EAI Endorsed Transactions
on Intelligent Systems and Machine Learning Applications Research Article EAI.EU

EyeFusionNet: A Hybrid CNN-Transformer Network for Automated Diagnosis of Eye Diseases using Colour Fundus Imaging

Akshaj Singh Bisht[1], Sagar Singh Chauhanr[2], Armaano Ajay[1], Rishi Manda[3] and Prasanna Bharathi S[4,*]

[1]School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India
[2] Department of Computer Science, Virginia Tech, Blacksburg, Virginia, USA
[3]Electrical and Computer Engineering, National University of Singapore, Singapore
[4] School of Electronics Engineering, Vellore Institute of Technology, Chennai, India

Abstract

INTRODUCTION: The timely and accurate diagnosis of eye conditions, such as myopia, cataracts, and diabetic retinopathy, is critical for preventing vision loss. Conventional diagnostic techniques depend on the manual examination of eye fundus images, a process that is often labour-intensive and susceptible to human errors.
OBJECTIVES: The proposed work aims to develop an automated deep learning-based solution for classifying eye diseases using the newly released Eye Disease Image Dataset with 5,335 images and 10 classes, addressing limitations in traditional manual diagnostic techniques.
METHODS: The proposed EyeFusionNet employs a dual-track architecture combining DenseNet169 for detailed local feature extraction and Transformer-iN-Transformer for capturing global context and long-range dependencies. The outputs are fused and refined using efficient channel attention to focus on important regions within the fundus images, with explainable artificial intelligence used to provide visualisations to doctors to validate the diagnosis.
RESULTS: The EyeFusionNet achieved an accuracy of 89% on the publicly available dataset of eye fundus images, outperforming state-of-the-art CNNs and transformer models, showcasing the potential of deep learning for eye fundus image diagnosis.
CONCLUSION: The proposed EyeFusionNet introduces a dual-track architecture for eye disease classification, providing a strong foundation for advancing automated diagnostic tools in clinical settings with the help of deep learning and explainable artificial intelligence to empower clinicians.

Keywords: Deep Learning, Eye fundus, Eye disease classification, CNN, Ophthalmology

Received on 05 November 2025, accepted on 20 February 2026, published on 25 February 2025

doi: 10.4108/eetismla.10805

*Corresponding author. Email: prasannabharathi.s@vit.ac.in

1. Introduction

Eye diseases represent a major global health issue, impacting millions across the world and leading to vision loss or blindness. In 2020, approximately 1 billion people worldwide were affected by visual impairment that could have been prevented or treated, with nearly 90% of them located in lower-income and middle-income nations [1]. There are numerous types of eye diseases, each presenting unique challenges, such as pterygium, glaucoma, and myopia. Symptoms of these conditions can vary but often include blurred vision, floaters, visual distortion, or a gradual loss of vision, significantly impacting daily life and overall well-being. Additionally, in 2020, approximately 1.07 million individuals experienced blindness, and around 3.28 million suffered from visual impairment, specifically as a result of Diabetic Retinopathy (DR) [2].

The after-effects of eye diseases can be profound, affecting an individual's quality of life, independence, and ability to work or perform everyday tasks. Vision impairment can lead to a loss of productivity, increased risk of injury, and even emotional and psychological distress [3]. Eye diseases are typically diagnosed through clinical examinations and imaging techniques like fundoscopy, optical coherence tomography, and visual field tests. However, traditional diagnostic methods have limitations, such as dependency on the examiner's expertise, potential delays in detecting subtle changes, and limited access to specialised equipment in some areas [4]. These challenges highlight the need for automated solutions that can offer more efficient, accurate, and accessible eye disease detection and management.

Automated diagnostic systems have significantly improved the field of medical diagnosis by providing faster, more accurate, and accessible methods for detecting and diagnosing eye diseases. These systems rely on advanced algorithms and Machine Learning (ML) models to process medical images, detect patterns, and generate predictions with limited need for human involvement. By automating the diagnostic process, these systems can reduce the workload on medical professionals and help ensure that patients receive timely care [5]. ML, particularly, has shown great promise in improving diagnostic accuracy and efficiency. ML models can learn from vast amounts of data, improving their ability to detect even subtle changes in medical images, which may otherwise go unnoticed by the human eye [6]. The benefits of ML include faster diagnosis, enhanced accuracy, and improved patient outcomes [7].

However, ML has a few limitations, particularly when it comes to medical image diagnosis. ML models often require large, labelled datasets to achieve high accuracy, and they can be prone to errors if the data used for training is biased or incomplete [8]. Additionally, traditional ML models may struggle to capture the complex, hierarchical features found in medical images, limiting their effectiveness in certain applications. Deep learning (DL) overcomes many of these limitations by employing multi-layered neural networks that autonomously extract features from raw data and learn complex patterns and representations [9]. In medical image diagnosis, DL excels at recognising intricate details within images, such as fine textures or subtle abnormalities, which are often critical in detecting conditions like DR or glaucoma. Unlike traditional ML methods, DL does not require extensive feature engineering, as it can learn directly from the data, making it particularly effective for analysing large and complex datasets [10]. In light of these capabilities, we are proposing a DL framework designed to alleviate the stress faced by the medical industry in diagnosing eye diseases. The proposed work makes the following key contributions:

- A dual-track architecture combining the strengths of DenseNet169 and TNT instead of a single path CNN-based or Vision Transformer architecture. DenseNet169 effectively captures fine-grained local features, while TNT models global context and long-range dependencies. The fused and enhanced feature maps overcome the limitations of traditional single-track architectures, providing a more comprehensive representation of eye fundus images.
- Lightweight attention-based refinement of the combined feature maps using ECA. This mechanism selectively emphasizes the most informative channels, helping the model focus on critical regions of the image while reducing redundancy and computational overhead.
- Explainable AI integration using SHAP to visualize and interpret the decision-making process of the network. This enhances transparency, reduces the "black box" nature of DL, and supports clinical trust and adoption.

2. Literature Survey

DL-based techniques have been extensively utilized for classifying eye diseases using fundus images. This section offers an overview of the various DL approaches employed in the detection of eye diseases from eye fundus images. Convolutional Neural Networks (CNNs) have played a pivotal role in transforming the field of image recognition and classification, enabling more accurate and efficient analysis of visual data [11]. CNNs are employed to automatically extract a hierarchical set of features, which can subsequently be used for classification, eliminating the need for manual feature extraction. CNNs have played an integral role in the detection and classification of various eye conditions like glaucoma, myopia, and DR using colour fundus images. Techniques such as Transfer Learning (TL) and Ensemble Learning (EL) have been frequently applied in image analysis of colour eye fundus images.

In one such attempt, Shamsan et al. developed an automated system for the classification of Colour Fundus Photography (CFP) images, leveraging hybrid techniques for feature extraction and fusion [12]. The features were extracted through a combination of DenseNet-121 and MobileNet networks, with Artificial Neural Network (ANN) used for classification. Additionally, Principal Component Analysis (PCA) was employed to reduce high-dimensional and redundant features. Ali et al. developed a method for classifying DR by employing ResNet-50 and InceptionV3 for feature extraction [13]. The extracted features were then concatenated using a custom CNN for final classification. The developed model showcased improved performance compared to various state-of-the-art networks and previously reported methods. Similarly, Al-Fahdawi et al. developed the Fundus-DeepNet, which used fundus images for detecting multiple ocular diseases [14]. The system is capable of identifying various conditions like age-related macular degeneration, cataracts, glaucoma, myopia, and hypertension. The SENet block was incorporated to enhance feature representation, while a Discriminative Restricted Boltzmann Machine (DRBM)-based block was utilised for advanced classification. Bernabé et al. designed a CNN-based model for the classification of glaucoma and DR, employing a K-Fold Cross-Validation approach for performance evaluation [15]. The proposed method outperformed several ML

techniques, including k-Nearest Neighbors (kNN), Decision Trees (DT), Naïve Bayes (NB), ANNs, Support Vector Machines (SVM), and Random Forests (RF), as reported in existing studies.

EL is a technique that improves predictive performance by combining the outputs of multiple networks [16]. Employing ensemble models aims to reduce prediction generalization error. Zhao et al. introduced a binary DR classification architecture named Bilinear Attention Net (BA-Net) [17]. The approach leverages a pre-trained ResNet-50, which is processed through a custom attention network comprising a three-layer CNN, called the RA-Net. Two identical RA-Net streams are trained concurrently to construct the bilinear model. The proposed architecture demonstrated superior performance compared to various existing methods and prior studies. Kim et al. proposed an ensemble model comprising two state-of-the-art CNNs, VGG19 and EfficientNet-B7, combined with an SVM classifier [18]. The network classified fundus images into four levels of visual acuity and demonstrated better performance when it was benchmarked against state-of-the-art CNN architectures. Qummar et al. designed a DR detection network to classify DR into 5 levels: normal, mild, moderate, severe, and Proliferative Diabetic Retinopathy (PDR) [19]. An ensemble of five deep CNN models, which included ResNet-50, InceptionV3, Xception, DenseNet-121, and DenseNet-169 was utilized to encode extracted features and improve classification performance across various stages of DR. Jiang et al. employed an ensemble model consisting of InceptionV3, ResNet-152, and Inception-ResNetV2 for DR classification [20]. The features extracted from the three CNNs were combined using the AdaBoost algorithm. The integrated model provided a classified score along with the weighted Class Activation Map (CAM).

TL is a technique that involves leveraging weights learnt from one task and dataset to reduce the learning costs associated with the target task [21]. Furthermore, TL often results in enhanced performance, particularly when the tasks are related, by allowing the model to fine-tune its knowledge to the particularities of the new target task. Chen et al. applied TL by utilizing a pre-trained InceptionV3 model [22]. Additionally, a new pre-processing algorithm was implemented to enhance the quality and consistency of the input retinal images. Zannah et al. proposed an automated diagnosis system that combines TL-based CNNs with an ML classifier for accurate eye disease classification [23]. EfficientNet, VGG16, VGG19, DenseNet, and ResNet50 were considered, with EfficientNet being selected as the final CNN. The system also employed an ensemble of ML classifiers, including Logistic Regression (LR), RF, Gaussian Naive Bayes (GNB), DT, k-NN, SVM, AdaBoost, and XGBoost. Features extracted from the EfficientNet model were reduced using PCA and passed through the ML ensemble classifier, with a Bayesian Optimization algorithm used for hyperparameter tuning. Wan et al. employed TL and hyperparameter tuning for four state-of-the-art models, which included AlexNet, VGGNet, GoogleNet, and ResNet, to classify DR images [24]. Aranha et al. developed a TL approach for diagnosing eye-related diseases using low-quality fundus images [25]. An ensemble of state-of-the-art CNNs was employed, with VGG16 serving as the base model. Rieck et al. proposed a hybrid CNN–transformer architecture for multi-class eye disease classification from colour fundus images [26]. The model integrates EfficientNet-B4 for local feature extraction with Swin Transformer modules to capture global contextual information and was evaluated on a real-world dataset comprising nine eye diseases and a healthy class using stratified five-fold cross-validation and achieved promising results. These works showcase the potential of DL to automate and fast-track eye disease diagnosis.

However, despite the effectiveness of these methods, several limitations remain in the existing literature. Most prior works rely primarily on transfer learning or ensemble learning using generic, pre-trained CNN models that are not specifically tailored to capture the integral and disease-specific features present in retinal images. Furthermore, many studies adopt single-track CNN architectures that emphasize only local feature extraction, thereby overlooking global contextual information that is critical for accurate fundus image interpretation. Existing approaches also lack the integration of lightweight attention mechanisms that could guide the model to focus on the most informative retinal regions. These limitations highlight the need for a hybrid CNN–Transformer architecture capable of jointly modelling both local and global features for improved classification performance.

3. Methodology

This section describes the eye disease dataset utilised in the proposed study in detail. In the subsequent section, the architecture of the developed system and its components are discussed briefly.

3.1. Dataset Description

This research utilizes a dataset comprising 5,335 images of both healthy and diseased eyes, sourced from Anwara Hamida Eye Hospital and BNS Zahurul Haque Eye Hospital located in Faridpur, Bangladesh [27]. The images were captured using the TRC-50DX and TL-211 digital fundus cameras, which were connected to the D7500 and D90 Nikon DSLR cameras with the help of hospital authorities between July 2023 and February 2024. The dataset includes images with different dimensions, stored in JPG format. The dataset is divided into 10 categories: Disc Edema, Glaucoma, Healthy, Macular Scar, Diabetic Retinopathy, Retinitis Pigmentosa, Retinal Detachment, Central Serous Chorioretinopathy, Pterygium, and Myopia. The classification was assisted by a domain expert from a healthcare institute to ensure accuracy. Prior to model training, several preprocessing steps were applied to standardize and enhance the images. All images were resized to 224×224 pixels to maintain uniform input dimensions.

Pixel intensity values were normalized to improve model convergence. To improve generalizability and mitigate overfitting, augmentation techniques were applied, including horizontal and vertical flipping, rotation (±15°), zooming (up to 10%), and width and height shifting by 10%. This increased the dataset size to 16,242 images. Table 1 summarizes the dataset distribution after being divided into training, validation, and test sets with a ratio of 70%, 20%, and 10%, and was done randomly to ensure unbiased representation. A few sample images are shown in Figure 1.

Table 1. Dataset Split of the Eye Color Fundus Images Dataset

Class	Split		
	Train	Valid	Test
Disc Edema	533	152	77
Glaucoma	2015	576	289
Healthy	1873	535	268
Macular Scar	1355	387	195
Diabetic Retinopathy	2410	688	346
Retinitis Pigmentosa	583	166	85
Retinal Detachment	525	150	75
Central Serous Chorioretinopathy	424	121	61
Pterygium	71	20	11
Myopia	1575	450	226

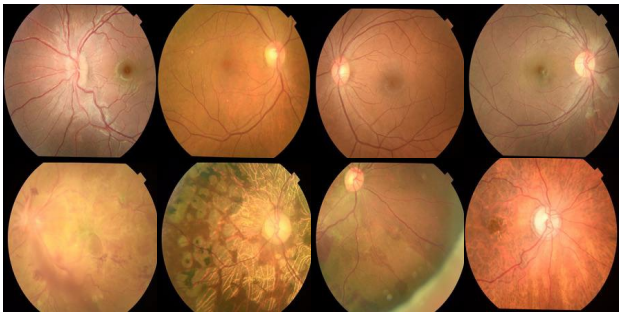


Figure 1. Sample Images from the Eye Color Fundus Images Dataset.

3.2. EyeFusionNet

The proposed EyeFusionNet combines the strengths of DenseNet169 and the Transformer-iN-Transformer (TNT) model to develop a powerful dual-track architecture for the classification of eye diseases using colour fundus images. In this framework, the classification layers of both DenseNet169 and TNT are removed, and only their respective feature extraction backbones are retained. The two networks operate as parallel tracks to extract complementary representations from the same input image. DenseNet169, with its densely connected convolutional blocks, effectively captures hierarchical patterns, texture information, and fine-grained spatial details. Its dense connectivity promotes feature reuse and stable gradient flow, making it well-suited for extracting localized disease-specific features. In contrast, TNT leverages self-attention mechanisms to model global context and long-range dependencies across the image, overcoming the locality limitations of convolutional architectures. The output feature maps from the final feature extraction layers of both networks are then fused through feature-level concatenation. Specifically, TNT produces a 384-dimensional global feature representation, while DenseNet169 generates a 1664-dimensional deep convolutional feature vector. These outputs are concatenated along the channel dimension, resulting in a unified 2048-dimensional multi-scale, feature-rich representation that integrates both local structural details and global contextual information. This fusion strategy enables EyeFusionNet to overcome the limitations of traditional single-track architectures by jointly leveraging CNN-based spatial representations and transformer-based contextual modelling.

To further refine the fused features and reduce redundancy, an Efficient Channel Attention (ECA) module is incorporated. The concatenated feature maps are passed through the ECA layer, which adaptively emphasizes the most informative channels while suppressing less relevant ones with minimal computational overhead. The refined

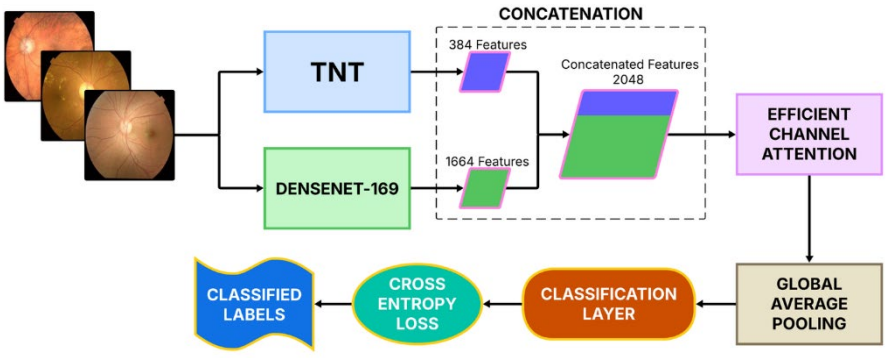


Figure 2. Visual Representation of the dual-track architecture of the proposed EyeFusionNet.

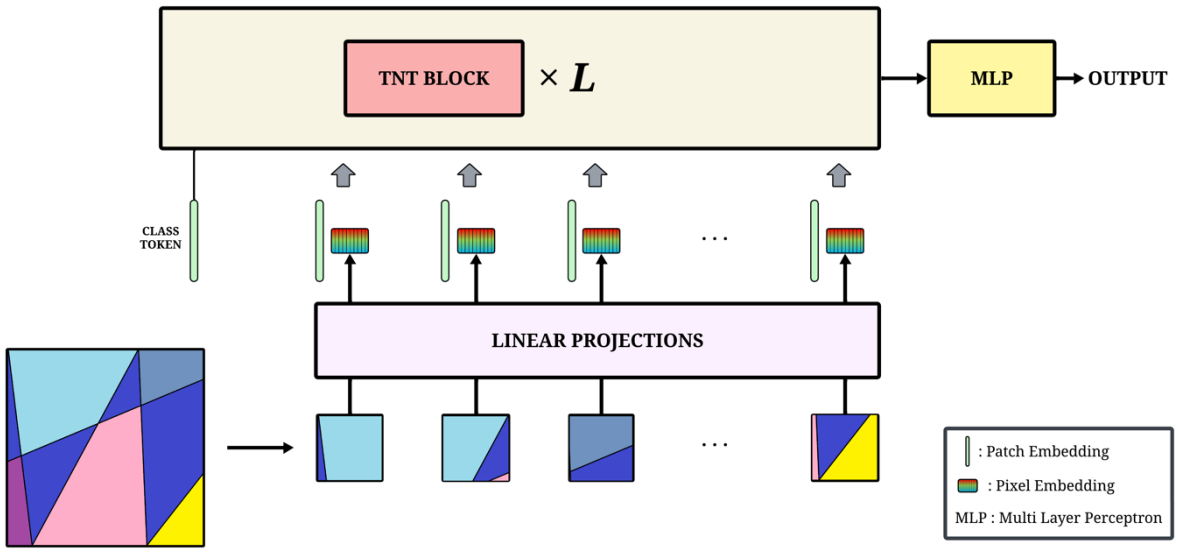


Figure 3. Structural Overview of the TNT Network Path in the Proposed EyeFusionNet

features are then subjected to Global Average Pooling (GAP) to reduce dimensionality while preserving the most discriminative information. Finally, focal loss is employed as the optimization objective to handle class imbalance and enhance classification performance. The proposed EyeFusionNet outputs the final eye disease classification along with explainable visualizations generated using XAI techniques, supporting clinical interpretability and validation. The overall architecture of EyeFusionNet is illustrated in Figure 2.

DenseNet-169

This section outlines the architecture and role of DenseNet169 in the EyeFusionNet for extracting features relevant to eye diseases like DR, glaucoma, optic disc edema, and retinitis pigmentosa from fundus images. DenseNet169 is a deep CNN based on densely connected blocks [28]. Unlike traditional architectures, where each layer feeds its output only to the next layer, DenseNet incorporates dense connectivity, enabling each layer to utilize the outputs of all preceding layers as inputs while also sharing its feature maps with all following layers. This structure promotes the reuse of extracted features, reducing redundancy and enhancing the learning of critical patterns in the data. For eye fundus images, DenseNet169 excels in capturing fine-grained spatial features and patterns essential for identifying disease-specific markers. For instance, it can effectively detect microaneurysms and haemorrhages in DR, optic nerve swelling in glaucoma, and pigmentation changes in retinitis pigmentosa. Its ability to learn from both local and mid-range dependencies makes it an invaluable component of the proposed network. DenseNet complements the TNT by providing detailed, localized feature maps. While DenseNet169 focuses on spatial patterns and texture details, TNT excels at capturing global context and long-range relationships. Together, they offer a holistic understanding of fundus images, where DenseNet enriches the network with localized structural information, and TNT adds semantic depth by modelling dependencies across the entire fundus image. This synergy ensures that EyeFusionNet effectively leverages the strengths of both approaches for superior classification performance.

Transformer-iN-Transformer

This section discusses the architecture of the TNT model and its significance in EyeFusionNet. TNT is a self-attention-based neural architecture designed to extract both local and global features from images, addressing limitations in traditional visual transformers that focus solely on patch-level information [29]. By viewing an image as a sequence of patches and further subdividing each patch into smaller sub-patches, TNT enhances feature representation at multiple granularities. In the proposed configuration, the input image is first divided into non-overlapping patches of size 16×16 pixels, forming the “outer” patch tokens, with an outer transformer dimension of 384 used for patch-level embeddings that capture global context. Each outer patch is further subdivided into smaller sub-patches with an inner transformer dimension of 24, encoding fine-grained local details. TNT employs multi-head attention, with 6 heads in the outer block to model interactions between patch embeddings and 4 heads in the inner block to model interactions among sub-patch embeddings. This allows the model to focus on multiple aspects of the image simultaneously at each granularity. The architecture stacks 12 TNT blocks sequentially, each containing an inner and an outer transformer, enabling hierarchical feature extraction. The inner block processes relationships among visual words

within a patch, capturing subtle details such as microaneurysms or pigment changes, while the outer block models interactions between visual sentences across the entire image to capture broader abnormalities like vascular distortions or optic disc swelling. For analyzing eye fundus images, TNT's dual-level attention mechanism proves particularly effective in identifying intricate patterns. In EyeFusionNet, TNT complements DenseNet169 by providing a global perspective that aligns with the localized spatial features extracted by the convolutional network. This complementary relationship ensures that both detailed local information and broader contextual patterns are leveraged, resulting in a more holistic and robust feature representation for the classification of eye diseases. The structure of the TNT model and TNT block is illustrated in Figures 3 and 4.

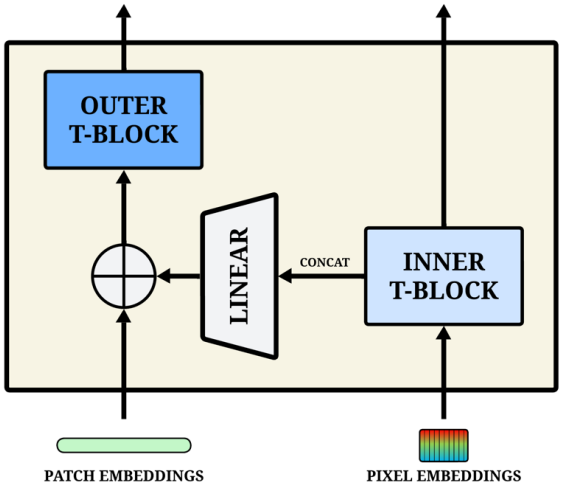


Figure 4. Schematic Overview of the TNT Block from the TNT track of the EyeFusionNet.

Efficient Channel Attention

The ECA module plays an integral role in refining the fused feature maps generated by EyeFusionNet's dual-track architecture [30]. By selectively emphasizing the most informative channels, ECA enhances the network's ability to focus on critical features that aid in detecting eye diseases such as DR, glaucoma, and optic disc edema. Unlike traditional channel attention mechanisms, which rely on computationally expensive fully connected layers, ECA employs a lightweight 1D convolution with an adaptively determined kernel size to achieve efficient and precise attention. The process begins with GAP to compute a descriptor vector Z_c for each channel c, summarizing the spatial features across the feature map. This descriptor is calculated using Eq. (1):

$$Z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{c,i,j} \quad (1)$$

Where H and W denote the height and width, and $X_{c,i,j}$ denotes the value at position i and j for channel c. The descriptor vector Z_c encodes the global information for each channel. Instead of using fully connected layers for channel weighting, ECA employs a 1D convolution operation with kernel size k, which is determined dynamically based on the number of channels c, using a mapping function φ as illustrated in Eq. (2)

$$k = \varphi(C) \quad (2)$$

This approach ensures that the kernel size adjusts to the feature map's dimensionality, balancing expressiveness and computational efficiency. The convolution output is then passed through a sigmoid activation function σ to produce the channel attention weights α as mentioned in Eq. (3):

$$\alpha = \sigma(Conv1D(Z_c, k)) \quad (3)$$

These attention weights are used to refine the original feature maps by scaling each channel with its corresponding weight, resulting in the updated feature map X'_c as given in Eq. (4):

$$X'_c = \alpha_c . X_c \quad (4)$$

By applying the ECA module, EyeFusionNet prioritizes the most relevant channels, allowing the model to focus on informative regions in fundus images while suppressing redundant information. This targeted refinement enhances the diagnostic capability of the network, improving its overall accuracy and robustness. The lightweight nature of ECA ensures minimal computational overhead, making it an efficient and effective choice for channel attention.

4. Results

This section provides a detailed summary of the hyperparameter optimization, model evaluation, SHAP visualisation, and a comparative study of the EyeFusionNet against existing CNN and transformer networks.

4.1. Hyper-parameter tuning

The performance of DL models can be significantly improved by optimizing hyperparameters, allowing for better generalization and learning capacity. In this work, several key hyperparameters were fine-tuned, including the learning rate, weight decay, batch size, gradient optimizer, loss function, and the learning rate scheduler step. A range of values was tested for each parameter: learning rates and weight decay values from 1 to 0.000001, and batch sizes between 16 and 64. Both the ADAM and SGD optimizers were evaluated for their efficiency in model convergence, along with cross-entropy and focal loss. The optimal configuration was achieved when the learning rate was set to 0.0001, along with a weight decay of 0.0001, and a batch size of 32. The Adam optimizer was found to outperform SGD, and focal loss was

used as the loss function. The learning rate scheduler used a step size of 25 to ensure stable learning. The proposed EyeFusionNet was trained for 100 epochs using cloud-based computational resources on Google Cloud, equipped with an NVIDIA T4 GPU with 16 GB GPU memory and 12 GB CPU RAM. All experiments were implemented using the PyTorch DL framework (version 2.6.0). This configuration enabled efficient training and stable convergence of the proposed model.

4.2. Evaluation of the EyeFusionNet

This section examines the performance of the proposed dual-track network for eye disease classification. The dual-track model, which combines DenseNet-169 and TNT, was enhanced with ECA to refine the feature maps extracted by both tracks. ECA helps focus on the most informative channels, reducing redundancy and improving the ability of the model to capture key features relevant to the classification of eye diseases. During testing, the model recorded an accuracy of 89%, a precision of 88%, a sensitivity (recall) of 90%, and an F1-score of 89%. These results indicate that the dual track approach with ECA enhanced the ability of the model to accurately classify eye fundus images and provided a more robust solution for detecting eye diseases. The performance of various sub-components used in the EyeFusionNet is summarized in Table 2.

Table 2. Summary of the Components of EyeFusionNet.

Model	Performance Metrics			
	Accuracy	Precision	Sensitivity	F1
DenseNet-169 Track	83%	86%	87%	86%
TNT Track	80%	84%	82%	83%
Dual Track	87%	88%	89%	88%
Dual Track with ECA	**89%**	**88%**	**90%**	**89%**

Table 3. 5-Fold Cross-Validation Results

Fold	Performance Metrics			
	Accuracy	Precision	Sensitivity	F1
1	89%	90%	89%	90%
2	88%	89%	90%	90%
3	89%	89%	88%	89%
4	89%	88%	90%	89%
5	90%	89%	90%	90%
Mean ± Std	**89.0 ± 0.7**	**89.0 ± 0.7**	**89.4 ± 0.9**	**89.6 ± 0.5**

The precision-recall and ROC curves are provided in Figures 5 and 6, respectively. To offer a more comprehensive overview of the model's performance, the t-SNE, prediction confidence, and inference time plots have been included in Figures 7, 8, and 9. The t-SNE plot visualizes the high-dimensional feature representations learned by the model in a two-dimensional space. This helps in understanding how well the model clusters different eye fundus image classes such as normal, diabetic retinopathy, or glaucoma based on their underlying visual patterns. Clear separation in t-SNE space reflects the model's ability to learn discriminative features crucial for accurate diagnosis. The class-wise precision, recall, and F1-score are reported in Table 4, providing a detailed quantitative evaluation of the model's performance across individual retinal disease categories. Furthermore, the confusion matrix shown in Figure 10 highlights class-level misclassifications, offering insight into prediction errors among visually and clinically similar conditions.

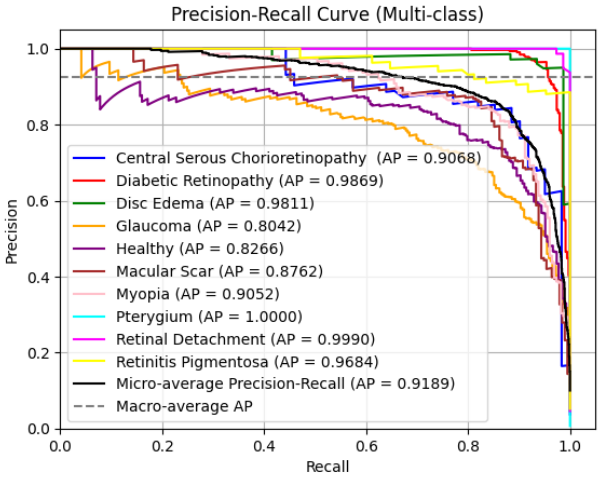


Figure 5. Precision-Recall Curves for the EyeFusionNet

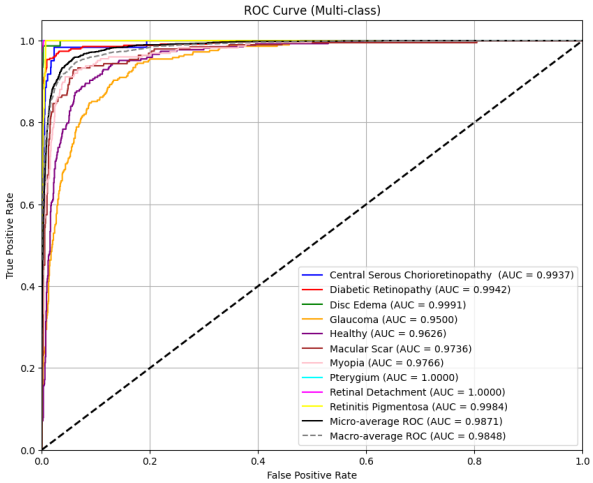


Figure 6. ROC Curves for the Proposed System

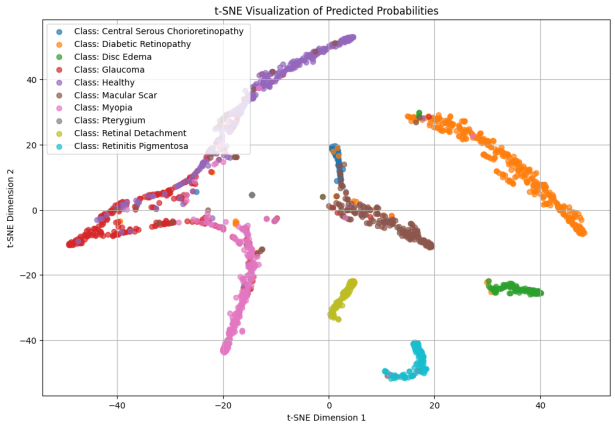


Figure 7. t-SNE plot for EyeFusionNet

Figure 10. Confusion Matrix for EyeFusionNet

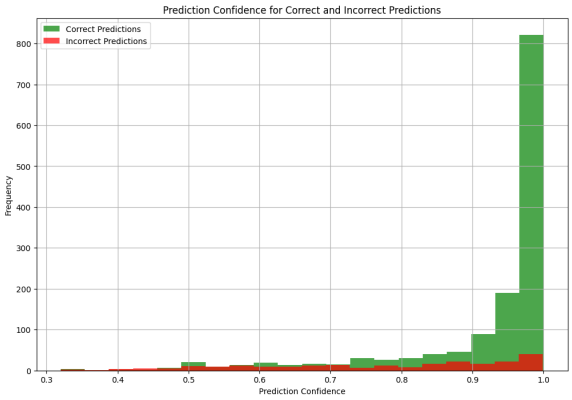


Figure 8. Prediction Confidence Distribution for EyeFusionNet

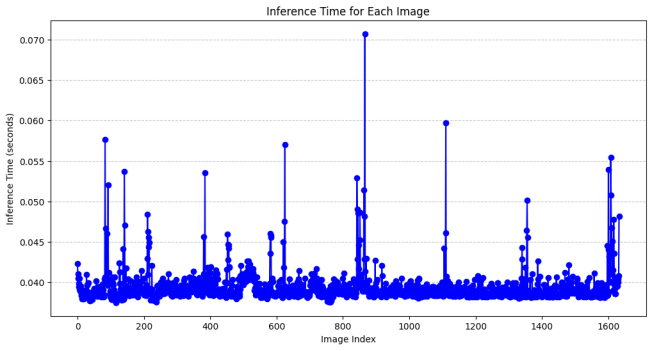


Figure 9. Distribution of Inference Time During Testing

Table 4. Summary of the Components of EyeFusionNet

Class	Performance Metrics			
	Accuracy	Precision	Sensitivity	F1
Central Serous Chorio-Retinopathy	89	90	89	89
Diabetic Retinopathy	91	94	91	92
Disc Edema	92	86	92	89
Glaucoma	86	89	86	88
Healthy	85	83	85	84
Macular Scar	87	82	87	85
Myopia	88	84	88	86
Pterygium	100	100	100	100
Retinal Detachment	93	100	93	97
Retinitis Pigmentosa	92	100	92	96

To ensure unbiased and generalizable performance, 5-fold cross-validation was conducted. This method splits the dataset into five subsets, using four for training and one for validation in each iteration, with all subsets serving as validation once. The proposed system operates by analyzing the input eye fundus image provided by the clinician and returns the final classification outcome along with XAI visualizations to enhance interpretability. As shown in Table 3, the overall results support the effectiveness of the proposed model. Furthermore, Table 4 presents SHAP-based XAI visualizations for selected retinal disease classes, including Disc Edema, Macular Scar, Diabetic Retinopathy, and Retinitis Pigmentosa. SHAP assigns importance scores to image regions based on cooperative game theory, enabling the decomposition of the model's prediction into localized, interpretable contributions. These visualizations highlight the specific regions of the fundus image that most strongly influenced the model's classification decision [31].

For Disc Edema, the SHAP maps show concentrated activations around the optic disc region, which is clinically consistent with optic nerve head swelling and blurred disc margins observed in affected patients. In the case of Macular Scar, high-attribution regions are localized around the macula, capturing irregular texture patterns and scarring that are characteristic of macular damage. For Diabetic Retinopathy, SHAP visualizations emphasize vascular structures and lesion-prone areas, including regions corresponding to microaneurysms, hemorrhages, and exudates, reflecting the model's reliance on pathological vascular cues. Similarly, for Retinitis Pigmentosa, the highlighted regions are predominantly distributed across peripheral retinal areas, aligning with the progressive peripheral degeneration that defines the disease.

The strong correspondence between the SHAP-highlighted regions and established ophthalmic biomarkers demonstrates that the model's predictions are grounded in clinically relevant visual evidence rather than spurious correlations. This interpretability significantly reduces the black-box nature of the deep learning framework and enhances its reliability and suitability for deployment in real-world clinical screening and diagnostic workflows.

4.3. Performance Comparison with Top DL Networks

Within this section, the performance of the EyeFusionNet is assessed on the publicly available Eye Fundus dataset. The dataset was released in 2024, hence there are no existing works available for comparison. Therefore, this is the first work leveraging this dataset, and the proposed network has been compared with state-of-the-art CNN and transformer networks. For a fair comparison, all models were fine-tuned and trained on the dataset. Among the CNN architectures, DenseNet-169 achieved the highest accuracy of 83%, making it the preferred choice for the first feature extraction track in EyeFusionNet. Similarly, transformer-based models such as Tokens-To-Token Transformer (T2T), Conv-Attentional Image Transformer (CaiT), Swin Transformer (SwinT), Data-Efficient Image Transformer (DeiT), Convolution-Enhanced Image Transformer (CeiT), and Transformer-in-Transformer (TNT) were evaluated. Among these, TNT attained the highest accuracy of 80%. EyeFusionNet surpassed all compared CNN and transformer models, achieving an accuracy of 89%, along with improved precision, sensitivity (recall), and F1-score. The combined performance metrics are summarized in Table 5, highlighting the robustness and effectiveness of the proposed model in eye disease classification.

5. Limitations & Future-Work

The proposed model for eye disease classification has some limitations that highlight areas for future improvement and research:

(i) While the model has shown good results, it has yet to be tested in real-world clinical settings. To assess its effectiveness and adaptability in healthcare environments, clinical trials or pilot studies involving healthcare professionals would be crucial. Real-world feedback could help fine-tune the model for practical deployment and improve its overall performance.

(ii) The model currently uses only eye fundus images for classification. Incorporating additional data types, such as patient demographics, medical history, or genetic factors, could provide a more holistic and accurate diagnostic tool. Future research should explore the integration of multi-modal data, which could enhance the robustness of the model and generalize better across different demographic groups.

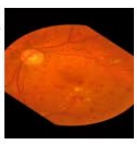
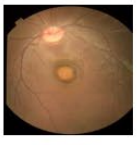
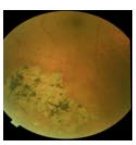
(iii) As quantum computing technology advances, there is potential for integrating quantum ML techniques to enhance the model's computational efficiency. This could be especially useful in handling large datasets or complex feature sets, potentially leading to faster training times and improved scalability. Future work could explore how quantum ML can be leveraged to optimize model performance.

(iv) The dataset used in this study exhibits class imbalance, with certain eye diseases being underrepresented, which may affect the model's ability to generalize uniformly across all classes. Future work can explore GAN-based data augmentation techniques to generate synthetic samples for minority classes and mitigate data imbalance. In addition, external validation on independent and multi-center datasets is required to assess the robustness and generalizability of the proposed model under diverse imaging conditions.

Table 5. Comparison of the EyeFusionNet against DL Networks

Model	Performance Metrics			
	Accuracy	**Precision**	**Sensitivity**	**F1**
EfficientNet-B0	56%	60%	57%	58%
MobileNet	60%	63%	62%	62%
ShuffleNet	68%	71%	70%	71%
T2T	70%	72%	71%	71%
CaiT	76%	81%	77%	78%
EfficientNet-B0	56%	60%	57%	58%
MobileNet	60%	63%	62%	62%
SqueezeNet	79%	83%	77%	77%
CeiT	79%	83%	81%	82%
ResNet-18	79%	83%	81%	82%
ResNet-50	79%	82%	82%	82%
ResNet-34	79%	83%	82%	82%
XceptionNet	79%	84%	82%	83%
TNT	80%	84%	82%	83%
AlexNet	81%	84%	83%	83%
DenseNet-121	82%	85%	84%	84%
DenseNet-201	83%	86%	85%	85%
DenseNet-169	83%	86%	87%	86%
EyeFusionNet	**89%**	**88%**	**90%**	**89%**

Table 6. XAI Visualization using SHAP

Class	XAI Visualisation			
	Disc Edema	Macular Scar	Diabetic Retinopathy	Retinitis Pigmentosa
Central Serous Chorioretinopathy				
Diabetic Retinopathy				
Disc Edema				
Glaucoma				
Healthy				
Macular Scar				
Myopia				
Pterygium				
Retinal Detachment				
Retinitis Pigmentosa				

6. Conclusion

Eye disease diagnosis, particularly using colour fundus images, is integral for the timely detection of conditions such as DR, glaucoma, and macular degeneration. However, manual analysis by specialists can be labour-intensive and susceptible to inconsistencies, particularly in resource-limited settings. To address these challenges, this work introduces EyeFusionNet, an automated DL-based solution for the classification of eye diseases using fundus images. EyeFusionNet integrates the strengths of CNNs and transformer models in a dual-fusion framework, enhancing the extraction of high and low-level features. The proposed system integrates DenseNet169, renowned for its capability to extract detailed hierarchical features, with TNT, which captures long-range relationships and high-level context within the images. These two models operate in parallel, extracting complementary features, which are then fused to create a more comprehensive representation of the image.

The dataset used in this work is the first of its kind for eye disease classification, focusing on colour eye fundus images across 10 disease classes. After evaluating state-of-the-art CNN and transformer models, the best-performing models were selected to form the EyeFusionNet architecture. The model's feature maps are refined using ECA, which enhances the most informative features while reducing redundancy. GAP is then applied to minimize dimensionality, followed by optimization with focal loss. The EyeFusionNet obtained an accuracy of 89%, outperforming several state-of-the-art CNN and transformer networks in eye disease classification. While the results demonstrate the effectiveness of the proposed dual-track fusion strategy, further validation on larger and more diverse datasets, as well as prospective clinical studies, is necessary before real-world deployment. Nonetheless, this work establishes a strong methodological foundation for future research in automated eye disease classification and highlights the potential of hybrid DL architectures in advancing medical image analysis.

References

[1] Forrest SL, Mercado CL, Engmann CM, Stacey AW, Hariharan L, Khan S, et al. Does the current Global Health Agenda Lack Vision? Global Health: Science and Practice 2023;11. https://doi.org/10.9745/GHSP-D-22-00091.

[2] Curran K, Peto T, Jonas JB, Friedman D, Kim JE, et al. Global estimates on the number of people blind or visually impaired by diabetic retinopathy: a meta-analysis from 2000 to 2020. Eye 2024;38:2047–57. https://doi.org/10.1038/s41433-024-03101-5.

[3] Gupta P, Fenwick EK, Man REK, Gan ATL, Sabanayagam C, Quek D, et al. Different impact of early and late stages irreversible eye diseases on vision-specific quality of life domains. Sci Rep 2022;12. https://doi.org/10.1038/s41598-022-12425-9.

[4] R.M. B, Vardhan KB, Nidhish M, Kiran C. S, Nahid Shameem D, Sai Charan V. Eye Disease Detection Using Deep Learning Models with Transfer Learning Techniques. ICST Transactions on Scalable Information Systems 2024;11. https://doi.org/10.4108/eetsis.5971.

[5] Kaur I, Ali A. A Complete Study on Machine Learning Algorithms for Medical Data Analysis. Fog Computing for Intelligent Cloud IoT Systems 2024:137–72. https://doi.org/10.1002/9781394175345.ch7.

[6] Suresh NV, Sridhar J, Selvakumar A, Catherine S. Machine Learning Applications in Healthcare. Advances in Healthcare Information Systems and Administration 2024:1–9. https://doi.org/10.4018/979-8-3693-7452-8.ch001.

[7] ALI L, Gun TC, Alhasan W. Comparative Analysis of Machine Learning Algorithms in Enhancing Healthcare Outcomes. EMSJ 2024;8:606–18. https://doi.org/10.59573/emsj.8(3).2024.38.

[8] Akter S, Dwivedi YK, Sajib S, Biswas K, Bandara RJ, Michael K. Algorithmic bias in machine learning-based marketing models. Journal of Business Research 2022;144:201–16. https://doi.org/10.1016/j.jbusres.2022.01.083.

[9] Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 2021;8. https://doi.org/10.1186/s40537-021-00444-8.

[10] Sarker IH. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN COMPUT SCI 2021;2. https://doi.org/10.1007/s42979-021-00592-x.

[11] Taye MM. Theoretical understanding of Convolutional Neural Network: Concepts, architectures, applications, Future Directions. Computation 2023;11:52. doi:10.3390/computation11030052.

[12] Shamsan A, Senan EM, Shatnawi HS. Automatic classification of colour fundus images for prediction eye disease types based on hybrid features. Diagnostics 2023;13:1706. https://doi.org/10.3390/diagnostics13101706.

[13] Ali G, Dastgir A, Iqbal MW, Anwar M, Faheem M. A hybrid convolutional neural network model for automatic diabetic retinopathy classification from fundus images. IEEE Journal of Translational Engineering in Health and Medicine 2023;11:341–50. https://doi.org/10.1109/jtehm.2023.3282104.

[14] Al-Fahdawi S, Al-Waisy AS, Zeebaree DQ, Qahwaji R, Natiq H, Mohammed MA, et al. Fundus-DeepNet: Multi-label Deep Learning Classification system for enhanced detection of multiple ocular diseases through data fusion of fundus images. Information Fusion 2024;102:102059. https://doi.org/10.1016/j.inffus.2023.102059.

[15] Bernabe O, Acevedo E, Acevedo A, Carreno R, Gomez S. Classification of eye diseases in fundus images. IEEE Access 2021;9:101267–76. https://doi.org/10.1109/access.2021.3094649.

[16] Mahajan P, Uddin S, Hajati F, Moni MA. Ensemble learning for disease prediction: A Review. Healthcare 2023;11:1808. https://doi.org/10.3390/healthcare11121808.

[17] Zhao Z, Zhang K, Hao X, Tian J, Heng Chua MC, Chen L, et al. Bira-net: Bilinear attention net for diabetic retinopathy grading. 2019 IEEE International Conference on Image Processing (ICIP) 2019:1385–9. https://doi.org/10.1109/icip.2019.8803074.

[18] Kim JH, Jo E, Ryu S, Nam S, Song S, Han YS, et al. A deep learning ensemble method to visual acuity measurement using fundus images. Applied Sciences 2022;12:3190. https://doi.org/10.3390/app12063190.

[19] Qummar S, Khan FG, Shah S, Khan A, Shamshirband S, Rehman ZU, et al. A deep learning ensemble approach for diabetic retinopathy detection. IEEE Access 2019;7:150530–9. https://doi.org/10.1109/access.2019.2947484.
[20] Jiang H, Yang K, Gao M, Zhang D, Ma H, Qian W. An interpretable ensemble deep learning model for diabetic retinopathy disease classification. 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) 2019. https://doi.org/10.1109/embc.2019.8857160.
[21] Iman M, Arabnia HR, Rasheed K. A review of deep transfer learning and recent advancements. Technologies 2023;11:40. https://doi.org/10.3390/technologies11020040.
[22] Chen H, Zeng X, Luo Y, Ye W. Detection of diabetic retinopathy using Deep Neural Network. 2018 IEEE 23rd International Conference on Digital Signal Processing (DSP) 2018. https://doi.org/10.1109/icdsp.2018.8631882.
[23] Zannah TB, Abdulla-Hil-Kafi Md, Sheakh MdA, Hasan MdZ, Shuva TF, Bhuiyan T, et al. Bayesian optimized machine learning model for automated eye disease classification from Fundus Images. Computation 2024;12:190. https://doi.org/10.3390/computation12090190.
[24] Wan S, Liang Y, Zhang Y. Deep convolutional neural networks for diabetic retinopathy detection by image classification. Computers & Electrical Engineering 2018;72:274–82. https://doi.org/10.1016/j.compeleceng.2018.07.042.
[25] Aranha GD, Fernandes RA, Morales PH. Deep transfer learning strategy to diagnose eye-related conditions and diseases: An approach based on low-quality fundus images. IEEE Access 2023;11:37403–11. https://doi.org/10.1109/access.2023.3263493.
[26] Rieck C, Mai C, Eisentraut L, Buettner R. A Novel Transformer-CNN Hybrid Deep Learning Architecture for Robust Broad-Coverage Diagnosis of Eye Diseases on Color Fundus Images. IEEE Access 2025;13:156285–300. https://doi.org/10.1109/access.2025.3606334.
[27] Sharmin S, Rashid MR, Khatun T, Hasan MZ, Uddin MS, Marzia. A dataset of color fundus images for the detection and classification of eye diseases. Data in Brief 2024;57:110979. https://doi.org/10.1016/j.dib.2024.110979.
[28] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely Connected Convolutional Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017. https://doi.org/10.1109/cvpr.2017.243.
[29] Han K, Xiao A, Wu E, Guo J, Xu C, Wang Y. Transformer in Transformer 2021. https://doi.org/10.48550/ARXIV.2103.00112.
[30] Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2020:11531–9. https://doi.org/10.1109/cvpr42600.2020.01155.
[31] Lundberg S, Lee S-I. A Unified Approach to Interpreting Model Predictions 2017. https://doi.org/10.48550/ARXIV.1705.07874.