

	Negative	BK		Negative	FK		Negative	AK
Negative	68.2	28.6	Negative	241.8	73.0	Negative	287.2	46.0
BK	85.9.6	230.1	FK	21.7	76.3	AK	6.6	73.0

Table 6.25: Confusion Matrix for BK-FK-AK Infections - Multitask v1 with Clinical loss and threshold.

	Healthy	BK	FK	AK	BK+FK	FK+AK	BK+AK	BK+FK+AK
Healthy	0	0	0	0	0	0	0	0
BK	12	117	25	15	29	3	31	0
FK	1	9	36	0	11	0	1	0
AK	0	2	1	25	0	0	8	0
BK+FK	0	7	13	0	20	0	0	0
FK+AK	0	0	0	0	0	0	0	0
BK+AK	0	0	0	24	0	1	19	0
BK+FK+AK	0	0	0	0	0	0	0	0

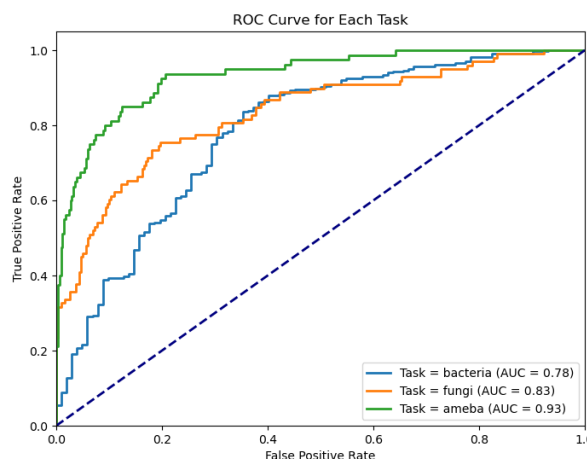


Figure 6.16: ROC Curve of one fold of Multitask v1 with clinical loss and threshold.

6.5.2.2 V2

Because the ROC curve of V2 had a similar aspect to the V1s one in previous tasks, the same behaviour was expected. Comparing the V2 with clinical loss and adaptive threshold to the V2 with the clinical loss by analysing the confidence intervals available in Table 6.26, a big decrease in the Bacteria recall can be observed reaching a confidence interval of 0.6869 - 0.7777, the opposite happens for fungi, achieving 0.7636 - 0.8343, ameba performance, 0.8780 - 0.9208, also increases but since the FN of this tasks were already so low, with only 8 FN (see Table 6.27) the improvement is not as noticeable as the fungi. Looking at precision, the expected decrease in fungi was observed, reaching 0.4773 - 0.5847 and the same for ameba, with 0.6593 - 0.7210. The improvement of TP in both fungi and ameba tasks, reaching 78.3 for fungi and 71,6 for ameba (see Table 6.27) resulted in an improvement in the Balanced accuracy, achieving 0.7585 - 0.808 for fungi and 0.8881 - 0.9133 for ameba. For bacteria the improvement was also seen since this metric gives equal weight to both classes and the performance of class 0 improved a lot with a very high threshold because a lot of TN were diagnosed. Nevertheless, an analysis of the Bacteria performance in the extended confusion matrix (see Table 6.28) allowed the conclusion that, similarly to what happened in V1, the implementation of this method a ROC curve that was closer to the ideal literature shape.

6.5.2.3 V1 vs V2

Similarly to before, V2 performed better, having a better confidence interval for all the metrics of interest in all tasks, except for the Fungi precision, which is approximately the same.

6.5.2.4 Youden's statistics evaluation

Due to what was reported in both models, the threshold was not a good tactic for the specific case.

Table 6.26: Confidence Intervals for Selected Subsets, Tasks, and Metrics on Multitask with Optimised Loss and Threshold v2.

Subset	Metric	Task	Confidence Interval
General	F1	Bacteria	0.7751 - 0.8260
General	F1	Fungi	0.5945 - 0.6707
General	F1	Ameba	0.7572 - 0.8027
General	Recall	Bacteria	0.6869 - 0.7777
General	Recall	Fungi	0.7636 - 0.8343
General	Recall	Ameba	0.8780 - 0.9208
General	Precision	Bacteria	0.8762 - 0.9018
General	Precision	Fungi	0.4773 - 0.5847
General	Precision	Ameba	0.6593 - 0.7210
General	BA	Bacteria	0.6946 - 0.7311
General	BA	Fungi	0.7585 - 0.8081
General	BA	Ameba	0.8881 - 0.9133
General	ACC	Bacteria	0.6971 - 0.7512
General	ACC	Fungi	0.7399 - 0.8100
General	ACC	Ameba	0.8898 - 0.9134
General	AUROC	Bacteria	0.7378 - 0.7864
General	AUROC	Fungi	0.8284 - 0.8840
General	AUROC	Ameba	0.9482 - 0.9652

Table 6.27: Average Confusion Matrices for Different Tasks with Multitask v2 Clinical loss and Threshold.

	Negative	BK		Negative	FK		Negative	AK
Negative	67.5	29.3	Negative	241.4	73.4	Negative	300.6	32.6
BK	84.6	231.4	FK	19.7	78.3	AK	8.0	71.6

Table 6.28: Confusion Matrix for BK-FK-AK Infections - Multitask v2 with Clinical loss and threshold.

	Healthy	BK	FK	AK	BK+FK	FK+AK	BK+AK	BK+FK+AK
Healthy	0	0	0	0	0	0	0	0
BK	7	113	17	10	48	1	16	0
FK	0	3	23	0	20	0	0	0
AK	0	2	0	17	0	0	9	0
BK+FK	0	10	6	0	31	2	2	1
FK+AK	0	0	0	0	0	0	0	0
BK+AK	1	0	0	29	0	1	21	0
BK+FK+AK	0	0	0	0	0	0	0	0

6.5.3 Best Predictive Model

All in all, the best performing model was V2, distinguishing itself as the one that was capable of retrieving satisfactory metrics for all tasks and allowing the multi-label diagnosis that was not offered by the also good performing single-task approaches.

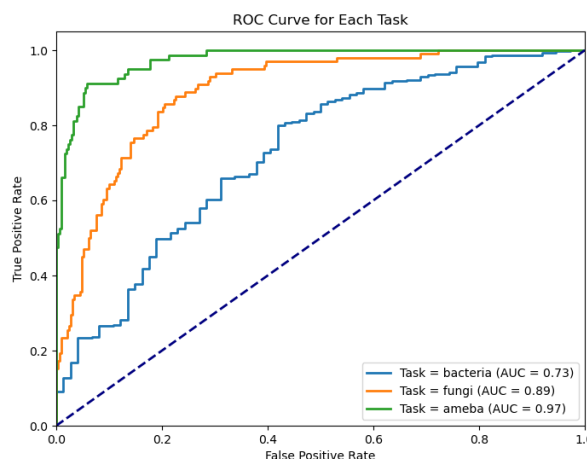


Figure 6.17: ROC Curve of a random fold of Multitask v2 with clinical loss and threshold.

6.6 Identity Features' Impact

The best-performing methodology was chosen for the identity feature evaluation. Since the sex and age prediction models had such good performance, meaning the patient's identity features could be detected by the model, the goal behind this task was to assess if the classification model was making its prediction based on any of these features, meaning that a bias was being introduced. The confidence intervals of every subset,task,metric combination can be found on the Appendix A, but only the p-values after correction will be analysed. The corrected p-values obtained for the other methodologies are available in the Appendix.

6.6.1 Sex Influence

Starting by the sex analysis (see Table 6.29), it was possible to see that the sex seems to have an impact on the Fungi Accuracy, with the male sex having a lower accuracy performance. Regarding Ameba, a significant difference was found for F1, Recall and Balanced Accuracy, again, looking at the confidence interval by subset table is possible to see that males have lower confidence intervals for all this metrics. This could mean that the architecture is using some features related with the sex of the patient to make predictions, mainly for ameba, which are obviously not related with the disease phenotype, suggesting that disentanglement of this features from the corneal eye photograph might result in a performance boost and avoid this bias. In bacteria no significance was seen.

6.6.2 Age Impact

Looking at the age table, 6.30, unlike what was expected by the dataset analysis, significant differences were seen for all metrics across ages, with the exception of Balanced Accuracy, meaning

Table 6.29: Corrected p-values for Sex.

Metric	Bacteria	Fungi	Ameba
F1	1.000	0.2510	0.0035
Recall	1.000	0.7070	0.0010
Precision	1.000	0.1320	0.2580
BA	1.000	1.000	0.0042
ACC	1.000	0.0003	0.6840
AUROC	1.000	1.000	0.4810

that age features are maybe being taken into account when making the prediction. The analysis of the Fungi values reported the same as bacteria, with all metrics exhibiting significant differences across age groups. Focusing on ameba, that was only found for ACC. The AUROC was classified as NAN because of the lack of data in at least one age group (usually 0-18 year olds) for that task and because of that, it was excluded from the table. Because the class imbalance regarding age is so high on the dataset, it is very strong to affirm that disentanglement is needed. Instead, an evaluation with a bigger for balanced dataset regarding age should be done and a conclusion taken based on that.

Table 6.30: Corrected p-values for Age.

Metric	Bacteria	Fungi	Ameba
F1	0.0001	7.70×10^{-16}	0.4555
Recall	0.0153	1.63×10^{-12}	0.9099
Precision	0.0029	2.93×10^{-20}	0.4946
BA	0.7693	1.24×10^{-13}	0.2418
ACC	0.0006	6.83×10^{-16}	0.0055

By analysing the results for the other frameworks, this patterns were also detected, making clear the need for further investigation of both sex and age in a larger dataset.

Chapter 7

Looking into the Future

7.1 Goals' Satisfaction

Looking back, three main goals were established at the start of the project. First, the creation of a comprehensive database of corneal eye images labelled by infection type, along with an in-depth analysis to ensure the dataset's quality and relevance. Second, the implementation of several deep learning algorithms to accurately classify these images according to different types of keratitis. Third, the assessment of the statistical impact of correlated features, such as sex and age, on disease presentation. These goals were successfully achieved, demonstrating the effectiveness and potential of deep learning solutions in advancing keratitis diagnosis and contributing to broader healthcare improvements. The proposed solution achieved state-of-the-art metrics for a multi-label diagnosis of bacterial keratitis, fungal keratitis, and acanthamoeba keratitis, with 95% confidence intervals for AUROC of 0.7413 - 0.7740, 0.8395 - 0.8725, and 0.9448 - 0.9617, respectively. When looking at the sex and age performance, excellent AUROC values are found, 0.8790- 0.8994 for sex and 0.8331-0.8624 for age. The capability of the model to predict the sex and age based on the photo raised a awareness towards bias, and so the infection metrics on several subsets were evaluated and compared. A significant impact of sex for Ameba and Age for Bacteria and Fungi was found, meaning a bias associated with patient features could be implied.

7.2 Future Work

Although the proposed solution has yielded very good results, some improvements could be done in the future to reduce the False Positives of Bacteria and False Negatives of Fungi and Amoeba.

As explained before, we think the problem is caused by the lack of data of Fungi and Amoeba representations, which leads them to not being correctly predicted for this class. When that happens, and because the model is not trained with any healthy patients, the eye exams are automatically diagnosed as bacteria infections, causing the False Positives. So the first proposed step, regarding future work, would be to either acquire more data or artificially generate new images. In addition to the acquisition of new images being very time and patient-dependent, it also comes

as an economical burden since each image needs to be associated with the laboratory test for the etiologies of suspicion. Also fungi and ameba are way less common as it was previously discussed and so this method has its limitations. The other solution includes using Generative Adversarial Networks, the so called GANs to artificially generate new positive images from the fungi and amoeba etiologies based on the ones already present in the dataset [144]. GANs are a type of deep learning framework where two neural networks, the generator and the discriminator, are put against each other. The generator creates synthetic data instances, such as images, from random noise with the goal of fooling the discriminator into classifying them as real data. Simultaneously, the discriminator learns to distinguish between real data and fake data generated by the generator. Through this adversarial process, we hypothesise that GANs could generate high-quality and diverse outputs that closely mimic the real data distribution. This way, the model could have access to a broader, larger and more balanced dataset, which will likely lead to an improvement in the tasks performance.

Another interesting test would be to evaluate the performance of the model on segmented images of the cornea, meaning only the actual eye, and no skin, would be inputted in the framework. That can yield significant improvements because, looking at some of the saliency maps, results suggest that the model is basing its prediction in some skin features. Those skin features can introduce noisy assumptions into the model and so an image where that area is not present could be ideal.

In this dissertation, VGG16, ResNet50 and DenseNet121 were tested, but we believe it would be interesting to benchmark other architectures as well (e.g. EfficientNet [145]).

Another proposed improvement would be to make the algorithm a two step framework where firstly the images would be tested for BK and not BK and then for FK vs AK. This could be interesting since fungi and ameba have similar datasets representation and so the positive class distribution would be similar. In addition, this solution allowed the usage of the Youden's classification benefits for fungi and amoeba classification without penalising the bacteria performance, which would be an ideal solution.

Lastly, the developed work has proven that it is possible to correctly predict the sex and age of the patient based on an eye photograph. This could mean that those identity features are somehow present in the images and used as a way for the model to predict a certain infection, introducing new biases to our methodology. The p-value studied that was performed pointed in that direction. To avoid those connections, disentanglement of identity features from the eye corneal photographs could be relevant to make sure the predictions are based on only medical features.

7.3 Final Remarks

In this dissertation, we presented a deep learning framework for the diagnosis of keratitis. The ability of deep learning algorithms to analyse eye photographs with high precision and accuracy represents a major advancement in the early detection of keratitis, potentially preventing the economical burden associated with the usual diagnosis technique.

Beyond the immediate benefits for keratitis diagnosis, AI has the potential to reduce health-care inequalities between countries. By providing advanced diagnostic tools that are accessible and scalable, AI can bridge the gap in regions with limited medical resources and expertise, like the ones affected by Keratitis. This democratisation of healthcare technology ensures that more people, regardless of their geographic location or economic status, can receive timely and accurate medical care, aligning itself with the United Nations Sustainable Development Goals.

References

- [1] GeeksforGeeks, “Introduction to deep learning.” <https://www.geeksforgeeks.org/introduction-deep-learning/>, Accessed: 2024-06-18.
- [2] National Cancer Institute, “Blood flow through blood vessels.” <https://visualsonline.cancer.gov/details.cfm?imageid=7161>, 2024s. Image ID: 7161.
- [3] All About Vision, “Neurotrophic keratitis: Symptoms, causes and treatments,” 2023. Accessed:2024-05-30.
- [4] L. Brown, A. K. Leck, M. Gichangi, M. J. Burton, and D. W. Denning, “The global incidence and diagnosis of fungal keratitis,” *The Lancet Infectious Diseases*, vol. 21, no. 3, pp. e49–e57, 2021.
- [5] E. J. Topol, “High-performance medicine: the convergence of human and artificial intelligence,” *Nature Medicine*, vol. 25, no. 1, pp. 44–56, 2019.
- [6] C. Stewart, “Artificial intelligence (ai) in healthcare market size worldwide from 2021 to 2030,” *Statista*, Sept. 2023.
- [7] W. Lu, Y. Tong, Y. Yu, Y. Xing, C. Chen, and Y. Shen, “Applications of artificial intelligence in ophthalmology: General overview,” *Eye Center, Renmin Hospital of Wuhan University, Eye Institute of Wuhan University, Wuhan, Hubei, China*, vol. 2, 2018.
- [8] B. Basok and I. Deveci Kocakoç, *ARTIFICIAL INTELLIGENCE IN HEALTH SCIENCES*, pp. 1–18. 12 2023.
- [9] M. Kuo, B. Hsu, Y. Yin, *et al.*, “A deep learning approach in diagnosing fungal keratitis based on corneal photographs,” *Scientific Reports*, vol. 10, p. 14424, 2020.
- [10] Z. Zhang, H. Wang, S. Wang, *et al.*, “Deep learning-based classification of infectious keratitis on slit-lamp images,” *Therapeutic Advances in Chronic Disease*, vol. 13, 2022.
- [11] G. Learning, “Everything you need to know about vgg16,” 2021. Accessed: 2024-05-27.
- [12] D. Genius, “Resnet50,” 2021. Accessed: 2024-05-27.
- [13] N. Radwan, *Leveraging Sparse and Dense Features for Reliable State Estimation in Urban Environments*. PhD thesis, 06 2019.
- [14] L. Zhao and Z. Zhang, “An improved pooling method for convolutional neural networks,” *Nature*, vol. 14, p. 1589, 2024. Received: 31 July 2023; Accepted: 02 January 2024; Published: 18 January 2024.

- [15] S. Ayon, “Dropout in neural networks,” *Towards Data Science*, 2019.
- [16] F. VL, “Interpreting training, validation accuracy and loss,” *Medium*, 2024.
- [17] GeeksforGeeks, “Backpropagation in neural network.” <https://www.geeksforgeeks.org/backpropagation-in-neural-network/>, 2023. [Accessed 2024-06-23].
- [18] J. Doe, “Understanding confusion matrix,” *Towards Data Science*, 2023.
- [19] Google Developers, “Machine learning crash course.” <https://developers.google.com/machine-learning/crash-course>, 2022.
- [20] J. Devpblog, “Cv 2: Gaussian and median filter, separable 2d filter.” <https://medium.com/jun94-devpblog/cv-2-gaussian-and-median-filter-separable-2d-filter-2d11ee022c66>, 2020. Accessed: 2024-06-12.
- [21] Database Camp, “Understanding overfitting in machine learning,” 2024. Accessed: 2024-06-17.
- [22] Devopedia, “Receiver operating characteristic (roc) curve.” <https://devopedia.org/roc-curve>. Accessed: 2024-06-17.
- [23] Johns Hopkins Medicine, “Keratitis.” <https://www.hopkinsmedicine.org/health/conditions-and-diseases/keratitis>, 2024. [Accessed 2024-05-30].
- [24] D. S. J. Ting, M. S. Y. Lim, A. H. Shek, S. Y. Lee, L. Y. M. Yip, K. L. Loh, C. Z. X. Wong, M. Ang, D. S. L. Ting, L. Y. K. Ting, and S. Sanjay, “Infectious keratitis: An update on epidemiology, causative microorganisms, risk factors, and antimicrobial resistance,” *Eye (London, England)*, vol. 35, no. 4, pp. 1084–1101, 2021.
- [25] A. Austin, T. Lietman, and J. Rose-Nussbaumer, “Update on the management of infectious keratitis,” *Ophthalmology*, vol. 124, pp. 1678–1689, 2017.
- [26] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
- [27] T. Davenport and R. Kalakota, “The potential for artificial intelligence in healthcare,” *Future Healthcare Journal*, vol. 6, no. 2, pp. 94–98, 2019.
- [28] U. Nations, “Sustainable development goals.”
- [29] “Mit-mgb ai cures conference,” 2024. Accessed: 2024-06-22.
- [30] “Ijup – investigação jovem da u.porto,” 2024. Accessed: 2024-06-22.
- [31] “Ieee bibm 2024,” 2024. Accessed: 2024-06-22.
- [32] American Academy of Ophthalmology, “The role of the ophthalmologist and ophthalmology.” <https://www.aao.org/about/policies/role-ophthalmology-ophthalmologist>, 2024. Accessed: 2024-05-30.
- [33] J. J. Kanski, *Clinical Ophthalmology: A Systematic Approach*. London: Elsevier, 8th ed., 2015.

- [34] A. Y. Lee, R. T. Yanagihara, C. S. Lee, L. D. Minor, S. Luo, and D. S. W. Ting, "Ethical and privacy issues of ai in ophthalmology," *Asia-Pacific Journal of Ophthalmology*, vol. 9, no. 4, pp. 393–397, 2020.
- [35] World Health Organization, "Blindness and vision loss." https://www.who.int/health-topics/blindness-and-vision-loss#tab=tab_3, 2024. Accessed: 2024-05-30.
- [36] "Global ophthalmology market – industry trends and forecast to 2030," Jan. 2023.
- [37] B. R. Straatsma, *Ophthalmology: Essays, Abstracts and Reviews, Volume 2*. New York: Springer Science & Business Media, 2011.
- [38] H. Kolb, E. Fernandez, and R. Nelson, *Webvision: The Organization of the Retina and Visual System*. University of Utah Health Sciences Center, 2019. Accessed: 2024-05-30.
- [39] A. C. Guyton and J. E. Hall, *Textbook of Medical Physiology*. Philadelphia: Elsevier, 13th ed., 2015.
- [40] K. Boyd, "Ophthalmology subspecialists," Feb. 2023.
- [41] W. H. Organization, "Blindness and visual impairment." <https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment>, 2023. Accessed: 2024-05-31.
- [42] Mayo Clinic, "Keratitis." <https://www.mayoclinic.org/diseases-conditions/keratitis/symptoms-causes/syc-20374110>, 2024. [Accessed 23-11-2023.
- [43] Prevent Blindness, "Keratitis." Accessed:2024-05-30.
- [44] Cleveland Clinic, "Keratitis: Types, symptoms & treatment," 2023. Accessed: 2024-06-01.
- [45] L. Ung, P. J. M. Bispo, S. S. Shanbhag, M. S. Gilmore, and J. Chodosh, "The persistent dilemma of microbial keratitis: Global burden, diagnosis, and antimicrobial resistance," *Surv Ophthalmol*, vol. 64, pp. 255–271, May-Jun 2019. Epub 2018 Dec 24.
- [46] D. S. J. Ting, C. Shan Ho, R. Deshmukh, D. G. Said, and H. S. Dua, "Infectious keratitis: an update on epidemiology, causative microorganisms, risk factors, and antimicrobial resistance," *Nature*, Jan. 2021.
- [47] S. Tan, A. Walkden, L. Au, C. Fullwood, A. Hamilton, A. Qamruddin, *et al.*, "Twelve-year analysis of microbial keratitis trends at a uk tertiary hospital," *Eye (Lond)*, vol. 31, pp. 1229–1236, 2017.
- [48] D. Ting, C. Settle, S. Morgan, O. Baylis, and S. Ghosh, "A 10-year analysis of microbiological profiles of microbial keratitis: the north east england study," *Eye (Lond)*, vol. 32, pp. 1416–1417, 2018.
- [49] M. Green, N. Carnt, A. Apel, and F. Stapleton, "Queensland microbial keratitis database: 2005-2015," *Br J Ophthalmol*, vol. 103, pp. 1481–1486, 2019.
- [50] R. Kaye, A. Kaye, H. Sueke, T. Neal, C. Winstanley, M. Horsburgh, *et al.*, "Recurrent bacterial keratitis," *Investig Ophthalmol Vis Sci*, vol. 54, pp. 4136–4139, 2013.

- [51] A. Tam, E. Côté, M. Saldanha, A. Lichtinger, and A. Slomovic, "Bacterial keratitis in toronto: a 16-year review of the microorganisms isolated and the resistance patterns observed," *Cornea*, vol. 36, pp. 1528–1534, 2017.
- [52] A. Cariello, R. Passos, M. Yu, and A. Hofling-Lima, "Microbial keratitis at a referral center in brazil," *Int Ophthalmol*, vol. 31, pp. 197–204, 2011.
- [53] J. Hernandez-Camarena, E. Graue-Hernandez, M. Ortiz-Casas, A. Ramirez-Miranda, A. Navas, L. Pedro-Aguilar, *et al.*, "Trends in microbiological and antibiotic sensitivity patterns in infectious keratitis: 10-year experience in mexico city," *Cornea*, vol. 34, pp. 778–785, 2015.
- [54] M. Yu, A. Höfling-Lima, and G. Furtado, "Microbiological and epidemiological study of infectious keratitis in children and adolescents," *Arq Bras Oftalmol*, vol. 79, pp. 289–293, 2016.
- [55] J. Kalamurthy, C. Kalavathy, P. Parmar, C. Nelson Jesudasan, and P. Thomas, "Spectrum of bacterial keratitis at a tertiary eye care centre in india," *Biomed Res Int*, vol. 2013, p. 181564, 2013.
- [56] Y. Zhang, Z. Wang, and X. Sun, "Etiological analysis and in vitro drug sensitivity of bacterial keratitis in northern china in the period of 2006-2015," *Zhonghua Yan Ke Za Zhi*, vol. 53, pp. 662–667, 2017.
- [57] M. Cabrera-Aguas, P. Khoo, C. George, M. Lahra, and S. Watson, "Antimicrobial resistance trends in bacterial keratitis over 5 years in sydney, australia," *Clin Exp Ophthalmol*, vol. 48, pp. 183–191, 2019.
- [58] S. Tavassoli, G. Nayar, K. Darcy, M. Grzeda, J. Luck, O. Williams, *et al.*, "An 11-year analysis of microbial keratitis in the south west of england using brain-heart infusion broth," *Eye (Lond)*, vol. 33, pp. 1619–1625, 2019.
- [59] M. Politis, D. Wajnsztajn, B. Rosin, C. Block, and A. Solomon, "Trends of bacterial keratitis culture isolates in jerusalem; a 13-years analysis," *PLoS ONE*, vol. 11, p. e0165223, 2016.
- [60] G. Castano, A. G. Elnahry, and P. K. Mada, *Fungal Keratitis*. Treasure Island (FL): StatPearls Publishing LLC., 2020. StatPearls.
- [61] W. B. Khor, N. Prajna, P. Garg, J. S. Mehta, L. Xie, Z. Liu, *et al.*, "The asia cornea society infectious keratitis study: a prospective multicenter study of infectious keratitis in asia," *Am J Ophthalmol*, vol. 195, pp. 161–170, 2018.
- [62] L. Keay, E. Gower, A. Iovieno, R. Oechsler, E. Alfonso, A. Matoba, *et al.*, "Clinical and microbiological characteristics of fungal keratitis in the united states, 2001-2007: a multicenter study," *Ophthalmology*, vol. 118, pp. 920–926, 2011.
- [63] D. Ting, C. Ho, J. Cairns, A. Elshahn, M. Al-Aqaba, T. Boswell, *et al.*, "12-year analysis of incidence, microbiological profiles and in vitro antimicrobial susceptibility of infectious keratitis: the nottingham infectious keratitis study," *Br J Ophthalmol*, 2020.
- [64] P. Maier, P. K. Betancor, and T. Reinhard, "Contact lens-associated keratitis— an often underestimated risk," *Dtsch Arztebl Int*, Oct. 2022.

- [65] S. McLeod, A. Kolahdouz-Isfahani, K. Rostamian, C. Flowers, P. Lee, and P. McDonnell, "The role of smears, cultures, and antibiotic sensitivity testing in the management of suspected infectious keratitis," *Ophthalmology*, pp. 23–28, 1996.
- [66] C. Dalmon, T. Porco, T. Lietman, N. Venkatesh Prajna, L. Prajna, M. Das, *et al.*, "The clinical differentiation of bacterial and fungal keratitis: A photographic survey," *Investigative Ophthalmology & Visual Science*, vol. 53, p. 1787, 2012.
- [67] P. Singh, A. Gupta, and K. Tripathy, *Keratitis*. Treasure Island (FL): StatPearls Publishing, 2024.
- [68] B. H. Jeng, D. C. Gritz, A. B. Kumar, D. S. Holsclaw, T. C. Porco, S. D. Smith, *et al.*, "Epidemiology of ulcerative keratitis in northern california," *Arch Ophthalmol*, vol. 128, pp. 1022–1028, 2010.
- [69] B. Khan, H. Fatima, A. Qureshi, S. Kumar, A. Hanan, J. Hussain, and S. Abdullah, "Drawbacks of artificial intelligence and their potential solutions in the healthcare sector," *Biomedical Materials & Devices*, Feb. 2023.
- [70] R. Fakoor, F. Ladhak, A. Nazi, and M. Huber, "Using deep learning to enhance cancer diagnosis and classification," in *A conference presentation*, The 30th International Conference on Machine Learning, 2013.
- [71] BBC News, "Nhs ai test spots tiny cancers missed by doctors," *BBC News*, March 2024. Retrieved on June 4, 2024.
- [72] S. Lee, S. Celik, B. Logsdon, and *et al.*, "A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia," *Nat Commun*, vol. 9, p. 42, 2018.
- [73] S. Berg, "Nudge theory explored to boost medication adherence." American Medical Association, Chicago, 2018.
- [74] K. Utermohlen, "Four robotic process automation (rpa) applications in the healthcare industry." Medium, 2018.
- [75] N. N. Khanna, M. A. Maindarkar, V. Viswanathan, J. F. E. Fernandes, S. Paul, M. Bhagawati, P. Ahluwalia, Z. Ruzsa, A. Sharma, R. Kolluri, I. M. Singh, J. R. Laird, M. Fatemi, A. Alizad, L. Saba, V. Agarwal, A. Sharma, J. S. Teji, M. Al-Maini, V. Rathore, S. Naidu, K. Liblik, A. M. Johri, M. Turk, L. Mohanty, D. W. Sobel, M. Miner, K. Viskovic, G. Tsoulfas, A. D. Protogerou, G. D. Kitis, M. M. Fouda, S. Chaturvedi, M. K. Kalra, and J. S. Suri, "Economics of artificial intelligence in healthcare: Diagnosis vs. treatment," *Healthcare (Basel)*, vol. 10, no. 12, p. 2493, 2022. Published online 2022 Dec 9.
- [76] S. Ji, Q. Gu, H. Weng, Q. Liu, P. Zhou, Q. He, R. Beyah, and T. Wang, "De-health: all your online health information are belong to us," *arXiv preprint*, 2019.
- [77] M. Baowaly, C. Lin, C. Liu, and K. Chen, "Synthesizing electronic health records using improved generative adversarial networks," *Journal of the American Medical Informatics Association*, vol. 26, no. 3, pp. 228–241, 2019.
- [78] U. o. P. Human Research Protection Office, "The european union (eu) general data protection regulation (gdpr)," 2024. Accessed: 2024-06-3.
- [79] U. of Galway, "Health research - university of galway," 2024. Accessed: 2024-06-3.

- [80] FDA, “Fda permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems.” <https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-based-device-detect-certain-diabetes-related-eye>, 2018. Accessed: 2024-06-04.
- [81] D. S. Char, N. H. Shah, and D. Magnus, “Implementing machine learning in health care – addressing ethical challenges,” *New England Journal of Medicine*, vol. 378, pp. 981–983, 2018.
- [82] J. Wolff, J. Pauling, A. Keck, and J. Baumbach, “The economic impact of artificial intelligence in health care: systematic review,” *Journal of Medical Internet Research*, vol. 22, no. 2, p. e16866, 2020.
- [83] McKinsey Global Institute, “A future that works: automation, employment, and productivity,” 2017.
- [84] T. H. Davenport and J. Kirby, *Only humans need apply: Winners and losers in the age of smart machines*. New York: HarperBusiness, 2016.
- [85] T. H. Davenport and K. Dreyer, “Ai will change radiology, but it won’t replace radiologists,” *Harvard Business Review*, 2018.
- [86] S. E. Shladover, “The truth about “self-driving” cars,” *Scientific American*, vol. 314, no. 6, pp. 53–57, 2016.
- [87] D. Insights, “State of ai in the enterprise,” 2018. Accessed: 2024-06-03.
- [88] T. H. Davenport and J. Glaser, “Just-in-time delivery comes to knowledge management,” *Harvard Business Review*, 2002.
- [89] Y.-B. Chou, A. U. Kale, P. Lanzetta, T. Aslam, J. Barratt, C. Danese, B. Eldem, N. Eter, R. Gale, J.-F. Korobelnik, I. Kozak, X. Li, X. Li, A. Loewenstein, P. Ruamviboonsuk, T. Sakamoto, D. S. Ting, P. van Wijngaarden, S. M. Waldstein, D. Wong, L. Wu, M. A. Zapata, and J. Zarranz-Ventura, “Current status and practical considerations of artificial intelligence use in screening and diagnosing retinal diseases: Vision academy retinal expert consensus,” *Current Opinion in Ophthalmology*, vol. 34, pp. 403–413, September 2023.
- [90] U.S. Food and Drug Administration, “Artificial intelligence and machine learning in software as a medical device.” <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices>. Accessed: June 4, 2024.
- [91] “Retinalscreenings.” <https://retinalscreenings.com/>. Accessed: June 5, 2024.
- [92] RightEye LLC, “The righteye vision system receives breakthrough device designation from u.s. fda for parkinson’s disease test.” Accessed: June 5, 2024.
- [93] Carl Zeiss Meditec, Inc., “Clarus 700.” <https://www.zeiss.com/meditec/en/products/retinal-cameras/clarus-700.html>. Accessed: June 5, 2024.
- [94] E. Inc., “Eyenuk’s eyeart ai system is the first and only fda-cleared ai technology for use with multiple retinal cameras,” 2021. Accessed: 2024-06-05.

- [95] AEYE Health, “Aeye diagnostic screening.” <https://www.aeyehealth.com/aeye-diagnostic-screening>. Accessed: June 5, 2024.
- [96] M. Kuo, B. Hsu, Y. Lin, *et al.*, “Comparisons of deep learning algorithms for diagnosing bacterial keratitis via external eye photographs,” *Scientific Reports*, vol. 11, p. 24227, 2021.
- [97] M. A. Dahlgren, A. Lingappan, and K. R. Wilhelmus, “The clinical diagnosis of microbial keratitis,” *American Journal of Ophthalmology*, vol. 143, no. 6, pp. 940–944.e1, 2007.
- [98] T. K. Redd, N. V. Prajna, M. Srinivasan, P. Lalitha, T. Krishnan, R. Rajaraman, A. Venugopal, N. Acharya, G. D. Seitzman, T. M. Lietman, J. D. Keenan, J. P. Campbell, and X. Song, “Image-based differentiation of bacterial and fungal keratitis using deep convolutional neural networks,” *Ophthalmol Sci*, vol. 2, p. 100119, Jan 2022.
- [99] T. K. Redd, L. Della Santina, N. V. Prajna, P. Lalitha, N. Acharya, and T. Lietman, “Automated differentiation of bacterial from fungal keratitis using deep learning,” *Investigative Ophthalmology & Visual Science*, vol. 62, p. 2161, June 2021.
- [100] A. K. Ghosh, R. Thammasudjarit, P. Jongkhajornpong, J. Attia, and A. Thakkestian, “Deep learning for discrimination between fungal keratitis and bacterial keratitis: Deepkeratitis,” *Cornea*, vol. 41, no. 5, pp. 616–622, 2022.
- [101] N. Hung, A. K. Shih, C. Lin, M. T. Kuo, Y. S. Hwang, W. C. Wu, C. F. Kuo, E. Y. Kang, and C. H. Hsiao, “Using slit-lamp images for deep learning-based identification of bacterial and fungal keratitis: Model development and validation with different convolutional neural networks,” *Diagnostics (Basel)*, vol. 11, p. 1246, Jul 2021.
- [102] Y. Xu, M. Kong, W. Xie, R. Duan, Z. Fang, Y. Lin, Q. Zhu, S. Tang, F. Wu, and Y.-F. Yao, “Deep sequential feature learning in clinical image classification of infectious keratitis,” *Engineering*, vol. 7, no. 7, pp. 1002–1010, 2021.
- [103] A. Koyama, D. Miyazaki, Y. Nakagawa, *et al.*, “Determination of probability of causative pathogen in infectious keratitis using deep learning algorithm of slit-lamp images,” *Scientific Reports*, vol. 11, p. 22642, 2021. Received: 27 August 2021, Accepted: 02 November 2021, Published: 22 November 2021.
- [104] S. Hu, Y. Sun, J. Li, P. Xu, M. Xu, Y. Zhou, Y. Wang, S. Wang, and J. Ye, “Automatic diagnosis of infectious keratitis based on slit lamp images analysis,” *Journal of Personalized Medicine*, vol. 13, no. 3, 2023.
- [105] R. Sarayar, Y. D. Lestari, A. A. A. Setio, and R. Sitompul, “Accuracy of artificial intelligence model for infectious keratitis classification: a systematic review and meta-analysis,” *Frontiers in Public Health*, vol. 11, 2023.
- [106] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [107] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2015.
- [108] MyGreatLearning, “Everything you need to know about vgg16,” 2020. Accessed: 2024-06-08.

- [109] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.
- [110] N. Kundu, “Exploring resnet50: An in-depth look at the model architecture and code implementation,” 2023. Accessed: 2024-06-08.
- [111] D. AI, “Implementing densenet-121 in pytorch: A step-by-step guide,” 2020.
- [112] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2017.
- [113] F. Reveriano, “The advantages of densenet,” 2018. Accessed: 2024-05-27.
- [114] J. Brownlee, “Pooling layers for convolutional neural networks,” 2023. Accessed: 2024-06-13.
- [115] J. Brownlee, “Batch normalization for training of deep neural networks,” 2023. Accessed: 2024-06-13.
- [116] Vergotten, “Loss functions: Comprehensive guide to loss functions in various machine learning domains,” *Medium*, 2023. Accessed: 2024-06-06.
- [117] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2017.
- [118] JavaTPoint, “Performance metrics in machine learning.” Accessed: June 5, 2024.
- [119] Educative.io, “What is balanced accuracy?,” 2023. Accessed: 2024-06-05.
- [120] R. Poplin, A. V. Varadarajan, K. Blumer, Y. Liu, M. V. McConnell, G. S. Corrado, L. Peng, and D. R. Webster, “Predicting sex from retinal fundus photographs using automated deep learning,” *Scientific Reports*, vol. 11, no. 1, p. 10236, 2021.
- [121] ImageNet, “Imagenet: A large-scale hierarchical image database,” 2024.
- [122] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems* (F. Pereira, C. Burges, L. Bottou, and K. Weinberger, eds.), vol. 25, pp. 1097–1105, Curran Associates, Inc., 2012.
- [123] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.
- [124] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?,” in *Advances in Neural Information Processing Systems* (Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, eds.), vol. 27, pp. 3320–3328, Curran Associates, Inc., 2014.
- [125] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [126] P. Simard, D. Steinkraus, and J. Platt, “Best practices for convolutional neural networks applied to visual document analysis,” pp. 958–962, 01 2003.

- [127] A. Howard, “Some improvements on deep convolutional neural network based image classification,” 12 2013.
- [128] Coursesteach, “Computer vision part 25: Gaussian filter,” 2020. Accessed: 2024-06-12.
- [129] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Z. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “Pytorch: An imperative style, high-performance deep learning library,” *CoRR*, vol. abs/1912.01703, 2019.
- [130] Y. LeCun, L. Bottou, G. B. Orr, and K. R. Müller, *Efficient BackProp*, pp. 9–50. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998.
- [131] J. Brownlee, “A gentle introduction to k-fold cross-validation.” <https://machinelearningmastery.com/k-fold-cross-validation/>, 2020. Accessed: 2024-06-13.
- [132] scikit-learn developers, “sklearn.model_selection.stratifiedkfold.” https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html, 2024. Accessed: 2024-06-13.
- [133] scikit-multiflow Development Team, “scikit-multiflow: A machine learning package for streaming data.” <http://scikit.ml/index.html>, 2024. Accessed: 2024-06-13.
- [134] G. Culture, “How does batch size impact your model learning?,” *Medium*, 2021. Accessed: 2024-06-13.
- [135] scikit-learn developers, “sklearn.utils.class_weight.compute_class_weight.” https://scikit-learn.org/stable/modules/generated/sklearn.utils.class_weight.compute_class_weight.html, 2024. Accessed: 2024-06-13.
- [136] Keras Team, “Transfer learning and fine-tuning.” https://keras.io/guides/transfer_learning/. Accessed: 2024-06-17.
- [137] J. Brownlee, “Threshold moving for imbalanced classification.” <https://machinelearningmastery.com/threshold-moving-for-imbalanced-classification/>, 2020. Accessed: 2024-06-18.
- [138] C. Team, “Saliency in captum.” <https://captum.ai/api/saliency.html>, 2023. Accessed: 2024-06-18.
- [139] K. Szczepankiewicz, A. Popowicz, K. Charkiewicz, *et al.*, “Ground truth based comparison of saliency maps algorithms,” *Scientific Reports*, vol. 13, p. 16887, 2023. Received 28 November 2022, Accepted 16 September 2023, Published 06 October 2023.
- [140] K. Academy, “Normal distributions review.” <https://www.khanacademy.org/math/statistics-probability/modeling-distributions-of-data/normal-distributions-library/a/normal-distributions-review>. Accessed: 2024-06-18.
- [141] National University, “Independent samples.” <https://resources.nu.edu/statsresources/IndependentSamples>. Accessed: 2024-06-17.

- [142] Boston University School of Public Health, “Hypothesis testing: ANOVA.” https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_hypothesistesting-anova/bs704_hypothesistesting-anova_print.html. Accessed: 2024-06-17.
- [143] W. Haynes, *Holm’s Method*, pp. 902–902. New York, NY: Springer New York, 2013.
- [144] J. Brownlee, “What are generative adversarial networks (gans)?” <https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/>, 2020. Accessed: 2024-06-22.
- [145] M. Tan and Q. V. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *International Conference on Machine Learning*, PMLR, 2019.

Appendix A

Appendix

A.1 Presented Poster

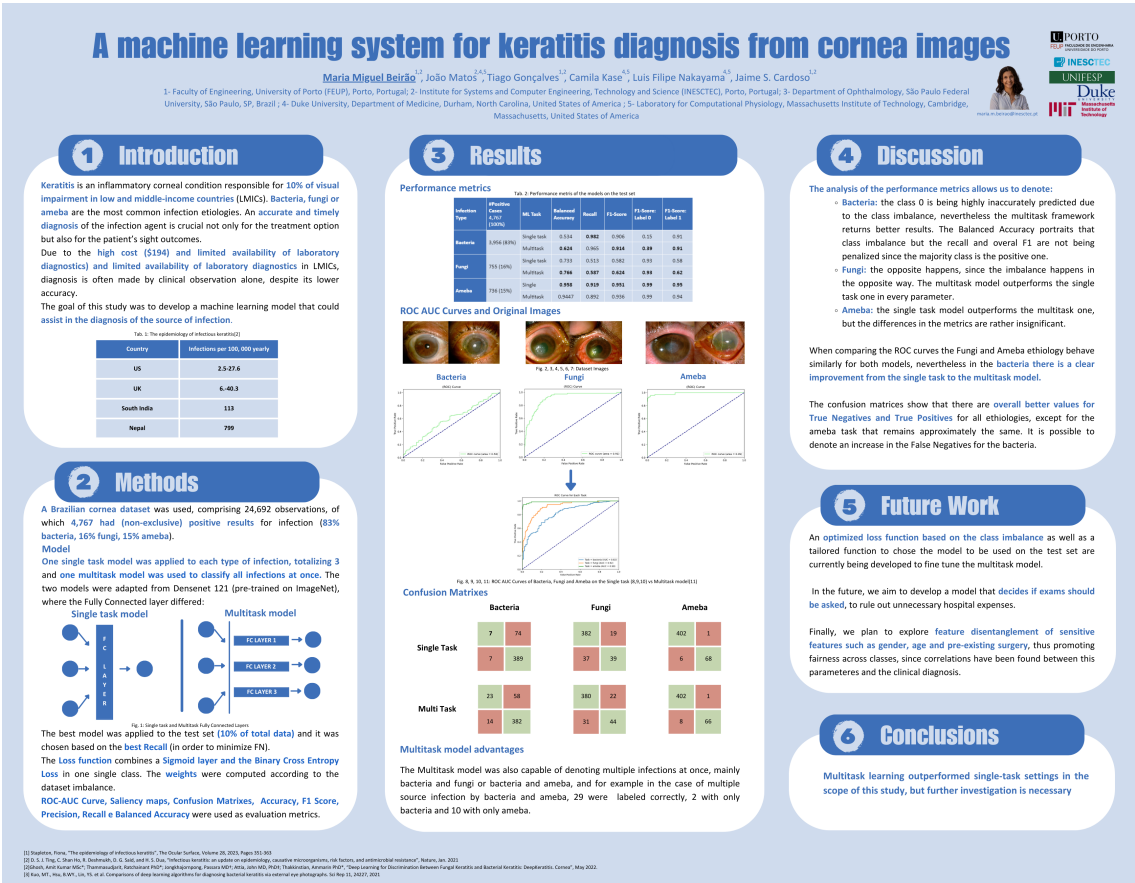


Figure A.1: Poster Presented in IJUP and AI Cures Conference.

A.2 Optimal Architecture and Parameters

A.2.1 Best Model Analysis

Starting by analysing the performance of ResNet50, VGG16 and DenseNet121 on a small 50 images subset, with a very simple classification model and no specific image transformations, it was possible to see that DenseNet121 not only performed better but also had the fastest time per epoch (see Table A.1). Since time constrictions are often a problem when dealing with Deep Learning tasks, DenseNet121 was considered our model of choice.

Table A.1: Assessment of the best baseline models.

Architecture	AUROC	Time per epoch(s)
DenseNet121	0.65	0.21
ResNet50	0.53	0.27
VGG16	0.44	1.16

A.2.2 Epochs

We runned several experiments to understand the number of epochs required for the model to converge.

A.2.2.1 Sex and Age

The effect of 100, 200 and 300 epochs on the loss can be seen on Figure A.2. Looking at the validation loss in the 100 epoch plot, it is unclear if a lower validation loss can be obtained or not and although the training loss is still decreasing, its stabilisation is still unclear. Comparing to the 200 epochs, the validation loss clearly reaches its minimum value at around 120 and then it starts gradually increasing, showing a clear sign of overfitting to the train data. Plotting the 300 epochs to understand if we were witnessing a local maximum and not just overfit, it was easy to conclude that our assumption was correct and it was indeed a case of overfitting. With that in mind, 200 epochs was considered the optimal value for sex and age. Nevertheless it was possible to see that the validation loss is still very unstable, which shows the already diversity of the dataset in hands and the difficulty of the model to adapt to it.

A.2.2.2 Infection Prediction

Looking now into the epochs influence on the disease prediction, the tests were performed on the multitask_v1 model and applied to all. Since this was a more difficult task, only 200 and 300 epochs were tested. In the 200 epochs graph the relative stabilisation of the validation loss, with some instability like it was seen on the previous examples, allows us to conclude that the best validation loss will probably be achieved at 150, nevertheless 300 epochs were tested, and as expected the validation loss started to increase, showing a clear case of overfit to the training data,

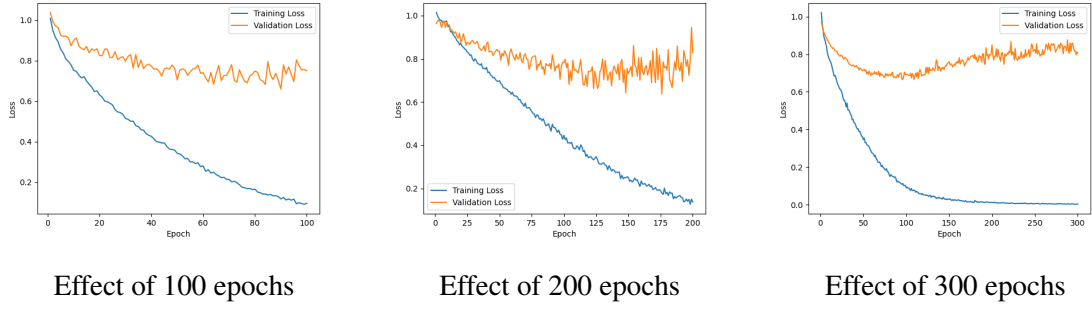


Figure A.2: Effect of epochs on Loss Curve of the sex prediction.

with difficulty generalising to the validation set. Looking at the train loss, we could also see that although it does not fully stabilise in 200 epochs, it reaches very low values of around 0.1, proving that the training is effective.

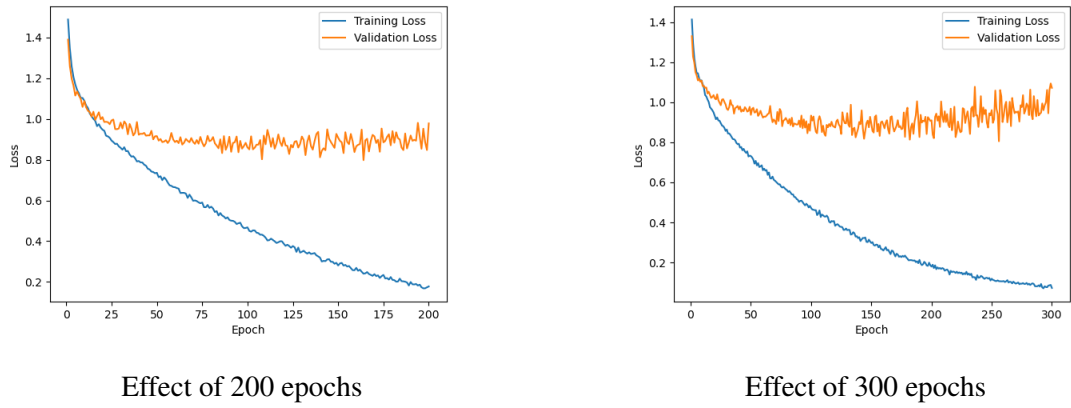


Figure A.3: Effect of epochs on the Multitask Loss Curve.

A.2.3 Learning Rate

The impact of the learning rate and weight decay was also evaluated by testing in which epoch the model had its lowest validation loss and its value. As it was previously seen, if the learning rate is too high, the validation loss might start decreasing initially but then increase very fast, and if it is too little, the validation loss decreases very slowly over time, since it does not learn enough information at a time. Looking at the testing of several learning rates in our model (see Table A.2) Looking at the table both described cases can be seen, with the $1e^{-5}$ achieving its minimum performance in the fourth epoch and rapidly increasing, proving that it is too high, and $1e^{-7}$ having its minimum loss almost in the end of the run time, meaning the model never fully stabilised. Because of that, $1e^{-6}$ was selected as the optimal learning rate. The weight decay was adapted according to the Learning rate, following the usual $1e^{-2}$ rule.

Table A.2: Learning Rate and weight decay impact on minimum validation loss.

Learning Rate	Weight Decay	Min Val Loss	Epoch
$1e^{-5}$	$1e^{-7}$	0.97	4
$1e^{-6}$	$1e^{-8}$	0.84	105
$1e^{-7}$	$1e^{-9}$	0.94	293

A.3 Loss Functions

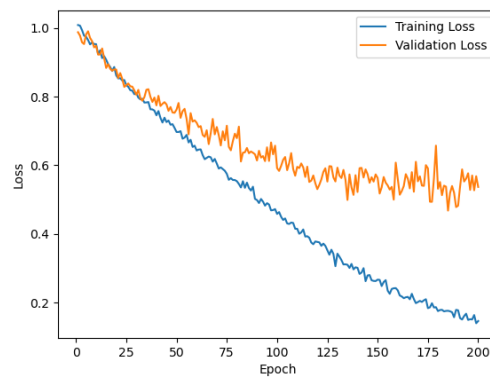


Figure A.4: Loss evolution over epochs for sex model.

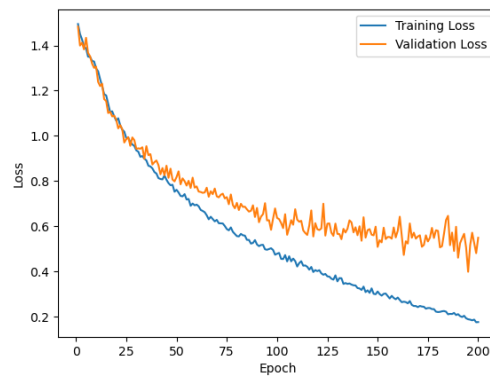


Figure A.5: Loss evolution over epochs for age model.

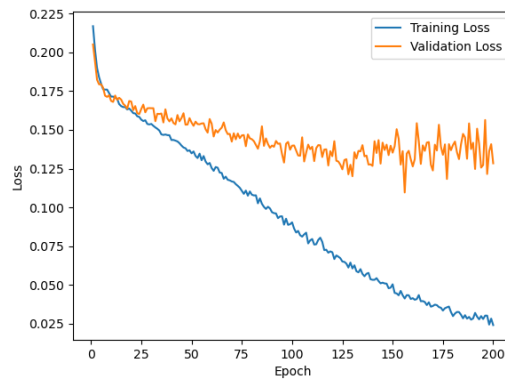


Figure A.6: Loss evolution over epochs for single task bacteria model.

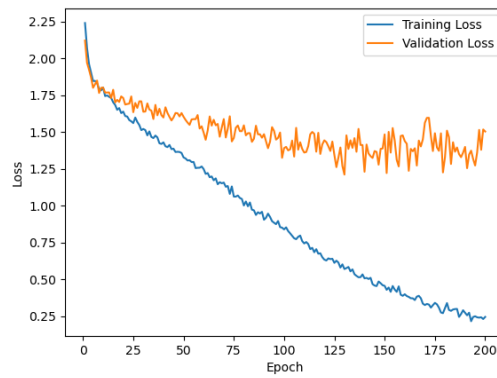


Figure A.7: Loss evolution over epochs for single task fungi model.

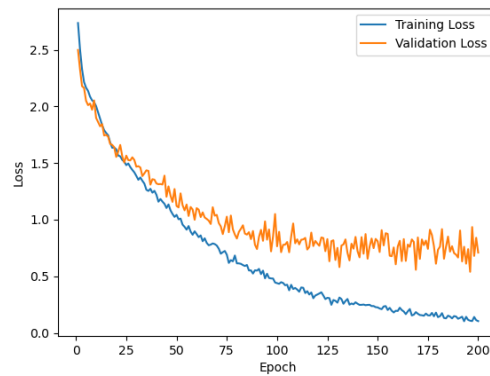


Figure A.8: Loss evolution over epochs for single task ameba model.

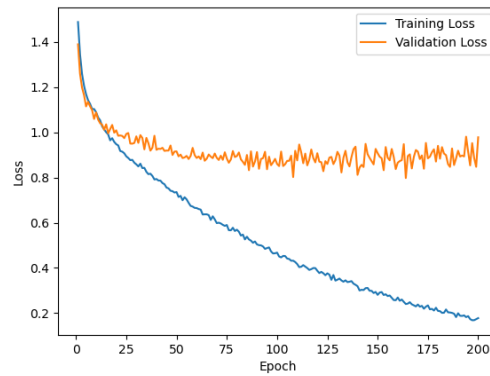


Figure A.9: Loss evolution over epochs for Multitask v1 with baseline loss.

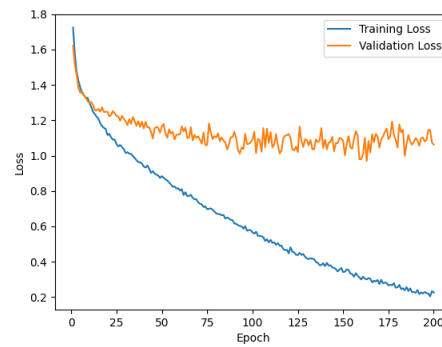


Figure A.10: Loss evolution over epochs for Multitask v2 with baseline loss.

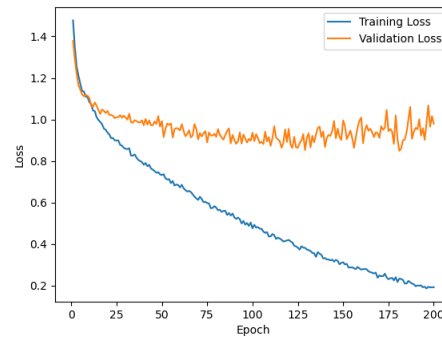


Figure A.11: Loss evolution over epochs for Multitask v1 with clinical loss (represents the loss of both results exploration cases, since the adaptive threshold method uses the same loss function).

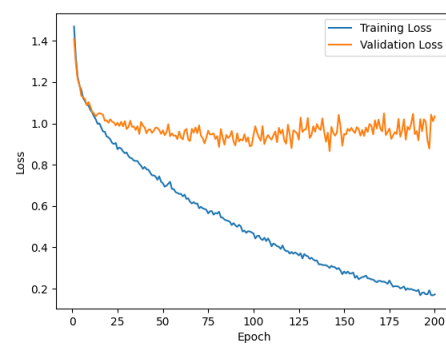


Figure A.12: Loss evolution over epochs for Multitask v2 with clinical loss (represents the loss of both results exploration cases, since the adaptive threshold method uses the same loss function).

A.4 Confidence Intervals for the best methodology

Table A.3: Confidence Intervals for Sex Subsets, Tasks, and Metrics of Multitask v2 with clinical loss

Subset	Metric	Task	Confidence Interval
Female Sex	F1	Bacteria	0.8655 - 0.8925
Female Sex	F1	Fungi	0.4172 - 0.5764
Female Sex	F1	Ameba	0.8144 - 0.8541
Female Sex	Recall	Bacteria	0.9403 - 0.9670
Female Sex	Recall	Fungi	0.3549 - 0.5319
Female Sex	Recall	Ameba	0.7830 - 0.8729
Female Sex	Precision	Bacteria	0.7921 - 0.8408
Female Sex	Precision	Fungi	0.5044 - 0.6629
Female Sex	Precision	Ameba	0.8296 - 0.8619
Female Sex	BA	Bacteria	0.5910 - 0.6651
Female Sex	BA	Fungi	0.6525 - 0.7442
Female Sex	BA	Ameba	0.8638 - 0.9031
Female Sex	ACC	Bacteria	0.7767 - 0.8215
Female Sex	ACC	Fungi	0.8671 - 0.9101
Female Sex	ACC	Ameba	0.8952 - 0.9170
Female Sex	AUROC	Bacteria	0.7141 - 0.8072
Female Sex	AUROC	Fungi	0.8083 - 0.8947
Female Sex	AUROC	Ameba	0.9500 - 0.9709
Male Sex	F1	Bacteria	0.8778 - 0.8912
Male Sex	F1	Fungi	0.5871 - 0.6762
Male Sex	F1	Ameba	0.6025 - 0.7269
Male Sex	Recall	Bacteria	0.9378 - 0.9683
Male Sex	Recall	Fungi	0.5099 - 0.6095
Male Sex	Recall	Ameba	0.5650 - 0.6746
Male Sex	Precision	Bacteria	0.8107 - 0.8416
Male Sex	Precision	Fungi	0.6929 - 0.7679
Male Sex	Precision	Ameba	0.6415 - 0.8040
Male Sex	BA	Bacteria	0.6180 - 0.6754
Male Sex	BA	Fungi	0.7061 - 0.7580
Male Sex	BA	Ameba	0.7631 - 0.8225
Male Sex	ACC	Bacteria	0.7977 - 0.8210
Male Sex	ACC	Fungi	0.7775 - 0.8140
Male Sex	ACC	Ameba	0.9126 - 0.9351
Male Sex	AUROC	Bacteria	0.7412 - 0.7757
Male Sex	AUROC	Fungi	0.8228 - 0.8602
Male Sex	AUROC	Ameba	0.9196 - 0.9522

Table A.4: Confidence Intervals for Age Subsets, Tasks, and Metrics of Multitask v2 with clinical loss.

Subset	Metric	Task	Confidence Interval
0-18 yo	F1	Bacteria	0.9066 - 0.9820
0-18 yo	F1	Fungi	nan - nan
0-18 yo	F1	Ameba	0.6123 - 0.9613
0-18 yo	Recall	Bacteria	0.9762 - 1.0072
0-18 yo	Recall	Fungi	nan - nan
0-18 yo	Recall	Ameba	0.6224 - 1.0176
0-18 yo	Precision	Bacteria	0.8419 - 0.9756
0-18 yo	Precision	Fungi	nan - nan
0-18 yo	Precision	Ameba	0.6100 - 0.9767
0-18 yo	BA	Bacteria	0.6247 - 0.9087
0-18 yo	BA	Fungi	0.9762 - 1.0072
0-18 yo	BA	Ameba	0.8791 - 0.9767
0-18 yo	ACC	Bacteria	0.8374 - 0.9657
0-18 yo	ACC	Fungi	0.9762 - 1.0072
0-18 yo	ACC	Ameba	0.8724 - 0.9716
0-18 yo	AUROC	Bacteria	nan - nan
0-18 yo	AUROC	Fungi	nan - nan
0-18 yo	AUROC	Ameba	nan - nan
18-40 yo	F1	Bacteria	0.8442 - 0.8747
18-40 yo	F1	Fungi	0.4401 - 0.5589
18-40 yo	F1	Ameba	0.7980 - 0.8563
18-40 yo	Recall	Bacteria	0.9120 - 0.9553
18-40 yo	Recall	Fungi	0.4241 - 0.5617
18-40 yo	Recall	Ameba	0.7529 - 0.8379
18-40 yo	Precision	Bacteria	0.7694 - 0.8282
18-40 yo	Precision	Fungi	0.4479 - 0.5837
18-40 yo	Precision	Ameba	0.8371 - 0.8935
18-40 yo	BA	Bacteria	0.6230 - 0.6855
18-40 yo	BA	Fungi	0.6870 - 0.7593
18-40 yo	BA	Ameba	0.8340 - 0.8777
18-40 yo	ACC	Bacteria	0.7555 - 0.8018
18-40 yo	ACC	Fungi	0.8983 - 0.9239
18-40 yo	ACC	Ameba	0.8498 - 0.8874
18-40 yo	AUROC	Bacteria	0.7503 - 0.7987
18-40 yo	AUROC	Fungi	0.8628 - 0.9077

Subset	Metric	Task	Confidence Interval
18-40 yo	AUROC	Ameba	0.9249 - 0.9491
40-65 yo	F1	Bacteria	0.8481 - 0.8760
40-65 yo	F1	Fungi	0.6115 - 0.6929
40-65 yo	F1	Ameba	0.4701 - 0.6494
40-65 yo	Recall	Bacteria	0.9347 - 0.9737
40-65 yo	Recall	Fungi	0.5176 - 0.6182
40-65 yo	Recall	Ameba	0.4674 - 0.6516
40-65 yo	Precision	Bacteria	0.7638 - 0.8118
40-65 yo	Precision	Fungi	0.7317 - 0.8208
40-65 yo	Precision	Ameba	0.4757 - 0.6765
40-65 yo	BA	Bacteria	0.5995 - 0.6487
40-65 yo	BA	Fungi	0.7058 - 0.7617
40-65 yo	BA	Ameba	0.7113 - 0.8076
40-65 yo	ACC	Bacteria	0.7564 - 0.7973
40-65 yo	ACC	Fungi	0.7462 - 0.8022
40-65 yo	ACC	Ameba	0.9082 - 0.9413
40-65 yo	AUROC	Bacteria	0.6929 - 0.7551
40-65 yo	AUROC	Fungi	0.8008 - 0.8640
40-65 yo	AUROC	Ameba	0.8918 - 0.9492
+65 yo	F1	Bacteria	0.9247 - 0.9486
+65 yo	F1	Fungi	0.4451 - 0.6062
+65 yo	F1	Ameba	0.2928 - 0.7595
+65 yo	Recall	Bacteria	0.9629 - 0.9841
+65 yo	Recall	Fungi	0.3755 - 0.5570
+65 yo	Recall	Ameba	0.2938 - 0.8145
+65 yo	Precision	Bacteria	0.8829 - 0.9234
+65 yo	Precision	Fungi	0.5868 - 0.6778
+65 yo	Precision	Ameba	0.2996 - 0.7871
+65 yo	BA	Bacteria	0.5839 - 0.6646
+65 yo	BA	Fungi	0.6480 - 0.7257
+65 yo	BA	Ameba	0.7656 - 0.9727
+65 yo	ACC	Bacteria	0.8668 - 0.9051
+65 yo	ACC	Fungi	0.7779 - 0.8111
+65 yo	ACC	Ameba	0.9641 - 0.9916
+65 yo	AUROC	Bacteria	0.7014 - 0.7784
+65 yo	AUROC	Fungi	0.7486 - 0.8298
+65 yo	AUROC	Ameba	nan - nan

A.5 P-values for the other methodologies

Table A.5: P-values for Task Bacteria by Sex.

Metric	P-value
F1	1.0000
Recall	1.0000
Precision	1.0000
BA	1.0000
ACC	1.0000
AUROC	1.0000

Table A.6: P-values for Task Bacteria by Age.

Metric	P-value
F1	1.0000
Recall	1.0000
Precision	0.1561
BA	1.0000
ACC	1.0000
AUROC	NaN

Table A.7: P-values for Task Fungi by Sex.

Metric	P-value
F1	0.0228
Recall	0.1514
Precision	0.1335
BA	0.4581
ACC	0.0003
AUROC	0.6748

Table A.8: P-values for Task Fungi by Age.

Metric	P-value
F1	1.9409e-14
Recall	1.7515e-09
Precision	3.0711e-14
BA	1.1107e-06
ACC	6.5465e-10
AUROC	NaN

Table A.9: P-values for Task Ameba by Sex.

Metric	P-value
F1	1.0000
Recall	1.0000
Precision	1.0000
BA	1.0000
ACC	1.0000
AUROC	1.0000

Table A.10: P-values for Task Ameba by Age.

Metric	P-value
F1	NaN
Recall	NaN
Precision	NaN
BA	NaN
ACC	NaN
AUROC	NaN

Table A.11: P-values for Tasks by Sex for Multitask V1.

Metric	Bacteria	Fungi	Ameba
F1	1.0000	0.0095	0.0140
Recall	1.0000	0.0536	0.0897
Precision	1.0000	0.0779	0.1346
BA	1.0000	0.2799	0.2094
ACC	1.0000	0.0001	0.2685
AUROC	1.0000	1.0000	1.0000

Table A.12: P-values for Tasks by Age for Multitask V1.

Metric	Bacteria	Fungi	Ameba
F1	0.0057	1.7757e-12	0.9424
Recall	0.1564	7.5369e-11	1.0000
Precision	0.0055	7.2705e-13	0.6279
BA	0.9424	1.7155e-11	1.0000
ACC	0.0152	8.7081e-14	0.0243
AUROC	NaN	NaN	NaN

Table A.13: P-values for Tasks by Sex for Multitask V2.

Metric	Bacteria	Fungi	Ameba
F1	1.0000	0.0078	0.0010
Recall	1.0000	0.3441	0.0068
Precision	1.0000	0.0031	0.1331
BA	1.0000	0.9866	0.0147
ACC	1.0000	0.0001	1.0000
AUROC	1.0000	1.0000	0.1954

Table A.14: P-values for Tasks by Age for Multitask V2.

Metric	Bacteria	Fungi	Ameba
F1	0.0000	0.0000	0.0548
Recall	0.0010	0.0000	0.1260
Precision	0.0001	0.0000	0.1295
BA	1.0000	0.0000	1.0000
ACC	0.0000	0.0000	0.0326
AUROC	NaN	NaN	NaN

Table A.15: P-values for Tasks by Sex for Multitask V1 with clinical loss.

Metric	Bacteria	Fungi	Ameba
F1	1.0000	0.0141	0.0046
Recall	1.0000	0.1126	0.1144
Precision	1.0000	0.0252	0.0471
BA	1.0000	0.3514	0.1836
ACC	1.0000	0.0035	0.3416
AUROC	1.0000	1.0000	0.7414

Table A.16: P-values for Tasks by Age for Multitask V1 with clinical loss.

Metric	Bacteria	Fungi	Ameba
F1	0.0473	1.2518e-12	0.0494
Recall	0.3486	2.1886e-09	0.0219
Precision	0.0484	6.0171e-13	0.7093
BA	1.0000	2.1322e-08	0.3482
ACC	0.0585	1.5599e-07	0.0040
AUROC	NaN	NaN	NaN

Table A.17: P-values for Tasks by Sex for Multitask V1 with clinical loss and adaptive threshold.

Metric	Bacteria	Fungi	Ameba
F1	1.0000	0.0000	0.0227
Recall	1.0000	0.8684	1.0000
Precision	1.0000	0.0097	0.0233
BA	1.0000	1.0000	1.0000
ACC	1.0000	1.0000	1.0000
AUROC	1.0000	1.0000	1.0000

Table A.18: P-values for Tasks by Age for Multitask V1 with clinical loss and adaptive threshold.

Metric	Bacteria	Fungi	Ameba
F1	0.0173	0.0000	0.1295
Recall	0.5870	0.0000	0.9693
Precision	0.0000	0.0000	0.1243
BA	0.0001	0.0240	1.0000
ACC	0.2973	0.0171	0.9661
AUROC	NaN	NaN	NaN

Table A.19: P-values for Tasks by Sex for Multitask V2 with clinical loss and adaptive threshold.

Metric	Bacteria	Fungi	Ameba
F1	1.0000	3.08e-11	9.50e-11
Recall	1.0000	NaN	NaN
Precision	1.0000	2.43e-11	2.15e-10
BA	1.0000	NaN	NaN
ACC	1.0000	2.35e-11	2.07e-10
AUROC	1.0000	1.0000	1.0000

Table A.20: P-values for Tasks by Age for Multitask V2 with clinical loss and adaptive threshold.

Metric	Bacteria	Fungi	Ameba
F1	1.0000	2.98e-26	3.30e-08
Recall	1.0000	0.0000	1.0000
Precision	0.0029	3.81e-24	1.56e-05
BA	1.0000	0.0000	0.9452
ACC	1.0000	3.70e-24	1.50e-05
AUROC	NaN	NaN	NaN