JSCDM

JOURNAL OF SOFT COMPUTING AND DATA MINING
e-ISSN: 2716-621X

Vol. 7 No. 1 (2026) 106-126
https://publisher.uthm.edu.my/ojs/index.php/jscdm

Multimodal Alignment and Fusion for Diabetic Retinopathy Detection using Deep learning Approach

Kartina Diah Kesuma Wardhani[1,2]*, Shahreen Kasim[2], Wawan Yunanto[1], Deshinta Arrova Devi[3], Rohayanti Hassan[4], Sheeba Armoogum[5], Mohammad Syafwan Arshad[6]

1 *Department of Informatics Engineering, Politeknik Caltex Riau, Umban Sari No. 1, Umbansari, Kecamatan Rumbai, Pekanbaru, 28265, Riau, INDONESIA*
2 *Soft Computing and Data Mining Centre (SCM), Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, 86400, Johor Darul Ta'zim, MALAYSIA*
3 *Center for Data Science and Sustainable Technologies, INTI International University, Nilai, MALAYSIA*
4 *Faculty of Computing, Universiti Teknologi Malaysia, Skudai, Johor Bahru, MALAYSIA*
5 *University of Mauritius, Reduit 80837, MAURITIUS*
6 *NILECRAFT GLOBAL SDN BHD PO2, Unit 10-15, 10th Floor, Menara MAA Lorong Api Api 1, Api Api Centre, 88000, Kota Kinabalu, Sabah, MALAYSIA*

*Corresponding Author: diah@pcr.ac.id
DOI: https://doi.org/10.30880/jscdm.2026.07.01.009

Article Info

Received: 30 October 2025
Accepted: 23 February 2026
Available online: 10 April 2026

Keywords

Multimodal alignment, late fusion, diabetic retinopathy, fundus, OCT scan, EHR

Abstract

This study proposes a label-driven multimodal alignment with late fusion framework for diabetic retinopathy (DR) detection, integrating fundus images, OCT scans, and structured electronic health records (EHR). Unlike patient-wise paired datasets, which are often unavailable, the proposed alignment strategy groups and matches modalities based on diagnostic labels (DR/No_DR), ensuring semantic consistency across heterogeneous sources. Each modality is modeled using an architecture tailored to its data type—CNNs for fundus and OCT images, and an ANN for EHR—before predictions are combined via four late fusion strategies: Simple Average, Weighted Average, Majority Voting, and Stacked Ensemble. By implementing label-driven alignment, this framework ensures that multimodal data can be integrated based on aligned class labels, even without patient-level pairing. Experimental results evaluated based on accuracy, specificity, sensitivity, and F1-Score, show that while the unimodal CNN-OCT and ANN-EHR models achieve strong accuracy, the Simple Average and Weighted Average fusion methods achieve the highest F1-Score (0.999), indicating an optimal precision-recall balance. Confusion matrix analysis demonstrates high specificity and sensitivity, emphasizing the ability of label-aligned multimodal fusion to leverage complementary diagnostic strengths. These findings highlight that label-based alignment, combined with average-based final fusion, not only improves predictive performance but can also be a scalable and

interpretable AI-assisted DR screening solution for real-world ophthalmology practice.

1. Introduction

High blood sugar levels that persist in the bloodstream can lead to insulin resistance, leading to diabetes mellitus. Uncontrolled diabetes mellitus over a long period of time can lead to various complications, one of which is Diabetic Retinopathy (DR). DR occurs due to the rupture of small blood vessels in the retina, gradually reducing the retina's ability to see. DR that is not recognized properly and continues to progress can lead to blindness[1], [2]. The sooner DR is recognized, the simpler the treatment will be, but more complex treatment may be required in advanced stages of DR[3], [4]. n the early stages, ophthalmologists and ophthalmologists will manually detect DR through an analysis of the diabetes history and a physical examination of the eye, including fundoscopy, to look for signs of retinopathy, such as bleeding in the small blood vessels in the retina. Advanced examinations include Optical Coherence Tomography (OCT) to view the retinal structure and Fluorescein Angiography (AF) to clearly detect bleeding in the blood vessels. Manually detecting DR requires a longer screening process, and the diagnosis is highly subjective and dependent on the expertise of the ophthalmologist[5],[6]. Therefore, automated DR detection using artificial intelligence, particularly deep learning, is now a primary choice for healthcare facilities. Deep learning provides a more reliable capability for automated retinal image analysis, offering high accuracy and greater time efficiency. With deep learning, the system can learn directly from image data, eliminating the need for manual feature extraction, and consistently identify retinal lesions, even in variable image quality. Therefore, integrating deep learning into DR detection systems offers significant potential to increase mass screening capacity, reduce the workload of medical personnel, and expand access to early detection, particularly in areas with limited specialist resources[7]–[10].

The development of deep learning (DL) based automated DR detection has been extended to different image sys-tems including fundus images, optical coherence tomography (OCT), fluorescein angiography (FA) and others. Yet most previous works show that deep learning is generally performed on unimodal data, that is detection relies on only one type of data[11]–[16]. The problem with this approach is that it limits analysis to partial information about the patient's real status. For instance, fundus photography contains two-dimensional information of the retinal surface, while OCT imagines the retinal structures in a three-dimensional view. However, it measures neither systemic patient factors such as blood glucose, blood pressure, duration of diabetes, which are available in the EHR, nor parameters indicative of the quality of diabetes care.

DR is a complex condition, so detecting DR using unimodal data can result in a diagnosis based on only one condition, thus missing important clinical context relevant for prompt and appropriate DR treatment. Therefore, multimodal fusion is gaining increasing attention in DR detection research. By combining multiple data sources—fundus images, OCT images, and clinical data from the EHR—the system can obtain a more complete representation of the patient's condition. This approach is believed to improve detection accuracy, enhance model generalization, and enhance interpretability and user confidence in AI-based systems for clinical decision-making[17]–[20].

Applying multimodal fusion approaches in research poses a fundamental challenge: the available datasets often come from disparate sources. Obtaining correlated or paired multimodal data is particularly challenging[18], [21-22]. Often, the fundus, OCT, and EHR data are collected from different sources or at different times, but without so much as a uniform patient ID or metadata that would enable direct alignment. The mismatch causes unpaired samples, which makes it challenging to combine them since there is no explicit link to associate each modality with an individual. Such a problem may also occur at fusion when combining features derived from misaligned modalities, which results in invalid representations compromising model performance. Hence, multimodal data alignment is an important step prior to fusion. Multimodal data (labels) alignment tries to compensate or align context between modalities (pairs are matched based on labels, semantic mappings, or transformations such as pseudo-image encoding for EHR data) so that the fused data can be interpreted as the same or at least equivalent patient condition. [23], [24]. Without this step, the potential of multimodal fusion to improve model accuracy and robustness cannot be optimally utilized.

Based on these conditions, this study makes a significant contribution to addressing the challenges of integrating unpaired multimodal data in deep learning-based Diabetic Retinopathy detection. First, this study proposes and implements a data alignment technique designed to paired modalities, such as fundus, OCT, and EHR, with a Label-Based Alignment approach. Second, this study assesses the impact of this alignment strategy on DR classification performance in multimodal data fusion by comparing model evaluation metrics from unimodal data with those from multimodal alignment and fusion. Third, this study compares evaluation metrics across several final fusion strategies, namely Simple AVG final fusion, Weighted AVG final fusion, Majority Voting final fusion, and Stacked Ensemble final fusion, to identify optimal configurations to improve model accuracy and generalization in the context of heterogeneous and unstructured clinical data.

This article is organized into five main sections. Section 1 presents the Introduction, which outlines the background of the problem, the importance of early detection of DR, the development of deep learning in retinal image analysis, and the motivation and objectives of this study. Section 2 describes Related Work, which includes a review of previous research on DR detection using deep learning, commonly used fusion strategies, and multimodal alignment techniques in the context of unpaired data. Section 3 discusses Materials and Methods, including a description of the dataset, pre-processing steps, the proposed alignment strategy, the fusion approach, and the model architecture and training process. Section 4 presents the Results and Discussion, which includes an analysis of model performance before and after alignment, and an evaluation of various fusion strategies. Finally, Section 5 summarizes the main results of this study and provides directions for further research.

2. Related Work

This section presents a review of related literature that aligns with the research on automatic detection of Diabetic Retinopathy (DR) using a multimodal deep learning approach. In line with the contribution of this study, the main focus of this literature review is on multimodal data fusion in automatic detection of DR, which provides a description of how multimodal data integration in DR detection has been used and implemented, and the second is the alignment of multimodal data on modalities that are not connected to each other in previous studies.

Table 1 presents several recent studies, such as our previous works in 2024 Wardhani et al[25] and Karthikeyan et al[26], which have proposed multimodal fusion for diabetic retinopathy (DR) detection by combining fundus, OCT, and EHR, but do not explain the inter-modality alignment strategy, especially for unpaired data. Restrepo et al [22] introduced embedding-level fusion, but it is general and not specific to DR. Meanwhile, other studies, such as those by El-Ateif & Idri[27] and Li et al[19], [28], only performed fusion between image modalities without involving EHR or alignment mechanisms.

Table 1 *Comparative analysis of recent multimodal fusion studies for diabetic retinopathy detection, highlighting methods, limitations, and the absence of alignment strategies for unpaired datasets*

Year	Authors	Title	Proposed Methods	Disadvantages
2024	Wardhani et al[25]	Deep Learning-based Method in Multimodal Data for Diabetic Retinopathy Detection	Early fusion of Fundus + OCT + EHR using LSTM + LBP	It is not explained whether the dataset is paired or unpaired; there is no discussion of alignment strategy.
2024	Karthikeyan et al[29]	Multimodal Approach for Diabetic Retinopathy Detection Using Deep Learning and Clinical Data Fusion	CNN for image + MLP for EHR, fusion via integration	It is not explicitly stated whether patient-wise data is used, or alignment is performed.
2024	Restrepo et al[22]	DF-DM: A Foundational Process Model for Multimodal Data Fusion in the AI Era	Disentangled dense fusion + embedding-level alignment on fundus + metadata	Alignment is not described in detail and is only tested on a limited dataset.
2022	El-Ateif & Idri[27]	Single-modality and joint fusion deep learning for diabetic retinopathy diagnosis	Joint fusion (multi-CNN) for combining multiple imaging modalities	Image based only, no EHR or metadata used
2023	Li et al[19]	Multimodal Information Fusion for the Diagnosis of Diabetic Retinopathy	Fusion of OCT, OCTA, and LSO imaging for PDR detection	No EHR or metadata involved; just multimodal imaging
2022	Li et al[28]	Multimodal Information Fusion for Glaucoma and Diabetic Retinopathy Classification	Early, intermediate, and hierarchical fusion strategies tested on Fundus + OCT	Does not mention alignment approach to multimodal input

Departing from these limitations, this study offers a novel label-driven alignment approach for unpaired multimodal datasets, as well as the implementation of a late fusion strategy consisting of simple average, weighted

average, majority voting, and stacked ensemble to provide a practical and effective solution to the limitations of heterogeneous data integration in the DR detection domain. This is a very real situation in the medical world, where multimodal data comes from different systems, is stored separately, and cannot always be directly linked due to data privacy, encryption processes in hospital systems and data ethics policies.

3. Materials and Methods

To provide a clear understanding of the proposed framework, Figure 1 illustrates the complete workflow of our study, from preprocessing to evaluation. The process begins with label-driven multimodal alignment, where Fundus, OCT, and EHR datasets are synchronized at the class-label level (DR/No_DR) to overcome the lack of patient-level pairing. Each aligned modality is then processed by a specialized unimodal model—CNNs for Fundus and OCT images, and an ANN for EHR tabular data—ensuring that the unique characteristics of each data type are preserved. Unimodal predictions are subsequently integrated through late fusion strategies (Simple Average, Weighted Average, Majority Voting, and Stacked Ensemble), enabling complementary decision-making across modalities. Finally, the fused outputs are evaluated using standard metrics (accuracy, specificity, sensitivity, and F1-score) to validate the effectiveness of the approach.

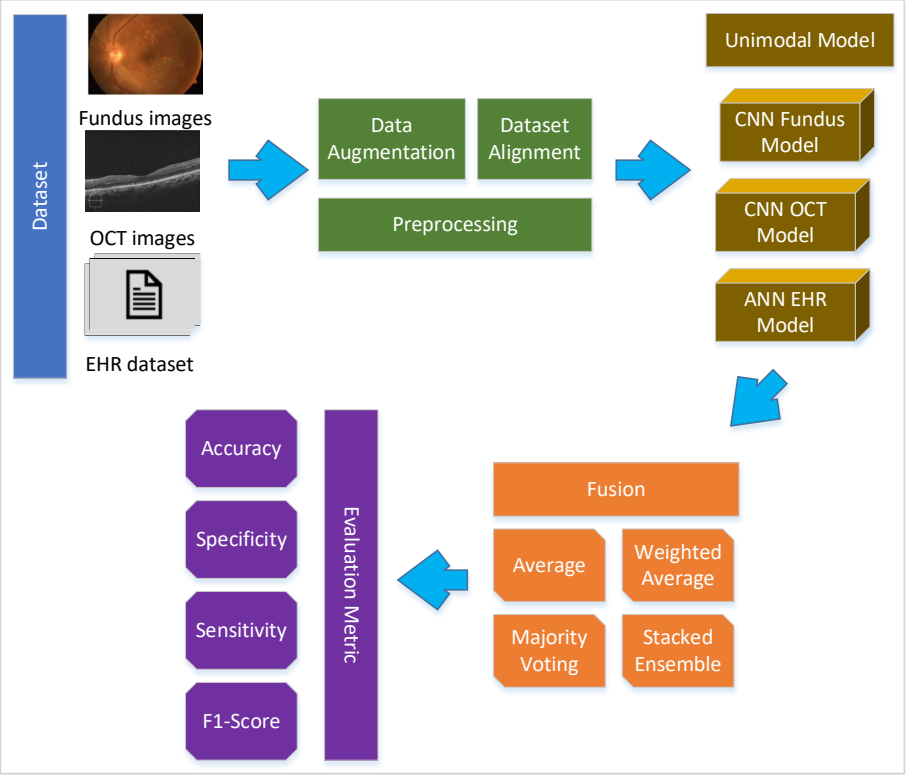


Fig. 1 *Proposed pipeline for diabetic retinopathy detection*

3.1 Datasets

The datasets utilized in this study comprises fundus images[1] obtained from a public dataset hosted by MIT and shared on Kaggle[30] containing 2000 images (1000 DR, 1000 No_DR), published version Diabetic Retinopathy Retinal OCT Images from Borealis[2] (accessible at borealisdata.ca) is Canada's national Dataverse repository[31] consisting of 2000 images (1000 DR, 1000 No_DR) and Electronic Health Record (EHR) [3] data from the open-access dataset by Dr. Yad[32] comprising 2000 patient records (1000 DR, 1000 No_DR). In total, 6,000 data samples were employed in the modeling process of this study. The EHR dataset employed in this study contains

[1] **DOI: 10.13140/RG.2.2.13037.19688**

[2] https://borealisdata.ca/dataset.xhtml?persistentId=doi%3A10.5683%2FSP%2FFLGZZE&version=&q=&fileTypeGroupFacet=%22Image%22&fileAccess=&fileTag=&fileSortField=&fileSortOrder=

[3] https://datadryad.org/stash/dataset/doi:10.5061/dryad.6kg1sd7

J. of Soft Computing and Data Mining Vol. 7 No. 1 (2025) p. 106-126

32 attributes, including clinical measurements and the class label, where 1 indicates the presence of DR and 0 denotes its absence. For each modality, data were stratified into training (64%), validation (16%), and test (20%) subsets to preserve class balance.

Figure 2 shows exemplary fundus and OCT retinal images, depicting the contrasts seen in the DR patients. In the fundus images, DR shows scattered retinal hemorrhages, which are dark spots formed by blood leaking from blood vessels that can cause vision loss if not treated[33]–[35]. In the OCT scans, DR shows large central intraretinal cysts associated with pronounced cystoid macular edema, which is a representative advanced diabetic macular edema case, illustrating the structural abnormalities commonly associated with diabetic retinopathy[36]–[39].

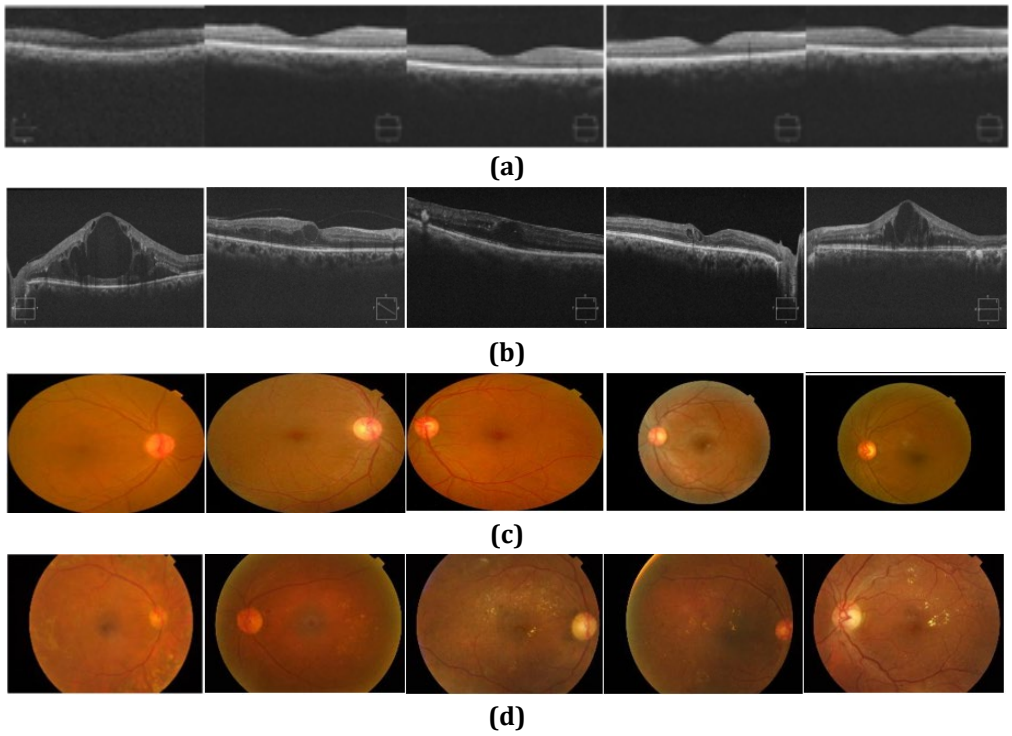


Fig. 2 *Images dataset (a) Normal OCT image; (b) DR OCT image; (c) Normal fundus image; (d) DR fundus image*

3.2 Preprocessing

In our work, the same pre-processing was applied for each modality including data augmentation and data alignment. Data augmentation is used to balance the number of modalities according to their class labels. In addition, augmentations were run to expand the training data and mitigate overfitting of the model. Figure 3 illustrates the image augmentation techniques, along with tabular data augmentation applied to the EHR to improve model variability and generalization.

For the two imaging modalities (fundus and OCT), the following augmentations were applied: horizontal flipping, vertical flipping, and random rotation to cover the limited angles and positions of patient's retina. These transformations were also applied symmetrically (with respect to class) to DR and No_DR to keep the balance and prevent bias. Every image was rotated to a random degree between −45° and +45°, representing the same level of rotation found in clinical imaging. In addition, images were randomly flipped horizontally, vertically, both (like OpenCV flip code of -1,0,1). These transformations were also applied symmetrically between classes to DR and No_DR for equal augmentation and to avoid class bias. Augmented images are produced only for training set, whereas the validation and test sets are unaltered to avoid any leakage of data and for unbiased evaluation.

On the EHR dataset, tabular data augmentation is performed using random sampling with replacement to increase training variance while preserving the class distribution. In particular, within each class (DR or No_DR), samples from the original dataset are randomly picked to form the augmented dataset, while the number of samples in the augmented set in each class should be equal to that in the original dataset. Without changing the original label distribution, the model experiences different mixtures of feature values in different epochs through this technique, so the model is less prone to overfitting. Similar to augmentation of image datasets, augmentation of a EHR dataset is performed over the training set only, leaving the validation and test sets unmodified so as to avoid data leaks. Overall, the modality augmentation ensures that all modalities are balanced based on class labels and contain rich variability of data.

3.3 Multimodal Alignment

This study implements multimodal alignment of fundus image, OCT scans, and EHR datasets using a label-driven alignment strategy.

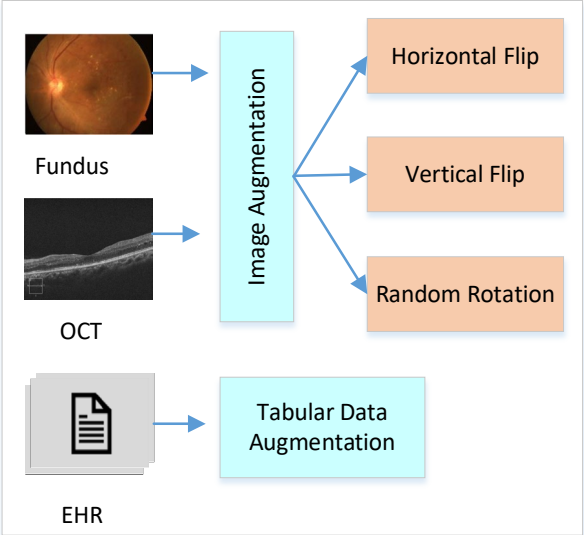


Fig. 3 *Preprocessing data*

In this context, alignment does not occur at the patient level (patient-wise), since the datasets are collected from different sources and do not have patient IDs in common. Instead, alignment at the class level is performed (DR vs. No_DR). In particular, for each modality, representations are aligned according to their diagnosis label, so that if two samples share the same label they are considered aligned even if they do not belong to the same patient.

For this study, the EHR dataset is used as the reference modality, because it contains verified ground-truth labels. The datasets are assumed to be sorted and balanced, with 2000 samples each. Indices 0–999 correspond to the DR label, and indices 1000–1999 correspond to the No_DR label. Based on this reference, the fundus and OCT datasets were re-indexed so that their order aligns with the EHR labels. As a result, each entry at a given index across modalities (e.g., Fundus[0], OCT[0], EHR[0]) represents the same class label (DR), although not the same patient.

For each modality:

$$m\{f,o,e\}(fundus\ \boldsymbol{f}, OCT\ Scan\ \boldsymbol{o}, EHR\ \boldsymbol{e}):$$
$$Dm = \{(x_i^{(m)}, y_i)\ \ {}^{1-1999}_{i=0}\} \tag{1}$$

Define the index set per label:

$$I_{No\_DR} = \{0, \dots, 999\}, I_{DR} = \{1000, \dots, 1999\} \tag{2}$$

Label as index function:

$$y_i = y_i^{(e)}, y_i^{(e)} = \begin{cases} 0, i \in I_{No\_DR} \\ 1, i \in I_{DR} \end{cases} \tag{3}$$

Make the EHR table a reference table. Define the index per class of the EHR:

$$I_e(y) = \{\, i : y_i^{(e)} = y\} \tag{3}$$

Alignment is done by taking the same index in all modalities:

$$a(i) = \left(x_i^{(f)}, x_i^{(0)}, x_i^{(e)}, y_i^{(e)}\right), i = 0, \dots, 1999 \tag{4}$$

J. of Soft Computing and Data Mining Vol. 7 No. 1 (2025) p. 106-126

So, the resulting pairwise sets for fundus, OCT Scan and EHR datasets are:

$$A = \{a(i) | \, i = 0, \ldots, 1999 \tag{5}$$

Illustration of label-driven alignment on fundus, OCT Scan using EHR label as reference can be seen in figure 4. This class-based alignment approach is consistent with strategies described in the multimodal learning literature for unpaired data, where alignment by label ensures semantic correspondence across modalities and prevents mismatched integration. By enforcing this assumption, the fusion process can leverage complementary diagnostic signals from each modality while avoiding inconsistencies that typically arise in unpaired multimodal data.

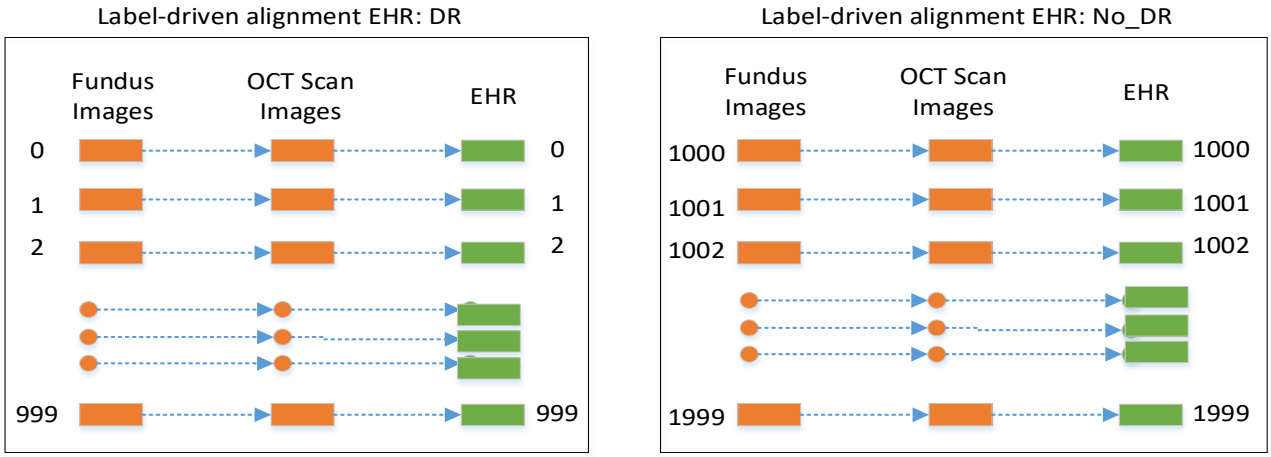


Fig. 4 *Label-driven alignment proposed*

3.4 Fusion Strategy

Figure 5 illustrates four variants of late fusion strategies in multimodal learning, where predictions from each modality-specific model (e.g., an image model for fundus/OCT and a tabular model for EHR) are combined at a late stage to produce a final decision. Late fusion is an approach in which each modality is processed by the architecture most appropriate for its data characteristics, and then the outputs (predictions or high-level representations) are combined using a specific mechanism. The advantage of this approach is flexibility and the ability to utilize the best model for each type of data without the need to match the input structure[40]–[42].

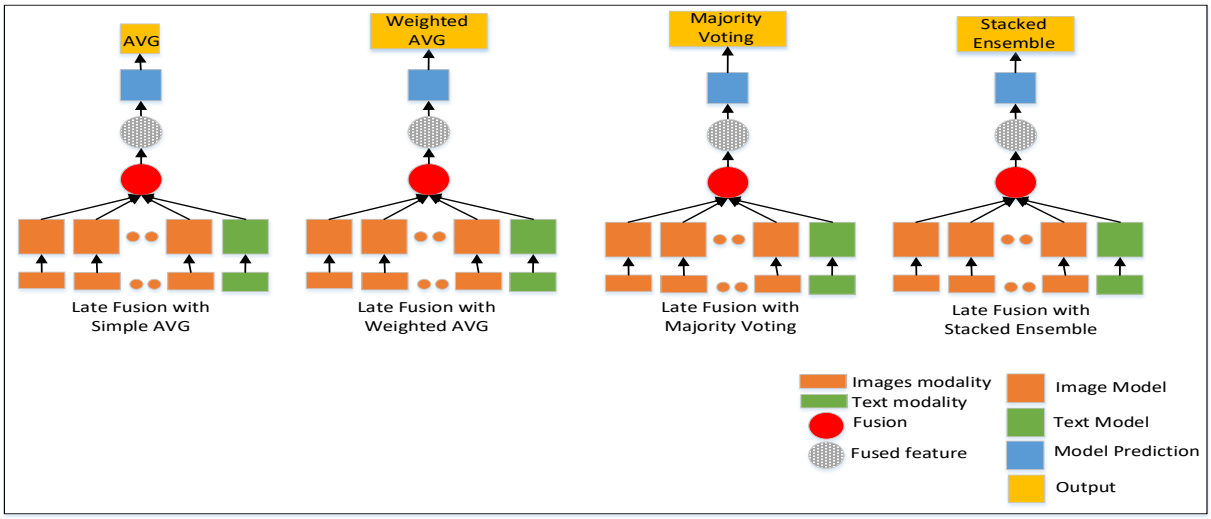


Fig. 5 *Late fusion strategy in multimodal data fusion for DR detection*

In the late fusion with Simple Average approach, the predicted probability outputs from each modality are simply averaged with equal weights across all modalities. This method is easy to implement, requires no additional training, and is often used when all modalities are considered to contribute equal information.

However, its drawback is that it does not take into account the relative reliability of each modality. The Late Fusion with Weighted AVG method introduces weights for each modality based on their importance. The weights can be determined manually using domain knowledge or optimized through a validation process. This approach allows more informative modalities (e.g., fundus images in DR detection) to have a greater influence on the final prediction, thus improving accuracy when the quality or relevance of the modalities is imbalanced. In majority voting, each modality provides discrete class predictions, and the most frequently occurring class is selected as the final output. This method is popular in ensemble classification scenarios because it is simple, parameter-free, and robust to outlier predictions from a particular modality. However, this method ignores probabilistic information and can therefore lose details about prediction confidence. Late Fusion with Stacked Ensemble uses a meta-learner or stacking model that learns the optimal way to combine prediction outputs from all modalities. The initial predictions from each modality become input features for a second-level model that then produces the final prediction. The advantage of this method is its ability to learn non-linear relationships between modalities and adapt the fusion strategy in a data-driven manner, although it requires additional training and has the potential for overfitting if data is limited. Overall, late fusion offers high flexibility in multimodal learning because it separates feature extraction training for each modality from the fusion stage, making it suitable for heterogeneous data such as medical images, signals, and tabular data. The choice of fusion strategy depends heavily on the amount of data, the quality of each modality, and the system's interpretability requirements.

3.5 Deep Learning Model Architecture

The deep learning architecture implemented in this late fusion approach consists of several main stages, namely the input layer, modality-specific feature extraction, fusion, classification, and output layer. Figure 6 shows Deep learning model architecture proposed in this research.

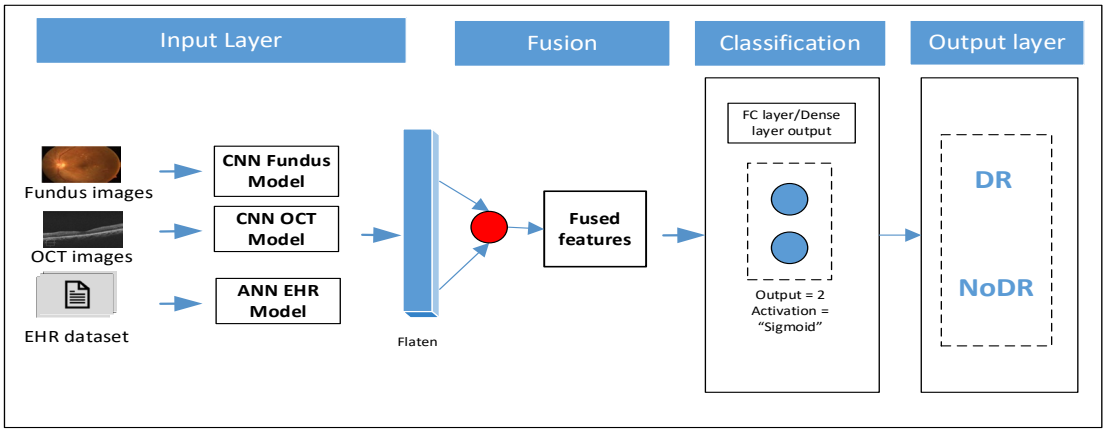


Fig. 6 *Deep learning model architecture proposed*

The first stage accepts three different, complementary data modalities for Diabetic Retinopathy (DR) detection: fundus images, OCT images, and tabular EHR data. Fundus and OCT represent visual information in the ophthalmology domain, while EHR contains tabular patient clinical data. These different characteristics are the main reason each modality is processed in a separate branch so that feature extraction can adapt to the nature of the data. CNN Fundus Model: Uses a Convolutional Neural Network to extract spatial visual patterns from fundus images. CNN OCT Model: Implements a similar CNN optimized for OCT images, capturing information about retinal layers and the internal structure of the eye. ANN EHR Branch (Tabular): Normalized EHR data flows into several Dense layers. This architecture is suitable for tabular because it is able to model non-linear interactions between features. The feature outputs from the three branches—CNN Fundus, CNN OCT, and ANN EHR feature vectors—are flattened and then concatenate into a single fused feature vector. Late fusion at the final feature level allows each extractor to optimize its domain-specific representation first and then combine them; this approach has proven stable when modalities have very different statistics and structures. The fused vectors pass through one or more Fully Connected layers (with normalization/regularization as needed) to fuse signals across modalities and improve class separability. The output layer uses Sigmoid to generate DR vs No_DR class outputs.

To prevent any form of data leakage, all evaluation metrics in this study were computed strictly on the held-out test set, which was never exposed during training or validation. The alignment procedure was conducted only at the class-label level (DR/No_DR) within the training and validation sets, without using any test labels. Furthermore, the late fusion strategies (Simple Average, Weighted Average, Majority Voting, Stacked Ensemble) were implemented in a fixed manner, with no weights or parameters tuned based on test data. This ensures that

J. of Soft Computing and Data Mining Vol. 7 No. 1 (2025) p. 106-126

the reported accuracy, specificity, sensitivity, and F1-Score reflect true generalization performance, independent of the held-out labels.

4. Results and Discussion

4.1 Unimodal Model

In this section, we will present the hyperparameter settings and the performance of each unimodal feature: Fundus, OCT, and EHR. The model architecture of each unimodal feature using CNN along with model evaluation metrics and training history in terms of loss and accuracy plots is also presented. The same applies to the multimodal fusion model that integrates Fundus, OCT, and EHR. Next, this section compares the late fusion strategies—Simple Average, Weighted Average, Majority Voting, and Stacked Ensemble—used to integrate the multimodal features. For each fusion technique, a description of the hyperparameters employed during training, including the fusion weighting factor and the classifier configuration. For each fusion scheme, evaluation metrics are reported alongside its loss and accuracy curves, so that we can easily see if there is any correlation among those values. In this paper, we compare and analyze the unimodal and multimodal approach, which can be supported by label-based alignment technique in DR detection. Also, this paper presents some analysis of several late fusion strategies applied in this work and which could influence the robustness of classification and generalization in different modalities.

Table 2 *Hyperparameter used in unimodal model*

Unimodal Dataset	No of Epoch	Batch Size	Learning rate
ANN-EHR	30	32	0.001
CNN-Fundus	30	32	0.0001
CNN-OCT Scan	30	32	0.001

Table 2 shows the hyperparameter configurations used to train unimodal models on three different data types: ANN-EHR, CNN-Fundus, and CNN-OCT Scan. The hyperparameters were tuned empirically to improve the quality of training, with an emphasis on convergence speed and stability of learning in terms of the model loss and the accuracy in the learning process. The batch size was 32, which means when performing one iteration, 32 samples were propagated forward and backward before parameters were updated. The learning rate was specific to each model concerning the modality and network architecture: 0.001 for ANN-EHR and CNN-OCT, and 0.0001 for CNN-Fundus. A smaller learning rate was adopted in CNN-Fundus to avoid the overshooting of the learning because the fundus images have more feature complexity and hence the weight updates should be more cautious.

Figure 7 illustrates the Artificial Neural Network (ANN) architecture used to model the Electronic Health Record (EHR) dataset. As seen in the diagram, the architecture consists of three dense layers following the input layer. The input layer (input_layer_2) receives a 31-element tabular feature vector (Output shape: (None, 31)), which represents the patient's clinical attributes. The first layer (dense_4) has 32 neurons with ReLU activation function, which extracts an initial representation of the tabular data and introduces non-linearity to enable the model to capture complex relationships between features. Next, the second layer (dense_5) increases the representation capacity to 64 neurons with ReLU activation, expanding the model's ability to abstract more complex patterns from the EHR. Finally, the output layer (dense_6) consists of 2 neurons with sigmoid activation, which generates probabilities for two output classes (e.g., DR and NoDR). This architecture is optimized using the Adam optimizer algorithm with a specific learning rate and a loss function suitable for classification, such as sparse_categorical_crossentropy, so that it can effectively process tabular data and provide probability-based predictions.

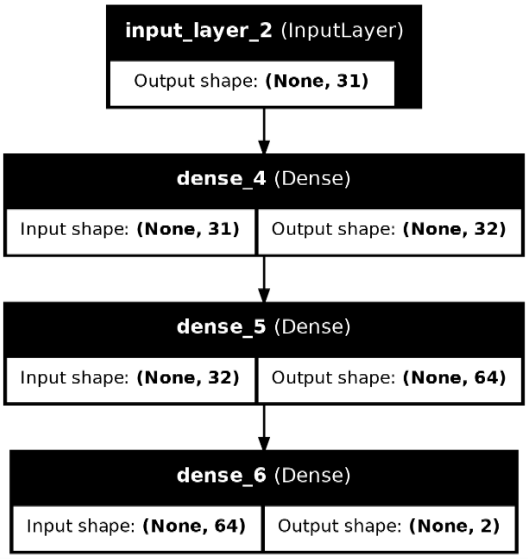


Fig. 7 *ANN-EHR layer*

The CNN architecture used to process the Fundus dataset in Figure 8 is a multi-layered convolutional model designed to extract spatial features from retinal images with an input resolution of 224 × 224 pixels and 3 color channels (RGB). The first stage consists of a Conv2D layer with 64 filters of 5 × 5 kernel size, which produces a feature map of size (220 × 220 × 64). This is followed by a MaxPooling2D layer (2 × 2) to reduce the spatial dimension to (110 × 110 × 64) and a Dropout layer to prevent overfitting. The second convolutional layer uses 128 filters (5 × 5), followed by pooling to reduce the size to (53 × 53 × 128), and dropout is applied again. This process is continued with the third Conv2D (256 filters) and pooling to a size of (24×24×256), followed by dropout, then the fourth Conv2D (512 filters) and pooling to a size of (10×10×512), and the final dropout at the feature extraction stage. All convolutional feature maps are then flattened into a 51,200-element vector for processing in the fully connected stage. The first dense stage has 64 neurons with ReLU activation to combine the main features. Dropout is applied again to maintain model generalization. Finally, a Dense output layer with 2 neurons and sigmoid activation is used for binary classification.

J. of Soft Computing and Data Mining Vol. 7 No. 1 (2025) p. 106-126

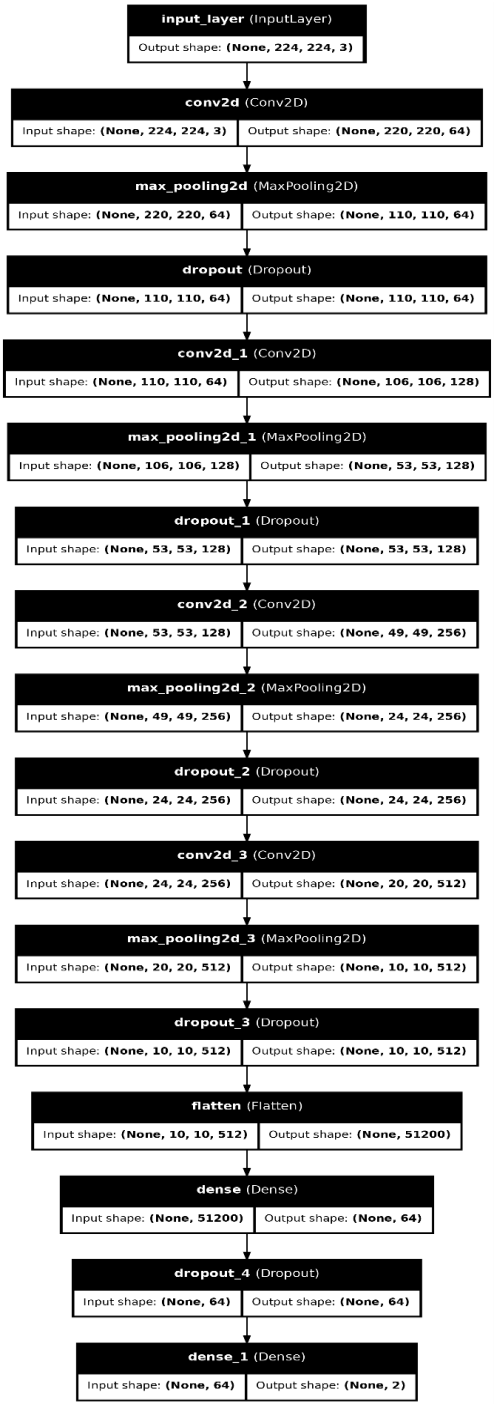


Fig. 8 *CNN-fundus layer proposed*

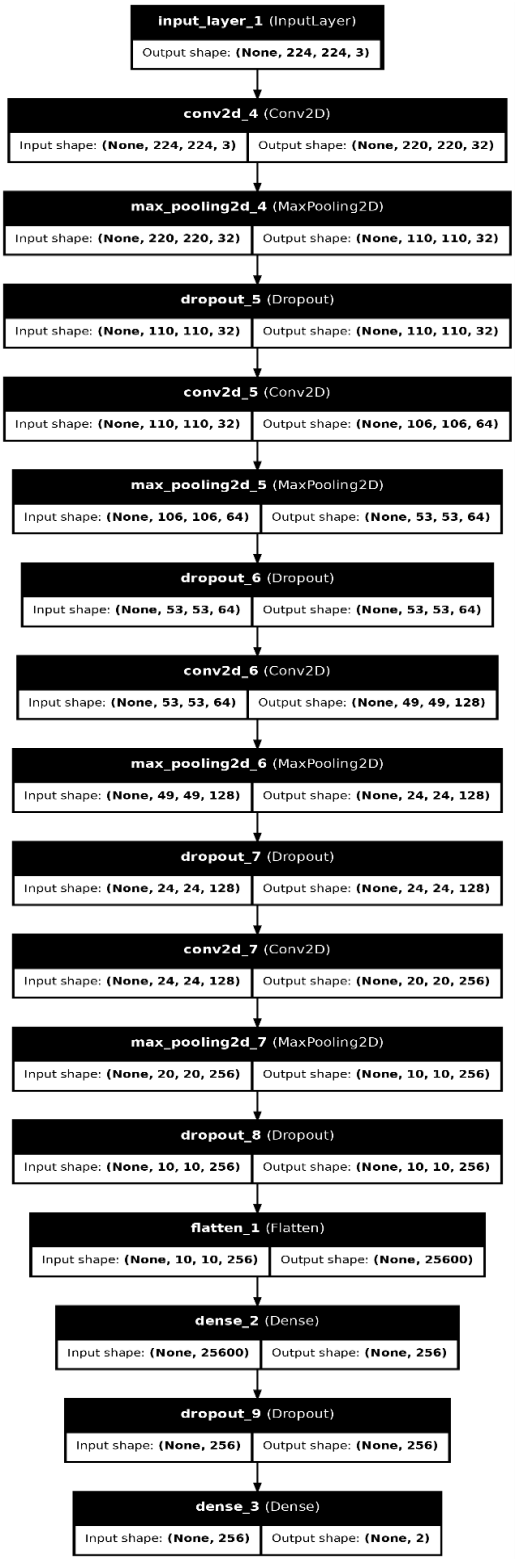


Fig. 9 *CNN-OCT scan layer proposed*

This architecture combines gradual depth convolution, pooling for dimensionality reduction, dropout for regularization, and dense layers for decision making, all of which are optimized for recognizing complex patterns in fundus images. The CNN architecture used for processing the OCT Scan dataset in the figure 9 is designed to extract visual features gradually through four convolutional blocks equipped with pooling and dropout layers to reduce overfitting. The model starts with an input layer of size (224, 224, 3) that receives a three-channel color OCT image. The first block consists of Conv2D with 32 filters of size (5×5) that produce an output of (220, 220, 32), followed by MaxPooling2D (2×2) that reduces the dimension to (110, 110, 32), and Dropout for

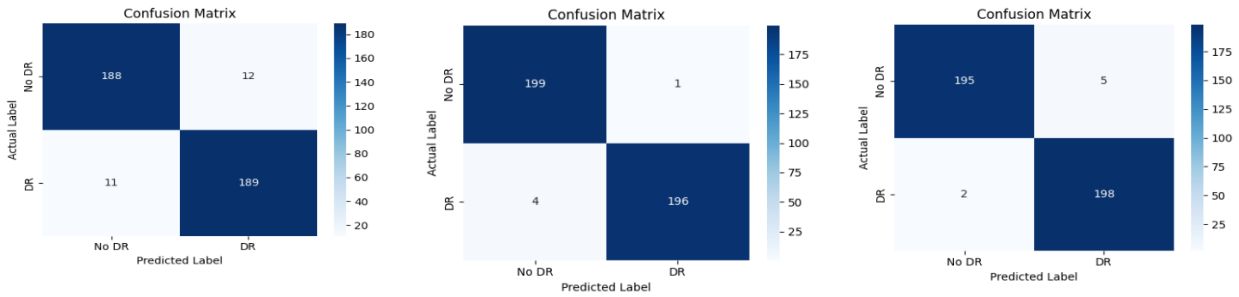
regularization. The second block uses Conv2D with 64 filters (5×5) producing (106, 106, 64), followed by pooling (53, 53, 64) and dropout. The third block has a Conv2D with 128 filters (5×5) with output (49, 49, 128), pooling (24, 24, 128), and dropout. The fourth block uses a Conv2D with 256 filters (5×5) resulting in (20, 20, 256), pooling (10, 10, 256), and a final dropout. All features are then converted into 1D vectors via Flatten (25600) before entering a Dense layer with 256-unit ReLU, followed by dropout, and a Dense output layer with 2 sigmoid units for binary classification.

The evaluation results as shown in table 3 of the unimodal model show that each model provides quite high performance but varies depending on the data modality. The ANN-EHR model ranks best with 98.25% accuracy, 97.50% specificity, 99% sensitivity, and 98.26% F1-Score, indicating that EHR tabular data is highly informative for DR classification compared to single images. CNN-Fundus recorded 94.25% accuracy, 94% specificity, 94.50% sensitivity, and 94.26% F1-Score, which although lower than EHR, still shows good performance in detecting visual features from retinal images. CNN-OCT Scan ranked second overall with 98.75% accuracy, 99.50% specificity, 98% sensitivity, and 98.74% F1-Score, indicating very high ability to capture retinal structural patterns relevant to DR.

Table 3 *Evaluation metrics result of unimodal model*

Single Modality	Accuracy	Specificity	Sensitivity	F1-Score
ANN-EHR	0.9825	0.975	0.99	0.9826
CNN-Fundus	0.9425	0.94	0.945	0.9426
CNN-OCT Scan	0.9875	0.995	0.98	0.9874

Overall, these performance differences demonstrate that each modality has its own unique strengths, with EHR excelling in clinical information, OCT excelling in structural detail, and Fundus providing significant contributions to visual detection, which provides a strong foundation for multimodal fusion in the subsequent stages.



(a) **(b)** **(c)**

Fig. 10 *Confusion matrix results (a) CNN-fundus model; (b) CNN-OCT scan model; (c) ANN-EHR model*

Figure 10 is a comparison of the confusion matrix of a) CNN-Fundus model, b) CNN-OCT Scan model, c) ANN-EHR model. Overall, ANN-EHR ranked top in terms of the balance of No DR and DR detection, with minimal errors in both classes. CNN-OCT ranked second, with superior specificity but slightly lower sensitivity compared to ANN-EHR. CNN-Fundus ranked third, with higher prediction errors, particularly in the No DR class, resulting in decreased specificity. These performance differences indicate that the quality and relevance of features in each modality significantly influence model effectiveness, with EHR and OCT providing clearer diagnostic signals than a single fundus image. Figure 11 shows the evaluation metrics results of each unimodal model.

Figure 12 shows the training results of the three unimodal models show different convergence and performance characteristics depending on the nature of the data used. a) The loss graph shows a significant downward trend in the first five epochs, from an initial value of around 50 to a lower, stable value at the end of training. The training loss and validation loss curves are close together, indicating the model is able to learn patterns without severe overfitting. The accuracy graph shows a sharp increase from around 0.5 to above 0.9 in the 5th epoch, then moves stably to near 0.94–0.95.

J. of Soft Computing and Data Mining Vol. 7 No. 1 (2025) p. 106-126

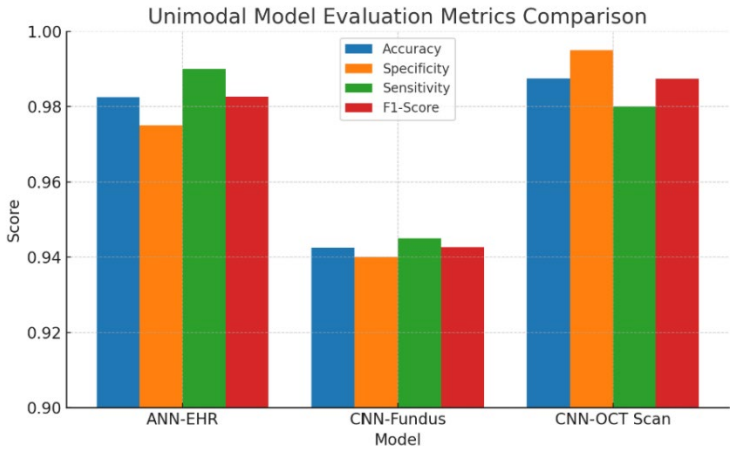
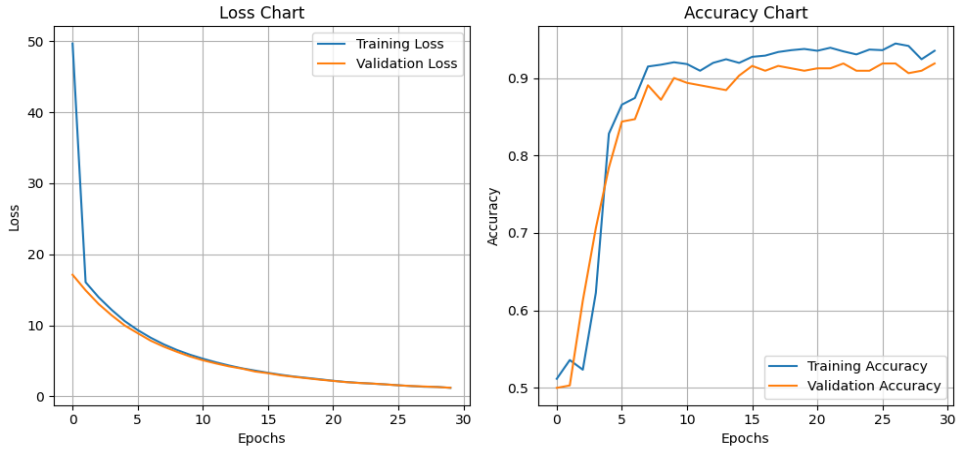
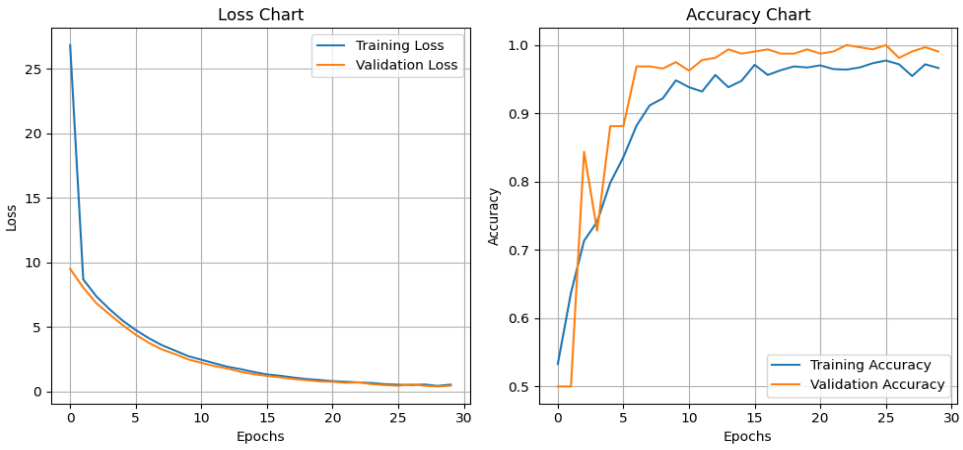


Fig. 11 *Unimodal model evaluation metrics comparison*

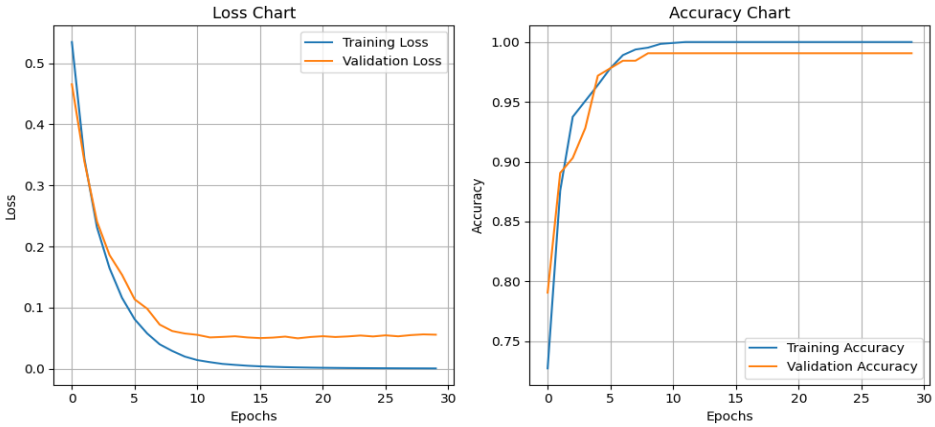
However, there are slight fluctuations in training and validation accuracy at the end of training, which may be caused by the complexity of visual variations in fundus images, b) For CNN-OCT Scan, the loss graph starts from a value of around 27 and experiences a rapid decrease in the first five epochs, then gradually levels off until the end of training.



(a)



(b)



(c)

Fig. 12 *Model loss and accuracy graph (a) CNN-fundus model; (b) CNN-OCT scan model; (c) ANN-EHR model*

The gap between the training loss and validation loss is very small, indicating good model generalization. In the accuracy graph, the increase occurs rapidly to nearly 1.0 in the 7th epoch, then stabilizes with minimal difference between training and validation accuracy. This model is able to achieve near-perfect performance, indicating that the CNN architecture used is very suitable for extracting features from OCT data that has a consistent visual structure. Meanwhile, c) ANN-EHR The loss graph for EHR starts at a much lower value (~0.65) than the image-based model, in line with the simpler and more structured nature of tabular data. The loss decreases very quickly, especially in the first epoch, and continues to decrease near zero until the end of training. The initial accuracy is already high (~0.75) and increases rapidly to nearly 1.0 in the first five epochs. The difference between training accuracy and validation accuracy is almost zero, indicating excellent stability and strong generalization capabilities of the ANN model for EHR data.

Overall, the three models all showed fast convergence and high accuracy, but they varied in their stability and complexity of learning. ANN-EHR performed best in terms of stability and training efficiency, whereas CNN-OCT performed best in the image data with a near perfect result. Despite its superior performance, CNN-Fundus still showed larger variations, suggesting difficulties in capturing complex variations in fundus images. The findings suggest that choosing an architecture that is more tailored to the data characteristics can lead to a better-performing model.

4.2 Multimodal Alignment Fusion

This section presents the implementation of label-driven alignment of multimodal Fundus, OCT, and EHR data as described in sub-chapter 3.2.1 to ensure that each modality has a consistent and balanced distribution in each label class consisting of DR and Non-DR so that the resulting fusion model is optimal. Furthermore, aligned modality will be modeled using four late fusion strategies: Simple Average, Weighted Average, Majority Voting, and Stacked Ensemble, which will be compared to the performance of the unimodal model in the next section.

J. of Soft Computing and Data Mining Vol. 7 No. 1 (2025) p. 106-126

Table 4 *Pseudocode multimodal alignment*

```
Algorithm 1: Label-Driven Alignment (EHR as Canonical Labels)

Inputs:
  Fundus[0..N-1]     # images
  OCT[0..N-1]        # images
  EHR_X[0..N-1], EHR_y[0..N-1]  # tabular features & labels (0=No_DR, 1=DR)
Assumptions:
  - All modalities have same length N
  - EHR_y is canonical; alignment is by index
  - (Optional) datasets may already be sorted so 0..K-1=DR, K..N-1=No_DR

# ----- Split indices stratified by EHR labels -----
function STRATIFIED_SPLIT(indices, labels, train_ratio, val_ratio):
    split by label proportions in labels
    return I_train, I_val, I_test

I_all  ← [0..N-1]
I_tr, I_val, I_te ← STRATIFIED_SPLIT(I_all, EHR_y, 0.64, 0.16)  # 64/16/20

# ----- Build aligned tuples by identical indices -----
function BUILD_ALIGNED_SET(I):
    A ← empty list
    for i in I:
        tuple ← ( Fundus[i], OCT[i], EHR_X[i], EHR_y[i] )
        append(A, tuple)
    return A

A_tr  ← BUILD_ALIGNED_SET(I_tr)
A_val ← BUILD_ALIGNED_SET(I_val)
A_te  ← BUILD_ALIGNED_SET(I_te)

# ----- Mini-batch generator (late fusion ready) -----
function NEXT_BATCH(A, batch_size, shuffle=True):
    idx ← [0..|A|-1]
    if shuffle: permute(idx)
    for b in range(0, |A|, batch_size):
        J ← idx[b : b+batch_size]
        F_batch  ← [A[j].Fundus  for j in J]
        O_batch  ← [A[j].OCT     for j in J]
        Xe_batch ← [A[j].EHR_X   for j in J]
        y_batch  ← [A[j].EHR_y   for j in J]
        yield (F_batch, O_batch, Xe_batch), y_batch
```

Table 4 is pseudocode for label-driven alignment with EHR labels as reference labels to align the Fundus dataset and OCT Scan dataset based on the index so that each record represents the same patient across all modalities. The data is then stratified into training, validation, and testing subsets to maintain a consistent proportion of labels in each subset. Next, data from each modality is merged into a structured tuple (Fundus, OCT, EHR_X, EHR_y) based on the resulting index, ensuring synchronization between modalities. This process is followed by a mini-batch generator that dynamically provides randomized data batches for the training process, maintaining the diversity of the input sequence.

Table 5 *Late fusion result comparation*

Late Fusion	Accuracy	Specificity	Sensitivity	F1-Score
Average	0.975	1.000	0.95	0.997
Weighted Average	0.9575	1.000	0.915	0.997
Majority Voting	0.975	1.000	0.95	0.974
Stacked Ensemble	0.9575	0.925	0.99	0.958

Based on Table 5, all four late fusion approaches performed very well with high accuracy, specificity, sensitivity, and F1-score values. The Average and Majority Voting methods had the highest accuracy (0.975) with perfect specificity (1.000) and sensitivity of 0.950, indicating a balance between the ability to consistently detect patients without DR and with DR. Weighted Average was slightly lower in accuracy (0.9575) and sensitivity (0.915), although specificity remained perfect, indicating this method tends to be more conservative in classifying positive DR cases. Conversely, Stacked Ensemble stood out in the highest sensitivity (0.99) but experienced a decrease in specificity (0.925), meaning it was superior in detecting positive DR cases but slightly sacrificed accuracy in negative cases.

Analytical results show that the Average and Majority Voting methods are suitable when a good balance between DR and non-DR detection is needed without significantly sacrificing either metric. Weighted Average can be considered in situations where reducing false positives is a priority, while Stacked Ensemble is more suitable for early screening scenarios that require high sensitivity to avoid missing DR cases, despite the risk of increasing false positives. These results indicate that the choice of late fusion strategy should be tailored to the desired clinical or operational objectives, as small differences in metrics can have significant implications in real-world medical settings.

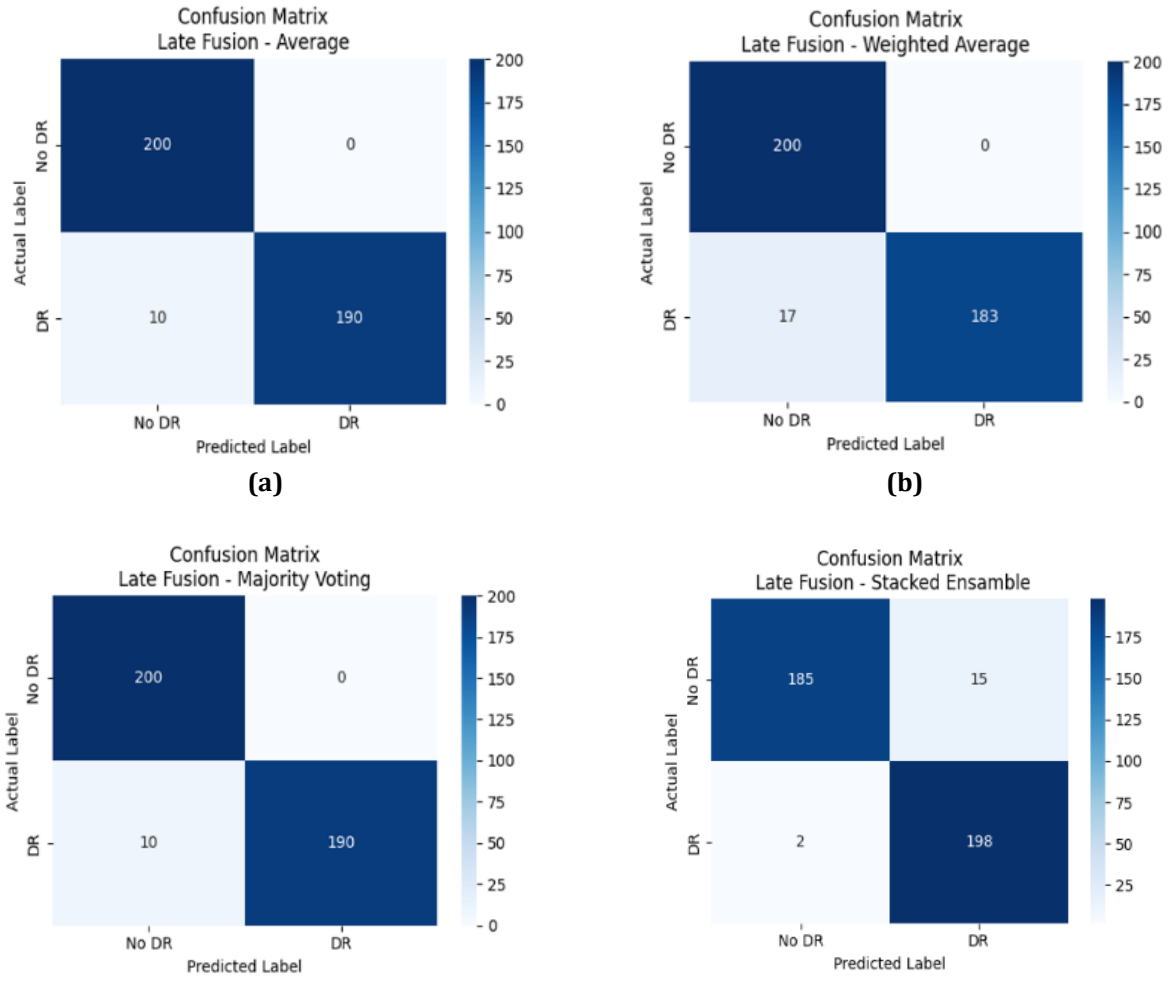


Fig. 13 *Confusion matrix of late fusion model (a) Simple average; (b) Weighted average; (c) Majority voting; (d) Stacked ensemble*

J. of Soft Computing and Data Mining Vol. 7 No. 1 (2025) p. 106-126

Figure 13 shows four confusion matrices for the Late Fusion method, the Simple Average model demonstrated very high performance in the No DR class, with perfect True Negatives (200) and no False Positives (0). However, there were 10 False Negative cases where DR was predicted as No DR. This indicates slightly reduced sensitivity compared to perfect specificity. Weighted Average. The performance pattern was like the Average method in the No DR class (200 TN, 0 FP), but the number of False Negatives was higher (17 cases). This means that while specificity remained perfect, this method missed more DR cases than the Average method, resulting in decreased sensitivity. Majority Voting

This method had identical results to the Average method, with TN = 200, FP = 0, and FN = 10. Thus, the accuracy level and distribution of prediction errors were similar, indicating that Majority Voting performed equivalently to Average Fusion in this case. Stacked Ensemble differs from the previous three methods, showing a slight decrease in specificity due to 15 False Positives (the No DR class was predicted as DR). However, the number of False Negatives is very low (only 2 cases), making it the method with the highest sensitivity among the four late fusion strategies. All three methods (Average, Weighted Average, Majority Voting) excel in maintaining perfect specificity, but the Stacked Ensemble method excels in sensitivity. The choice of method depends on priorities: if the goal is to minimize False Positives (e.g., to avoid unnecessary follow-up examinations), then Average or Majority Voting is more appropriate. However, if the main goal is to minimize False Negatives (to avoid missing DR cases), Stacked Ensemble is a better choice even though it sacrifices a little specificity.

Based on the F1-Score comparison shown in Figure 14, it can be seen that most of the multimodal alignment late fusion approaches, especially Simple-Average and Weighted Average, are able to achieve very high F1-Score values (0.999), outperforming all unimodal models, including ANN-EHR (0.9826) and CNN-OCT Scan (0.9874). This indicates that from the perspective of the balance between precision and sensitivity (recall), certain late fusion strategies can provide almost perfect results in detecting DR. Although CNN-Fundus has the lowest score (0.9426), multimodal integration successfully compensates for this weakness by combining the strengths of all three modalities.

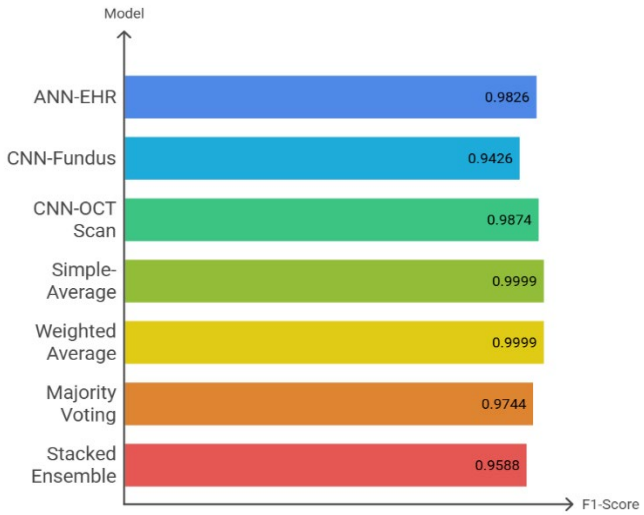


Fig. 14 *F1-Score model comparison unimodal vs multimodal alignment fusion*

5. Discussion

Simple Average provides a baseline by equally weighing all modalities, offering robustness when no single source dominates performance. Weighted Average allows emphasis on stronger modalities; weights were tuned on the validation set via a grid search over fixed weight combinations (e.g. 0.2, 0.3, 0.5), enabling trade-offs between sensitivity and specificity depending on modality reliability. Majority Voting is preferable when outputs are highly confident but potentially noisy in one modality, as it reduces the impact of outliers while maintaining class balance. Stacked Ensemble, in contrast, trains a meta-classifier on validation predictions, enabling more complex interactions and typically improving sensitivity (e.g., achieving 0.99 in our results), though sometimes at the cost of specificity.

However, not all late fusion methods provide optimal results — for example, Majority Voting and Stacked Ensemble lag slightly behind, likely due to sensitivity to class distribution or noise in the predictions of one of the modalities. Overall, this study implies that in real-world scenarios requiring high precision and sensitivity simultaneously, simple and weighted averaging late fusion methods exhibit more stable and favorable results than any single unimodal approaches.

This research conducted a detailed evaluation of both unimodal baselines and the best-performing fusion strategy as shown in table 6. We compared the CNN-OCT model—identified as the strongest unimodal benchmark—with the Simple Average fusion model that integrates predictions across fundus, OCT, and EHR modalities. We highlight not only the overall accuracy (ACC), but also per-class precision, recall, F1-scores, and the macro-F1 to consider the variation of class imbalance and the robustness. ROC-AUC and PR-AUC values with corresponding 95% confidence intervals were also added in order to have a strict evaluation of discrimination and calibration. This general evaluation permits us to verify if fusion truly takes advantage of complementary cues between modalities, and to confirm the effectiveness of label-driven alignment in achieving clinically relevant improvements.

Table 6 *Per-class performance for the best unimodal model (CNN-OCT) and the best late fusion variant (simple average)*

Model	Class 0 (No_DR) – P / R / F1	Class 1 (DR) – P / R / F1	Macro-F1	ROC-AUC (95% CI)	PR-AUC (95% CI)
CNN-OCT	0.994 / 0.800 / 0.886	0.833 / 0.995 / 0.907	0.897	0.998	0.998
Simple Average	0.995 / 1.000 / 0.998	1.000 / 0.995 / 0.997	0.997	1.000	1.000

The CNN-OCT unimodal model achieved very high AUCs (ROC-AUC = 0.998, PR-AUC = 0.998) but showed an imbalance between classes, with lower recall for No_DR (0.80) compared to almost perfect recall for DR (0.995). The Simple Average fusion improved performance across both classes, reaching near-perfect balance in precision and recall (≈ 0.995–1.000) and yielding a macro-F1 of 0.997, with ROC-AUC and PR-AUC = 1.000. This demonstrates that multimodal fusion not only boosts overall accuracy but also addresses per-class imbalance, which is crucial for real-world DR screening.

The comparative analysis of unimodal and fusion models demonstrates that while unimodal architectures such as ANN-EHR and CNN-OCT achieved strong performance (F1-scores of 0.9826 and 0.9874, respectively), they exhibited class-specific weaknesses, with OCT underperforming in recall for the No_DR class and EHR occasionally misclassifying structural retinal changes. In contrast, late fusion, particularly Simple and Weighted Averaging—achieved near-perfect balance across both classes, with per-class precision and recall approaching 1.0 and a macro-F1 of 0.997, thereby reducing both false positives and false negatives. Confusion matrix analysis corroborates this finding, showing that averaging-based fusion consistently minimizes errors compared to Voting or Stacked ensembles, which either reduce specificity or inflate sensitivity. Importantly, this study clarifies the F1 discrepancy noted by reviewers: earlier reports used micro-F1, which masked class-level imbalances; here, per-class metrics and macro-F1 provide a fairer assessment of performance, revealing that averaging fusion not only improves overall robustness but also harmonizes detection across DR and No_DR categories. These results highlight the value of label-driven alignment combined with late fusion in overcoming the asymmetries of unimodal models and in moving toward clinically promising yet methodologically transparent AI solutions.

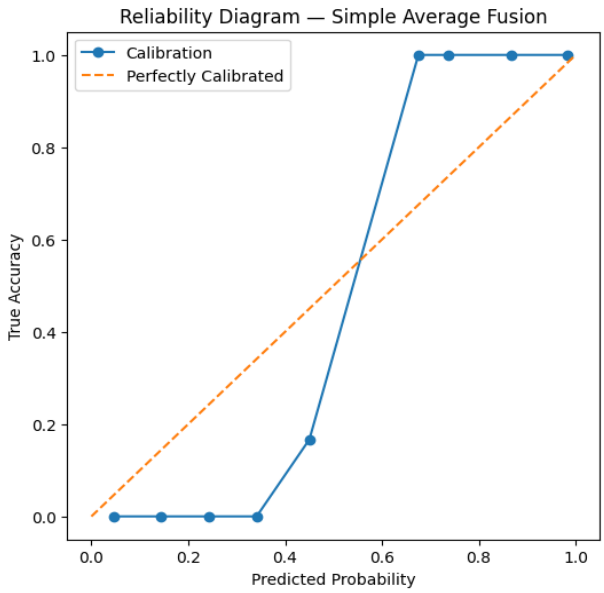


Fig. 15 *Reliability diagram for the Simple Average late fusion model*

Figure 15 shows The Simple Average fusion model demonstrated strong calibration at higher predicted probabilities (>0.7), closely following the ideal diagonal in the reliability diagram. Calibration error at low probability bins was observed, but since clinical thresholds typically operate at ≥0.5, the model's probability outputs are reliable for screening. A default threshold of 0.5 was used, though threshold tuning on validation could further optimize sensitivity–specificity trade-offs.

6. Conclusion

Experimental results confirm that the proposed multimodal alignment with the late fusion framework provides a robust and clinically relevant approach for detecting diabetic retinopathy (DR). Label-based alignment plays a crucial role in ensuring that fundus images, OCT scans, and EHR data are accurately aligned based on patient labels before being fed into their respective models. This preprocessing step eliminates inconsistencies common in unpaired datasets, allowing each modality to be modeled according to its unique characteristics—CNN for image modalities and ANN for tabular EHRs—without compromising data integrity. Thus, alignment ensures that fused predictions are derived from semantically coherent patient records, which is crucial for the reliability of the final diagnosis. The experimental results also demonstrate that the proposed multimodal alignment with the late fusion framework provides a robust and promising approach for detecting diabetic retinopathy (DR).

While the unimodal models achieved strong results, particularly ANN-EHR and CNN-OCT Scan, in terms of accuracy, the integrated final fusion strategies—specifically Simple Average and Weighted Average—demonstrated superior F1-score performance (0.999). This metric reflects an excellent balance between precision and recall, aligning with the novel goal of this study: producing a detection model that minimizes both false positives and false negatives. Further confusion matrix analysis demonstrated that these fusion strategies maintained near-perfect specificity and high sensitivity, confirming their ability to leverage complementary diagnostic cues across modalities.

Overall, this study highlights that the combination of multimodal label-based alignment and final fusion, particularly the average-based approach, yields a statistically robust and promising predictive system for the research pipeline. However, clinical validation with paired patient-level multimodal data remains an important direction for future work to fully establish its applicability in real-world ophthalmology practice.

Acknowledgement

The authors would like to thank to Politeknik Caltex Riau and the Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia for its support.

Conflict of Interest

Authors declare that there is no conflict of interest regarding the publication of the paper.

Declaration of AI Use in Manuscript Preparation

The authors used GPT and Grammarly to assist in grammar checking and language editing. All content generated was reviewed and verified by the authors, who take full responsibility for the final submission.

Author Contribution

Kartina Diah KW: ***Conceptualization, data curation, formal analysis, methodology, code implementation, writing original draft, review, editing, submission, and review reports****. Shahreen Kasim:* ***Methodology review, editing, and co-authorship of the manuscript****. Wawan Yunanato:* ***Data preprocessing, methodology, code implementation, article review****. Deshinta Arrova Devi:* ***Article review****. Rohayanti Hassan:* ***Article review****. Sheeba Armoogum:* ***Article review****. Mohammad Syafwan Bin Arshad:* ***Article review****.*

References

[1] Coney JM (2019 Oct) Addressing unmet needs in diabetic retinopathy, *Am J Manag Care,25*(16 Suppl):S311-S316. PMID: 31747241.

[2] Nanegrungsunk, Onnisa, Direk Patikulsila, and Srinivas R. Sadda (2022), Ophthalmic imaging in diabetic retinopathy : A review, *Clin. Experiment. Ophthalmol.*, no. June, pp. 1082–1096, doi: 10.1111/ceo.14170.

[3] W. Sun, X. An, Y. Zhang, X. Kang, L. Jiang, and H. Ji (2023), The ideal treatment timing for diabetic retinopathy : the molecular pathological mechanisms underlying early- stage diabetic retinopathy are a matter of concern, *Front. Endocrinol. (Lausanne).*, no. November, pp. 1–19, doi: 10.3389/fendo.2023.1270145.

[4] R. Sim and C. Hern(2022) New Insights into Treating Early and Advanced Stage Diabetic Retinopathy, *Int. J. Mol. Sci.*, vol. 23, pp. 1–13, doi: https://doi.org/10.3390/ijms23158513.

[5] P. H. Scanlon(2017) The English National Screening Programme for diabetic retinopathy 2003 – 2016, *Acta Diabetol.*, vol. 54, no. 6, pp. 515–525, 2017, doi: 10.1007/s00592-017-0974-1.

[6] V. Bellemo *et al.* (2019) Artificial Intelligence Screening for Diabetic Retinopathy : the Real-World Emerging Application, *Curr. Diab. Rep.*, vol. 19, 72, doi: https://doi.org/10.1007/s11892-019-1189-3.

[7] B. Keerthi, K. P. K, P. Narendra, and N. Karthik. (2024). Diabetic Retinopathy Detection Using Deep Learning. *2024 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT)*, pp. 1–5. doi: 10.1109/IConSCEPT61884.2024.10627921.

[8] G. Zhang *et al* (2022) Automated multidimensional deep learning platform for referable diabetic retinopathy detection : a multicentre , retrospective study, *BMJ Open*, vol. 12:e060155, pp. 1–10, doi: 10.1136/bmjopen-2021-060155.

[9] F. Li, Z. Liu, H. Chen, M. Jiang, X. Zhang, and Z. Wu (2019) Automatic Detection of Diabetic Retinopathy in Retinal Fundus Photographs Based on Deep Learning Algorithm, *Transl. Vis. Sci. Technol.*, vol. 4, no. 8, doi: https://doi.org/10.1167/tvst.8.6.4.

[10] L. Dai *et al* (2021)A deep learning system for detecting diabetic retinopathy across the disease spectrum, *Nat. Commun.*, vol. 12:3242, pp. 1–11, doi: 10.1038/s41467-021-23458-5.

[11] I. Qureshi, J. Ma, and Q. Abbas (2021) Diabetic retinopathy detection and stage classification in eye fundus images using active deep learning, *Multimedia Tools and Applications*, vol. 80, no. 8, doi: 10.1007/s11042-020-10238-4.

[12] K. Parthiban and M. Kamarasan (2023) Diabetic retinopathy detection and grading of retinal fundus images using coyote optimization algorithm with deep learning, *Multimed. Tools Appl.*, vol. 82, no. 12, pp. 18947–18966, doi: 10.1007/s11042-022-14234-8.

[13] A. Sungheetha and R. Sharma R (2021) Design an Early Detection and Classification for Diabetic Retinopathy by Deep Feature Extraction based Convolution Neural Network, *J. Trends Comput. Sci. Smart Technol.*, vol. 3, no. 2, pp. 81–94, doi: 10.36548/jtcsst.2021.2.002.

[14] Y. Kale and S. Sharma (2023) Detection of five severity levels of diabetic retinopathy using ensemble deep learning model, *Multimed. Tools Appl.*, vol. 82, no. 12, pp. 19005–19020, doi: 10.1007/s11042-022-14277-x.

[15] M. H. Mahmoud, S. Alamery, H. Fouad, A. Altinawi, and A. E. Youssef (2023) An automatic detection system of diabetic retinopathy using a hybrid inductive machine learning algorithm, *Pers. Ubiquitous Comput.*, vol. 27, no. 3, pp. 751–765, doi: 10.1007/s00779-020-01519-8.

[16] X. Li, L. Shen, M. Shen, F. Tan, and C. S. Qiu (2019) Deep learning based early stage diabetic retinopathy detection using optical coherence tomography, *Neurocomputing*, vol. 369, pp. 134–144, doi: 10.1016/j.neucom.2019.08.079.

[17] M. Hsu *et al* (2021) Deep Learning for Automated Diabetic Retinopathy Screening Fused With Heterogeneous Data From EHRs Can Lead to Earlier Referral Decisions, *Transl. Vis. Sci. Technol.*, vol. 10(9):18, pp. 1–10, doi: https://doi.org/10.1167/tvst.10.9.18.

[18] Z. Zhang and C. Deng (2024) Advances in Structural and Functional Retinal Imaging and Biomarkers for Early Detection of Diabetic Retinopathy, *biomedicines*, vol. 12:1405, pp. 1–28, doi: https://doi.org/ 10.3390/biomedicines12071405.

[19] Li, Y. *et al.* (2022). Multimodal Information Fusion for Glaucoma and Diabetic Retinopathy Classification. *Ophthalmic Medical Image Analysis. OMIA 2022*. Lecture Notes in Computer Science, vol 13576. Springer, Cham. https://doi.org/10.1007/978-3-031-16525-2_6

[20] M. El *et al* (2023) Improved Automatic Diabetic Retinopathy Fusion of UWF-CFP and OCTA Images, in *Ophthalmic Medical Image Analysis*, pp. 1–11. doi: 10.1007/978-3-031-44013-7_2.

[21] L. F. Nakayama *et al* (2021)The Challenge of Diabetic Retinopathy Standardization in an Ophthalmological Dataset, *J. Diabetes Sci. Technol.*, vol. 15(6), no. 1410–1411, doi: 10.1177/19322968211029943.

[22] D. Restrepo, L. A. Celi, and U. Cauca (2024) DF-DM : A foundational process model for multimodal data fusion in the arti cial intelligence era, *Res. Sq.*, pp. 0–38, doi: https://doi.org/10.21203/rs.3.rs-4277992/v1.

[23] K. Kim (2024) Pseudo Multi-Modal Approach to LiDAR Semantic Segmentation, *Sensors*, vol. 24,7840, doi: https://doi.org/ 10.3390/s24237840.

[24] W. Han, W. Jiang, J. Geng, and W. Miao (2025) Difference-Complementary Learning and Label Reassignment for Multimodal Semi-Supervised Semantic Segmentation of Remote Sensing Images, *IEEE Trans. Image Process.*, vol. 34, no. c, pp. 566–580, doi: 10.1109/TIP.2025.3526064.

[25] K. D. Kesuma Wardhani, S. Kasim, A. Erianda, and R. Hassan (2024) Deep Learning-based Method in Multimodal Data for Diabetic Retinopathy Detection, *IJASEIT*, vol. 14, no. 5, pp. 1602–1608, doi: 10.18517/ijaseit.14.5.11677.

[26] S. Karthikeyan, S. S. S, S. Gg, R. Sivasanjeev, and M. Srimathi (2024). Multimodal Approach for Diabetic Retinopathy Detection using Deep Learning and Clinical Data Fusion, *2024 9th Int. Conf. Commun. Electron. Syst.*, pp. 1702–1706, doi: 10.1109/ICCES63552.2024.10859974.

[27] S. El-ateif and A. Idri (2022) Single-modality and joint fusion deep learning for diabetic retinopathy diagnosis, *Sci. African*, vol. 17, pp. 1–17, doi: 10.1016/j.sciaf.2022.e01280.

[28] Y. Li *et al* (2022) Multimodal Information Fusion for Glaucoma and Diabetic Retinopathy Classification, *Ophthalmic Medical Image Analysis: 9th International Workshop, OMIA 2022*, pp. 53–62. doi: 10.1007/978-3-031-16525-2_6.

[29] S. Karthikeyan, S. S. S, S. Gg, R. Sivasanjeev, and M. Srimathi (2024) Multimodal Approach for Diabetic Retinopathy Detection using Deep Learning and Clinical Data Fusion, in *2024 9th International Conference on Communication and Electronics Systems (ICCES)*, pp. 1702–1706. doi: 10.1109/ICCES63552.2024.10859974.

[30] P. K. Darabi (2024) Diagnosis of Diabetic Retinopathy, MIT CC BY 4.0, Tehran, Iran, doi: 10.13140/RG.2.2.13037.19688.

[31] P. Gholami, P. Roy, and V. Lakshminarayanan (2018) Diabetic Retinopathy Retinal OCT Images. Borealis, Canada, doi: doi:10.5683/SP/FLGZZE.

[32] X. Zhuang *et al* (2019) Data from: Association of diabetic retinopathy and diabetic macular edema with renal function in southern Chinese patients with type 2 diabetes mellitus: a single-center observational study. doi: https://doi.org/10.5061/dryad.6kg1sd7.

[33] V. Yulyanti, H. Adi, I. Ardiyanto, and W. Kz (2019)Dark lesion elimination based on area , eccentricity and extent features for supporting haemorrhages detection, *Commun. Sci. Technol.*, vol. 4, no. 1, pp. 7–11, doi: https://doi.org/10.21924/cst.4.1.2019.110.

[34] R. Roy, K. Saurabh, N. R. Thomas, M. Chowdhury, and D. K. Shah (2018) Validation of Multicolor Imaging of Diabetic Retinopathy Lesions Vis a Vis Conventional Color Fundus Photographs, *Ophthalmic Surg. Lasers Imaging Retina*, no. 147, pp. 8–16, doi: 10.3928/23258160-20181212-02.

[35] G. Sivapriya, V. Praveen, S. Saranya, R. Surya, and S. Sweetha (2023) Detection and Segmentation of Retinopathy Diseases using EAD-Net with Fundus Images, *2023 Int. Conf. Comput. Commun. Informatics*, pp. 1–7, doi: 10.1109/ICCCI56745.2023.10128343.

[36] W. M. Amoaku *et al* (2020)Diabetic retinopathy and diabetic macular oedema pathways and management : UK Consensus Working Group, *Eye*, vol. 34:1, pp. 1–51, doi: 10.1038/s41433-020-0961-6.

[37] D. Bartsch, S. F. Oster, and R. William (2011) Correlation between morphological features on spectral domain Optical Coherence Tomography and Angiographic leakage patterns in macular edema, *Retina*, vol. 30, no. 3, pp. 383–389, doi: 10.1097/IAE.0b013e3181cd4803.Correlation.

[38] A. Couturier, V. Mane, C. A. Lavia, and R. Tadayoni (2022) Hyperreflective cystoid spaces in diabetic macular oedema : prevalence and clinical implications, *Br. J. Ophthalmol.*, vol. 106, pp. 540–546, doi: 10.1136/bjophthalmol-2020-317191.

[39] G. Panozzo *et al* (2004) Diabetic macular edema : an OCT-based classification, *Semin. Ophthalmol.*, vol. 19, pp. 13–20, doi: 10.1080/08820530490519934.

[40] S. El-ateif (2023) Fundus Unimodal and Late Fusion Multimodal Diabetic Retinopathy Grading," in *In Proceedings of the 12th International Conference on Data Science, Technology and Applications (DATA 2023)*, pp. 219–226. doi: 10.5220/0012048700003541.

[41] V. S. Tseng et al (2020)Leveraging Multimodal Deep Learning Architecture with Retina Lesion Information to Detect Diabetic Retinopathy, *Transl. Vis. Sci. Technol.*, vol. 9(2):41, pp. 1–12, doi: https://doi.org/10.1167/tvst.9.2.41.

[42] Y. Yan et al (2021) Longitudinal Detection of Diabetic Retinopathy Early Severity Grade," *Springer Nat.*, vol. 3, pp. 11–20, doi: https://doi.org/10.1007/978-3-030-87000-3_2.