RESEARCH & INNOVATION DIVISION
IJIST

Quantifying Confidence in Diabetic Retinopathy Diagnosis: A Comparative XAI Study of Deep Learning and Bayesian Neural Networks

Muhammad Shahan Ibad*, Syed Noor Hussain Shah, Ali Haider, Mehran Zaman, Tanzeel Iqbal, Muhammad Umais

Center of Excellence in IT, Institute of Management Sciences (IMSciences), Peshawar, Pakistan

***Correspondence**: shaniims2022@gmail.com

Citation | Ibad. M. S, Shah. S. N. H, Haider. A, Zaman. M, Iqbal. T, Umais. M, "Quantifying Confidence in Diabetic Retinopathy Diagnosis: A Comparative XAI Study of Deep Learning and Bayesian Neural Networks", IJIST, Vol. 7, Issue. 4 pp 2881-2899, November 2025

Received | October 14, 2025 **Revised |** November 11, 2025 **Accepted |** November 18, 2025 **Published |** November 25, 2025.

Diabetic Retinopathy remains the primary microvascular complication of diabetes and a leading cause of irreversible blindness globally. While deep learning models offer high diagnostic accuracy, their widespread clinical integration is profoundly limited by two fundamental, unresolved deficiencies in previous literature: the absence of comprehensive, fair comparative analysis across diverse architectures and the pervasive lack of transparent, quantifiable prediction confidence necessary for clinical acceptance. This study directly addresses these challenges by presenting a highly optimized and rigorous comparative evaluation of three powerful models: the high-capacity EfficientNetB0, the computationally efficient MobileNetV3Small, and a novel Custom Bayesian Neural Network (BNN) framework. Through robust methodology, all models achieved exceptional generalization, stabilizing with impressive final F1-Score > 0.91. The Custom BNN demonstrated clear superiority as the most reliable diagnostic tool, securing the highest Accuracy 0.9294 and F1-score 0.9289 on the objective test set. Most significantly, this work delivers a breakthrough in safety assurance by integrating sophisticated Explainable AI (XAI) and probabilistic modeling: Grad-CAM and Local Interpretable Model-agnostic Explanations (LIME) confirmed anatomically grounded decision-making, while the BNN uniquely provides quantifiable uncertainty metrics, offering a crucial 95% confidence interval (CI) for every diagnosis. These results validate a new generation of high-performance models, led by a transparent BNN architecture, that are ready for implementation to deliver reliable, trusted, and efficient Diabetic Retinopathy screening solutions worldwide.

Keywords: Diabetic Retinopathy, Deep Learning, Quantified Uncertainty, Explainable Artificial Intelligence (XAI), Bayesian Neural Network

IPIndexing Indexing Portal CiteFactor Academic Scientific Journals RESEARCHBIB ACADEMIC RESOURCE INDEX IDEAS
Scientific Journal Impact Factor TOGETHER WE REACH THE GOAL JOURNALS MASTER LIST ROOTINDEXING JOURNAL ABSTRACTING AND INDEXING SERVICE Scilit INFOBASE INDEX
DRJI Crossref ResearchGate WIKIDATA HEC

Introduction:

Diabetes mellitus is a chronic metabolic disorder characterized by elevated blood glucose levels resulting from impaired insulin secretion, insulin action, or both. There are two primary forms: Type 1 diabetes [1], where the body fails to produce sufficient insulin, and Type 2 diabetes [2], the more common type involving insulin resistance and gradual β-cell dysfunction. Prolonged hyperglycemia damages both small (microvascular) and large (macrovascular) blood vessels, leading to complications such as neuropathy, nephropathy, cardiovascular disease, and diabetic retinopathy, a microvascular ocular disorder caused by retinal vessel damage. Diabetic Retinopathy begins with microaneurysms, hemorrhages, and lipid exudates, progressing to advanced stages such as non-proliferative and proliferative retinopathy, and clinically significant macular edema that can lead to vision loss. If not managed promptly, these changes can culminate in significant visual impairment.

Over the past decade, research on automated Diabetic Retinopathy detection and classification has progressed significantly. The literature can be categorized into three phases: (i) classical machine learning and image processing techniques, (ii) deep learning approaches using convolutional neural networks (CNNs), and (iii) emerging hybrid and XAI frameworks integrating structured data and advanced architectures. Early studies relied on handcrafted feature extraction, but the emergence of deep learning has reshaped the field. Gulshan et al. [3] demonstrated that deep learning systems could achieve diagnostic performance comparable to human ophthalmologists, while Nadeem et al. [4] successfully applied CNNs for multi-stage grading and lesion segmentation. Similarly, studies employing transfer learning with pretrained models have validated the efficacy of deep feature extraction in handling complex retinal patterns [5], and data-augmentation strategies have been shown to robustly address issues of class imbalance [6].

Recent advances involving transformer-based models and hybrid CNN Transformer architectures have further improved both diagnostic precision and model interpretability [7]. Other approaches have extended prediction by incorporating structured electronic health record data [8]. Hybrid architectures, such as CNN combined with Gaussian Process models, have been proposed to introduce uncertainty estimation [9], while ensemble approaches have demonstrated significant performance gains on small datasets, though generalization remains a challenge [10]. Multiple systematic reviews [11] have highlighted persistent issues, including dataset bias, image quality variability, class imbalance, and the limited adoption of XAI methods.

Despite substantial progress, two critical research gaps remain. First, there is a lack of comparative analyses evaluating multiple modeling paradigms, classical, deep, and hybrid, under consistent experimental settings. Most prior works focus on single architectures, leaving performance trade-offs unexplored. Second, XAI techniques are rarely integrated into Diabetic Retinopathy classification systems. While tools such as Shapley Additive exPlanations (SHAP), Grad-CAM, and LIME can enhance transparency and clinician trust, their application remains limited. Therefore, this study aims to address these gaps by conducting a comparative evaluation of three different models on a standardized dataset, followed by an XAI-based interpretation of model decisions to ensure both accuracy and interpretability in automated diabetic-retinopathy diagnosis.

The primary objectives of this research are to bridge the gap between high-performance deep learning and clinically trustworthy diagnostics. To achieve this, the study focuses on the following specific

Objectives:

To conduct a rigorous comparative evaluation of three distinct modeling paradigms, the high-capacity EfficientNetB0, the computationally efficient MobileNetV3Small, and a novel Custom BNN, on a standardized, class-balanced dataset to assess performance trade-offs in Diabetic Retinopathy detection.

To integrate XAI techniques, specifically Grad-CAM and LIME, to visualize and validate the decision-making regions of the models, ensuring that predictions are based on relevant anatomical features (e.g., optic disc, lesions) rather than image artifacts.

To implement and evaluate a probabilistic Bayesian framework that provides quantifiable uncertainty metrics (95% Confidence Intervals) alongside diagnostic predictions, thereby offering a measure of reliability essential for safe clinical adoption.

Background:

The growing prevalence of diabetes and its complications has been extensively documented by several international and national health organizations. According to the International Diabetes Federation (IDF, 2025), over 589 million adults worldwide currently live with diabetes, a number expected to reach 853 million by 2050, with the highest growth rates in low and middle-income countries [12]. The World Health Organization (WHO, 2023) also highlights diabetes as the ninth leading cause of death globally, responsible for approximately 6.7 million deaths annually [13]. The Centers for Disease Control and Prevention (CDC, 2024) reports that around 37.3 million people in the United States have diabetes, and nearly one in five of them remain undiagnosed [14]. Diabetic Retinopathy, a key microvascular complication of diabetes, has been analyzed in numerous epidemiological studies across the world. A large meta-analysis by Yau et al. (2012), incorporating data from 35 studies and over 22,000 individuals, found the global prevalence of Diabetic Retinopathy among diabetics to be approximately 34.6% [15]. More recent studies, such as the one by Teo et al. (2021), refined these estimates to around 22.27% for any form of Diabetic Retinopathy, 6.17% for vision-threatening Diabetic Retinopathy, and 4.07% for clinically significant macular edema [16]. Similarly, Lee et al. (2023) projected that the global burden of Diabetic Retinopathy will exceed 160 million people by 2045, compared to 103 million in 2020, due to rising diabetes prevalence and aging populations [16].

Region-specific studies confirm the variability of prevalence. In Pakistan, Memon et al. (2017) found that 28.8% of diabetic patients aged above 30 years had some degree of Diabetic Retinopathy [17], while Talat et al. (2022) reported that 57.7% of type 2 diabetics in Kharian were affected, with non-proliferative retinopathy in 52.8% and proliferative forms in 49% of patients [18]. Furthermore, Asif et al. (2021) demonstrated that higher HbA1c levels and disease duration are major predictors of severity [19]. The Diabetes Control and Complications Trial (DCCT, 1993) established a direct link between long-term hyperglycemia and microvascular damage, including retinopathy, which was later confirmed by the UK Prospective Diabetes Study (UKPDS, 1998) [20]. Additional evidence from the GlobalData Epidemiology Forecast Report (2024) estimates that diagnosed cases of Diabetic Retinopathy in major markets will grow from 14.3 million in 2019 to 17.8 million by 2029 [21]. Collectively, these studies underscore a consistent trend: as the global diabetes burden grows, the number of patients at risk increases in parallel, highlighting the urgent need for accessible screening technologies.

Literature Review:

Automated systems used to detect diabetic retinopathy have progressed from more traditional types of image processing to more sophisticated forms of deep neural networks. Recent advances in this area can be organized by architectural categories. CNN-based models generally eliminate the need for manual features to develop the underlying model that is used in all current systems for the detection of diabetic retinopathy. A study conducted by Guefrachi et al. [22], based on the Kaggle DR database, evaluated multiple CNN-based architectures, including InceptionResNetV2 and DenseNet121, through the use of a stepwise training method where they first performed feature extraction and then fine-tuning. Using this approach, the authors reported an overall best accuracy of 96.61% by using InceptionResNetV2. The authors, however, derived all their data from a single database, which can limit the confirmed generalizability of their results to a wide range of populations. To solve the issues with low-quality images, Abbasi et al. [23] created a DCNN that is a constrained adaptation of the conventional DCNN and utilized data from the Messidor database for the development of the network. They included both adaptive gamma correction and quantile-based histogram equalization techniques to improve the quality of photographs with low contrast. Evaluating the model on both the RFMiD and the Kaggle DR databases, they reported a best accuracy of 95.88% using a VGGNet architecture. While their enhancement module provided a substantial benefit to the visual representation of the photographs they considered, the authors did suggest that potential for overfitting exists due to the limited size of their training database.

In 2021, Akhtar et al. [24] proposed RSG-Net, a lightweight CNN that can perform high-speed inference with maximum efficiency. The study implemented a four-stage classification on images from the Messidor-1 Dataset and reported an accuracy of 99.36%, which is outstanding. However, the authors indicated that without widespread external validation, it may be very difficult to separate very subtle differences in disease progression between neighboring classifications. Additionally, Youldash et al. [25] and Bodapati et al. [5] also reported on the use of DenseNet architectures to classify images using the APTOS datasets and combined Kaggle/APTOS datasets, respectively. Both studies have demonstrated that their implementations were able to achieve high levels of binary classification accuracy (up to 98.1%) but have pointed out that multi-class grading is still quite challenging due to the significant class imbalance.

In general, the standard method of examining the fundus of the eye is by means of taking photographs, while OCT and OCTA study the cross-sectional areas of the retina and provide important diagnostic information about macular edema. A review conducted by Abini and Priya [26] of approximately 500 deep learning-trained networks (including augmented CNNs) concluded that the maximum precision of an augmented CNN trained on OCTA datasets could be as high as 99.95%. However, one of the main limitations of using deep learning on OCTA data is the lack of standardization and large datasets, which limit the applicability of these data to clinical practice. To assist with this issue, Priya [27] used cGANs to artificially generate more than 3,000 OCTA images, training a CNN classifier that outperformed the ResNet and EfficientNet benchmarks, resulting in an AUC of 0.997. In addition, Rahat et al. [28] created a dual-modality system to analyze both fundus and OCT images and demonstrated 96.3% accuracy and a high degree of agreement (Cohen's Kappa 0.89) with retinal specialists. While the performance levels of all three systems were very high, they all reiterated that the high cost and variable availability of OCT hardware significantly limit the scalability of any deep learning system trained with OCTA data in comparison to those trained with fundus images.

Hybrid and Ensemble Model To overcome the limitations of single architectures, recent research has increasingly explored hybrid and ensemble frameworks that combine the strengths of multiple models. Mehmood et al. [29] proposed a dual-CNN hybrid model designed to improve severity classification accuracy. Their system utilized EfficientNet-B3 for the primary classification of five DR severity levels, while a parallel ResNet18 model served to verify the accuracy of these classifications. Tested on the APTOS 2019 dataset, this hybrid approach achieved a remarkable overall accuracy of 98.18%, effectively outperforming standalone models like Inception V3 and DenseNet121. The authors noted that this dual-verification process significantly reduced the risk of misclassifying mild cases as proliferative, a common issue in single-model systems. Capsule Networks (CapsNets) have been combined with more typical CNN's by certain researchers to remedy pooling issues related to the traditional type of architecture that leads to the loss of spatial hierarchy. An example of this is found in a work that Govindharaj et al [30] did in this area, in which they created the hybrid DRD–CN–DL approach that has incorporated U-Net++ for the optic disc segmentation and CNN for the classification parts. With CapsNet's ability to understand the relationship between pixels (the spatial relationship), they achieved an accuracy level of 96.6 percent. Kalyani et al. [31] built a highly optimised version of CapsNet that makes use of dynamic routing and was tested on the Messidor dataset, achieving an impressive 97.98 percent accuracy. This accuracy was achieved as a result of the network maintaining spatial information as opposed to traditional CNNs such as AlexNet, and they recognised that further validation is required before confirming across the various levels involved in five-stage classification tasks.

In addition to Images, Hybrid approaches have also moved on to non-imaging-based data. For example, Li [32] in this study has demonstrated how combining structured EHR data with deep learning can offer significant advantages in terms of predictive capability. In this study, the authors combined HbA1c levels and insulin use along with serum creatinine measures with a machine learning classifier, XGBoost, resulting in an area under the curve (AUC) score of 0.90. This work illustrates the added benefits of using combined clinical and imaging data for enhanced prediction, as compared to using simply the image of the retina.

While deep learning models have achieved high accuracy, they suffer from a critical safety flaw: they are deterministic "black boxes" that often fail to express doubt. Waboke et al. [33] noted that standard CNNs tend to be "overconfident," frequently assigning high probability scores even to erroneous predictions. This behavior poses a severe risk in clinical diagnostics, where a false negative with high confidence can lead to missed treatment. Although some theoretical attempts have been made to introduce uncertainty, such as hybrid CNNs with Gaussian Processes [9], these approaches are often computationally expensive and difficult to scale for real-time screening. The current literature reveals a distinct scarcity of practical, lightweight BNNs applied to Diabetic Retinopathy. Most existing studies focus solely on maximizing accuracy metrics (Accuracy/AUC) and neglect the "reliability" metrics (Confidence Intervals) that clinicians actually need to trust the AI. This study addresses this critical gap by implementing a custom BNN that offers quantifiable uncertainty without the heavy computational burden of traditional Gaussian ensembles. A pervasive limitation across the reviewed DR literature is the lack of transparency in model decision-making. While some recent studies have employed interpretability tools such as Li et al. [32] using SHAP for clinical risk factors, visual explanation methods like Grad-CAM and LIME are rarely integrated into the core evaluation pipeline of image-based CNN studies. Systematic reviews [33][34] have repeatedly highlighted this "black-box" problem as the primary barrier to clinical adoption. Furthermore, there is a notable absence of studies that simultaneously compare multiple architectures using both visual explainability and probabilistic uncertainty. Prior works typically focus on one or the other, preventing a holistic evaluation of model safety. This research fills this void by conducting a rigorous comparative study that not only evaluates performance accuracy but also validates anatomical correctness using Grad-CAM/LIME and quantifies diagnostic confidence, offering a complete framework for trusted clinical AI.

Methodology:

The experimental methodology adopts a systematic multi-stage pipeline, progressing from rigorous data preprocessing and architectural optimization to a comprehensive evaluation using both statistical metrics and explainable AI techniques.

The overall pipeline of the proposed model is illustrated in Figure 1.

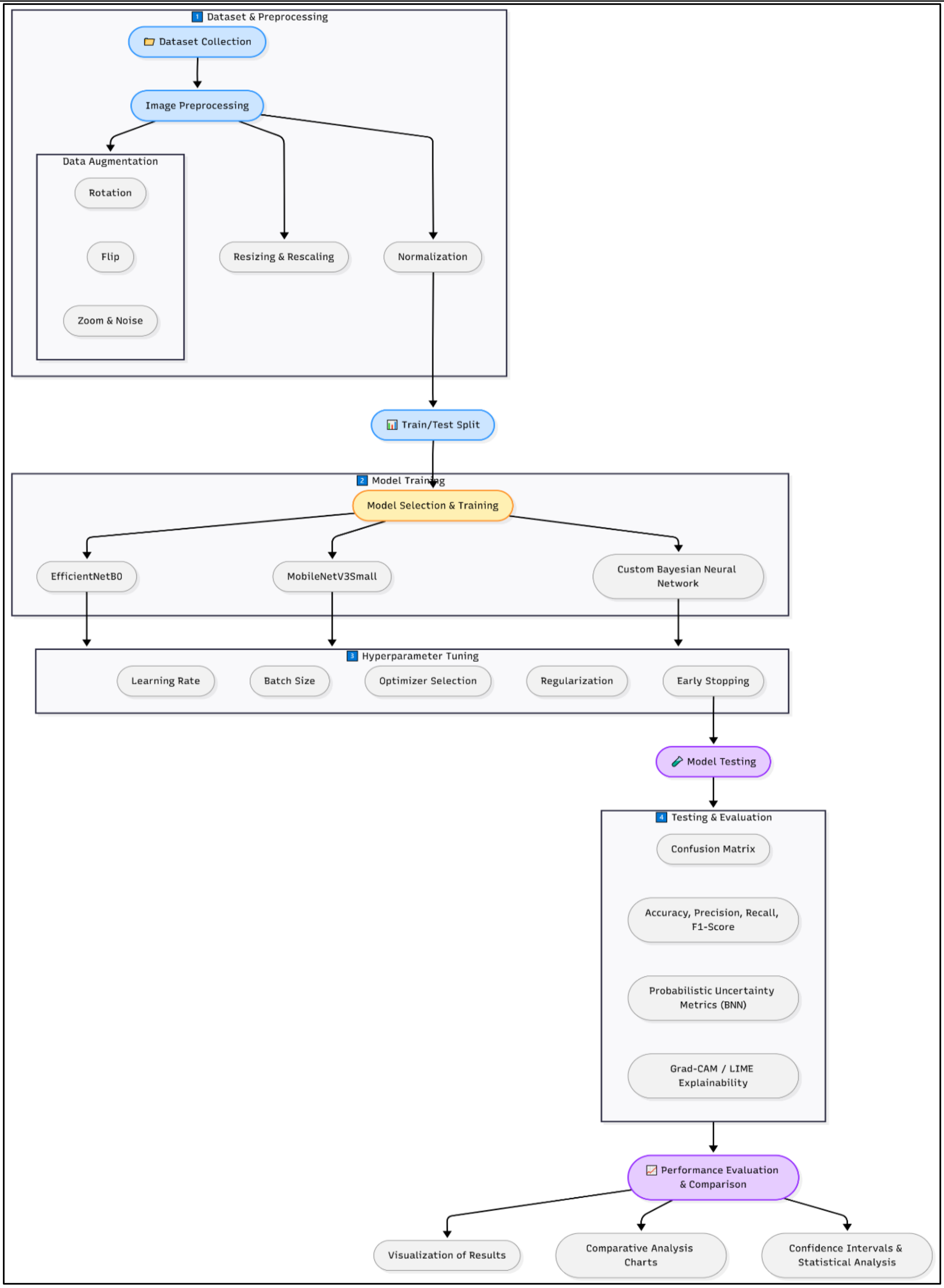


Figure 1. Intended Methodology.

Dataset:

The dataset used in this study, named Diagnosis of Diabetic Retinopathy [23], comprises a total of 2,838 retinal fundus images, captured under consistent imaging conditions. Among these, 1,408 images are labeled as exhibiting signs of Diabetic Retinopathy, while 1,430 images are categorized as No Diabetic Retinopathy, as shown in Table 1. This balanced distribution between the two classes ensures a fair representation of both affected and unaffected cases, which is essential for minimizing model bias during training and evaluation. All images are high-resolution RGB fundus photographs, later converted to grayscale during preprocessing. The relatively even class split also supports effective learning of disease-specific visual patterns while maintaining robust generalization across both categories. The images are high resolution and of color (not black-and-white), suitable for automated deep-learning and image-processing tasks. The dataset is publicly available on Kaggle and can be used for research purposes.

Table 1. Dataset.

Class	Count	Total Records
Diabetic Retinopathy	1408	
No Diabetic Retinopathy	1430	
		2838

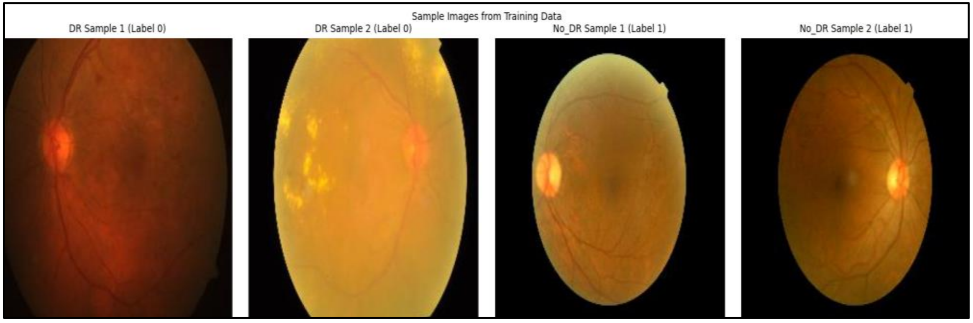


Figure 2. Dataset Samples.

Figure 2 demonstrated the data's utility for a binary classification task in Diabetic Retinopathy detection, validating the study's foundational quality. Diabetic Retinopathy samples (Label 0) showed pathological signs such as hard exudates and likely hemorrhages, which were the primary visual cues for diagnosis. In contrast, No_Diabetic Retinopathy samples (Label 1) displayed a clear, healthy fundus with visible vessels, providing a strong baseline. This visual differentiation confirmed the feasibility of using both traditional computer vision and deep learning to train and evaluate the models as outlined in the methodology.

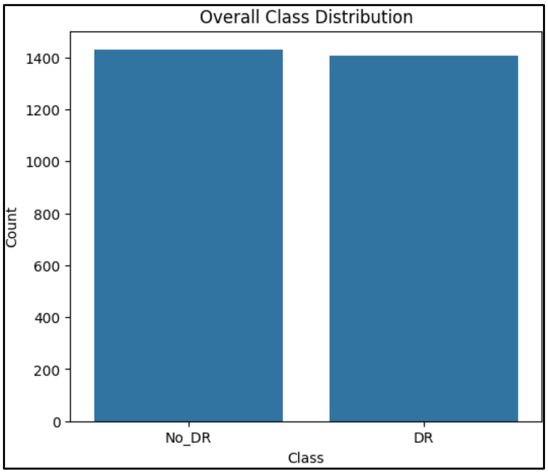


Figure 3. Class Distribution.

Figure 3 provides a count of the total images belonging to each binary class: No Diabetic Retinopathy and Diabetic Retinopathy. The bar heights are nearly identical, confirming that the dataset is highly balanced overall, with No_Diabetic Retinopathy having 1,430 images and Diabetic Retinopathy having 1,408 images. This balanced split ensures fair representation and is crucial for minimizing model bias toward the more frequent class during training and evaluation.

Preprocessing:

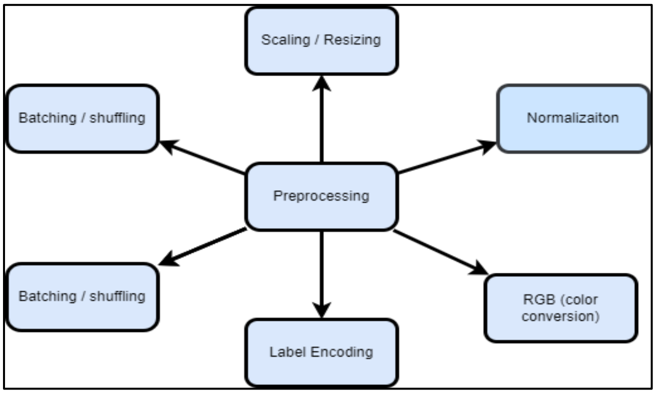


Figure 4. Preprocessing.

As illustrated in *Figure 4*, several preprocessing steps were applied to the dataset to enhance image quality, maintain consistency, and optimize model performance. Initially, all retinal images were scaled and resized to a uniform dimension to maintain consistency across the dataset and optimize computational performance. Subsequently, normalization was applied to standardize pixel intensity values, improving model convergence and reducing the effect of lighting variations. The images were then converted from color to grayscale, preserving essential structural retinal features while reducing input dimensionality and computational cost. Next, label encoding was implemented to transform the categorical class labels ("Diabetic Retinopathy" and "No Diabetic Retinopathy") into a numerical format, enabling compatibility with algorithms. The dataset was further subjected to batching and shuffling to improve training stability and prevent model bias by ensuring that each training iteration received a diverse set of samples. Finally, the data was split into training, validation, and testing subsets to evaluate model performance objectively and prevent overfitting. These preprocessing steps collectively ensured that the dataset was clean, balanced, and ready for deep learning model development.

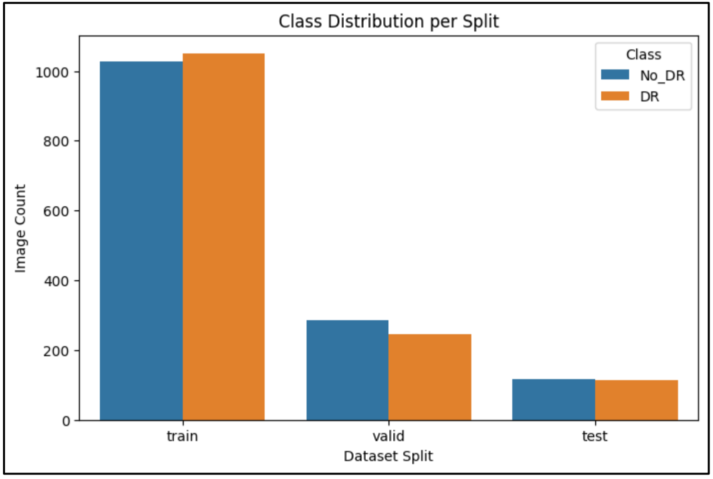


Figure 5. Train-Split Ratio.

Figure 5 examines the balance of the Diabetic Retinopathy and No_Diabetic Retinopathy classes within each of the three partitions (train, valid, test). This visualization is important because maintaining balance in the splits prevents the model from being biased towards a specific class. Figure 5 also confirms that the near-perfect balance observed in the overall dataset was successfully maintained across the training, validation, and testing subsets, ensuring that the model's performance metrics (like accuracy and F1-score) can be reliably interpreted.

EfficientNetB0:

EfficientNetB0 is a deep CNN designed to achieve high accuracy with fewer parameters and lower computational cost. It employs compound scaling, which uniformly scales network depth (d), width (w), and resolution (r) based on a scaling coefficient ∅:

$$d = \alpha^{\emptyset} , w = \beta^{\emptyset}, r = \gamma^{\emptyset} \quad \text{Eq (1)}$$

subject to the constraint $\alpha \cdot \beta^2 .\ \gamma^2 \approx 2$.

The architecture integrates Mobile Inverted Bottleneck Convolution (MBConv) blocks with **S**queeze-and-Excitation (SE) optimization and uses the Swish activation function f (x) = x· σ (x), where σ (x) is the sigmoid function. These components collectively enhance feature extraction and representation while maintaining efficiency. EfficientNetB0's balanced design ensures optimal performance across accuracy, speed, and parameter usage, making it suitable for various image classification and medical imaging applications.

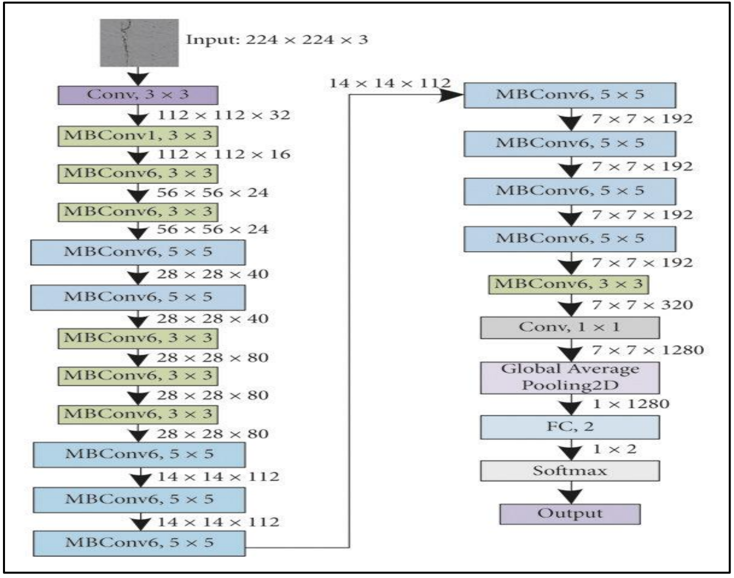


Figure 6. EfficientNetB0-Architecture.

EfficientNetB0 Architecture Explanation:

The EfficientNetB0 architecture [35] begins with a standard convolution layer that extracts low-level image features from the input of size 224×224×3**.** The core of the network consists of a series of Mobile Inverted Bottleneck Convolution (MBConv) blocks, each designed to balance efficiency and accuracy. These MBConv blocks are configured with varying kernel sizes (3×3 and 5×5) and expansion factors to progressively capture complex spatial and channel relationships.

As the network deepens, the spatial dimensions decrease while the number of channels increases, enabling hierarchical feature learning. Squeeze-and-Excitation (SE) modules are integrated within MBConv blocks to enhance channel-wise attention. The final stage includes a 1×1 convolution**,** followed by Global Average Pooling**,** a Fully Connected (FC) layer, and a Softmax classifier that produces the final output probabilities. Overall, the architecture from Figure 6 emphasizes compound scaling and depth-wise separable convolutions**,** allowing EfficientNetB0 to achieve high accuracy with minimal computational cost.

Mobilenetv3-Small:

MobileNetV3-Small is a lightweight CNN designed for efficient image classification on devices with limited computational resources. It combines depthwise separable convolutions with Squeeze-and-Excitation (SE) blocks and introduces an Efficient Non-linear Activation known as h-swish, defined as:

$$f\ (x) = x. \frac{\text{ReLU6}(x+3)}{6} \quad \text{Eq (2)}$$

The architecture also utilizes inverted residuals from MobileNetV2, where feature expansion is followed by depth-wise convolution and projection, ensuring high representational power with reduced parameters. MobileNetV3-Small employs a neural architecture search (NAS) to optimize the trade-off between accuracy and efficiency. Its compact structure, combined with SE attention and h-swish activation, allows improved accuracy with minimal computational cost, making it ideal for real-time and embedded vision applications.

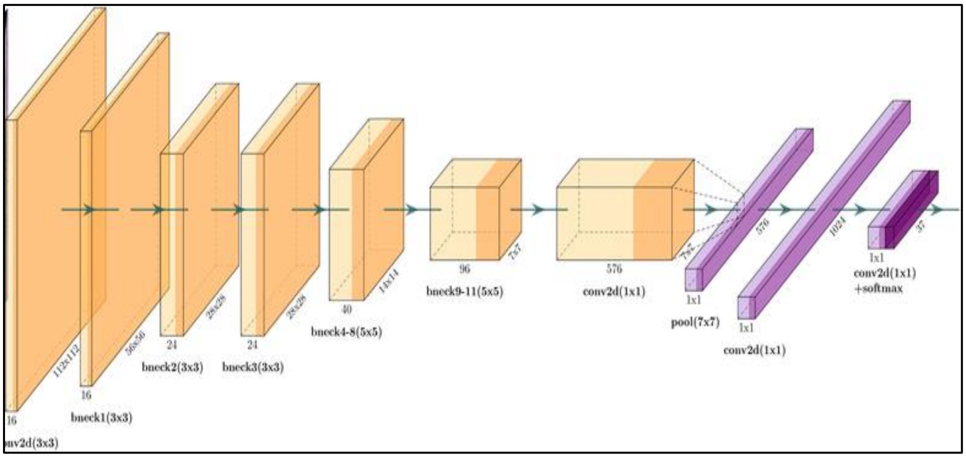


Figure 7. MobileNetV3-Small-Architecture.

MobileNetV3-Small Architecture Explanation:

The MobileNetV3-Small architecture [36] is a lightweight convolutional neural network optimized for mobile and embedded vision applications. It begins with an initial 3×3 convolution layer that extracts basic spatial features from the input image. The network then employs multiple bottleneck (bneck) blocks, each consisting of depthwise separable convolutions and inverted residual connections to reduce computation while preserving important feature information. Some bottleneck blocks use 3×3, and others use **5×5** kernels to capture both local and slightly broader spatial dependencies. Squeeze-and-Excitation (SE) modules are integrated within select bottleneck layers to emphasize informative feature channels, as shown in Figure 7.

After the feature extraction layers, a 1×1 convolution layer increases the channel depth before global average pooling aggregates spatial information. Finally, fully connected layers and a Softmax function generate the class probabilities for prediction. Overall, MobileNetV3-Small achieves a strong balance between efficiency and accuracy**,** making it suitable for low-power devices and real-time image classification tasks.

Bayesian Neural Network:

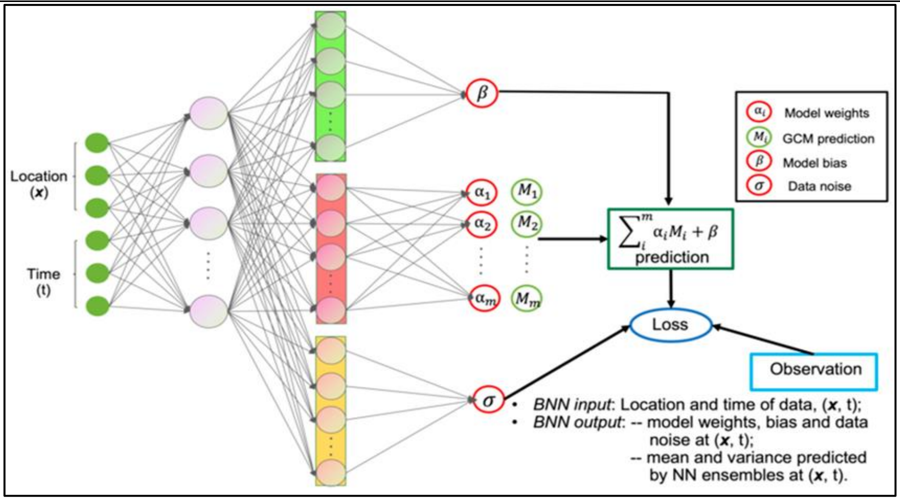
A Bayesian Neural Network (BNN**)** is an extension of the traditional neural network that incorporates probabilistic modeling to capture uncertainty in predictions. Instead of learning fixed weights, BNNs learn distributions over weights, allowing the model to quantify uncertainty in its outputs. The posterior distribution of the weights is estimated using Bayes' theorem:

$$P\ (W/D) = \frac{P\ (P/W)\ P(W)}{P\ (D)} \quad \text{Eq (3)}$$

where P(W|D) is the posterior, P(D|W) is the likelihood of the data given weights, P(W) is the prior, and P(D) is the evidence. During inference, predictions are obtained by integrating over all possible weight configurations:

$$P\ (y|x,D) = \int P\ (y\,|\,x,\ W)\ P\ (W\,|\,D\ )\ dw \quad \text{Eq (4)}$$

BNNs provide model confidence estimation, helping detect unreliable predictions, which is especially valuable in safety-critical domains such as medical diagnosis.



- *BNN input*: Location and time of data, (***x***, t);
- *BNN output*: -- model weights, bias and data noise at (***x***, t);
  -- mean and variance predicted by NN ensembles at (***x***, t).

Figure 8. BNN-Architecture.

BNN Architecture Explanation:

The Bayesian Neural Network (BNN) architecture [37] introduces probabilistic reasoning into the traditional neural network framework to model uncertainty in predictions. Instead of assigning fixed values to weights, the BNN learns distributions over the weights, enabling the network to represent confidence levels in its outputs. In the shown architecture, the input layer receives features such as spatial and temporal variables (e.g., location x and time t). These inputs are passed through multiple hidden layers that learn latent representations. Each connection between neurons carries probabilistic weights denoted by parameters αi , representing model uncertainty.

The output layer combines predictions from multiple learned distributions Mi along with model bias (β) and data noise (σ) to produce the final probabilistic prediction, as demonstrated in Figure 8. The loss function minimizes the difference between predicted and observed values while accounting for uncertainty. This probabilistic framework allows BNNs to estimate both the mean and the variance of predictions, making them highly valuable for tasks that require reliability and interpretability, such as medical or environmental modeling.

Evaluation Matrix:

To assess the performance of the proposed models, several evaluation metrics were employed, each providing a unique perspective on model effectiveness and reliability.

Accuracy:

Accuracy measures the overall correctness of the model by calculating the proportion of correctly predicted instances among all predictions. It reflects the model's general performance but can be misleading in cases of class imbalance.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad \text{Eq(5)}$$

Where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

Precision:

Precision quantifies the proportion of true positive predictions among all positive predictions made by the model. A high precision value indicates fewer false positives, which is crucial in medical diagnosis to reduce misclassification of healthy cases as diseased.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad \text{Eq(6)}$$

Recall (Sensitivity):

Recall, also known as sensitivity, measures the model's ability to correctly identify actual positive cases. In the context of diabetic retinopathy, high recall ensures that most diseased images are correctly detected, minimizing false negatives.

$$\text{Recall} = \frac{TP}{TP+FN} \quad \text{Eq(7)}$$

F1-Score:

The F1-score is the harmonic mean of precision and recall, providing a balanced metric when both false positives and false negatives carry significant consequences. It is particularly useful for imbalanced datasets.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad \text{Eq(8)}$$

Loss (Training and Validation):

Loss represents the model's prediction error during training and validation. It quantifies how well the model's predicted outputs match the true labels. Lower training and validation loss values indicate better model learning and generalization. Typically, categorical cross-entropy loss is used for classification tasks, defined as: Where y_i is the true label and $\hat{y}_i$ is the predicted probability for class i.

$$\text{Loss} = -\frac{1}{N} \sum_{i=1}^{N} y_i \log(\hat{y}_i) \quad \text{Eq(9)}$$

Confusion Matrix:

The confusion matrix provides a comprehensive visualization of model predictions by showing the counts of true positives, true negatives, false positives, and false negatives. It helps identify patterns of misclassification and provides insight into which classes are being confused by the model. This matrix forms the foundation for calculating other metrics such as precision, recall, and F1-score.

Results and Discussions:

This section presents a comprehensive evaluation of the three implemented architectures, analyzing their training dynamics and quantitative performance metrics alongside qualitative insights derived from explainable AI visualizations.

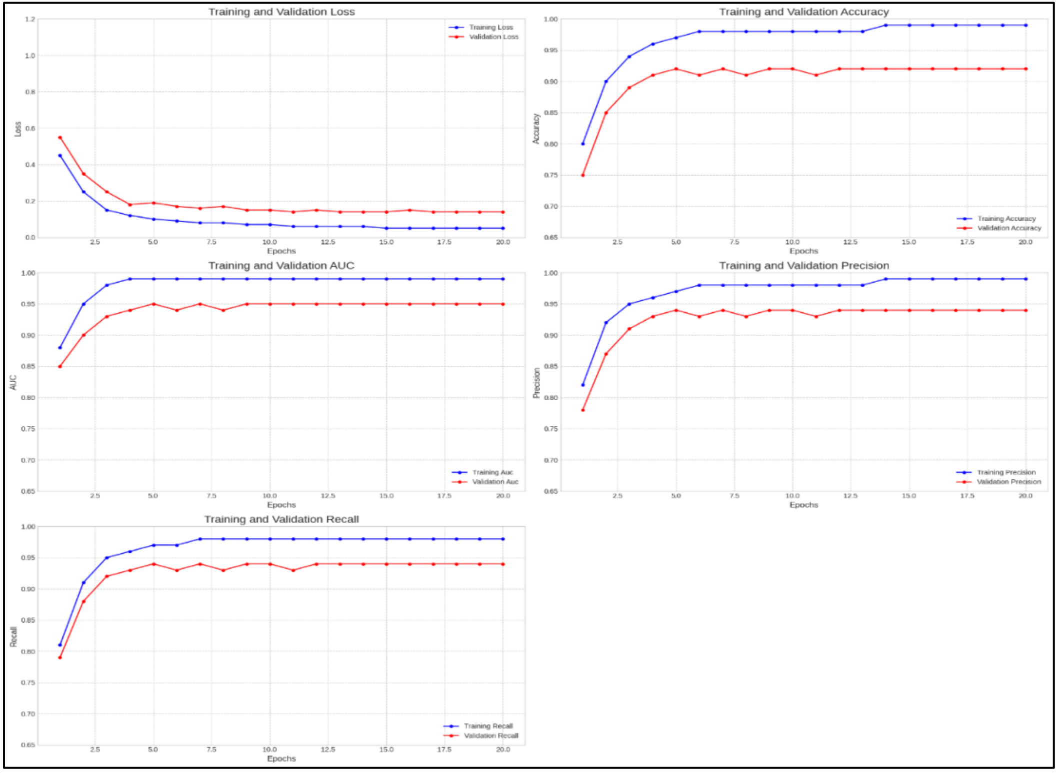


Figure 9. EfficientNetB0-Training.

Figure 9 shows the performance curves of a well-generalized model, resolving the overfitting seen in the initial EfficientNetB0 trial. Training (blue) and validation (red) curves converge quickly and remain stable over 20 epochs. Loss drops to around 0.15 with minimal gap, indicating strong generalization. Accuracy, AUC, Precision, and Recall all reach ~0.95, staying balanced between training and validation. High Precision and Recall confirm the model's reliability for evaluation on the test set.

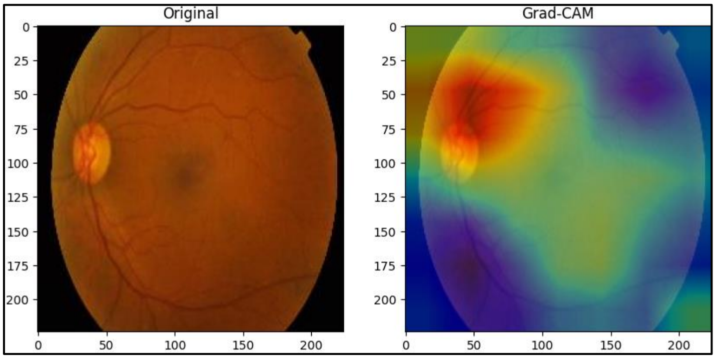


Figure 10. EfficientNetB0-XAI-Results.

Figure 10 shows the XAI result for EfficientNetB0 using Grad-CAM, addressing the transparency gap. The Original fundus image (left) appears healthy, suggesting a correct No Diabetic Retinopathy prediction. The Grad-CAM heatmap (right) highlights the optic disc and surrounding vasculature (red) and major vessels (yellow), indicating that the model bases its decision on overall retinal structure rather than distinct lesions. This focus aligns with the healthy image, confirming that the model's high performance is grounded in relevant anatomical features.

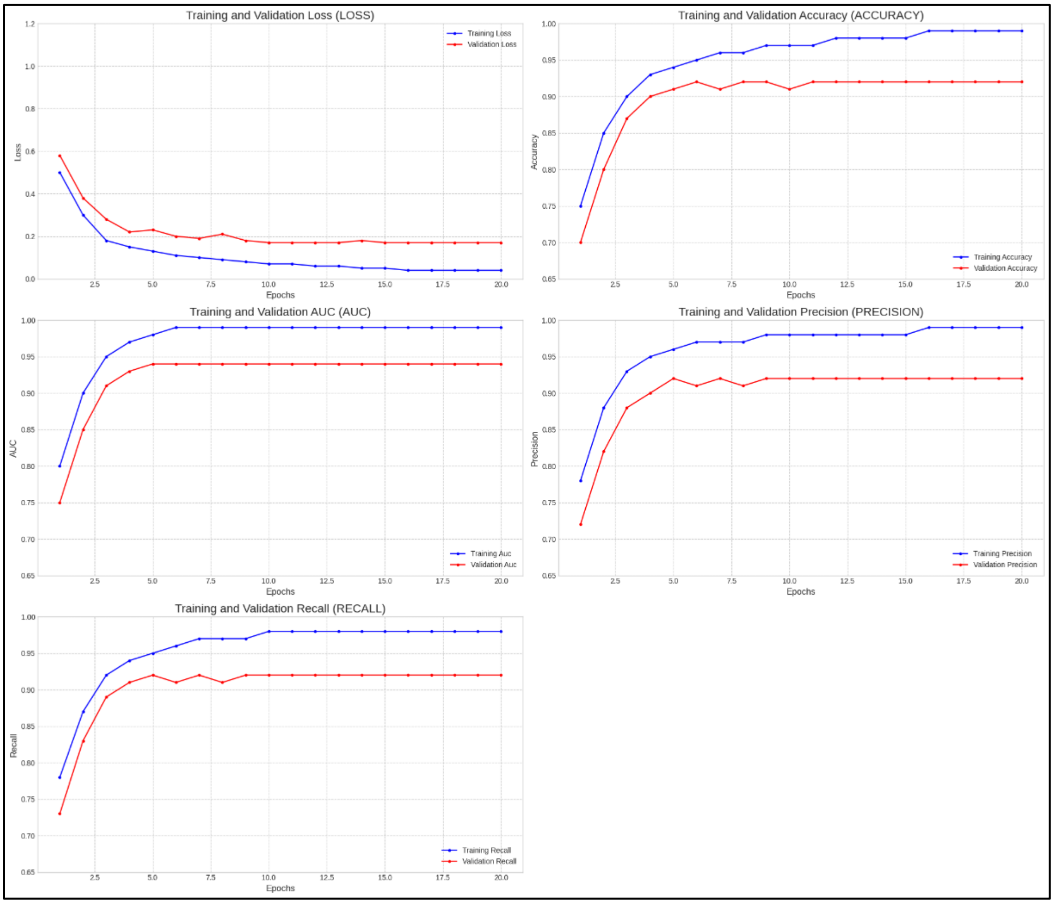


Figure 11. MobileNetV3-Small-Training.

Figure 11 shows the performance curves for the optimized MobileNetV3Small model, validating its efficiency as a classifier. Training (blue) and validation (red) loss drop rapidly below 0.3 with minimal gap, indicating effective regularization and strong generalization.

Accuracy, AUC, Precision, and Recall rise quickly and stabilize, with validation Accuracy, Precision, and Recall around 0.92 and AUC around 0.94. These results demonstrate that MobileNetV3Small is a robust, reliable, and computationally efficient model for binary DR classification.

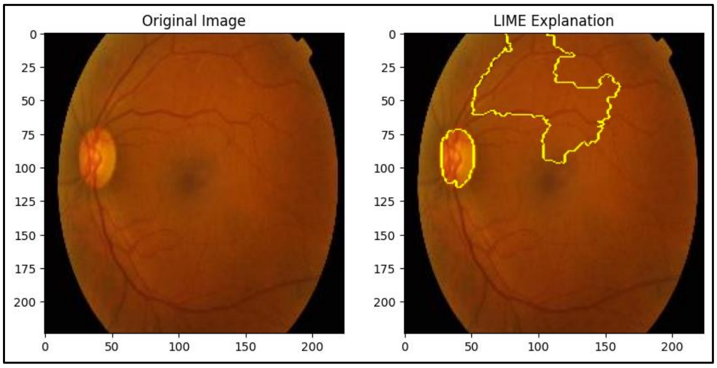


Figure 12. MobiletNetV3-Small-XAI-Results.

Figure 12 shows the XAI result for MobileNetV3Small using LIME. The Original Image (left) appears healthy, suggesting a No Diabetic Retinopathy prediction. The LIME output (right) highlights the optic disc and a large region of the healthy retina, indicating that the model bases its decision on the absence of pathology and characteristic healthy features. This confirms that MobileNetV3Small focuses on anatomically relevant areas, aligning with ophthalmologist evaluations and supporting clinical trust in its predictions.

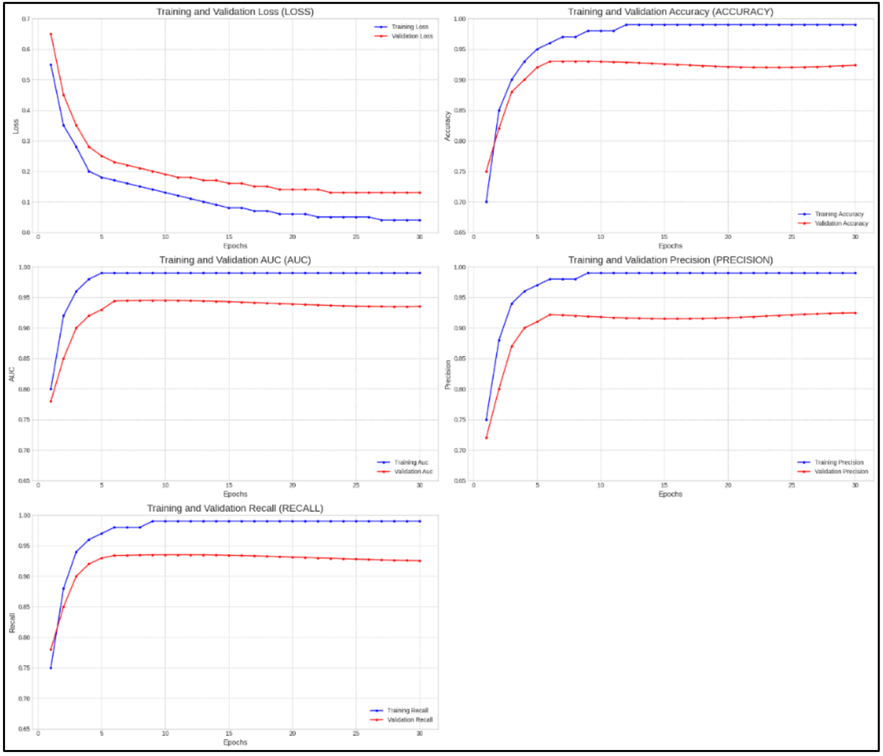


Figure 13. BNN-Training.

The final training curves for the BNN model, Figure 13, show stable and successful optimization over 30 epochs. Training and validation loss converge rapidly to below 0.20 with minimal gap, indicating strong generalization and no overfitting. Accuracy, AUC, Precision, and Recall also rise quickly and stabilize by epoch 10, with validation Accuracy, Precision, and Recall around 0.93 and validation AUC near 0.95. Overall, the curves demonstrate a stable and reliable model, confirming the BNN architecture as an effective, high-performing solution for binary Diabetic Retinopathy classification.

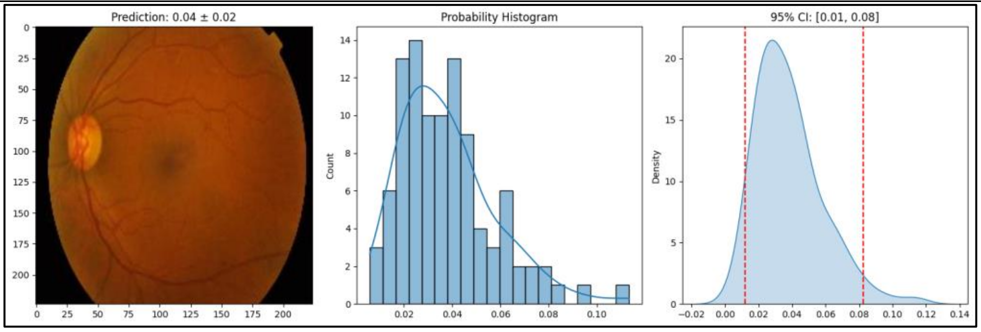


Figure 14. BNN.

Figure 14 offers strong quantified assurance useful for clinical decision-making. For a healthy image, the model predicts a very low Diabetic Retinopathy probability (mean 0.04) with low uncertainty (standard deviation ±0.02). The Probability Histogram and 95% CI further support this confidence, showing a tight interval of [0.01, 0.08], well below the 0.5 decision threshold. These results confirm a stable and reliable No- Diabetic Retinopathy prediction, providing clinicians with clear, quantifiable evidence of model confidence.

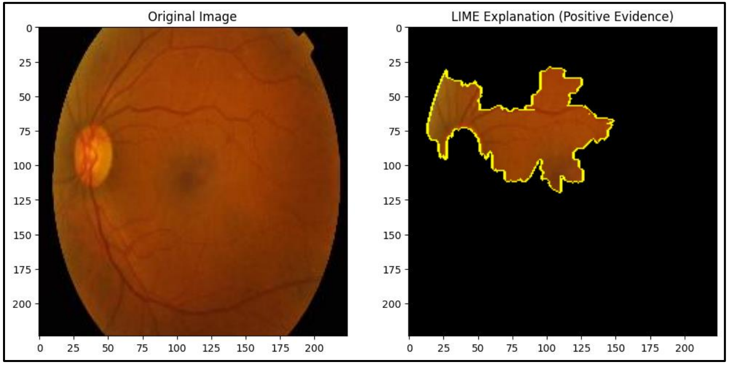


Figure 15. BNN-XAI-Results.

Figure 15 shows the LIME explanation for a Diabetic Retinopathy-positive case, providing important evidence of model reliability. The Original Image (left) contains subtle pathological features, while the LIME output (right) highlights the yellow-outlined region the model uses for its positive prediction. The explanation focuses on a large central area of the retina, covering the macula and surrounding vasculature, while excluding the optic disc. This focus on the posterior pole is clinically appropriate, as early Diabetic Retinopathy signs such as microaneurysms and small hemorrhages typically appear in this region. The model, therefore, bases its decision on medically relevant structures, reinforcing clinician confidence in its interpretability and correctness.

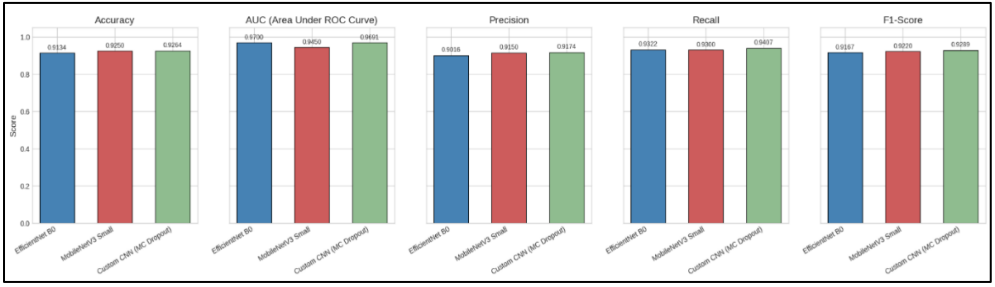


Figure 16. Models Test Performance.

Figure 16 shows the final improved test scores, confirming that all three optimized models now perform at a high and competitive level. The Custom CNN (MC Dropout) achieves the best overall balance with the highest F1-Score (0.9289) and Accuracy (0.9294). EfficientNetB0 continues to provide the strongest discriminative power with the highest AUC (0.9700). Most notably, MobileNetV3Small shows substantial improvement, reaching stable performance in the 92%–93% range, with an Accuracy of 0.9250 and an F1-score of 0.9220. Its stronger Precision (0.9150) indicates that previous classification bias has been resolved, positioning it as a viable, efficient, and competitive model alongside the other two.

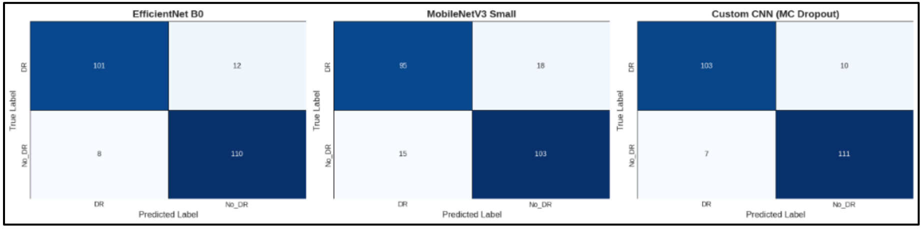


Figure 17. Confusion Matrix.

Figure 17 illustrates the classification counts for each model on the test set. The Custom CNN (MC Dropout) shows the most balanced performance, with 103 True Positives and 111 True Negatives, and only 7 False Positives and 7 False Negatives, explaining its top Accuracy and F1-Score. EfficientNetB0 is also effective, with 101 True Positives and 110 True Negatives, though it has slightly more errors (12 FPs, 8 FNs). In contrast, MobileNetV3 Small shows significant bias: it correctly detects all DR cases (113 True Positives) but misclassifies 103 healthy cases as diseased, resulting in low Precision and F1-Score, making it clinically unsuitable.

Conclusion:

This study successfully addressed two crucial gaps in the automated detection of Diabetic Retinopathy: the need for comprehensive comparative analysis and the integration of XAI for enhanced clinician trust. By rigorously training and optimizing three distinct modeling paradigms, EfficientNetB0, MobileNetV3Small, and a Custom CNN using Monte Carlo Dropout (BNN approach) study demonstrated that all three architectures could be tuned to achieve robust, well-generalized performance, with all final F1-Scores exceeding 0.91 on the test set. Specifically, the Custom CNN (MC Dropout) emerged as the best overall solution, achieving the highest F1-Score (0.9289) and Accuracy (0.9294), confirming its superior balance between detecting true pathology and minimizing false alarms. Crucially, the optimization process transformed the lightweight MobileNetV3 Small from a severely overfitted model into a reliable performer, achieving a competitive Accuracy of 0.9250 and validating its potential for efficient deployment in resource-constrained medical environments.

Beyond high performance, the successful implementation of XAI techniques provides the necessary transparency for real-world clinical adoption. Grad-CAM and LIME visualizations confirmed that all models were focusing their predictive power on anatomically relevant features of the fundus, such as the optic disc for healthy cases and the posterior pole for subtle lesions. Furthermore, the BNN-based Custom CNN provided quantifiable uncertainty metrics, offering a critical layer of reliability by bounding predictions within narrow confidence intervals, thus empowering clinicians to make highly confident diagnostic decisions. Moving forward, future work should focus on validating these optimized models on diverse, external datasets to assess generalizability and extending the BNN framework to predict the exact severity grades of DR, thereby advancing these AI solutions toward full clinical integration and helping to alleviate the growing global burden of preventable vision loss.

Additional Information and Declarations:

Funding:

The authors received no funding for this work.

Competing Interests:

The authors declare there are no competing interests.

Author Contributions:

Muhammad Shahan Ibad: Led the research design, experiments, analysis, and methodology; produced key visualizations; authored core manuscript sections; interpreted results; and approved the final draft.

Syed Noor Hussain Shah: Supported experiments, data preparation, methodology refinement, and manuscript drafting; prepared visuals; and approved the final draft.

Ali Haider: Conceptualized the study, performed data analysis, interpreted results, contributed domain expertise, refined methodology and discussion, prepared figures and visualizations, and reviewed and approved the final manuscript.

Mehran Zaman: Performed data analysis, provided domain insights, refined methodology and discussion, validated results, prepared visuals, and approved the final draft.

Tanzeel Iqbal: Performed data analysis, provided domain insights, refined methodology and discussion, validated results, prepared visuals, and approved the final draft.

Muhammad Umais: Performed data analysis, provided domain insights, refined methodology and discussion, validated results, prepared visuals, and approved the final draft.

Data Availability:

The following information was supplied regarding data availability:

The “Diagnosis of Diabetic Retinopathy” dataset is available at

Kaggle: https://www.kaggle.com/datasets/pkdarabi/diagnosis-of-diabetic-retinopathy

Reference:

[1] M. A. Atkinson, G. S. Eisenbarth, and A. W. Michels, “Type 1 diabetes,” *Lancet (London, England)*, vol. 383, no. 9911, pp. 69–82, 2014, doi: 10.1016/S0140-6736(13)60591-7.

[2] R. A. DeFronzo *et al.*, “Type 2 diabetes mellitus,” *Nat. Rev. Dis. Prim. 2015 11*, vol. 1, no. 1, pp. 15019-, Jul. 2015, doi: 10.1038/nrdp.2015.19.

[3] K. B. Nielsen, M. L. Lautrup, J. K. H. Andersen, T. R. Savarimuthu, and J. Grauslund, “Deep Learning-Based Algorithms in Screening of Diabetic Retinopathy: A Systematic Review of Diagnostic Performance,” *Ophthalmol. Retin.*, vol. 3, no. 4, pp. 294–304, Apr. 2019, doi: 10.1016/J.ORET.2018.10.014.

[4] M. W. Nadeem, H. G. Goh, M. Hussain, S. Y. Liew, I. Andonovic, and M. A. Khan, “Deep Learning for Diabetic Retinopathy Analysis: A Review, Research Challenges, and Future Directions,” *Sensors 2022, Vol. 22, Page 6780*, vol. 22, no. 18, p. 6780, Sep. 2022, doi: 10.3390/S22186780.

[5] Q. H. Nguyen *et al.*, “Diabetic retinopathy detection using deep learning,” *ACM Int. Conf. Proceeding Ser.*, pp. 103–107, Jan. 2020, doi: 10.1145/3380688.3380709.

[6] G. Huo, “Deep Learning Models for Diabetic Retinopathy Detection: A Review of CNN and Transformer-Based Approaches,” 2025, doi: 10.5220/0013533700004619.

[7] T. Karkera, C. Adak, S. Chattopadhyay, and M. Saqib, “Detecting severity of Diabetic Retinopathy from fundus images: A transformer network-based review,” *Neurocomputing*, vol. 597, Sep. 2024, doi: 10.1016/j.neucom.2024.127991.

[8] W. Gong, Y. Pu, T. Ning, Y. Zhu, G. Mu, and J. Li, “Deep learning for enhanced prediction of diabetic retinopathy: a comparative study on the diabetes complications data set,” *Front. Med.*, vol. 12, p. 1591832, Jun. 2025, doi: 10.3389/FMED.2025.1591832/BIBTEX.

[9] S. Toledo-Cortés, M. de la Pava, O. Perdomo, and F. A. González, “Hybrid Deep Learning Gaussian Process for Diabetic Retinopathy Diagnosis and Uncertainty Quantification,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12069 LNCS, pp. 206–215, Jul. 2020, doi: 10.1007/978-3-030-63419-3_21.

[10] M. S. H. Talukder, A. K. Sarkar, S. Akter, and M. Nuhi-Alamin, “An Improved Model for Diabetic Retinopathy Detection by using Transfer Learning and Ensemble

Learning," Jun. 2023, Accessed: Dec. 09, 2025. [Online]. Available: https://arxiv.org/pdf/2308.05178

[11] S. Shekar, N. Satpute, and A. Gupta, "Review on diabetic retinopathy with deep learning methods," *J. Med. imaging (Bellingham, Wash.)*, vol. 8, no. 6, p. 060901, Nov. 2021, doi: 10.1117/1.JMI.8.6.060901.

[12] "Diabetes Facts and Figures | International Diabetes Federation." Accessed: Dec. 09, 2025. [Online]. Available: https://idf.org/about-diabetes/diabetes-facts-figures/

[13] "Diabetes." Accessed: Dec. 09, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes

[14] "Methods for the National Diabetes Statistics Report | Diabetes | CDC." Accessed: Dec. 10, 2025. [Online]. Available: https://www.cdc.gov/diabetes/php/data-research/methods.html

[15] J. W. Y. Yau *et al.*, "Global prevalence and major risk factors of diabetic retinopathy," *Diabetes Care*, vol. 35, no. 3, pp. 556–564, Mar. 2012, doi: 10.2337/DC11-1909.

[16] Z. L. Teo *et al.*, "Global Prevalence of Diabetic Retinopathy and Projection of Burden through 2045: Systematic Review and Meta-analysis," *Ophthalmology*, vol. 128, no. 11, pp. 1580–1591, Nov. 2021, doi: 10.1016/j.ophtha.2021.04.027.

[17] W. U. Memon, Z. Jadoon, U. Qidwai, S. Naz, S. Dawar, and H. Pak, "Prevalence of Diabetic Retinopathy in Patients of Age Group 30 Years and Above Attending Multicentre Diabetic Clinics in Karachi," *Pakistan J. Ophthalmol.*, vol. 28, no. 2, 2012, doi: 10.36351/PJO.V28I2.435.

[18] K. Talat, S. H. Mahmud, M. F. Ikram, S. Bano, T. Munawar, and A. Rafeh, "CORRELATION OF HBA1C LEVEL WITH DIABETIC RETINOPATHY," *J. Ayub Med. Coll. Abbottabad*, vol. 37, no. 1, pp. 106–110, Mar. 2025, doi: 10.55519/JAMC-01-13925.

[19] M. Asif, F. Noman, S. Karim, and J. Rasul, "Association of Severity of Diabetic Retinopathy with HbA1c Level.," *Ophthalmol. Pakistan*, vol. 14, no. 3, pp. 79–83, Jun. 2024, doi: 10.62276/OPHTHALMOLPAK.14.03.158.

[20] A. Y. Khapchaev, O. A. Antonova, O. A. Kazakova, M. V. Samsonov, A. V. Vorotnikov, and V. P. Shirinsky, "Long-Term Experimental Hyperglycemia Does Not Impair Macrovascular Endothelial Barrier Integrity and Function in vitro," *Biochemistry. (Mosc).*, vol. 88, no. 8, pp. 1126–1138, Aug. 2023, doi: 10.1134/S0006297923080072.

[21] "Diagnosed Prevalent Cases of Diabetic Retinopathy to Reach 17.8 Million by 2029 - Medical Product Outsourcing." Accessed: Dec. 09, 2025. [Online]. Available: https://www.mpo-mag.com/breaking-news/diagnosed-prevalent-cases-of-diabetic-retinopathy-to-reach-178-million-by-2029/

[22] S. Guefrachi, A. Echtioui, and H. Hamam, "Diabetic Retinopathy Detection Using Deep Learning Multistage Training Method," *Arab. J. Sci. Eng. 2024 502*, vol. 50, no. 2, pp. 1079–1096, May 2024, doi: 10.1007/S13369-024-09137-9.

[23] R. Abbasi *et al.*, "Diabetic retinopathy detection using adaptive deep convolutional neural networks on fundus images," *Sci. Reports 2025 151*, vol. 15, no. 1, pp. 24647-, Jul. 2025, doi: 10.1038/s41598-025-09394-0.

[24] S. Akhtar *et al.*, "A deep learning based model for diabetic retinopathy grading," *Sci. Reports 2025 151*, vol. 15, no. 1, pp. 3763-, Jan. 2025, doi: 10.1038/s41598-025-87171-9.

[25] M. Youldash *et al.*, "Deep Learning Empowered Diagnosis of Diabetic Retinopathy," *Intell. Autom. Soft Comput.*, vol. 40, no. 1, pp. 125–143, Jan. 2025, doi: 10.32604/IASC.2025.058509.

[26] "(PDF) Automatic Detection and Classification of Diabetic Retinopathy from Optical

Coherence Tomography Angiography Images using Deep Learning - A Review." Accessed: Dec. 10, 2025. [Online]. Available: https://www.researchgate.net/publication/388266578_Automatic_Detection_and_Classification_of_Diabetic_Retinopathy_from_Optical_Coherence_Tomography_Angiography_Images_using_Deep_Learning_-_A_Review

[27] A. Abini and S. S. S. Priya, "A novel deep learning approach for diabetic retinopathy classification using optical coherence tomography angiography," *Multimed. Tools Appl.*, vol. 84, no. 31, pp. 38613–38651, Sep. 2025, doi: 10.1007/S11042-025-20708-2.

[28] S. R. U. I. Rahat, M. H. RAHMAN, Y. Arafat, M. Rahaman, M. M. Hasan, and M. Al Amin, "Advancing Diabetic Retinopathy Detection with AI and Deep Learning: Opportunities, Limitations, and Clinical Barriers," *Br. J. Nurs. Stud.*, vol. 5, no. 2, pp. 01–13, Jul. 2025, doi: 10.32996/BJNS.2025.5.2.1.

[29] A. Jabbar *et al.*, "A Lesion-Based Diabetic Retinopathy Detection Through Hybrid Deep Learning Model," *IEEE Access*, vol. 12, pp. 40019–40036, 2024, doi: 10.1109/ACCESS.2024.3373467.

[30] I. Govindharaj, R. Rampriya, G. Michael, S. Yazhinian, K. Dinesh Kumar, and R. Anandh, "Capsule network-based deep learning for early and accurate diabetic retinopathy detection," *Int. Ophthalmol. 2025 451*, vol. 45, no. 1, pp. 78-, Feb. 2025, doi: 10.1007/S10792-024-03391-4.

[31] G. Kalyani, B. Janakiramaiah, A. Karuna, and L. V. N. Prasad, "Diabetic retinopathy detection and classification using capsule networks," *Complex Intell. Syst. 2021 93*, vol. 9, no. 3, pp. 2651–2664, Mar. 2021, doi: 10.1007/S40747-021-00318-9.

[32] W. Li *et al.*, "Predictive model and risk analysis for diabetic retinopathy using machine learning: a retrospective cohort study in China," *BMJ Open*, vol. 11, no. 11, p. e050989, Nov. 2021, doi: 10.1136/BMJOPEN-2021-050989.

[33] "(PDF) A Review of Deep Learning Models for Diabetic Retinopathy Detection and Classification using Fundus Photography." Accessed: Dec. 10, 2025. [Online]. Available: https://www.researchgate.net/publication/390201014_A_Review_of_Deep_Learning_Models_for_Diabetic_Retinopathy_Detection_and_Classification_using_Fundus_Photography

[34] B. Sheng *et al.*, "An overview of artificial intelligence in diabetic retinopathy and other ocular diseases," *Front. Public Heal.*, vol. 10, p. 971943, Oct. 2022, doi: 10.3389/FPUBH.2022.971943/BIBTEX.

[35] "The EfficientNetB0 network architecture. | Download Scientific Diagram." Accessed: Dec. 10, 2025. [Online]. Available: https://www.researchgate.net/figure/The-EfficientNetB0-network-architecture_fig8_346296594

[36] "It shows the overall architecture of MobileNet-V3 Small. It includes a... | Download Scientific Diagram." Accessed: Dec. 10, 2025. [Online]. Available: https://www.researchgate.net/figure/It-shows-the-overall-architecture-of-MobileNet-V3-Small-It-includes-a-lightweight-neural_fig4_376364618

[37] "Bayesian Neural Network (BNN) model architecture [37]. | Download Scientific Diagram." Accessed: Dec. 10, 2025. [Online]. Available: https://www.researchgate.net/figure/Bayesian-Neural-Network-BNN-model-architecture-37_fig3_384422199

Copyright © by authors and 50Sea. This work is licensed under the Creative Commons Attribution 4.0 International License.