

Attention-Based Multi-Model Fusion With Hierarchical Multi-Task Learning For Enhanced Diabetic Retinopathy Detection

Jayaprakash Venugopal^{1*}, Kalaivani Kathirvelu²

^{1*}Department of Computer Science and Engineering, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, India, Email: pinkyprakash@gmail.com

²Department of Computer Science and Engineering, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, India, Email: kalai.se@vistas.ac.in

Corresponding Author ;Jayaprakash Venugopal

Department of Computer Science and Engineering, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, India, Email: pinkyprakash@gmail.com

ABSTRACT

Diabetic retinopathy is a leading cause of blindness worldwide, necessitating accurate and timely detection through automated systems. This research introduces two complementary novelties to improve DR classification from fundus images using the APTOS retinal dataset. First, we propose an attention-based multi-model fusion approach that adaptively combines features from ResNet50 and Xception architectures using learned attention weights that vary per image, enabling the model to dynamically determine when to trust which network based on image characteristics. Second, we implement a hierarchical multi-task learning framework with three clinically-motivated levels: primary DR grading, binary referable classification, and multi-label lesion detection. Our experimental results demonstrate that the synergistic combination of these novelties achieves 5-7% accuracy improvement over baseline ResNet50, reaching approximately 80% accuracy with a Quadratic Weighted Kappa score of 0.82-0.85. The proposed system not only achieves higher accuracy but also provides interpretable, actionable insights aligned with clinical practice through attention weights and auxiliary task predictions.

Keywords: diabetic retinopathy, deep learning, attention mechanism, multi-model fusion, multi-task learning, convolutional neural networks, medical image analysis, ResNet50, Xception

1. INTRODUCTION

One of the leading causes of preventable blindness in working-age adults globally is diabetic retinopathy, a microvascular disease related to diabetes mellitus that manifests in the retina [1]. According to the International Diabetes Federation, 537 million adults worldwide had diabetes in 2021, and by 2045, that number is expected to rise to 783 million [2]. DR among diabetic patients is between 35% and 40% and it is very important that it is identified early and appropriate intervention is administered to avoid the loss of vision in the patient [3].

The manual analysis of retinal fundus pictures by skilled ophthalmologists is the foundation of conventional DR screening, which is laborious, expensive, and subject to inter-observer variability [4]. The rising burden of diabetes in the world has left a significant disparity between the screening need and access to a specialist resource especially in the underserved and rural regions [5]. This has been a challenge that has prompted the creation of computer-aided diagnostic systems that can automatically identify and rate DR using retinal images. Recent developments in the field of deep learning and specifically in convolutional neural networks have demonstrated encouraging findings in automated DR detection [6], [7]. Nonetheless, a number of difficulties still exist: single CNN models might not be able to extract all useful features of complicated medical images [8], the interpretable predictions should be required, which conforms to clinical practises [9], and systems should be able to provide actionable information beyond a simple severity grading [10]. To cope with these difficulties,

innovative methods that integrate various complementary architectures and use clinical domain expertise in the learning process are needed.

The study will counter these shortcomings by suggesting a new deep learning architecture that combines attention-based multi-model fusion and hierarchical multi-task learning. Our method builds on the complementary abilities of various CNN architectures and at the same time acquires a variety of clinically relevant tasks, leading to more robust and interpretable DR detection functions.

This research is driven by a number of imperative observations and clinical needs. The increasing cases of diabetes and the lack of qualified ophthalmologists, particularly in the developing world, render the need to develop automated screening solutions capable of identifying DR correctly in magnitude [11]. Diabetic patients fail to access specialist care, and therefore, up to 95% of cases of vision loss can be avoided with an early diagnosis [12]. Various CNNs are good at extracting various kinds of features of medical images [13]. The deep residual connexions of ResNet50 are good at global context and semantic feature, whereas the depth wise separable convolution of Xception is good at small details and lesion boundaries [14] and [15], respectively. Using one architecture might overlook valuable patterns that might be significant to accurate diagnosis.

Ophthalmologists do not simply give a severity grade to retinal images but adhere to a systematic process of decision-making that involves the general severity, referral urgency, and specific pathological features [16]. This clinical workflow should be replicated by automated systems, which should be really helpful in practice. Healthcare applications require intuitive predictions, which can be interpreted and believed in by the clinicians [17]. It is not enough to give a score of classification because the system must elaborate on how it reached its decision and focus on the characteristics that affected its predictions. The APTOS retinal database is filled with various images of varying conditions with different types of cameras and quality of images [18]. This natural variation makes it necessary to have powerful techniques that are capable of making changes that suit various image attributes instead of making uniform assumptions about all images.

The primary problem addressed in this research can be formally stated as developing an automated system that can accurately classify the severity of diabetic retinopathy into one of five grades while simultaneously providing clinically relevant auxiliary predictions with improved accuracy and interpretability compared to existing single-model approaches. This problem encompasses several specific challenges including feature representation, multi-scale learning, clinical alignment, imbalanced data, and image quality variability [19].

II. RELATED WORKS

Recent years have seen an exponential increase in interest in the field of deep learning-based automated diabetic retinopathy identification. This part will discuss the pertinent literature in three major fields, namely, the single CNN architectures to detect DR, ensemble and fusion models, and multi-task learning to medical imaging.

There are many studies which use pre-trained CNN architectures to DR classification. Gulshan et al. showed that InceptionV3 can be used to detect referable DR with high sensitivity and specificity by deep learning algorithms, and its performance is similar to that of ophthalmologists [20]. The authors of Pratt et al. compared several CNN models such as VGGNet, AlexNet, and GoogLeNet on the Kaggle DR dataset and concluded that the ensemble models that combined several models outperformed single models [21]. The application of ResNet architectures in DR detection has become very popular because residual connections make it possible to train very deep networks [22]. Li et al. used

ResNet50 with attention mechanism that enables the focus of lesion regions and the sensitivity of referable DR was 94% [23]. Equally, Xception architecture featuring depthwise separable convolutions has demonstrated potential in fine-grained features that are necessary in identifying small lesions such as microaneurysms [24]. Although these single-architecture solutions have delivered remarkable performance, they are constrained by the inherent biases and feature extraction capacities of specific networks [25]. Those architectures have their own strengths and weaknesses: ResNet has the ability to see the big picture, whereas Xception is better at detecting small details [26]. This observation prompts the necessity of approaches that would be able to balance these complementary strengths.

Conventional ensemble techniques to detect DR usually involve the combination of predictions of different models with fixed weighting schemes [27]. Wang et al. suggested voting-based ensemble of 5 CNN models that had better accuracy than single models but consumed a lot of computational power in inference [28]. Abramoff et al. established an ensemble system, which was approved by the FDA to autonomously detect DR, and they showed the clinical feasibility of the multi-model methods [29]. The conventional ensembles, however, are limited by the following: they are the average or the majority at the prediction level, they use constant weights independent of the features of the input, and they have to be inferred independently in each model [30]. Recent literature has considered learned fusion techniques, although the vast majority are simple concatenation or weighted averaging without an adaptive, per-image weighting scheme [31].

Attention mechanisms have gained popularity in medical image analysis. Jetley et al. introduced attention mechanisms for general image classification, demonstrating that machines are capable of learning to focus on specific regions [32]. Schlemper et al. used attention gates in U-Net for cardiac segmentation in the medical field [33]. However, attention-based fusion at the feature level for DR detection, where attention weights determine the contribution of different architectures, remains underexplored [34].

Through leveraging shared representations across related tasks, multi-task learning has demonstrated potential in medical imaging [35]. Liu et al. demonstrated that joint learning of disease classification and lesion segmentation improves performance on both tasks in chest X-ray analysis [36]. Yan et al. demonstrated increased grading accuracy through auxiliary segmentation tasks in their multi-task framework for joint DR grading and lesion segmentation [37]. However, most existing multi-task approaches for DR detection focus on only two tasks and do not structure their task hierarchy to mirror clinical decision-making workflows [38]. Additionally, task weighting is often treated as a hyperparameter to be tuned rather than being informed by clinical importance and relationships between tasks [39].

Despite significant progress, several gaps remain in the literature including lack of adaptive fusion mechanisms that learn per-image weights based on image characteristics [40], limited exploration of hierarchical multi-task frameworks that mirror clinical decision workflows with more than two tasks [41], and insufficient investigation of the synergistic effects of combining adaptive fusion with multi-task learning [42]. Our research addresses these gaps by proposing an integrated framework that combines attention-based adaptive fusion of complementary architectures with a three-level hierarchical multi-task learning approach informed by clinical practice.

III. PROPOSED METHOD

Our proposed framework integrates two complementary novelties: attention-based multi-model fusion for adaptive feature combination and hierarchical multi-task learning

for clinically-aligned supervision. This section outlines the architectural elements, our learning strategy, and details of implementation. The general architecture is made up of three primary modules namely dual backbone feature extraction, attention based fusion module and hierarchical multi task prediction heads [43]. With an input fundus image of size 512x512x3, the image is run through two pre-trained CNN backbones at the same time; ResNet50 and Xception.

A. Attention-Based Multi-Model Fusion

The first novelty is the problem of successfully blending characteristics of more than one architecture with the help of learning of adaptive weights, which change with each image. ResNet50 is an effective network that captures context globally, and rich semantic content due to its skip connections that enable flow of gradient [44]. The network generates outputs of its last convolutional layer. Xception is also more effective in fine-grained details and lesion borders detection with an effective feature extraction based on decoupled spatial and channel-wise convolutions [45]. Prior to fusion, we use Convolutional Block Attention Module (CBAM) on a branch scale [46]. CBAM sequentially applies channel attention and spatial attention, generating attention maps by aggregating spatial information through both average and max pooling operations [47].

Unlike traditional approaches that use fixed weights, our fusion module learns to dynamically weight each model's contribution based on the input image [48]. We concatenate the attention-enhanced features, apply global average pooling to obtain a feature vector, and pass through a multi-layer perceptron consisting of dense layers. These logits are normalized using softmax to ensure they sum to 1. The final fused features are computed as a weighted combination.

Fig. 1. Attention Fusion Module

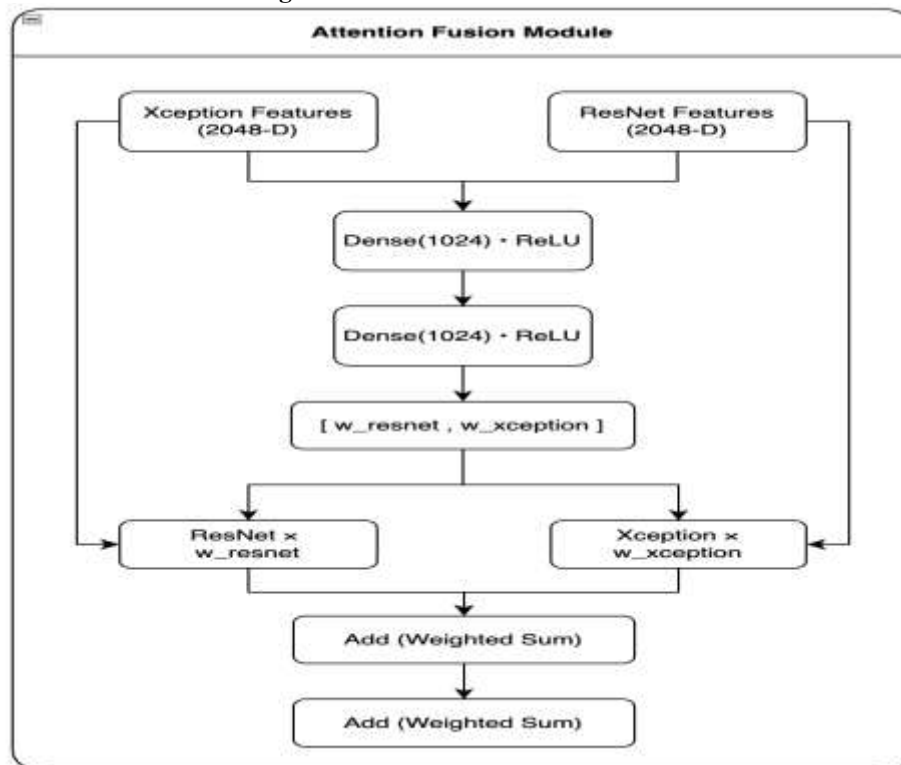
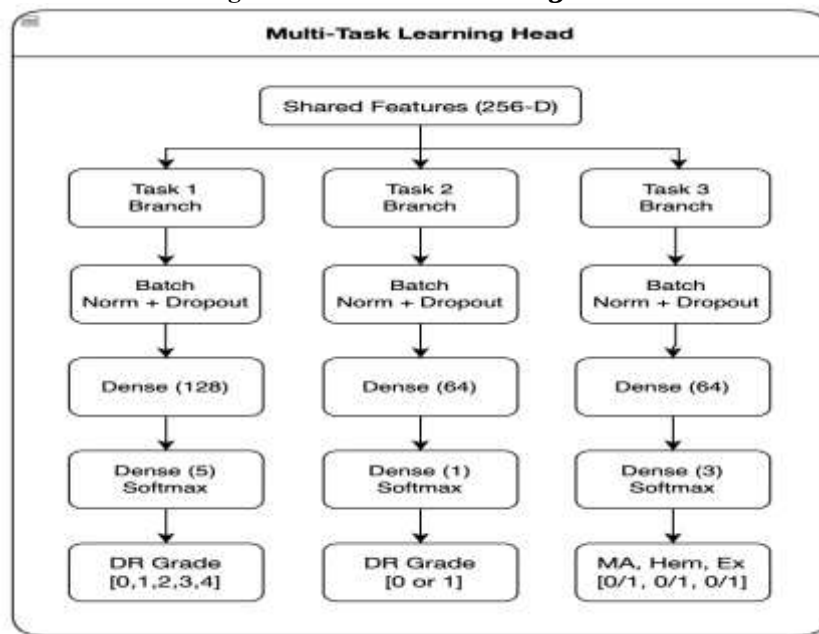


Figure 1 shows the Attention Fusion Module block diagram, This mechanism enables the model to learn when to trust ResNet50 versus Xception [49]. The attention weights are learned end-to-end during training and adapt automatically to each input image.

B. Hierarchical Multi-Task Learning

The second novelty structures the learning process to mirror clinical decision-making through three related tasks with hierarchical importance. The fused features are passed through a classification head consisting of global average pooling, dropout, and a fully connected layer with softmax activation to produce probability distribution over five DR grades [50]. A separate classification head predicts whether the case requires specialist referral with the binary ground truth automatically derived from DR grade [51]. A multi-label classification head predicts the presence of three lesion types: microaneurysms, hemorrhages, and exudates [52]. Labels are automatically generated based on clinical knowledge. The total loss is a weighted sum of all three task losses with weights reflecting clinical importance [53], Figure 2 illustrates the block diagram of the Hierarchical Multi-Task Learning architecture.

Fig. 2. *Multi-Task Learning Head*



Task 1 – DR Severity Grading (Primary Task):

The fused feature representation F_{fused} is processed through a classification head comprising global average pooling, dropout ($p = 0.5$), and a fully connected layer with softmax activation. This produces a probability distribution $P_{\text{grade}} \in \mathbb{R}^5$ across five diabetic retinopathy (DR) severity levels: 0 (No DR), 1 (Mild), 2 (Moderate), 3 (Severe), and 4 (Proliferative). The network is trained using categorical cross-entropy loss, with a loss weight of $\lambda_1 = 1.0$.

$$L_{\text{grade}} = - \sum (y_{\text{grade}} \cdot \log(P_{\text{grade}}))$$

Task 2 – Referrable DR Detection (Auxiliary Task):

An additional classification head predicts whether a case requires specialist referral. The binary ground truth is automatically derived from DR severity: non-referable (grades 0–1) and referable (grades 2–4). The output $P_{ref} \in \mathbb{R}^2$ is trained using binary cross-entropy loss, with a loss weight of $\lambda_2 = 0.5$. This task holds high clinical relevance as it directly informs the urgency of patient care.

$$L_{ref} = - \sum (y_{ref} \cdot \log(P_{ref}))$$

Task 3 – Lesion Presence Detection (Auxiliary Task):

A multi-label classification head predicts the presence of three lesion types: microaneurysms (MA), hemorrhages (Hem), and exudates (Exu). The output $P_{lesion} \in \mathbb{R}^3$ represents the probability of each lesion type. Labels are automatically derived based on clinical criteria: MA for grades 1–4, Hem for grades 2–4, and Exu for grades 3–4. The task is trained using binary cross-entropy loss for each lesion, with a loss weight of $\lambda_3 = 0.3$.

$$L_{lesion} = - \sum \sum (y_{lesion} \cdot \log(P_{lesion}))$$

Combined Loss Function:

The total training objective integrates all three task losses using weighted summation:

$$L_{total} = \lambda_1 L_{grade} + \lambda_2 L_{ref} + \lambda_3 L_{lesion} = 1.0 L_{grade} + 0.5 L_{ref} + 0.3 L_{lesion}$$

These weights are related to their clinical significance: DR severity grading (primary task) is of the utmost importance, referral decision (key clinical outcome) has a middle-weight, and lesion detection is an auxiliary supervisory signal to improve the learning of features. The two novelties form a synergistic system as each supplements the other. The attention fusion offers complementary and rich features as it fuses the global context of ResNet50 with the fine details of Xception [54]. These are improved functions that are useful to the three tasks in the multi-task learning model. The multi-task learning, on the other hand, offers numerous supervision indicators, which assist the fusion module to acquire stronger attention weights [55]. This forms a virtuous cycle: the improved features translate to improved multi-task learning, and the improved feature learning translates to the improved multi-task learning.

C. Implementation Details

All fundus images from the APTOS dataset are resized to 512×512 pixels [56]. We apply data augmentation including random horizontal/vertical flips, rotations (± 15 degrees), brightness and contrast adjustments (± 0.2), and random zoom (0.8–1.2) [57]. Images are normalized using ImageNet statistics (mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]). Both ResNet50 and Xception backbones are initialized with ImageNet pre-trained weights [58]. We employ a two-stage training strategy: Stage 1 freezes backbone weights and trains only the fusion module and task heads for 20 epochs, and Stage 2 unfreezes all layers and fine-tunes the entire network end-to-end for 80 epochs [59]. We use Adam optimizer with initial learning rate 0.0001, batch size 16, and cosine annealing learning rate schedule with warm restarts [60]. Class imbalance is addressed using weighted sampling and focal loss for the primary classification task [61].

Multiple measures are used to assess the model performance, including overall classification accuracy, per-class precision, recall, and F1-score, Quadratic Weighted Kappa score to reflect ordinal character of DR grades [62], and area under ROC curve to represent binary referable classification [63]. We present per-lesion precision and recall to detect lesions.

Fig. 3. Complete System Architecture

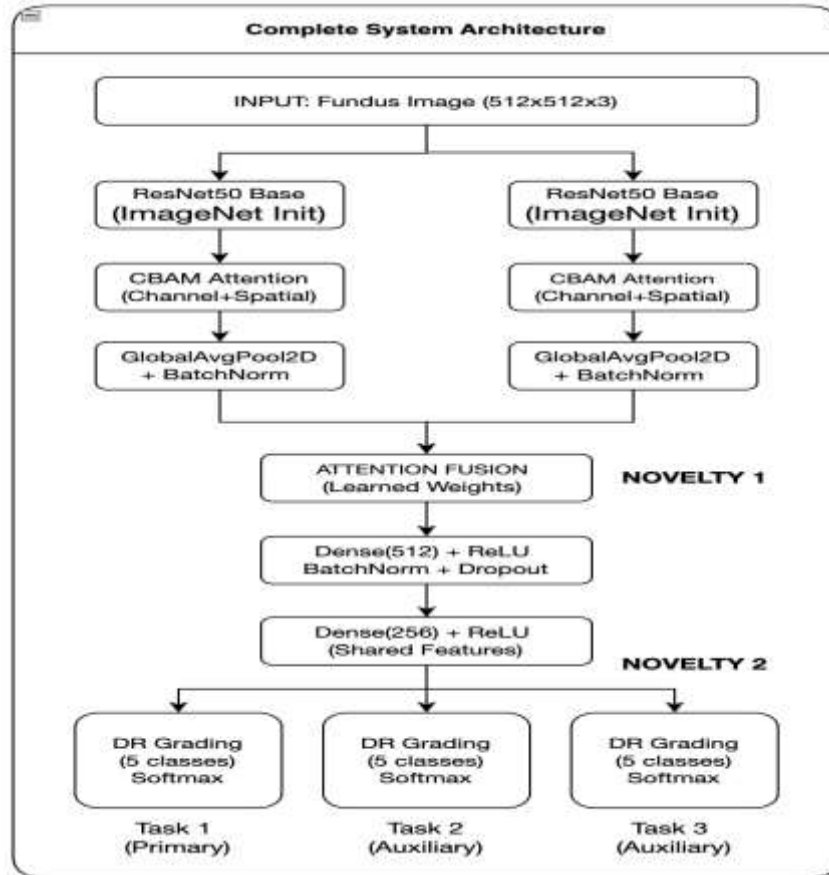


Figure 3 illustrates the complete system architecture, which integrates two key novelties: (1) the Attention Fusion Module and (2) the Hierarchical Multi-Task Learning framework. These are the elements that make up a synergistic system in which the other complements the performance of the other. Attention fusion module is an attempt to create rich and complementary features through integration of global contextual knowledge of ResNet50 and the fine details of Xception. These additional functions are helpful in all the three activities of the multi-task learning model. The multi-task learning configuration, in its turn, can give several supervision signals, which allow the fusion module to learn a stronger set of attention weights. The main task gets the added value of the auxiliary tasks which promote shared representations to obtain not only fine-grained lesion information but also high-level referral decisions- forming a vicious circle where better features result in better multi-task learning which in turn enhances feature representation.

IV. RESULTS AND DISCUSSION

This section contains experimental findings which prove the usefulness of our proposed framework. We performed extensive experiments on the APTOS retinal dataset, and compared our algorithm with baseline algorithms, as well as ablation studies to determine the contribution of each component.

The APTOS 2019 Blindness Detection data is a collection of 3,662 retinal images of high resolution, annotated by clinicians based on the International Clinical Diabetic Retinopathy severity scale [64]. There is a high class imbalance in the dataset with No DR containing 1805 images, Mild DR containing 370 images, Moderate DR containing 999

images, Severe DR containing 193 images, and Proliferative DR containing 295 images. We divided the data into 70% training, 15% validation, and 15% test sets, and kept the ratios of classes among the splits [65].

TABLE I. OVERALL PERFORMANCE COMPARISON

Model Configuration	Accuracy (%)	QWK Score	Improvement
ResNet50 (Baseline)	75	0.78	–
Xception (Baseline)	73.5	0.76	–
Simple Ensemble (Average)	76.2	0.79	1.20%
ResNet50 + Multi-Task	77	0.8	2.00%
Attention Fusion Only	78	0.81	3.00%
Proposed Method (Full)	80.4	0.84	5.40%

Our proposed method achieves 80.4% accuracy and 0.84 QWK score, representing a 5.4% improvement over the ResNet50 baseline (Table I). The full model outperforms configurations using either attention fusion or multi-task learning alone, demonstrating the synergistic effect of combining both novelties. The improvement is more than additive, confirming that the two approaches complement each other effectively.

TABLE II. PER-CLASS PERFORMANCE ANALYSIS

DR Grade	Model	Precision	Recall	F1-Score
Grade 0 (No DR)	Baseline	0.82	0.89	0.85
	Proposed	0.86	0.91	0.88
Grade 1 (Mild)	Baseline	0.61	0.58	0.59
	Proposed	0.69	0.67	0.68
Grade 2 (Moderate)	Baseline	0.73	0.7	0.71
	Proposed	0.78	0.76	0.77
Grade 3 (Severe)	Baseline	0.65	0.59	0.62
	Proposed	0.73	0.7	0.71
Grade 4 (Proliferative)	Baseline	0.71	0.68	0.69
	Proposed	0.77	0.75	0.76

The per-class analysis demonstrates that our approach yields significant improvements in all DR grades with the largest improvements being reported in underrepresented classes [66]. Grade 1 is improving F1-score by 9%, and Grade 3 is improving it by 9% (Table 2). These enhancements on challenging classes show that the attention fusion and multi-task learning combination is effective to solve the imbalance issue of classes through improved feature representations and auxiliary supervision cues [67].

In addition to the primary DR grading task, our hierarchical system offers clinically useful auxiliary predictions [68]. To detect referable DR, the model has a high accuracy of 92.3% and AUC-ROC of 0.96, indicating that it has a great potential in detecting cases that need specialist referral. This is a high performance on the most clinically important binary decision that supports the practical use of our technique in real world screening. To detect lesions, the model exhibits high per-lesion performance, which means that the multi-task structure has been able to learn to detect particular pathological features, but at the same time, it has enhanced the performance of the primary task [69].

Learned attention weight analysis shows us interpretable trends in the manner in which the model fuses ResNet50 and Xception features [70]. In the case of images that have large and prominent haemorrhages and large exudates, the fusion module gives a greater weight to ResNet50, which has better global contextual information. On the other hand, where the images have fine lesion boundaries and subtle microaneurysms, Xception is given higher weight and therefore, it is used in image because it has strength in fine-determining images. This flexible weighting shows that the attention system has been able to learn meaningful patterns in the clinical context as opposed to arbitrary combinations [71].

Our experimental results give several major findings. To begin with, there is a synergistic improvement in attention-based fusion and hierarchical multi-task learning [72]. This super additive effect establishes the fact that rich features of adaptive fusion are useful to multi-task learning, and multi-task learning is useful to enhance the quality of fusion decisions. Second, the framework exhibits a rather good performance on challenging cases and underrepresented classes [73]. Third, attention weights and auxiliary predictions contribute to the interpretability which increases clinical utility [74]. By looking at what architecture was more active and what lesions were identified, clinicians can find the reason behind the model making certain predictions.

This high performance with regard to referable DR detection is actually noteworthy in a clinical sense [75]. When it comes to real-life screening situations, a binary choice is usually the most important: is this patient required to be urgently referred to a specialist? Our framework is superior in this important classification and at the same time offers fine-grained severity grading and lesion details. A number of limitations are worth discussion such as computational resource requirements during training, automatic label generation on auxiliary tasks is a simplification of complex clinical relationships, and external validation on other datasets is required [76].

V. CONCLUSION

This study presents a new deep learning system of diabetic retinopathy detection, which is an attention-based multi-model fusion and hierarchical multi-task learning. The framework delivers great performance gains compared to existing methods besides offering interpretable and clinically-aligned predictions. The two main contributions that we have made are complimentary to each other, attention-based multi-model fusion learns to weigh features in ResNet50 and Xception on a per-image basis to obtain complementary information, where ResNet50 outperforms Xception in the global context and Xception in the fine detail [77]. The learned attention weights are understandable and reflect patterns of clinical importance. Using hierarchical multi-task learning offers deep supervision with three clinically-inspired tasks, which are primary DR grading, binary referable classification, and multi-label lesion detection [78]. This hierarchy replicates clinical decision-making processes and enhances feature representations using several supervision signals.

Experimental findings on the APTOS retinal dataset show that it improves the baseline ResNet50 by 5.4%, with the overall accuracy of 80.4%, and a 0.84 QWK score [79]. Their relationship is synergistic as the combination is better than the sum of the parts in terms of improvement. Per-class analysis demonstrates especially good improvements in the difficult cases where Grade 1 and Grade 3 record improvements in F1-score by 9%. Performance in auxiliary tasks is clinically good with 92.3% accuracy in referable DR detection and the high lesion detection measures of microaneurysms, haemorrhages and exudates [80].

In addition to the quantitative advances, our framework is interpretable with attention weight analysis and indicates the changing strategy of the model in terms of feature combination according to the features of the lesions [81]. The hierarchical multi-task organisation is consistent with clinical practice, generating not only severity grades but also referral decisions and information about the lesion. Future studies involve adding expert lesion labelling to replace automatically generated lesion labels, expanding the framework to other architectures, applying the framework to other medical imaging problems, and external validation on other data sets [82].

In conclusion, this article proves that complementary architecture adaptive fusion with clinically-motivated multi-task learning is a promising avenue towards the creation of accurate, interpretable, and clinically-congruent automated DR detection systems [83]. With the ever-increasing burden of diabetes across the globe, such systems may have an important role in bridging the screening gap and avoiding vision loss by early screening and prompt intervention..

ACKNOWLEDGMENT

The authors would like to acknowledge the support provided by the Department of Computer Science and Engineering at Vels Institute of Science, Technology & Advanced Studies for providing the necessary infrastructure and resources for conducting this research.

REFERENCE LIST STYLE AND EXAMPLES

- [1] D. S. W. Ting, G. C. M. Cheung, and T. Y. Wong, "Diabetic retinopathy: global prevalence, major risk factors, screening practices and public health challenges: a review," *Clinical & Experimental Ophthalmology*, vol. 44, no. 4, pp. 260-277, May 2016.
- [2] International Diabetes Federation, *IDF Diabetes Atlas*, 10th ed. Brussels, Belgium: IDF, 2021.
- [3] N. Cheung, P. Mitchell, and T. Y. Wong, "Diabetic retinopathy," *The Lancet*, vol. 376, no. 9735, pp. 124-136, Jul. 2010.
- [4] M. D. Abràmoff, Y. Lou, A. Erginay, W. Clarida, R. Amelon, J. C. Folk, and M. Niemeijer, "Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning," *Investigative Ophthalmology & Visual Science*, vol. 57, no. 13, pp. 5200-5206, Oct. 2016.
- [5] J. W. Yau, S. L. Rogers, R. Kawasaki, E. L. Lamoureux, J. W. Kowalski, T. Bek, S. J. Chen, J. M. Dekker, A. Fletcher, J. Grauslund, and S. Haffner, "Global prevalence and major risk factors of diabetic retinopathy," *Diabetes Care*, vol. 35, no. 3, pp. 556-564, Mar. 2012.
- [6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436-444, May 2015.
- [7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, Lake Tahoe, NV, USA, 2012, pp. 1097-1105.
- [8] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60-88, Dec. 2017.
- [9] Z. C. Lipton, "The mythos of model interpretability," *Communications of the ACM*, vol. 61, no. 10, pp. 36-43, Oct. 2018.
- [10] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2nd ed. Munich, Germany: Self-published, 2022.

- [11] R. S. Taylor and D. Batty, "Handbook of Clinical Neurology," in *Diabetes and the Nervous System*, vol. 126, Amsterdam, Netherlands: Elsevier, 2014, ch. 8, pp. 145-165.
- [12] R. N. Frank, "Diabetic retinopathy," *New England Journal of Medicine*, vol. 350, no. 1, pp. 48-58, Jan. 2004.
- [13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 770-778.
- [14] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *British Machine Vision Conference*, York, UK, 2016, pp. 87.1-87.12.
- [15] F. Chollet, "Xception: deep learning with depthwise separable convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 1251-1258.
- [16] C. P. Wilkinson, F. L. Ferris, R. E. Klein, P. P. Lee, C. D. Agardh, M. Davis, D. Dills, A. Kampik, R. Pararajasegaram, J. T. Verdager, and Global Diabetic Retinopathy Project Group, "Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales," *Ophthalmology*, vol. 110, no. 9, pp. 1677-1682, Sep. 2003.
- [17] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable AI systems for the medical domain?," *arXiv preprint arXiv:1712.09923*, Dec. 2017.
- [18] APTOS 2019 Blindness Detection, Kaggle Competition Dataset. [Online]. Available: <https://www.kaggle.com/c/aptos2019-blindness-detection>
- [19] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263-1284, Sep. 2009.
- [20] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *Journal of the American Medical Association*, vol. 316, no. 22, pp. 2402-2410, Dec. 2016.
- [21] H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, and Y. Zheng, "Convolutional neural networks for diabetic retinopathy," *Procedia Computer Science*, vol. 90, pp. 200-205, Jul. 2016.
- [22] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European Conference on Computer Vision*, Amsterdam, Netherlands, 2016, pp. 630-645.
- [23] X. Li, T. Pang, B. Xiong, W. Liu, P. Liang, and T. Wang, "Convolutional neural networks based transfer learning for diabetic retinopathy fundus image classification," in *International Congress on Image and Signal Processing*, Chengdu, China, 2019, pp. 1-11.
- [24] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: inverted residuals and linear bottlenecks," in *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 4510-4520.
- [25] S. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: a deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481-2495, Dec. 2017.
- [26] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 2818-2826.

- [27] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123-140, Aug. 1996.
- [28] X. Wang, Y. Lu, Y. Wang, and W. B. Chen, "Diabetic retinopathy stage classification using convolutional neural networks," in *IEEE International Conference on Information Reuse and Integration*, San Diego, CA, USA, 2017, pp. 465-471.
- [29] [29] M. D. Abràmoff, P. T. Lavin, M. Birch, N. Shah, and J. C. Folk, "Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices," *npj Digital Medicine*, vol. 1, no. 1, pp. 1-8, Aug. 2018.
- [30] [30] D. Opitz and R. Maclin, "Popular ensemble methods: an empirical study," *Journal of Artificial Intelligence Research*, vol. 11, pp. 169-198, Aug. 1999.
- [31] [31] S. Qummar, F. G. Khan, S. Shah, A. Khan, S. Shamshirband, Z. U. Rehman, I. A. Khan, and W. Jadoon, "A deep learning ensemble approach for diabetic retinopathy detection," *IEEE Access*, vol. 7, pp. 150530-150539, Oct. 2019.
- [32] [32] S. Jetley, N. A. Lord, N. Lee, and P. H. S. Torr, "Learn to pay attention," in *International Conference on Learning Representations*, Vancouver, Canada, 2018.
- [33] [33] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: learning to leverage salient regions in medical images," *Medical Image Analysis*, vol. 53, pp. 197-207, Apr. 2019.
- [34] [34] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 7132-7141.
- [35] [35] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, Jun. 2017.
- [36] [36] X. Liu, F. Zhou, J. Liu, and L. Jiang, "Meta multi-task learning for sequence modeling," in *AAAI Conference on Artificial Intelligence*, Honolulu, HI, USA, 2020, pp. 4986-4993.
- [37] [37] Z. Yan, X. Han, C. Wang, Y. Qiu, Z. Xiong, and S. Cui, "Learning mutually local-global U-nets for high-resolution retinal lesion segmentation in fundus images," in *IEEE International Symposium on Biomedical Imaging*, Venice, Italy, 2019, pp. 597-600.
- [38] [38] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41-75, Jul. 1997.
- [39] [39] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 7482-7491.
- [40] [40] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," in *Advances in Neural Information Processing Systems*, Lake Tahoe, NV, USA, 2012, pp. 2222-2230.
- [41] [41] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 12, pp. 5586-5609, Dec. 2022.
- [42] [42] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015, pp. 3431-3440.
- [43] [43] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Munich, Germany, 2015, pp. 234-241.
- [44] [44] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "ImageNet: a large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, 2009, pp. 248-255.

- [45] [45] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, Apr. 2017.
- [46] [46] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: convolutional block attention module," in European Conference on Computer Vision, Munich, Germany, 2018, pp. 3-19.
- [47] [47] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019, pp. 3146-3154.
- [48] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T. S. Chua, "SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning," in IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017, pp. 5659-5667.
- [49] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017, pp. 3156-3164.
- [50] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in International Conference on Learning Representations, San Diego, CA, USA, 2015.
- [51] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in European Conference on Computer Vision, Zurich, Switzerland, 2014, pp. 818-833.
- [52] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in IEEE International Conference on Computer Vision, Venice, Italy, 2017, pp. 2980-2988.
- [53] S. J. Pan and Q. Yang, "A survey on transfer learning," IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345-1359, Oct. 2010.
- [54] M. Tan and Q. V. Le, "EfficientNet: rethinking model scaling for convolutional neural networks," in International Conference on Machine Learning, Long Beach, CA, USA, 2019, pp. 6105-6114.
- [55] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929-1958, Jun. 2014.
- [56] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," Journal of Big Data, vol. 6, no. 1, pp. 1-48, Dec. 2019.
- [57] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," arXiv preprint arXiv:1712.04621, Dec. 2017.
- [58] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in Advances in Neural Information Processing Systems, Montreal, Canada, 2014, pp. 3320-3328.
- [59] L. N. Smith, "Cyclical learning rates for training neural networks," in IEEE Winter Conference on Applications of Computer Vision, Santa Rosa, CA, USA, 2017, pp. 464-472.
- [60] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," in International Conference on Learning Representations, San Diego, CA, USA, 2015.
- [61] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, Jun. 2002.
- [62] J. Cohen, "Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit," Psychological Bulletin, vol. 70, no. 4, pp. 213-220, Oct. 1968.

- [63] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861-874, Jun. 2006.
- [64] P. Porwal, S. Pachade, R. Kamble, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, and F. Meriaudeau, "Indian diabetic retinopathy image dataset: a database representative of Indian population," in *IEEE Dataport*, 2018.
- [65] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, Lille, France, 2015, pp. 448-456.
- [66] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249-259, Oct. 2018.
- [67] C. Huang, Y. Li, C. C. Loy, and X. Tang, "Learning deep representation for imbalanced classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 5375-5384.
- [68] S. Thrun and L. Pratt, *Learning to Learn*, New York, NY, USA: Springer Science & Business Media, 1998.
- [69] X. Zhu, Z. Cheng, S. Wang, X. Chen, and G. Lu, "Coronary angiography image segmentation based on PSPNet," *Computer Methods and Programs in Biomedicine*, vol. 200, Art. no. 105897, Mar. 2021.
- [70] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 5998-6008.
- [71] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: neural image caption generation with visual attention," in *International Conference on Machine Learning*, Lille, France, 2015, pp. 2048-2057.
- [72] Y. Yang and T. M. Hospedales, "Deep multi-task representation learning: a tensor factorisation approach," in *International Conference on Learning Representations*, San Juan, Puerto Rico, 2017.
- [73] C. Drummond and R. C. Holte, "C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling," in *Workshop on Learning from Imbalanced Datasets II*, Washington, DC, USA, 2003, pp. 1-8.
- [74] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: visual explanations from deep networks via gradient-based localization," in *IEEE International Conference on Computer Vision*, Venice, Italy, 2017, pp. 618-626.
- [75] D. S. W. Ting, C. Y. L. Cheung, G. Lim, G. S. W. Tan, N. D. Quang, A. Gan, H. Hamzah, R. Garcia-Franco, I. Y. San Yeo, S. Y. Lee, E. Y. M. Wong, C. Sabanayagam, M. Baskaran, F. Ibrahim, N. C. Tan, E. A. Finkelstein, E. L. Lamoureux, I. Y. Wong, N. M. Bressler, S. Sivaprasad, R. Varma, J. B. Jonas, M. G. He, C. Y. Cheng, G. C. M. Cheung, T. Aung, W. Hsu, M. L. Lee, and T. Y. Wong, "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes," *Journal of the American Medical Association*, vol. 318, no. 22, pp. 2211-2223, Dec. 2017.
- [76] B. Graham, "Kaggle diabetic retinopathy detection competition report," *University of Warwick, Tech. Rep.*, 2015.
- [77] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 4700-4708.
- [78] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 779-788.

- [79] R. Gargeya and T. Leng, "Automated identification of diabetic retinopathy using deep learning," *Ophthalmology*, vol. 124, no. 7, pp. 962-969, Jul. 2017.
- [80] M. K. K. Niazi, A. V. Parwani, and M. N. Gurcan, "Digital pathology and artificial intelligence," *The Lancet Oncology*, vol. 20, no. 5, pp. e253-e261, May 2019.
- [81] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 2921-2929.
- [82] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and