



Diabetic Retinopathy Detection with Uncertainty scores: A Combined Approach Using Transfer Learning and Ensemble Calibration

Preeti Verma^{a,b}, Dr. Sivasankar Elango^a, and Dr. Kunwar Singh^a

^a Department of CSE National Institute of Technology, Tiruchirappalli, Tamilnadu, India, 620015

^b Department of CSE(AI), KIET Group of Institution, Delhi- NCR, Ghaziabad, Uttar Pradesh, 201206

✉ Preeti.verma02@gmail.com, Sivasankarelango@gmail.com, Kunwar2081@gmail.com

KEYWORD

calibration;
deep models;
transfer
learning;
diabetic
retinopathy;
uncertainty

ABSTRACT

The rising prevalence of diabetes has made Diabetic Retinopathy (DR) a major cause of blindness, underscoring the necessity for a computer-aided diagnostic system that can support clinical diagnoses without requiring extensive human effort. Many researchers have turned to deep learning to create automated screening and diagnostic tools for DR. However, for such systems to be truly effective in clinical practice, they must provide highly accurate assessments and well-calibrated estimates of uncertainty. Unfortunately, deep neural networks often tend to be overconfident in their predictions and are not easily amenable to probabilistic approaches. In our study, we introduce a novel approach for evaluating diagnostic uncertainty in DR predictions by employing ensemble-based calibration techniques. What sets our approach apart from cutting-edge convolutional neural network models is our use of the EfficientNet architecture, which offers superior accuracy through transfer learning. We then apply a set of post-calibration techniques to transform the model's probabilistic output into a confidence level. To gauge the uncertainty of our forecasts, we compute the entropy of the calibrated confidence value. This approach greatly assists users in determining whether it is necessary to seek a second opinion. Our model achieves an impressive accuracy score of 96 %, and our ensemble technique exhibits a notable reduction in Expected Calibration Error (ECE), in addition to providing a reassuring uncertainty score.

1. Introduction

Diabetes presents a substantial and growing public health challenge. According to the International Diabetes Federation's projections, the number of people living with diabetes is expected to rise to 643 million by 2035 and a staggering 783 million by 2045 (Guariguata et al., 2014). Among the complications associated with diabetes, diabetic retinopathy (DR) stands out as a common and serious concern that can lead to significant vision loss and impairment within this population (Chen et al., 2023). The consequences of DR extend beyond the individual, with considerable economic implications for society as a whole (Kwee et al., 2022). Fortunately, timely intervention and treatment of DR can substantially reduce the risk of visual impairment. To achieve this, regular DR screening is crucial for early detection and appropriate care. However, the clinical diagnosis of DR through the examination of fundus images is a labour-intensive and time-consuming process. Given the increasing number of DR patients and a shortage of ophthalmologists, the likelihood of missed diagnoses looms large. This is where computer-aided diagnosis (CAD) steps in, effectively overcoming these limitations by reducing the labour and time required while maintaining a high degree of accuracy (Atwany et al., 2022; Ayhan et al., 2020).

In various domains, deep learning has made remarkable strides, from voice recognition and object identification to self-driving cars and medicine (Prasath, 2022; Pandey et al., 2022; Wen et al., 2022; Zaidi et al., 2022). In addition to applications in image recognition, speech processing, and healthcare, deep learning has recently made strides in solving mathematical problems, such as differential equations, symbolic computation, and function approximation (Bahramian et al., 2024; Dana Mazraeh & Parand, 2024; 2025a; 2025b; 2025c). In the medical field specifically, Convolutional Neural Networks (CNNs) have played a pivotal role, propelling groundbreaking advancements in areas such as retinal vascular classification, glaucoma screening, and the categorization of cancer subtypes (Mirzania et al., 2021; Tian et al., 2021; Uhm et al., 2021). Researchers have introduced various CNN algorithms, including VGGNet (Simonyan and Zisserman, 2014), ResNet (He et al., 2016), and GoogLeNet (Szegedy et al., 2015). Nevertheless, training these networks with limited medical data presents its own set of challenges.

This is where transfer learning comes into play as a strategic solution. Transfer learning allows for the transfer of knowledge gained from pre-trained input data, typically a vast collection of images, to a new dataset. This approach proves highly effective in enhancing the reliability and stability of neural networks when working with limited data (Tan et al., 2018). In our proposed method, we chose to leverage EfficientNet, as it demonstrated superior performance compared to other state-of-the-art CNN models when applied within the context of transfer learning.

Despite the significant strides made in the field of Deep Neural Networks (DNNs), their probabilistic outputs often fall short of providing reliable estimates of the true likelihood of correct predictions. Ideally, a classifier's probabilistic outputs should accurately reflect the genuine probability of its predictions being correct, essentially representing its accuracy. A well-calibrated classifier is one that readily interprets its outputs as confidence scores (Abdar et al., 2021; Ghoshal et al., 2022; Guo et al., 2017; Ye et al., 2022). In contrast, DNNs frequently suffer from miscalibration, displaying excessive confidence in their predictions. Consequently, they fail to produce well-calibrated and accurate uncertainty estimates for their judgments (Ayhan et al., 2020).

To rectify this miscalibration issue, we can employ various calibration methods, such as temperature scaling, Platt scaling, histogram binning, and the like. These methods serve to transform the model's probabilistic output into genuine confidence scores, thereby enhancing the reliability of the model's predictions.

In our innovative approach, we tackle the critical challenge of accurately gauging the confidence of a model's predictions and quantifying the associated uncertainty. Our solution involves an ensemble of calibration methods that collectively estimate the model's confidence score. By combining the results from various calibration methods, we enhance the precision of confidence score estimation, instilling greater trust in the model's output. This not only aids users in better comprehending the model's predictions but also equips them with a measure of uncertainty for each prediction. By providing a transparent understanding of the model's confidence and uncertainty, users can make more informed decisions about relying on the model's output or seeking further validation or analysis. In summary, our ensemble of calibration methods represents a promising avenue for enhancing the interpretability and reliability of deep learning models across different domains.

The primary contributions of our research can be summarised as follows:

- Introduction of a post-hoc model calibration method that leverages an ensemble of diverse calibration techniques.
- Exploration of how incorporating uncertainty calculations can assist users in understanding a classifier's predictions, facilitating their decisions regarding the need for a second opinion.
- The proposed approach can be seamlessly integrated as a post-processing module for pre-trained models, eliminating the need for retraining.

The structure of the paper is as follows: Section 2 provides a comprehensive review of relevant literature in the context of DR classification, calibration, and uncertainty. Section 3 delineates our novel approach for achieving reliable and well-calibrated predictions for diabetic retinopathy using our model. In Section 4, we furnish details about the APTOS datasets, share experimental findings, and conduct a performance comparison with previous research. Finally, in Section 5, we present our concluding remarks.

2. Related Work

In the medical industry, deep learning has grown in prominence as a research topic due to its flexibility, adaptability, and unprecedented accuracy in image classification applications. Its success has led to its adoption in various fields such as speech recognition (Mukhamadiyev et al., 2022; Roßbach et al., 2022), pattern recognition (Bijalwan et al., 2022; Jing et al., 2022), and drug discovery (Bai et al., 2022; Pandey et al., 2022). In medical image analysis and classification, deep learning has also been successfully implemented, with several studies focusing on diabetic retinopathy screening (Ayhan et al., 2020; Suganyadevi et al., 2022; Tsuneki, 2022; Van der Velden et al., 2022). For instance, Shu-I Pao et al. proposed a bichannel CNN that processes both DR grey level images and green component images simultaneously, achieving an accuracy of 87.83 %, sensitivity of 77.81 %, specificity of 93.88 %, and Area Under the Curve (AUC) of 0.93 (Pao et al., 2020). Sraddha Das et al. developed a CNN model that used segmented blood vessels from fundus images as input and used a combination of squeeze, excitation, and bottleneck layers to detect DR, achieving 98.7 % accuracy and 97.2 % precision on the DIARETDB1 dataset (Das et al., 2021). Gabriel TozattoZago et al. employed a dual convolutional neural network approach, with one network used for selecting patches from the training image and the other for training the selected patches. Their model showed promising results in diabetic retinopathy (DR) screening, achieving a receiver operating characteristic curve of 0.912–95 %CI (0.897–0.928)

and a sensitivity of 0.940–95 %CI (0.921–0.959) on the Messidor dataset (Zago et al., 2020). Ling Dai et al. developed a comprehensive deep learning system, called DeepDR, that can identify the various stages of DR. The model consisted of three sub-networks: the image quality assessment sub-network, the lesion-aware sub-network, and the DR grading sub-network. DeepDR achieved impressive area under the receiver operating characteristic curves, with values of 0.901, 0.941, 0.954, and 0.967 for detecting microaneurysms, cotton-wool spots, hard exudates, and haemorrhages, respectively (Dai et al., 2021).

Several recent research studies have investigated the use of deep learning techniques, particularly transfer learning, to detect diabetic retinopathy. For example, Ayaka Sugeno et al. employed transfer learning to develop a DR severity grading system using a state-of-the-art convolutional neural network, EfficientNet-B3, and the publicly available Kaggle Asia Pacific Tele-Ophthalmology Society (APTOS) 2019 training dataset. The trained model demonstrated high specificity and sensitivity values of 0.98 for detecting DR retinas. Additionally, the accuracy values for severity grading for the first, second, and third predicted labels were 0.84, 0.95, and 0.98, respectively (Sugeno et al., 2021). In another study, Francisco J. Martinez-Murcia et al. employed transfer learning with a ResNet50-based architecture and achieved an AUC of 0.97 for binary classification (Martinez-Murcia et al., 2021).

Despite their high accuracy, deep learning models are not yet effective in clinical practice because their probabilistic output does not accurately reflect the true likelihood of the prediction (Zadrozny & Elkan, 2001). Most models incorrectly interpret softmax output as a genuine confidence score, which can result in inaccurate predictions. This is particularly problematic in the medical field, where the risk of missing an infected patient is extremely hazardous. To address this issue, researchers have employed various calibration approaches, such as temperature scaling, to calibrate the model's confidence scores to their true values. For example, Murat Seçkin Ayhan et al. achieved an adaptive expected calibration error of 0.01 using temperature scaling (Ayhan et al., 2020). However, a single calibration method may not always produce well-calibrated results, prompting researchers to propose an ensemble of calibration methods to increase or decrease the model's prediction confidence. Chuan Guo presented a comparative table of various calibration methods, including histogram binning, Isotonic regression, temperature scaling, vector scaling, matrix scaling, and Bayesian Binning into Quantiles (BBQ), that have been applied as post-processing techniques without requiring model training (Guo et al., 2017). These calibration methods have been applied as post-processing where no model training is required, and we discovered that a single method does not always produce a well-calibrated result, so motivated by this limitation, we proposed an ensemble of calibration methods to decide whether to increase or decrease the confidence of the model prediction.

3. Methodology

In Figure 1, the architecture diagram illustrates the construction of the DR prediction model, which is primarily composed of the following steps: (1) The fundus images from the dataset were utilised to train the EfficientNet CNN model using transfer learning. (2) Upon completion of the training process, the softmax probability was generated on the validation set. (3) The generated softmax probabilities from the validation dataset were employed to train the calibration model. During the training process, the function learned to map the softmax probability to the actual probability. After learning the calibration model, our final diagnostic model was ready to predict the test data. (4) For a test image, the

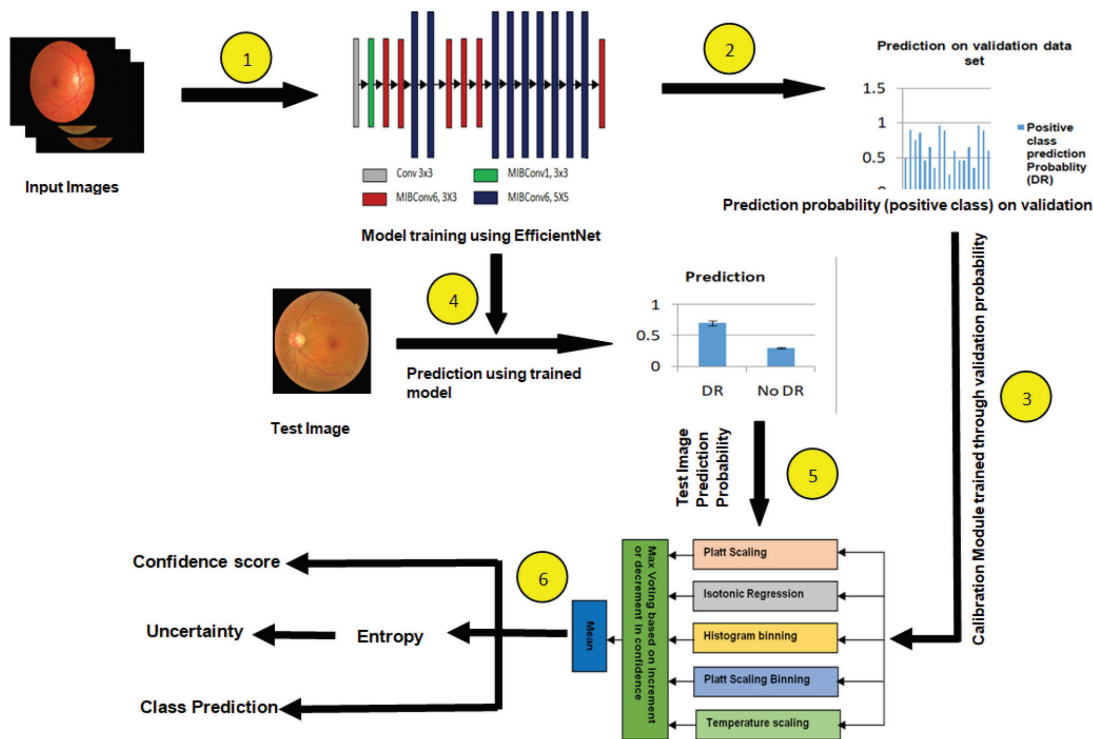


Figure 1. Architecture of the proposed DR detection system

trained model was used to generate the softmax output. (5) The generated softmax output was sent to the trained calibration module to calibrate the confidence score. (6) The confidence output was then used to estimate the uncertainty and finally, output the prediction result along with the confidence and uncertainty scores.

3.1. State of the Art CNN Based Models

To benchmark the performance of our proposed EfficientNet B3 model, we compare it against other state-of-the-art CNN architectures such as DenseNet, ResNet, InceptionNet, and XceptionNet. These models were selected due to their established performance in medical image analysis and their relevance as baselines in diabetic retinopathy detection tasks.

3.1.1. EfficientNet

The EfficientNets (EfficientNet-B0 through EfficientNet-B7) are a collection of pre-trained CNN models that were trained using the ImageNet dataset. The EfficientNet model is based on the assumption that increasing network dimensions such as depth (D), width (W), and resolution (R) will increase network performance and efficiency. As a result, Tan & Le, (2019) suggested a compound scaling

strategy, as depicted in Eq. (1), in which they evenly scaled up the network's size using compound coefficients.

$$\begin{aligned}
 \text{depth (D)} &= a^{\phi} \\
 \text{width (W)} &= b^{\phi} \\
 \text{Resolution (R)} &= c^{\phi} \\
 \text{s.t. } a \cdot b^2 \cdot c^2 &\approx 2 \\
 a \geq 1, b \geq 1, c &\geq 1
 \end{aligned} \tag{1}$$

Where D, W, and R are grid-search-derived constants. EfficientNet-B0 was scaled up to develop EfficientNet-B1 through B7 while maintaining the D, W, and R constants in Eq (1).

3.1.2. DenseNet

The name «DenseNet» was coined at the CVPR Conference for the densely connected convolutional networks that Gao Huang, Zhuang Liu, and his coworkers developed in 2017 (Huang et al., 2017). The denseNet architecture connects every layer to every other layer, which is referred to as a densely connected convolutional network. For K layers, there are $K(K+1)/2$ direct connections. Each layer takes as inputs all of the feature maps from the preceding layers, as well as that layer's own feature maps for each following layer and its own feature maps for all succeeding layers. The removal of the vanishing-gradient problem, enhanced feature propagation, higher feature reuse, and much fewer parameters are just a few of the compelling benefits offered by dense networks.

3.1.3. Residual Networks (ResNet)

The ResNet architecture, which was proposed by Microsoft Research specialists in 2015, presented a revolutionary design (He et al., 2016). This design developed the Residual Blocks concept to deal with the vanishing/exploding gradient issue. This network employs a process known as skip connections. The residual blocks connect the activations of the previous layers to the subsequent layers by bypassing certain layers in between. This technique is also known as skip connection technique. This type of skip connection has the advantage that regularisation will skip any layer that impairs architectural performance. As a result, it is feasible to train a very deep neural network without running into problems with gradients that are expanding or dissipating.

3.1.4. InceptionNet

A deep neural network with an architectural structure made up of recurring elements called Inception modules is referred to as an inception network (Szegedy et al., 2015). A strong deep neural network has a large number of layers in its network, which causes two issues. The first is that increasing the number of layers in a neural network may result in overfitting, especially if we have insufficient labelled training data, and the second is increased processing cost. Given the aforementioned issue, the Inception module design enables the use of filters of various sizes without affecting the network's depth. Instead of connecting each filter completely one after the other, they are added in parallel. A concatenation layer, a max pooling layer, a 1x1 convolution layer, a 3x3 convolution layer, a 5x5 convolution layer, and an input layer are all included in an inception module. The 1x1 convolution layer can assist to minimise model size, which reduces computing cost.

3.1.5. XceptionNet

The term «extreme inception», or Xception, takes the concepts of Inception to their logical conclusion. In Inception, each depth space was processed using a different filter based on the input spaces after the original input was compressed using 1x1 convolutions. Xception just reverses this process. Instead, it first applies the filters to every depth map separately before condensing the input space using 1x1 convolution. This is also known as depth wise separable convolutions. Skip connection is a ResNet design feature that is also used by Xception networks, which increases computational performance (Chollet, 2017).

3.2. Transfer Learning

Transfer learning is a powerful technique in machine learning and deep learning where knowledge learned from one task or domain is transferred to another related task or domain. Instead of training a model from scratch for a new task, transfer learning enables leveraging pre-trained models that have already learned features from large amounts of data (Iman et al., 2022; Weiss et al., 2016). This allows for faster training with smaller datasets and often results in improved performance, as the pre-trained model has already learned relevant representations that can be generalized to the new task. Transfer learning has been widely used in various applications, including image recognition, natural language processing, and speech recognition, and has shown to be a valuable approach for building accurate and efficient machine learning models with limited data or computational resources (Kim et al., 2022).

3.3. Calibration

Calibration aims to identify the optimal function that best captures the relationship between the predicted probabilities and the actual likelihood of the specific event occurring. In our scenario, we calibrate the softmax probabilities derived from the CNN model to align with the true probabilities of diabetic retinopathy.

A well-calibrated classifier is particularly reliable, especially in critical domains such as healthcare, where sound decisions are paramount. When our classifier is well-calibrated, any prediction it makes with low confidence can prompt a referral to a healthcare professional, facilitating more informed decisions. To achieve this, we have implemented an ensemble of various methods for calibrating the probabilities to determine genuine confidence scores.

The selected calibration techniques—Platt Scaling, Isotonic Regression, Histogram Binning, Platt-Histogram Binning, and Temperature Scaling—represent both parametric and non-parametric approaches, providing a diverse range of strategies for improving calibration. These methods have been widely used in literature due to their simplicity, interpretability, and effectiveness in post-hoc calibration without requiring model retraining.

3.3.1. Platt Scaling

Platt Scaling is a parametric approach (Platt, 1999), where the probabilities of the validation (z_i) set are used to obtain the calibrated probabilities. In this method, a logistic regression model is used to learn two parameters, i.e, $x, y \in \mathbb{R}$, and outputs the calibrated probabilities $P_{\text{platt}} = \sigma(xz_i + y)$. Where σ is the sigmoid function. This expression can be equivalently written as equation (2).

$$P_{\text{platt}} = \frac{1}{1 + e^{xz_i + y}} \quad (2)$$

3.3.2. Isotonic Regression

Isotonic Regression is a non-parametric approach (Zadrozny & Elkan, 2002), where a piecewise constant function h learns from the validation dataset to output scores that minimise the squared loss $\sum_{i=1}^n (h(p_i^{\text{pred}}) - y_i^{\text{true}})^2$. The optimization problem can be written as equation (3):

$$P_{\text{IR}} = \min_{\substack{\theta_1 \leq \theta_2 \leq \dots \leq \theta_k \\ a_1 \leq a_2 \leq \dots \leq a_{k+1}}} \sum_{i=1}^k \sum_{j=1}^n l(a_k \leq p_j^{\text{pred}} < a_{k+1}) (\theta_k - y_j^{\text{true}})^2 \quad (3)$$

Subject to $0 = a_1 \leq a_2 \leq \dots \leq a_{k+1} = 1$

Where k is the number of intervals, $a_1, a_2, \dots, a_{k+1}$ are the interval boundaries; and $\theta_1, \dots, \theta_k$ are the function values.

3.3.3. Histogram Binning

Histogram Binning is an approach that is a less strict version of isotonic regression, where the prediction space is divided into either equal-width intervals or intervals containing an equal number of samples, and a piecewise constant function h is learned over these bins. The predictions θ_k are chosen to minimise the bin-wise squared loss (equation 4):

$$P_{\text{HB}} = \min_{\theta_1 \leq \theta_2 \leq \dots \leq \theta_k} \sum_{i=1}^k \sum_{j=1}^n l(a_k \leq p_j^{\text{pred}} < a_{k+1}) (\theta_k - y_j^{\text{true}})^2 \quad (4)$$

3.3.4. Platt Scaling and Histogram Binning

Platt Scaling and Histogram Binning are combined in this method. We first fit a Platt Scaling function, f , and then bin the input samples similarly to Histogram Binning. The primary change is that instead of using the initially predicted probability, we bin the samples using the Platt Scaling output. The calibrated forecast for each bin is the average of the scaling function, f , rather than the average number of positive samples. The motivation behind this strategy is that the output from our scaling function falls inside a narrower range than the original label values, resulting in a lesser calibration error.

$$p^{\text{calib}} = \frac{1}{1 + e^{xz_i + y}}$$

$$P_{\text{PCB}} = \min_{\theta_1 \leq \theta_2 \leq \dots \leq \theta_k} \sum_{i=1}^k \sum_{j=1}^n l(a_k \leq p_j^{\text{pred}} < a_{k+1}) (\theta_k - y_j^{\text{true}})^2 \quad (5)$$

3.3.5. Temperature Scaling

Temperature Scaling (Guo et al., 2017) is a post-processing calibration method that operates directly on the logits produced by the final layer of a neural network, without initially applying the

softmax function. A scalar temperature parameter T is optimised on a held-out validation set to adjust (or scale) the logits (z_i), thereby producing calibrated probability estimates.

$$p_i^{\text{calib}} = \frac{e^{z_i/T}}{\sum_{j=1}^K e^{z_j/T}} \quad (6)$$

T stands for temperature, and $T > 1$ «softens» the softmax (i.e. increases the output entropy). The probability p_i^{calib} approaches $1/K$, indicating greater uncertainty, as T increases. We revert to the initial probability with $T = 1$. The probability collapses to a point mass as T approaches zero. On the validation set, T is optimised concerning Negative Log-Likelihood (NLL). The class prediction remains unaltered since parameter T does not affect the softmax function's maximum. To put it another way, temperature scaling has no bearing on the model's accuracy.

Where z is the final output of the last layer without using the softmax function called logit, and T is the learned parameter. We learn this parameter on a validation set, where T is chosen to minimise negative log-likelihood (NLL).

3.4. Ensemble of Calibration Methods

The reason for utilising an ensemble of calibration methods is because it has been found that one technique of calibration sometimes raises the confidence value when it should be lowered and sometimes reduces when it should be increased. Motivated by this limitation, we proposed an ensemble of methods that individually calibrate the confidence value, and then use maximum voting on methods to determine whether the maximum method increased the confidence; if so, the maximum of all calibrated confidence is taken; otherwise, the minimum value of all methods is taken.

3.5. Accessing Uncertainty

In terms of deep learning, uncertainty is the measurement of the worthiness of the prediction given by the classifier. If the uncertainty is high, the predicted result is not trustworthy, on the other hand, if the uncertainty is low, we can trust our classifier's decision. In the proposed model, the uncertainty associated with each prediction is quantified using the entropy of the obtained calibrated confidence value.

$$\begin{aligned} H[p(y|x; f^\omega)] &= -\sum_k p(y = c|x; f^\omega) \log p(y = c|x; f^\omega) \\ &= -\sum_k p_k \log p_k \end{aligned} \quad (7)$$

For our model we are doing binary classification so the above equation is reduced to $-\text{plogp} - (1-p)\log(1-p)$. Where p is the calibrated confidence value obtained from the ensemble of calibration methods.

3.6. Proposed Model Algorithm

Algorithm 1: Proposed Model Algorithm	
Input:	<i>Training set</i> = $\{x, y\}_{i=0}^{N_{train}}$, <i>Validation set</i> = $\{x', y'\}_{i=0}^{N_{val}}$, <i>Test Image</i> = (x^*, y^*) , <i>Pretrained EfficientNet</i> = f^ϕ , C_1 = Platt scaling, C_2 = Isotonic regression C_3 = Histogram binning, C_4 = Platt histogram, C_5 = Temperature scaling
Output:	<i>Confidence score</i> : y_{conf}^* <i>Predictive Uncertainty</i> : y_u^*
1:	// Step 1: Training using transfer learning
2:	$f^\phi \leftarrow \text{Train}(f^\phi, \{(x_i, y_i)\} \text{ for } i = 1 \text{ to } N_{train})$
3:	// Step 2: Fit Calibration function on validation set
4:	for each calibration method $C_k \in \{C_1, C_2, C_3, C_4, C_5\}$ do
5:	$h_k \leftarrow \text{FitCalibration}(C_k, \{P_{valj}\} \text{ for } j = 1 \text{ to } N_{val})$
6:	end for
7:	// Step 3: Predict softmax probability for the test image
8:	$P_{test} \leftarrow \text{PredictProbability}(y^* = 1 \mid x^*, f^\phi)$
9:	// Step 4: Apply all calibrators to get calibrated probabilities
10:	for each calibration function h_k do
11:	$P_{test_k} \leftarrow h_k(P_{test})$
12:	end for
13:	// Step 5: Ensemble decision on confidence score
14:	if Majority (h_k) increases confidence then
15:	$y_{conf}^* \leftarrow \text{Max}(P_{test1}, P_{test2}, P_{test3}, P_{test4}, P_{test5})$
16:	else
17:	$y_{conf}^* \leftarrow \text{Min}(P_{testC1}, P_{testC2}, P_{testC3}, P_{testC4}, P_{testC5})$
18:	end if
19:	// Step 6: Predictive uncertainty
20:	$\text{Uncertainty}(y_u^*) = \frac{1}{2} (y_{conf}^* \log(y_{conf}^*) - (1 - y_{conf}^*) \log(1 - y_{conf}^*))$

In essence, the algorithm trains a model using transfer learning, then calibrates prediction scores using five different post-hoc methods. For each test sample, all calibration methods are applied, and their outputs are aggregated using a majority vote to determine whether to use the

highest or lowest confidence. This calibrated confidence score is then used to compute uncertainty via entropy.

4. Results and Discussion

4.1. Dataset

For the classification experiments in this study, the researchers utilised the APTOS 2019 dataset, which was obtained from the Kaggle website [41]. This dataset, provided by the Aravind Eye Hospital in India, comprises of 3662 high-resolution images that have been assessed by qualified medical professionals, along with accompanying samples. All images in the dataset have been rated based on the International Clinical Diabetic Retinopathy Severity Scale (Pan et al., 2013), with each image being assigned a score on a scale of 0 to 4, representing the severity of diabetic retinopathy as follows: 0 - No DR, 1 - Mild, 2 - Moderate, 3 - Severe, and 4 - Proliferative DR.

The dataset used for binary classification contained 1857 DR and 1805 No-DR images, showing minimal imbalance. Therefore, oversampling or weighting techniques were not applied in this study. For future studies involving more skewed datasets or multiclass problems, techniques such as SMOTE, class weighting, or balanced batch generation will be considered.

4.2. Evaluation Metrics

In this research, we harnessed the power of transfer learning to train a DNN based on the EfficientNet B3 architecture for the purpose of predicting diabetic retinopathy. To assess the performance of EfficientNet and the other state-of-the-art CNN models detailed in this paper, we employed an array of performance metrics, including accuracy, sensitivity, specificity, precision, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). The metrics presented in sections 4.2.2 through 4.2.6 were derived by considering various indices such as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), which were extracted from the confusion matrix generated during the classification process. TP represents the count of correctly identified unhealthy images, while TN signifies the tally of correctly identified healthy images. FN denotes the number of misclassified unhealthy images, and FP indicates the quantity of inaccurately categorised healthy images. The model's calibration was assessed through the computation of the Expected Calibration Error (ECE), which involved dividing the predicted confidence values into equal-length bins and quantifying the disparity between the predicted and actual confidence values.

4.2.1. Calibration Metrics: Expected Calibration Error (ECE)

We can calculate the calibration error by dividing the predictions into k bins of the same interval. Then, the expected calibration error can be calculated by estimating the difference between the accuracy and confidence of the predictions that fall into different bins. The accuracy of a bin can be calculated using Eq. (8)

$$\text{Accuracy} (B_k) = \frac{1}{|B_k|} \sum_{j \in B_k} 1(Y_j^{\text{pred}} - Y_j^{\text{true}}) \quad (8)$$

Where B_k is the k th bin and we do the summation of the differences between the actual output (Y^{true}) and predicted output (Y^{pred}). The confidence of a bin can be calculated using Eq. (9)

$$\text{Confidence}(B_k) = \frac{1}{|B_k|} \sum_{j \in B_k} p^{\text{pred}} \quad (9)$$

Where p^{pred} is the probability of the positive class of a binary classifier.

The expected calibration error (ECE) can be calculated using Eq. (10) by obtaining the difference in accuracy and confidence. The well-calibrated classifier is where the average accuracy and average confidence of each bin should be zero.

$$\text{ECE} = \sum_{i=1}^k \frac{|B_k|}{d} |\text{accuracy}(B_i) - \text{confidence}(B_i)| \quad (10)$$

Maximum Calibration Error (MCE) captures the worst-case difference between accuracy and confidence across all bins.

$$\text{MCE} = \max_{i \in \{1, \dots, k\}} |\text{accuracy}(B_i) - \text{confidence}(B_i)| \quad (11)$$

The Brier Score (BS) measures the mean squared difference between predicted probabilities and true labels.

$$\text{BS} = \frac{1}{|d|} \sum_{i=1}^d Y_i^{\text{pred}} - Y_i^{\text{true}} \quad (12)$$

Where d is the number of samples and B_k is the bin size of each bin. A lower Brier Score indicates better calibrated and more accurate probabilistic predictions.

4.2.2. ROC-AUC

Receiver Operating Characteristic – Area Under the Curve is a performance metric used in machine learning to evaluate the effectiveness of a binary classification model. It is useful in situations where the distribution of positive and negative instances in the data is imbalanced. It provides a single number that summarises the model's ability to correctly classify positive and negative instances at different classification thresholds.

4.2.3. Accuracy

The percentage of correct predictions out of the overall number of predictions.

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

4.2.4. Precision

The percentage of correctly identified positive cases.

$$\text{Precision} = \frac{TP}{TP + FP}$$

4.2.5. Sensitivity (Recall)

It measures the proportion of positive instances that were correctly identified by the model out of all the actual positive instances in the data. Mathematically, recall is defined as:

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

4.2.6. Specificity

It measures the proportion of actual negative instances that are correctly identified as negative by the model, out of all the actual negative instances in the data.

$$\text{Sepecificity} = \frac{TN}{TN + FP}$$

4.2.7. F1 Score

It sits at a harmonic midpoint between precision and recall, with a range of [0, 1]. This metric commonly reflects the accuracy (successful classification of examples) and reliability of the system (minimal instances missed by our classifier).

$$\text{F1 - Score} = 2 * \frac{\text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

The experiment comprises two separate tasks. The first task focuses on binary classification, specifically identifying DR. The second task involves calibrating the model to determine its authentic confidence level, thereby facilitating the estimation of the model's uncertainty.

4.3. Task1: Building a Model Using Transfer Learning

The primary aim of the proposed model is to identify the most effective state-of-the-art approach for predicting diabetic retinopathy in a binary classification context. This challenge necessitates the use of a dataset containing both healthy (No-DR) and diseased (DR) images. The APTOS 2019 dataset is segregated into two categories, comprising 1857 diseased images in one class and 1805 healthy images in the other. The data is divided into training, validation, and testing sets in a 7:1:2 ratio, with 2626 images allocated for training, 657 for validation, and 365 for testing.

For the classification of fundus photos for diabetic retinopathy using transfer learning, this research employs established CNN architectures, as presented in Table 1. Given the limitations of a small dataset and the need to reduce training time for deep learning models, pre-trained weights from state-of-the-art CNN models on ImageNet were employed. To align with the specific problem, the final fully connected (FC) layers of all models, originally designed for 1000 outputs, were adapted for binary outputs. Table 1 provides details on the number of top layers set to be trainable, while keeping all other layers non-trainable. The activation function chosen for the final layer was Softmax, and the loss function used was categorical cross-entropy. All models utilised the Adam optimization method with a learning rate of 0.00001. As part of this study, every pixel value in the datasets was normalised by dividing it by 255. To accommodate hardware constraints, the images were resized to 224x224 pixels.

Table 2 presents a comparative analysis of evaluation metrics for various CNN networks utilised in a binary classification task. As per the data in the table, EfficientNet B3 stands out as the top performer, excelling in accuracy, precision, F1 score, AUC, and recall, boasting respective scores of 96 %, 96.1 %, 95.2 %, 98 %, and 97.2 %. Notably, all models achieved similar accuracy scores, as shown in Table 2. In addition, the EfficientNet B3 model exhibited the advantage of having fewer trainable parameters (Table 1) in comparison to all other models, solidifying its position as the leader in binary classification accuracy.

Concerning the ROC curve, Figure 2(f) illustrates the ROC curves of various state-of-the-art models, including our EfficientNet B3 model. The ROC curve offers insights into the accuracy of the model's predictions, with the AUC value positioned in the top left corner of the ROC graph, denoting the area under the ROC curve and ranging from 0 to 1. A larger AUC signifies the model's greater ability to distinguish between positive and negative instances, signifying superior overall performance. In the

Table 1. Configuration details for each CNN architecture used in the study

CNN Architecture	No. of Epochs	Input Size	Learning rate	Unfrozen top layers of the models	Total parameters (params)
EfficientNet B3	20	224X224	0.00001	20 layers	Total: 18,592,865 Trainable: 18,467,658 Non-trainable: 125,207
DenseNet201	30	224X224	0.00001	10 layers	Total: 19,306,562 Trainable: 19,077,506 Non-trainable: 229,056
ResNet152	40	224X224	0.00001	20 layers	Total: 59,421,058 Trainable: 59,269,634 Non-trainable: 151,424
XceptionNet	20	224X224	0.00001	10 layers	Total: 21,911,594 Trainable: 21,857,066 Non-trainable: 54,528
InceptionNetV3	20	224X224	0.00001	10 layers	Total: 22,852,898 Trainable: 22,818,466 Non-trainable: 34,432

Table 2. Comparison of classification performance metrics across different CNN models on the binary DR classification task

CNN Architecture	Accuracy	Specificity	Sensitivity	Precision	Area Under the Curve	F1- Score	Expected Calibration Error
EfficientNet	96	94.5	97.2	96.1	98	95.2	0.060
DenseNet201	95	93.2	96.7	95.2	98	94.2	0.069
ResNet152	95	95.2	95.3	93.4	98	94.3	0.056
XceptionNet	93	91.9	94.9	91.9	97	91.9	0.077
InceptionNetV3	92	88.5	94.5	0.92	97	90.3	0.072

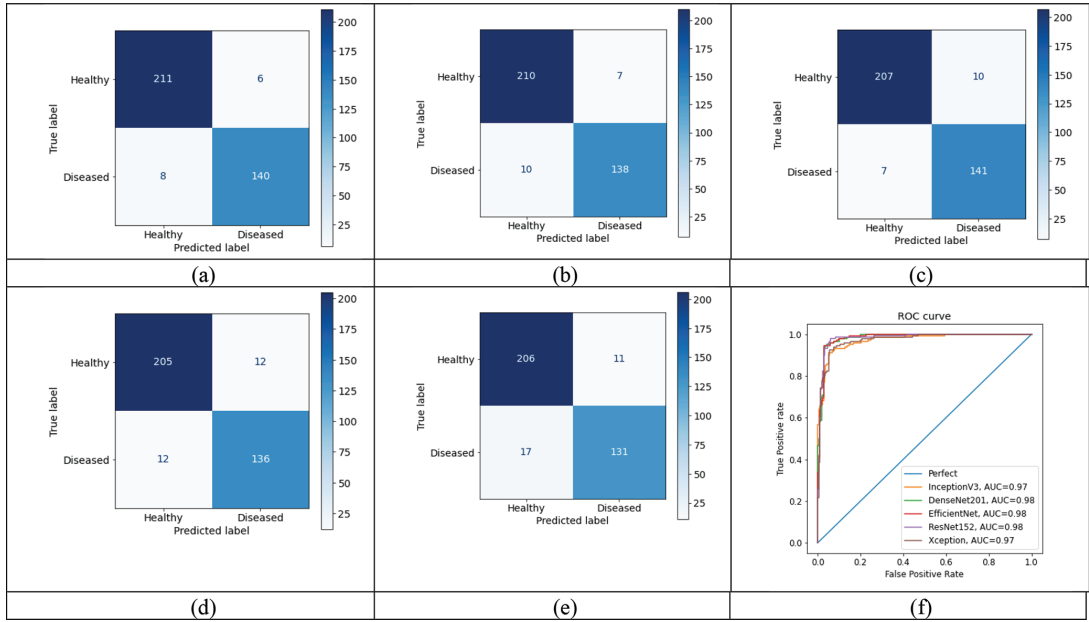


Figure 2. Confusion matrices and ROC curves for five CNN architectures. Subfigures (a–e) show true vs predicted labels; (f) shows ROC curves and AUC values for evaluating classifier performance

results, the AUC values for InceptionV3 and Xception models were 97 %, while the EfficientNet B3, DenseNet201, and ResNet152 models achieved higher AUC values of 98.00 %.

To assess the classifier's accuracy, a confusion matrix is employed, with predicted classes represented in columns and actual classifications in rows. The matrix's diagonal line corresponds to cases where the predicted and actual values align, and the colour intensity signifies the accuracy of the predictions. In Figure 2 (a, b, c, d, e), the confusion matrix for all models used in the binary classification task is displayed. EfficientNetB3, for instance, achieved 351 correct predictions out of 365 test photos,

Table 3. Performance comparison of the proposed EfficientNet B3 model with recent state-of-the-art methods using the APTOS 2019 dataset

paper	Method	Dataset	Sensitivity	Precision	AUC	F1-Score	Accuracy
(Ayaka Sugeno et.al. [32])	EfficientNet-B3 with enhanced images	APTOS	98.0	-	94.0 (average)		84.0
(Kobat et al., 2022)	pre-trained DenseNET	APTOS	92.67	92.77	-	92.70	93.85
(Adak et al., 2023)	Ensembled transformer	APTOS	92.88	90.55	-	91.67	95.75
(Oulhadj et al., 2022)	deep learning and deformable registration	APTOS	77.0	71.0	-	83.0	83.0
Proposed model	EfficientNet-B3	APTOS	97.2	96.1	98.0	95.2	96

with 14 misclassifications. Similarly, DenseNet201, ResNet152, Xception, and Inception V3 delivered accurate predictions of 348, 348, 341, and 337, respectively.

Several studies have utilised the APTOS 2019 dataset to classify DR. To better understand the performance of our EfficientNet model, we compared it with existing techniques in the literature. Table 3 shows that our proposed technique outperforms existing methods, achieving an accuracy of 96 %.

4.4. Task 2: Optimization of confidence score using ensemble approach and estimation of prediction uncertainty

The primary objective of this study is to ensure that the softmax probability of a deep learning model is calibrated to its true confidence score, resulting in more reliable predictions. This means that if the model predicts a 30 % chance of a patient having diabetic retinopathy, the actual frequency of diabetic retinopathy in that group of patients should be 30 %. Many previous studies have attempted to calibrate deep learning models using various methods (Ayhan et al., 2020; Van der Velden et al., 2022), however, relying on a single method can lead to incorrect adjustments in confidence scores. In Table 4, for instance (3rd row), the model predicted a negative label as positive with a high softmax probability of 0.904664, but Platt scaling increased the confidence score to 0.95, which was incorrect. Similarly, the model predicted the correct label (2nd row) but with low confidence of 0.72322, and applying calibration methods increased the confidence score.

To address these issues, the authors propose an ensemble method that combines several calibration techniques, including Platt scaling, temperature scaling, isotonic regression, histogram binning, and Platt histogram binning to calibrate the model's confidence score. The method employs a majority voting approach to determine whether to increase or decrease the confidence level based on the outputs of all calibrators. For instance, if the majority of calibrators decrease the confidence value, the model finalises the lowest confidence level obtained from all calibrators. Using this ensemble approach, the authors aim to improve the accuracy and reliability of deep learning models in predicting medical conditions such as diabetic retinopathy.

Table 4. Example output probabilities and confidence values for individual and ensemble calibration methods on selected test samples

True class label	Predicted class label	Test probability (Softmax Probability)	Calibrated Probability (Confidence) of positive class					Ensemble
			Platt scaling	Isotonic Regression	Histogram binning	Platt histogram	Temperature scaling	
0	0	0.010952	0.0311	0.034211	0.020903	0.020224	0.036609	0.036609
1	1	0.72322	0.818182	0.825858	0.750897	0.804185	0.79978	0.825858
0	1	0.904664	0.95	0.825858	0.906518	0.804185	0.87164	0.804185
1	0	0.382798	0.446429	0.474934	0.450266	0.460929	0.458139	0.474934

Figure 3 (shown below) illustrates the performance of the ensemble method across all models used in this study. To showcase the effectiveness of the calibration method, both a reliability diagram and a histogram map were employed. A reliability diagram is constructed by plotting the relative frequency of positive instances against the average predicted probability for each bin, while a histogram illustrates the utilization of bin values in computation. In Figure 3(a), the reliability graph for the EfficientNet model is presented, showcasing the application of all calibration functions to the EfficientNet data. By examining this graph, it becomes evident that without any calibration, the predicted calibration error stands at 0.060, as estimated using equation 11. However, through the ensemble approach, this error is reduced to 0.057.

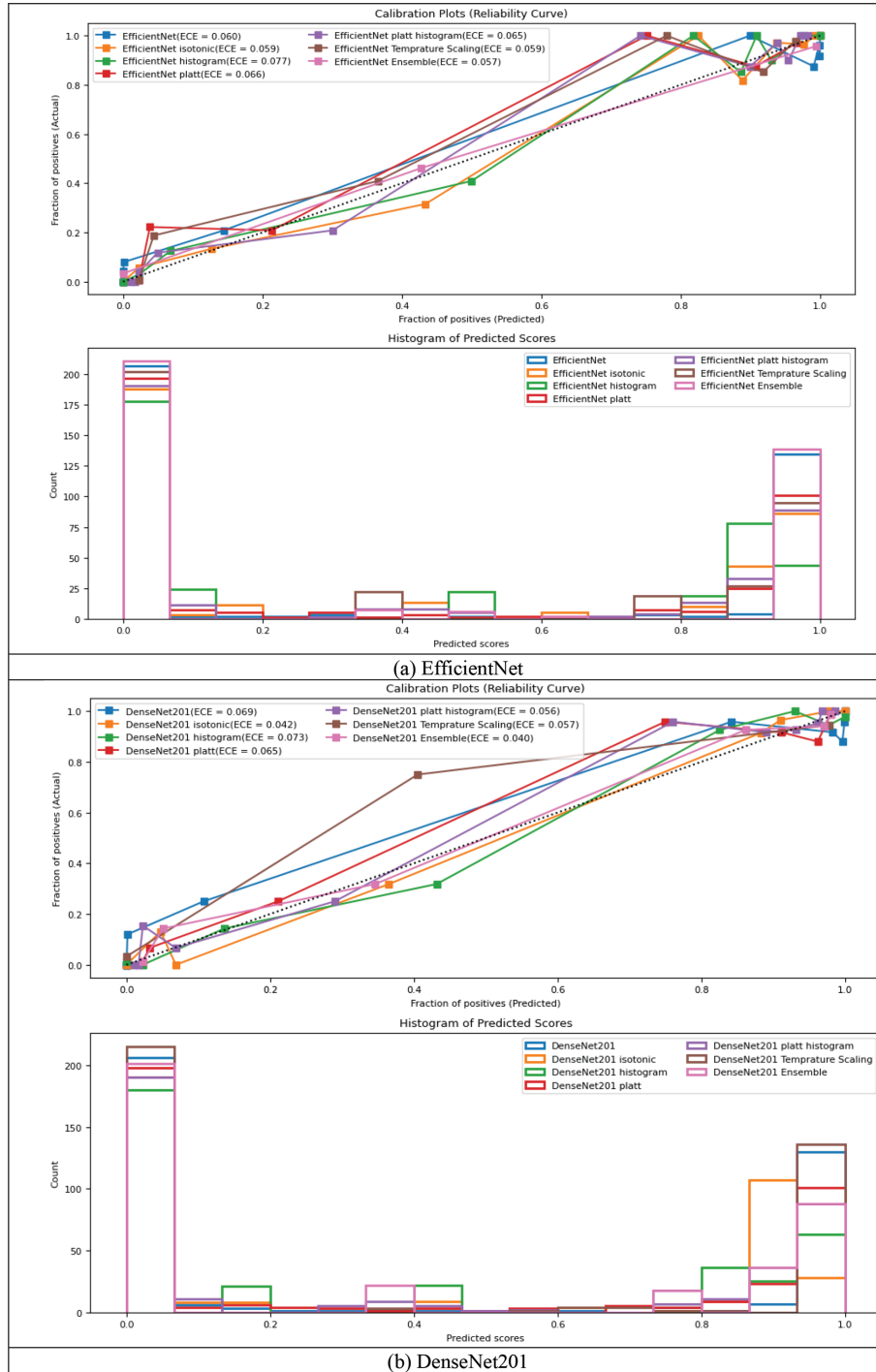
Similarly, the Densenet network (as seen in Table 5, 2nd row) initially yields an Expected Calibration Error (ECE) of 0.69, but after applying the ensemble calibration approach, it decreases to 0.40. With enhanced prediction confidence, it becomes feasible to assess uncertainty using the entropy method for binary classification. Figure 4 portrays how the calibration of a model's confidence has elevated the uncertainty associated with erroneous predictions.

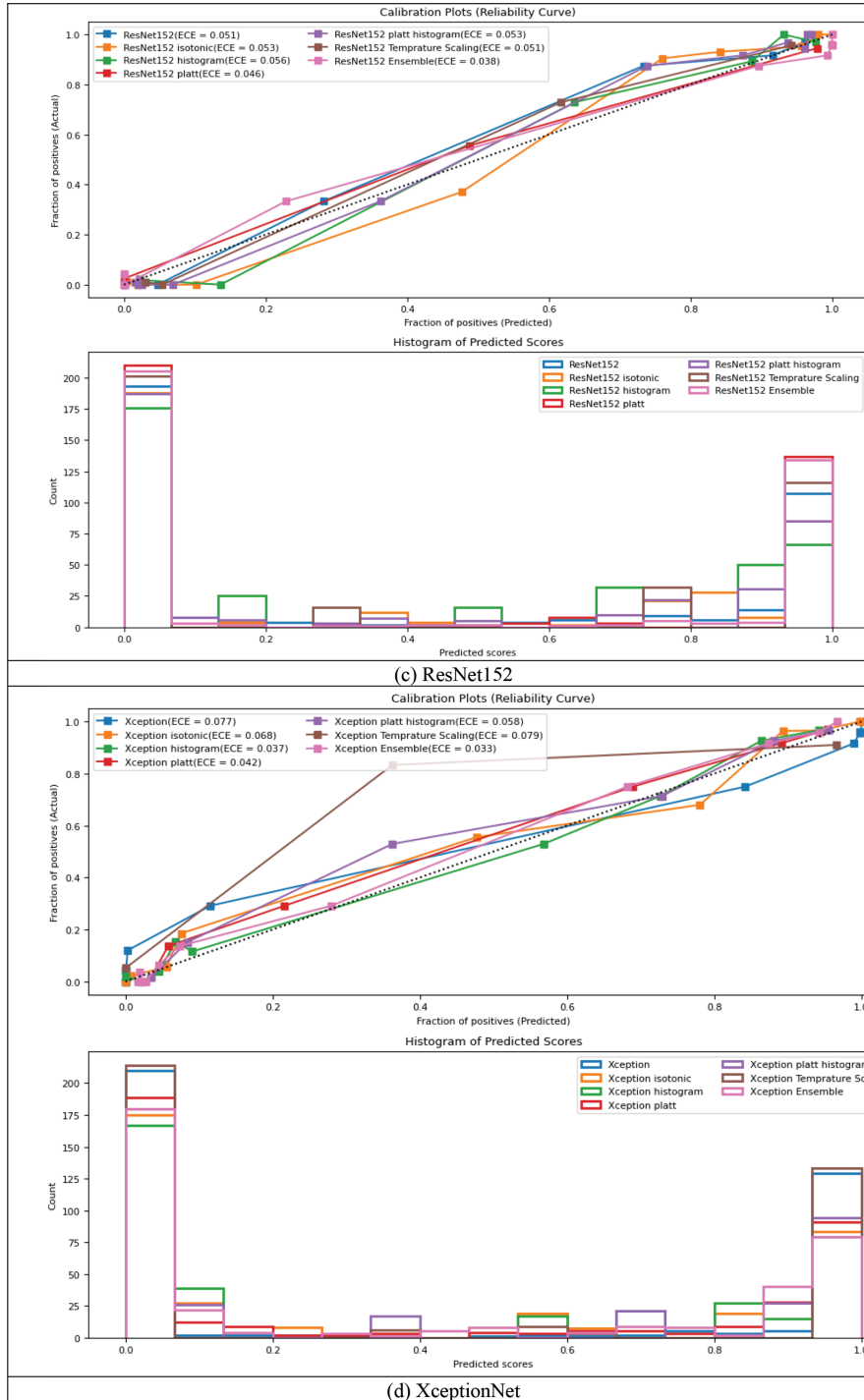
Table 5 demonstrates that various other models have also displayed encouraging outcomes when utilizing the ensemble method for calibration. By calibrating the model's confidence to their true values, ECE was reduced, thereby enhancing the model's reliability.

In addition, Table 5 reveals that for each CNN architecture, different calibration techniques have resulted in less ECE compared to not utilizing the calibration module. For example, for DenseNet201 (Table 5, 2nd row), the Platt histogram yields the least ECE mean, whereas for ResNet152, the histogram method results in the lowest error. However, the ensemble method has consistently resulted in less error across all models.

In addition to ECE, we computed the Maximum Calibration Error (MCE) and Brier Score to provide a more holistic assessment of calibration. The MCE measures the worst-case deviation between predicted and true probabilities, while the Brier Score captures the mean squared error between predicted probabilities and actual outcomes. These values are summarised in Table 6 and further support the effectiveness of the ensemble calibration approach.

Upon obtaining the calibrated confidence score, the uncertainty of the predicted outcome can be computed. Figure 4(c) depicts an instance in which the test image is classified as a DR image with a softmax probability of 0.98, despite being an image without DR. After calibration, the model's prediction confidence score was reduced to 0.80, and by utilizing equation 7, the model calculated an





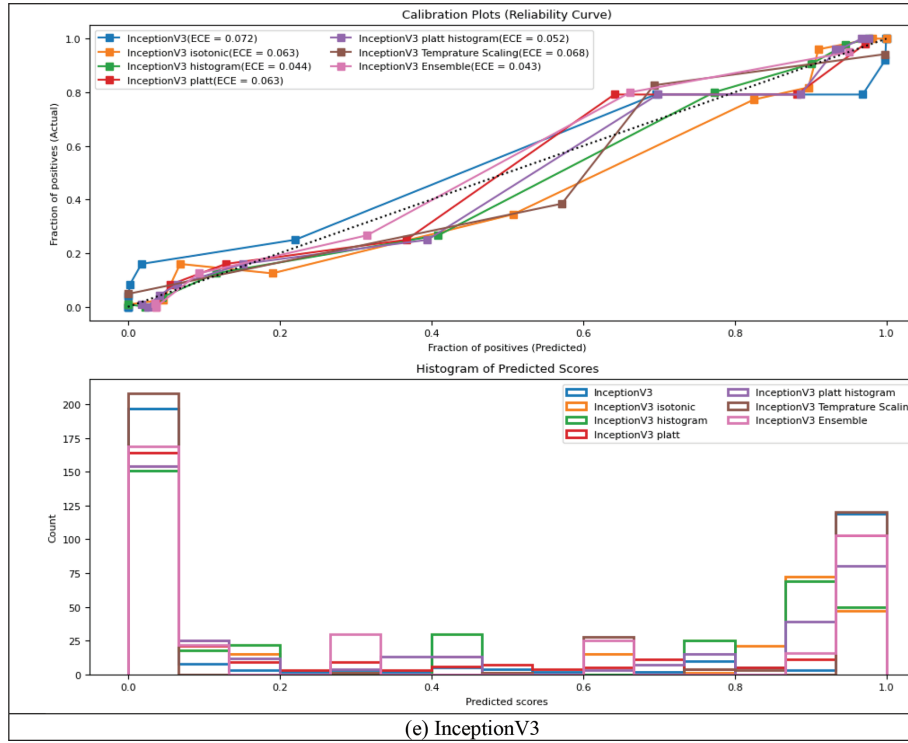


Figure 3. Reliability diagrams for different CNN models, showing how predicted probabilities align with actual outcomes before and after applying ensemble calibration. A closer alignment with the diagonal line indicates better calibration

Table 5. Expected Calibration Error (ECE) values before and after applying individual calibration techniques and the proposed ensemble approach for different CNN architectures

Architecture	ECE	ECE after Platt Scaling	ECE after Isotonic Regression	ECE after Histogram Binning	ECE after Platt scaling Binning	ECE after Temperature Scaling	ECE after Ensemble method
EfficientNet	0.060	0.066	0.059	0.077	0.065	0.059	0.057
DenseNet201	0.069	0.065	0.042	0.073	0.056	0.057	0.040
ResNet152	0.051	0.046	0.053	0.056	0.053	0.051	0.038
XceptionNet	0.077	0.042	0.068	0.037	0.058	0.079	0.033
InceptionNetV3	0.072	0.063	0.063	0.044	0.052	0.068	0.043

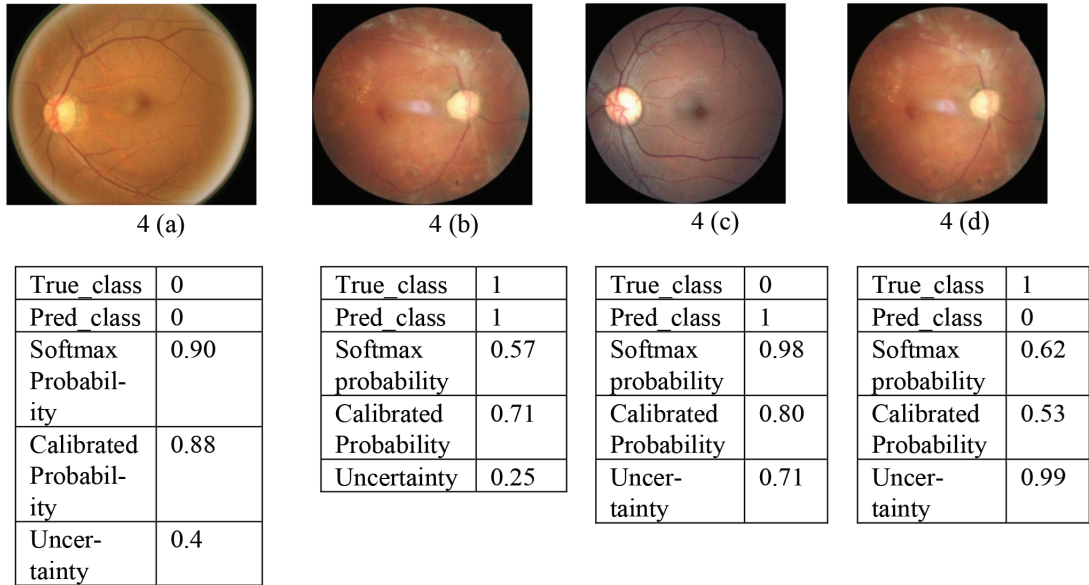


Figure 4. Example test predictions with associated softmax probabilities, calibrated confidence scores, and uncertainty estimates. Demonstrates how entropy-based uncertainty highlights uncertain predictions needing second opinion

Table 6. Calibration performance comparison using Expected Calibration Error (ECE), Maximum Calibration Error (MCE), and Brier Score for each CNN model. Lower values indicate better calibration quality

Model	ECE	MCE	Brier Score
EfficientNet	0.060	0.110	0.045
DenseNet201	0.069	0.123	0.051
ResNet152	0.056	0.102	0.049
XceptionNet	0.077	0.131	0.057
InceptionNetV3	0.072	0.128	0.053

uncertainty value of 0.71, which is relatively high. Therefore, it can be concluded that the prediction outcome requires a second opinion.

The ensemble method benefits from the complementary strengths of diverse calibration techniques. While individual calibrators may occasionally over- or under-estimate confidence, the ensemble reduces this risk by averaging across methods, thus improving robustness. However, in datasets with limited calibration samples or extreme class imbalance, ensemble performance may be less effective due to noisy consensus.

5. Conclusion

Diabetic retinopathy is a leading cause of blindness worldwide, and early detection and accurate prediction of its progression are critical for effective management and prevention of vision loss. The use of EfficientNet B3, a state-of-the-art convolutional neural network (CNN), in conjunction with calibrated probabilities has shown promising results in the diagnosis and management of diabetic retinopathy. EfficientNet B3, known for its superior performance in image recognition tasks, has demonstrated its potential in diabetic retinopathy screening due to its deep architecture and efficient use of computational resources. The calibrated probabilities approach, which adjusts the predicted probabilities to align with the true probabilities of the outcomes, helps to mitigate overconfidence or underconfidence issues commonly associated with deep learning models.

By leveraging the power of EfficientNet B3 and calibrated probabilities, healthcare professionals can improve the accuracy and reliability of diabetic retinopathy diagnosis. This can lead to early identification of the disease, timely referral to ophthalmologists, and appropriate management strategies to prevent vision loss in patients with diabetes.

Further research and validation of this approach using large-scale datasets and real-world clinical settings are needed to fully establish the clinical utility of EfficientNet B3 and calibrated probabilities in diabetic retinopathy screening. Although this study focuses on binary classification (No DR vs DR), future work will extend this approach to multiclass classification (e.g., mild, moderate, severe, proliferative DR), which would provide more clinically relevant insights and benefit fine-grained risk assessment.

CRedit authorship contribution statement

Preeti Verma: Data curation, Conceptualization, Investigation, Methodology, Validation, Formal analysis, Writing – original draft. **Dr. Sivasankar Elango:** Conceptualization, Investigation, Methodology, Validation, Formal analysis, Writing – review & editing, Supervision. **Dr. Kunwar Singh:** Conceptualization, Investigation, Methodology, Validation, Formal analysis, Writing – review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., ... & Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76, 243-297. <https://doi.org/10.1016/j.inffus.2021.05.008>
- Adak, C., Karkera, T., Chattopadhyay, S., & Saqib, M. (2023). Detecting Severity of Diabetic Retinopathy from Fundus Images using Ensembled Transformers. *arXiv preprint arXiv:2301.00973*. <https://doi.org/10.48550/arXiv.2301.00973>

- Atwany, M. Z., Sahyoun, A. H., & Yaqub, M. (2022). Deep learning techniques for diabetic retinopathy classification: A survey. *IEEE Access*. <https://doi.org/10.1109/ACCESS.2022.3157632>
- Ayhan, M. S., K hlew in, L., Aliyeva, G., Inhoffen, W., Ziemssen, F., & Berens, P. (2020). Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection. *Medical Image Analysis*, 64, 101724. <https://doi.org/10.1016/j.media.2020.101724>
- Bahramian, M., Azimzadeh Irani, A., Pourgholi, R., & Aliyari Boroujeni, A. (2024). Recognition of English Handwritten Digit using Convolutional Neural Network (CNN). *Soft Computing Journal*. <https://doi.org/10.22052/scj.2024.243236.1021>
- Bai, Q., Liu, S., Tian, Y., Xu, T., Banegas-Luna, A. J., P rez-S nchez, H., ... & Yao, X. (2022). Application advances of deep learning methods for de novo drug design and molecular dynamics simulation. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 12(3), e1581. <https://doi.org/10.1002/wcms.1581>
- Bijalwan, V., Semwal, V. B., & Gupta, V. (2022). Wearable sensor-based pattern mining for human activity recognition: deep learning approach. *Industrial Robot: the international journal of robotics research and application*, 49(1), 21-33. <https://doi.org/10.1108/IR-09-2020-0187>
- Chen, Y., Zhu, Z., Cheng, W., Bulloch, G., Chen, Y., Liao, H., ... & Guangzhou Diabetic Eye Study group. (2023). Choriocapillaris flow deficit as a biomarker for diabetic retinopathy and diabetic macular edema: 3-year longitudinal cohort. *American journal of ophthalmology*, 248, 76-86. <https://doi.org/10.1016/j.ajo.2022.11.018>
- Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1251-1258).
- Dai, L., Wu, L., Li, H., Cai, C., Wu, Q., Kong, H., ... & Jia, W. (2021). A deep learning system for detecting diabetic retinopathy across the disease spectrum. *Nature communications*, 12(1), 1-11. <https://doi.org/10.1038/s41467-021-23458-5>
- Das, S., Kharbanda, K., Suchetha, M., Raman, R., & Dhas, E. (2021). Deep learning architecture based on segmented fundus image features for classification of diabetic retinopathy. *Biomedical Signal Processing and Control*, 68, 102600. <https://doi.org/10.1016/j.bspc.2021.102600>
- Ghoshal, B., & Tucker, A. (2022). On calibrated model uncertainty in deep learning. *arXiv preprint arXiv:2206.07795*. <https://doi.org/10.48550/arXiv.2206.07795>
- Guariguata, L., Whiting, D.R., Hambleton, I., et al. (2014). 'Global estimates of diabetes prevalence for 2013 and projections for 2035'. *Diabetes Research and Clinical Practice* 103(2), pp. 137–149. <https://doi.org/10.1016/j.diabres.2013.11.002>
- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017, July). On calibration of modern neural networks. In *International conference on machine learning* (pp. 1321-1330). PMLR.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).
- Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700-4708).
- Iman, M., Rasheed, K., & Arabnia, H. R. (2022). A review of deep transfer learning and recent advancements. *arXiv preprint arXiv:2201.09679*. <https://doi.org/10.3390/technologies11020040>
- Jing, Q., Yan, J., Lu, L., Xu, Y., & Yang, F. (2022). A novel method for pattern recognition of gis partial discharge via multi-information ensemble learning. *Entropy*, 24(7), 954. <https://doi.org/10.3390/e24070954>

- [dataset] Kaggle APTOS 2019 Blindness Detection competition, Available: <https://www.kaggle.com/c/aptos2019-blindness-detection/data>. [Accessed: 2023-06-12].
- Kim, H. E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M. E., & Ganslandt, T. (2022). Transfer learning for medical image classification: A literature review. *BMC medical imaging*, 22(1), 69. <https://doi.org/10.1186/s12880-022-00793-7>
- Kobat, S. G., Baygin, N., Yusufoglu, E., Baygin, M., Barua, P. D., Dogan, S., ... & Acharya, U. R. (2022). Automated diabetic retinopathy detection using horizontal and vertical patch division-based pre-trained DenseNET with digital fundus images. *Diagnostics*, 12(8), 1975. <https://doi.org/10.3390/diagnostics12081975>
- Kwee, A., Teo, Z. L., & Ting, D. S. W. (2022). Digital health in medicine: Important considerations in evaluating health economic analysis. *The Lancet Regional Health–Western Pacific*, 23. <https://doi.org/10.1016/j.lanwpc.2022.100476>
- Martinez-Murcia, F. J., Ortiz, A., Ramírez, J., Górriz, J. M., & Cruz, R. (2021). Deep residual transfer learning for automatic diagnosis and grading of diabetic retinopathy. *Neurocomputing*, 452, 424-434. <https://doi.org/10.1016/j.neucom.2020.04.148>
- Mazraeh, H. D., & Parand, K. (2024). GEPINN: An innovative hybrid method for a symbolic solution to the Lane–Emden type equation based on grammatical evolution and physics-informed neural networks. *Astronomy and Computing*, 48, 100846. <https://doi.org/10.1016/j.ascom.2024.100846>
- Mazraeh, H. D., & Parand, K. (2025a). A three-stage framework combining neural networks and Monte Carlo tree search for approximating analytical solutions to the Thomas–Fermi equation. *Journal of Computational Science*, 87, 102582. <https://doi.org/10.1016/j.jocs.2025.102582>
- Mazraeh, H. D., & Parand, K. (2025b). An innovative combination of deep Q-networks and context-free grammars for symbolic solutions to differential equations. *Engineering Applications of Artificial Intelligence*, 142, 109733. <https://doi.org/10.1016/j.engappai.2024.109733>
- Mazraeh, H. D. & Parand, K. (2025c). Approximate symbolic solutions to differential equations using a novel combination of Monte Carlo tree search and physics-informed neural networks approach. *Engineering with Computers*, 1-29. <https://doi.org/10.1007/s00366-025-02127-x>
- Mirzania, D., Thompson, A. C., & Muir, K. W. (2021). Applications of deep learning in detection of glaucoma: a systematic review. *European Journal of Ophthalmology*, 31(4), 1618-1642. <https://doi.org/10.1177/1120672120977346>
- Mukhamadiyev, A., Khujayarov, I., Djuraev, O., & Cho, J. (2022). Automatic speech recognition method based on deep learning approaches for Uzbek language. *Sensors*, 22(10), 3683. <https://doi.org/10.3390/s22103683>
- Oulhadj, M., Riffi, J., Chaimae, K., Mahraz, A. M., Ahmed, B., Yahyaouy, A., ... & Tairi, H. (2022). Diabetic retinopathy prediction based on deep learning and deformable registration. *Multimedia Tools and Applications*, 81(20), 28709-28727. <https://doi.org/10.1007/s11042-022-12968-z>
- Pan, C. W., Cheung, C. Y., Aung, T., Cheung, C. M., Zheng, Y. F., Wu, R. Y., ... & Saw, S. M. (2013). Differential associations of myopia with major age-related eye diseases: the Singapore Indian Eye Study. *Ophthalmology*, 120(2), 284-291.
- Pandey, M., Fernandez, M., Gentile, F., Isayev, O., Tropsha, A., Stern, A. C., & Cherkasov, A. (2022). The transformational role of GPU computing and deep learning in drug discovery. *Nature Machine Intelligence*, 4(3), 211-221. <https://doi.org/10.1038/s42256-022-00463-x> <https://doi.org/10.1016/j.opthta.2012.07.065>

- Pao, S. I., Lin, H. Z., Chien, K. H., Tai, M. C., Chen, J. T., & Lin, G. M. (2020). Detection of diabetic retinopathy using bichannel convolutional neural network. *Journal of Ophthalmology*. <https://doi.org/10.1155/2020/9139713>
- Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3), 61-74.
- Prasath, A. (2022). Design of an integrated learning approach to assist real-time deaf application using voice recognition system. *Computers and Electrical Engineering*, 102, 108145. <https://doi.org/10.1016/j.compeleceng.2022.108145>
- Roßbach, J., Kollmeier, B., & Meyer, B. T. (2022). A model of speech recognition for hearing-impaired listeners based on deep learning. *The Journal of the Acoustical Society of America*, 151(3), 1417-1427. <https://doi.org/10.1121/10.0009411>
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. <https://doi.org/10.48550/arXiv.1409.1556>
- Suganyadevi, S., Seethalakshmi, V., & Balasamy, K. (2022). A review on deep learning in medical image analysis. *International Journal of Multimedia Information Retrieval*, 11(1), 19-38. <https://doi.org/10.1007/s13735-021-00218-1>
- Sugeno, A., Ishikawa, Y., Ohshima, T., & Muramatsu, R. (2021). Simple methods for the lesion detection and severity grading of diabetic retinopathy by image processing and transfer learning. *Computers in Biology and Medicine*, 137, 104795. <https://doi.org/10.1016/j.combiomed.2021.104795>
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9).
- Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., & Liu, C. (2018). A survey on deep transfer learning. In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings*, Part III 27 (pp. 270-279). Springer International Publishing. https://doi.org/10.1007/978-3-030-01424-7_27
- Tan, M., & Le, Q. (2019). EfficientNet, Rethinking model scaling for convolutional neural networks. In: *International Conference on Machine Learning, PMLR*, pp. 6105–6114.
- Tian, J., Smith, G., Guo, H., Liu, B., Pan, Z., Wang, Z., & Fang, R. (2021). Modular machine learning for Alzheimer's disease classification from retinal vasculature. *Scientific Reports*, 11(1), 1-11. <https://doi.org/10.1038/s41598-020-80312-2>
- Tsuneki, M. (2022). Deep learning models in medical image analysis. *Journal of Oral Biosciences*. <https://doi.org/10.1016/j.job.2022.03.003>
- Uhm, K. H., Jung, S. W., Choi, M. H., Shin, H. K., Yoo, J. I., Oh, S. W., ... & Ko, S. J. (2021). Deep learning for end-to-end kidney cancer diagnosis on multi-phase abdominal computed tomography. *NPJ precision oncology*, 5(1), 54. <https://doi.org/10.1038/s41698-021-00195-y>
- Van der Velden, B. H., Kuijf, H. J., Gilhuijs, K. G., & Viergever, M. A. (2022). Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis*, 102470. <https://doi.org/10.1016/j.media.2022.102470>
- Wen, L. H., & Jo, K. H. (2022). Deep learning-based perception systems for autonomous driving: A comprehensive survey. *Neurocomputing*. <https://doi.org/10.1016/j.neucom.2021.08.155>
- Ye, T., Si, S., Wang, J., Cheng, N., & Xiao, J. (2022). Uncertainty Calibration for Deep Audio Classifiers. *arXiv preprint arXiv:2206.13071*. <https://doi.org/10.48550/arXiv.2206.13071>

- Zadrozny, B., & Elkan, C. (2001, June). Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Icml, 1*, pp. 609-616.
- Zadrozny, B., & Elkan, C. (2002, July). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 694-699). <https://doi.org/10.1145/775047.775151>
- Zago, G. T., Andreão, R. V., Dorizzi, B., & Salles, E. O. T. (2020). Diabetic retinopathy detection using red lesion localization and convolutional neural networks. *Computers in biology and medicine*, 116, 103537. <https://doi.org/10.1016/j.compbimed.2019.103537>
- Zaidi, S. S. A., Ansari, M. S., Aslam, A., Kanwal, N., Asghar, M., & Lee, B. (2022). A survey of modern deep learning based object detection models. *Digital Signal Processing*, 103514. <https://doi.org/10.1016/j.dsp.2022.103514>
- Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, 3(1), 1-40. <https://doi.org/10.1186/s40537-016-0043-6>

