Machine Learning Classification of Bacterial vs Fungal Keratitis from Photographs

Kevin Supakkul (supakkul)
Kimmy Chang (kchang08)
Ryan Beauchamp (rmb87)
Mo Tiwari (PhD student: advisor)

1 Introduction

Infectious keratitis is a leading cause of blindness around the world, affecting 2 million people every year [1]. Such infections are treatable if the causative organism is diagnosed early and correctly [2]. Gold-standard clinical diagnoses such as Gram and Giemsa stains, however, are only accurate about 65% of the time [3]. The goal of this project was to develop a classification model to accurately distinguish bacterial infections from fungal infections from photographs, thereby enabling accurate treatment for patients. The input to the model was an image of an eye with an active keratitis infection. A neural network was used to output a prediction as to whether the infection was bacterial or fungal.

1.1 Prior work

Kuo et al. differentiated fungal from non-fungal keratitis using a deep learning model with a dataset of 114 fungal vs 174 non-fungal images [4]. The study was able to achieve 70% validation accuracy, surpassing the accuracy of clinical diagnosis. Their research verified that ML is indeed applicable for this use case, and they incorporated valuable insight from medical professionals. However, since they gathered their own images from patients, the model generalization beyond their dataset is unknown. Additionally, the 70% accuracy reported was a result of k-fold cross-validation ($k = 5$) and lacked a statistic detailing model generalization on a test set that was withheld from hyperparameter tuning; this may introduce bias as the hyperparameters may overfit to the validation set.

Xu et al. differentiated bacterial, fungal, and herpes simplex virus stromal keratitis [5]. The study included 387 bacterial images and 519 fungal images and was able to achieve 53.33% bacterial keratitis and 83.33% fungal keratitis validation recall. One compelling technique used in this article was sequential-level feature learning to identify and feed lesion-specific features into a deep learning model, though one downside is that it requires more manual labor for labeling datasets (from a trained medical provider).

This work makes the following novel contributions:

1. keratitis types - focused on fungal vs bacterial keratitis, whereas prior studies focused on multiple non-fungal varieties,
2. dataset - included a more diverse dataset than the single hospital used for collection in Kuo et al. [4], and experimented with approximately equal representation from each class which is different from Xu et al. [5],
3. methods - experimentation with different ML models, optimization, and regularization techniques than those cited in the articles,
4. statistical analysis - investigation of confidence intervals via bootstrap sampling,
5. results - the goal was for the lower bound of the 95% confidence interval to exceed the 70% validation accuracy seen in Kuo et al. [4] and have a higher bacterial recall than in Xu et al. [5].

2 Dataset

It was nontrivial to find a large quantity of high-quality photographs accurately labeled as active cases of bacterial or fungal keratitis. In total, 438 images were gathered consisting of 214 fungal and 224 bacterial photos. 113 of the images were received from Stanford Health Care, and the rest of the images were collected by searching through medical journals, medical case reports, and image search engines. In addition, the database was deduplicated to avoid the possibility of an image appearing in the train and test sets. Overall, this dataset is larger than Kuo et al. but smaller than Xu et al., though it is more evenly split than both [4, 5].

The dataset was randomly partitioned into a 60%: 20%: 20% train/validation/test split. Each trial evaluated the model's performance based on a randomly sampled 20% validation set. With the help of advisor, Mo Tiwari, there are plans to eventually publish this work pending further improvements; thus, this report does not touch the test dataset so that it can be kept for a final evaluation in the future (a decision that was confirmed by TA Ian Tullis). Thus, all the results presented in this report are validation results derived from the training/validation set.

3 Methods

3.1 Baseline Models

Prior work suggests deep learning would be the most effective model for accurately predicting bacterial vs fungal keratitis from images [4, 5]. Non-deep learning techniques including logistic regression and support vector machines were used as a quick baseline in the early stage. Here, pixels were flattened into a vector and used in the logistic regression and support vector machine models from the scikit-learn library.

3.1.1 Logistic Regression

The logistic regression model calculates the probability that an image is bacterial or fungal. It utilizes the sigmoid function:

$$g(z) = \frac{1}{1 + e^{-z}}$$

Parameters for logistic regression are fit according to maximimum likelihood via stochastic gradient descent until convergence.

3.1.2 Support Vector Machine

Similarly, raw image data was used to train a support vector machine using the scikit-learn library, as this is another classical approach to solving classification problems. Support Vector Machines (SVMs) operate under a different framework. Specifically, the SVM model tries to define a linear decision boundary between the classification of an image as bacterial or fungal that allows for the maximization of the number of correct images and degree of confidence (distance from the decision boundary) for the predictions on the training images. This can be framed as an optimization problem that attempts to maximize the minimum geometric margin[1] of each training image. Formally, this can be expressed as:

$$\max_{\gamma,w,b} \gamma$$
$$s.t. \quad y^{(i)}(w^T x^{(i)} + b) \geq \gamma, i = 1, ..., n$$
$$||w|| = 1$$

where γ is the functional margin, $y^{(i)}(w^T x^{(i)} + b)$ is the geometric margin for each training image, and the $||w||$ constraint ensures that the geometric margins are at least γ.

3.1.3 Results of Baseline Models

The upper bound for the 95% confidence intervals for these two methods was 60.8%. Since this is below the gold standard, a different approach was needed to achieve the modeling goals.

[1] The functional margin represents the correctness and confidence in an image's classification. The geometric margin is functional margin scaled by $||w||$ where w is the vector orthogonal to the decision boundary.

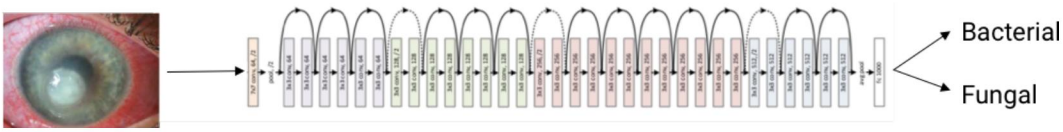

Figure 1: Eye image is passed into the ResNet-34 neural network, which features special skip connections and utilizes batch normalization [6]. Two fully connected layers were added at the end of the model with dropout to combat overfitting. The model provides one output for each class, with the larger output determining classification into bacterial or fungal keratitis.

3.2 Data Preprocessing

Image preprocessing was required since flattened pixel vectors do not preserve information about the spatial relationship between image features. Thus, a pretrained convolutional neural network (ResNet) was used in order to obtain a more robust feature vector.

Data augmentations were also used to extend the impact of the dataset. This included random horizontal flips and random rotations from -45 degrees to +45 degrees. Both of these approaches preserved the orientation of the eye and were optimized as hyperparameters for the model. The dimensions of 448 x 448 pixels were used instead of ResNet's standard 224 x 224 pixels to provide more details of the eye for the model. Additionally, normalization of the images was experimented with to standardize the input using the mean and standard deviation of the images. This was not ultimately used due to variance resulting from the small dataset.

3.3 Proposed Method

The ResNet set of models were chosen based on their high-performance within the convolutional neural network architectures [6]. After testing the performance of each ResNet model against the success metrics below, the ResNet-34 convolutional neural network was selected since it proved most resistant to overfitting while accurately making predictions on the small dataset (see Figure 1) [7].

After selecting ResNet-34, the following hyperparameters were tuned with the help of an automated script, and then performance was evaluated using k-fold cross validation: optimizer, learning rate, scheduler, batch size, dropout probability, L2 norm regularization, and epochs.

In general, reducing overfitting was the greatest challenge in this project since the dataset contained only 438 images. This was approached in two main ways: model design and optimization techniques. A dropout probability of 0.5 was added to the final fully connected layer of the ResNet-34 model, which helped substantially. To further combat overfitting, L2 norm regularization was experimented with to add a penalty term to the model weights during backpropagation, though this term was ultimately not included after hyperparameter tuning. In addition, the batch size was tuned to 32 and the model was optimized using stochastic gradient descent (SGD), which outperformed other optimizers including Adam, AdamW, and RMSprop.

Another challenge was guiding the model to find a better minimum in the loss function. Using a one cycle learning rate scheduler, the model was ramped up from a learning rate near 0 up to a maximum learning learning rate of 0.0025, and then slowly stepped down over time. The goal was to allow stochastic gradient descent to tune initially before ramping up the step size to give the model flexibility to find a better local minimum, and then to decrease the learning rate to allow gradient descent to settle at the minimum [13].

4 Results

Accuracy was the primary metric that was evaluated since the dataset had close to an even data split between classes. Accuracy was evaluated with two approaches: k-fold cross-validation and bootstrap sampling. For the deep learning model, k-fold cross-validation ($k = 5$) yielded a 73.7% accuracy, and bootstrap sampling of 20 trials yielded a 95% confidence interval of [66.6%, 79.8%] for validation accuracy. This represents a significant improvements over the baseline logistic regression confidence interval of [58.1%, 60.8%] and SVM confidence interval of [50.4%, 55.5%].

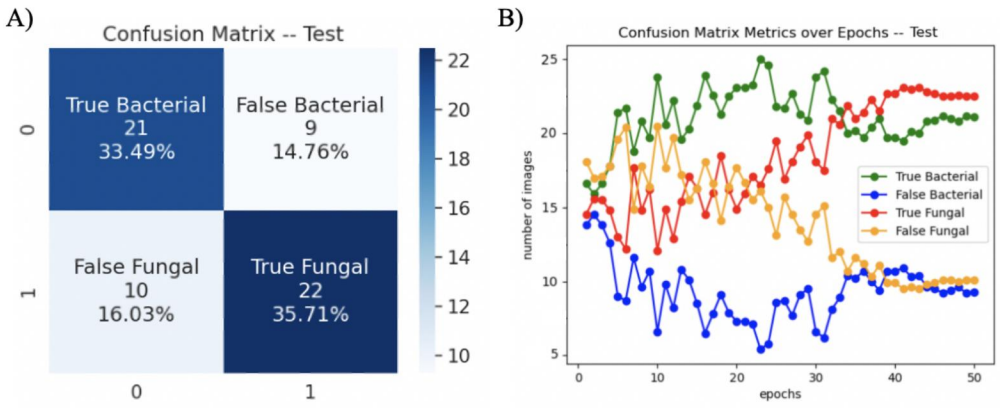

Figure 2: **(A)** Confusion matrix suggests model classifies fungal images more accurately than bacterial. **(B)** Confusion matrix metric trends over 50 epochs for the validation set. Note: 0 represents bacterial classification, 1 represents a fungal classification; x-axis is predictions, y-axis is labels.

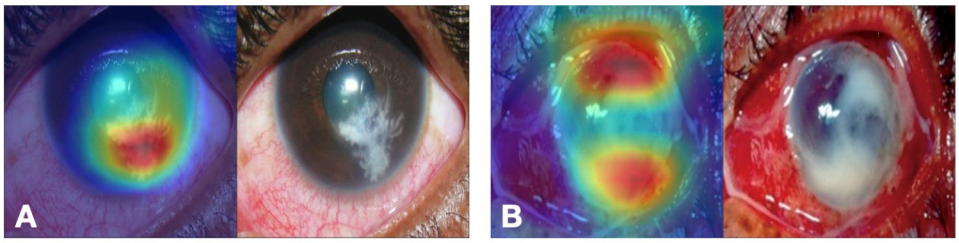

Figure 3: Grad-CAM heat map of the indicative features generally illustrates that the model focuses on medically relevant features, such as ulcers. This is true for both types of infections: **(A)** fungal infection (original image from [10]), and **(B)** bacterial infection (original image from [11]).

Using the deep learning model, a confusion matrix was constructed for the validation set (see Figure 2). For the validation set, there was a 71.0% fungal recall and 67.7% bacterial recall. The confusion matrix metrics over the 50 epochs shows that the number of true bacterial & true fungal images increases and false bacterial & false fungal images decreases as training progresses (see Figure 2).

Additionally, Grad-CAM was used to visualize heatmaps for the model, which show that the model typically focuses on relevant medical features that can be indicative of keratitis type, such as ulcers[2]. Example visualizations can be seen in Figure 3.

5 Discussion

5.1 Interpretation of results

These results demonstrate a statistically significant improvement in prediction accuracy of bacterial vs fungal keratitis compared to clinical diagnoses, as the lower bound of the 95% confidence interval was above the 65% accuracy from gold-standard clinical diagnoses [3]. In addition, the k-fold cross-validation accuracy of 73.7% is greater than the 70% accuracy shown in prior literature. However, since the 95% confidence interval contains 70%, improved performance over prior literature cannot be concluded. Nonetheless, the lowest single-tail p-value over the 20 bootstrap trials was $p = 0.106$ (computed using the bootstrapped distribution), demonstrating notable progress towards this goal (see Section 6.1 on future work).

Furthermore, the results demonstrate higher fungal recall than bacterial recall, which is significant because fungal keratitis is the more dangerous of the two types, so misdiagnosing a fungal infection is undesirable [4]. The results also demonstrate higher bacterial recall than seen in Xu et al., though

[2]Grad-CAM images were selected to illustrate a phenomenon and may not be representive of all the data.

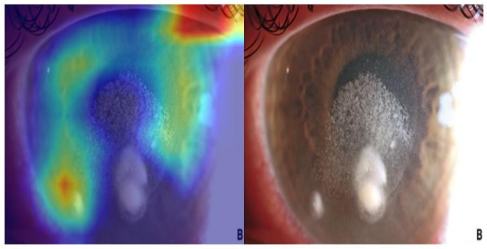

Figure 4: Grad-CAM heat map illustrates that the model sometimes focuses on spurious features such as eyelashes (original image from [12]).

the prior work had more types of keratitis to classify which affects this metric [5]. Additionally, the results from this study are from a diverse dataset, providing evidence that machine learning models can generalize on images from different hospitals and cameras for this classification task.

5.2 Error Analysis

The small number of images in the dataset is the most likely cause of performance variability (see Section 6.1 for more details). The model is susceptible to small changes in hyperparameters, suggesting that the global minimum is difficult to find. With a small dataset, it is possible that the model overfits to spurious features that are not relevant to infectious keratitis diagnosis.

As seen in the Grad-CAM images in Figure 3, the model correctly focuses on features such as corneal ulcers in fungal and bacterial photos. However, as seen in Figure 4, for some images the model focuses on spurious features such as eyelashes, which is behavior also noted in prior research [4]. This highlights that there is room for improvement in the model, so the current results might not represent the full potential of a deep learning approach to classifying bacterial and fungal keratitis.

6 Conclusion

This work presents a machine learning model that demonstrates a statistically significant improvement in prediction accuracy of bacterial vs fungal keratitis compared to gold-standard clinical diagnoses. It also showed a k-fold cross-validation accuracy of 73.7%, which is greater than the 70% cross-validation accuracy shown in prior literature. While the confidence interval of [66.6%, 79.8%] has not yet demonstrated statistically significant improvement over prior literature, there exists potential to further improve the model.

6.1 Future Work

Future work involves gathering additional images for the dataset to help with overfitting, and then revisiting model and optimization choices that were made specifically for a small dataset [9]. The hypothesis that performance variability is directly related to the size of the dataset will be tested. In particular, an accuracy vs number of images graph will be plotted and analyzed. Additionally, image pre-processing techniques can improve the features the model focuses on [5]. Initial experimentation with bounding box techniques to remove spurious features such as eyelashes has been started, and can be further developed to improve model accuracy.

Exceeding 70% accuracy in the lower bound of the confidence interval would lead to publishable results in medical and / or technical journals. This model could then be explored as part of a mobile app that can be used to provide a predicted diagnoses to assist patients who may not have immediate access to medical care.

7 Acknowledgements

Much thanks to Professor Moses Charikar, Professor Chris Ré, and the rest of the CS 229 staff for their support throughout the quarter. We are also immensely grateful to Mo Tiwari for his mentorship throughout this quarter as well as the other members of the Thrun lab.

Contributions

Kevin

Gathered images from medical case reports, PubMed, and reverse image search engines; created data loader for PyTorch and NumPy; created logistic regression model; programmed software architecture for NN with transfer learning from Resnet; used Grad-CAM to visualize gradients; programmed script for automated hyperparameter tuning; programmed script for image transformation visualization; experimented with automated bounding boxes; investigated L2 norm regularization; experimented with scheduled learning rate decay; investigated momentum; notably improved bacterial keratitis performance.

Kimmy

Gathered images from medical case reports and Google Image Search; deduplicated images across dataset; ran statistical analysis for the initial models (logistic regression, SVM) to create a 95% confidence interval to use as a baseline; implemented and investigated a range of data augmentation options for the models; investigated data preprocessing and cleaning to improve performance; programmed script for generating confusion matrix plot; programmed bootstrap sampling and confidence intervals for deep learning model; experimented with hyperparameters.

Ryan

Contacted and met with authors of ophthalmology journal articles, adding 113 images to the dataset from Stanford Health Care; gathered images from medical journals and image search engines; created SVM model; experimented with pre-processing of images via normalization; tuned hyperparameters, including the selection and programming of the ResNet-34 model, optimizer, one cycle learning rate scheduler, max learning rate, batch size, pixel resolution, and dropout rate; optimized data augmentation to maximize model performance; programmed the plotting of a loss curve relative to the learning rate to tune the max learning rate.

References

[1] Ung L, Bispo PJM, Shanbhag SS, et al. "The persistent dilemma of microbial keratitis: global burden, diagnosis, and antimicrobial resistance." Surv Ophthalmol, 2019, 64(3):255–271, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7021355/.

[2] "Keratitis." Mayo Clinic, https://www.mayoclinic.org/diseases-conditions/keratitis/symptoms-causes/syc-20374110.

[3] Austin, Ariana, Tom Lietman, and Jennifer Rose-Nussbaumer. "Update on the management of infectious keratitis." Ophthalmology 124.11 (2017): 1678-1689.

[4] Kuo MT, Hsu BWY, Yin YK, et al. "A deep learning approach in diagnosing fungal keratitis based on corneal photographs." Scientific Reports, 2020, 10: 14424, https://www.nature.com/articles/s41598-020-71425-9.

[5] Xu Y, Kong M, Xie w, et. al."Deep Sequential Feature Learning in Clinical Image Classification of Infectious Keratitis." Engineering, 2020, https://www.sciencedirect.com/science/article/pii/S2095809920301776.

[6] CS231n Convolutional Neural Networks for Visual Recognition, https://cs231n.github.io/convolutional-networks/.

[7] Fast.ai / PyTorch: Transfer Learning using Resnet34 on a self-made small dataset (262 images), https://medium.com/analytics-vidhya/fast-ai-pytorch-transfer-learning-using-resnet34-on-a-self-made-small-dataset-262-images-17003c9af3ce.

[8] Finding Good Learning Rate and The One Cycle Policy, https://towardsdatascience.com/finding-good-learning-rate-and-the-one-cycle-policy-7159fe1db5d6.

[9] Handling overfitting in deep learning models, https://towardsdatascience.com/handling-overfitting-in-deep-learning-models-c760ee047c6e.

[10] Prajna, Venkatesh N et al. "Fungal keratitis: The Aravind experience." Indian journal of ophthalmology vol. 65,10 (2017): 912-919. doi:10.4103/ijo.IJO_821_17.

[11] Srinivasan, M et al. "Distinguishing infective versus noninfective keratitis." Indian journal of ophthalmology vol. 56,3 (2008): 203-7. doi:10.4103/0301-4738.40358.

[12] Huang, Sonia MBBS; Sun, Michelle T. MBBS, PhD; Gupta, Aanchal FRANZCO Unilateral Streptococcus pneumoniae microbial keratitis after small-incision lenticule extraction, JCRS Online Case Reports: April 2020 - Volume 8 - Issue 2 - p e00013 doi: 10.1097/j.jcro.0000000000000013.

[13] Finding Good Learning Rate and The One Cycle Policy, https://towardsdatascience.com/finding-good-learning-rate-and-the-one-cycle-policy-7159fe1db5d6.