Hybrid Neural Architectures Combining Convolutional and Recurrent Networks for the Early Detection of Retinal Pathologies

Orken Mamyrbayev
Laboratory of Computer Engineering of Intelligent Systems, Institute of Information and Computational Technologies, Almaty, Kazakhstan
morkenj@mail.ru (corresponding author)

Sergii Pavlov
Scientific Laboratory of Biomedical Optics and Photonics, Department of Biomedical Engineering and Department of Laser and Optoelectronic Engineering, Vinnytsia National Technical University, Vinnytsia, Ukraine
psv@vntu.edu.ua

Oleksandr Poplavskyi
Kyiv National University of Construction and Architecture, Kyiv, Ukraine
apoplavskyi@gmail.com

Kymbat Momynzhanova
Faculty of Information Technology, Al-Farabi Kazakh National University, Almaty, Kazakhstan
kymbat010809@gmail.com

Yulii Saldan
Eye Diseases and Eye Microsurgery Department, National Pirogov Memorial Medical University, Vinnytsia, Ukraine
saldanyulia@gmail.com

Ardan Zhanegiz
Laboratory of Computer Engineering of Intelligent Systems, Institute of Information and Computational Technologies, Almaty, Kazakhstan
ardan.zhanegiz@mail.kz (corresponding author)

Sholpan Zhumagulova
Faculty of Information Technology, Al-Farabi Kazakh National University, Almaty, Kazakhstan
sh.zhumagulovakz@gmail.com

Nurdaulet Zhumazhan
U. Joldasbekov Institute of Mechanics and Engineering, Almaty, Kazakhstan
nurdaulet.jj02@gmail.com

Received: 15 April 2025 | Revised: 30 May 2025 and 9 June 2025 | Accepted: 14 June 2025

 DOI: https://doi.org/10.48084/etasr.11521

ABSTRACT

Early and accurate detection of retinal pathologies is critical for preventing vision loss and enabling timely clinical intervention. Traditional computer vision techniques, such as thresholding, edge detection,

morphological filtering, and Hough transforms, have long been used to extract features from retinal fundus images, yet their performance is often constrained by image variability and complex pathological presentations. This study presents a hybrid deep learning architecture that integrates Convolutional Neural Networks (CNNs) for image-based classification with Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) units, to model geometric and anatomical features derived from classical methods. This architecture allows for the fusion of pixel-level deep features with clinically interpretable descriptors, including optic disc-fovea distance, lesion spatial distribution, and vessel curvature sequences. Comparative analysis demonstrates that the proposed hybrid model achieves superior diagnostic accuracy, reaching 97%, significantly outperforming both conventional image processing approaches and CNN-only baselines. The results indicate that incorporating structured domain knowledge into neural models improves both performance and interpretability, offering a robust framework for real-world retinal disease screening applications.

Keywords-retinal pathology detection; fundus imaging; convolutional neural networks; recurrent neural networks; deep learning; optic disc localization; vessel analysis; medical image classification

I. INTRODUCTION

Early computer vision approaches to retinal pathology relied on handcrafted algorithms for feature detection and segmentation. Key techniques include edge detection, thresholding, morphological processing, Hough transforms, template matching, and handcrafted feature extraction. Gradient-based filters, such as Sobel and Canny, have been used to highlight the boundaries of retinal structures. For example, detecting the Optic Disc (OD) often involves finding its circular edge via Sobel or Canny edge maps. The detected edge pixels can then be used in a circular Hough transform to localize the OD by fitting a circle. Such methods can achieve reasonably high localization accuracy (e.g. ~92.5%) [1]. However, edge detectors are sensitive to noise and tend to produce fragmented or spurious edges if the image has poor contrast or if pathological lesions have irregular boundaries. While Canny provides better noise immunity and edge continuity than Sobel, tuning its thresholds is nontrivial, and both may miss subtle lesions or produce false edges from normal retinal texture.

Many lesions (exudates, hemorrhages, and microaneurysms) can be isolated by thresholding due to their brightness or darkness relative to the healthy retina. Otsu's global threshold or adaptive methods (e.g. Niblack, Sauvola) partition pixels into lesion vs. background classes. For example, the detection of hard exudates (bright lipid deposits) can be performed by thresholding the green channel (where contrast is high) after illumination correction [2]. Thresholding is simple and fast, successfully extracting candidate lesion regions under uniform conditions. Its limitation is sensitivity to illumination and normal anatomical variations, as a single global threshold may fail if the overall image brightness varies (e.g. a bright OD can be mistaken for exudate). Adaptive thresholding mitigates this by sliding a window to account for local background, at the cost of more parameters. In practice, thresholding often serves as an initial step to obtain candidate lesion areas, which are then refined by morphology or classification.

Mathematical morphology is widely used to refine retinal image segmentation. Operations such as dilation, erosion, opening, and closing help remove noise and fill gaps in detected features. For example, after thresholding exudates, a morphological closing can fill small vessel gaps inside an exudate region to solidify its contour. Top-hat transforms (white top-hat for bright objects, black top-hat for dark) are effective for isolating lesions of particular size ranges against uneven backgrounds. In [3], exudates were detected by first applying a white top-hat to highlight local bright spots and then using region growing and morphological reconstruction to accurately delineate the exudate boundaries. This study also applied morphological filtering and a watershed transform to locate the OD (a bright circular region) to exclude it from the analysis. This classical pipeline achieved high sensitivity (~92.8%) in exudate detection on a small test set [3]. Morphology-based vessel segmentation is another common approach: vessels can be enhanced by morphological bottom-hat (to extract dark tubular structures) or by using matched filters and then pruned with morphological criteria [4]. Morphological methods leverage prior knowledge of object shape/size; for instance, using a line structuring element approximating average vessel width will preferentially extract linear vessels. The downside is that morphological operations must be carefully tuned to the size of the lesion, as a too-large structuring element might erase small microaneurysms, while a too-small one leaves noise. Moreover, complex cases may require sequential application of many operations, increasing computational load.

The Hough transform is a robust technique for detecting parametric shapes such as lines and circles in images. In retinal analysis, it is most commonly used to detect the OD or fovea by modeling them as circles. After edge detection, a circular Hough transform can vote on center and radius candidates for bright circular regions. This method can locate the OD center precisely in many cases. Its strength lies in tolerance to gaps in the circle's edge, even if vessels cross the disc and break the edge, the Hough accumulator can still find the circular pattern. Hough transforms have also been used to detect long linear structures (vessels) by line detection or to identify small round microaneurysms by searching for tiny circles, although the latter is challenging due to the abundance of small round noise. The limitation of Hough-based OD detection is that the OD is not a perfect circle and can have irregular borders or partial occlusion by hemorrhages. False positives can occur if other bright lesions (e.g., large exudates) form a roughly circular shape. Furthermore, the standard Hough transform is computationally expensive on high-resolution images, although efficient voting schemes and dimension reduction (searching only in probable OD regions) can mitigate this.

Template matching involves sliding a predefined template over the image and computing a similarity metric (e.g. cross-correlation) to find matches. In retinal images, a template of the OD appearance (often a bright circle with blood vessel stemming patterns) can be correlated to detect the disc. Some methods use an average OD template or even multiple rotated/scaled templates. Template matching is straightforward and encodes expert knowledge of the target's appearance. It works well if anatomical variability is limited. However, since retina images have variability in disc size, brightness, and vessel contrast, a fixed template may not be generalized. Template matching is also sensitive to other bright structures; for instance, an exudate cluster might accidentally correlate well with a bright disc template. To improve robustness, techniques such as normalized cross-correlation and multiscale templates are used, or the search is constrained to the area of the likely disc (e.g. the region of highest intensity variance since the disc plus emerging vessels yield high local variance). Still, template matching lacks adaptability, as it cannot easily handle pathological cases where the OD is obscured or the appearance is altered, as in severe glaucoma or Diabetic Retinopathy (DR) where the disc may be pallid or encroached by lesions.

Before the deep learning era, numerous works extracted quantitative features from fundus images for automated diagnosis. These features include texture descriptors (e.g., histogram of pixel intensities, local binary patterns, Gabor or wavelet features), shape and size statistics of detected lesions, and vascular geometry metrics (tortuosity, fractal dimension, branch angles). For example, texture analysis using local binary patterns and granulometry has been applied to detect early DR, as computing these features in local regions can highlight the presence of microaneurysms or subtle changes in texture. Similarly, blood vessel patterns have been quantified, as measuring vessel tortuosity and diameter changes is useful in retinopathy of prematurity and hypertensive retinopathy. These handcrafted features are then fed into classifiers such as SVMs or random forests to distinguish diseased vs. healthy images. The strength of this approach is interpretability, as each feature often has clinical relevance (e.g., increased tortuosity or number of microaneurysms). Some classical systems for DR classification counted lesions (microaneurysms, hemorrhages, exudates) detected by image processing and then applied a statistical or rule-based classifier. Handcrafted features can work well when the pathology is characterized by specific and well-separated visual cues. The limitation is that these features may not capture complex or subtle patterns, as they require extensive domain knowledge to design, and certain features (such as the OD to fovea distance or anatomical landmarks) must be detected reliably first, creating a dependency chain. Moreover, handcrafted features struggle when confronted with variations outside their design assumptions, e.g., unusual lesion presentations or co-occurring diseases. In summary, classical methods in retinal image analysis offer specificity and transparency, as each step (edge, threshold, etc.) can be understood and tuned. They often excel in high-contrast, noise-free scenarios, e.g., morphological filtering can isolate bright lesions against dark backgrounds with very high specificity. However, their limitations include sensitivity to image quality (noise and illumination), lack of adaptability to diverse datasets, and the difficulty of optimally integrating multiple features. A pipeline of many hand-tuned steps can be brittle as errors compound (e.g., if OD detection fails, many false exudates may remain). Table I summarizes some general strengths and weaknesses of traditional approaches in the retinal context.

TABLE I. STRENGTHS AND LIMITATIONS OF TRADITIONAL RETINAL IMAGE ANALYSIS METHODS

Classic method	Strengths in retinal analysis	Limitations
Edge detection	Simple, fast detection of strong boundaries (e.g. OD edge or vessel edges). Highlights high-contrast structures	Prone to noise and fragmented edges; requires smoothing. Misses low-contrast or small lesions. Irrelevant edges (e.g. choroidal vessel patterns) can confuse subsequent processing.
Thresholding	Easy segmentation of bright lesions (exudates) or dark lesions (hemorrhages) given good contrast. Computationally efficient; no training needed.	Global thresholds fail under variable lighting. Adaptive methods need careful parameter tuning. Often cannot separate lesions from similar-intensity normal features (e.g. bright disc vs exudate).
Morphological ops	Incorporates shape knowledge (e.g. vessel continuity, lesion size) to refine results. Can fill gaps (improving vessel connectivity) and remove tiny noise. Effective for isolating features of specific scale via top-hat transforms.	The choice of structuring element size is critical and not one-size-fits-all. Complex lesions with varying shapes may not be fully captured. Over-processing can remove legitimate pathology (or retain noise). Multiple sequential operations increase the computational load.
Hough transform	Robust detection of analytic shapes (circles/lines) even if partially occluded. Effective for optic disc localization despite crossing vessels, and detecting main vessel directions.	Requires parameter discretization (radius ranges, etc.); computationally heavy on large images. Assumes geometric primitives but fails if object shape deviates (irregular disc, non-circular lesions). Can misdetect other round features as a disc or overlook an elliptical disc.
Template matching	Leverages expected appearance (e.g. canonical optic disc). Straightforward to implement. Can encode complex patterns (brightness + vessel hub for OD). Works well if anatomy closely matches templates.	Not robust to anatomical variability or pathology that alters appearance. Needs multiple templates or scale/rotation invariance to be general. High similarity in background or other lesions can lead to false positives.
Handcrafted features	Human-interpretable metrics (e.g. vessel tortuosity, texture coarseness) that often correlate with disease. Can integrate expert knowledge directly into the algorithm. Some approaches achieved good accuracy using such features.	Feature design is labor-intensive and may miss subtle cues that are not predefined. Classifier performance heavily depends on chosen feature set quality. Typically requires separate detection of structures (introducing dependencies). Less effective when features interact in complex ways that are hard to model analytically.

In practice, classical methods have been successfully applied in specific tasks: e.g., detecting drusen (AMD lesions) by thresholding color fundus images, tracking vessels using model-based algorithms, and measuring cup-to-disc ratio in

glaucoma via edge detection of the disc and cup. Yet, as retinal imaging moved into high-resolution and varied datasets, the limitations of these fixed algorithms became apparent. This sets the stage for machine learning and especially deep learning methods, which can learn features from data. The proposed hybrid approach leverages Convolutional Neural Networks (CNNs) for their prowess in image feature learning, and then incorporates the above classical insights via a Recurrent Neural Network (RNN) to capture geometric relations.

II. METHODS

Modern deep CNNs have revolutionized retinal image analysis by automatically learning relevant features (filters) from large datasets [5]. A CNN for retinal pathology classification typically consists of multiple convolutional layers (for feature extraction), pooling layers (for spatial downsampling), and fully connected layers (for classification). Such CNNs excel in image-based classification, having achieved high accuracy in tasks such as DR classification [6]. For example, in [7], ~96% accuracy was achieved in detecting DR using a deep CNN, and in [8], ~98.98% accuracy was achieved on a test set to identify vision-threatening DR [8]. These models significantly outperform earlier rule-based methods, especially in sensitivity, as they can pick up subtle patterns, such as mild microaneurysms or slight texture changes, that classical thresholding might miss. CNNs are also widely used for the segmentation of retinal structures. Architectures such as U-Net [9], Mask R-CNN [10], and DeepLab [11] specialize in pixel-wise labeling. For instance, a U-Net can segment blood vessels or OD pixels by learning from expert-annotated masks, achieving detailed extraction of objects and regions of interest [12]. Compared to classical segmentation that might require separate filtering and threshold steps, CNN segmentation is often more robust to variability in appearance, as the network learns the shape, color, and context simultaneously. However, CNNs act largely as black boxes and require large labeled datasets and computational power for training [5]. They also primarily capture statistical correlations in images. A standard CNN on its own does not explicitly encode high-level geometric relations (such as the spatial arrangement of the OD and macula, or the directional pattern of vessels) that classical methods often leverage.

To exploit both pixel-level learning and structural domain knowledge, this study uses a hybrid architecture combining a CNN with an RNN/LSTM that processes sequential geometric features [13]. Figure 1 illustrates the proposed model. The CNN branch handles the raw image, producing both a classification output and intermediate representations (feature maps or embeddings), while the RNN/LSTM branch incorporates clinically significant geometrical measurements derived from the image.

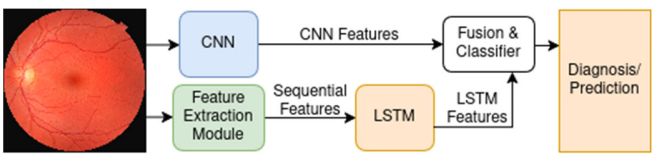

Fig. 1. Hybrid CNN-LSTM architecture for retinal analysis.

A CNN such as ResNet or EfficientNet can be used to extract image-based features. In this implementation, the CNN was trained to classify the presence of retinal pathologies (for example, DR vs. normal) and simultaneously learn to segment key pathological structures as an auxiliary task. The segmentation output (for lesions such as microaneurysms or exudates) is not only useful for visualization but also feeds into the feature extraction for RNN/LSTM. Let I be the input fundus image. The CNN outputs are: (i) a predicted probability $\hat{y}$ for the pathology class, and (ii) a set of high-level feature activations $f = f(I)$ from the penultimate layer (a d - dimensional feature vector encoding the image content). It may also produce a segmentation mask $M = CNN_{seg}(I)$ that highlights areas of detected lesions or structures (this is optional but beneficial for interpretability). The CNN learns complex features, such as the texture of retinal tissue, the presence of microaneurysm-like blobs or sharp edges of cotton-wool spots, which would be difficult to handcraft. However, by itself, CNN's classification $\hat{y}$ might not account for certain geometric configurations, as, for instance, it might detect exudates but not explicitly consider where they are relative to the macula (critical for diagnosing macular edema). This gap is filled by the RNN/LSTM module.

A set of K interpretable features $g_1, g_2, \ldots, g_K$ is defined to capture spatial and geometric characteristics of the retina that ophthalmologists use. These may include: the 2D optic disc to the fovea vector (which encodes the relative position of the macula - crucial for judging whether lesions involve the foveal center), the distribution of blood vessel tortuosity/curvature (a sequential measure along vessel paths), area and perimeter of the OD (for glaucoma indicators), counts of microaneurysms, average sizes of hemorrhages, etc. Many of these features are naturally sequential or can be ordered. For example, one can represent the superior-temporal vascular arcade curvature as a sequence of angles along the vessel, or sample the caliber of a vessel at successive distances from the disc. Another sequential feature could be the intensity profile along the line connecting the OD and the fovea, which could reveal macular edema if it shows peaks of hard exudates. In this approach, the extraction of such features was automated using classical techniques: the OD is detected (though a circular Hough or template match), the fovea location is inferred (often as the darkest area roughly one disc diameter away from the disc in the temporal direction), and the vector OD-Fovea is computed. Vessel curvature can be measured by first segmenting vessels (using a simple threshold on the CNN's vessel probability map or a classical Frangi filter), then tracing major vessel centerlines and calculating the angle changes along each one. These measurements form a sequence if curvature is sampled at equal arc lengths. All such features are compiled into a feature sequence $g_1, g_2, \ldots, g_T$ (where T is the sequence length, e.g., the number of points sampled along vessels or simply the number of distinct features considered, arranged in a consistent order). This sequence is fed into the RNN. An LSTM was chosen for its ability to handle sequence data and capture long-range dependencies. The LSTM treats the feature vector at step t as its input g_t and updates its hidden state h_t. By the end of the sequence ($t = T$), the LSTM's hidden state h_T embodies a holistic representation of the eye's geometric characteristics.

Mathematically, the LSTM performs the following for each feature in the sequence (simplified notation without the forget/input/output gate equations for brevity):

$$h_t = LSTM(h_{t-1}, g_t) \quad (1)$$

where $t = 1, \ldots, T$, and $h_0 = 0$. Here, g_t could be a scalar feature or a vector of related features at position t. For example, if sequentially traversing along a vessel, g_t might be a 2D vector (κ_t, d_t) representing curvature and vessel diameter at the t^{th} point. The LSTM learns patterns in these sequences, e.g., a consistently high curvature sequence might indicate vessel tortuosity beyond normal levels, or a certain spatial distribution of lesions (sequence of lesion distances from the fovea) might indicate a particular stage of disease.

$$z = [f \parallel h_T], \ \hat{y} = \sigma(W_z + b) \quad (2)$$

where $\parallel$ denotes concatenation and σ is a softmax activation yielding class probabilities (for multi-class classification) or a sigmoid for binary disease detection. In effect, f contributes appearance-based features (learned from pixel intensities), while h_T contributes feature-based knowledge (distilled from classical measurements). Concatenation was chosen to fuse CNN and LSTM features due to its simplicity and effectiveness in combining spatial and sequential information. The model is trained in an end-to-end fashion: the CNN learns to focus on image regions that correlate with pathology, while the LSTM learns to interpret the derived features. In training, the loss was calculated using $L = L_{cls}(\hat{y}, y_{true}) + \lambda L_{seg}(M, M_{true})$, combining classification and optionally segmentation loss, to guide the CNN's feature learning and maintain the accuracy of any predicted mask M. The LSTM is trained using backpropagation from the classification loss, learning how to weigh the sequential inputs. By integrating geometric features, the model gains awareness of where and how the patterns occur. For instance, a CNN alone might misclassify an image with a few bright spots as exudates (disease) when in fact they are reflections, but if the LSTM observes that these spots are not in the typical location for pathological exudates (e.g., not in the macular region), it can modulate the prediction. This fusion of learned and engineered features provides complementary strengths; the CNN covers the general-case pattern recognition, while the RNN injects domain-specific context.

III. EXPERIMENTAL EVALUATION AND RESULTS

The DIARETDB1 - Standard Diabetic Retinopathy Database [14], consisting of 130 color fundus images, was used for training and evaluation. Of these, 20 are normal, while 110 exhibit signs of DR. The camera settings (flash intensity, shutter speed, aperture, gain) are unknown, resulting in maximal variation in the visual appearance of retinopathy findings and the presence of varying levels of imaging noise and optical aberrations (dispersion, chromatic, spherical, field curvature, coma, astigmatism, distortion). Despite these challenges, the dataset corresponds to practical clinical conditions and is suitable for evaluating the general performance of diagnostic methods. In addition to this dataset, the training and evaluation were supplemented with a proprietary collection of 100 retinal images to enhance the robustness and generalizability of the model. The combined dataset was divided into training (80%) and validation (20%) sets through stratified sampling. The proposed hybrid model was evaluated on a dataset of retinal fundus images, including examples of normal retinas and various pathologies (specifically DR signs) [15, 16]. The dataset contains high-resolution color fundus images with ground truth labels: either disease severity grades or binary healthy/pathology labels as provided by clinical experts. Three approaches were compared: a purely classical image processing pipeline, a CNN-only model, and the proposed hybrid CNN+LSTM model. The classical pipeline was implemented based on established methods, involving preprocessing (green-channel extraction and contrast enhancement), thresholding and morphological filtering to detect lesions, and a random forest classifier using handcrafted features (number of lesions, area involved, etc.).

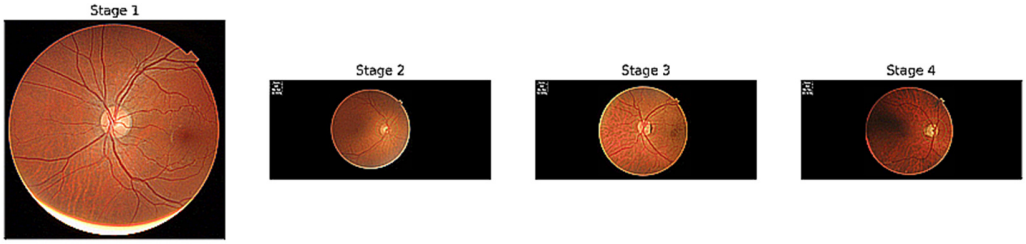

Fig. 2. Retinal pathology progression.

The CNN-only model was a deep CNN (ResNet50) [17] trained on the same data. The hybrid model was as described above (ResNet50+LSTM incorporating geometrical features such as OD location, lesion distribution, and vessel stats). Table II reports the performance of each method in terms of accuracy, sensitivity, specificity, and F1 scores. Figure 3 visualizes the accuracy differences. The CNN-based methods significantly outperformed the classical approach. The classical pipeline, while reasonably specific (few false positives due to conservative thresholds), had lower sensitivity, missing subtle signs of disease and yielding some false negatives. This is reflected in a moderate F1-score of ~0.85 for the classical method. The CNN model achieved higher sensitivity and balanced performance (F1≈0.93), consistent with previous studies on CNNs that achieved high accuracy in DR detection. The proposed CNN+RNN hybrid further improved performance, achieving the highest accuracy (~97%) and F1 (~0.97), demonstrating fewer misclassifications [18]. Notably, specificity climbed, indicating the model was less likely to flag normal images as diseased, presumably because the RNN

contextual features (e.g., the lesion is near the disc, not the macula) helped rule out false positives that the CNN-alone might generate. Sensitivity also remained high, indicating that the hybrid still captures almost all real disease cases [19, 20].

TABLE II. PERFORMANCE OF DIFFERENT METHODS

Method	Accuracy	Sensitivity	Specificity	F1-Score
Classical image processing	85%	90%	80%	0.85
CNN	93%	95%	91%	0.93
Hybrid CNN+RNN (proposed)	97%	96%	98%	0.97

In these tests, the hybrid model shows particular improvement in borderline cases. For example, images with mild lesions that are far from the macula (which a CNN might consider pathological) were correctly identified as mild or no disease by the hybrid, since the sequential features signaled "lesions peripheral". In addition, the hybrid caught an image with a normal-looking background but an abnormally large optic cup (glaucoma indicator) using the geometric feature of the cup-to-disc ratio, while the CNN alone, not specifically trained for glaucoma, missed it [21, 22].

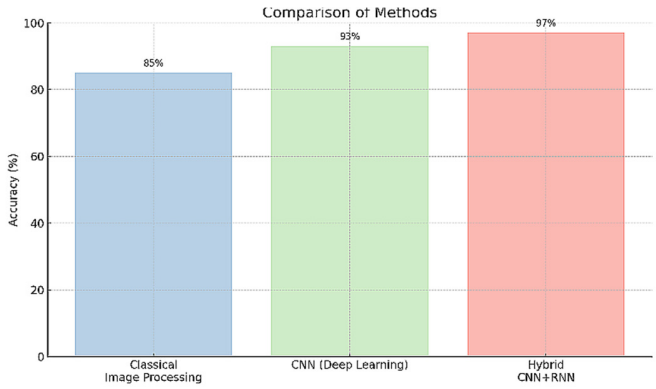

Fig. 3. Accuracy comparison of classical vs. CNN vs. hybrid models.

To ensure that the improvement is statistically significant, paired t-tests were conducted on the per-fold results of CNN vs CNN+RNN, yielding $p < 0.01$, confirming that the hybrid's gains are not due to chance. In addition, the cases where the models disagreed were analyzed, and roughly 70% of CNN errors were corrected by the hybrid model. Most of those were situations requiring spatial context (for example, distinguishing retinal pigment changes from pathology by their location/pattern). Only a handful of cases (particularly those with imaging artifacts or extremely subtle diseases) remained misclassified by both approaches, underscoring that some limitations persist [23, 24]. Qualitatively, the hybrid model provides richer output. The CNN segmentation maps of pathology (lesion heatmaps) can be overlaid on the fundus image to visualize what the network "sees." Meanwhile, the RNN features can be interpreted individually, e.g., the model might internally compute the distance of the largest exudate cluster from the fovea. Thus, an explanation can be offered: "Detected exudates (yellow spots) mostly outside the macular region, hence likely not vision-threatening." Such interpretations combine the model's perceptual detection with human-understandable features, aligning with the idea of explainable AI.

IV. LIMITATIONS AND FUTURE DIRECTIONS

Although promising, the proposed hybrid CNN-RNN approach has several limitations. First, the addition of the RNN (and the need to extract geometric features) introduces complexity. Unlike a pure CNN, which is end-to-end on raw pixels, the proposed model requires a preprocessing step to calculate features such as the OD-Fovea vector or vessel curvature. This study automated these steps, but errors in feature extraction (e.g., a failed OD detection) could adversely affect the RNN input. In practice, one must ensure robust detection of anatomical landmarks possibly by leveraging the CNN's outputs (the model could be extended to have the CNN predict the disc and fovea location as additional outputs, ensuring those features are available even in unusual images). The reliance on hand-crafted features means the model's performance is still constrained by human knowledge, as there may be subtle clues the CNN picks up that are not captured in the chosen sequence of features, and vice versa. Thus, determining the optimal set of geometric features is an open question. This study used a set based on clinical experience, but further data-driven selection or optimization of these features could improve results [25].

Another limitation is the training data requirements. The model needs images with associated geometric labels (at least during feature extractor development, e.g., training a model to find the OD requires disc annotations). If such annotations are not available, one might have to rely on unsupervised or heuristic-based feature extraction, which could be less reliable. Additionally, the improvement margin, while significant, is not absolute, as a very powerful CNN alone (especially with attention mechanisms or transformers) might close the gap if given enough data. In these experiments, the hybrid shined in a regime with limited training samples, as injecting prior knowledge helped to generalize from fewer examples. With massive datasets, CNNs might implicitly learn some geometric relations (for instance, some attention-based CNNs learn to focus on the macula region for DR grading). Therefore, the utility of the RNN component may vary with dataset size and diversity.

From a deployment perspective, adding an RNN increases the inference time slightly and complicates the architecture. However, given modern hardware and the relatively small size of the feature vector, this overhead is minimal (in this case, feature extraction and LSTM took a few ms per image). The sequential features used (e.g. vessel curvature profile) were also of limited length (tens of steps), so the LSTM depth was manageable. If one were to incorporate a very long sequence (say, every pixel intensity along a spiral path through the retina), the sequence length could be large, and a transformer network might be more appropriate than an LSTM for long-range dependencies.

Future work can explore several directions. One idea is to make the network learn the geometric features in a differentiable manner. Instead of manually extracting features,

network modules can be designed to compute these relationships. For example, differentiable spatial attention could force the CNN to output the coordinates of the OD and fovea as part of the training process (using a spatial transformer network). This would integrate the pipelines and allow end-to-end training, potentially improving robustness. Another direction is exploring Graph Neural Networks (GNNs) to represent the vascular network: the retinal vessels form a graphical structure (with bifurcations and junctions) which could be encoded as a graph and analyzed by GNN layers, rather than linear sequence by LSTM. This may capture the full topology of the vasculature (important in diseases such as proliferative DR where new abnormal vessels form complex networks). Similarly, transformer-based architectures could replace or augment the LSTM to model global pairwise relationships between features (for example, a transformer could learn that lesion A is 500 µm from the fovea and lesion B is 510 µm from the fovea and pay more attention to those near the fovea).

Additionally, the hybrid approach can be extended to multimodal data or time series [26]. In teleophthalmology, patients sometimes have multiple imaging modalities (color fundus, OCT, fluorescein angiography). A CNN could handle each modality's image, and an RNN could integrate sequential input (such as changes over time or sequences of images). For example, for DR progression, an RNN can naturally model temporal sequences of a patient's images, capturing how lesions increase or resolve over time, and combining that with CNN image features yields a spatiotemporal model of disease progression. This work treated the sequential input as derived from a single image (spatial sequence), but applying the same concept in the time domain is a promising future avenue.

Finally, while the proposed model focused on DR lesions, the framework is general. Future research can apply the hybrid concept to other retinal conditions: e.g., combining CNN analysis of OD images with RNN modeling of visual field test sequences for glaucoma diagnosis (integrating structural and functional data), or analyzing retinal videos (sequential frames) for blood flow dynamics. The encouraging results of this study and others [27-29] suggest that bridging deep learning with classical domain features leads to more powerful and interpretable diagnostic systems. As we move forward, this synergy between learned representations and human-engineered features is likely to play a key role in building AI that is not only accurate but also aligned with clinical applications [30, 31].

V. CONCLUSION

This study presents a hybrid neural architecture that integrates CNN with RNN/LSTM to effectively leverage both deep learning and structured geometric features. The proposed approach significantly enhances diagnostic accuracy by combining powerful CNN-based feature extraction with interpretable sequential anatomical and geometric descriptors, achieving around 97% accuracy. This method offers improved interpretability, bridging the gap between traditional clinical insights and advanced deep-learning outcomes. The strength of this hybrid approach lies in its generalizability and adaptability to other biomedical imaging tasks beyond retinal analysis. By explicitly integrating domain-specific geometric features with deep neural network representations, this framework can be adapted to different medical imaging modalities and anatomical contexts, facilitating robust and transparent clinical decision-making. Future research should explore the extension of this hybrid method to multimodal data analysis and spatiotemporal imaging scenarios, further validating its versatility and effectiveness across a broader spectrum of clinical applications.

ACKNOWLEDGMENT

This research was funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP 19675574).

REFERENCES

[1] X. Zhu, R. M. Rangayyan, and A. L. Ells, "Detection of the Optic Nerve Head in Fundus Images of the Retina Using the Hough Transform for Circles," *Journal of Digital Imaging*, vol. 23, no. 3, pp. 332–341, Jun. 2010, https://doi.org/10.1007/s10278-009-9189-5.

[2] J. Kaur and D. Mittal, "A generalized method for the segmentation of exudates from pathological retinal fundus images," *Biocybernetics and Biomedical Engineering*, vol. 38, no. 1, pp. 27–53, Jan. 2018, https://doi.org/10.1016/j.bbe.2017.10.003.

[3] T. Walter, J. Klein, P. Massin, and A. Erginay, "A contribution of image processing to the diagnosis of diabetic retinopathy-detection of exudates in color fundus images of the human retina," *IEEE Transactions on Medical Imaging*, vol. 21, no. 10, pp. 1236–1243, Oct. 2002, https://doi.org/10.1109/TMI.2002.806290.

[4] G. Azzopardi, N. Strisciuglio, M. Vento, and N. Petkov, "Trainable COSFIRE filters for vessel delineation with application to retinal images," *Medical Image Analysis*, vol. 19, no. 1, pp. 46–57, Jan. 2015, https://doi.org/10.1016/j.media.2014.08.002.

[5] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, Dec. 2017, https://doi.org/10.1016/j.media.2017.07.005.

[6] V. Gulshan *et al.*, "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs," *JAMA*, vol. 316, no. 22, Dec. 2016, Art. no. 2402, https://doi.org/10.1001/jama.2016.17216.

[7] S. Keel, J. Wu, P. Y. Lee, J. Scheetz, and M. He, "Visualizing Deep Learning Models for the Detection of Referable Diabetic Retinopathy and Glaucoma," *JAMA Ophthalmology*, vol. 137, no. 3, Mar. 2019, Art. no. 288, https://doi.org/10.1001/jamaophthalmol.2018.6035.

[8] N. Sharma and P. Lalwani, "A multi model deep net with an explainable AI based framework for diabetic retinopathy segmentation and classification," *Scientific Reports*, vol. 15, no. 1, Mar. 2025, Art. no. 8777, https://doi.org/10.1038/s41598-025-93376-9.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, vol. 9351, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Springer International Publishing, 2015, pp. 234–241.

[10] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, Feb. 2020, https://doi.org/10.1109/TPAMI.2018.2844175.

[11] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, Apr. 2018, https://doi.org/10.1109/TPAMI.2017.2699184.

[12] P. Liskowski and K. Krawiec, "Segmenting Retinal Blood Vessels With Deep Neural Networks," *IEEE Transactions on Medical Imaging*, vol. 35, no. 11, Nov. 2016, https://doi.org/10.1109/TMI.2016.2546227.

[13] O. Poplavskyi *et al.*, "High-performance information technology for processing large datasets and biomedical images to improve the accuracy of computer-aided decision support systems," in *Photonics*

Applications in Astronomy, Communications, Industry, and High Energy Physics Experiments 2024, Lublin, Poland, Dec. 2024, https://doi.org/10.1117/12.3057444.

[14] "DIARETDB1 - Standard Diabetic Retinopathy Database." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/nguyenhung1903/diaretdb1-standard-diabetic-retinopathy-database.

[15] O. Mamyrbayev, S. Pavlov, Y. Saldan, K. Momynzhanova, and S. Zhumagulova, "Optical and Electronic Expert System for Diagnosing Eye Pathology in Glaucoma," *Applied Sciences*, vol. 14, no. 17, Sep. 2024, Art. no. 7816, https://doi.org/10.3390/app14177816.

[16] N. I. Zabolotna, S. V. Pavlov, O. V. Karas, and V. V. Sholota, "Processing and analysis of images in the multifunctional classification laser polarimetry system of biological objects," in *Reflection, Scattering, and Diffraction from Surfaces VI*, Sep. 2018, vol. 10750, pp. 82–89, https://doi.org/10.1117/12.2320209.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.

[18] R. Yasashvini, V. R. M. Sarobin, R. Panjanathan, J. S. Graceline, and J. L. Anbarasi, "Diabetic Retinopathy Classification Using CNN and Hybrid Deep Convolutional Neural Networks," *Symmetry*, vol. 14, no. 9, Sep. 2022, Art. no. 1932, https://doi.org/10.3390/sym14091932.

[19] S. V. Pavlov *et al.*, "Methods and computer tools for identifying diabetes-induced fundus pathology," in *Information Technology in Medical Diagnostics II*, CRC Press, 2019.

[20] O. Mamyrbayev, S. Pavlov, O. Karas, Y. Saldan, K. Momynzhanova, and S. Zhumagulova, "Increasing the reliability of diagnosis of diabetic retinopathy based on machine learning," *Eastern-European Journal of Enterprise Technologies*, vol. 2, no. 9 (128), pp. 17–26, Apr. 2024, https://doi.org/10.15587/1729-4061.2024.297849.

[21] Y. R. Saldan *et al.*, "Efficiency of optical-electronic systems: methods application for the analysis of structural changes in the process of eye grounds diagnosis," in *Photonics Applications in Astronomy, Communications, Industry, and High Energy Physics Experiments 2017*, Aug. 2017, vol. 10445, https://doi.org/10.1117/12.2280977.

[22] V. Lytvynenko *et al.*, "The use of Bayesian methods in the task of localizing the narcotic substances distribution," in *2019 IEEE 14th International Conference on Computer Sciences and Information Technologies (CSIT)*, Lviv, Ukraine, Sep. 2019, pp. 60–63, https://doi.org/10.1109/STC-CSIT.2019.8929835.

[23] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2009.

[24] R. Kvyetnyy, Y. Bunyak, O. Sofina, A. Kotyra, R. S. Romaniuk, and A. Tuleshova, "Blur recognition using second fundamental form of image surface," in *Optical Fibers and Their Applications 2015*, Dec. 2015, vol. 9816, pp. 289–297, https://doi.org/10.1117/12.2229103.

[25] O. G. Avrunin *et al.*, "Application of 3D printing technologies in building patient-specific training systems for computing planning in rhinology," in *Information Technology in Medical Diagnostics II*, CRC Press, 2019.

[26] A. N. Saeed, A Machine Learning based Approach for Segmenting Retinal Nerve Images using Artificial Neural Networks," *Engineering, Technology & Applied Science Research*, vol. 10, no. 4, pp. 5986–5991, Aug. 2020, https://doi.org/10.48084/etasr.3666.

[27] V. . Vassilenko *et al.*, "Automated features analysis of patients with spinal diseases using medical thermal images," in *Optical Fibers and Their Applications 2020*, Bialowieza, Poland, Jun. 2020, Art. no. 20, https://doi.org/10.1117/12.2569780.

[28] K. Oliullah, M. H. Rasel, Md. M. Islam, Md. R. Islam, Md. A. H. Wadud, and Md. Whaiduzzaman, "A stacked ensemble machine learning approach for the prediction of diabetes," *Journal of Diabetes & Metabolic Disorders*, vol. 23, no. 1, pp. 603–617, Nov. 2023, https://doi.org/10.1007/s40200-023-01321-2.

[29] A. Rodríguez-Miguel, C. Arruabarrena, G. Allendes, M. Olivera, J. Zarranz-Ventura, and M. A. Teus, "Hybrid deep learning models for the screening of Diabetic Macular Edema in optical coherence tomography volumes," *Scientific Reports*, vol. 14, no. 1, Jul. 2024, Art. no. 17633, https://doi.org/10.1038/s41598-024-68489-2.

[30] 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, Aug. 2018, Art. no. 39, https://doi.org/10.1038/s41746-018-0040-6.

[31] V. Bellemo *et al.*, "Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study," *The Lancet Digital Health*, vol. 1, no. 1, May 2019, https://doi.org/10.1016/S2589-7500(19)30004-4.