

# Classification of Keratitis from Eye Corneal Photographs using Deep Learning

Maria Miguel Beirão  
*University of Porto*  
*INESC TEC*  
 Porto, Portugal  
 maria.m.beirao@inesctec.pt

Luis Filipe Nakayama  
*MIT*  
*Universidade Federal de São Paulo*  
 Cambridge, MA, USA  
 luis.naka@mit.edu

João Matos  
*MIT*  
 Cambridge, MA, USA  
 jcmatos@mit.edu

Tiago Gonçalves  
*University of Porto*  
*INESC TEC*  
 Porto, Portugal  
 tiago.f.goncalves@inesctec.pt

Camila Kase  
*Universidade Federal de São Paulo*  
 São Paulo, Brazil  
 camila.kase@unifesp.br

Denise de Freitas  
*Universidade Federal de São Paulo*  
 São Paulo, Brazil  
 dfreitas.epm.unifesp@gmail.com

Jaime S. Cardoso  
*University of Porto*  
*INESC TEC*  
 Porto, Portugal  
 jsc@fe.up.pt

**Abstract**—Keratitis is an inflammatory corneal condition responsible for 10% of visual impairment in low- and middle-income countries (LMICs), with bacteria, fungi, or amoeba as the most common infection etiologies. While an accurate and timely diagnosis is crucial for the selected treatment and the patients' 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. In this study, we investigate and compare different deep learning approaches to diagnose the source of infection: 1) three separate binary models for infection type predictions; 2) a multitask model with a shared backbone and three parallel classification layers (Multitask V1); and, 3) a multitask model with a shared backbone and a multi-head classification layer (Multitask V2). We used a private Brazilian cornea dataset to conduct the empirical evaluation. We achieved the best results with Multitask V2, with an area under the receiver operating characteristic curve (AUROC) confidence intervals of 0.7413-0.7740 (bacteria), 0.8395-0.8725 (fungi), and 0.9448-0.9616 (amoeba). A statistical analysis of the impact of patient features on models' performance revealed that sex significantly affects amoeba infection prediction, and age seems to affect fungi and bacteria predictions.

**Index Terms**—computer vision, deep neural networks, keratitis, machine learning, multitask learning

## I. INTRODUCTION

According to the World Health Organization (WHO) [12], corneal blindness impacts 6M people worldwide, with *keratitis* as its leading cause [19], [20], with different consequences in low, middle and high-income countries [1]. Keratitis (or corneal ulcer) is an inflammation of the cornea caused by microorganisms (e.g., bacteria (BK), viruses (VK), fungi (FK), amoeba (AK), or herpes simplex keratitis (HSK)) or by non-infectious factors (e.g., trauma or chemical exposure) [11]. This disease is more common in LMICs, with its risk factors (e.g., compromised immune systems, poor ocular hygiene,

contact lens wear, eye trauma, and surgery, exposure to contaminated water and soils and agricultural work) [9], [14] being correlated with the context of rural areas of these geographies [19]. The symptoms of keratitis include eye redness, eye pain, excess tears or discharge, foreign body sensation, difficulty opening the eyelid, blurred vision, light sensitivity, and watery eyes [9]. While the symptoms are well-defined, the main challenge with this disease is related to the diagnosis of the etiology, a crucial detail for the treatment of the condition and prevention of vision loss. The clinical pipeline for determining the etiology includes techniques such as corneal scrapping with gram stain, microscopy, and culture analysis. However, given their substantial costs (i.e., approximately \$200 per patient in Brazil), these resources are often unavailable in LMICs. The alternative approach relies on performing a clinical examination with a slit lamp, with bright light and enough magnification to observe the infection. Nevertheless, distinguishing between the types of infection remains difficult and might lead to a misdiagnosis, worsening the symptoms and requiring an eye surgery [1], [4], [18]. In this paper, we explore the potential of deep learning to classify keratitis etiologies using eye corneal photographs, making the following scientific contributions:

- 1) Proposal and implementation of deep learning algorithms to classify different etiologies of keratitis from corneal eye photographs, using a new private Brazilian cornea dataset with clinical attributes and annotations;
- 2) Assessment of the statistical impact of correlated features (i.e., age and sex) on disease presentation.

The code related to the implementation of this work is publicly available in a GitHub repository<sup>1</sup>.

This work was financed by National Funds through the Portuguese funding agency, Portuguese Foundation for Science and Technology (FCT) through the Ph.D. Grant “2020.06434.BD”.

<sup>1</sup>[https://github.com/mariamiguel01/Keratitis\\_Classifier](https://github.com/mariamiguel01/Keratitis_Classifier)

## II. RELATED WORK

### A. Literature Review on Keratitis Classification

In this section, we present a literature review on the state-of-the-art deep learning algorithms applied to keratitis classification in external eye photographs and organize our findings divided by predictive task: a) *BK versus FK*, and b) *Multiple Infections*.

a) *BK versus FK*: Reed et al. [16] developed a solution that uses a data augmentation pipeline and regularization techniques to fine-tune different convolutional neural network (CNN) architectures. After hyper-parameter optimization, their best model achieved an AUROC of 83%. In another study conducted on a different Indian dataset comprising 50% BK and 50% FK, Reed et al. [15] leveraged transfer learning on a pre-trained deep learning system and reported an accuracy of 76%.

b) *Multiple Infections*: Koyama et al. [10] developed a deep learning algorithm to classify BK, FK, AK, and HSK using pre-trained CNNs and a dataset containing 4,306 images from clinical cases and web sources under different illumination conditions. Under a k-fold cross-validation strategy, they trained a hierarchical two-step CNN classifier to handle class imbalance, with the first CNN discriminating BK vs. non-bacterial cases and the second categorizing non-bacterial cases into AL, FK, or HSK. The final classifier, a gradient-boosting decision tree (GBDT), uses the CNN scores.

### B. Ophthalmologists versus Artificial Intelligence

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 that of the deep learning system and much more dependent on other variables like years of expertise and familiarity with the disease. Since this performance is very dependent on the dataset, several studies compared the model and the experts' performance. Redd et al. [16] compared the performance of 66 international expert cornea specialists in the same Indian dataset used in [15]. The panel of specialists achieved an AUROC of 76%, which is significantly lower ( $p < 0.01$ ) than the one found by the deep learning model (83%).

## III. METHODOLOGY

This study was conducted in accordance with the principles of the Declaration of Helsinki and was approved by the Research Ethics Committee of the Federal University of São Paulo (UNIFESP), Brazil.

### A. Dataset

1) *Description*: Two ophthalmologist co-authors (Camila Kase and Luis Nakayama) acquired the data at UNIFESP. This dataset contains healthy and non-healthy patients for keratitis (i.e., AK, BK, and FK), confirmed by corneal scraping culture. The dataset contains 24,692 exams, belonging to 3,296 patients (i.e., multiple exams per patient), associated with the patient's clinical data and the corresponding eye corneal images. From the total exams, we used the subset that had

medical information and tested positive to keratitis (4,767 exams). After validation with the clinicians, we realized that the same image corresponded to different patient visits, and that an image is labeled as positive for an infection if any test for that infection is positive. These details allowed us to clean the dataset from redundant or non-relevant entries, and achieve a final dataset of 2,585 unique cases. We also excluded the cases with no infection (512 cases), resulting in a final dataset of 2,064 cases, the proportions of each class are present in Table I. Figures 1 - 6 show some examples of the dataset.



Fig. 1. Bacteria Dataset Example.



Fig. 2. Fungi Dataset Example.

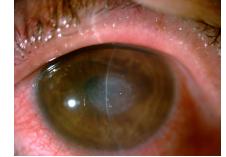


Fig. 3. Amoeba Dataset Example.



Fig. 4. Bacteria Dataset Example.



Fig. 5. Fungi Dataset Example.



Fig. 6. Amoeba Dataset Example.

TABLE I  
NUMBER AND PROPORTION OF CASES FOR THE POSSIBLE COMBINATIONS OF INFECTIONS (ABSENT COMBINATIONS DID NOT OCCUR IN THE DATASET).

Bacteria	Amoeba	Fungi	# Cases	% Cases
Pos.	Neg.	Neg.	1176	56.98%
Neg.	Neg.	Pos.	277	13.42%
Pos.	Neg.	Pos.	213	10.32%
Neg.	Pos.	Neg.	207	10.03%
Pos.	Pos.	Neg.	191	9.26%

2) *Exploratory Data Analysis*: Following clinical advice, we focus our exploratory data analysis on *age* and *sex*. We analyzed the age distribution using a kernel density estimation, observing a high peak of ages 20-40 and 40-65, which can be related to the two most common sources of infection (i.e., inadequate use of contact lenses and agricultural work). For implementation purposes, we grouped age values into four bins: 0 – 18 (label 0), 18 – 40 (label 1), 40 – 65 (label 2), and > 65 (label 3). The majority of cases exist in label 2 (39.39%) and label 1 (36.05%), while label 0 is underrepresented (2.62%). Regarding sex, our analysis showed that there are 1,203 *male* patients (58.28%) and 861 *female* patients (41.72%) (i.e., balanced dataset for sex). We also performed a correlation analysis between the clinical features and the etiologies to understand the intrinsic biases of the dataset. Results showed positive and negative correlations between the etiologies and the clinical features (e.g., sex or

age). Naturally, it is important to denote that there is a high correlation (both positive and negative) between the different types of infections.

### B. Implementation

1) *Models*: We propose and implement two different approaches to achieve deep learning models to predict keratitis from eye cornea photographs:

- **Baseline approach**: three simple binary deep learning models, one for each type of infection. Diagnosis for all infections requires the use of the three models.
- **Multi-task approach**: a multi-label deep learning model that predicts all the infections in a single forward pass, reducing runtime and computational costs. Within this approach, we tested two different architectures: 1) *Multi-task V1*, containing three binary classification layers, one per infection type; and 2) *Multi-task V2*, containing a classification layer with three neurons.

Besides, following the correlation analysis between clinical features and the clinical outcomes, we also developed models to predict age (multi-class model) and sex (binary model) from the same images to understand if we could mitigate potential biases [13]. To select an appropriate backbone (i.e., the feature extractor) for the models, we trained three candidates (i.e., DenseNet121 [8], ResNet50 [6] and VGG16 [17]) in a small subset (i.e., 50 images) of the dataset and recorded their runtime and performance. We selected the backbone with a lower runtime and better performance. After the backbone, we added a 2D average pooling layer with a kernel of  $3 \times 3$ , stride of  $2 \times 2$ , and padding of 1, followed by a batch normalization layer and a dropout layer with a probability of 0.3. Figure 7 shows the classification layer of each of the approaches.

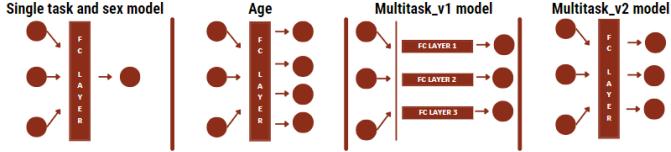


Fig. 7. Schematic overview of the fully connected layers of each model, from left to right: **Baseline approach** and **Sex prediction** - a fully connected layer with a single neuron (and a sigmoid activation function); **Age prediction** - a fully connected layer with four neurons (and a softmax activation function); **Multitask\_v1** - three parallel fully connected layers, each layer with a single neuron (and a sigmoid activation function), i.e., each layer predicting the binary outcome of a given infection; **Multitask\_v2** - a fully connected layer with three neurons (and a sigmoid activation function), i.e., each neuron predicting the binary outcome of a given infection.

2) *Data Preparation and Augmentation*: Given the reduced number of usable cases (i.e., 2,064 out of 24,692), we started by artificially creating photographs of “left” and “right” eyes per patient, by applying a flip operation on the  $y$ -axis to the available images in the dataset (each patient only had an image of an eye). Besides increasing the size of the dataset, it also allows the model to be robust against images of both eyes. Afterwards, we resized all the images in the dataset to  $224 \times$

$224$ . We present the conversion of labels for the models in Table II.

TABLE II  
CONVERSION OF THE CLINICAL ATTRIBUTES INTO NUMERICAL LABELS.

Clinical Attribute	Transformed (Original) Data
Age	0 (0-17), 1 (18-39), 2 (40-64), 3 ( $>65$ )
Infection Test Result	0 (Negative), 1 (Positive)
Sex	0 (Male), 1 (Female)

We applied online data augmentation during training. The pipeline employed in this work consisted of random rotations (up to 20 degrees), random vertical flips (with a probability of 0.5), Gaussian blur ( $kernel = 5 \times 5$ ,  $0.1 < \sigma < 2$ ) and random color transformations (i.e., brightness, contrast, saturation, and hue adjustments) with the same parameters described in the literature [7]. We also applied a z-score normalization to the images using ImageNet’s mean ([0.485, 0.456, 0.406]) and standard deviation ([0.229, 0.224, 0.225]).

### C. Training Details

1) *K-fold Strategy*: We implemented a k-fold strategy ( $k = 10$ ) for all the experiments. For the single-task models, we preserved the class distribution in the splits, while for the multi-task (i.e., multi-label), we ensured that every class and combination of classes were properly balanced. For each fold, we retained 10% of the dataset as the test set to validate the best model, and, we split the remaining dataset to achieve, 80% for train and 10% for validation. This division was done in the dataset without the flipped images, and those were later added to the subset where the non-flipped image was, to avoid data leakage.

2) *Training and Validation Pipeline*: We trained the models for a certain number of epochs (between 100 and 300) and defined the optimal case for each. We initialized the models with ImageNet pre-trained weights. We have frozen the backbone of the models during the first ten epochs. Regarding loss functions, we used the binary cross-entropy for the binary classification and multi-task (multi-label) models, and cross-entropy for the multi-class models (age). To mitigate the issues related to class imbalance, we used class weights in the loss functions for all models. We used the adaptive moment estimation (Adam) optimizer and defined the best hyper-parameters (e.g., learning rate and weight decay) for each model by monitoring the validation loss (i.e., the minimum). We saved the best model according to the best value in the validation loss and an “early-stopping” routine. We used a batch-size of 16 for all the models.

3) *Multi-task Optimization*: Apart from the details described above, we added a few improvements to the multi-task models. We complemented the binary cross-entropy loss with a new term that takes into account that the cost of misidentifying a high-cost infection should be higher than that of misidentifying a low-cost infection. Adding this term to the loss makes sense because keratitis is common in LMICs, meaning that an incorrect treatment plan will have severe

financial impacts on the patient and the healthcare providers. The average cost of each infection was computed using the insights provided by our clinical partners. To do so, we multiplied the prices of the medicines used in each infection (i.e., R\$45.2 for bacterial keratitis, R\$203 for fungal keratitis, and R\$95.5 for amoeba keratitis) by the number of flasks of each medicine used and the maximum duration of the treatment in months. We then divided these values by the cost of all infections. The values obtained were considered the hospital costs. After that, the combined loss was calculated using Equation 1:

$$\mathcal{L}_{total} = 0.8 \cdot (C_w \cdot \mathcal{L}_{BCE}) + 0.2 \cdot (H_w \cdot \mathcal{L}_{BCE}), \quad (1)$$

where  $\mathcal{L}_{total}$  is the total loss,  $C_w$  is the class weights,  $\mathcal{L}_{BCE}$  is the binary cross-entropy loss, and  $H_w$  is the hospital weights (described above). Finally, we integrated an adaptive threshold solution to consider a certain probability of a prediction. In this case, the optimal threshold would maximise the true positive rate (TPR) and minimise the false positive rate (FPR). To find this threshold, we computed Youden's J for every threshold value, comprising the receiver operating characteristic (ROC) curves of each task. The threshold that maximised J was then considered the optimal classification threshold for that infection [3].

#### D. Evaluation Details

1) *Performance Metrics*: To evaluate the predictive performance of the models, we computed the accuracy (ACC), AUROC, balanced accuracy (BA), F1-score, mean absolute error (MAE), precision, and recall [2]. During the training phase, we used these metrics and the evolution of the loss (in both training and validation splits) to understand the evolution of the model. In the test set, we computed these metrics, along with the confusion matrices (CFs) and the F1-score and recall for each class (within the same label). For the multi-task models, we also generated the joint CF of all infections. Lastly, to understand the TPR and FPR, we plotted the ROC curves. To ensure and validate the models' diagnosis, we generated saliency maps for every fold and one example of each class's positive and negative case.

2) *Statistical Significance*: Since we implemented a k-fold strategy, we computed the average for all these metrics (and confusion matrices) across folds. We computed a 95% normal confidence interval (CI) for the results of each metric and task, across models. After that, and only on models related to keratitis classification, we studied the statistical significance of the sex and age for the diagnosis. 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, we grouped the age and sex subsets. For the sex analysis, since only two independent groups were being compared, we used the t-test, whereas, for the age analysis, Analysis of Variance (ANOVA) was used. To account for the family-wise error

rate of conducting multiple tests on the same dataset, Holm-Bonferroni Correction was implemented [5].

3) *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, ACC, BA, and MAE were computed by sex and age 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.

## IV. RESULTS AND DISCUSSION

We obtained the best results with the DenseNet121 backbone, trained for 200 epochs, using Adam optimiser with a learning rate of  $1 \times 10^{-6}$ , and a weight decay of  $1 \times 10^{-8}$ .

#### A. Sex and Age Prediction

For the tasks of sex and age prediction, our results suggest that the age bin is less predictable than the sex from the eye photograph (see Figure 8). On the other hand, unexpectedly, results for both tasks were fairly good, meaning that sex and age can be derived from eye corneal photos, and potentially influencing the prediction for keratitis, confirming the need for subset analysis. A look at the AUROCs, we can see that for sex the AUROC lies in  $0.8790 - 0.8994$  and for age in  $0.8331 - 0.8624$ . For the MAE across folds, a CI of  $0.2891 - 0.3114$  was found, which is low. This potentially means that when an incorrect classification happens, it does not occur for the most erroneous case. To validate this assumption, we looked into the CFs.

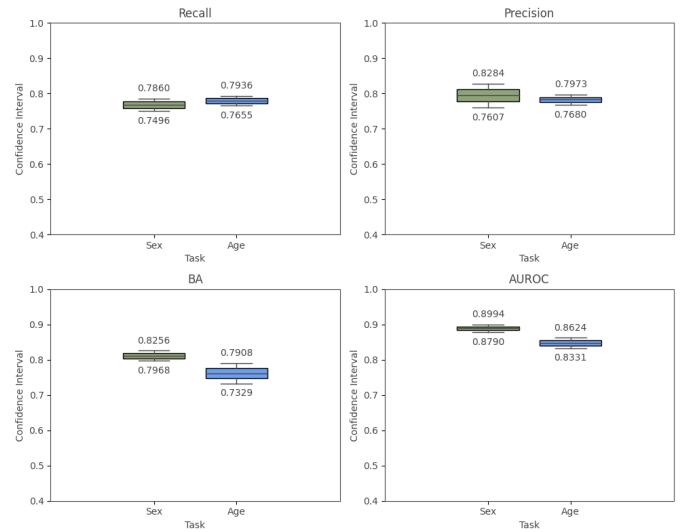


Fig. 8. Results for the sex and age prediction models.

#### B. Keratitis Classification

In the analysis of the results for the keratitis classification models, we focused on the precision and recall for all the etiologies because those were the metrics where the most

differences were found. We excluded the baseline approach (single-task) from the pool of best models, since it needed to run three times to output a complete diagnosis, and did not use potential correlations between the etiologies to make the prediction, which could jeopardize the multi-infection case. From an attentive look at Figure 9, we realize that both Multi-task V1 and V2 models had the worst results for amoeba when compared to the models that had the optimal loss optimization. Besides, we observe an increase in incorrect classification (for all tasks) for the models that had the adaptive threshold, portrayed by a decrease of recall for bacteria and amoeba and a decrease in precision for fungi. Multi-task V1 and V2 with the clinical loss (see Equation 1) attained very similar results, with the V2 being slightly better for recall in fungi and amoeba, and precision in bacteria. This model obtained AUROCs of

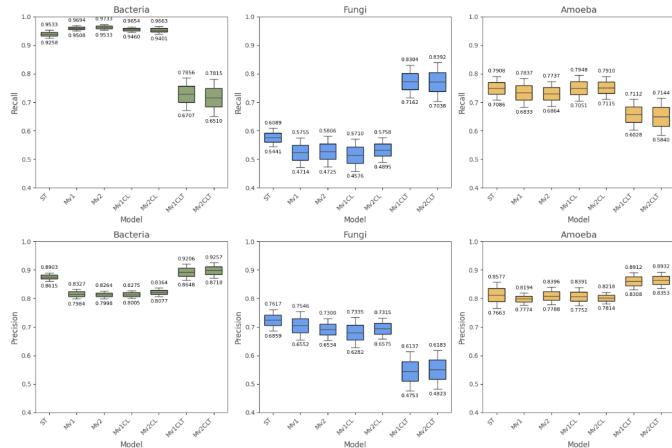


Fig. 9. Recall and Precision of all models. Note: **ST**-Single task, **Mv1**-MultitaskV1, **Mv2**-MultitaskV2, **Mv1CL**-MultitaskV1 with clinical loss, **Mv2CL**-MultitaskV2 with clinical loss, **Mv1CLT**-MultitaskV1 with clinical loss and adaptive threshold, **Mv2CLT**-MultitaskV2 with clinical loss and adaptive threshold

0.9448 – 0.9617 for amoeba, 0.8395 – 0.8725 for fungi, and 0.7413 – 0.7740 for bacteria. If we focus our attention on recall and precision, we observe that recall was especially relevant for amoeba and fungi, and precision for bacteria (possibly explained by the data imbalance). For instance, the low values found for fungi mean that there are several false negatives (FNs). This analysis is complemented with the CF (see Table IV-B), which shows us that the model struggles with predicting the positive class correctly (once again, possibly due to data imbalance). Please note that we can derive similar conclusions for amoeba and bacteria with respect to the false positives (FPs). Looking at the extended CF (see Table IV), we can see that incorrect classification often happens because the model can not distinguish between a single or combined infection: for instance, from all the fungi infections, 19 were predicted as bacteria infections, 12 as fungi infections, and 21 as a combined infection of bacteria and fungi. These results support the need for more data.

TABLE III  
AVERAGE CFs FOR THE MULTI-TASK V2 MODEL TRAINED WITH THE CLINICAL LOSS (SEE EQUATION 1), FOR THE DIFFERENT TASKS.

Task	Actual/Predicted	Negative	Positive
<b>Amoeba</b>	Negative	318.4	14.8
	Positive	19.8	59.8
<b>Bacteria</b>	Negative	31.1	65.7
	Positive	14.8	301.2
<b>Fungi</b>	Negative	292.0	22.8
	Positive	45.8	52.2

TABLE IV  
CONFUSION MATRIX FOR MULTI-TASK V2 MODEL TRAINED WITH THE CLINICAL LOSS. NOTE: **H**=HEALTHY, **B**=BACTERIA, **F**=FUNGI, **A**=AMOEBA, **B,F**=COMBINED INFECTION OF BACTERIA AND FUNGI, **F,A**=COMBINED INFECTION OF FUNGI AND AMOEBA, **B,A**=COMBINED INFECTION OF BACTERIA AND AMOEBA, **B,F,A**=COMBINED INFECTION OF BACTERIA, FUNGI AND AMOEBA.

Label\Pred	H	B	F	A	B,F	F,A	B,A	B,F,A
<b>H</b>	0	0	0	0	0	0	0	0
<b>B</b>	4	193	2	2	24	0	9	0
<b>F</b>	0	19	12	0	21	0	2	0
<b>A</b>	4	13	1	11	0	0	11	0
<b>B,F</b>	0	20	0	0	24	0	0	0
<b>F,A</b>	0	0	0	0	0	0	0	0
<b>B,A</b>	1	5	0	2	0	0	30	0
<b>B,F,A</b>	0	0	0	0	0	0	0	0

### C. Influence of Patient Features

We chose the best-performing methodology for the identity feature evaluation. Given the interesting performance of the sex and age prediction models (i.e., the model was detecting identity features), we sought to assess if the classification model was using these features to output its predictions (i.e., the introduction of bias).

1) *Influence of Sex*: Table V presents the summary of the performance metrics for sex (i.e., p-values after correction). We can observe that sex has an impact on the ACC of the detection of fungi infection, with male sex having lower values. Regarding the detection of amoeba infection, we found a significant difference for F1-score, recall and BA. This can potentially mean that the model is using features related to the sex of the patient in the final predictions (especially for amoeba), meaning that a disentanglement of these features from the corneal eye photograph could boost the performance of the model.

2) *Impact of Age*: Table V presents the summary of the performance metrics for age (i.e., p-values after correction). We found significant differences for all metrics across all age groups, with the exception of BA (i.e., age features may play a role in the prediction). The analysis of the values for the detection of fungi infection reported the same as the detection of bacteria infection, with all metrics exhibiting significant differences across age groups. Regarding the detection of amoeba infection, we found significant value for ACC. Please note that we do not report AUROC values for this task because

of the lack of data on at least one age bin. Moreover, given the class imbalance, we argue that feature disentanglement is needed for this case, although an evaluation with a bigger and balanced dataset should also be done.

TABLE V  
CORRECTED P-VALUES FOR AGE AND SEX BY INFECTION TYPE.

	Bacteria		Fungi		Amoeba	
Metric	Age	Sex	Age	Sex	Age	Sex
F1-score	<b>0.0001</b>	1.000	<b>7.70 × 10<sup>-16</sup></b>	0.2510	0.4555	<b>0.0035</b>
Recall	<b>0.0153</b>	1.000	<b>1.63 × 10<sup>-12</sup></b>	0.7070	0.9099	<b>0.0010</b>
Precision	<b>0.0029</b>	1.000	<b>2.93 × 10<sup>-20</sup></b>	0.1320	0.4946	0.2580
BA	0.7693	1.000	<b>1.24 × 10<sup>-13</sup></b>	1.000	0.2418	<b>0.0042</b>
ACC	<b>0.0006</b>	1.000	<b>6.83 × 10<sup>-16</sup></b>	<b>0.0003</b>	<b>0.0055</b>	0.6840
AUROC	-	1.000	-	1.000	-	0.4810

An analysis of the results for the other frameworks revealed similar patterns, making clear the need for further investigation of both sex and age in a larger dataset.

## V. CONCLUSIONS AND FUTURE WORK

Although the proposed solution has yielded good results, several improvements could be made in the future to reduce FPs for the detection of bacteria infection and FNs for the detection of fungi and amoeba infections. As explained earlier, we believe that the problem is caused by the lack of data on fungi and amoeba infections, causing the model to not predict these classes correctly. Since we do not train the model with healthy patients, when this happens, the eye exams are often classified as bacterial infections, causing FPs. Hence, the first step in future work would be to either acquire more data or artificially generate new images with generative adversarial networks (GANs) or diffusion models. Another interesting test would be to evaluate the performance of the model on segmented images of the cornea, as well as test other state of the art backbones. It would also be interesting to introduce the sex and age of that patient as auxiliary inputs to the prediction model and evaluate the performance. Lastly, the developed work has proven that it is possible to predict the sex and age of the patient based on an eye photograph. This could mean that those identity features are present in the images, introducing new biases to our methodology. The obtained p-values pointed in that direction. To avoid those connections, the disentanglement of identity features from the eye corneal photographs could be relevant.

In conclusion, in this work, we demonstrated 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 amoeba 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, interesting AUROC values were found, 0.8790 – 0.8994 for sex and 0.8331 – 0.8624 for age. A significant impact of sex for amoeba infections and age for

bacteria and fungi infections was found, meaning a bias associated with patient features could be implied.

## REFERENCES

- [1] A Austin, T Lietman, and J Rose-Nussbaumer. Update on the management of infectious keratitis. *Ophthalmology*, 124:1678–1689, 2017.
- [2] Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.
- [3] Jason Brownlee. Threshold moving for imbalanced classification. <https://machinelearningmastery.com/threshold-moving-for-imbalanced-classification/>, 2020. Accessed: 2024-06-18.
- [4] C Dalmon, TC Porco, TM Lietman, N Venkatesh Prajna, L Prajna, MR Das, et al. The Clinical Differentiation of Bacterial and Fungal Keratitis: A Photographic Survey. *Investigative Ophthalmology & Visual Science*, 53:1787, 2012.
- [5] Haynes, Winston. Holm's Method. In Dubitzky, Werner and Wolkenhauer, Olaf and Cho, Kwang-Hyun and Yokota, Hiroki, editor, *Encyclopedia of Systems Biology*, pages 902–902. Springer New York, New York, NY, 2013.
- [6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [7] Shaodan Hu, Yiming Sun, Jinhao Li, Peifang Xu, Mingyu Xu, Yifan Zhou, Yaqi Wang, Shuai Wang, and Juan Ye. Automatic diagnosis of infectious keratitis based on slit lamp images analysis. *Journal of Personalized Medicine*, 13(3), 2023.
- [8] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [9] Johns Hopkins Medicine. Keratitis. <https://www.hopkinsmedicine.org/health/conditions-and-diseases/keratitis>, 2024. [Accessed 2024-05-30].
- [10] 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*, 11:22642, 2021. Received: 27 August 2021, Accepted: 02 November 2021, Published: 22 November 2021.
- [11] Mayo Clinic. Keratitis. <https://www.mayoclinic.org/diseases-conditions/keratitis/symptoms-causes/syc-20374110>, 2024. [Accessed 23-11-2023].
- [12] World Health Organization. Blindness and visual impairment. <https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment>, 2023. Accessed: 2024-05-31.
- [13] Ryan Poplin, Avinash V. Varadarajan, Kevin Blumer, Yun Liu, Michael V. McConnell, Greg S. Corrado, Lily Peng, and Dale R. Webster. Predicting sex from retinal fundus photographs using automated deep learning. *Scientific Reports*, 11(1):10236, 2021.
- [14] Prevent Blindness. Keratitis. <https://preventblindness.org/keratitis/>. Accessed:2024-05-30.
- [15] Travis K. Redd, Luca Della Santina, N Venkatesh Prajna, Prajna Lalitha, Nisha Acharya, and Thomas Lietman. Automated differentiation of bacterial from fungal keratitis using deep learning. *Investigative Ophthalmology & Visual Science*, 62(8):2161, June 2021.
- [16] Travis K. Redd, N. Venkatesh Prajna, Muthiah Srinivasan, Prajna Lalitha, Tiruvengada Krishnan, Rengaraj Rajaraman, Anand Venugopal, Nisha Acharya, Gabriel D. Seitzman, Thomas M. Lietman, Jeremy D. Keenan, Jonathan P. Campbell, and Xian Song. Image-based differentiation of bacterial and fungal keratitis using deep convolutional neural networks. *Ophthalmol Sci*, 2(2):100119, Jan 2022.
- [17] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition, 2015.
- [18] Pratibha Singh, Anshu Gupta, and Krishna Tripathy. *Keratitis*. StatPearls Publishing, Treasure Island (FL), 2024.
- [19] Darren Shu Jeng Ting, Michelle Sze Yee Lim, Aina Hong Shek, Shu Yen Lee, Leonard Yan Ming Yip, Kha Leong Loh, Charline Zhi Xuan Wong, Marcus Ang, Desmond Shern Lung Ting, Lawrence Yuen Ka Ting, and Srinivasan Sanjay. Infectious keratitis: An update on epidemiology, causative microorganisms, risk factors, and antimicrobial resistance. *Eye (London, England)*, 35(4):1084–1101, 2021.
- [20] World Health Organization. Blindness and vision loss. [https://www.who.int/health-topics/blindness-and-vision-loss#tab=tab\\_3](https://www.who.int/health-topics/blindness-and-vision-loss#tab=tab_3), 2024. Accessed: 2024-05-30.