Hybrid Transformer Network for Diabetic Retinopathy Severity Grading and Lesion Segmentation

Conor King

Student ID: 201677856

MSc Bioinformatics
5th September 2023

Lay Summary

Diabetic retinopathy (DR) is an ocular disease and is the leading cause of adult blindness worldwide. It is a common complication of diabetes which is becoming more prevalent due to the steady increase in the number of people suffering from diabetes worldwide. As a result, ophthalmologists are struggling to keep up with the growing demand for DR screening, which involves the identification of retinal lesions in retinal images. Consequently, machine learning methods have been employed to automate this time-consuming process. Typically, the machine learning methods used for this purpose are designed to perform either lesion segmentation or severity grade classification. However, the current research proposes a novel model to carry out joint classification and segmentation (JCS) for DR diagnosis. JCS methods are able to provide lesion information as well as severity grade predictions, offering greater explanatory power for ophthalmologists to make more informed diagnoses. While other JCS methods do exist for this task, the novelty of our model is in the type of model that we are using. We used a hybrid model which incorporates elements from Convolutional Neural Networks (CNNs), the most commonly used models for image analysis, as well as the Vision Transformer (ViT). The ViT model is an adaptation of a Natural Language Processing model which is used in computer vision tasks and has not yet been applied in a JCS framework for DR diagnosis. We compared the results of our model to existing state-of-the-art methods, achieving competitive results which validate the performance of our model.

Abstract

DR is a common complication of diabetes and is the leading cause of blindness in adults worldwide. It is characterised by various retinal lesions which are difficult to detect, and with the global increase in diabetes incidences, ophthalmologists are struggling to meet the growing demand for DR screening. Recently, machine learning methods have been developed to automate this process. Majority of the existing methods are designed for either lesion segmentation or severity grade classification, with very few methods carrying out both tasks due to the increased computational demand. However, the ability of JCS methods to provide lesion information with a severity grade prediction makes them more clinically appropriate. In addition, most models use a CNN framework which struggle to accurately model the variation of lesions between patients. As a result, we have proposed a novel JCS model which utilises a hybrid Transformer backbone, inspired by the H2Former model. The hybrid Transformer encoder allows us to effectively capture local features and model long-range dependencies by incorporating the feature encoding strategies of both CNNs and Transformers. Using a fine-grained annotated dataset, our model is compared to multiple state-of-the-art segmentation models which we adapted to have a JCS framework. These experiments demonstrated that our model is able to match the lesion segmentation performance and exceed the classification performance of the established methods. Additionally, we established that the addition of a classification branch does not inhibit the segmentation performance of existing models, proving the efficacy of JCS methods.

1 Introduction

Diabetic retinopathy (DR) is a common complication of diabetes and is attributed with being the leading cause of blindness among adults worldwide (Li et al., 2019). The global prevalence of diabetes has been steadily increasing, reaching approximately 537 million individuals in 2021, and is projected to rise to 783 million by 2045 (International Diabetes Federation, 2021). With DR affecting roughly one in three people suffering from diabetes, ophthalmologists are struggling to meet the growing demand for DR screening (Li et al., 2019). DR is characterised by damage to the blood vessels in the retina as a result of high blood pressure and blood glucose levels, leading to the formation of retinal lesions (Gillow et al., 1999; Zhou et al., 2021). These lesions can present as small red dots or even yellow-white deposits, varying in shape and size (Zhou et al., 2021). Ophthalmologists use the presence and abundance of lesions to diagnose the severity of DR as some lesions are associated with mild DR, while the presence of other lesion types suggest more severe DR (Zhou et al., 2019). Due to the variety of these lesions, it is difficult and time-consuming for ophthalmologists to accurately detect them and determine the severity of DR. As a result, the development of automated DR screening methods which utilise machine learning techniques have been a prominent subject of recent research on medical imaging.

Recently, the use of machine learning methods in medical image analysis tasks has become more widespread. This is largely due to the increase in availability of big data, the advances in processing speed brought about by the introduction of graphics processing units (GPUs), and the development of machine learning algorithms (Shen et al., 2017). These improvements have allowed machine learning methods to detect subtle patterns and anomalies in medical images, making them capable of outperforming medical professionals in certain tasks (Rajpurkar et al., 2017). Additionally, they have reduced many barriers to access professional analysis and diagnosis (Gulshan et al., 2016). While there are many

different types of machine learning models, deep learning methods such as CNNs have become the most dominant in medical image analysis. This is because of their abilities to handle more complex data types, and to capture more complex relationships in data (LeCun et al., 2015).

Conventionally, CNNs have been at the forefront of medical image analysis. They are renowned for their ability to effectively capture local features and discover localised patterns, allowing them to excel in both classification and segmentation tasks (Apivanichkul *et al.*, 2023; Castiglioni et al., 2021; LeCun et al., 2015). Most notably, the U-Net architecture (Ronneberger et al., 2015) has been widely adapted for image segmentation tasks, showcasing immense versatility, and achieving state-of-the-art performance across multiple domains. However, CNNs are limited in their ability to capture long-range dependencies within image datasets (Zheng et al., 2021). Due to the variable nature of retinal lesions, long-range dependencies are necessary to capture global information for effective semantic segmentation (He et al., 2023). To address this limitation, Transformer-based methods, known for their success in natural language processing, have recently gained popularity (Zheng et al., 2021). ViT, the first Transformer-based model used for computer vision tasks, introduced the self-attention mechanism which better captures long-range information (Dosovitskiy et al., 2020). Although initially designed for image classification, ViT has since achieved state-of-the-art results for segmentation tasks (Zheng et al., 2021). Nevertheless, the Transformer is not without pitfalls of its own as it requires images to be embedded into a series of 1-dimensional (1D) tokens for input (Dosovitskiy et al., 2020). Thus, the spatial information of the images is not retained and the Transformer's ability to learn local features is inhibited (He et al., 2023). Moreover, features are learned through a token-wise attention mechanism, leading to ineffective modelling of multi-scale channel-wise feature information (He et al., 2023). Given the complementary nature of the drawbacks associated with CNNs and ViT, a

novel hybrid model known as H2Former was introduced by He et al. (2023). The principal innovation of the H2Former model is the encoder, which utilises a hybrid Transformer block (Figure 2), that aims to combine the strengths of CNNs, Transformers, and Multi-Scale Channel Attention (MSCA). Since its release, the H2Former has only been used for image segmentation, and is yet to be applied to classification tasks. Consequently, the use of the H2Former in JCS models is also yet to be explored.

Although not widely utilised, JCS models have recently gained traction in medical image analysis due to their comprehensive outputs (Girard et al., 2019; Mehta et al., 2018; Zhou et al., 2019). JCS models are neural networks which carry out both classification and segmentation tasks in a single framework. In the context of DR, JCS models produce an output which provides the predicted DR grades along with the identified retinal lesions that are present. This provides a more transparent and comprehensive output compared to most CNNs, which provide predictions with little-to-no explanation, often resembling a black box (Wu et al., 2021). JCS models are of much greater use to ophthalmologists, who are able to make more informed diagnoses when using computer vision methods to streamline their practice.

The current research on computer-aided methods in DR diagnosis aims to make several key contributions. Firstly, we propose a novel JCS model which employs an H2Former backbone, making use of the hybrid Transformer encoder from He et al. (2023). This contribution stems from the scarcity of hybrid Transformer methods in DR research. Additionally, this study aims to extend the capabilities of the H2Former model, which has currently only been used for segmentation tasks, by applying it in a JCS capacity. The proposed model will be compared to established state-of-the-art methods which will serve as baselines. Furthermore, the baseline models will be adapted from their initial pure

segmentation versions into JCS models, allowing for further investigation into the efficacy of JCS frameworks.

2 Methods

In this section, the overall architecture of the proposed model and its various components, will be described. Thereafter, the data that is used in this research will be described, followed by details on training of the proposed model and the comparisons that were made with established methods.

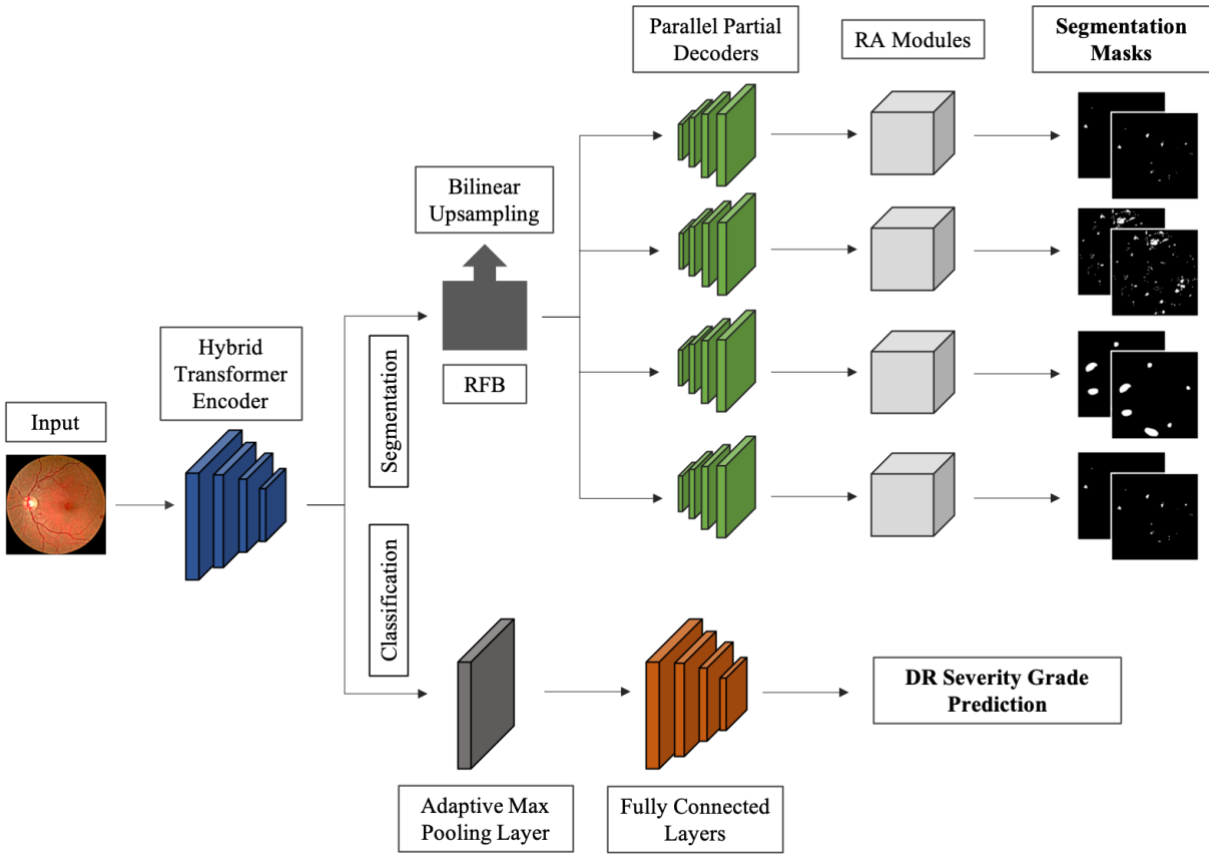

Figure 1. *Structure of the proposed hybrid Transformer joint classification and segmentation model.*

2.1 Model Architecture

The proposed model, visualised in Figure 1, incorporates an H2Former backbone, utilising the four-stage encoder from the H2Former model (He et al., 2023). Thereafter, the model splits into two branches, one for segmentation and the other for classification. The segmentation branch, which employs the parallel partial decoders and reverse attention

modules from the PraNet model (Fan et al., 2020), is responsible for locating and identifying retinal lesions. While the classification branch, which is made up of an adaptive max pooling layer and a fully connected layer, provides predictions of the DR grade for each image.

Hybrid Transformer Encoder

First, the input images are fed through a convolutional stem, comprised of a convolutional layer and a Max Pooling layer, which divides the images into patches. The patches are then embedded into image tokens and passed through the encoder in a sequential manner. Each stage of the encoder consists of a hybrid Transformer block (He et al., 2023) which integrates three components: a convolutional block, MSCA block, and a Transformer block. Additionally, a patch merging layer exists between each stage of the encoder, which is implemented as a 3x3 convolutional layer with a stride of 2, to reduce the number of tokens and expand their dimensions between each stage (He et al., 2023). As can be seen in the structure of the hybrid Transformer block, depicted in Figure 2 below, the embedded image tokens are first passed through the convolutional block and the MSCA block. The convolutional block extracts local features from each of the tokens while introducing a strong inductive bias. Meanwhile, the MSCA captures multi-scale relationships across different channels from the image tokens. These features are then integrated and used as input for the Transformer block which encodes long-range dependencies between the tokens. The combination of these complementary strategies allows the hybrid Transformer block to encode local and global information, enhancing the model's feature representations. Furthermore, the spatial resolution of the generated feature maps is reduced in each stage of the encoder by 1/4, 1/8, 1/16, and 1/32 of the original input image size, respectively.

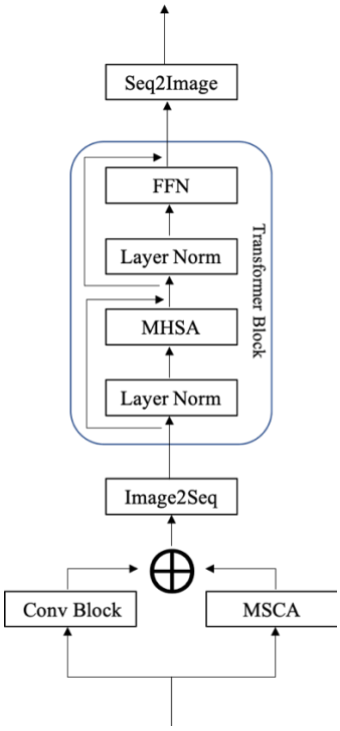

Figure 2. *Hybrid Transformer block from the H2Former encoder, which is used in our model. This image is a recreation of the image in the H2Former publication (He et al., 2023). The Transformer block is comprised of a Layer Normalisation, multi-head self-attention (MHSA) module, and a Feed Forward Network (FFN).*

Lesion Segmentation

As the encoder performs downsampling operations for feature extraction, the resulting feature map is first upsampled using bilinear interpolation with a scale factor of four. Bilinear interpolation takes the weighted average of the four nearest pixels to each pixel in the feature map, estimating the new pixel values to produce a feature map with greater spatial resolution (Parsania & Kumar, 2016). Receptive Field Block (RFB) modules are then employed to improve segmentation accuracy by enhancing the deep features that were extracted by the encoder (Liu et al., 2018). The RFB modules create four branches, each corresponding to one of the four retinal lesions: microaneurysms (MA), haemorrhages (HE), hard exudates (EX), and soft exudates (SE). Each of the enhanced feature maps are then passed on to the parallel partial decoders (PPD) and Reverse Attention (RA) modules from the PraNet model (Fan et al., 2020). The PPDs aggregate high-level features to create a global map which provides

rough spatial location information which is used by the RA modules. Only high-level features are used as low-level features make little contribution towards segmentation performance and are computational expensive due to their larger spatial resolution (Wu et al., 2019). The global maps are then processed by the RA modules, which serve a dual purpose. The RA modules refine the identified regions of interest in the global map by detecting precise lesion boundaries. Additionally, they ignore the foreground objects identified in the global map to focus the attention on other areas in which lesions may have been missed. This results in improved segmentation performance and the output of accurate segmentation masks which are resized to match the original input image size of 128x128.

Classification

The classification branch, which is used to provide predictions of the DR grade for each image, is comprised of an adaptive max pooling layer followed by a series of fully connected layers and dropout layers. The adaptive max pooling layer produces a 1D representation of the features that have been extracted by the model's encoder. This is done to standardise the input size for the fully connected layer, as well as to reduce the computational complexity of the classification task. The 1D feature representations are then fed into a series of fully connected layers interspersed with dropout layers. The dropout layers are included as a regularisation technique used to improve the generalisation of the model and to reduce overfitting (Baldi & Sadowski, 2013). A dropout probability of 0.2 is used, meaning that 20% of the output from the first two fully connected layers are randomly deactivated and excluded from backpropagation on every iteration of the training process. This allows the network to learn a more robust and generalised representation of the data (Baldi & Sadowski, 2013). Meanwhile, the fully connected layers incrementally reduce the number of channels present in the input from 512 to five, corresponding to the five DR grades. The output from the final fully connected layer serves as the DR grade prediction.

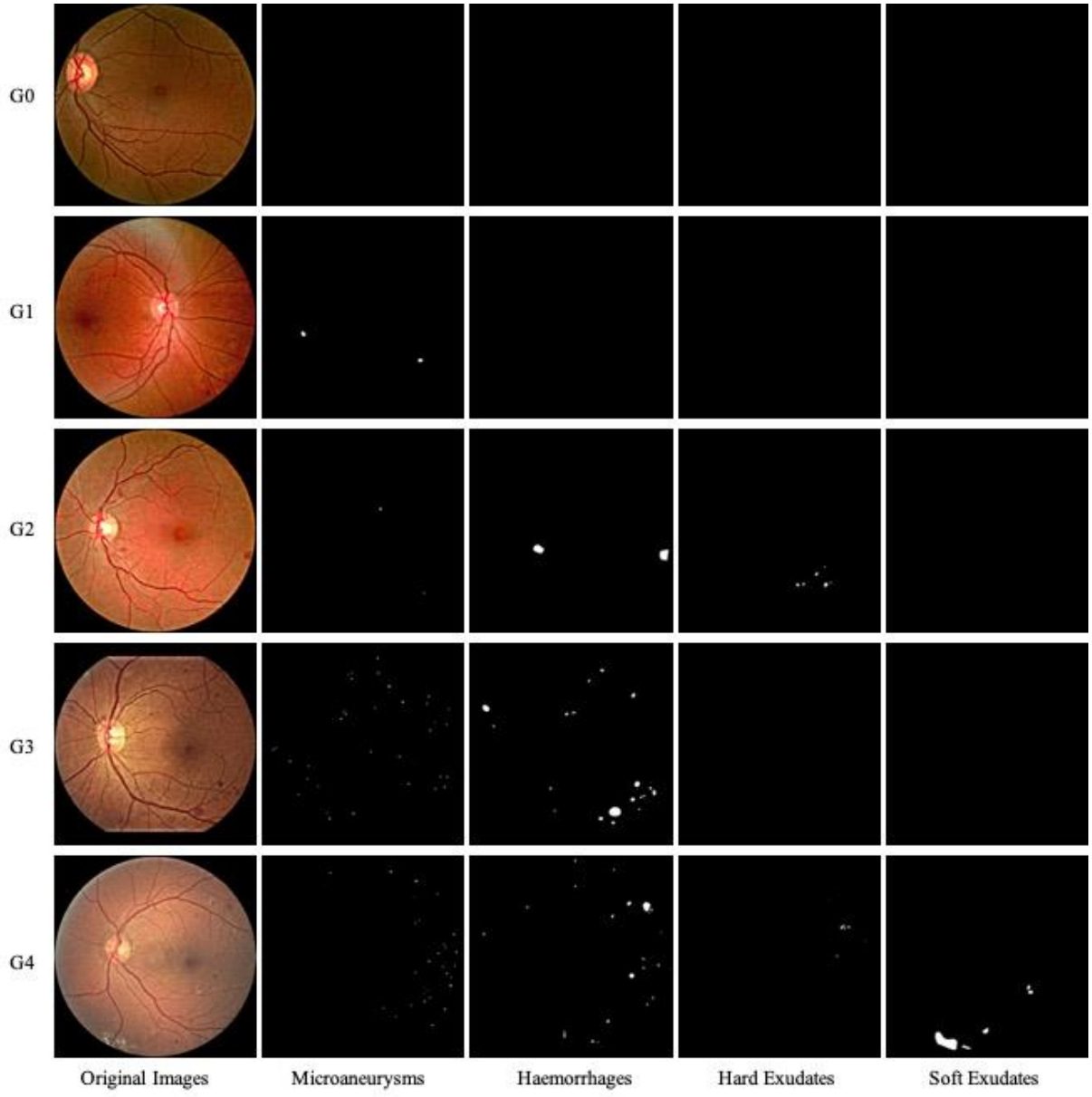

Figure 3. *Qualitative results of lesion segmentation for MA, HE, EX, and SE. Each row corresponds to the five severity grades, as labelled on the left, showing the presence or absence of each of the four lesion types for the severity grades.*

2.2 Datasets

There are several publicly available fundus image datasets for DR grading. However, there are significantly fewer that contain pixel-level annotations for DR lesion segmentation. As a result, we used the FGADR Seg-set (Zhou et al., 2021) which contains 1,842 images with both image-level and pixel-level annotations. The pixel-wise annotated lesions include MA, HE, EX, and SE, which are visualised in Figure 3. Meanwhile, intra-retinal microvascular

abnormalities (IRMA) and neovascularisations (NV) were excluded due to missing ground truth masks. The image-level annotations adhere to the internationally recognised DR severity grades (Wilkinson et al., 2003) which range from no retinopathy to proliferative DR on a scale of 0 – 4. A data split of 60-20-20 was used to divide the Seg-set into training, validation, and testing sets respectively, and all images were resized to 128x128 during pre-processing.

2.3 Implementation and Experiments

The overall parameters of the proposed model were optimised using stochastic gradient descent (SGD) with a momentum of 0.9. This model, and all baseline models, were trained over 10,000 iterations using an NVIDIA GeForce GTX 1650 12GiB GPU. The hyperparameters for each model were individually optimised with the proposed model achieving its best performance with a batch size of 16, base learning rate of 1e-2, and a learning rate decay of 0.99 every 200 epochs. The following overall loss function was applied to the training of all models:

$$Loss = Seg_{loss} + (0.1 \times Grade_{loss}), \tag{1}$$

where Seg_{loss} is the summation of the loss functions for the segmentation of each lesion type, while $Grade_{loss}$ is the loss function for the classification task. The segmentation loss functions for each lesion are defined as the sum of Binary Cross Entropy (BCE) loss and Dice loss, both of which are defined in the equations below, as per Jadon (2020), while the grading loss function is solely comprised of BCE loss:

$$BCE\ Loss = -[target * \log(output) + (1 - target) * \log(1 - output)], \tag{2}$$

$$Dice\ Loss = 1 - Dice, \tag{3}$$

in which *target* represents the binary pixel value of the ground truth masks and *output* is the predicted probability of the binary pixel value. *Dice* represents the Dice coefficient which is subsequently defined as an evaluation metric.

The performance of the models was measured using a series of both segmentation and classification metrics. The Dice coefficient was used to measure the segmentation performance and is calculated by measuring the overlap between the predicted segmentation masks and the provided ground truths using the following formula, obtained from Zhou et al. (2021):

$$Dice = \frac{2|S \cap G| + 1}{|S| + |G| + 1}, \tag{4}$$

where S and G represent the predicted segmentation masks and the ground truth masks, respectively. Meanwhile, to evaluate the classification performance we used precision, specificity, sensitivity, Area Under the Curve of the Receiver Operating Characteristic (AUC-ROC), and F1-score, the formula for which is provided below (Zhou et al., 2021):

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}, \tag{5}$$

In order to accurately compare the performance of the proposed model, several established methods were adapted to serve as baseline models. The baseline models include U-Net (Ronneberger et al., 2015), U-Net++ (Zhou et al., 2018), Attention U-Net (Oktay et al., 2018), and Dense U-Net (Li et al., 2018), all of which were originally segmentation models created for other segmentation tasks. This means that the architecture of each of these models had to be slightly modified in order to fit the data and requirements of this study. These models were selected as baselines because Zhou et al. (2021) demonstrated that U-Net frameworks performed significantly better than non-U-Net frameworks on the FGADR dataset. Once the models were adapted to carry out the desired lesion segmentation task, their performance was evaluated before adding a classification branch to each of them. Thereafter, they were re-evaluated as JCS models, allowing us to assess the impact that adding a classification branch has on segmentation performance, while also testing the proposed hybrid Transformer model.

3 Results and Discussion

In this section, the results of the aforementioned models and experiments are presented. As was previously mentioned, the performance of the baseline models was assessed before and after the classification branch for DR grading was introduced. We conducted 4,000 iterations to test all models on 20% of the total dataset, corresponding to 368 images, with hyperparameters individually optimised for each model.

3.1 Segmentation Results

The segmentation results from the baseline models before adding the classification branch are presented in Table 1. From these results, we can see relatively consistent segmentation within each lesion type across all baseline methods. However, there is significant variation between the lesion types, with Dice scores having a range of up to 37%, indicating a possible systemic issue in the data. Additionally, all methods produced considerably lower Dice scores than those reported by Zhou et al. (2021). This is most likely because Zhou et al. (2021) trained each model for single segmentation of each lesion, whereas in the current study, we had each model perform multi-lesion segmentation simultaneously. Despite the reduced performance, accurate comparisons can still be carried out between the models as the differences between the lesion types are consistent across all methods.

Table 1. Segmentation results of baseline models before classification branch for DR grading is added. The best results for each lesion type are highlighted in bold.

Methods	*SE*		*EX*		*MA*		*HE*	
	Dice	MAE	Dice	MAE	Dice	MAE	Dice	MAE
U-Net	**0.5992**	**0.0018**	0.3963	0.0074	**0.3598**	**0.0042**	0.2286	**0.0105**
Attention U-Net	0.3820	0.0043	0.4414	0.0070	0.3475	0.0044	**0.2371**	0.0106
Dense U-Net	0.3001	0.0056	**0.4960**	**0.0061**	0.3398	0.0044	0.2126	0.0112
U-Net++	0.3012	0.0050	0.4772	0.0068	0.3566	0.0044	0.2019	0.0109

The U-Net model, despite having the simplest architecture, drastically outperformed the other baseline methods in SE segmentation with a Dice score of 59.92%. While it also achieved the best Dice score for MA, the U-Net model was considerably worse than the other baseline methods in EX segmentation. Attention U-Net was the most consistent method, only

marginally surpassing the other methods for HE, and outperforming at least one other method on the remaining three lesion types. Similar can be said for U-Net++ which was close to obtaining the best overall results for EX and MA segmentation. On the other hand, it scored the lowest for HE segmentation and only surpassed Dense U-Net in SE segmentation by 0.1%. Dense U-Net, the most computationally demanding of the baseline models, attained the highest Dice score for EX segmentation. However, it significantly underperformed in SE segmentation compared to U-Net and Attention U-Net.

Table 2. Segmentation results of hybrid Transformer model and baseline models after the classification branch was added. The best results for each lesion type are highlighted in bold.

Methods	*SE*		*EX*		*MA*		*HE*	
	Dice	MAE	Dice	MAE	Dice	MAE	Dice	MAE
U-Net	**0.6590**	**0.0017**	0.3794	0.0076	**0.3598**	**0.0042**	**0.2520**	**0.0104**
Attention U-Net	0.4082	0.0041	0.4256	0.0069	0.3532	0.0045	0.2171	0.0106
Dense U-Net	0.4000	0.0027	0.4110	0.0071	**0.3598**	**0.0042**	0.2212	0.0106
U-Net++	0.2547	0.0062	**0.4740**	**0.0063**	0.3293	0.0046	0.1877	0.0110
H2Former (Ours)	0.6292	0.0021	0.3768	0.0078	0.3584	0.0046	0.2493	0.0105

Table 2 shows the segmentation performance of the baseline models after the classification branch for DR grading was added. Initially, we encountered overfitting issues with the U-Net model after adapting it for JCS. However, after increasing the dropout probability of the dropout layers in the classification branch to 50%, we were able to extract satisfactory performance. Overall, the addition of the classification branch to the baseline models had a limited effect on their segmentation performance for each lesion. With that being said there were consistent decreases in EX segmentation across all baseline models. Meanwhile, there were relatively consistent increases in SE segmentation, with the exception of U-Net++ which decreased by 4.65%. In most cases, the performance only differed by a few percent at most, despite a few notable exceptions. The performance of U-Net and Dense U-Net for SE segmentation improved by approximately 6.00% and 10.00% respectively. Meanwhile, the EX segmentation performance of Dense U-Net decreased by 8.49%. With this result and the change in U-Net++ SE segmentation being the only notable decreases in performance, it is

evident that the JCS adaptations are able to closely match the performance of their pure segmentation counterparts. As a result, we have shown that JCS frameworks are viable methods for clinical applications, and that their only remaining limitation is the increased computational demand.

The segmentation performance of our proposed hybrid Transformer model is also presented in Table 2. Despite it not achieving the highest performance for any of the lesion types, it was able to match or exceed at least one of the baseline models for SE, MA, and HE segmentation. Although our method recorded the lowest performance for EX segmentation, U-Net was only able to outperform it by 0.26%. Nevertheless, this performance is commendable as the baseline models are all adaptations of established pure segmentation models. Therefore, we can confidently state that our proposed model was able to achieve state-of-the-art performance in the DR lesion segmentation task.

3.2 DR Grading Results

We evaluated the classification performance in DR severity grading for each of the models, the results of which are shown in Table 3. For this task, our hybrid Transformer model obtained dominant performance across nearly all metrics, outperforming all of the baseline methods with the exception of U-Net++ achieving a higher AUC-ROC score. The performance achieved by our model is most likely a result of the superior feature extraction attained by combining the feature encoding strategies of CNNs, MSCA, and ViT (He et al., 2023). The simplicity of the architecture of the classification branch makes the classification performance more reliant on the ability of the encoder compared to the segmentation branch which incorporates RFB and RA modules to refine the extracted features.

Despite all metrics being lower than initially expected the ability of our method to outperform such renowned methods shows great promise. The U-Net model performed particularly poorly in comparison to the other methods. This is likely due to the previously

mentioned increase in the dropout probability reducing the number of features that are learnt by the U-Net model, compounded with the simplicity of the model's architecture. Conversely, the Attention U-Net and U-Net++ models were almost able to match the performance of our method across all of the metrics. The Attention U-Net model employs a self-attention gating module, not dissimilar to the self-attention mechanism of ViT, allowing it to focus on relevant features in images and capture contextual information (Oktay et al., 2018). Furthermore, attention gates are commonly used in natural language processing (Oktay et al., 2018), which is where Transformer architectures were established. These similarities are likely to explain the Attention U-Net's comparative performance to our hybrid Transformer model. As for the U-Net++ model, specially designed skip connections are used to selectively combine feature maps from different layers to help the model focus on foreground objects (Zhou et al., 2018). While this is distinctly different to the self-attention mechanism of the hybrid Transformer model, it has a similar effect in allowing the model to focus on informative regions. Furthermore, the skip connections and deep supervision of U-Net++ allow it to extract features at multiple scales (Zhou et al., 2018), resembling the MSCA block from the hybrid Transformer encoder. Once again, the commonalities between U-Net++ and the H2Former model in feature extraction are likely to have caused similar performance in the classification metrics.

Table 3. Classification performance results for DR severity grade predictions. The best results are highlighted in bold.

Methods	*F1*	*Precision*	*Specificity*	*Sensitivity*	*AUC-ROC*
U-Net	0.4504 (0.394 - 0.507)	0.3945 (0.339 - 0.456)	0.6356 (0.585 – 0.678)	0.5571 (0.505 – 0.609)	0.7628 (0.734 – 0.790)
Attention U-Net	0.6199 (0.569 – 0.674)	0.6226 (0.571 – 0.683)	0.8342 (0.805 – 0.863)	0.6413 (0.595 – 0.690)	0.8510 (0.823 – 0.879)
Dense U-Net	0.5490 (0.492 – 0.605)	0.5801 (0.502 – 0.663)	0.7959 (0.764 – 0.824)	0.5897 (0.538 – 0.641)	0.8385 (0.812 – 0.864)
U-Net++	0.6192 (0.566 – 0.675)	0.6154 (0.561 – 0.675)	0.8344 (0.806 – 0.861)	0.6413 (0.592 – 0.690)	**0.8578** (0.829 – 0.885)
H2Former (Ours)	**0.6320** (0.579 – 0.683)	**0.6288** (0.576 – 0.685)	**0.8446** (0.816 – 0.870)	**0.6495** (0.601 – 0.698)	0.8564 (0.830 – 0.882)

As was previously mentioned, the results achieved in both the segmentation and classification tasks were lower than initially expected. As this is consistent across all of the models, this is thought to be a consequence of the data that was used in the study. While the FGADR dataset contains fine-grained annotations and performed well in single lesion segmentation (Zhou et al., 2021), the current research shows that it has struggled to meet the requirements for effective multi-lesion segmentation. Additionally, the dataset does not contain the same number of images for each lesion type or for each severity grade, resulting in class imbalances which can negatively impact the results. This highlights one of the fundamental problems with the use of machine learning in medical image analysis, which is the lack of large datasets containing accurate annotations. Medical image datasets require expert annotation by medical professionals, which are expensive and time consuming to obtain (Tajbakhsh et al., 2016; Zhou et al., 2021). Additionally, the size of datasets can often be limited by the scarcity of the disease being investigated (Tajbakhsh et al., 2016). Consequently, the use of larger and more complex datasets in the current research may produce significantly improved results.

Nevertheless, the setup of the current study allows us to oversee these limitations. The use of multiple baseline models which have previously produced commendable results, allows us to fairly assess the potential of our model. Likewise, our decision to include multiple metrics to measure the classification performance helps in further mitigating the effects of the class imbalance and the small size of the dataset that was used. Using various metrics, each of which measure performance in a unique way, ensures us that any notable differences that are seen in the results are less likely to have been caused by skewed or inconsistent data.

4 Conclusions

In this study we proposed a novel hybrid Transformer model with a JCS framework for DR lesion segmentation and severity grading. We compared our model to multiple state-of-the-art methods which were all adapted for this particular task. Our model was able to match the segmentation performance of the baseline models, while surpassing them in classification performance. As a result, we have successfully demonstrated that the proposed model can be effectively used for automated DR diagnosis. Furthermore, our results show that JCS models are able to meet the performance benchmark of pure segmentation models in the context of DR lesion segmentation.

References

Apivanichkul, K., Phasukkit, P., Dankulchai, P., Sittiwong, W., & Jitwatcharakomol, T. 2023. Enhanced deep-learning-based automatic left-femur segmentation scheme with attribute augmentation. *Sensors*, **23**(12), 5720. https://doi.org/10.3390/s23125720

Baldi, P., & Sadowski, P. 2013. Understanding dropout. *Advances in Neural Information Processing Systems*, **26**, 2814–2822. https://proceedings.neurips.cc/paper_files/paper/2013/file/71f6278d140af599e06ad9bf1ba03cb0-Paper.pdf

Castiglioni, I., Rundo, L., Codari, M., Di Leo, G., Salvatore, C., Interlenghi, M., Gallivanone, F., Cozzi, A., D'Amico, N. C., & Sardanelli, F. 2021. AI applications to medical images: From machine learning to deep learning. *Physica Medica*, **83**, 9–24. https://doi.org/10.1016/j.ejmp.2021.02.006

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv (Cornell University)*. https://doi.org/10.48550/arXiv.2010.11929

Fan, D., Ji, G., Zhou, T., Chen, G., Fu, H., Shen, J., & Shao, L. 2020. PraNet: Parallel reverse attention network for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Vol. **12266**). Springer, Cham. https://doi.org/10.1007/978-3-030-59725-2_26

Gillow, J. T., Gibson, J. M., & Dodson, P. M. 1999. Hypertension and diabetic retinopathy---what's the story? *British Journal of Ophthalmology*, **83**(9), 1083–1087. https://doi.org/10.1136/bjo.83.9.1083

Girard, F., Kavalec, C., & Cheriet, F. 2019. Joint segmentation and classification of retinal arteries/veins from fundus images. *Artificial Intelligence in Medicine*, **94**, 96–109. https://doi.org/10.1016/j.artmed.2019.02.004

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., & Webster, D. A. 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, **316**(22), 2402. https://doi.org/10.1001/jama.2016.17216

He, A., Wang, K., Li, T., Du, C., Xia, S., & Fu, H. 2023. H2Former: An efficient hierarchical hybrid transformer for medical image segmentation. *IEEE Transactions on Medical Imaging*. https://doi.org/10.1109/tmi.2023.3264513

International Diabetes Federation. 2021. *IDF Diabetes Atlas* (10th ed.). https://www.diabetesatlas.org

Jadon, S. 2020. A survey of loss functions for semantic segmentation. In *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE, pp. 1–7. https://doi.org/10.1109/cibcb48159.2020.9277638

LeCun, Y., Bengio, Y., & Hinton, G. E. 2015. Deep learning. *Nature*, **521**(7553), 436–444. https://doi.org/10.1038/nature14539

Li, T., Gao, Y., Wang, K., Guo, S., Liu, H., & Kang, H. 2019. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Information Sciences*, **501**, 511–522. https://doi.org/10.1016/j.ins.2019.06.011

Li, X., Chen, H., Qi, X., Dou, Q., Fu, C., & Heng, P. 2018. H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE Transactions on Medical Imaging*, **37**(12), 2663–2674. https://doi.org/10.1109/tmi.2018.2845918

Liu, S., Huang, D., & Wang, Y. 2018. Receptive field block net for accurate and fast object detection. In *Computer Vision - ECCV 2018*. Springer, Cham, pp. 04–419. https://doi.org/10.1007/978-3-030-01252-6_24

Mehta, S., Mercan, E., Bartlett, J., Weaver, D. L., Elmore, J. G., & Shapiro, L. G. 2018. Y-Net: Joint Segmentation and Classification for diagnosis of Breast Biopsy Images. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. Lecture Notes in Computer Science* (Vol. **11071**). Springer, Cham, pp. 893–901. https://doi.org/10.1007/978-3-030-00934-2_99

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M. C. H., Heinrich, M. P., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., Glocker, B., & Rueckert, D. 2018. Attention U-Net: Learning where to look for the pancreas. *arXiv (Cornell University)*. https://arxiv.org/abs/1804.03999

Parsania, P. S., & Kumar, A. V. S. 2016. A comparative analysis of image interpolation algorithms. *International Journal of Advanced Research in Computer and Communication Engineering*, **5**(1), 29–34. https://doi.org/10.17148/ijarcce.2016.5107

Ronneberger, O., Fischer, P., & Brox, T. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (Vol. **9351**). Springer, Cham, pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D. Y., Bagul, A., Langlotz, C. P., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. 2017. CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv (Cornell University)*. https://arxiv.org/pdf/1711.05225.pdf

Shen, D., Wu, G., & Suk, H. 2017. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, **19**(1), 221–248. https://doi.org/10.1146/annurev-bioeng-071516-044442

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C., Gotway, M. B., & Liang, J. 2016. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*, **35**(5), 1299–1312. https://doi.org/10.1109/tmi.2016.2535302

Wilkinson, C. P., Ferris, F. L., Klein, M. L., Lee, P. P., Agardh, C., Davis, M. D., Dills, D., Kampik, A., Pararajasegaram, R., & Verdaguer, T. J. 2003. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*, **110**(9), 1677–1682. https://doi.org/10.1016/s0161-6420(03)00475-5

Wu, Y., Gao, S., Mei, J., Xu, J., Fan, D., Zhang, R., & Cheng, M. 2021. JCS: An explainable COVID-19 diagnosis system by joint classification and segmentation. *IEEE Transactions on Image Processing*, **30**, 3113–3126. https://doi.org/10.1109/tip.2021.3058783

Wu, Z., Su, L., & Huang, Q. 2019. Cascaded partial decoder for fast and accurate salient object detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3902–3911. https://doi.org/10.1109/cvpr.2019.00403

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P. H. S., & Zhang, L. 2021. Rethinking semantic segmentation from a Sequence-to-Sequence perspective with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/cvpr46437.2021.00681

Zhou, Y., He, X., Huang, L., Liu, L., Zhu, F., Cui, S., & Shao, L. 2019. Collaborative learning of semi-supervised segmentation and classification for medical images. In

2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2074–2083. https://doi.org/10.1109/cvpr.2019.00218

Zhou, Y., Wang, B., Huang, L., Cui, S., & Shao, L. 2021. A benchmark for studying diabetic retinopathy: Segmentation, grading, and transferability. *IEEE Transactions on Medical Imaging*, **40**(3), 818–828. https://doi.org/10.1109/tmi.2020.3037771

Zhou, Z., Siddiquee, M. R., Tajbakhsh, N., & Liang, J. 2018. UNet++: A nested U-Net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. DLMIA ML-CDS 2018*. Springer, Cham, pp. 3–11. https://doi.org/10.1007/978-3-030-00889-5_1