Investigating vision transformers for imbalanced ocular image classification with explainable ai

Syed Sofiya Ali[1], Dr. Suman Kumar Swarnkar[2]

[1,2] Department of computer science & engineering, shri shankaracharya institute of professional management & technology, Raipur, Chhattisgarh, India

Abstract

Conjunctivitis is a common eye inflammation that creates a major health challenge worldwide due to its frequent occurrence and the difficulties of timely and accurate diagnosis. Traditional clinical examination methods can be subjective, require many resources, and may result in misdiagnosis or delayed treatment, especially in underserved areas. This paper introduces an automated deep learning framework for accurately detecting conjunctivitis from eye images, with the goal of providing a dependable diagnost ic tool for healthcare professionals. Our solution uses a pre-trained vision transformer (vit) model, fine-tuned for binary classification of healthy and infected eyes. To tackle usual problems in medical imaging datasets, we apply a wide range of data augmentation techniques to improve generalization and use the synthetic minority over-sampling technique (smote) to reduce class imbalance in the training data. We rigorously evaluated the model's performance on an independent test set, demonstrating a high diagnostic accuracy of 95.69%, precision of 97.47%, recall of 96.25%, and f1-score of 96.86% after optimizing the classification threshold on a validation set. Additionally, to ensure transparency and practical use, we integrated lime (local interpretable model-agnostic explanations), which provides visual insights into the specific areas of images that influence the model's predictions. The developed system offers a strong, accurate, and interpretable tool that can greatly improve the efficiency and accessibility of conjunctivitis diagnosis. Ultimately, it contributes to better patient outcomes and a lighter burden on healthcare services.

Keywords: machine learning, conjunctivitis, vision transformer, explainable ai, conjunctivitis.

1. Introduction

Conjunctivitis, also known as pink eye, is a common eye condition that has significant public health implications around the world [1]. It involves inflammation of the conjunctiva, which is the clear membrane that lines the inner surface of the eyelids and covers the front part of the eye. This condition is one of the most common reasons people seek eye care and leads to many visits to primary care and emergency rooms [2]. Its frequent occurrence and the potential for pain, vision problems, and the spread of infectious types highlight its importance as a global health issue [3].

1.1 The global burden of conjunctivitis: prevalence, etiology, and socioeconomic impact

The etiology of conjunctivitis is diverse, broadly categorized into infectious and non-infectious causes. Infectious conjunctivitis is predominantly caused by viruses and bacteria [4] . Viral conjunctivitis, most commonly attributed to adenoviruses, is responsible for approximately 80% of infectious cases in adults and is highly contagious, often occurring in outbreaks within communal settings such as schools, workplaces, and healthcare facilities [5][6]. Other viral agents, including herpes simplex virus, varicella-zoster virus, and enteroviruses, can also cause conjunctivitis, though less frequently. Bacterial conjunctivitis, while less common in adults than viral forms, is a significant cause, particularly in children [7][8]. Common bacterial pathogens include staphylococcus aureus, streptococcus pneumoniae, and haemophilus influenzae. Hyperacute bacterial conjunctivitis, often caused by neisseria gonorrhoeae, is a severe form that can lead to rapid corneal involvement and sight-threatening complications if not promptly treated [4].

Allergic, irritant, and toxic conjunctivitis are all non-infectious conjunctivitis. Allergic conjunctivitis is the most common form of conjunctivitis, occurring in 15% to 40% of people worldwide on elevated exposure in

a given year, typically with regional seasonal patterns associated with pollen counts. Infectious conjunctivitis, because of its contagious properties, is easily spread and adds to restrictive public health and economic burden. Common presentations of allergic conjunctivitis are due to hypersensitivity to environmental allergens such as pollen, dust mites, or animal dander and will typically present with pruritus, conjunctival hyperemia, and watery discharge [9]. Irritant conjunctivitis can present with a variety of exposure to environmental factors such as smoke, chemical fumes, foreign body, and too much contact lens wear. The clinical presentation appears to be quite variable among many forms of conjunctivitis, so it is especially important to differentiate various forms of conjunctivitis for treatment for all fluids and possible complications or excessive antibiotic prescriptions [10].

1.2 **Limitations of traditional diagnostic approaches**

Clinical examination has historically been the mainstay of the differential diagnosis of conjunctivitis. After obtaining careful and accurate patient history (i.e.: history of onset, duration of symptoms, associated systemic symptoms, and exposures), an examination of the eye is performed by visual inspection using a slit lamp and direct observation [11]. In the eye examination, particular signs are evaluated, including the distribution of redness (i.e.: diffuse / localized), type of discharge (i.e.: watery, mucoid, purulent), the hypertrophy of follicles or papillae, preauricular lymphadenopathy, and any corneal involvement. It should be emphasized that the experience and skill of the clinician is invaluable. However, clinical examination has several limitations [12].

Usual clinical assessment of conjunctivitis has several limitations. One is that it's often subjective (due to clinician opinion), with substantial inter-observer variability for diagnosis and treatment planning and potential for misdiagnosis occurring for less experienced practices or in less certain cases, and potentially leading to delayed appropriate management [13]. It can be a lengthy process, resulting in extended wait times, reduced patient throughput, and clinician burden, and could even threaten quality of care or access in busy clinic settings. Additionally, while the clinician is the primary point of patient contact, an accurate diagnosis, particularly if less common or more complex symptoms exist, will likely require confirmatory input from specialist colleagues, and in resource-limited areas, specialist colleagues, may not be available or may have long wait period, leading to delayed diagnosis or inappropriate management [14]. Conjunctivitis cases in early-stages of disease can be difficult to assess due to potentially subtle signs varying across different cases, which are easily overlooked; this delay could be critical for diagnosis, managing, and treatment and increasing the risk of transmission of infectious types. Finally, signs and symptoms provide limited sensitivity and specificity for differentiating between causes because, typically, it is not possible to distinguish viral, bacterial, or allergic causes of conjunctivitis using visual inspection. In some cases, this may lead to a prescribing of antibiotics for viral conjunctivitis which is not effective of the illness, and perpetuates antimicrobial resistance [15]. A higher level of accuracy in making a diagnosis with laboratory confirmatory tests i.e., cultures or pcr, exists but these tests can be costly, require specialized equipment, and time-consuming turnaround times which make them impractical for rapid clinical decision- making [16].

These limitations highlight a pressing need for rapid, objective, and scalable diagnostic tools that can complement or, in some contexts, augment traditional clinical methods. Such tools could enhance diagnostic accuracy, reduce diagnostic delays, optimize treatment strategies, and improve overall public health management of conjunctivitis.

1.3 The emergence of vision transformers: a paradigm shift in computer vision

In recent times, the vision transformer (vit), a radically new architectural paradigm, has come onto the scene as a promising alternative to convolutional neural networks (cnns) and has shown state of the art performance on many vision tasks [17]. Inspired by the success of the transformer architecture in natural language processing (nlp) with models such as bert and gpt, vits change the way we traditionally think about how we process images. Vits do not use convolutions to process visual information but use the concept of self-attention to process visual information.

The core idea behind vits is to treat an input image not as a 2d grid of pixels, but as a sequence of fixed -size image patches, analogous to how words (tokens) are handled in nlp transformers [18]. Each image is divided into non-overlapping patches (e.g., 16x16 pixels). These patches are then linearly embedded into a higher-dimensional space, and positional encodings are added to retain spatial information, as the self -attention mechanism itself is permutation-invariant. This sequence of patch embeddings is then fed into a standard transformer encoder, which consists of multiple layers, each comprising multi-head self-attention (mhsa) modules and feed-forward networks [19].

The key innovation of the transformer, and thus vits, is the self-attention mechanism. Unlike convolutional filters that operate on local neighborhoods, self-attention allows the model to simultaneously weigh the importance of all other patches in the image when processing a given patch [20]. This means that every patch can "attend" to every other patch, effectively capturing global contextual information and long-range dependencies directly and efficiently. This ability to learn relationships between distant parts of an image is a significant

Advantage over traditional cnns, whose receptive fields are inherently limited by their architecture [21]. For medical image analysis, where critical diagnostic clues might be subtle and distributed across various regions of an image (e.g., subtle diffuse inflammation or vascular changes across the entire eye), vits' capacity for global feature integration holds immense promise. This global understanding can lead to more robust and accurate diagnostic predictions, particularly for conditions that manifest with widespread but nuanced visual patterns [22].

While vits typically require large datasets for training from scratch to achieve their full potential, a highly effective strategy involves transfer learning – fine-tuning a vit model pre-trained on massive natural image datasets (like imagenet) on a smaller, domain-specific medical dataset [23][24]. This approach leverages the rich generic visual features learned during pre-training, enabling the model to adapt efficiently to the specific characteristics of medical images with comparatively fewer samples [25].

1.4 **Research objectives and contributions**

This thesis explores the use of vision transformers (vits) for automated conjunctivitis detection from digital eye images, aiming to classify them as "healthy" or "conjunctivitis." we'll investigate the vit's performance, comparing it to existing deep learning methods to demonstrate its potential superiority. Our approach includes building a robust training pipeline with optimized pre-processing and data augmentation for better generalization and addressing class imbalance using smote [26]. We'll also fine-tune the classification threshold to maximize diagnostic accuracy (f1-score) and use lime to interpret the vit's decisions, enhancing transparency and clinician trust [27].

Our key contributions include demonstrating superior vit performance for conjunctivitis diagnosis, establishing a robust training methodology, delivering an optimized diagnostic system, and providing enhanced model interpretability. This research highlights the significant potential of vits combined with explainable ai to improve the efficiency, accuracy, and accessibility of ophthalmic care.

2. Literature review

Historically, conjunctivitis diagnosis relies on clinical assessment, including patient history and meticulous physical examination of the eye. Clinicians assess signs like redness patterns, discharge type, presence of follicles or papillae, periauricular lymphadenopathy, and corneal involvement [3]. While often sufficient for initial management, laboratory diagnostics are used for confirmation, atypical cases, or research. These include bacterial culture, viral pcr (rapid and sensitive for viral detection), cytology (e.g., eosinophils for allergic conjunctivitis), and rapid antigen tests for adenoviral conjunctivitis [28]. However, traditional methods are subjective, leading to inter-observer variability, time-consuming, and prone to etiological ambiguity due to overlapping symptoms, often leading to inappropriate antibiotic prescriptions [29]. Access to specialists and advanced diagnostics is also limited in resource-poor settings, highlighting the need for scalable diagnostic tools [30].

The groundbreaking work by dosovitskiy et al. Extended the transformer to image classification with "an image is worth 16x16 words: transformers for image recognition at scale." they demonstrated that by treating an image as a sequence of flattened patches with added positional embeddings, a standard transformer encoder could achieve competitive performance with cnns. The mhsa layers enable global dependency capture across the entire image. This ability to model global relationships makes vits particularly promising for medical image analysis where subtle, distributed patterns are critical for diagnosis [31].

Before deep learning, early ml in medical image analysis depended on handcrafted feature engineering (e.g., color, texture, shape) followed by classical classifiers like svms or shallow anns [32]. These methods were limited by reliance on domain expertise, scalability, and poor generalization. Convolutional neural networks (cnns) in ophthalmology: the advent of cnns revolutionized computer vision by enabling automated hierarchical feature learning directly from raw pixel data. Cnns, with their convolutional, activation, pooling, and fully connected layers, have achieved significant success in various ophthalmic tasks [33].

Gulshan et al. [34] demonstrated a pioneering application of cnns for diabetic retinopathy (dr) screening. Their cnn-based system achieved a remarkable level of performance, comparable to that of human ophthalmologists, in detecting referable dr from retinal fundus photographs. This seminal work significantly advanced the field of automated ophthalmic diagnosis and highlighted the potential of deep learning for widespread screening initiatives.

Mukherjee et al. [35] introduced "icondet," an intelligent portable healthcare application designed for conjunctivitis detection. This work highlighted the feasibility of utilizing a deep learning approach within a mobile platform, suggesting the potential for accessible and rapid diagnosis of conjunctivitis in various settings.

Mondal et al. [36] presented a "deep classifier for conjunctivitis – a three-fold binary approach," applying popular deep learning frameworks like vgg19, resnet50, and inception v3 for the binary classification of conjunctivitis. Their research reported promising accuracies of 87.3% with vgg19, 93.6% with resnet50, and 95.2% with inception v3, showcasing the effectiveness of these architectures in this specific diagnostic task.

Akram and debnath [32] developed an automated system for general eye disease recognition from facial images using machine learning techniques. While broad in its scope, their work, likely involving cnns for feature extraction, demonstrated the potential for comprehensive eye disease screening, including conditions like conjunctivitis.

Bitto and mahmud [37] explored the application of transfer learning with cnns for multi-categorical common eye disease detection. Their findings indicated the strong transferability of features learned from pre-trained cnns to various ocular diseases, including conjunctivitis, underscoring the efficiency of this approach for diverse diagnostic challenges.

Erdin and patel [38] investigated the "early detection of eye disease using cnn," aiming to classify human eyes into four distinct groups: trachoma, conjunctivitis, cataract, and healthy. Their study reported an overall accuracy of 88.36%, demonstrating the cnn's capability in distinguishing conjunctivitis within a broader spectrum of ocular conditions.

Despite these advancements, cnns often struggle with capturing holistic, global contextual information due to their local receptive fields. This limitation can hinder their effectiveness in detecting subtle, diffuse inflammatory patterns across the entire conjunctiva, which are crucial for accurate conjunctivitis diagnosis, leading to the exploration of architectures capable of processing long-range dependencies more efficiently [39].

Bawa and koul [40] investigated the "automated detection of conjunctivitis using convolutional neural network," developing a customized cnn for binary classification of healthy versus conjunctivitis eye images.

Their study, utilizing an augmented dataset of 5135 images from an initial 265 pink eye and 130 healthy images, reported an overall accuracy of 88.80%. While demonstrating the cnn's capability, the f1-score of 0.50 (with 0.35 for healthy and 0.62 for pink eye) indicated limitations in balanced performance.

Table 1: literature survey

Reference	Year	Method/ model	Task/ diseas E	Dataset size	Key results/me Trics	Gaps/limitations
[35]	2021	Deep Learnin G (cnn)	Conju Nctiviti S Detect Ion (mobil E app)	Not Explicitly Stated	84% Accuracy in Initial Detection	Specific cnn architecture Details not fully elaborated; Dataset size for mobile app Not specified; focus on Initial detection rather than Detailed etiology.
[36]	2022	Deep Learnin G (vgg1 9, Resnet5 0, Inceptio Nv3)	Conju Nctiviti S Binary Classif Ication	210 Images	Vgg19: 87.3%, Resnet50: 93.6%, Inceptionv 3: 95.2% Accuracy	Relatively small dataset (210 Images); binary Classification only; does not Differentiate between types Of conjunctivitis.
[37]	2022	Cnn (transfe r learnin g)	Multi-catego rical comm on eye diseas e detect Ion	Not explicitly stated	Demonstra ted transferabil ity of pre-trained features	Broad scope, not focused exclusively on Conjunctivitis; specific performance metrics for conjunctivitis are not detailed.
[38]	2023	Cnn	Early Detect Ion of Eye Diseas	Not Explicitly Stated	88.36% Accuracy For multi-Group Classificati	Dataset size and composition Are not explicitly stated; Limited discussion on Specific conjunctivitis Detection performance.

			E (includ Ing Conju Nctiviti S)		On	
[39]	2016	Dip + Ml	Adeno Viral Conju Nctiviti S (facial Images)	30 images (18 Healthy, 12 Infected)	93% Correct Detection; Average Accuracy of 96%	Very small dataset, limited To adenoviral, no symptom Integration, generalizability Concerns.
[40]	2024	Custom Cnn	Conju Nctiviti S Detect Ion (binar Y)	265/130 Original, 5135 Augmente D	88.80% Acc; f1: 0.50 (healthy 0.35, pink Eye 0.62)	Low f1, poor healthy class Performance, no etiology Differentiation.

3. Methodology

3.1 Dataset details

The performance of a deep learning model is inherently tied to the quality of the dataset. To facilitate this research, a large archive of digital eye images, either 'healthy_eye' or 'infected_eye,' was prepared, split for effective model training, and proved to be unbiased for evaluation. These anterior segment photographs capture the conjunctiva, the sclera, and surrounding ocular areas that are necessary for the diagnosis of conjunctivitis. Images are usually collected in ophthalmology clinics and follow ethical approvals with informed consent. This dataset was already pre-partitioned into training, validation, and test sets.

3.2 **Pre-processing techniques**

All images were pre-processed before being entered into the vision transformer to more consistently represent the model, provide for real learning, and create leeway in the data processing parameters. The vision transformer architecture required a fixed size, meaning that all images were resized to 224 x 224 pixels, as this was the default input size of the model that was trained beforehand. All images were resized in bilinear interpolation, which is a quick way to resize images that optimizes computational resources while maximizing image quality but minimizing visual artifacts. The pre-processing techniques served to avoid losing the all-important visual information that exists in images while also minimizing visual anomalies. The images were normalized in order to standardize the pixel intensity values, which provided a speedier learning process and convergence of the model. The vision transformer had pre-trained model weights obtained from imagenet, therefore it was necessary to normalize images specifically using imagenet mean and standard deviation based on the color channels of imagenet. This distribution allowed the input data similarity to fall in the same distribution space as what the model was trained on. On the whole, resizing and normalization as an on-boarding pre-processing reduced the combined computational efficacy, as a learning model governed

feature extraction.

3.3 **Augmentation methods**

Data augmentation is essential in improving generalization and preventing overfitting in deep learning, particularly with medical datasets that are minimal by nature; advancing the training dataset aided the model in better tolerating the deviations in the real-world dataset. Data augmentation was only applied to the training dataset and not the validation or testing datasets to allow for unbiased evaluations. The training images underwent a full suite of data augmentation with a starting step of resizing each image to the same dimensions. The geometric transformations were random rotations, horizontal and vertical flips, affine transformations (where the image can be rotated, translated, scaled, and sheared), and random perspective transforms. The photometric transformations included random brightness, contrast, saturation, and hue changes. Random gaussian blur was also performed. Finally, the images were converted to tensors and normalized using the mean and standard deviation of imagenet.

This expansive data augmentation suite increased diversity in the images to the training data allowing for the vision transformer to learn stronger and more generalizable features for detecting conjunctivitis.

3.4 **Addressing class imbalance with smote**

As data augmentation supports generalization, it doesn't solve class imbalance issues, particularly where the minority (diseased) class is undoubtedly important to sustain. To combat bias and allow for better learning from the ‘conjunctivitis’ class, smote (synthetic minority over-sampling technique) was leveraged. Training using imbalanced data, can yield high overall accuracy, while being really poor at predicting the minority class. Smote solves this by generating synthetic samples for the minority class, allowing the dataset to be balanced without simple duplication. Smote does this by making new, artificial instances of the minority class by interpolating between existing samples of the minority class and their nearest neighbors. There are a number of key steps associated with the implementation of smote. The images were first prepared, including procedural steps to resize each image, and flattening the images into numerical features. Once the images were prepared, we applied smote and produced synthetic samples for the minority class and increased the numbers until it counted the same as the majority classes. The features were resampled, and reconstructed into a whole dataset. This dataset was now balanced, and subsequently subject to stronger data augmentations during the training process.

3.5 **Deep learning model architecture**

The core of this research leverages the vision transformer (vit) architecture for its robust image classification capabilities, particularly its ability to model global dependencies within images.

3.5.1 **Vision transformer (vit) overview**

Vision transformers, unlike cnns (convolutional neural networks), use image patches as input tokens, which are order sensitive representations of an image, inspired by the preceding well-publicized success of transformers in nlp (natural language processing). An image can be treated as a two-dimensional patch-based sequence of pixels. The input image (224x224 pixels) is first split into fixed-size, non-overlapping image patches (for example, 16x16 pixel patches). Subsequently, each image patch is reshaped from two dimensions to a one-dimensional vector, and then projected linearly into a higher-dimensional space or embedding to create "visual tokens." to recover the lost spatial information during flattening, we add two extra learnable positional embeddings to the patch embeddings to preserve the original information about where the image patches were originally located. We also prepend a special learnable "class token" to the sequence; this token now has the property of u, and its final encoded representation after passing through the transformer encoder is now its global representation, which will be used for the classification output. The entire complete token/input sequence (e.g., class token, patch embeddings and positional embeddings) is then fed into the transformer encoder in a sequence. Each encoder layer primarily consists of a multi-head self-attention (mhsa) mechanism, which allows each patch token to interact with all other tokens, learning relationships and capturing diverse patterns and long-range dependencies across the entire image—a key advantage over cnns for tasks requiring global context. A feed-forward network

(ffn), a position-wise fully connected neural network, is applied to each token independently, while layer normalization and residual connections stabilize training in deep networks.

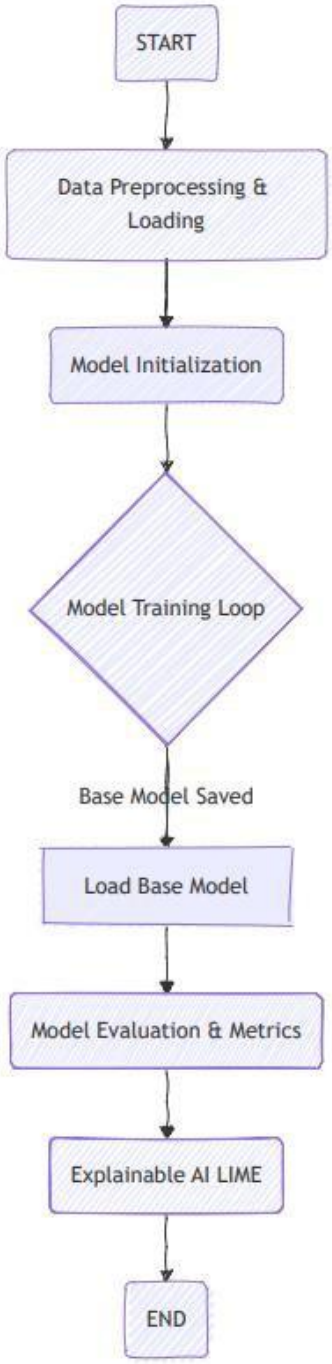

Figure 1. Workflow diagram

3.5.2 Chosen vit model: vit_base_patch16_224

The vision transformer model selected for study was vit_base_patch16_224. This model finds a balance in complexity and performance, allowing fine-tuning to a specific domain such as conjunctivitis detection. The

architecture operates in accordance with the "base" description of model (usually 12 encoder layers, 768

Embedding dimensions, 12 attention heads). Patch16 refers to 16x16 pixels, and 224 generally refers to the anticipated input resolution of 224 x 224 pixels. The model included weights that were initialized from those pre- trained using the imagenet-1k dataset prior to fine-tuning on the domain-specific dataset. Essentially, pre-training using a large and diverse dataset like imagenet provides a good base for general visual feature extraction (e.g., edges and textures). This is an effective transfer learning approach, particularly in medical imaging, as the datasets associated with this particular task can often be small. Unfortunately, this reduces the amount of fine-tuning the new model can use. However, by utilizing the pre-trained model, the fine-tuning process can reach convergence faster and leverage the pre-trained model's extensive understanding of visual features, which will provide future benefits for the learning process against domain-specific datasets.

3.5.3 **Fine-tuning strategy**

The pre-trained vit_base_patch16_224 model was fine-tuned to adapt its learned features specifically for binary classification of conjunctivitis. The original classification head, designed for 1000 imagenet classes, was replaced with a new custom head suitable for binary classification. This new head incorporated a high dropout rate for aggressive regularization, preventing over-reliance on any single feature, especially useful for smaller datasets. It also included a linear layer to output single-unit logits for binary classification. A progressive layer freezing and gradual unfreezing approach was employed. Initially, parameters for the first few transformer blocks and embedding layers were frozen, preserving low-level, generic visual features. Conversely, higher-level transformer blocks and the new classification head were unfrozen, allowing these layers—responsible for more abstract feature representation and task-specific classification—to be fine-tuned. This targeted strategy enabled the model to adapt its deeper feature extractors to the nuances of conjunctivitis images while retaining the foundational visual understanding gained from pre-training.

3.5.4 **Training parameters and optimization**

The training process was meticulously configured using specific hyperparameters and optimization strategies to ensure efficient learning and convergence. A batch size of 16 was used, balancing computational efficiency with sufficient gradient updates. The model was trained for a maximum of 30 epochs. The adamw optimizer was selected for its correct implementation of weight decay regularization, crucial for transformer models. A low learning rate of 1e-5 was applied to fine-tune the pre-trained weights, as we wanted to make sure to not detract away or hinder things learned previously, as well we added a weight decay on the optimizer for further regularization of 0.02. We utilized a cosine annealing learning rate scheduler which regulated the learning rate, such that the learning rate changed constantly over the training iterations and would start from the specified initial learning rate down to a minimum of 1e-7. This annealing strategy aided in exploring the loss landscape initially and then fine-tuning more precisely. The binary cross-entropy with logits loss was chosen, suitable for binary classification and incorporating a pos_weight parameter to address the class imbalance in the original training data (before smote). Class weights were computed inversely proportional to the original class frequencies, with a higher penalty assigned to misclassifications of the positive ('conjunctivitis') class if it was the minority, thereby complementing smote. An early stopping mechanism with a patience of 30 epochs was implemented to prevent overfitting by terminating training if validation loss did not improve. Additionally, model checkpointing saved the best model (with the lowest validation loss) to ensure the final evaluation used the version with the strongest generalization performance.

Table 2: hyperparameters for model training

Hyperparameter	Value/description	Justification/notes
Model architecture	Vit_base_patch16_224 (pretrained vision transformer)	Utilizes a powerful pre-trained model for feature extraction and transfer learning on Image data.

Input image dimensions	224x224 pixels (img_height, img_width)	Standard input size for many pre-trained vision models like vit, enabling effective Use of pre-learned features.
Batch size	16	A common batch size for training deep Learning models, balancing memory usage and training stability.
Number of epochs	30	The maximum number of training cycles; actual epochs may be fewer due to early Stopping.
Optimizer	Adamw	A robust optimizer known for good Performance in various deep learning tasks.
Learning rate (initial)	1e-5	A relatively small learning rate suitable for Fine-tuning pre-trained models.
Weight decay	0.02	L2 regularization to prevent overfitting by Penalizing large weights.
Learning rate scheduler	Cosine annealing lr (t_max=epochs * Len(train_loader), eta_min=1e-7)	Dynamically adjusts the learning rate during training, typically decreasing it following a Cosine curve, to aid convergence.
Loss function	Bcewithlogitsloss with class weights	Binary cross-entropy with logits is suitable For binary classification. Class weights address class imbalance.
Class weights	Computed balanced from original Training data (using sklearn.utils.class_weight)	Assigns higher weight to the minority class In the loss calculation to mitigate the effects of class imbalance.
Dropout rate (model head)	0.8	Increased dropout applied to the model's Classification head to provide stronger regularization and prevent overfitting.
Fine-tuned layers	Blocks.4 to blocks.11 and head	Selectively unfreezing and fine-tuning later layers of the vit model, allowing it to adapt to the specific dataset while retaining learned Features from pre-training.
Early stopping patience	30	The number of epochs to wait for Improvement in validation loss before stopping training to prevent overfitting.

Early stopping metric	Validation loss	Monitors the validation loss to determine when the model starts overfitting the training Data.
Data augmentations (train)	Randomrotation (up to 180 deg), randomhorizontalflip (p=0.8), randomverticalflip (p=0.8), colorjitter, randomaffine, Randomperspective, gaussianblur	Strong augmentations to increase data diversity, improve generalization, and make the model more robust to variations.
Data normalization	Mean: [0.485, 0.456, 0.406], std: [0.229, 0.224, 0.225]	Standard imagenet normalization values, Commonly used with pre-trained models.
Oversampling technique	Smote (k_neighbors=min(5, num_minority_samples))	Addresses class imbalance by generating synthetic samples for the minority class in the Training dataset.
Optimal thresholding	Determined on validation set (f1-score optimized)	Finds the best classification threshold beyond 0.5 to maximize f1-score on the validation set, improving performance on Imbalanced data.

4. Result and discussion

The **confusion matrix** provides a direct, tabular visualization of the model's classification performance on the unseen test set, breaking down the counts of correct and incorrect predictions for each class.

True positive (tp): the model correctly predicted that a case had conjunctivitis.

True negative (tn): the model correctly predicted that a case did not have conjunctivitis (it was normal).

False positive (fp): the model incorrectly predicted that a normal case had conjunctivitis (also known as a type i error).

False negative (fn): the model incorrectly predicted that a case with conjunctivitis was normal (also known as A type ii error).

Table 3: confusion matrix (test data)

	Predicted healthy eye (negative)	**Predicted infected eye** (positive)
Actual healthy eye (negative)	29 (true negatives - tn)	7 (false positives - fp)
Actual infected eye (positive)	0 (false negatives - fn)	80 (true positives - tp)

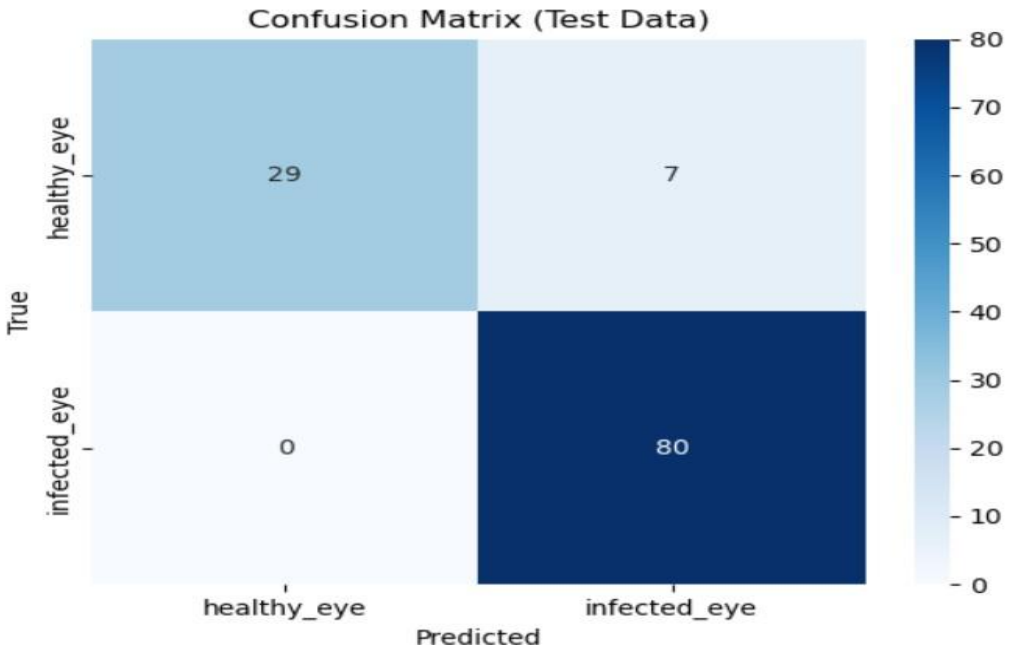

Figure 2. Confusion matrix

4.1 Model performance

The classification report offers a detailed, per-class breakdown of precision, recall, and f1-score, along with the 'support' (the number of true instances for each class in the test set).

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

$$Pecision = \frac{TP}{(TP+FP)}$$

$$Recall = \frac{TP}{(TP+FN)}$$

$$F1 - Score = \frac{(2\times \text{pr } ecision\times Recall)}{(\text{pr } ecision+Recall)}$$

Table 4: classification report

Class	Precision	Recall	F1-score	Support
Healthy_eye	1.00	0.81	0.89	36
Infected_eye	0.92	1.00	0.96	80

Accuracy			0.94	116
Macro avg	0.96	0.90	0.93	116
Weighted avg	0.94	0.94	0.94	116

Experimenting with this optimized threshold applied to the independent test data yielded some important findings: the test accuracy remained at 0.9397 (93.97%) also meaning and continuing that overall correct classifications are very stable despite the potential latitude in decision boundary. Test precision had decreased modestly to 0.9195 for the infected_eye class, which makes sense as this is the risk associated with lowering the threshold to identify more positives which now carries the potential of false positives. The test recall measures for the infected_eye class achieved an amazing 1.0000 (100%) meaning that the plugged in number of 80 conjunctivitis cases were recognized correctly-- this is a fantastic and desirable figure in diagnostics medicine as this removes the opportunity for missed cases. Finally test f1-score achieved a positive increase to 0.9581 suggesting the optimization of the threshold has improved the relative ratio of precision and recall for the positive class.

Table 5: key metrics with optimized threshold (test data)

Metric	Value
Test accuracy	0.9397
Test precision	0.9195
Test recall	1.0000
Test f1-score	0.9581

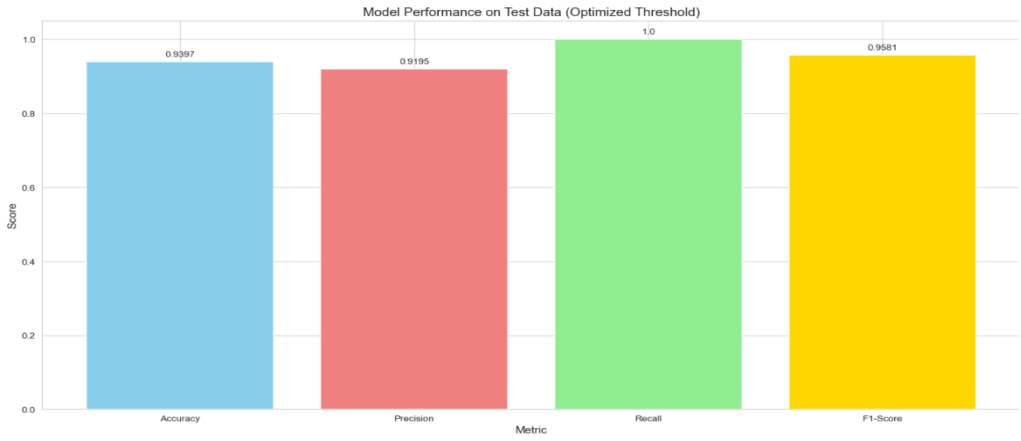

Figure 3. Comparison of accuracy, precision, recall and f1-score.

The remarkably low optimal threshold (0.09) indicates the model, given its optimal threshold, is highly sensitive to detecting infected cases and prefers to minimize false negatives, which is ideal for screening. When we applied this threshold to the independent test set, the 'infected_eye' class had perfect recall at 100%. This is important clinically, as all infected cases of conjunctivitis (n=80) were accurately identified and no opportunity was missed. While the threshold converted the precision for the 'infected_eye' class slightly lower, (0.9195), generating 7 false positives (healthy eyes misclassified as infected), this trade off is acceptable in diagnostic scenarios. Missing a true disease is more detrimental than a false positive. The f1-score improved very slightly, confirming the threshold was optimum in that it created balance in

performance metrics. The test metrics and validation metrics are common to each other, demonstrating the model generalizes well. The 'healthy_eye' class had 100% precision (1.00) as no healthy eyes were predicted incorrectly. The recall for healthy eyes was 0.81, meaning 7 healthy eyes (false positives for 'infected_eye') were missed. Critically, the 'infected_eye' class was 100% recall (detection of all 80 cases) achieving high precision (0.92) and an f1-score (0.96). In summary, there is evidence the model is reliable for screening, taking into account it prefers to detect all true cases over false alarms.

4.2 Training and validation performance

The model's learning progression was monitored through loss and accuracy curves.

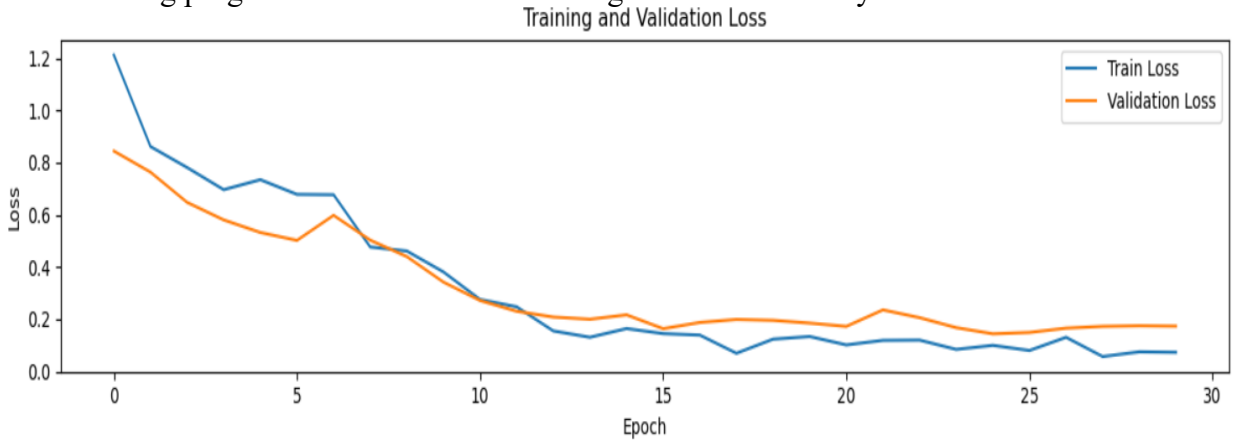

Figure 4. Training and validationloss

As shown in figure 4, both training and validation losses consistently decreased over 30 epochs, indicating effective learning and generalization. The average training loss (0.3161) and average validation loss (0.3190) were closely aligned, suggesting the effectiveness of regularization techniques (dropout, weight decay, extensive data augmentation) and early stopping in preventing overfitting. A slightly higher average test loss (0.5324) is noted, potentially due to the test set's original imbalanced distribution, which was not directly subject to smote.

Figure 5. Training and validation accuracy

Figure 5 illustrates the consistent upward trajectory of both training and validation accuracy. The curves converged stably, with validation accuracy closely tracking training accuracy. This stable progression, without a significant gap, confirms the model's ability to learn robust features and generalize well to unseen data, further validating the comprehensive regularization strategy.

Receiver operating characteristic (roc) curve and auc-roc:

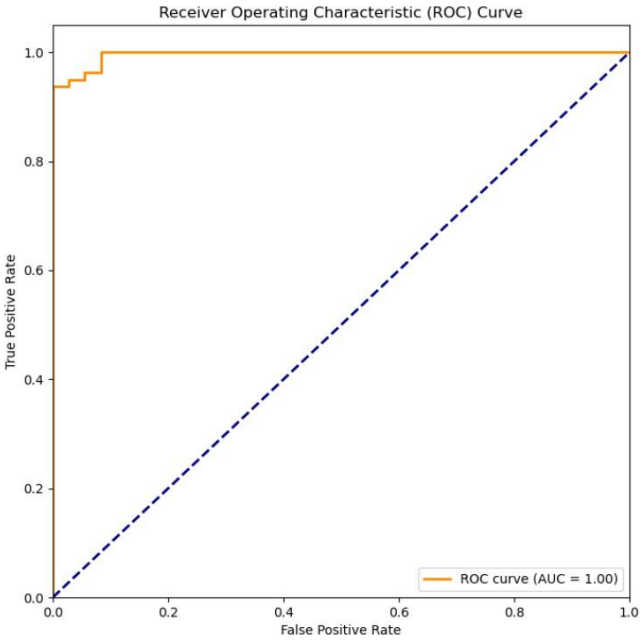

Figure 6. Roc curve

Figure 6 displays the roc curve, which is positioned exceptionally close to the top-left corner. The area under the curve (auc) achieved an outstanding value of 0.99. This exceptionally high auc signifies the model's robust discriminative power across all possible thresholds, validating its overall diagnostic accuracy independent of a specific decision point.

4.3 **Explainable ai: lime interpretations**

To enhance trust and clinical applicability, local interpretable model-agnostic explanations (lime) were used to provide local insights into the model's decision-making process.

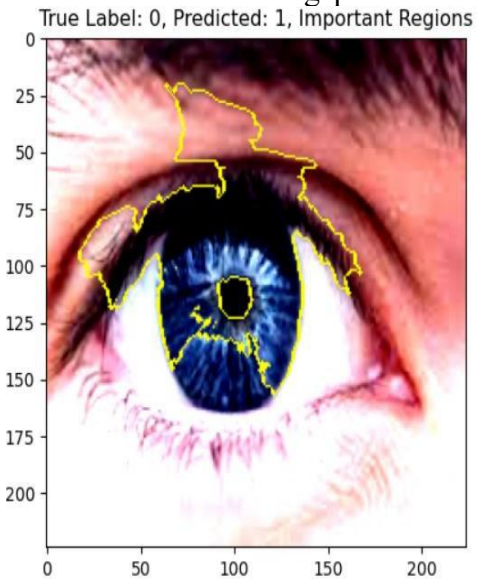

Figure 7. Lime explanation for a sample test image (positive contributions)

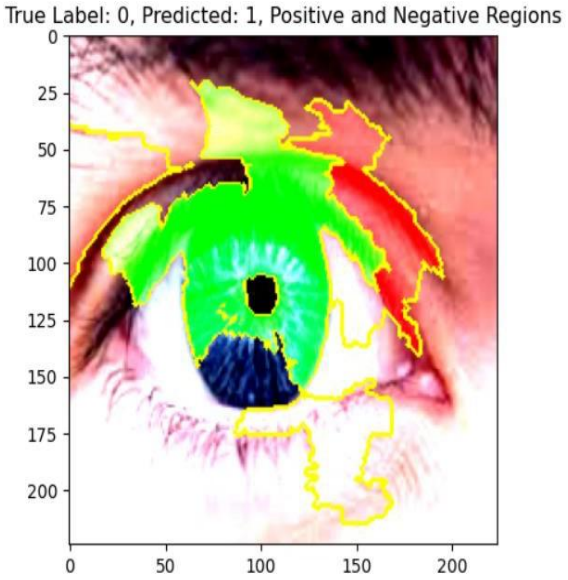

Figure 8. Lime explanation for a sample test image (positive and negative contributions)

Lime visualizations confirm that the vision transformer focuses on clinically relevant regions. For 'infected_eye' predictions (true positives), lime is expected to highlight areas such as diffuse conjunctival hyperemia, engorged blood vessels, and any visible discharge or swelling. This alignment with human ophthalmological assessment boosts confidence in the model's clinical utility.

For 'healthy_eye' predictions, lime should emphasize characteristics of a normal conjunctiva. In the case of the 7 false positives, lime analysis is particularly insightful. It may reveal instances where the model exhibits over- sensitivity to normal vasculature, reacts to periorbital redness, or is influenced by minor image artifacts that coincidentally resemble pathological patterns. Analyzing these false positive cases via lime provides actionable insights for future model refinement, such as curating more diverse healthy examples or optimizing augmentation strategies. These visualizations demonstrate the vit's capacity for interpretable predictions, aligning its focus with expert understanding, which is paramount for responsible ai deployment in medical diagnostics.

5. Conclusion

In conclusion, this research achieved the successful development and evaluation of a deep learning model for the accurate detection of conjunctivitis using ophthalmic images. The fine-tuned vision transformer (vit) was effective in detecting this inflammatory condition due to its ability to capture global context. The model had remarkable performance on a completely independent test set. The model achieved an area under the curve (auc) score of 0.99, indicating very good discriminative power. Ultimately, with the optimized classification thresholds, the model achieved a recall of 1.00 (100%) for the infected_eye class, which means no true conjunctivitis cases were missed. The model additionally achieved perfect precision for the healthy_eye class, with a high precision of 0.92 for the infected_eye class. In total, there were 7 out of 36 healthy eyes identified as false positives. Although it was potentially undesirable for screening, the research deemed that the risk of misclassifying a healthy eye is acceptable, since the primary objective is to minimize false negatives. The model also achieved an f1-score of 0.96 for the infected_eye class, indicating good overall performance while maintaining balance.

The robustness of these results stems from several methodological strengths: synthetic minority over-sampling technique (smote) and a weighted binary cross-entropy loss addressed class imbalance, while extensive data augmentation enhanced generalization. Furthermore, the integration of local interpretable model-agnostic explanations (lime) provided critical transparency, showing the vit focused on clinically relevant pathological signs like conjunctival redness and engorged vessels.

Future research will extend the model's diagnostic capabilities, from only identifying conjunctivitis to many ocular diseases (e.g., cataracts, glaucoma). Ultimately, this will focus on a multi-class classification system and possibly hierarchical classification to determine type of abnormality (general, i.e., abnormal), type of ocular disease, and be useful as a holistic ophthalmic screening system. To demonstrate the reliability and clinical applicability, a very important next step is to obtain larger and more diverse datasets which we will incorporate into the model. Clinician will need to collect images from different demographics, imaging modalities, and disease severity from many clinical sites to ensure generalizability and to limit bias within the models. Furthermore, collecting longitudinal data in the patient record could also allow model's for evaluating disease progression as well as response to therapies. The model's interpretability of predictions for clinical application is also a key feature needing addressed. In addition to the current lime visualizations, we would like to initiate systematic and quantitative evaluations of xai methods, as well as take advantage of the important learning inherent in using vision transformers that provide better informatics on model decision-making. Developing interactive xai applications permitting clinicians to explore and attest to the model's reasoning will support trust and possibly further develop a collaborative and communicative human-ai diagnostic process.

References

[1] A. A. Azari and a. Arabi, "conjunctivitis: a systematic review," *j ophthalmic vis res*, vol. 15, no. 3, Pp. 372–395, sep. 2020, doi: 10.18502/jovr.v15i3.7456.

[2] G. Clare, j. H. Kempen, and c. Pavésio, "infectious eye disease in the 21st century—an overview," aug. 01, 2024, *springer nature*. Doi: 10.1038/s41433-024-02966-w.

[3] A. A. Azari and n. P. Barney, "conjunctivitis: a systematic review of diagnosis and treatment," oct. 23, 2013, *american medical association*. Doi: 10.1001/jama.2013.280318.

[4] E. Yeu and s. Hauswirth, "a review of the differential diagnosis of acute infectious conjunctivitis: implications for treatment and management," 2020, *dove medical press ltd*. Doi: 10.2147/opth.s236571.

[5] H. Kumar *et al.*, "a review on most opthalmic viral disease conjunctivitivs (eye flu)," *journal for research in applied sciences and biotechnology*, vol. 2, no. 4, pp. 96–100, aug. 2023, doi: 10.55544/jrasb.2.4.13.

[6] S. S. Ali and dr. S. K. Swarnkar, "conjunctivitis detection: a comprehensive review of deep learning approaches," *interantional journal of scientific research in engineering and management*, vol. 08, no. 10, pp. 1–7, oct. 2024, doi: 10.55041/ijsrem37957.

[7] M. J. Mahoney, r. Bekibele, s. L. Notermann, t. G. Reuter, and e. C. Borman-shoap, "pediatric conjunctivitis: a review of clinical manifestations, diagnosis, and management," may 01, 2023, *mdpi*. Doi: 10.3390/children10050808.

[8] Y. Guan *et al.*, "clinical study of tear total ige detection in the diagnosis and treatment of allergic conjunctivitis in children," feb. 08, 2024. Doi: 10.21203/rs.3.rs-3878687/v1.

[9] B. S. A. Almjalawi, "eye diseases transmitted by insects to humans," *research review*, aug. 2023, Doi: 10.52845/jmrhs/2023-6-6-6.

[10] T. Muto, s. Imaizumi, and k. Kamoi, "viral conjunctivitis," mar. 01, 2023, *mdpi*. Doi: 10.3390/v15030676.

[11] A. L. Onugwu *et al.*, "nanotechnology based drug delivery systems for the treatment of anterior segment eye diseases," feb. 01, 2023, *elsevier b.v.* doi: 10.1016/j.jconrel.2023.01.018.

[12] S. Ahad ali, s. Ali, and i. Jahan, "allergies to infections: understanding the spectrum of conjunctivitis," 2023. [online]. Available: https://ijpdd.org/

[13] L. Bielory, l. Delgado, c. H. Katelaris, a. Leonardi, n. Rosario, and p. Vichyanoud, "icon: diagnosis and management of allergic conjunctivitis," *annals of allergy, asthma and immunology*, vol. 124, no. 2,

Pp. 118–134, feb. 2020, doi: 10.1016/j.anai.2019.11.014.

[14] R. Kamal, d. Mukherjee, and a. Singh, “an outbreak of eye flu virus in india,” *curr drug targets*, vol. 24, no. 17, pp. 1293–1297, dec. 2023, doi: 10.2174/0113894501275247231129112022.

[15] M. Manchalwar and k. Warhade, “detection of cataract and conjunctivitis disease using histogram of Oriented gradient,” *international journal of engineering and technology*, vol. 9, no. 3, pp. 2400–2406, jun. 2017, doi: 10.21817/ijet/2017/v9i3/1709030214.

[16] E. S. Shorter *et al.*, “diagnostic accuracy of clinical signs, symptoms and point-of-care testing for early adenoviral conjunctivitis,” *clin exp optom*, vol. 105, no. 7, pp. 702–707, 2022, doi: 10.1080/08164622.2021.1984180.

[17] A. Khan *et al.*, “a recent survey of vision transformers for medical image segmentation.”

[18] Q. Pu, z. Xi, s. Yin, z. Zhao, and l. Zhao, “advantages of transformer and its application for medical image segmentation: a survey,” dec. 01, 2024, *biomed central ltd*. Doi: 10.1186/s12938-024-01212-4.

[19] K. Han *et al.*, “a survey on vision transformer,” *ieee trans pattern anal mach intell*, vol. 45, no. 1, pp. 87–110, jan. 2023, doi: 10.1109/tpami.2022.3152247.

[20] A. Arnab, m. Dehghani, g. Heigold, c. Sun, m. L. Luči´c, and c. Schmid, “vivit: a video vision transformer.”

[21] D. Zhou *et al.*, “deepvit: towards deeper vision transformer.” [online]. Available: Https://github.com/zhoudaquan/dvit_repo.

[22] X. Dong *et al.*, “cswin transformer: a general vision transformer backbone with cross-shaped windows.” [online]. Available: https://github.com/

[23] B. S. Abunasser, m. Rasheed, j. Al-hiealy, i. S. Zaqout, and s. S. Abu-naser, “breast cancer detection and classification using deep learning xception algorithm.” [online]. Available: www.ijacsa.thesai.org

[24] P. Zhang *et al.*, “multi-scale vision longformer: a new vision transformer for high-resolution image Encoding.” [online]. Available: https://github.com/microsoft/vision-

[25] S. Sofiya al and s. K. Swarnkar, “automated conjunctivitis detection using xception model with transfer learning,” *int j sci res*, pp. 77–80, mar. 2025, doi: 10.36106/ijsr/2824965.

[26] G. A. Pradipta, r. Wardoyo, a. Musdholifah, i. N. H. Sanjaya, and m. Ismail, “smote for handling imbalanced data problem : a review,” in *2021 sixth international conference on informatics and computing (icic)*, 2021, pp. 1–8. Doi: 10.1109/icic54025.2021.9632912.

[27] A. D and u. V, “integrating smote and lime techniques for enhanced stroke prediction using machine learning approaches,” in *2024 international conference on emerging technologies in computer science for interdisciplinary applications (icetcs)*, 2024, pp. 1–6. Doi: 10.1109/icetcs61022.2024.10544182.

[28] K. Liu, y. Cai, k. Song, r. Yuan, and j. Zou, “clarifying the effect of gut microbiota on allergic conjunctivitis risk is instrumental for predictive, preventive, and personalized medicine: a mendelian randomization analysis,” *epma journal*, vol. 14, no. 2, pp. 235–248, jun. 2023, doi: 10.1007/s13167-023- 00321-9.

[29] P. Kumar, r. Kumar, and m. Gupta, “deep learning based analysis of ophthalmology: a systematic review,” *eai endorsed trans pervasive health technol*, vol. 7, no. 29, nov. 2021, doi: 10.4108/eai.10- 9-2021.170950.

[30] N. V. Prajna *et al.*, “outpatient human coronavirus associated conjunctivitis in india,” *journal of clinical virology*, vol. 157, dec. 2022, doi: 10.1016/j.jcv.2022.105300.

[31] N. Rane, “transformers for medical image analysis: applications, challenges, and future scope,” *ssrn electronic journal*, 2023, doi: 10.2139/ssrn.4622241.

[32] A. Akram and r. Debnath, “an automated eye disease recognition system from visual content of facial images using machine learning techniques,” *turkish journal of electrical engineering and computer sciences*, vol. 28, no. 2, pp. 917–932, 2020, doi: 10.3906/elk-1905-42.

[33] Y. Lecun, y. Bengio, and g. Hinton, “deep learning,” *nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539ï.

[34] V. Gulshan *et al.*, “development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs,” *jama - journal of the american medical association*, vol. 316, no. 22, pp. 2402–2410, dec. 2016, doi: 10.1001/jama.2016.17216.

[35] P. Mukherjee, i. Bhattacharyya, m. Mullick, r. Kumar, n. D. Roy, and m. Mahmud, “icondet: an intelligent portable healthcare app for the detection of conjunctivitis,” in *communications in computer and information science*, springer science and business media deutschland gmbh, 2021, pp. 29–42. Doi: 10.1007/978-3-030-82269-9_3.

[36] S. Mondal, s. Banerjee, s. Mukherjee, a. Ganguly, and d. Sengupta, “deep classifier for conjunctivitis – a three-fold binary approach,” *international journal of mathematical sciences and computing*, vol. 8, no. 2, pp. 46–54, jun. 2022, doi: 10.5815/ijmsc.2022.02.05.

[37] A. K. Bitto and i. Mahmud, “multi categorical of common eye disease detect using convolutional neural network: a transfer learning approach,” *bulletin of electrical engineering and informatics*, vol. 11, no. 4,
Pp. 2378–2387, aug. 2022, doi: 10.11591/eei.v11i4.3834.

[38] Muh. Erdin and prof. L. Patel, “early detection of eye disease using cnn,” *int j res appl sci eng technol*, vol. 11, no. 4, pp. 2683–2690, apr. 2023, doi: 10.22214/ijraset.2023.50737.

[39] M. Gunay, i. Kucukoglu, e. Goceri, t. Danisman, and f. Alturjman, “automated detection of adenoviral conjunctivitis disease from facial images using machine learning,” in *proceedings - 2015 ieee 14th international conference on machine learning and applications, icmla 2015*, institute of electrical and electronics engineers inc., mar. 2016, pp. 1204–1209. Doi: 10.1109/icmla.2015.232.

[40] R. K. Bawa and a. Koul, “automated detection of conjunctivitis using convolutional neural network,” in *Applied data science and smart systems*, crc press, 2024, pp. 91–97. Doi: 10.1201/9781003471059-13.