İSTANBUL UNIVERSITY PRESS

Acta Infologica
acin 2023, 7 (2), 281–292
DOI: 10.26650/acin.1173465

Research Article / Araştırma Makalesi

A Deep Learning-Based Classification Study for Diagnosing Corneal Ulcers on Ocular Staining Images

Oküler Boyama Görüntülerinde Kornea Ülserinin Teşhisi İçin Derin Öğrenmeye Dayalı Bir Sınıflandırma Çalışması

Onur Sevli[1]

[1](Assoc. Prof. Dr.) Burdur Mehmet Akif Ersoy University, Computer Engineering Department, Burdur, Turkiye

Corresponding author : Onur SEVLİ
E-mail : onursevli@mehmetakif.edu.tr

ABSTRACT

Corneal ulcer is a common disease worldwide and is one of the leading causes of corneal blindness. Diagnosis of the disease requires expertise, and the number of experienced ophthalmologists is not sufficient, especially in underdeveloped countries. For this reason, it is necessary to develop technology-based decision support systems in the diagnosis of the disease. However, the number of studies on this subject is not sufficient. In this study, CNN-based classifications were performed using corneal ulcer images obtained by an ocular staining technique, consisting of 712 samples and three classes. In addition to the AlexNet and VGG16 state-of-the-art architectures, which are widely used in the literature, a CNN model proposed for this study was used for classification. In the classifications performed by applying data augmentation, 95.34% accuracy with AlexNet, 98.14% with VGG16, and 100% accuracy with the proposed model was obtained. The findings were compared with similar studies in the literature. It was concluded that the accuracy rates obtained with all of the models used in the study were generally higher than similar studies in the literature, and the accuracy obtained with the proposed CNN model was higher than all of the peers. In addition, the success of the proposed model compared to other models with more complex structures revealed that it is not always necessary to use complex architectures for high accuracy.

Keywords: Corneal ulcer diagnosis, convolutional neural network, classification

ÖZ

Kornea ülseri dünya genelinde yaygın görülen bir hastalık olup kornea körlüğünün önce gelen nedenlerindendir. Hastalığın teşhisi uzmanlık gerektirmekte olup, özellikle az gelişmiş ülkelerde tecrübeli oftalmolog sayısı yeterli sayıda değildir. Bu durum hastalığın teşhisinde etkin ve uzmanlara destek sistemlerin oluşturulmasını gerekli kılmaktadır. Ancak henüz bu konuda yapılmış olan çalışmaların sayısı yeterli düzeyde değildir. Bu çalışmada 712 adet ve 3 türden oluşan, oküler boyama tekniği ile elde edilen kornea ülser görüntüsü kullanılarak CNN tabanlı sınıflandırmalar gerçekleştirilmiştir. Literatürde yaygın kullanılan AlexNet ve VGG16 daha derin state-of-art mimarileri yanında bu çalışma için önerilen bir CNN modeli kullanılmıştır. Veri arttırımı uygulanarak gerçekleştirilen sınıflandırmalarda AlexNet ile 95.34%, VGG16 ile 98.14%, ve önerilen model ile 100% doğruluk elde edilmiştir. Elde edilen bulgular literatürdeki benzer çalışmalarda karşılaştırılmıştır. Tüm modeller ile elde edilen doğruluk oranlarının literatürdeki çalışmaların genelinden yüksek olduğu, önerilen CNN modeli ile elde edilen doğruluğun ise emsallerin tamamından yüksek olduğu sonucuna ulaşılmıştır. Ayrıca önerilen modelin daha karmaşık yapıdaki diğer modellere nazaran da yüksek başarı sergilemiş olması, daha minimal mimarilerle de yüksek başarı elde edilebileceğini ortaya koymuştur.

Anahtar Kelimeler: Kornea ülseri teşhisi, evrişimsel sinir ağı, sınıflandırma

Submitted : 10.09.2022
Revision Requested : 06.10.2022
Last Revision Received : 04.04.2023
Accepted : 12.06.2023
Published Online : 14.08.2023

This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

1. INTRODUCTION

The cornea is the hard and transparent layer located in front of the iris, belonging to the dioptric system of the eye. The fibers of the cornea, which has a fibrous structure consisting of collagen, are located in the stroma layer, which forms a large part of its thickness. It is the first layer of the eye that refracts incoming light and accounts for about two-thirds of the total refractive power of the eye. In addition, thanks to its hard structure, it protects other parts of the eye (Maurice, 1957). A corneal ulcer is a condition that occurs as a result of deterioration of the epithelial layer or corneal stroma of the cornea due to inflammatory or infective causes (Chen & Yuan, 2010). Ocular surface diseases, damage caused by corneal surgery or contact lens use, adnexal diseases, and other traumas are among the risk factors for the formation of corneal ulcers (Amescua et al., 2012). Damage to corneal tissues due to viral, bacterial, or fungal sources causes corneal ulcers Studies reported that viral cases in the formation of corneal ulcers were more common in developed countries, and bacterial and fungal cases were more common in developing countries (Garg & Rao, 1999).

Corneal ulcer is a common eye problem worldwide and is the second leading cause of ocular morbidity (Song et al., 2014). Corneal ulcers can seriously damage eye health, causing cornea scars, perforation, endophthalmitis, and visual trouble. Corneal ulcer is among the leading causes of corneal blindness (Katara et al., 2013). Failure to diagnose the disease in a timely and correct manner and to apply the correct treatment on time may cause irreversible damage to the eye (Diamond et al., 1999; Cohen et al., 1987).

Corneal ulcer is one of the important problems threatening eye health, especially in developing countries, and the annual average corneal ulcer cases in these countries reaches 1.5 million (Basak et al., 2005). Diagnosis of a corneal ulcer is critical and is performed by experienced professionals. However, the number of experienced ophthalmologists around the world, especially in geographical regions with limited resources, is not sufficient and this makes the early diagnosis of the disease difficult. While early diagnosis increases the success of treatment, correct analysis of the morphological structure resulting from the disease is effective in determining the correct treatment procedures. An accurate distinction must be made between different ulcer stages and types to reduce the risk of permanent vision damage or blindness.

The ocular staining technique is used in the diagnosis of corneal ulcers as well as in the diagnosis of various eye diseases. In this technique, topical dyes are widely used to characterize ocular surface diseases and to quantify their severity (Bron et al., 2015). Quantitative analysis of corneal disorders is made more easily by examining colored eye surfaces under a slit lamp microscope. Although the manual diagnosis of a corneal ulcer is reliable, it requires high sensitivity, takes time, and the results obtained may vary in terms of the reviewers. In this case, the right treatment decision may not be made, or the treatment process could be delayed. Delayed or incorrect/incomplete treatment causes progression of the disease and the formation of irreversible defects. For this reason, it becomes necessary to develop intelligent support systems that will help experts make decisions effectively, quickly and with high accuracy.

With the development of technology, artificial intelligence techniques are widely used in the medical field as well as in many other fields. Machine learning, a sub-discipline of artificial intelligence, provides stable predictions about new situations by learning from existing data. Deep learning, which is a machine learning technique, can successfully reveal the complex hierarchy in the nature of data with the help of deep neural networks. The Convolutional Neural Network (CNN) is a deep learning method used especially in computer vision. In the literature, there are studies using different machine learning techniques and CNN for the diagnosis of corneal ulcers. However, the number of these studies using artificial intelligence for corneal ulcer diagnosis is still limited.

Noting the number of corneal ulcer cases in developing countries, Saini et al. (2003) collected a total of 106 corneal ulcer images from patients living in India for their study. The study achieved 90.7% accuracy in the classification study with artificial neural networks (ANNs) for corneal ulcer diagnosis, using the dataset consisting of the images collected. Akram and Debnath (2019), captured images of faces with a digital camera and then segmented the eye region on these images. A study was carried out to detect the presence of corneal ulcers using the fragmentary images. By using data augmentation on a total of 513 images, a binary classification was performed as a corneal or non-corneal ulcer with the proposed CNN model. The average accuracy value obtained for the two classes as a result of 40 epochs is 98.99%. Kim et al. (2019), proposed a CNN-based diagnostic model to determine the degree of corneal ulceration in dogs. They performed classifications with three different degrees normal, superficial, and deep on a total of 1,040 images collected at Korea Konkuk University Veterinary Medical Teaching Hospital. A 92% accuracy was achieved with the ResNet50 model with their classifications using different transfer learning models.

In the literature, the SUSTech-SYSU dataset is widely used in addition to the study-specific datasets on the detection of corneal ulcers. In this dataset, there are 712 eye images obtained using the ocular staining technique. These images belong to three different types of corneal ulcers: flaky corneal ulcers (FCU), point-like corneal ulcers (PCU), and

point-flaky mixed corneal ulcers (PFCU). Different applications were made for segmentation and classification with this dataset, which was used frequently in recent studies.

Segmentation is defined as determining the boundaries of the target region on the image. Corneal ulcer segmentation on ocular staining images is important for the quantitative assessment of ocular surface defects. To realize this critical and challenging task, Wang et al. (2021) performed a segmentation study based on the Adjacent Scale Fusion method. In this study, which was carried out on the SUSTech-SYSU dataset, the Dice Coefficient value of 80.73% was reached. Portela et al. (2021) performed a segmentation study for corneal ulcer detection with a dataset of ocular staining images specific to their study. Using U-NET and DexiNet architectures, they obtained an average of 70.50% Dice Coefficient in the study with a total of 449 FCU type disease images. The PFCU and FCU type disease images are more difficult to distinguish and this results in reduced diagnostic success. Wang et al. (2021) proposed a segmentation network to distinguish FCU and PFCU type images with higher success using the SUSTech-SYSU dataset. They reached a Dice Coefficient of 89.14% with this network called CU-SegNet, which was based on the encoder-decoder structure. Diagnosis of corneal ulcers becomes more difficult due to large differences in shape, blurred borders, and noise interference. Addressing this problem, Wang et al. (2021) performed a segmentation study on the SUSTech-SYSU dataset with a semi-supervised GAN using the Semi-MsST-GAN. A Dice Coefficient of 90.93% was reached with this model, which was then compared with different techniques.

In the literature, besides the segmentation studies on the SUSTech-SYSU dataset, there are classification studies performed with different techniques. Tang et al. (2020), performed a classification on this dataset using a modified VGG network. Eighty-eight . eighty-nine percent accuracy, 92.27% precision, and 71.93 recall values were the results obtained from the classification of images consisting of three different classes; the FCU, PCU, and PFCU. Gross et al. (2021)proposed a specific CNN model for the classification of the same dataset. The highest accuracy value reached with a proposed model was 92.73%. Teeyapan (2021) performed classifications using the SUSTech-SYSU dataset with the transfer learning method. In the study where different architectures were tested, the ResNet50 model provided the highest result with 95.10% accuracy.

Li et al. (2021), suggested a deep learning-based method for early and accurate diagnosis, noting that corneal ulceration is one of the major causes of corneal blindness worldwide. Classifications were performed with the DenseNet121, InceptionV3 and ResNet50 models on a dataset of 6,567 samples, consisting of corneal images with normal cornea, ulcerated cornea and other abnormalities. The highest success was obtained with the DenseNet121 as 96% Cohen's kappa coefficient.

Diagnosis of corneal ulcers can be challenging for specialists. Sajeev and Prem Senthil (2021) proposed a CNN-based method for classifying corneal ulcers of bacterial and viral origin. They classified the dataset consisting of a total of 446 corneal ulcer images belonging to these two classes, with different input sizes and CNN architectures with two or three convolution layers. The highest accuracy obtained was 81.2% with the model with 64x64 input size and three convolution layers.

Xu et al. (2021), stated that corneal ulcer is an emergency that needs to be treated quickly, so a study was completed with a classification study using the deep learning on images with ulcers. Classifications were done on 115,408 microscopic images collected from 10,609 patients, using the VGG16, GoogLeNet and DenseNet models, at the image-level and patch-level. With DenseNet, the most successful model, they achieved an accuracy of 61.04% at image-level and 66.30% at patch-level. The results of the study were compared with the diagnoses of ophthalmologists and it revealed that the method they proposed gave more successful results.

This study performed a CNN-based application for the successful diagnosis of corneal ulcers using ocular staining images. The SUSTech-SYSU dataset, which is widely used in literature, was the preferred method of work. Classification studies were performed on three different types of images using two known state-of-the-art architectures, as well as a less complex proposed CNN model for this study. The main motivation of this study is to achieve a higher success compared to similar studies in literature and to present a more effective solution. In addition, the sub-objective is to demonstrate that a less complex model proposed for this study outperforms the more complex models. The following sections contain the dataset information, the classification methods used, findings and discussion.

2. MATERIAL AND METHOD

2.1. Dataset

The dataset used in the study included eye images obtained by the ocular staining technique, created by Deng et al. (2020) for the detection of corneal ulcers. The images in the dataset were obtained from patients with various types and grades of corneal ulcers at the Zhongshan Ophthalmic Center of Sun Yat-sen University, China. No distinction

was made regarding external conditions such as age and gender of the patients whose eye images were taken. In this data set, which included a total of 712 images, there were data samples belonging to three different classes of corneal ulcers. These classes, in which the data samples belong, are flaky corneal ulcers (FCU), point-flaky mixed corneal ulcers (PFCU), and point-like corneal ulcers (PCU). There were 91, 263, and 358 images in each class, respectively. The graph showing the dataset class distributions is given in Fig. 1.

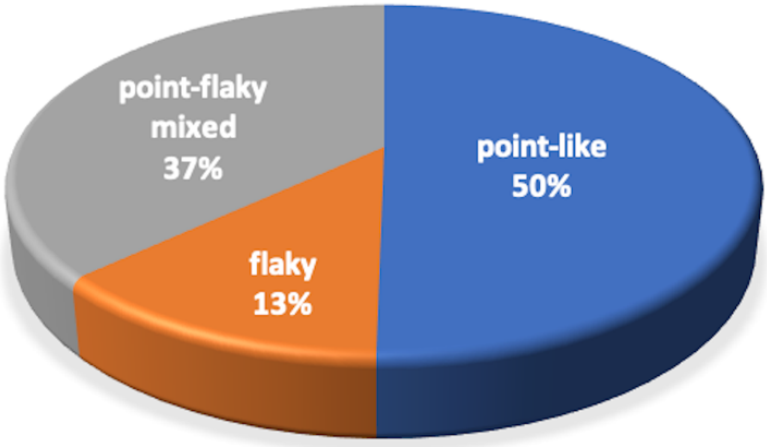

Figure 1. Dataset class distributions

The type of cases for the data set were divided as follows; PCU 50%, PFCU 37%, and FCU 17%. The dataset was not balanced in terms of the number of images in these classes. Image samples of each class in the dataset are given in Fig. 2.

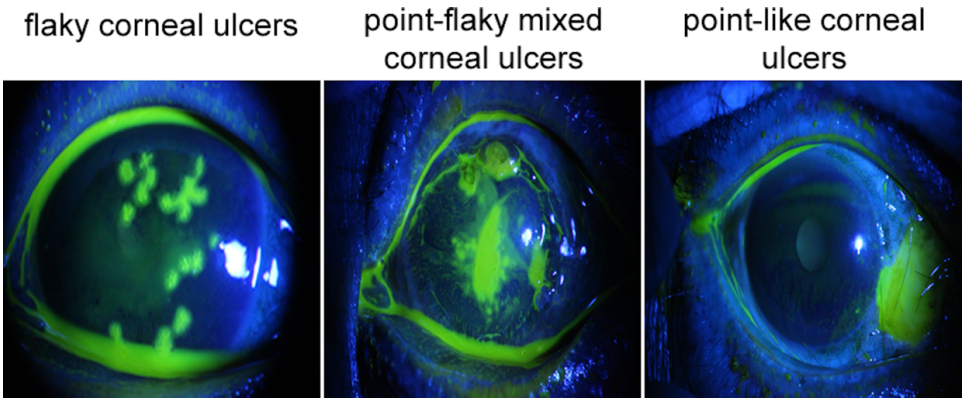

Figure 2. Sample images from the dataset

2.2. Models Used for Classification

Deep learning, which is increasingly used in many fields, is a method of machine learning performed with artificial neural networks consisting of many layers. Deep neural network models, which can contain different numbers of layers and processing units depending on the structure of the problem to be solved (Aksoy, 2021), have a wider learning capacity than classical machine learning techniques. Deep learning models provide high success in revealing complex hierarchical structures and making consistent classifications.

The Convolutional Neural Network (CNN) is a deep learning method that is widely used in solving classification and regression problems in image analysis and has gained popularity with its high success. Input, convolution, pooling, fully connected, and classification layers make up a typical CNN.

The CNN model's inputs are the pixels of the image to be processed. In the convolution layer, feature detectors, also known as filters, are stridden over the input pixels to reveal a subset of features. Convolution is the main operation of a CNN model that enables feature extraction from the image. Dimension reduction is achieved in the pooling stage

using filters applied to the input matrix. The reduction action is carried out by a filter window, also known as a pool, which takes the maximum, minimum, or average of the remaining pixel values in the pool. Activation functions in the weighted layers of the neural network increase nonlinearity.

In the training of machine learning models, the problem of overfitting a certain class may arise due to the structure of the dataset. One of the methods that can be applied to alleviate this problem is the dropout operation. In the dropout process, a certain percentage of neurons in a neural network are randomly disabled during training, increasing the adaptability of the network to different situations. The fully connected layer is involved in the transition to the classification stage. The model in the classification layer tries to predict which class the input sample belongs to.

The number of layers in a CNN model and the number of processing units in each layer are two configurable parameters that change depending on the situation. Training a neural network model with an appropriate amount and variety of data is one of the key steps to obtaining highly accurate results. State-of-the-art CNN models are frequently used in various studies as they successfully classify approximately 14 million images in an ImageNet dataset and provide highly accurate findings when applied to other fields. In this study, two architectures commonly referred to in the literature, AlexNet and VGG16 were used. It was created by improving the architecture.

AlexNet architecture is a CNN model developed by Alex Krishevsky et al. (2017) that provides high accuracy in classifying the ImageNet dataset. AlexNet architecture consists of 14 layers, eight of which are weighted. There are five convolution layers in the model, three of which are followed by a max-pooling layer. The model has a total of 65 thousand neurons and more than 60 million parameters. AlexNet was among the top five with only a 17% error rate in the ILSVRC-2012 image processing competition and outperformed its successor by 10.9

VGG16 architecture is a CNN model developed by Simonyan and Zisserman (2014). It was developed by enhancing the AlexNet model and using a significant number of 3x3 filters in place of filters with huge core sizes. VGG16 architecture has 13 convolutional layers and five max-pooling layers, and three dense layers in the classification part. It is called VGG16 because it has a total of 16 weighted layers. The VGG16 model, which contains over 138 million parameters, was among the top five models with the highest accuracy of 92.7% in the ImageNet dataset.

It may be necessary to increase the number of layers and components of a CNN model to increase classification accuracy. However, the perception that this increase will always increase the classification success of the model is not correct. The increase in the number of layers and components also increases the number of parameters that need to be calculated in training of the model. The more parameters, the longer the training period of the model. Ideally, the deep learning method is expected to create a model that provides the highest performance with the fewest parameters. In addition to two high complexity state-of-the-art architectures used in this study, a less complex CNN architecture was proposed. The block diagram showing the general structure of the proposed CNN model is given in Fig. 3.

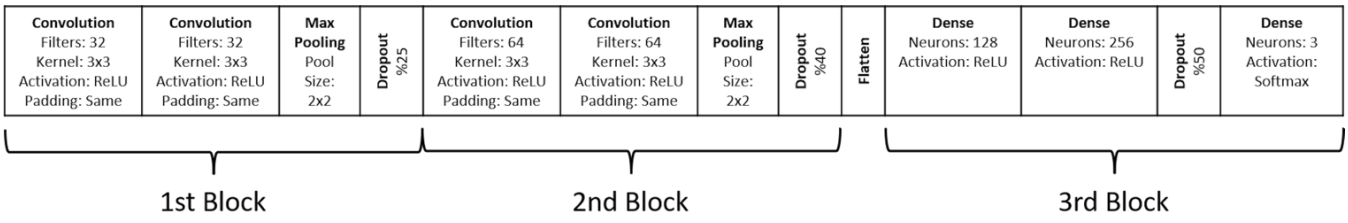

Figure 3. Block diagram of the proposed CNN model

The proposed model is considered in three main blocks. The first block contains two convolutions; a pooling, and a dropout layer. The first two layers, the convolution layers, each contain 32 filters of 3x3 size. In these layers, the ReLU activation function is used, and the same padding is applied. In the third layer, max-pooling is applied with a pool size of 2x2. Then 25% dropout is performed.

In the first two layers of the second block, there are convolution operations with 64 filters of 3x3 size. The ReLU is used as an activation function and the same padding is applied. These layers are followed by a max-pooling layer with a pool size of 2x2. Then a 40% dropout is performed.

After the first two blocks of the model, the extracted features are flattened and transferred to the classifier network, which is the third block. The classification process is performed with a neural network. In the first two layers of the classifier, there are dense layers containing 128 and 256 neurons using the ReLU activation function. Then a 50% dropout is performed. The output layer, a dense layer that contains as many neurons as the number of classes in the dataset, applies the Softmax activation function.

The classification layer used in the proposed CNN model was also used in the classification layer of the other two architectures used in this study. In the proposed model, there are 65,568 parameters excluding the classification layer. The number of parameters of the classification layer is 25,724,035. The total number of trainable parameters of the model, which includes a total of seven weighted layers, including the classification layer, is 25,789,603.

3. EXPERIMENTAL STUDY AND FINDINGS

The 712 corneal ulcer images, consisting of three types colored with the ocular staining technique, were classified with AlexNet, one of the state-of-the-art models widely used in the literature, and the VGG16 model, which was created by improving this architecture, as well as a less complex proposed CNN model used for this study. The state-of-art models used in the classification were fine-tuned. In the preprocessing stage before classification, image rescaling, the normalization of image pixels, and encoding of the labels were performed. Each image in the original dataset is colored and has a size of 2,592x1,728 pixels. Before classification, each image is resized to 224x224 pixels, which is ideal for the CNN architectures used. After rescaling, the pixel values forming the images were normalized with the min-max method. Each label in the dataset consisting of three different classes was numerically encoded. As a result of encoding, the FCU type was labeled as 0, PFCU type as 1, and PCU type as 2.

There were 91 FCU, 263 PFCU, and 358 PCU images in the dataset. In its original form, the dataset had an imbalanced class distribution. When working with imbalanced datasets, the classifier model tends to learn the dominant class and may be weak in learning minority classes. In order to overcome this problem data augmentation was applied using the original images in the dataset by providing the class balance of the dataset. Data augmentation was achieved by applying processes such as rotation, flipping, shifting, reflecting, and scaling on the original images at certain rates. The data augmentation in this study was done by applying 10% rotation, 10% zoom, and 10% shift horizontally and vertically. The processes applied and their ratios were experimental, and were preferred for this study because they gave good results.

The same structural neural network classifier was used after each of the 3 CNN architectures used in this study. This co-classifier has three dense layers. After the first two dense layers, there is a 50% dropout layer. There are 128 neurons in the first dense layer and 256 neurons in the second dense layer, and the ReLU was used as an activation function in both layers. In the dense layer at the output of the classifier, there are three neurons representing the number of classes in the dataset, and the Softmax was used as an activation function.

In the training phase of all three models, common parameters were used and all of them were trained under equal conditions. The values of the parameters were obtained experimentally in a way that would give ideal results for the problem that was to be solved in the study. The hyperparameters and their values used in the training of the models are given in Table 1.

Eighty percent of the data set was used as training and 20% as a test set. To evaluate the performance of each CNN model within an ideal period, the number of epochs was set as 100. As a result of the 100 epoch training, accuracy graphs, and confusion matrices that were obtained from each model were given. The success of each model was reported with different metrics obtained from the confusion matrices.

Table 1. Hyperparameters used in the training phase and their values

Parameter	Value
Batch size	16
Number of epochs	100
Optimizer	Adam
Optimizer parameters	lr=0.00001, beta1 = 0.9, beta_2=0.999, verbose=1, epsilon=None, decay=0.0
Learning rate (LR) reduction	ReduceLROnPlateau
LR reduction metrics	patience=3, verbose=1, factor=0.5, min_lr=0.00001

The confusion matrix provides detailed information about the extent to which the model used can distinguish between the classes in the dataset. A truly positive data sample is called True Positive (TP) if it is positively predicted by the classifier, and False Negative (FN) if it is negatively predicted by the classifier. Similarly, if the data sample with a negative class is predicted negatively by the classifier, it is called True Negative (TN), and if it is incorrectly predicted as a positive, it is called False Positive (FP). The metrics produced to express the performance of the model with these values are given in (1), (2), (3), and (4).

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \tag{1}$$

$$Precision = TP/(TP + FP) \tag{2}$$

$$Recall = TP/(TP + FN) \tag{3}$$

$$F1 - Score = 2 * (Precision * Recall)/(Precision + Recall) \tag{4}$$

The accuracy metric characterizes the overall success of the classifier. The precision is the hit rate on samples that the model classifies as positive. The recall (also called sensitivity) shows how many of the true positive values are correctly determined. The F1-Score refers to the balance between precision and recall.

The train and validation accuracy graphs and confusion matrix obtained after 100 epoch training and testing processes of the AlexNet model with the specified parameters are given in Fig. 4.

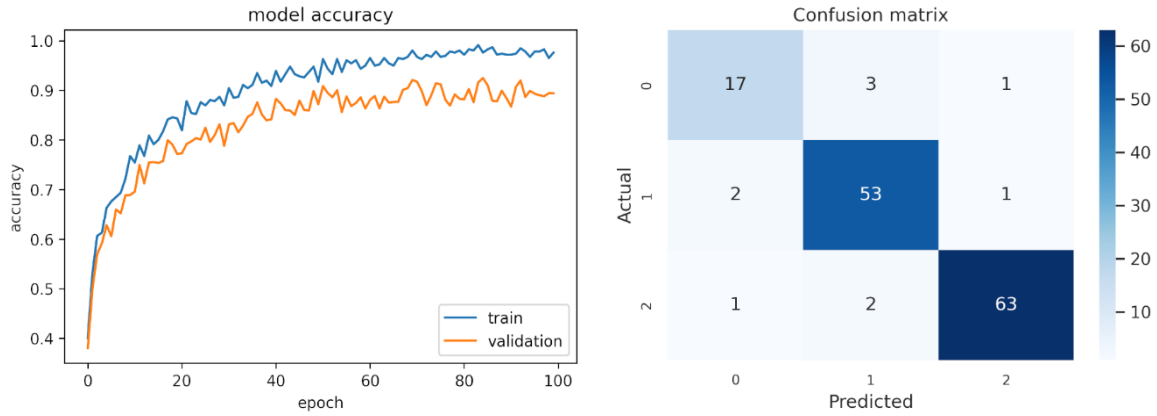

Figure 4. Accuracy graphs and confusion matrix obtained with the AlexNet model

When the accuracy graphs of the model were examined, it showed that the training and validation scores are increasing even though there are oscillations. Although the validation result was slightly lower, it increased in parallel with the training accuracy, indicating that the model did not fall into an overfit condition. Since the optimization of the AlexNet model is more limited compared to the other state-of-the-art model used, its performance is also relatively lower. With this model, the accuracy value reached as a result of 100 epochs was 95.34

When the confusion matrix of the model was examined, it showed that the rate of distinguishing the FCU type labeled as 0 is lower. This model was observed to have more difficulty in distinguishing between the FCU and PFCU classes. The model distinguished the PFCU class at a slightly higher rate, and it was observed that this class was confused with the FCU and to a lesser extent with the PCU class. The class that the model was able to distinguish clearly was the PCU labeled as 2. The model confused this class more with PFCU and less with FCU class. In the general evaluation, it was observed that the success of this model, which was built with AlexNet architecture, was limited compared to the other models. The metrics calculated according to the results obtained from the confusion matrix of the AlexNet model are given in Table 2.

Table 2. Measurements obtained with the AlexNet model

Label	Class	Precision	Recall	F1-Score
0	flaky_corneal_ulcers (FCU)	0.85	0.8095	0.8293
1	point_flaky_mixed_corneal_ulcers (PFCU)	0.9138	0.9464	0.9298
2	point_like_corneal_ulcers (PCU)	0.9692	0.9545	0.9618

When the measurements obtained with the AlexNet model were evaluated, it showed that the PCU is the class that can be distinguished at the highest rate. The precision obtained for this class is 96.92%, recall is 95.45% and F1-Score is 96.18%. The second class with the highest distinction rate was the PFCU. Precision 91.38%, recall 94.64% and F1-Score 92.98% for this class. The class in which the model has the lowest success in distinction is the FCU. The precision obtained for this class is 85%, recall 80.95%, and F1-Score 82.93%.

The accuracy graphs and confusion matrix obtained after 100 epoch training and testing processes of the VGG16 model with the specified parameters are given in Fig. 5.

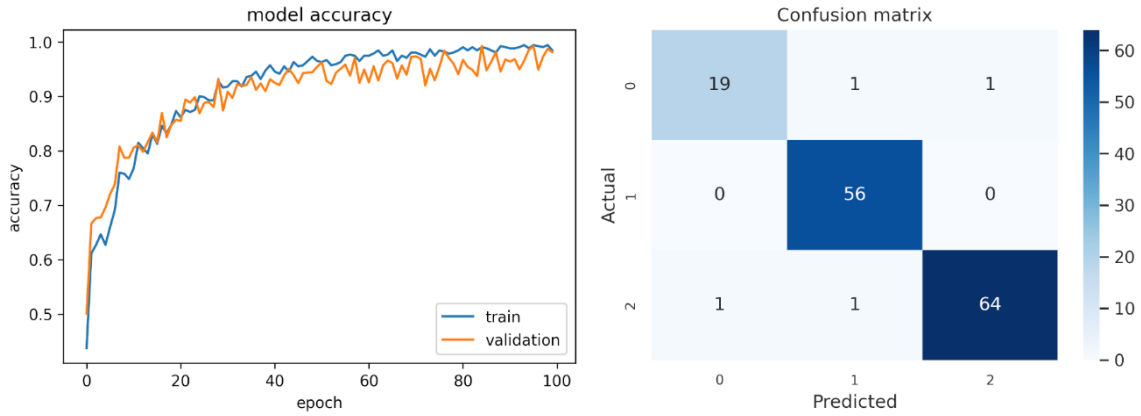

Figure 5. Accuracy graphs and confusion matrix obtained with the VGG16 model

When the accuracy graphs of the VGG16 model were examined, it showed that it is more stable than the AlexNet. The rate of increase in accuracy is relatively higher. Findings were more consistent as expected because the VGG16 architecture is an improved form of AlexNet. The fact that both training and validation accuracies are increased indicates that the model did not have an overfit condition. The accuracy reached with the VGG16 model was 98.14

When the confusion matrix was examined, it showed that the success of the model in distinguishing the FCU class labeled as 0 is relatively low. This class was confused with the PFCU and PCU classes in similar ratios. The model showed the highest success in distinguishing the samples belonging to the one labeled PFCU class. All data samples of the PFCU type were completely distinguished from other types. The PCU class with label 2 is highly distinguishable but confused with the PFCU and FCU classes in equal ratios. The metrics calculated according to the results obtained from the confusion matrix of the VGG16 model are given in Table 3.

Table 3. Measurements obtained with the VGG16 model

Label	Class	Precision	Recall	F1-Score
0	flaky_corneal_ulcers (FCU)	0.95	0.9048	0.9268
1	point_flaky_mixed_corneal_ulcers (PFCU)	0.9655	1.0	0.9825
2	point_like_corneal_ulcers (PCU)	0.9846	0.9697	0.9771

When the measurements obtained with the VGG16 model are examined, the most successful result in terms of the hit rate in the samples classified as positive by the model was obtained for the PCU with 98.46%. This is followed by the PFCU with 96.55% and FCU with 95%. In the correct determination of true positive values, the most successful result was obtained by PFCU at 100%. This is followed by PCU with 96.97% and FCU with 90.48%. In the F1-Score, which shows the balance of these two conditions, the highest success was obtained for PFCU with 98.25%, PCU with 97.71%, and FCU with 92.68%. The most successful results obtained with the VGG16 model were in the PFCU class, and the lowest successful results were in the FCU class.

The CNN model proposed for this study had a simpler architecture compared to the other two models used, AlexNet and VGG16. In this study, it was tested whether a minimal architecture model could compete with more complex models. The accuracy graphs and confusion matrix obtained after 100 epoch training and testing processes using the same hyperparameters as the other models are given in Fig. 6.

When the accuracy graphs of the proposed model were examined, it showed that it increases more rapidly and steadily than the other two models used. According to the accuracy graphs, there was no overfit situation for the proposed model.

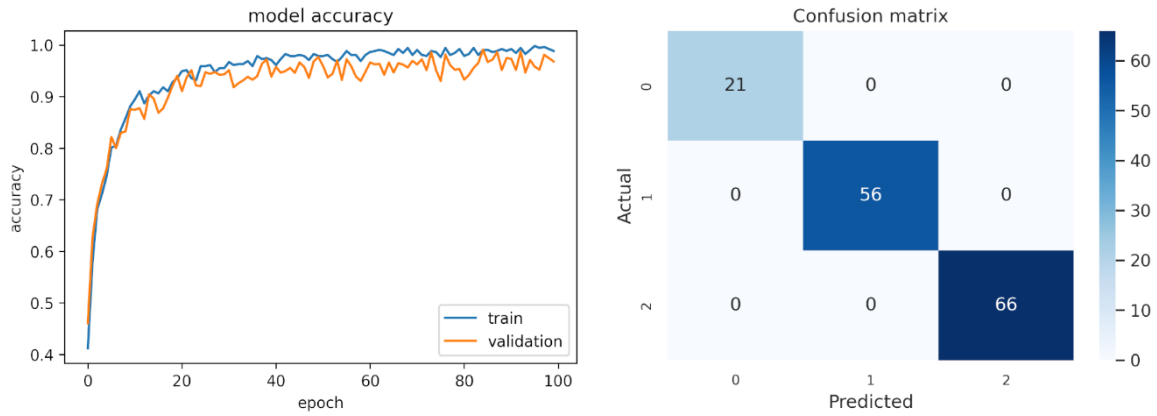

Figure 6. Accuracy graphs and confusion matrix obtained with the proposed model

The oscillations in the graph are less than in the other two models, and the curve tends to flatten in a shorter time. The accuracy value obtained with the proposed model was 100%.

When the confusion matrix of the proposed model was examined, it showed that all classes are clearly distinguished from each other. Any instance of a class was not confused with any other class. Other metrics calculated accordingly are given in Table 4.

Table 4. Measurements obtained with the proposed model

Label	Class	Precision	Recall	F1-Score
0	flaky_corneal_ulcers (FCU)	1.0	1.0	1.0
1	point_flaky_mixed_corneal_ulcers (PFCU)	1.0	1.0	1.0
2	point_like_corneal_ulcers (PCU)	1.0	1.0	1.0

When the measurements obtained were examined, it showed that all classes could be distinguished from each other with 100% success. The fact that the proposed model was less complex than other models and provided higher success showed that the idea about model complexity increasing success is not always true. However, this result is also related to the nature of the dataset used and the preprocessing performed in this study. Therefore, its performance in different problem situations should be tested.

The accuracy values obtained with all models used and the average values of other metrics for all classes are summarized in Table 5.

Table 5. Summary of measurements obtained with all models used

Model	Accuracy (%)	Precision(%)	Recall (%)	F1-Score (%)
AlexNet	95.34	91.10	90.35	90.72
VGG16	98.14	96.67	95.82	96.24
Proposed Model	100	100	100	100

When Table 5 is examined, it shows that the lowest classification success achieved was 95.34%. Although the AlexNet model, which delivered this accuracy is a complicated model, it is less sophisticated than the VGG16 model. The VGG16 is an improved model according to AlexNet and achieved the second highest accuracy rate of 98.14%.

When the examples that the models misclassified were examined, it showed that they belonged to PFCU type cases. Examples of misclassified images are given in Fig. 7.

The PFCU cases combined the characteristics of both PCU and FCU types. PCU typically occurred as small (1 mm or less), sharp-edged and low-depth spots. FCU, on the other hand, produced scratches or a crusty, scaly appearance, usually located in the middle or periphery of the cornea. Due to the coexistence of the features of the FCU and PCU types in the PFCU type, the fact that a feature of any type could suppress the other in general may cause the image class to be determined incorrectly. There are other studies in the literature that support this situation and report that it could be difficult to distinguish the PFCU type from the FCU and PCU types due to the common features (Wang et al., 2021).

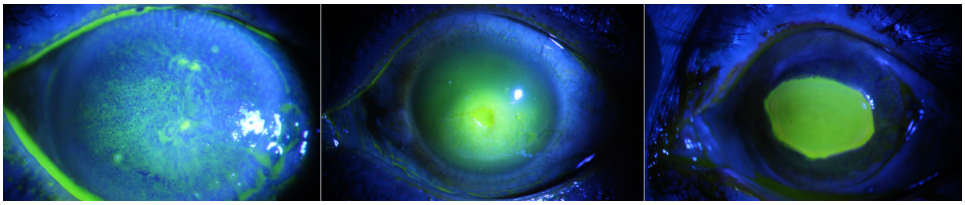

Figure 7. Misclassified sample images

However, both AlexNet and VGG16 models have a deeper and more complex structure, which requires the use and calculation of more parameters in model training. Having more parameters also complicates the optimization, these delays reaching high accuracy values and could cause more oscillations in the result graphs. Although there is a perception that model complexity increases classification accuracy, it was observed in this study that higher accuracy and higher performance could be achieved with a simpler model. The classifier models' goal is to obtain the highest accuracy possible, but a quick response from the model is also anticipated. A model with a simple structure demonstrated quicker reactions with fewer parameters. The proposed model had a simpler structure than the others and delivered the perfect mix between performance and classification success. The proposed model provided 100% accuracy with fewer parameters and reached high accuracy in a shorter time with better performance. This is shown on the model's accuracy graph. While having a high success rate for this problem condition, it should be mentioned that the proposed model has not yet been evaluated in other scenarios. This minimal model may be constrained if the dataset's number of classes and classification complexity rise.

The results obtained in this study with similar studies carried out for the classification of corneal ulcers on eye images colored with the ocular staining technique in the literature are given in Table 6 comparatively.

Table 6. Comparison of this study with similar studies in the literature

Reference	Dataset	Method	Best accuracy (%)
Saini et al. (2003)	study-specific (106 images)	ANN	90.7
Akram and Debnath (2019)	study-specific (513 images)	Proposed CNN	98.99
Kim et al. (2019)	study-specific (1040 images)	ResNet50	92
Tang et al. (2020)	SUSTech-SYSU	Modified VGG	88.89
Xu et al. (2021)	study-specific (115408 images)	DenseNet	66.30
Sajeev and Prem Senthil (2021)	study-specific (446 images)	Proposed CNN	81.2
Gross et al. (2021)	SUSTech-SYSU	Proposed CNN	92.73
Teeyapan (2021)	SUSTech-SYSU	ResNet50	95.10
	Literature average		88.24
This study	SUSTech-SYSU	AlexNet	95.34
		VGG16	98.14
		Proposed CNN	100

The number of studies on corneal ulcers and artificial intelligence in the literature is limited. Most of these studies were carried out in the last three years. Besides the SUSTech-SYSU dataset, the originally collected datasets were also used. The methods used were artificial neural networks and deep neural network models in different architectures. The most successful result in the table was obtained as 98.99% accuracy in the study performed by Akram and Debnath (2019). Although the result obtained with their proposed CNN model is close to the result of the VGG16 model used in this study, it is lower than the result obtained with the proposed CNN model in this study. In addition, the dataset used in that study includes 513 samples, which is less than the number of samples used in this study. Therefore, the generalizability of its success is lower.

Another study with high accuracy was carried out by Teeyapan (2021). In that study, the same dataset was used and an accuracy of 95.10% was obtained in the classification performed with the ResNet50 model. This 95.10% accuracy is lower than any classification in this study. Although both studies use the same data set, this study showed that the

data processing and classification techniques used outperformed that study. Another study using the ResNet50 model was carried out by Kim et al. (2019) on a data set approximately 1.5 times larger, but the accuracy rate remained at 92%. This rate was lower than all the results in this study.

Gross et al. (2021) used the same data set and performed classifications with a proposed CNN model. The accuracy rate of 92.73% achieved was lower than all the models used in this study, and it is approximately 7% lower when compared to the proposed model of this study. Another study using a proposed model was carried out by Sajeev and Prem Senthil (2021). The dataset size used in that study was about two-thirds of this study. The 81.2% accuracy they obtained is at least 14% lower than all the models used in this study, and the limited data set they used indicates lower generalizability.

Tang et al. (2020) performed classification on the same dataset with a variation of the VGG16 architecture used in this study. The accuracy of 88.89% obtained was approximately 10% lower than the accuracy obtained with the VGG16 model used in this study. In addition, it was up to 12% lower than the other two models used in this study. The study conducted by Xu et al. (2021) was carried out using the DenseNet model on a much larger data set compared to other studies in the literature. The accuracy rate remained at 66.30% due to the increase in the number of samples, the increase in the workload, and the performance limitation of the model used. This score they obtained is 30% lower than the general average of this study.

The average accuracy of other studies in literature is approximately 88.24%. In this study, even the lowest-performing AlexNet model has a 7% higher value than this rate. The 100% accuracy provided by the proposed model is 12% higher than the average and also higher than all other studies.

The proposed model of this study showed a higher performance with a minimal structure compared to the complex models both in this study and in the literature.

However, the proposed model yielded effective results in corneal ulcer classification on the ocular staining images, but its performance may vary for different problems. Although it is expected to show high success for similar problems, it should be tested in different problem situations. In addition, although the datasets used in this subject in the literature do not generally contain a very high number of data samples, the fact that the number of original images used in this study is not very high could be considered a limitation. In this study, data augmentation was applied to alleviate the problem.

4. CONCLUSION

A corneal ulcer is a common eye problem, and an accurate diagnosis of the disease reduces the risk of permanent eye damage. However, the diagnosis of the disease requires special expertise. Especially in undeveloped countries, the scarcity of experienced ophthalmologists increases the need for artificial intelligence-based decision support systems for accurate diagnosis. In this study, a deep learning-based approach is presented for the classification of corneal ulcers with high success through ocular staining images. Classifications were performed on the SUSTech-SYSU dataset, consisting of 712 samples of three different types of corneal ulcers, with two different state-of-the-art models and a proposed CNN model. An accuracy of 95.34% was obtained with the AlexNet model, 98.14% with the VGG16, and 100% with the proposed model. When the findings were compared with similar studies in the literature, it was found that the average of the three models used was higher than the other studies, and the proposed model gave better results than all of the existing studies. This study contributes to the literature containing a limited number of studies on this subject. It also revealed that high accuracy can be achieved with models with less complexity for certain problems. In future studies, the proposed model will be tested for performance with similar medical image analyses and for solving different problems.

Peer Review: Externally peer-reviewed.

Conflict of Interest: The author has no conflict of interest to declare.

Grant Support: The author declared that this study has received no financial support.

ORCID ID of the author / Yazarın ORCID ID'si

Onur Sevli 0000-0002-8933-8395

REFERENCES

Akram, A., & Debnath, R. (2019). An Efficient Automated Corneal Ulcer Detection Method using Convolutional Neural Network. 2019 22nd International Conference on Computer and Information Technology (ICCIT), 1–6.

Aksoy, B. (2021). Estimation of Energy Produced in Hydroelectric Power Plant Industrial Automation Using Deep Learning and Hybrid Machine Learning Techniques. *Electric Power Components and Systems, 49*(3), 213–232. https://doi.org/10.1080/15325008.2021.1937401

Amescua, G., Miller, D., & Alfonso, E. C. (2012). What is causing the corneal ulcer? Management strategies for unresponsive corneal ulceration. *Eye, 26*(2), 228–236. https://doi.org/10.1038/eye.2011.316

Basak, S. K., Basak, S., Mohanta, A., & Bhowmick, A. (2005). Epidemiological and microbiological diagnosis of suppurative keratitis in Gangetic West Bengal, eastern India. *Indian Journal of Ophthalmology, 53*(1), 17–22.

Bron, A. J., Argüeso, P., Irkec, M., & Bright, F. V. (2015). Clinical staining of the ocular surface: Mechanisms and interpretations. *Progress in Retinal and Eye Research*, 44, 36–61. https://doi.org/10.1016/j.preteyeres.2014.10.001

Chen, J., & Yuan, J. (2010). Strengthen the study of the ocular surface reconstruction. Chinese Journal of Ophthalmology, 46(1), 3–5.

Cohen, E. J., Laibson, P. R., Arentsen, J. J., & Clemons, C. S. (1987). Corneal ulcers associated with cosmetic extended wear soft contact lenses. *Ophthalmology, 94*(2), 109–114.

Deng, L., Lyu, J., Huang, H., Deng, Y., Yuan, J., & Tang, X. (2020). The SUSTech-SYSU dataset for automatically segmenting and classifying corneal ulcers. *Scientific Data, 7*(1), 23. https://doi.org/10.1038/s41597-020-0360-7

Diamond, J., Leeming, J., Coombs, G., Pearman, J., Sharma, A., Illingworth, C., Crawford, G., & Easty, D. (1999). Corneal biopsy with tissue micro homogenisation for isolation of organisms in bacterial keratitis. *Eye, 13*(4), 545–549.

Garg, P., & Rao, G. N. (1999). Corneal ulcer: Diagnosis and management. *Community Eye Health, 12*(30), 21–23. PubMed.

Gross, J., Breitenbach, J., Baumgartl, H., & Buettner, R. (2021). High-performance detection of corneal ulceration using image classification with convolutional neural networks. E 54th Hawaii International Conference on System Sciences, 3416–3425.

Katara, R. S., Patel, N. D., & Sinha, M. (2013). A clinical microbiological study of corneal ulcer patients at western Gujarat, India. Acta Medica Iranica, 399–403.

Kim, J. Y., Lee, H. E., Choi, Y. H., Lee, S. J., & Jeon, J. S. (2019). CNN-based diagnosis models for canine ulcerative keratitis. Scientific Reports, 9(1), 1–7.

Krizhevsky, A., Sutskever, I.,& Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84–90.

Li, Z., Jiang, J., Chen, K., Chen, Q., Zheng, Q., Liu, X., Weng, H., Wu, S., & Chen, W. (2021). Preventing corneal blindness caused by keratitis using artificial intelligence. *Nature Communications, 12*(1), 1–12.

Maurice, D. M. (1957). The structure and transparency of the cornea. *The Journal of Physiology, 136*(2), 263.

Portela, H. M. B., MS Veras, R. de, Vogado, L. H. S., Leite, D., Sousa, J. A. de, Paiva, A. C. de, & Tavares, J. M. R. (2021). A Coarse to Fine Corneal Ulcer Segmentation Approach Using U-net and DexiNed in Chain. Iberoamerican Congress on Pattern Recognition, 13–23.

Saini, J. S., Jain, A. K., Kumar, S., Vikal, S., Pankaj, S., & Singh, S. (2003). Neural network approach to classify infective keratitis. Current Eye Research, 27(2), 111–116.

Sajeev, S., & Prem Senthil, M. (2021). Classifying infective keratitis using a deep learning approach. 2021 Australasian Computer Science Week Multiconference, 1–4.

Simonyan, K., & Zisserman, A. (2014). Very Deep ConvNets for Large-Scale Image Recognition. CoRR.

Song, X., Xie, L., Tan, X., Wang, Z., Yang, Y., Yuan, Y., Deng, Y., Fu, S., Xu, J., Sun, X., & others. (2014). A multi-center, cross-sectional study on the burden of infectious keratitis in China. *PLoS One, 9*(12), e113843.

Tang, N., Liu, H., Yue, K., Li, W., & Yue, X. (2020). Automatic classification for corneal ulcer using a modified VGG network. 2020 International Conference on Artificial Intelligence and Computer Engineering (ICAICE), 120–123.

Teeyapan, K. (2021). Deep learning-based approach for corneal ulcer screening. The 12th International Conference on Computational Systems-Biology and Bioinformatics, 27–36.

Wang, T., Wang, M., Zhu, W., Wang, L., Chen, Z., Peng, Y., Shi, F., Zhou, Y., Yao, C., & Chen, X. (2021). Semi-MsST-GAN: A Semi-Supervised Segmentation Method for Corneal Ulcer Segmentation in Slit-Lamp Images. *Frontiers in Neuroscience*, 15.

Wang, T., Zhu, W., Wang, M., Chen, Z., & Chen, X. (2021). Cu-segnet: Corneal ulcer segmentation network. 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), 1518–1521.

Wang, Z., Lyu, J., Luo, W., & Tang, X. (2021). Adjacent Scale Fusion and Corneal Position Embedding for Corneal Ulcer Segmentation. *International Workshop on Ophthalmic Medical Image Analysis*, 1–10.

Xu, Y., Kong, M., Xie, W., Duan, R., Fang, Z., Lin, Y., Zhu, Q., Tang, S., Wu, F., & Yao, Y.-F. (2021). Deep sequential feature learning in clinical image classification of infectious keratitis. *Engineering, 7*(7), 1002–1010.

How cite this article

Sevli, O. (2023). A deep learning-based classification study for diagnosing corneal ulcers on ocular staining images. *Acta Infologica, 7*(2), 281-292. https://doi.org/10.26650/acin.1173465