International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

A Framework for Building a Balanced Corneal Image Dataset from Public Repositories for AI-Based Diagnosis

Amanda Grace Ndebele[1], Munyaradzi Charles Rushambwa[2], Rajkumar Palaniappan[3]

[1]Master of Technology Degree in Industrial Automation Student, School of Engineering, Electronic Engineering Department, Harare Institute of Technology
[2]Lecturer, School of Engineering, Electronic Engineering Department, Harare Institute of Technology
[3]Associate Professor, College of Engineering, Department of Mechatronics Engineering, University of Technology Bahrain

Abstract
Deep learning models' performance in medical imaging highly depends on the availability of well-balanced, high-quality annotated datasets, [1]. Nevertheless, several ophthalmic diseases, such as keratitis, do not have the public databases readily available. This paper provides a systematic approach to creating well-balanced corneal image datasets from scattered public databases such as Kaggle and GitHub. The proposed system is capable of tackling problems such as image heterogeneity, class imbalance, and format discrepancies. The manual provides step-by-step guidance on image selection, quality control, class redistribution, data pre-processing, and data augmentation. A pilot dataset of 400 images from the four classes, Bacterial Keratitis (BK), Fungal Keratitis (FK), Herpes Simplex Keratitis (HSK), and Normal Eyes, each class with 100 images, was generated using this framework, and experiments were conducted in MATLAB. The results show that the dataset can be used to train baseline convolutional neural networks (CNNs) with reproducible accuracy. This framework is a cost-effective, scalable approach to enable an AI-based diagnostics system for underserved medical domains.

Keywords: class balancing, corneal image datasets, data augmentation, dataset construction, deep learning, keratitis, low-resource AI, medical image curation

1. Introduction

The use of deep learning in medical diagnosis has experienced a surge in recent years, due in part to its success at learning intricate patterns from large collections of visual data, [2]. In ophthalmology, CNNs have had great success in diagnosing diabetic retinopathy, glaucoma, cataract, etc [3]. Nevertheless, such gains are limited to diseases with richly annotated image datasets. By comparison, disorders such as keratitis act as one of the major causes of blindness, yet remain underrepresented.

Keratitis is a condition in which the cornea becomes inflamed due to trauma, infection, improper use of contact lenses, or, infrequently, underlying diseases. Early diagnosis is vital as delayed intervention may result in irreversible blindness. Regrettably, clinical diagnosis in most cases is based on the slit-lamp

examination by ophthalmologist practitioners or lab-testing facilities, which in turn may not be available in a low-resource environment. Artificial intelligence (AI) may provide an answer to the latter problem.
Dataset construction is a major difficulty as there is limited data available on corneal images publicly, and the existing data may be informal, non-labeled, or inconsistent. Also, the number of normal cornea images is generally fewer than infected cases, which causes the imbalanced dataset issue that affects the performance and the generalization of the model.
This manuscript introduces a simple and practical method for compiling balanced sets of corneal images for AI diagnostic systems. The framework is motivated by the author's experience in curating images from various sources such as kaggle.com, github.com, academic image sets, etc. A description for a step-by-step dataset cleaning, class balancing, resizing, and augmentation, and recommendations for usability assessment of data in small-sized CNN experiments are provided.
By allowing reproducible dataset generation in low-resource setups, this work is a step towards democratizing AI development in the field of ophthalmology. This manuscript also fills an important gap in the literature by providing an open methodology to prepare data in low-data domains, which is a crucial basis for robust, interpretable, and deployable deep-learning systems.

2. Motivation and Background

Deep learning has demonstrated revolutionary promises in many medical image analysis problems, from tumor detection to skin lesion classification, [4]. However, the performance of such models relies on the quantity and the quality of the labeled training data. One area in which this reliance can be particularly observed is ophthalmology, where diseases such as diabetic retinopathy and age-related macular degeneration have been aided by large datasets such as EyePACS and Messidor. In contrast, infectious corneal diseases such as keratitis have fewer publicly available datasets.
Keratitis is a preventable and treatable cause of corneal blindness, yet it is one of the leading causes of vision-related disability due to blindness globally, particularly in low- and middle-income settings, [5]. An AI-driven automated model to diagnose keratitis is an essential tool in the decision-making process to cut down diagnostic delays and misclassifications. Yet for AI to be reliable in this setting, it must be trained on balanced, representative datasets that reflect differences among infection types and patient demographics. The major setback, however, is that there are too few publicly available large-scale datasets for infectious corneal diseases such as keratitis, which remains undeserved. [6]. The task of keratitis focuses on ocular imaging, and the existing public datasets either do not contain the keratitis case study or their subset of ocular images is small and has many labels or annotations. Moreover, these datasets may contain images in various resolutions, lighting conditions, orientations, and file formats, so that their use becomes challenging without much pre-processing. The class imbalance, particularly the under-represented normal corneal images, further compounds the problem.
These constraints are both a challenge and an opportunity. The hard part is assembling usable, clean datasets out of scattershot sources, which is a time-consuming and picky process that is seldom described in AI research. The best part is the ability to develop and share a general and reproducible framework for constructing datasets that can be utilized by scholarly researchers. The structure not only aids model training but also promotes transparency and replicability in AI. Such a constraint motivated the present paper to address this critical gap and suggest a methodology for obtaining, cleaning, balancing, and organizing corneal images, particularly for keratitis diagnosis. The resulting framework is intended to be platform-independent and executable with the common tools, such as MATLAB, Python, and even

worksheet-assistant file management, making a low entry barrier for researchers from different technical fields.

3. Dataset Acquisition Strategy

The author started with a comprehensive search across a variety of open-access resources, including Kaggle, GitHub, academic publications, and Ophthalmology research repositories. The primary purpose was to obtain a rich dataset of corneal images in infectious and non-infectious eye disease, and in particular to image specific keratitis subtypes: BK, FK, HSK, and normal corneal images. The identified public datasets were reviewed with the following criteria:

- Applicability for anterior eye or corneal imaging
- Labels or class information known or available
- Image resolution and clarity
- Permissions for Reuse of Academic Content

Nevertheless, the main challenge was that raw data from some of the sources was often disorganized, intermixed with non-corneal eye images (e.g., fundus maps, retinal angiographs), or stored in different file types (JPG, PNG, BMP). Filters were used to guarantee the integrity of the dataset and the clarity of the classes. Only high-resolution corneal images, i.e., images with frontal viewing and little occlusion, were included. All images had to be in JPG file format or transcoded with lossless compression from another file format. Images were automatically organized into one of the four classes based on metadata and/or filename tags. Duplicates, blurs, redundant and irrelevant scans were removed by hand. The images were then categorized into four distinct folders (BK, FK, HSK, Normal) that were to be independently reviewed. This manual procedure guaranteed quality for each class and less label noise, which is a consistent problem in AI-based medical image research.

Following cleaning, the curated dataset was balanced to possess precisely:

- 100 images per class
- 400 images in total

The dataset was intentionally kept at an academic scale to emulate settings in which training data is scarce or costly to annotate. The class distribution was uniformly distributed to avoid bias during model training, and this made the dataset very suitable for CNN experiments. The final dataset was saved in a folder structure so one can easily load it within a workspace, in this case, MATLAB, as follows:

Dataset/
├── BK/
├── FK/
├── HSK/
└── Normal/

4. Image Curation and Class Balancing

An image viewer and regular preview tools available in MATLAB were used to manually examine each dataset image. The aim was to eliminate images with poor resolution or pixelation, blurry or out-of-focus footage, overexposure or underexposure, anatomical structures such as eyelids and lashes, and non-frontal corneal perspectives. Such strict filtering guaranteed that only high-quality, optimally interpretable images in the clinical daily routine were kept for learning. About 30% of the raw images were rejected at this stage. To avoid overfitting the model and to avoid biased learning, duplicate images were removed.

Duplicates were identified by matching filenames, timestamps, and pixel-level similarity. Then, only one from each redundant image was retained, where possible, the one with the best focus and the highest contrast.

To address class imbalance, [2], a challenge faced by many scholars, the dataset was balanced by manually removing samples to obtain a 1:1:1:1 class distribution across the four classes: BK, FK, HSK, and Normal. Teacher folders were verified to ensure that each of the 100 unique images was included, essential for reducing bias within the network by constant class imbalance during training, learn with the same label equality opportunities, and to enable fair metric computation, precision, recall, and F1-score. A MATLAB validation script was implemented to programmatically count the images per folder and mark exceptions.

```
% Define base folder and subfolders
baseDir = 'SplitMiniDataset';
subsets = {'Train', 'Validation', 'Test'};
classes = {'BK', 'FK', 'HSK', 'Normal'};

% Create folder structure
fprintf('Creating folder structure (if not present)...\n');
for i = 1:length(subsets)
for j = 1:length(classes)
folderPath = fullfile(baseDir, subsets{i}, classes{j});
if ~exist(folderPath, 'dir')
mkdir(folderPath);
end
end
end

% Count and display image files
fprintf('\Image Count per Class in Each Subset:\n');
for i = 1:length(subsets)
fprintf('\n%s:\n', subsets{i});
for j = 1:length(classes)
folderPath = fullfile(baseDir, subsets{i}, classes{j});
imageFiles = dir(fullfile(folderPath, '*.jpg'));
imageFiles = [imageFiles; dir(fullfile(folderPath, '*.jpeg'))];
imageFiles = [imageFiles; dir(fullfile(folderPath, '*.png'))];
imageFiles = [imageFiles; dir(fullfile(folderPath, '*.bmp'))];
imageFiles = [imageFiles; dir(fullfile(folderPath, '*.tif'))];
imageFiles = [imageFiles; dir(fullfile(folderPath, '*.tiff'))];

fprintf(' %s: %d images\n', classes{j}, length(imageFiles));
end
end
```

All classes were saved to separate subfolders with the following directory structure:

SplitMiniDataset/
├── Train/
│ ├── BK/
│ ├── FK/
│ ├── HSK/
│ └── Normal/
├── Validation/
│ └── (same structure)
└── Test/
└── (same structure)

This was done to ensure the data was readily usable with MATLAB's imageDatastore as well as other frameworks such as TensorFlow/ Keras that pull labels from the file system.

5. Preprocessing and Augmentation Pipeline

After the dataset was cleaned and balanced, the images were preprocessed and augmented for model training. Preprocessing was made with the objectives of making image dimensions consistent, promoting the generalization of the model, and simulating various corneal imaging scenarios.

Image Resizing and Normalization

All the images were resized to 224 × 224 pixels, which was related to the input of the customized CNN. This distance helps balance between maintaining features and being computationally low, and not requiring a GPU. Images were also normalized and scaled pixel values to [0, 1]. This step normalizes the contrast in magnitudes and helps to speed up the convergence during the training.

Augmentation Techniques

To overcome the small dataset size and to introduce variation, the following data augmentation techniques were applied: using the imageDataAugmenter in MATLAB:

- Random Horizontal Flipping, which introduces variation in the left-right corneal symmetry
- Random Rotation (±15 degrees), which emulates eye-orientation changes during acquisition of an image
- Random brightness adjustment, which solves the problem of non-uniform illumination conditions.
- Zooming and Scaling, which emulates changes in zoom settings on the slit-lamp

These transformations were implemented on the training portion of the datasets and not on the validation and test sets, ensuring fairness while evaluating.

Augmentation Implementation in MATLAB

The following code snippet was applied

```
augmenter = imageDataAugmenter( ...
'RandRotation', [-15, 15], ...
'RandXReflection', true, ...
'RandXScale', [0.9, 1.1], ...
'RandYScale', [0.9, 1.1], ...
'RandBrightness', [0.8, 1.2]);

augimdsTrain = augmentedImageDatastore([224 224], imdsTrain, ...
'DataAugmentation',
```

augmenter);
This augmented image datastore was fed into the training loop and allowed the CNN to learn from a more diverse image distribution.

6. Dataset Evaluation and Usability Testing

To demonstrate the contribution of the curated and augmented dataset, a training experiment was conducted with a lightweight convoluted neural network (CNN). The challenge was not to surpass state-of-the-art accuracy on some specific tasks, but rather to assess the usability of the dataset and obtain a good balance in the readiness of the dataset for training AI models in low-resource setups. A custom CNN architecture was constructed with MATLAB R2023a with the Deep Learning Toolbox. The model included four convolutional layers followed by batch normalization, ReLU activation, and max pooling. A global average pooling layer and two fully-connected layers completed the model for four-class classification. The model was trained on a CPU-only machine with the following configuration:

- Epochs: 10
- Mini-Batch Size: 16
- Optimizer: Stochastic Gradient Descent with Momentum (SGDM)
- Initial Learning Rate: 0.001

The dataset was divided into a training set and a validation set in the ratio of 70-15-15 over four balanced classes. The performance of the trained model was assessed on the test set based on accuracy, precision, recall, F1-Score, and a Confusion Matrix. These were chosen in such a way that they take into account the overall performance of the models and the balance for each class.

Observed Results

The model obtained 60% accuracy on the test set with some variation between classes. Herpes Simplex Keratitis (HSK) was identified most consistently, in contrast to the well-known difficulties establishing Bacterial and Fungal Keratitis misclassification, even in clinical practice.
The confusion matrix revealed a relatively even misclassification distribution, which constitutes evidence of the structural balance and diversity of the dataset. These results validate the utility of the dataset in prototyping CNN models and performing small-scale diagnostic studies. The dataset accelerated training with simple hardware, promoted equality in all classes, and proved diagnostic difficulties, consistent with a complicated reality. Its balance between structure and reproducibility makes it ideal for further benchmarking practices and open-source AI courses.

7. Contribution and Impact

This work makes a concrete contribution to AI-powered medical diagnostics by focusing on the frequently neglected issue of dataset accessibility and preparation, in particular in an underrepresented domain such as keratitis, [6]. While most literature concerns itself with novel model architectures or performance gains, this paper fills a critical gap by providing a practical and reproducible template for dataset creation, a fundamental but seldom documented part of deep learning research.

Technical Contributions

- Open Framework for Dataset Generation: A well-defined and reusable pipeline was established to collect, curate, balance, and set up the corneal images from open-access platforms. This approach can be repeated or modified for other ocular or biomedical imaging applications.

- Designed dataset for academic use: The dataset with 400 labeled and balanced corneal images would be beneficial for students' projects, research experiments, and model prototyping, especially in hardware-constrained conditions.
- Dataset Compatibility Demonstration: The paper is also able to demonstrate dataset usability via small-scale CNN training and testing that high learning accuracy is still possible with the provided curated public dataset and basic machine resources.

Educational and Research Impact

- Pedagogical Value: The step-by-step documentation, from image retrieval to CNN testing, renders this study a useful tutorial for students or early-career researchers who need to learn how to create and maintain datasets for their projects.
- Dataset Repositories Ground: This work could motivate the development of a common open-source repository of corneal images, similar to what EyePACS and ISIC were built for retinopathy and skin lesions, [3].

Societal Relevance

- Increasing AI Access in Low-Resource Areas: By demonstrating that a usable dataset can be developed and tested with basic hardware or without obtaining software licenses specialized for the task, this work is providing the means to empower researchers and healthcare innovators in developing countries.
- Enabling Early Detection of Corneal Disease: Longer term, frameworks like this can be used to develop clinical tools to enable rapid triage and decision support in diagnosing infectious keratitis and other anterior eye diseases.

8. Conclusion

This paper proposes a pragmatic and repeatable methodology to create well-balanced corneal image datasets from various public repositories, aiming to develop AI-based diagnostic systems for keratitis. In these times where data availability and model reproducibility are as pivotal as algorithmic innovation, given that the lack of annotated, balanced, and standardized data for underrepresented conditions is a major bottleneck for medical AI research, this is a contribution focusing on such a case. Defining a full pipeline from sourcing images to augmenting them and testing their usability, the framework allows researchers to create diagnostic datasets without needing access to high-end computational power or proprietary tools. Via manual curation and class balancing, and basic pre-processing with MATLAB, a 400-image dataset is constructed, which facilitates the training of CNN and provides a relatively fair evaluation towards four classes of corneal conditions.

The evaluation results indicated that, despite being small, the dataset is balanced and diverse to enable academic research and proof-of-concept model development. Additionally, the meticulous recording of all the steps promotes transparency and can facilitate replication in both lab and classroom settings, as well as in settings where resources are limited.

This research underscores that ethical AI starts with not just the algorithm, but the integrity of data. In democratizing dataset creation for vision-based AI systems in ophthalmology, this work helps advance our pursuit of equitable, explainable, and scalable medical diagnosis.

References:

1. K. He, “Deep Residual Learning for Image Recognition”.
2. D. S. Kermany, M. Goldbaum, W. Cai, and M. A. Lewis, “Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning Resource Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning,” *Cell*, vol. 172, no. 5, pp. 1122-1131.e9, 2018, doi: 10.1016/j.cell.2018.02.010.
3. V. Gulshan *et al.*, “Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs,” vol. 94043, 2016, doi: 10.1001/jama.2016.17216.
4. P. Kauser Ahmed and M. D. Meraj Alam, “Brain tumor detection in medical imaging using soft computing techniques,” *Res. J. Pharm. Biol. Chem. Sci.*, vol. 7, no. 5, pp. 1170–1174, 2016.
5. R. Bourne *et al.*, “Trends in prevalence of blindness and distance and near vision impairment over 30 years: an analysis for the Global Burden of Disease Study,” *Lancet Glob. Heal.*, vol. 9, no. 2, pp. e130–e143, Feb. 2021, doi: 10.1016/S2214-109X(20)30425-3.
6. J. Lv *et al.*, “Deep learning-based automated diagnosis of fungal keratitis with in vivo confocal microscopy images,” vol. 8, no. Ivcm, pp. 1–9, 2020, doi: 10.21037/atm.2020.03.134.