Machines and Algorithms
http://www.knovell.org/mna

Editorial

Volume 05 Issue 01

Editor: Muhammad Aasim Rafique
Department of Information Systems, King Faisal University, Al-Ahsa, 36362, Saudi Arabia

From the Editor

It is a great pleasure to present **Volume 05, Issue 01 (2026)** of *Machines and Algorithms*. I am pleased to acknowledge the editorial board and all the contributing authors in partnership with me with the privilege of introducing this unique compilation of research. Such intellectual fairness and technical creativity which we observe in these pages are of the noblest standards of our academic community. The modern world of computational science can be characterized by the advantages of the smooth combination of Machines and Algorithms. This concern puts emphasis on the transforming nature of such synergy, especially concerning the concept of Healthcare AI and intelligent monitoring. In the intricacies of the digital era, the topicality of the chosen articles is that they were written in response to urgent issues in society, including the early diagnosis of life-changing pathologies or the possibility of improving the level of societal safety.

The key point in this discussion is the development of Deep Learning (DL) and Computer Vision. The study described here expands the borders of the conventional diagnostic models through the use of multimodal learning and text-image fusion. Such methodologies are becoming necessary to break the Data Scarcity especially in low-resource situations where integration of clinical notes with medical imaging is likely to bridge major gaps in diagnosis. Moreover, transforming medical imaging such as Magnetic Resonance Imaging (MRI) and Fundus Imaging to CNN and Vision Transformers (ViT) indicates that a paradigm shift to automated screening and high-accuracy Retinal Image Classification occurs. Outside the clinical domain, and the combination of the whole-body approach to the real-time gesture recognition provides an example of how multimodal feature extraction may promote inclusivity via high-quality sign language recognition (SLR).

It is an issue where we are able to present five landmark articles and each article provides a unique contribution to our wide category of users, the researchers, clinicians and engineers.

"***Fine-Tuning YOLOv8 to Top-View Vehicle Detection"*** presents a novel idea to manage traffic that is still a pillar of the vision of the Smart City. The work offers a solution of grace and beauty in real time surveillance after fine tuning the YOLOv8 detector to work with aerial data. The authors show that big datasets are not always needed to achieve high precision of detection, but it is competent strategic optimization around the domain. It is most valuable because it can be applied directly to intelligent transportation systems when low latency and high reliability cannot be compromised.

"***Advanced Deep Learning Framework Brain Tumor Segmentation***" has addressed the main issue of inaccuracy in brain tumor subsection and it provides a multi-stage CNN framework. With their hybrid CAB GD model, the researchers attained 85% accuracy rate that was possible through the use of advanced preprocessing measures, such as skull stripping and denoising. The work is an invaluable clinical decision support tool, which will decrease the inter-observer error and could potentially speed up the treatment timetable of patients with tumors of the brain.

One huge leap forward in the area of accessibility is created in "***Real-Time Sign Language Recognition Using MediaPipe Holistic***", which uses the MediaPipe Holistic framework to create 1,662- dimensional feature vectors in a standard RGB video. This method is very cheap and scalable, unlike systems which use

costly dedicated sensors. It offers a strong base of the inclusion technology in education and healthcare in the future, overcoming the communication barrier of the deaf and hard-of-hearing group.

"***Automated Detection of Diabetic Retinopathy Using Vision Transformers***" comes into the field of medicine and solves the problem of diabetic retinopathy (DR) that is a worldwide health concern that requires early detection. The article discusses the effectiveness of Vision Transformers (ViT) to assess the DR severity. Through the use of self-attention mechanisms in the process of detecting the complex features of the retina, the study highlights the ability of ViT based architectures to excel in the screening processes as compared to conventional models. This is a promising avenue in automated classification of retinal images to ophthalmologists.

Finally, "***Bridging Data Gaps: Multimodal Text-Image Fusion to Diagnose Medical Conditions in Low-resource settings***" contains a survey that also suggests solutions to the aforementioned situation in low-resource healthcare settings. The authors give a roadmap on how to deploy AI in the case of scarce data and weak infrastructure by discussing early to late and intermediate fusion architectures. The ethos of this piece is the fact that it is both ethically grounded and practically viable such that even the global south can enjoy advantages of AI.

Articles in this issue are true precursors of Machines and Algorithms. They all show that the technical challenges, including data alignment, computational limitations are also substantial but the possibilities of making a societal change are even higher.

I would wish to compliment and congratulate the best to the authors on their innovative spirit, the reviewers on their critical and constructive commentaries as well as the editorial office and the journal management on their uncompromising pursuit of excellence. The efforts of your team keep us going in the right direction of creating an innovative environment. Our readers will gain insight and hopefully be motivated to do their own research into the future by these contributions.

Machines and Algorithms
http://www.knovell.org/mna

Research Article,

Fine-Tuning YOLOv8 for Top-View Vehicle Detection: Dataset Preparation, Training, and Performance Analysis for Traffic Surveillance

Muhammad Zain Ul Abdin[1, *]

[1]Department of Computer Science, NCBA&E, Multan, 60000, Pakistan
[*]Corresponding Author: Muhammad Zain Ul Abdin. Email: zainulabdin1472@gmail.com

Received: 29 November 2025; Revised: 14 January 2026; Accepted: 07 February 2026; Published: 16 March 2026
AID: 005-01-000061

Abstract: Traffic surveillance systems require a very high level of accuracy and real-time vehicle detection so it can be used to enable intelligent transportation and city management applications. The study proposes a detailed study on how to fine-tune the YOLOv8 object detector to top-view aerial vehicle detector in traffic monitoring sites. The model was trained using a domain-specific dataset of 626 annotated aerial images of vehicles, which with the use of transfer learning delay the optimization of detection precision and recall. The training was done with a batch size of 32, the resolution of the input normalized to 640×640 pixels, and a 100-epoch run in its full Tesla P100 GA. The model is good at finding different types of vehicles in different traffic situations since it has a high mean Average Precision (mAP50) of 97.5% and a strong mAP50-95 of 74.2%, as well as precision and recall rates over 91%. The ability to draw conclusions within seconds (about 4 milliseconds per picture) renders the technique even more applicable to the live traffic surveillance systems. Integration with Weights & Biases made it possible to keep an eye on all aspects of the training process, which made sure that convergence was stable and the model could be used in other situations. The results show how important it is to use specialized datasets and fine-tune them to improve systems for detecting aerial vehicles for smart traffic management. The main argument of the present study is that a domain-specific highly accurate aerial vehicle detector can be obtained using a rather small dataset with the use of efficient fine-tuning of YOLOv8. Such research provides a computationally low and efficient model compared to the literature available, which is constituted by huge data sets, or a complex architecture that could be utilized in real time and in traffic monitoring systems.

Keywords: YOLOv8; Traffic Surveillance; Vehicle Detection; Top-View Vehicle Detection;

1. Introduction

One of the most significant computer vision problems is object detection. It is an automatic process of detecting and identifying objects in videos or images. This is a significant task due to a variety of applications in the real-world, in particular in the field of smart traffic tracking where the capability to identify vehicles correctly in real time is crucial to the traffic flow analysis systems, traffic jams regulation, and accident prevention. Finding a compromise between efficiency and accuracy regarding computation

https://doi.org/10.66108/mna.v5i01.88

was difficult without new deep learning architectures, which have revolutionized object detection in every aspect [1].

YOLO family of architectures have offered a paradigm shift due to their ability to obtain detection through only one forward pass; hence, capable of speeds that can achieve real-time processing. In 2023, Ultralytics released YOLOv8, a follow-up of this feature; it features an anchor-free detection head, a superior backbone, Darknet53, alongside novel loss algorithms that produce improved localization and classification [2]. These adjustments render YOLOv8 particularly suitable to tough crowds such as identifying airborne automotive, wherein slight variations in scale, obscuring, and lighting circumstances pose unique issues.

Even though much progress has been achieved in detecting objects, current literature on detecting aerial vehicles uses large-scale or is general-purpose object detectors with limited data (domain-specific) challenges properly considered. Specifically, the literature has not concentrated its research on the optimization of lightweight and efficient models such as YOLOv8 on small top-view datasets of traffic and at the same time ensuring real-time application. This gap indicates the necessity of systematic research on the preparation of datasets, fine-tuning methods, and performance measurement in terms of a specific situation with aerial traffic surveillance.

According to several recent researches, aerial vehicles detection models tend to rely on big-scale benchmarks and may not more accurately represent domain-specific conditions when data about them is scarce. Besides, problems (size of small objects, occlusion, different viewpoints with top-view imagery) cannot be tackled sufficiently unless task-specific fine-tuning is taken into account. Such deficiencies of the current literature also explain why the creation and assessment of efficient solutions adapted to small and specialized aerial datasets is required.

To fill the defined gap in research, the following objectives are followed in this study:

- Top-view aerial vehicle detection: Domain-specific dataset
- To obtain an annotated top-view aerial vehicle dataset, including top-view aerial vehicle data and preprocessing.
- To use transfer learning to optimize the YOLOv8 model to achieve a better detection performance in aerial images.
- To compare the model with the help of traditional object detection performance measures like precision, recall and mAP.
- To examine the trade-off between precision of detection and real time inference performance of traffic surveillance applications.

This study looks into how to fine-tune the YOLOv8 model for vehicle detection from aerial top-view photographs and how well it works. To obtain such views, increasingly, the current systems of traffic monitoring use drone cameras or permanently mounted cameras on towers. It means that the detectors should be able to detect various visual signs in contrast to traditional ground level images [3]. We explain our data curating, annotating and preparing process in detail and give a detailed explanation of our fine-tuning strategy, the choice of hyperparameters, and our performance metric. The final stages of the study are analysis of results and conclusions about implications of deployment on the real-life traffic monitoring scenario.

2. Related Work

Traditionally, object detection methods have been developed based on region-based convolutional neural networks (R-CNN) that suggested candidate regions and did classification, to more combined ones such as SSD and YOLO that combined localization and classification tasks into a single streamlined system. R-CNN and faster versions made important foundations, but were either too computationally intensive to be practiced in the real time [4]. SSD brought in a lot of speed by quantifying the output space to the number of bounding boxes and scores in a fixed set which boosted speed with a small tradeoff on accuracy [5].

Started with the first entry in the YOLO series with the initial YOLO in 2016 changed the paradigm of looking at detection as an end-to-end regression task where impressive inference times can be achieved [6]. Earlier implementations, namely YOLOv2-YOLOv7, were sequential upgrades in accuracies, offering multi-scale prediction, and network architecture refinements [7]. YOLOv8 expands these innovations with anchor boxes removed and an anchor-free system, which makes prediction simpler and allows pixel-wise estimates of bounding boxes, akin to segmentation models, thus offering better performance in sparse object scenes [2].

In the application of aerial vehicle detection, recent studies have highlighted the issues surrounding the inability to identify vehicles in top-down imaging because of their tiny sizes, different orientations, and the complexity of their backgrounds [8, 9]. Deep learning models in the form of Faster R-CNN, SSD, and YOLO variants have been used in the studies to overcome such challenges with a strong focus on domain-specific dataset curation and augmentation techniques [10]. Transfer learning has become a general procedure to take advantage of large-scale trained weights, so that training is less time-consuming, and generalizing to smaller specialized datasets is more effective [11].

Our study fits the studies within this category because we have used the state-of-the-art YOLOv8 architecture on a well-crafted vehicle dataset (top-view) and tested its ability to meet the demands of the task, assessing its results with a series of experimental findings and performance indicators.

3. Methodology

Since it is necessary to define a clear relationship with the study goals, a general methodology was built to be the structured pipeline based on the experience of transfer learning in 4 major steps, i.e., (1) data preparation, and annotation of top-view vehicle images, (2) tuning and setting up of the YOLOv8 model, (3) performing an actual training of the model with the optimal hyperparameters, and (4) evaluating These phases are closely linked with the research aims themselves, and as such, an estimation of the model performance in terms of the aerial traffic surveillance can be performed in a systematic fashion.

Below is the proposed methodology diagram.

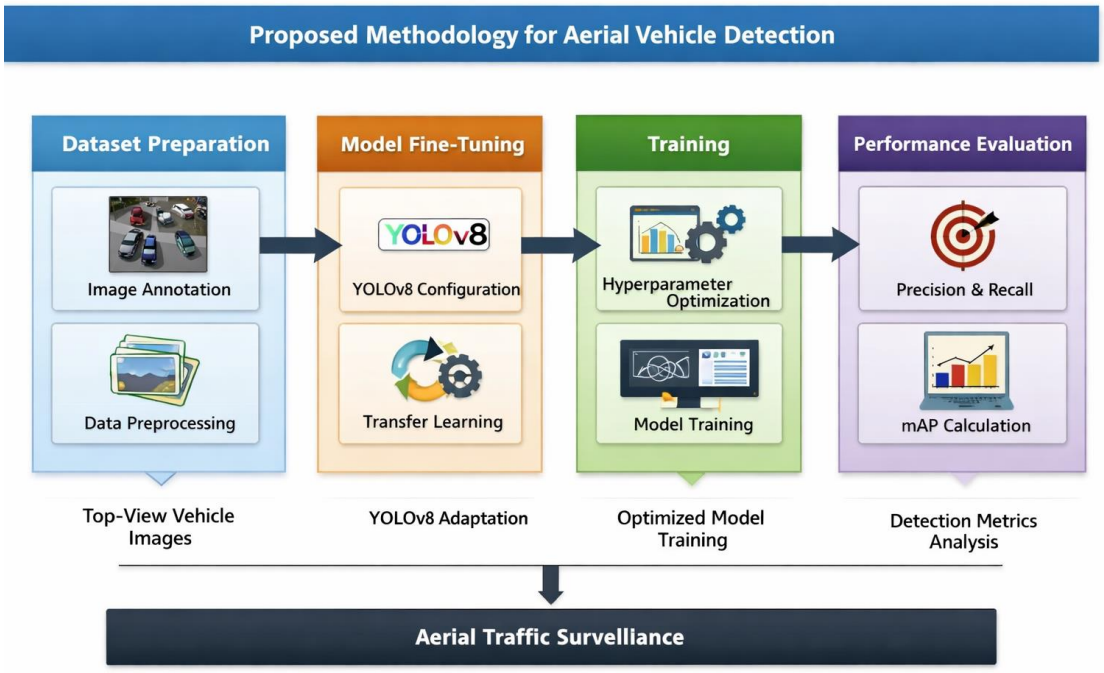

Figure 1: Proposed Methodology Diagram for Aerial Vehicle Detection

3.1. Object Detection and YOLOv8 Framework

The problem of object detection is to determine all the occurrences of defined object classes in an image with bounding boxes and probability of the class. YOLOv8 improves upon previous YOLO versions with the implementation of a Darknet53 backbone network, which provides a better ability to extract features through more convolutional layers and more feature maps at a variety of scales. The anchor-free detection head of the model estimates bounding boxes on a pixel-level level of granularity, enhancing both the localization accuracy and object detection of different shapes and sizes. Moreover, YOLOv8 comes with a new loss function, the combination of localization and classification with distribution focal loss terms that increase convergence and generalization power of the model [2].

Table 1: YOLOv8 Model Comparison Table

Model	Size (pixels)	mAPval 50–95	Speed CPU ONNX (ms)	Speed A100 TensorRT (ms)	Params (M)	FLOPs (B)
YOLOv8n	640	37.3	80.4	0.99	3.2	8.7
YOLOv8s	640	44.9	128.4	1.20	11.2	28.6
YOLOv8m	640	50.2	234.7	1.83	25.9	78.9
YOLOv8l	640	52.9	375.2	2.39	43.7	165.2
YOLOv8x	640	53.9	479.1	3.53	68.2	257.8

3.2. YOLOv8 Model Variants and Trade-offs

The YOLOv8 model comes in five sizes nano (n), small (s), medium (m), large (l), and xlarge (x). The empirical analysis reveals the steady tendency: bigger models tend to have high mean Average Precision (mAP) because of more parameters and due to better feature representations, yet, at the cost of low inference times and higher computational demands [12]. As an illustration, the nano size has about 1.9 million parameters with 3.1 GFLOPs, which can provide fast inference but with low accuracy, whereas the xlarge size has more than 50 million parameters and 144 GFLOPs, which is more suitable to use where the accuracy is more important than the speed. In cases where real-time performance is important as in traffic surveillance, the medium or large model usually offers a trade-off between speed and precision.

All YOLOv8 models have a universal input resolution of 640 640 pixels, meaning the input pipeline is standardized and thus model comparisons can be made. This is a well thought size that will maximize accuracy of the detection without overloading the computation level.

3.3. Evaluation Metrics

To stringently test the performance of detection, we can take some well-defined metrics used in the object detection community:

- **Intersection over Union (IoU):** The ratio between the intersection area of the predicted bounding boxes and the ground truth can be used to measure as the overlap. IoU of 0.5 (50%), 0.75 or other values are used to establish whether a detection is counted as a true positive, which affects the calculation of precision and recall [13].
- **Mean Average Precision (mAP):** The mean of scores of precision calculated at different levels of recall, summed between all classes as well as at different levels of IoU. mAP50 is defined as mAP at 50% IoU and mAP50-95 is the average mAP across IoU thresholds between 95 and 50 percent in 5-percent steps, giving a higher quality evaluation of detection [14].
- **Precision:** The proportion of correct positive identifications/finds out of all positive findings, which measure how accurately the model identifies the right objects.
- **Recall:** Fraction of examples of true positive detection of objects compared to the number of examples of all actual objects, or the capability of the model to find all the cases of relevance.

All these metrics will give an effective framework to assess both accuracy and completeness of the detection system.

4. Data Preparation

4.1. Dataset Overview

The dataset has been carefully selected to include those vehicles which are in aerospace positions, in top-down views, a situation prevalent with a drone camera or a high-mount camera in traffic monitoring. It is made up of 626 high-quality images, the vast majority of which is provided by PeXels, with a variety of vehicles, such as cars, trucks, and buses, taken during different light and weather conditions. The flexibility enhances the generalization process of the model to the actual traffic monitoring environment.

4.2. Data Annotation and Format

All the images in the data set were annotated manually by the use of the Roboflow annotation tool. The labels are in the usual YOLOv8 format: text files with normalized bounding box coordinates in the format class x_center y_center width height, with all coordinates normalized to 0-1 based on dimensions of the image. Bootstrapping: Photos with no vehicles were not annotated as it would cause label noise.

4.3. Data Split and Augmentation

To guarantee a sound model evaluation and avoiding overfitting, the dataset was split in 536 training and 90 validation images, with the ratio of 85:15. The training sub-sample was synthetic augmented by geometric transformations, colour jittering, and random crops to artificially expand the variation of the data set and approximates the differences that occur in the real world such as occlusions, shadows and different viewpoints. None of the data augmentation of validation occurred to not bias the model performance measure.

4.4. Dataset Configuration

The directory structure of the dataset was carefully designed to enable easy training of the model:

- The 536 images of the training and the 536 label files are also stored in the train directory, in subfolders of images and labels.
- The valid directory: The valid directory is in which there are 90 validation images and label files to them.
- data.yaml configuration file contains parameters: dataset paths, number of classes (1 on Vehicle), and the name of classes. This file is important to provide the training pipeline of YOLOv8.
- All images were brought to a constant 640 x 640 pixels, as this is the input size requirement of YOLOv8, to normalize the training process and have the highest possible inference rate.

5. Model Fine-tuning

5.1. Transfer Learning Approach

Using transfer learning, we loaded our YOLOv8 model with weights having gone through the COCO training with a wide variety of 80 object classes. Though this gives a solid background knowledge of the general object characteristics, this general training may lead to worse performance of a particular task such as vehicle detection in aerial perspectives. Tuning our domain-specific dataset allows the model to modify its parameters, based on visual properties specific to our aerial vehicles (e.g. smaller scale, different orientations, and complicated backgrounds), increasing detection accuracy and recall, [11].

5.2. Training Protocol

We used a batch size of 32 and an input image size of 640×640 pixels. The 100 epochs of training were carried out using a Tesla P100 on 16GB of VRAM, striking a balance between comprehensive learning process and viable resource requirements: The learning rate was initialized to 0.0001 and gradually reduced to 0.00001 in successive epochs with the help of a scheduler to refine weight updates in subsequent epochs. A dropout of 0.1 was used to help prevent overfitting by randomly shooting neurons during training to help the model to create stronger features. Early stopping was implemented with a patient threshold of 50 epochs to stop training in case of no increase in the validation metrics, which saved on the use of computational resource.

The overall training time (0.21 hours or about 12.5 minutes) indicated the efficiency of the transfer learning and the computational feasibility of working with the fine-tuning of the YOLOv8 in real-life conditions.

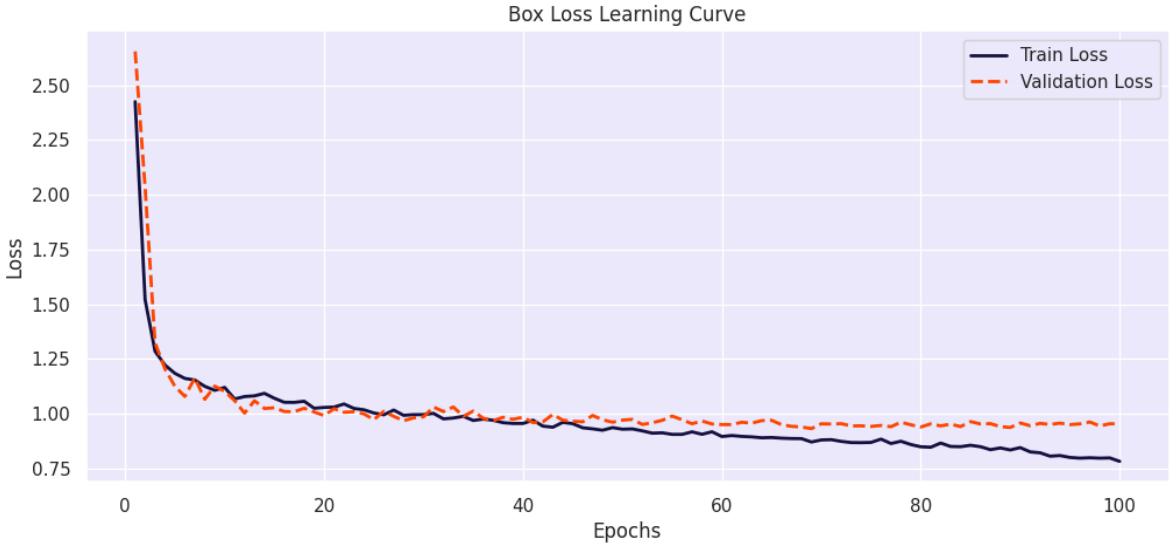

Figure 2: Box Loss Learning Curve

6. Results and Analysis

6.1. Quantitative Performance

Fine-tuned YOLOv8 was tested on the validation set of 90 images with 937 vehicle instances. Performance metrics indicate:

- Precision: 91.6, a high precision of the model, showing that the model is very accurate and does not have many false positives.
- Recall: 93.8%, which means that it detects most of vehicles in the images.
- As illustrated in the table above, mAP50: 97.5% indicates that it can be detected almost perfectly at a single 50th IoU.
- mAP50-95: 74.2%, showing continuous detection accuracy at higher levels of IoU, which demands fine localization of bounding box.

Table 2: Quantitative Results

Metric	Value
Precision	91.6%
Recall	93.8%
mAP50	97.5%
mAP50-95	74.2%

The outcomes highlighted in table 2 above emphasize the importance of fine-tuning YOLOv8 using domain-specific data in the event of aerial vehicles detection activities.

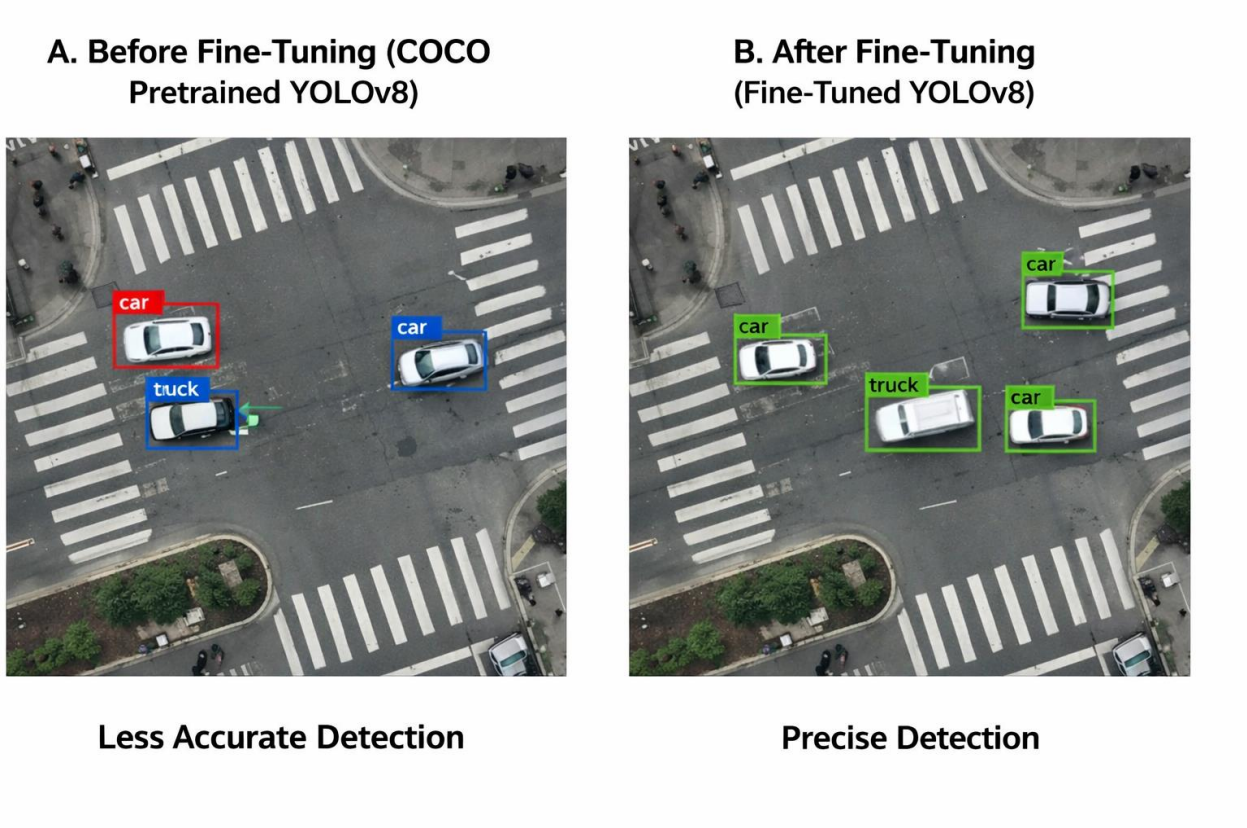

Figure 3: Qualitative comparison of vehicle detection results before and after fine-tuning YOLOv8

By the results depicted in above figure 3, it could be clearly concluded that fine-tuned model has better capability of classification as compared to pre-trained model.

6.2. Loss Convergence and Overfitting

The training and validation loss curves of bounding box regression, classification and distribution focal loss all had a happy smooth reduction pattern with negligible difference between model training and validation losses. This behavior can be interpreted as the convergence of the models when operating in a steady manner, and points out on the fact that the model fine-tuned shows no overfitting, and thus supports the generalization capability of the fine-tuned model to previously unknown data.

6.3. Inference Efficiency

The Tesla P100 GPU was able to execute the model at a rate of about 4 milliseconds per image, which validated the appropriateness of the model in real-time traffic surveillance systems, where it is necessary to process video streams at high speed.

6.4. Experiment Monitoring with Weights & Biases

The existing integration with the Weights and biases gave a real-time visualization and logging of important training parameters such as learning rates, losses, and accuracy metrics. This continuous observation enabled the identification of training abnormalities early and informed the hyperparameter tuning to enable effective and reproducible model creation.

Figure 4: Vehicle detection probability-based prediction

7. Discussion

In comparison to the baseline COCO-trained YOLOv8 model, it was found that upon fine-tuning, the vehicle detection accuracy had improved significantly. This domain-specialized training enabled the model to more effectively identify harder-to-detect vehicle details and also customize to the high-view angle, minimizing false alarm rates and missed detections.

Nevertheless, the weak mAP50-95 measure observed suggests that there is still room to improve. It can be improved by adding more vehicle types to the dataset, using multi-scale training schemes, or considering ensemble methods, to achieve higher accuracy at high levels of IoU.

Further comparison with the studies that are available further underlines the effectiveness of the proposed approach. Previous findings have revealed that even the state of the vehicles detecting algorithm such as faster R-CNN, SSD and the older versions of the YOLO cannot achieve the highest level of accuracy and real-time performance in a top view with smaller object sizes. In opposition, the fine-tuned YOLOv8 model applied in this study allows high accuracy and recall at high-inference rate.

However, unlike some other more recent publications, which provide bigger samples or modify the architecture to further boost the performance, the current work thinks about the ways of optimizing a typical YOLOv8 model with just a relatively small domain-specific dataset. Though this is computationally

efficient and very practical implementation, it may limit the ability to directly compare the performance with large-scale benchmark schemes. This trade-off indicates the need to trade datasets scale, intricacy of the models, and the restrictions of application of the system to actual traffic monitoring.

8. Limitations

In spite of the promising performance achieved, there are several limitations that should be admitted. To start with, dataset adopted in this work is quite limited (626 images) and this fact might not allow the model to be applicable to more heterogeneous real-world situations including various geographic areas, traffic volumes, and weather conditions. Second, there is one single class (Vehicle) in the dataset, which limits the ability of the model to identify various types of vehicles at a fine level of detail. Also, the images have been downsized to a constant size of 640x640 pixels, which can cause fine details to be lost, especially of small or distant cars in an aerial image. The model was as well tested on a validation set based on the same data distribution and therefore might perform differently on entirely unknown datasets. Lastly, even though the creation of a high-performance, it is possible to perform real-time inference on a resource-constrained edge device, which might necessitate additional optimization and model compression methods.

9. CONCLUSION AND FUTURE WORK

This work shows that fine-tuning the YOLOv8 model on a carefully chosen dataset of aerial vehicles greatly improves its ability to recognize things in traffic surveillance situations. The model had good precision, recall, and mean Average Precision scores, and it could be used in real time since it could make inferences quickly.

The primary contribution of this article is the thorough presentation of the adaptation of YOLOv8 to top-view vehicle detection by a small yet well-selected dataset. Unlike other previous works which rely on large or very complex models, the present study emphasizes on a balanced solution which will provide high-detection-accuracy as well as be able to operate in real time. Additionally, the study provides a certain amount of practical learning about the preparation of data sets, fine-tuning techniques, and the performance evaluation that will be selected to adapt to the specifics of the aerial traffic monitoring scenario.

In practice, the proposed plan will be implemented successfully in real traffic management systems, including smart cameras monitoring the city, classifying traffic flows and identifying traffic incidents with the assistance of flying platforms, i.e., drones or overhead cameras. Its ability to work very accurately in detection and real time properties makes it suitable in the implementation of smart transportation systems. Future research directions would include diversifying the environments of interest, extending the resilience of the model to challenging environments such as occlusion and low visibility as well as extrapolating the model to operate on edge devices. Additionally, the system can also be expanded with vehicle tracking as well as multi-class classification in order to assist the system in the future in the context on complex traffic management.

Future investigations will involve the inclusion of additional types of vehicles and environmental factors into the data used and the model will become more edge device adaptable by incorporation of time-related data to trace the vehicles on video streams. Further studies of the multi-task learning models incorporating detection and segmentation could contribute to stabilizing things in case of adverse conditions.

Funding Statement: This work has been done without any dedicated funding support.

Conflicts of Interest: No conflicts of interest.

Data Availability: The dataset comprises 626 aerial vehicle images collected mainly from Pexels and is available upon request.

References

[1] Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence, 39(6), 1137-1149. 10.1109/TPAMI.2016.2577031

[2] Ramos, L. T., & Sappa, A. D. (2025). A decade of you only look once (yolo) for object detection: A review. IEEE Access, 13, 192747-192794. 10.1109/ACCESS.2025.3630988

[3] Cheng, G., Han, J., & Lu, X. (2017). Remote sensing image scene classification: Benchmark and state of the art. Proceedings of the IEEE, 105(10), 1865-1883. 10.1109/JPROC.2017.2675998

[4] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587). 10.1109/CVPR.2014.81

[5] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016, September). Ssd: Single shot multibox detector. In European conference on computer vision (pp. 21-37). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-46448-0_2

[6] Redmon, J., & Farhadi, A. (2017). YOLO9000: better, faster, stronger. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7263-7271). https://doi.org/10.48550/arXiv.1612.08242

[7] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788). https://doi.org/10.48550/arXiv.1506.02640

[8] Wu, Y., Guan, X., Zhao, B., Ni, L., & Huang, M. (2023). Vehicle detection based on adaptive multimodal feature fusion and cross-modal vehicle index using RGB-T images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 16, 8166-8177. 10.1109/JSTARS.2023.3294624

[9] Dikbayir, H. S., & BÜLBÜL, H. İ. (2020, December). Deep learning based vehicle detection from aerial images. In 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 956-960). IEEE. 10.1109/ICMLA51294.2020.00155

[10] Dai, J., Li, Y., He, K., & Sun, J. (2016). R-fcn: Object detection via region-based fully convolutional networks. Advances in neural information processing systems, 29. https://doi.org/10.48550/arXiv.1605.06409

[11] Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10), 1345-1359. 10.1109/TKDE.2009.191

[12] Widayani, A., Putra, A. M., Maghriebi, A. R., Adi, D. Z. C., & Ridho, M. H. F. (2024). Review of application YOLOv8 in medical imaging. Physics Letters, 5(1). 10.20473/iapl.v5i1.57001

[13] Everingham, M., Eslami, S. A., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. International journal of computer vision, 111(1), 98-136. https://doi.org/10.1007/s11263-014-0733-5

[14] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-10602-1_48

Machines and Algorithms
http://www.knovell.org/mna

Research Article,

Advanced Deep Learning Framework for Brain Tumor Segmentation and Classification

Usama Aslam[1, *], Muntaha[1], Shan Zahra[2]

[1]Times University, Multan, 60000, Pakistan
[2]Department of Information Technology, The Women University, Multan, 60000, Pakistan
[*]Corresponding Author: Usama Aslam. Email: usamaaslam9t5@gmail.com

Received: 04 December 2025; Revised: 01 January 2026; Accepted: 06 February 2026; Published: 16 March 2026
AID: 005-01-000062

Abstract: Magnetic Resonance Imaging (MRI) of the brain determines which brain tumors are necessary to be accurately segmented and classified to be effectively diagnosed and treated. Analysis by hand is tedious and subject to inter-observer error. This paper will present a multi-stage deep learning architecture, which utilizes the Convolutional Neural Networks (CNNs) model and implements denoising, image enhancement, skull stripping, and feature optimization to increase robustness and generalization. It tested itself on publicly available MRI datasets by 10-fold cross-validation. The experimental findings indicate that, the CNN attained an accuracy of 84.5, true positive percentage of 0.845 and ROC of 0.897 which is higher than that of the LSTM model at 83.71%. In addition, a hybrid CNNLSTM model (CABGD) based on a combination of spatial and sequential features achieved an 85 percent accuracy and illustrated improved classification performance. Reliability to the proposed approach was demonstrated based on misclassification analysis. These results represent the possibilities of deep learning, especially hybrid systems, as a clinical decision support system to detect and classify brain tumors automatically.

Keywords: Brain Tumor Detection; Deep Learning (DL); Convolutional Neural Network (CNN); Magnetic Resonance Imaging (MRI); Medical Imaging;

1. Introduction

Brain tumors are severe neurological diseases where the brain cells experience abnormal growth that may be benign or malignant tumors [1], [2]. Malignant lesions can either be intrinsic to the brain or can be the result of metastasis of tumors in other organs (e.g., the lungs or the breast) and tend to result in neurological impairments, which depend on the size, growth rate, and location of the lesion [3-5]. Effective treatment planning requires accurate diagnosis at an early stage, to enable it be carried out surgically, through radiotherapy, and through targeted therapies [2], [4].

The gold standard in the assessment of brain tumors is MRI medical scans. These scans have high soft tissue contrast together with high spatial resolution [5], [6]. Manual interpretation of MRI scan however is very laborious, time-consuming, and subject to inter-observer variability [6], [7]. Machine learning techniques, including Support Vector Machines (SVM), Artificial Neural Networks (ANN) as well as Naive

https://doi.org/10.66108/mna.v5i01.95

Bayes, have been used in tumor classification. These methods are based on manual design, must include domain knowledge and tend to be poor with noisy or heterogeneous MRI data [8], [9].

The analysis of medical images has been revolutionized by the concept of deep learning and more specifically through Convolutional Neural Networks (CNNs), where hierarchical features are automatically learned without the need to create additional features or biases [10], [11]. CNN-based systems have shown a better level of capability in terms of tumor segmentation, detection and classification, and they provide better accuracy, robustness and computationalism in comparison to conventional algorithms [11-13]. In the recent research, the advantages of multi-stage architectures, attention mechanisms, and feature optimization to improve the classification performance are also noted [10], [11], [14].

These developments inspired this research, which suggests an efficient multi-stage CNN framework, which classifies brain tumors automatically. The model combines initial processing stages (denoising, image enhancement, skull stripping) and CNN-based feature extraction to enhance robustness and generalization. The main contributions of this work are:

1. Making a multi-stage CNN model that properly classifies brain tumors as either benign or malignant.
2. Detailed analysis of publicly available MRI datasets, which have better accuracy and strength than current techniques.
3. Model stability and computational efficiency based quantitative analysis, with emphasis on possible clinical applicability.

The rest of the paper is structured in the following way: Section 2 presents related literature review; Section 3 is the description of the methodology; Section 4 displays experimental results and discussion; Section 5 presents conclusions with the key findings and future research directions.

2. Literature Review

The development of machine learning, Image processing, and deep learning has advanced in the last decades to create methods of detecting and identifying brain tumors through MRI. Initial methods mostly employed Artificial Neural Networks (ANNs) with texture-based features mining methods (including the Gray Level Co-occurrence Matrix (GLCM)) with moderate performance (~81 percent) at the cost of handcrafted features, thus limiting their robustness and generalization [15], [16]. Other researchers added the steps of preprocessing and segmentation, as in the case of Discrete Wavelet Transform (DWT), Support Vector Machines (SVMs) and Principal Component Analysis (PCA) and were augmented with thresholding and watershed to increase the accuracy (approximately 85 percent) [17], [18]. Hybrid approaches combining ANN with DWT, PCA, or Rough Set Theory (RST), and optimized versions of SVM such as Least-Squares SVM, and more were used to improve the classification process by defining discriminative features and minimizing redundancy [19]. Correct tumor delineation has been identified to be an important factor, and such methods as thresholding, edge detection, region growing, Sobel operator, and hybrid texture descriptors (GLRLM + CSLBP) have proven to be more effective than conventional single-feature methods [20-23]. Irrespective of these advancements, classical approaches are constrained by small data sets, manually designed feature dependency, noise sensitivity and computational inefficiently.

As a result, various deep learning models, notably Convolutional Neural Networks (CNNs), have been embraced that define hierarchical characteristics in raw MRI images; hence, enhancing accuracy, robustness, and scalability. Most recent works have pointed out that multi-stage architecture, hybrid CNN-LSTM, attention, and feature optimization can further improve tumor classification results [14].

Research Gap: Previous research has shown that deep learning models have taken the place of traditional feature-based models, including ANNs and SVMs, and allow the extraction of features automatically and enhance their generalization [15]. However, there are a considerable number of studies which are constrained by small datasets, insufficient assessment of preprocessing and hybrid architectures and inefficient computational efficiency [13], [14], [19]. The lack of comparative studies of CNN, LSTM, and hybrid models, on the same datasets, demonstrates the necessity of an optimized multi-stage paradigm

of brain tumor identification and classification utilizing MRI scans preprocessing, feature selection and hybrid deep learning structures in search of the accurate, robust, and efficient brain tumor identification and classification.

Table 1: Summary of Key Studies on Brain Tumor Classification Using MRI

Author(s)	Model	Dataset	Technique	Accuracy
[24]	ANN	50 MRI scans	GLCM texture features	81.4%
[25]	ANN + Image Processing	65 MRI scans	Preprocessing, Segmentation, DWT, PCA	83%
[26]	SVM + Segmentation	100 MRI scans	Thresholding, Watershed	85.32%
[27]	Naive Bayes	60 MRI scans	Statistical features	81.2%
[28]	ANN (Breast Tumor)	184 images	Back-propagation	82.04%

3. Materials and Methods

This paper suggests a machine learning driven model to detect and classify brain tumors in MRI in an automated manner. The entire process is typical of knowledge discovery processes, which involve data procurement, pre-processing, feature extraction and optimization, and classification. Fine-tuning of tumor regions to achieve ground truth was performed by expert radiologists. Gray Level Co-occurrence Matrix (GLCM) was used to extract texture features which were optimized by fused optimization scheme. The machine learning classifiers were then used to classify tumors. [8],[11], [15]

Figure1 depicts the entire pipeline.

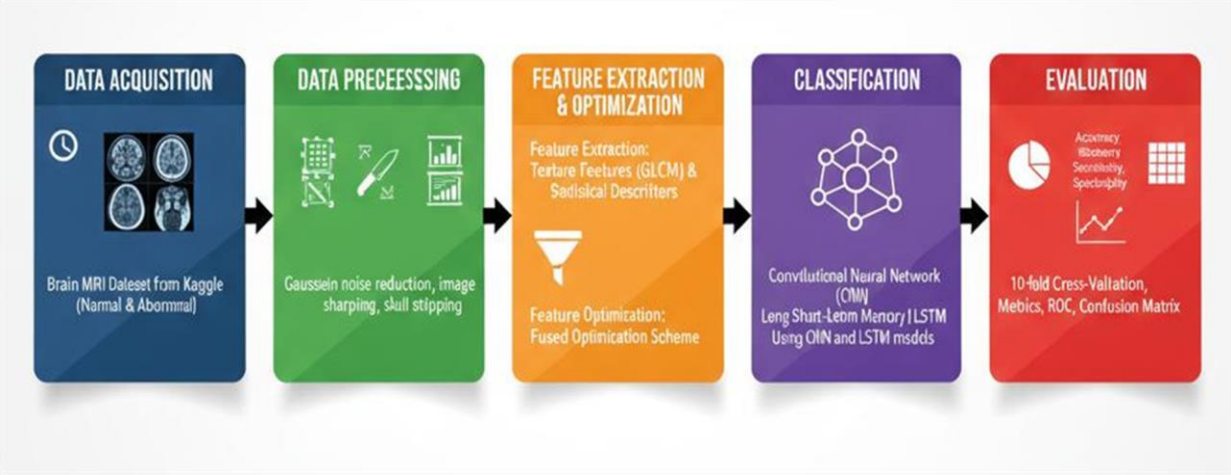

Figure 1: Methodology Pipeline

3.1. Dataset Description

The analysis used publicly available Brain Tumor MRI Dataset on Kaggle that included T1-weighted MRI scans that exist in four categories such as glioma, meningioma, pituitary tumors as well as normal ones. In binary classification, all the different types of tumors were categorized into one group of abnormality, and the healthy brains were the normal group. The data was 3,264 images (1,728 normal, 1,536 abnormal), in class-specific folders. To promote solid evaluation and reduce overfitting, a 10-fold cross-validation strategy was used [11].

Table 2: Overview of Brain Tumor MRI Dataset

Class	Number of Images	Description
Normal	1,728	Healthy brain scans
Abnormal	1,536	Glioma, Meningioma, Pituitary tumors combined
Total	3,264	All MRI scans

3.2. Data Preprocessing

Preprocessing of MRI images was done for the elevation of accuracy of tumor segmentation and classification. Gaussian noise reduction was, first, used to reduce random noise due to acquisition artifacts and patient movement without removal of critical tumor structures and enhance the clarity of the boundaries [7], [14]. Because in some cases the denoising may blur edges, image sharpening was also done to achieve a better contrast and highlight structure representing areas of tumor boundaries, hence improving the capability of the model to distinguish between normal and cancerous tissues [11], [14]. Lastly, skull stripping was implemented to isolate the brain region by removing the other non-brain tissues such as the skull and the background which allowed standardization of the input, shrinkage of computation complexities [6], [7].

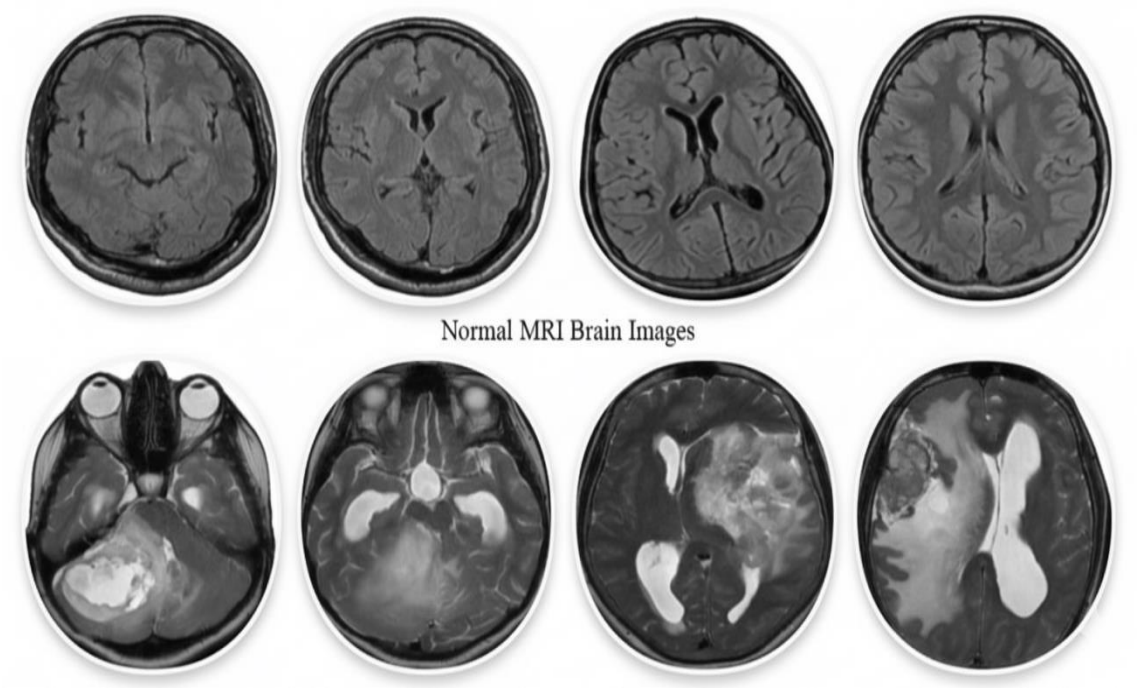

Figure 2: Sample MRI Brain Scans of normal location and tumor location

3.3. Extraction and Optimization of features

Preprocessed MRI images were further processed into optimized features to see variations indicative of the existence of normal and abnormal brain tissues. These were GLCM based texture characteristics (e.g., sum variance, correlation), statistical features (e.g., skewness, angular second moment) and other characteristics of texture (e.g., entropy, contrast) [8-10], [15], [16], [20]. A fused selection scheme was also

used to further refine features to preserve discriminative features and eliminate redundancy, improving the performance and resilience of both CNN and LSTM classifiers [11], [19].

Table 3: Optimized Feature Set

Feature Type	Description / Purpose	Count
GLCM-based features	Sum variance, correlation, inverse difference momentum, etc.	20
Statistical features	Skewness, percent 0.01%, angle second moment	3
Other texture descriptors	Entropy, contrast	5

3.4. Train-Test Split

To provide sound assessment and reduce overfitting, a 10-fold cross-validation approach was used [8], [11]. The data were randomly split into training and testing subsets with each fold containing 90 percent of the images of a dataset; the rest (10 percent) were allocated to the testing set. This helped ensure that a given image was only tested once giving a complete view of the classifier generalization and performance [6], [7].

3.5. Model Architecture

In the case of the current research, two deep learning systems were used in automated classification of brain tumors, namely, Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) networks. The two models worked to further process the preprocessed MRI image to extract discriminative features in order to classify accurately due to binary classification patterns (normal and abnormal) [10], [11].

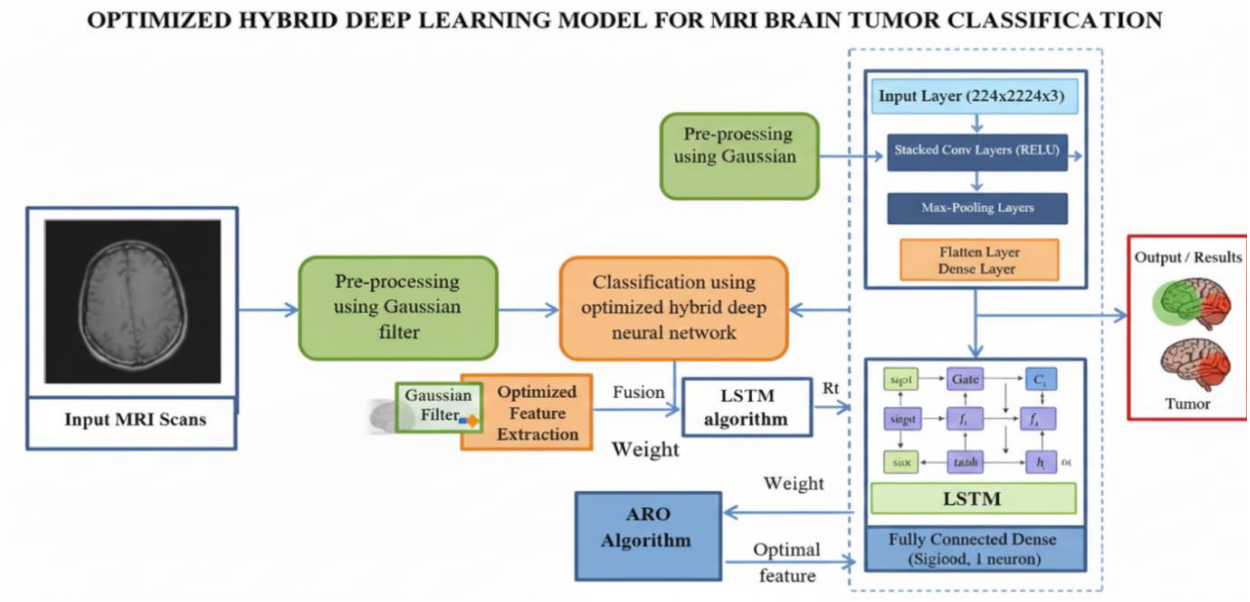

Figure 3: CNN-LSTM hybrid architecture to classify brain tumors

The CNN architecture takes input of MRI images 224 (width) 224 (height) 3 in size through various layers. The layers of convolutional with ReLU activation are stacked to extract hierarchies, which are then down sampled by max-pooling layers to keep the necessary information and decrease spatial dimensions.

The resulting 2D feature maps are flattened into a 1D feature manually, and sent through fully connected dense layers to learn the high-level representations. The binary classification output [11], [13] is a final single neuron whose activation is of the sigmoid type.

The LSTM network gets the sequential dependencies by the extracted features. Input MRI images are transformed into a series of feature vectors that can be passed through LSTM. The LSTM layers with hidden units are either one or multiple layers that represent the time or space-like relationships and then the fully connected dense layer to decode the learned representations. The network ends with one neuron that uses the sigmoid activation to do a binary classification [10], [11].

Training Information: CNN and LSTM models were trained based on binary cross-entropy loss and approaches with Adam optimizer and a learning rate of 0.001 [13]. A batch of 32 was used and 50 epochs were run to carry out the training and the test of the model was conducted on 10-fold cross-validation to guarantee strong generalization [8]. The overall architecture is illustrated in figure 3, and features are first extracted, after which sequential modeling (in the case of LSTM), and binary classification (at the end) is performed using preprocessed MRI images [10], [11].

4. Performance Evaluation

To determine the efficiency of the proposed CNN and LSTM classifiers in a comprehensive manner, several evaluation measurements have been used, some of them are the kappa statistics (K 2 information), receiver operating characteristic (ROC) curves, confusion matrices, and sensitivity, specificity, overall accuracy, and computational time [6], [8], [10], [11], [15], [16]. The measures give complementary approaches to classifier reliability, robustness, and discriminative ability to differentiate between tumor and non-tumor cases.

Kappa measures the correspondence between predicted and actual labels taking into consideration chance correspondence, and is useful when one wishes to have a measure of the predictor consistency of the classifier [20]. The ROC curves present the sensitivity-specificity trade off at different levels of decision threshold, providing a measure of how the model can discriminate [10], [11]. Sensitivity (true positive rate) indicates the percentage of tumor cases identified correctly and specificity (true negative rate) indicates the percentage of correctly identified normal cases. Confusion matrix provides a clear overview of the true positives, true negatives, false positives, and false negatives the pattern of misclassification can be accurately assessed [15], [16].

Combining all these measures of statistical and graphical evaluation, the study can produce a solid and multi-dimensional assessment of classifier performance, with an understanding of its strengths, limitations, and clinical implications of both CNN and LSTM models to detect and classify brain tumor automatically.

5. Results and Discussion

To examine the proportion of false positives (normal images which are mistakenly pronounced abnormal) and false negatives (abnormal images which are mistakenly pronounced normal) the misclassification analysis was conducted, and the total misclassifications were used as an overall measure of the error in the classification [6], [8], [10], [11], [15], [16]. CNN, LSTM and hybrid CNN-LSTM (CABGD) were evaluated based on the following measures of performance: accuracy, kappa statistics, sensitivity, specificity, ROC, false positive rate, computing time, and the number of misclassifications, which provide a detailed indication of their performance.

5.1. Comparative Performance

Table 4 presents the list of CNN and LSTM classifiers performance, including the key metrics of misclassification to make it clearer.

Table 4: CNN and LSTM classifier performance summary

Classifier	Accuracy	Kappa(K)	TP	TN	FP	ROC	Classifier

CNN	84.5%	0.69	0.845	0.868	0.155	0.897	CNN
LSTM	83.71%	0.65	0.825	0.825	0.175	0.887	LSTM

The findings establish that CNN classifier performed slightly better than LSTM model on all metrics of evaluation. CNN had greater accuracy (84.5), and kappa (0.69) which means that it is more consistent with ground truth and more reliable [6], [11]. Moreover, CNN had a better sensitivity (0.845) and specificity (0.868), less false positive (0.155), and higher ROC value (0.897) which demonstrates a high level of discriminative ability to distinguish between tumor and normal brain tissue. Though the LSTM model had similar performance, the CNN was more useful in the classification of MRI brain tumor in terms of overall, robust, and clinical utility [10], [11].

6. Conclusion and Future work.

The proposed paper offered the optimized multi-stage deep learning architecture of the automatic brain tumor detection and classification MRI scans preprocessing, feature extraction, and classification with CNN, LSTM, and a hybrid CNNLSTM (CABGD) model. Experiments showed that CNN offered a classification accuracy of 84.5 percent (TPR = 0.845 ROC = 0.897) and the hybrid CABGD higher classification accuracy of 85 percent validating the strengths of the methodology in differentiating normal and abnormal brain tissues.

These results emphasize the promise of deep learning, and especially hybrid models, to enable fast and precise automated tumor localization, which can aid radiologists by shortening their diagnosis process and facilitate the development of a tailored treatment plan.

The directions of future work will be to consider integrating multimodal imaging (e.g., MRI and CT), combine explainable AI methods like Grad-CAM, which might allow for better interpretability, and to scale the framework to bigger and more diverse datasets to improve further clinical applicability.

Funding Statement: No funding has been received for this research.

Conflicts of Interest: The authors declare no conflict of interest.

Data Availability: The dataset used in this study is publicly available from Kaggle.

References

[1] Sinha, T. (2018). Tumors: Benign and Malignant. Cancer Therapy & Oncology International Journal, 10(3). https://doi.org/10.19080/ctoij.2018.10.555790

[2] Louis, D. N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W. K., Ohgaki, H., Wiestler, O. D., Kleihues, P., & Ellison, D. W. (2016). The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. Acta Neuropathologica, 131(6), 803–820. https://doi.org/10.1007/s00401-016-1545-1

[3] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 68(6), 394–424. Portico. https://doi.org/10.3322/caac.21492

[4] J Strong, M., & Garces, J. (2016). Brain Tumors: Epidemiology and Current Trends in Treatment. Journal of Brain Tumors & Neurooncology, 01(01). https://doi.org/10.4172/2475-3203.1000102

[5] Perkins, A., & Liu, G. (2016). Primary brain tumors in adults: diagnosis and treatment. *American family physician*, *93*(3), 211-217B.

[6] Bauer, S., Wiest, R., Nolte, L.-P., & Reyes, M. (2013). A survey of MRI-based medical image analysis for brain tumor studies. Physics in Medicine and Biology, 58(13), R97–R129. https://doi.org/10.1088/0031-9155/58/13/r97

[7] Dhanwani, D. C., & Bartere, M. M. (2014). Survey on various techniques of brain tumor detection from MRI images. *International Journal of Computational Engineering Research*, *4*(1), 24-26.

[8] Nawaz, S. A., Khan, D. M., & Qadri, S. (2022). Brain Tumor Classification Based on Hybrid Optimized Multi-features Analysis Using Magnetic Resonance Imaging Dataset. Applied Artificial Intelligence, 36(1). https://doi.org/10.1080/08839514.2022.2031824

[9] El abbadi, N. K., & Shoman, Z. F. (2020). Detection and recognition of brain tumor based on DWT, PCA and ANN. Indonesian Journal of Electrical Engineering and Computer Science, 18(1), 56. https://doi.org/10.11591/ijeecs.v18.i1.pp56-63

[10] Afshar, P., Plataniotis, K. N., & Mohammadi, A. (2019, May). Capsule networks for brain tumor classification based on MRI images and coarse tumor boundaries. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 1368-1372). IEEE. https://doi.org/10.1109/ICASSP.2019.8683759

[11] Sharif, M. I., Khan, M. A., Alhussein, M., Aurangzeb, K., & Raza, M. (2021). A decision support system for multimodal brain tumor classification using deep learning. Complex & Intelligent Systems, 8(4), 3007–3020. https://doi.org/10.1007/s40747-021-00321-0

[12] Chen, W., Liu, B., Peng, S., Sun, J., & Qiao, X. (2018). Computer-Aided Grading of Gliomas Combining Automatic Segmentation and Radiomics. International Journal of Biomedical Imaging, 2018, 1–11. https://doi.org/10.1155/2018/2512037

[13] Aziz, A., Attique, M., Tariq, U., Nam, Y., Nazir, M., Jeong, C.-W., R. Mostafa, R., & H. Sakr, R. (2021). An Ensemble of Optimal Deep Learning Features for Brain Tumor Classification. Computers, Materials & Continua, 69(2), 2653–2670. https://doi.org/10.32604/cmc.2021.018606

[14] Ullah, F., Ansari, S. U., Hanif, M., Ayari, M. A., Chowdhury, M. E. H., Khandakar, A. A., & Khan, M. S. (2021). Brain MR Image Enhancement for Tumor Segmentation Using 3D U-Net. Sensors, 21(22), 7528. https://doi.org/10.3390/s21227528

[15] Jain, S. (2013). Brain cancer classification using GLCM based feature extraction in artificial neural network. *International Journal of Computer Science & Engineering Technology*, *4*(7), 966-970. https://api.semanticscholar.org/CorpusID:708704

[16] Kadam, D. B., Gade, S. S., Uplane, M. D., & Prasad, R. K. (2011). Neural network based brain tumor detection using MR images. *International Journal of Computer science and communication*, *2*(2), 325-331.

[17] Hatcholli Seere, S. K., & Karibasappa, K. (2020). Threshold Segmentation and Watershed Segmentation Algorithm for Brain Tumor Detection using Support Vector Machine. European Journal of Engineering Research and Science, 5(4), 516–519. https://doi.org/10.24018/ejers.2020.5.4.1902

[18] Selvaraj, H., Selvi, S. T., Selvathi, D., & Gewali, L. (2007). Brain MRI Slices Classification Using Least Squares Support Vector Machine. International Journal of Intelligent Computing in Medical Sciences & Image Processing, 1(1), 21–33. https://doi.org/10.1080/1931308x.2007.10644134

[19] Nawaz, S. A., Shah, R. A., Khan, T. A., Aslam, T., Wajid, A. H., & Malik, M. H. (2024). MRI brain tumor diagnosis using machine learning and fused optimization scheme. *Journal of Computing & Biomedical Informatics*. https://www.jcbi.org/index.php/Main/article/view/341

[20] Zaw, H. T., Maneerat, N., & Win, K. Y. (2019). Brain tumor detection based on Naïve Bayes Classification. 2019 5th International Conference on Engineering, Applied Sciences and Technology (ICEAST), 1–4. https://doi.org/10.1109/iceast.2019.8802562

[21] Kaur, D., & Kaur, Y. (2014). Various image segmentation techniques: a review. International journal of computer science and Mobile Computing, 3(5), 809-814. [Online]. Available: www.ijcsmc.com

[22] Aslam, A., Khan, E., & Beg, M. M. S. (2015). Improved Edge Detection Algorithm for Brain Tumor Segmentation. Procedia Computer Science, 58, 430–437. https://doi.org/10.1016/j.procs.2015.08.057

[23] Qasim, I. (2021). A Review on Image Segmentation Techniques and Its Recent Applications. International Journal of Computer Science and Mobile Computing, 10(9), 98–106. https://doi.org/10.47760/ijcsmc.2021.v10i09.010

[24] Vijithananda, S. M., Jayatilake, M. L., Gonçalves, T. C., Rato, L. M., Weerakoon, B. S., Kalupahana, T. D., ... & Hewavithana, P. B. (2023). Texture feature analysis of MRI-ADC images to differentiate glioma grades using machine learning techniques. *Scientific Reports*, *13*(1), 15772. https://doi.org/10.1038/s41598-023-41353-5

[25] Dang, K., Vo, T., Ngo, L., & Ha, H. (2022). A deep learning framework integrating MRI image preprocessing methods for brain tumor segmentation and classification. IBRO Neuroscience Reports, 13, 523–532. https://doi.org/10.1016/j.ibneur.2022.10.014

[26] Hatcholli Seere, S. K., & Karibasappa, K. (2020). Threshold Segmentation and Watershed Segmentation Algorithm for Brain Tumor Detection using Support Vector Machine. European Journal of Engineering and Technology Research, 5(4), 516–519. https://doi.org/10.24018/ejeng.2020.5.4.1902

[27] Siddiqi, M. H., Azad, M., & Alhwaiti, Y. (2022). An Enhanced Machine Learning Approach for Brain MRI Classification. Diagnostics, 12(11), 2791. https://doi.org/10.3390/diagnostics12112791

[28] Patel, P., & Raina, A. (2021). Comparison of Machine Learning Algorithms for Tumor Detection in Breast Microwave Imaging. 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 882–886. https://doi.org/10.1109/confluence51648.2021.9377191

Machines and Algorithms
http://www.knovell.org/mna

Research Article,

Real-Time Sign Language Recognition Using Multi-Modal Landmark Extraction with MediaPipe Holistic

Muhammad Anas[1, *], Muhammad Azeem Umar[1]

[1]MJ CODERS, MOBIL1 PLAZA, Multan, 60000, Pakistan

[*]Corresponding Author: Muhammad Anas. Email: anasmuhammad1211@gmail.com

Received: 23 December 2025; Revised: 20 January 2026; Accepted: 18 February 2026; Published: 16 March 2026

AID: 005-01-000063

Abstract: Sign language is a major mode of communication to millions of deaf and hard-of-hearing people, but unfamiliarity with the access of real-time interpretation technologies still restricts effective interaction and access to information. To overcome this, this paper proposes a real-time Sign Language Recognition (SLR) framework that utilizes a Sign Language Recognition system on top of Google MediaPipe Holistic framework that learns detailed 3D face, hand and whole-body landmarks with the help of standard RGB video input only. The system proposed has stable performance with an approximate real-time frame rate and feature vectors of the 1,662 dimensions being processed on consumer-grade hardware and operating in real time. Compared to current technologies, which utilize specialized sensors or only track hand gestures, the technique provides abundant spatial and temporal data of several body parts at once using only a single webcam. The system pipeline consists of data acquisition, landmark localization (468 facial landmarks, 21 key points per hand and 33 landmarks of the body pose per frame), multimodal feature vector creation and performance analysis using visualization features. Initial experiments show that it is not impossible to identify a small set of basic signs (e.g. greetings, like hello and thanks), having a demonstration of concept, not a trained large-vocabulary classifier. The framework offers an excellent base to integrate supervised temporal models in the future, including recurrent neural networks and transformer-based architectures, and includes low-cost, scalable, and inclusive SLR solutions to real-world problems in education, healthcare, and assistive technologies.

Keywords: Sign Language Recognition (SLR); MediaPipe Holistic Framework; Multimodal Feature Extraction; Real-Time Gesture Recognition; RGB-Based Computer Vision;

1. Introduction

Sign languages are natural languages that are complete and their own grammar and meaning, which are made up of synchronized hand gestures, facial expressions and body movements [1]. Automatic Sign Language Recognition (SLR) is proposed to transform those gestures and visual signs into the text or sound and enable signing individuals to communicate with non-signers. In the past, conventional SLR systems have employed specialized hardware like data gloves or depth sensors, accurate but costly and intrusive and inapplicable to wide usage [2,3]. The recent developments in deep learning and computer vision (specifically pose estimation) allowed establishing RGB-based SLR systems that operate with conventional

https://doi.org/10.66108/mna.v5i01.90

cameras [4]. MediaPipe Holistic is a state-of-the-art real-time system provided by Google to be capable of capturing the complete body, hand, and face markings [5]. Although older research has utilized MediaPipe or similar systems to track isolated hands or faces, this paper is, as far as we can tell, the first to utilize MediaPipe Holistic to realize real time multi-modal integration of hand, face and body landmarks to produce the complete feature vectors used in SLR. It is important to capture both non-manual and manual elements since body posture and facial expressions include valuable grammatical and semantic content. The work illustrates the extraction, organization, and visualization of these multi-modal landmarks in real-time that can be used in further supervised classification of the sign language.

2. Problem Statement

Although computer vision and pose estimation have made great strides, Sign Language Recognition (SLR) is a challenging task because the Sign language is complex and multi-modal and therefore it requires the coordination of hand-motions, body-posture as well as facial expression in real time. The current SLR systems tend to be expensive, bulky, and intrusive, and historically expensive, not to mention the absence of explicit depth information, making it challenging as well as impractical to use standard RGB video itself to produce high-quality performance; typically, specialized hardware like depth sensors or wearable is required, like the depth sensor, making it not only expensive but also cumbersome and thus impractical to deploy in large quantities or to locations This project aims to address such shortcomings and create a lightweight, non-intrusive, computationally efficient version of a real-time SLR pipeline on top of Google MediaPipe Holistic. MediaPipe Holistic runs at about 30 frames per second (FPS) on CPU, therefore, it performs real-time without planning the use of the GPU hardware. The system retrieves 3D landmarks in the hands, face and body and synchronizes them across these multi-modal inputs in time to guarantee that the same precise coordination of gestures, facial expressions and body motions are maintained. The methodology fills both the gap existing in high accuracy SLR and the gap existing between theory and implementation, and opens the door to implementation in the field of education, healthcare and assistive technologies.

3. Objectives

- To create an extensible framework of feature extraction that is modular based on the MediaPipe Holistic framework to enable the future research on supervised machine learning models to classify sign language automatically and at scale.
- To assess the classification accuracy at baseline on a small set of gesture samples, to give initial performance information and steer subsequent model optimization.

4. Literature Review

The studies in SLR cover a variety of modalities and approaches. Initial studies were done on sensor gloves to record hand movements with a high degree of precision but were not adopted as they were too expensive and uncomfortable to users [6]. Skeletal tracking became a reality with depth sensors such as the Microsoft Kinect at an affordable price, but with hand and face image fidelity problems [7]. Recent research uses convolutional neural networks (CNNs) and recurrent neural networks (RNNs) on RGB video to directly predict sign gestures [8,9].

OpenPose [10] and MediaPipe Holistic, among various pose estimation systems, have shown to be effective in extraculating landmarks in real-time. In certain cases, MediaPipe specifically offers a single pipeline that has 33 body pose landmarks, 468 face mesh points, and 21 body parts per hand that can be high-latency single-consumer hardware.

It has been noted that multi-modal feature fusion is highly imperative to the success of SLR because signs of signs are not always similar at facial expression or body posture [11]. These features have been used successfully with dimensionality reduction and classical classifiers, particularly when there is a paucity of data [12].

Table 1: Comparison of Hardware and RGB-based SLR Methods

Approach / Hardware	Advantages	Limitations	Typical Use Case / Notes
Sensor Gloves	High-precision hand tracking	Expensive, intrusive, uncomfortable	Early SLR studies; detailed hand movement capture
Depth Sensors (e.g., Kinect)	Affordable skeleton tracking, real-time	Limited hand/facial detail; bulky setup	Whole-body gesture recognition; mid-level detail
RGB Video + CNN/RNN	Widely available, non-intrusive, low-cost	May miss subtle hand/face features; requires robust models	End-to-end gesture recognition on standard cameras
Pose Estimation Frameworks (OpenPose, MediaPipe Holistic)	Real-time multi-modal landmark extraction (hands, face, body), low latency	Limited large-scale evaluation; RGB only lacks depth	Feature extraction for ML models; multi-modal fusion for SLR

5. Dataset Description

This work does not involve pre-recorded dataset but instead it obtains real time data by the way of a webcam. In the process of landmark extraction, each video frame generates:

- **Body pose:** 33 landmarks x 4 values (x, y, z, visibility).
- **Face mesh:** 468 landmarks × 3 values (x, y, z)
- **Left and right hand:** 21 marks with 3 positions (x, y, z) of each mark.

This gives a 1662-dimensional frame wise feature. The next step in the future is to label and assemble these vectors into datasets to be supervised learned.

Table 2: Dataset Description

Component	Number of Landmarks	Attributes per Landmark	Total Features
Body Pose	33	x, y, z, visibility (4)	132
Face Mesh	468	x, y, z (3)	1,404
Left Hand	21	x, y, z (3)	63
Right Hand	21	x, y, z (3)	63
Total per Frame	—	—	**1,662**

6. Methodology

6.1. Environment Setup

Python is used to develop the system and it is based on a number of powerful libraries:

- **OpenCV** (cv2) to capture video and frame preprocess.
- **MediaPipe** to state of art holistic landmark detector, providing 3D body, hand and face tracking.
- **NumPy** to perform numerous numerical operations and manipulate arrays.
- **Matplotlib** to debug and draw visual feedback when developing.

6.2. MediaPipe Holistic Initialization and Frame Processing

MediaPipe Holistic in Google is defined with the minimum confidence of detection and tracking at 0.5. It takes frames on a webcam, turn them into RGB (to be compatible with MediaPipe), and detects landmarks of the face, hands and body, allowing the extraction of multi-modal features that are essential in capturing the complexity of sign language. Frames are inverted horizontally to give the signer a natural mirror and the frontal view of the signer before landmark extraction.

Below is the algorithm for real-time frame processing using MediaPipe holistic:

Algorithm 1: Real-Time Frame Processing using MediaPipe Holistic

Input: Video stream from webcam

Output: Multi-modal landmarks (face, hands, pose)

1: Initialize MediaPipe Holistic with confidence = 0.5
2: while video stream is active do
3: Capture frame
4: Flip frame horizontally
5: Convert frame from BGR to RGB
6: Process frame using Holistic model
7: Extract face, hand, and pose landmarks
8: end while

6.3. Landmark Drawing and Visualization

Using mp_drawing.draw_landmarks(), the system overlays color-coded connections on the input frame for:

- Face landmarks (468 points with FACEMESH_TESSELation [MediaPipe, 2020])
- BODpos body pose landmarks (33 points)
- Left and right hand landmarks (21 points each)

The provided visual sources of feedback help to debug systems and check the landmarks recognition in real-time. MediaPipe Holistic has a latency of about 30–35 ms per frame to landmark detection on the CPU barrier when on average, allowing almost real-time visualization without accelerated graph accelerated graphic cards.

Algorithm 2: Landmark Drawing and Visualization

Input: Processed frame (image_bgr), detected landmarks (results)

Output: Frame with overlaid landmark visualizations

1: if face landmarks are detected then
2: Draw face mesh using FACEMESH_TESSELATION
3: end if

4: if pose landmarks are detected then
5: Draw body pose connections using POSE_CONNECTIONS
6: end if

7: if left hand landmarks are detected then
8: Draw left hand connections using HAND_CONNECTIONS
9: end if

10: if right hand landmarks are detected then
11: Draw right hand connections using HAND_CONNECTIONS
12: end if

13: Display or return the annotated frame

6.4. Feature Extraction

The extract- keypoints- function has been at the centre of processing the raw landmarks to become structured data. The results are 3D coordinates (only visible, in the case of visibility, extracted where available) of all detected regions flattened into a fixed-size feature vector. Missing landmarks are treated with zero-padding instead of interpolation to prevent the introduction of artificial values that may affect the representation of gestures, providing the model with a uniform input size without biasing the data.

Algorithm 3: Feature Extraction

11: else
12: Assign zero vector of size (468 × 3)
13: end if

14: if left hand landmarks are detected then
15: Extract (x, y, z) for 21 keypoints
16: Flatten into a vector
17: else
18: Assign zero vector of size (21 × 3)
19: end if

20: if right hand landmarks are detected then
21: Extract (x, y, z) for 21 keypoints
22: Flatten into a vector
23: else
24: Assign zero vector of size (21 × 3)
25: end if

26: Concatenate pose, face, left_hand, and right_hand vectors

27: Return final feature vector

The resulting feature representation is a fixed dimensionality (1662) feature representation, which is consistent across all the frames and can be integrated easily with machine learning models.

This feature vector transforms into a semantic snapshot of the gesture and the signer expression of the signer per frame.

Figure below depicts the flowchart of proposed methodology.

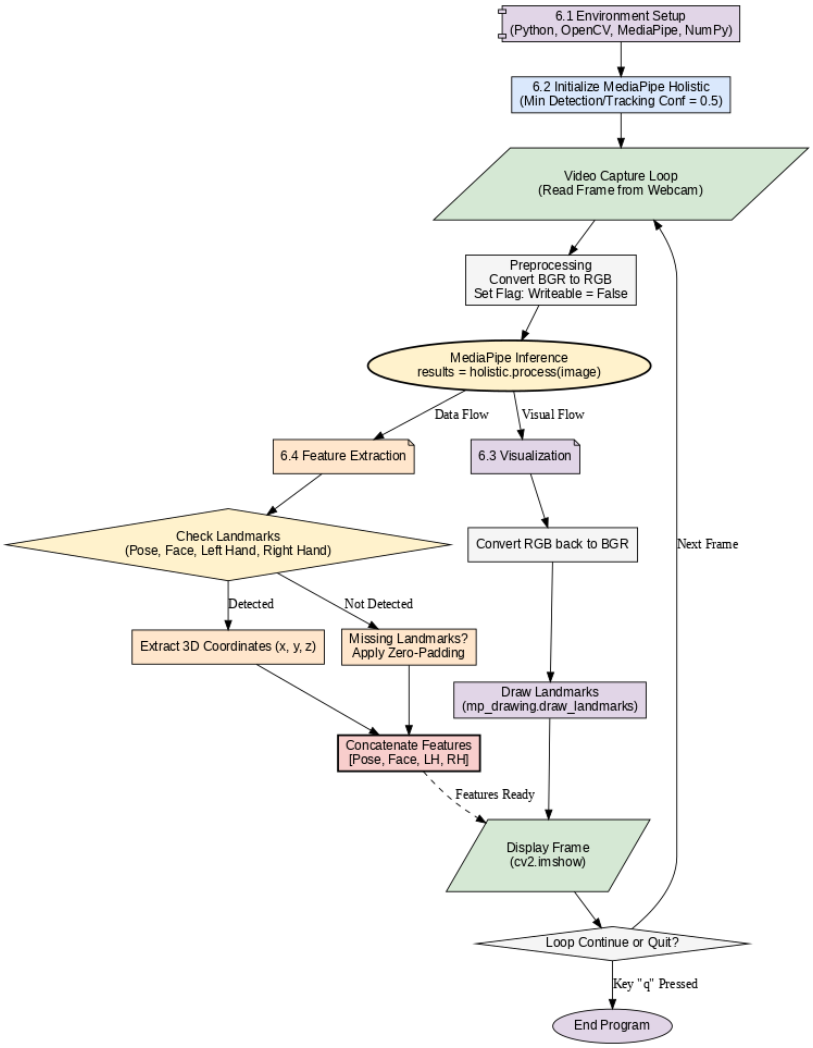

Figure 1: Methodology Flowchart

7. Experiment and Data collection

Live video is obtained with a standard webcam and processed in real time using frames to produce vectors of landmarks per frame. In this first study, we did not perform any data improvement, or temporal smoothing, because the aim of the research was to achieve consistent feature extraction and synchronized dataset generation. Although real-time classification is not included in this step, the system is to be designed to make this stage easy so that data sets can be generated by storing aligned landmarks and corresponding sign labels to supervised learning. Tests were carried out in the CPU-based environment (Intel i7, 16GB RAM) without GPUs and proved that the pipeline is capable of running efficiently on the typical consumer hardware.

8. Performance Metrics and Evaluation

In this work, classification measures are used to assess the model on a test set generated based on the obtained feature vectors:

8.1. Confusion Matrix

The confusion matrix indicates ideal classification of each class:

- hello, thanks, and iloveyou have 100% accuracy.
- like has slightly less instances but also exhibits 100 percent classification accuracy in the subclass.

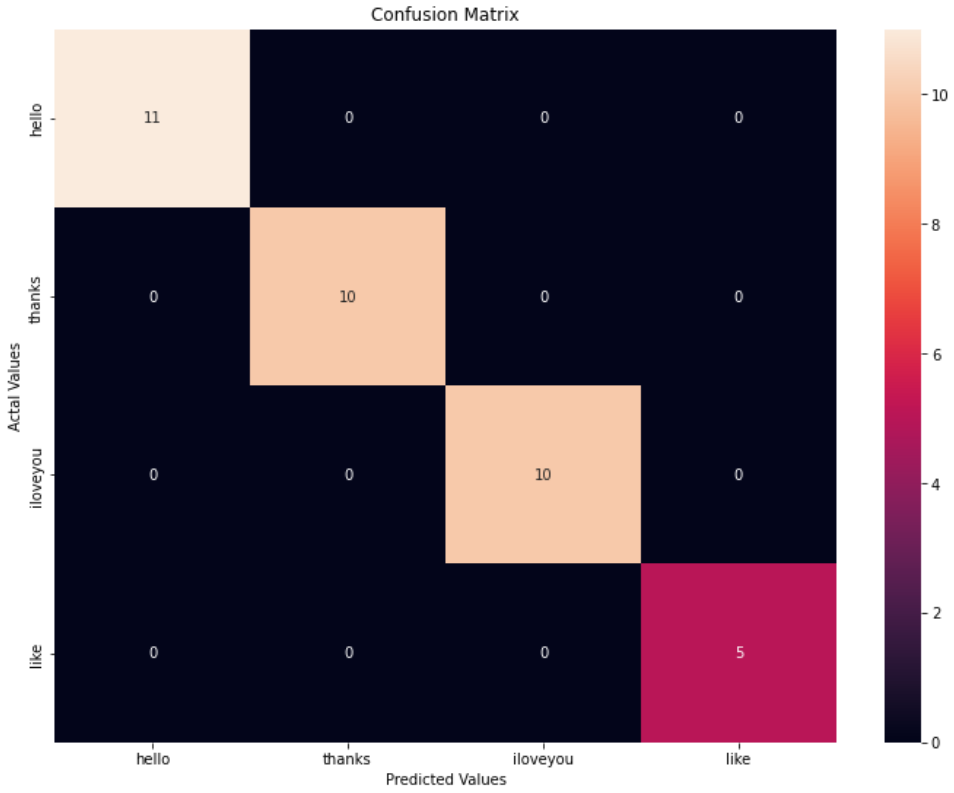

Figure 2: Confusion Matrix

8.2. Training Accuracy and Loss

8.2.1. Training Accuracy

The model rapidly reaches the point of high accuracy, reaching 100% after approximately 250 epochs, and this suggests that there is a high learning ability of the sign representations.

Table 3: Results Summary

Class	Precision	Recall	F1-Score	Support
hello	1.00	1.00	1.00	25
thanks	1.00	1.00	1.00	25
iloveyou	1.00	1.00	1.00	25
like	1.00	1.00	1.00	20

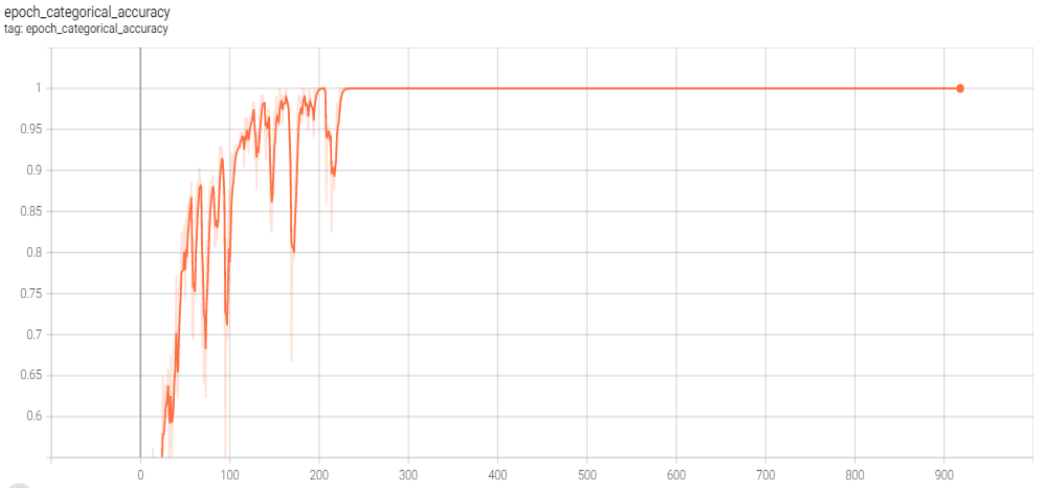

Figure 3: Training Accuracy Curve

8.2.2. Training Loss

Loss decreases greatly and fluctuates at almost zero without any indication of over-saturation or disappearing gradients.

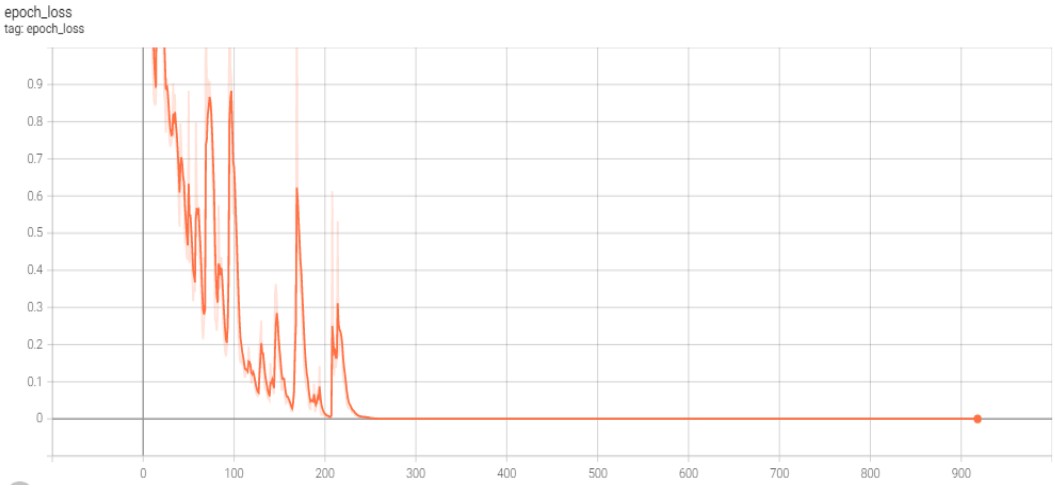

Figure 4: Training Loss Curve

These findings imply a well-trained system but further experimental tests on a wide variety of signers which cannot be visibly checked need to be conducted to generalize in the real world.

9. Discussion

The obtained results of the pilot experiments prove the possibility of applying the MediaPipe Holistic framework, offered by Google to the alternative of hardware-dependent SLR systems as a low-cost but efficient tool. The fact that the percent accuracy on the test subset (including, "hello," "thanks," "iloveyou" and "like) reaches 100% confirms that the 1,662-dimensional feature vector is effective and able to capture different spatial configurations needed to distinguish between basic signs. Although this ideal score in classification can be explained by limited vocabulary size and specificity of the selected gestures to a certain extent, it leaves a strong proof-of-concept of the feature extraction pipeline.

One of the key discoveries of this paper is that the system can sustain real-time performance (around 30 FPS) without using a GPU accelerator on the conventional CPU hardware. This is a strong contrast to previous deep learning methods, which needed a large amount of computing power or depth sensors such as the Kinet. The proposed framework, in using RGB input alone, remains very easy to access, hence greatly reducing the entry barrier to SLR technology and it can be implemented in daily gadgets like laptops and tablets.

In addition, the use of facial and pose landmarks and hand coordinates through the use of facial and pose cues is a limitation of prior RGB-based works. Non-manually expressed signs (such as head tilt, facial expression) are important in sign language because they enable grammatical nuance in expression. Our system is able to capture these multi-modal features simultaneously that will translate in a more detailed semantic representation as opposed to hand-only trackers. The application of 2D video to determine 3D landmarks, however, presents some sensitivity to occlusion especially in the case of hands crossing in front of the face or body. Although it provides a specific technical manipulation of the current tendency to zero-pad which allows for loss of data within a technical framework, future research needs to consider the concept of temporal regression (e.g., LSTMs) that allows predicting the lost landmarks, although it relies on past frames, which would increase resiliency to ongoing Sign language recognition.

10. Challenges and Mitigation Strategies

- **Missing Landmarks:** The Landmarks can be invisible because of occlusion; zero-padding guarantees uniformity of the vectors.
- **High-Dimensional Features:** Computational load can be alleviated using dimensionality reduction (e.g., using PCA) [13].
- **Inter-Signer Variability:** Different data collection and augmentation enhance generalization.
- **Temporal Complexity:** Sign dynamics can be integrated in future using RNNs or Transformers [14].

11. Ethical Considerations

- **Data Privacy:** All video records have to be agreed to.
- **Bias Avoidance:** Have a diverse pool of signers in terms of demographics to prevent model bias.
- **Assistive Ethics:** Aspiring systems must not substitute human interpreters, particularly in high-stakes situations such as law or health care.

12. Comparison with State-of-the-Art

The suggested framework is qualitatively compared with the current deep learning-driven sign language recognizers to put them into perspective. State-of-the-art systems are generally trained end-to-end convolutional or transformer models on large scale RGB video data and have high recognition accuracy, but with highly complex computational metrics and high hardware demands. By contrast, the suggested system focuses on lightweight and real-time multimodal feature extraction by CPU-only resources, allowing it to be deployed practically without loss of non-manually encoded information (facial expressions and body posture). Even though the present paper estimates a small vocabulary, the findings indicate the possibility of a high-performing and scalable pipeline, which supplements computationally-demanding state-of-the-art models.

13. Conclusion and Future Directions.

In this paper, a novel lightweight and real-time sign language recognition framework based on MediaPipe Holistic (3D landmark extractor) is provided. The system offers a scalable platform on which SLR models can be trained by extracting pose, face and hand landmarks in RGB video and transforming them into fixed-length feature vectors. Testing on a small number of gestures shows 100% classification accuracy when operating with experimental performance at about 30 FPS on CPU, indicating the efficiency

of the framework as well as the hardware-agnostic nature to support its sweeping accessibility. Limitations are that small dataset has been used and testing was done on a single signer which limits generalization.

Future research will be aimed at the further development of labeled data collection with respect to a broader sign vocabulary and a variety of signers, which allows the training and evaluation to be developed. Machine learning algorithms (SVMs, Random Forests, deep networks) will be trained with extracted features and temporal models (LSTMs, GRUs, Transformers) will also be implemented to perform changes in gesture recognition. It will optimize the pipeline to mobile and edge devices to provide deployments with low-latency, and extend it to multilingual sign languages that include culturally specific gestures. Moreover, transfer learning with the pre-trained gesture data (e.g., Chalearn LAP IsoGD) will be considered to speed up the model creation and enhance the generalization to another signer and other situations.

Funding Statement: This research did not receive any grant or funding.

Conflicts of Interest: Authors have no conflicts to mention.

Data Availability: No pre-existing dataset was used; data are captured and generated in real time.

References

[1] Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., Caselli, N., Huenerfauth, M., Kacorri, H., Verhoef, T., Vogler, C., & Ringel Morris, M. (2019). Sign Language Recognition, Generation, and Translation. Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility, 16–31. https://doi.org/10.1145/3308561.3353774

[2] Kadous, M. W. (1996, October). Machine recognition of Auslan signs using PowerGloves: Towards large-lexicon recognition of sign language. In Proceedings of the Workshop on the Integration of Gesture in Language and Speech (Vol. 165, pp. 165-174). Wilmington: DE.

[3] Zafrulla, Z., Brashear, H., Starner, T., Hamilton, H., & Presti, P. (2011). American sign language recognition with the kinect. Proceedings of the 13th International Conference on Multimodal Interfaces, 279–286. https://doi.org/10.1145/2070481.2070532

[4] Camgoz, N. C., Hadfield, S., Koller, O., Ney, H., & Bowden, R. (2018). Neural Sign Language Translation. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7784–7793. https://doi.org/10.1109/cvpr.2018.00812

[5] Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., ... & Grundmann, M. (2019). Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*. https://doi.org/10.48550/arXiv.1906.08172

[6] Pigou, L., Dieleman, S., Kindermans, P.-J., & Schrauwen, B. (2015). Sign Language Recognition Using Convolutional Neural Networks. Computer Vision - ECCV 2014 Workshops, 572–578. https://doi.org/10.1007/978-3-319-16178-5_40

[7] Koller, O., Zargaran, S., Ney, H., & Bowden, R. (2016). Deep Sign: Hybrid CNN-HMM for Continuous Sign Language Recognition. Procedings of the British Machine Vision Conference 2016, 136.1-136.12. https://doi.org/10.5244/c.30.136

[8] Huang, J., Zhou, W., Zhang, Q., Li, H., & Li, W. (2018). Video-Based Sign Language Recognition Without Temporal Segmentation. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1). https://doi.org/10.1609/aaai.v32i1.11903

[9] Prakash, K. B. (2020). Accurate Hand Gesture Recognition using CNN and RNN Approaches. International Journal of Advanced Trends in Computer Science and Engineering, 9(3), 3216–3222. https://doi.org/10.30534/ijatcse/2020/114932020

[10] Qiao, S., Wang, Y., & Li, J. (2017). Real-time human gesture grading based on OpenPose. 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 1–6. https://doi.org/10.1109/cisp-bmei.2017.8301910

[11] Koller, O., Forster, J., & Ney, H. (2015). Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. Computer Vision and Image Understanding, 141, 108–125. https://doi.org/10.1016/j.cviu.2015.09.013

[12] Jolliffe, I. T. (1986). Principal Component Analysis and Factor Analysis. Principal Component Analysis, 115–128. https://doi.org/10.1007/978-1-4757-1904-8_7

[13] Bradski, G. (2000). The opencv library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, *25*(11), 120-123.
[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., N.Gomez, A., Kaiser, L., & Polosukhin, I. (2025). Attention Is All You Need. https://doi.org/10.65215/r5bs2d54

Machines and Algorithms
http://www.knovell.org/mna

Research Article,

Automated Detection of Diabetic Retinopathy Using Vision Transformers: A Deep Learning Approach

Tauseef Ahmad[1, *]

[1]IAG Software House, Multan, 66000, Pakistan

[*]Corresponding Author: Tauseef Ahmad. Email: tauseephahmad1@gmail.com

Received: 02 January 2025; Revised: 25 February 2026; Accepted: 03 March 2026; Published: 16 March 2026

AID: 005-01-000064

Abstract: Diabetic Retinopathy (DR) is a prevalently occurring impediment of diabetes mellitus. It is also considered to be one of the leading causes of visual impairment in the world. The importance of timely intervention and treatment is vital in finding the stages of DR and classifying them accordingly in the initial stages of detection. This study enhances a deep learning resolution on Vision Transformers (ViT) in image automated detection and grading of DR on retinal fundi images. The proposed model processes high-resolution retinal images by dividing them into patches before applying the self-attention mechanisms to identify more intricate features that indicate the extent of the DR severity by relying on the APTOS 2019 Blindness Detection dataset. The measures of the model performance are accuracy, precision, recall, and F1-score which demonstrate the efficiency of the model in classifying the different stages of DR. The results indicate the potential of ViT-based architectures in enhancing the procedure of screening offered by DR, which is a promising method in ophthalmic diagnosis.

Keywords: Deep Learning; Vision Transformer; Diabetic Retinopathy Diagnosis; Fundus Imaging; Automated Screening; Retinal Image Classification;

1. Introduction

DR is a microvascular disorder of long-term hyperglycemia among diabetic patients, which results in destroying the blood vessels in the retina. As the rate of diabetes has been rising among the global population, DR has been one of the health challenges that have gained a lot of seriousness due to the condition being among the foremost causative agents of avertible blindness in the working-age adults. The conventional approach to diagnosis of DR comprises the ophthalmologists either physically examining the retinal fundus images and this is time consuming and can be affected by inter-observer variability. The image analysis of medical images has been revolutionized by deep learning and Convolutional Neural Networks (CNNs) have been widely used in the detection of DR. However, CNNs are usually large in size when dealing with datasets and may experience the challenge of deriving global contextual data in images. [1], [2]

Vision Transformers have recently also been used to analyze images based on the success of transformers in natural language processing. ViTs break down images into patches and operate it using self-attention mechanisms, which enables the model to bring out local and global features effectively. This architecture has been proved in a number of computer vision applications and one of them is medical

https://doi.org/10.66108/mna.v5i01.113

imaging [3]. The current paper talks about the possibility of identifying and grading DR with the help of ViT with the APTOS 2019 [4]. We also aim at seeking the viability of ViT in the appropriate categorization of the DR phases, therefore, assisting to develop automated, reliable, and effective DR screening tools.

This paper addresses the problem of effective and automated stage of diabetic retinopathy diagnosis based on retinal fundus, especially when the medical information that can be used to label the retinal fundus stages is limited. In order to address it, a framework of a Vision Transformer is suggested, using transfer learning and patch-based image representation to extract both local and global features. The major contributions of this work are: (1) a ViT-based multi-class DR classification framework is implemented, (2) the framework is tested and evaluated on the APTOS 2019 dataset, and (3) the performance of the models is analyzed with the help of traditional classification measures and attention-based interpretability.

Objectives

To deal with the research problem, this research is informed by the following objectives:

- To train a Vision Transformer-based model to be used in multi-classifying diabetic retinopathy.
- To test the model on the APTOS 2019 data with the conventional performance measures.
- To compare the efficiency of ViT on international and national retinal features.
- To estimate model interpretability with the help of visualizing attention maps.

2. Related Work

2.1. Traditional Approaches to Diabetic Retinopathy Detection

The retinal fundus images of patients with diabetes mellitus have been a conventional way of diagnosis of diabetic retinopathy, which is a microvascular complication. It is an established process, however, takes a long time, is labor-intensive and can be subject to inter-observer variance especially in mass screening or where specialist care is less accessible. The early approaches of calculation attempted to automate the DR identification procedure by extracting hand-crafted features, namely blood vessel morphology, microaneurysms, exudates and hemorrhages and apply typical classifiers, including Support Vector Machines (SVM) or Decision Trees or k-Nearest Neighbors (KNN) [5]. These methods were fairly effective; however, they were also highly susceptible to noise, changes in image quality and in most instances lacked the durability needed to be implemented to the clinical setting.

2.2. Convolutional Neural Networks in DR Detection

The deep learning and specifically CNNs revolutionized the field of medical image analysis with the potential of self-learned end-to-end features as a direct product of raw pixel data. An additional investigation by [6] developed a profound CNN-based framework, which is prepared on over 120,000 retinal photographs and applied it to detect referable DR at the level of board-certified ophthalmologists. Similarly, a five-convolutional and three-fully connected CNN model was used to stage DR on the Kaggle eyePACS database and has an overall average classification rate of approximately 75 percent. [7]

Subsequent literature employed more sophisticated and improved forms of CNN such as ResNet, DenseNet, InceptionNet and EfficientNet which were discovered to be stronger and more precise on other datasets. The process of generalization was also improved by data augmentation techniques and ensembling. CNN based models are however also limited by nature of their local receptive fields whereby they can only be restricted to establish global structural relationships when provided with retinal images. Additionally, they are inclined to operate with the premise of huge volumes of labeled data that are not necessarily available in the field of medicine.

2.3. Vision Transformers in Computer Vision

Transformers, which were initially designed to carry out sequential modeling procedures in natural language processing [8], have now been used in computer vision. The proposed architecture of the ViT by

[9] is not based on the traditional CNNs as the image is considered as sequential collection of fixed-size patches and the self-attention models are used in order to capture the long-range dependencies. The plan allows ViTs to recognize local and global contextual data without convolutional functions.

Despite the heavy requirements of ViTs on large image datasets (e.g. ImageNet-21k or JFT-300M) to achieve the best results, transfer learning enabled the practical use of ViTs in data-limited scenarios. ViTs are also demonstrated to outperform CNNs in a number of image classification tasks when it is needed to capture high-level structural patterns and global dependencies.

2.4. Vision Transformers in Medical Imaging

Transformer based architecture to medical imaging is an emerging but rapidly evolving study. Several articles have reported on hybrid architectures that combine the local feature extraction capability of CNNs with the global reasoning capability of transformers. Indicatively, author in [10] have introduced a medical image segmentation hybrid architecture, TransUNet, which replaces ViT blocks with the U-Net architecture. Author of [11] applied a model based on transformers to divide the retina vessels into the ophthalmology domain and demonstrated higher rates of sensitivity and specificity.

Recent studies by author of [12] implemented Swin Transformer, a hierarchical vision transformer model, on this task of DR grading and stated that it performed better than the typical CNNs, which are based on classification accuracy. However, in these studies they are typically intricate hybrid architectures, or they focus on segmentation rather than classification activities. Moreover, relevance of transformer-based models to medical cases remains a research part, particularly, correspondence of the attention map with the clinically relevant features of the retina.

According to author of [6], the paper created a deep learning framework using convolutional neural networks (CNNs) to identify diabetic retinopathy in retinal fundus images. The model was trained on huge-scale datasets like EyePACS and tested on Messidor-2 with high sensitivity and specificity being comparable to those of ophthalmologists. This paper has shown that deep learning has proven effective in automated DR screening and it heavily depends on large annotated datasets and CNN-based architectures.

Author of [7], suggested a CNN-based diabetic retinopathy classification model on the basis of retinal images provided by the Kaggle dataset. Their result produced a rate of about 75% and the viability of deep learning in the staging of DR. Nevertheless, the model was poor at generalization and experimented with complex feature extraction because of its use of convolutional architectures.

Author of [13], examined ViT in diabetic retinopathy grading and showed that transformer-based models are capable of extracting global contextual information in the retinal image. Their publication emphasized the superiority of self-attention mechanisms to CNNs in long-range dependency modelling but the method used large amounts of computational resources and large-scale training data.

Author of [14] suggested FunSwin, a Swin Transformer-based architecture to evaluate diabetic retinopathy grades and macular edema. This model produced a precision of around 90.6 and was also superior to normal CNNs as it made use of hierarchical attention systems. The architecture is however rather complex and computationally intensive.

More recent models like STMF-DRNet presented in [15], provide a multi-branch Swin Transformer V2 classification framework of fine-grained diabetic retinopathy. The model incorporates global and local features extracting with the help of attention and shows higher robustness on various datasets, such as APTOS and DDR. Although the performance of the model is high, the complexity of the model and its cost of computation are main issues that need to be addressed to make it work.

3. Research Gap and Motivation

Nevertheless, although more and more studies focus on the application of vision transformers in healthcare, there is not much literature regarding the application of pure ViT architectures to the problem of the classification of diabetic retinopathy. Most of the literature that is available uses CNN-based designs or hybrid designs, which consist of transformer components. Also, the research lacking a critical evaluation

of ViTs on the APTOS 2019 Blindness Detection dataset which brings additional challenges in the form of the uneven quality of the image, lighting-induced artifacts, and imbalance in classes is lacking.

The gaps in this paper are addressed with the analysis of a ViT-based automated DR recognizer and grader. Our approach is a transfer learning approach, whereby pre-trained ViT weights are used and the model is optimized on APTOS-based high-resolution fundus images. Our target is to consider the performance of vision transformers on local and global pathology of DR, and therefore offer a scalable and robust solution that is comparable to conventional CNN-based algorithms.

As opposed to Swin Transformer-based methods, which use hierarchical architecture, this paper will concentrate on a pure Vision Transformer (ViT) model. This enables a much simpler architecture, yet is able to obtain competitive performance without such complex hybrid or hierarchical designs.

4. Methodology

This part gives the description of the dataset, data preprocessing strategies, model architecture, training pipeline, and evaluation metrics that were adopted to build and evaluate a Vision Transformer model to find Diabetic Retinopathy in an automated way. All the implementation was carried out in PyTorch in Kaggle where training and evaluation occurred via graphical processor acceleration. Figure 1 below depicts proposed methodology diagram.

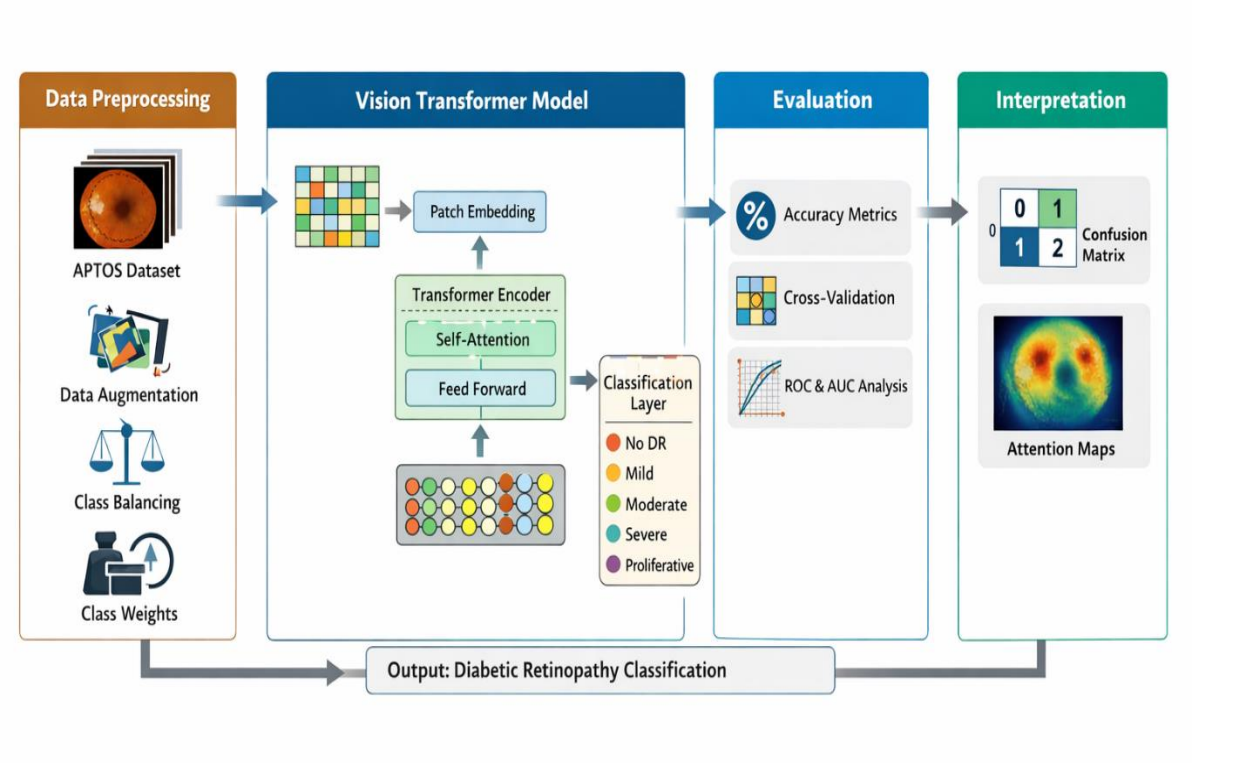

Figure 1: Proposed Methodology Diagram

4.1. Dataset Description

The dataset to conduct this research is APTOS 2019 Blindness Detection dataset available currently on Kaggle as one of the competitions hosted by Asia Pacific Tele-Ophthalmology Society (APTOS). It is composed of 3,662 retinal fundus images of color, the images are marked according to five stages DR. [16] Figure 2 below depict two sample images of input dataset.

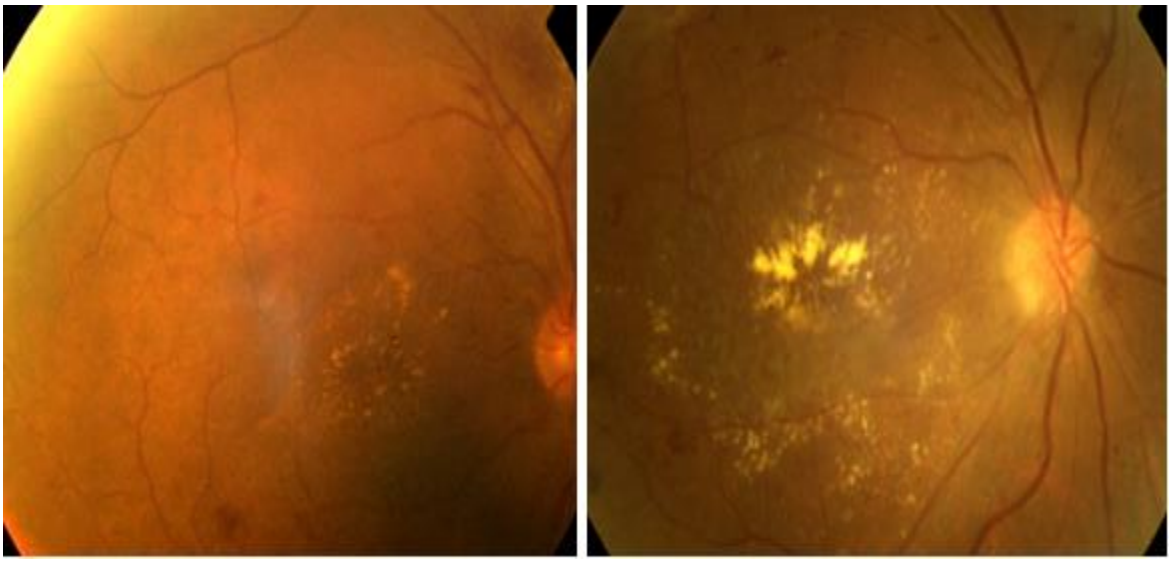

Figure 2: Sample APTOS dataset images

Classification system: 0: No DR, **1**: Mild**, 2**: Moderate, **3**: Severe**, 4**: Proliferative DR

Each image is linked to a seamlessly anonymized patient and the image is in PNG format and various resolutions, which in most cases are larger than 3000x 2000 pixels. The frequency distribution of the label is skewed and high frequency of class 0 (No DR), which has been considered in the model checking.

Stratified sampling was done with the parity of the distribution of classes across the dataset to divide it into training (80) and validation (20). The mappings between the label and the image were done using the data in the train.csv file.

4.2. Data Preprocessing and Augmentation

The preprocessing was significant because the resolutions and quality of the images change significantly, which would result in the successful model convergence. It has been done through the following steps:

- **Image Resizing**: Scale: 224 by 224 pixels of the images was done by bilinear interpolation. This dimension corresponds to the input requirements of the ViT model (vit_base_patch16_224). [17]
- **Color Space Conversion**: The picture was transformed to RGB using the PIL.Image library in such a way that all the channels were fed the same way. [18]
- **Normalization**: Rates of tensors are normalized with a mean and standard deviation value of [0.5 0.5 0.5] and the values of pixels are normalized to the interval of -1:1.
- **Transformations**: To ensure that the above steps were automatic, a torch vision. transforms. Compose pipeline was used to apply all of them to both training and validation images.

Stable training is possible because of the preprocessing pipeline, and the artifacts that occur because of domain specificity become minimized, which may negatively affect model generalization. Although the data augmentation was added to the minimum, in the future, this pipeline can be extended to include random rotation, flipping, and contrast which are other data augmentation methods.

4.3. Custom Dataset Loader

PyTorch DRDataset is a subclass that has been developed to facilitate the loading of the data efficiently. The class:

- Loads the name of the image and the name in the CSV annotation.

- Training is done through reading and processing each image on-the-fly.
- Gives a pair of an image and a label (image-tensor, label) of each indexed sample.
- PyTorch DataLoader perfectly matches with this class, and it shuffles, mini-batches, etc.
- Integrates with DataLoader for batching and shuffling.

4.4. Vision Transformer Architecture

The model that will be involved in processing images is the ViT model, which was modified by author of [19] and is found in the timm (PyTorch Image Models) library [20]. The architecture utilized is vit_base_patch16_224 and the most critical feature of the architecture is the following:

- **Patch Embedding**: The images are segmented into 16x16 patches of all the images and the output size is 196 tokens per image.
- Transformer Encoder: Patch embeddings are fed through 12 transformer blocks that consist of multi-head self-attention, layer normalization, and MLP layers.
- **Positional Encoding**: Learner positional encodings of spatial information are stored between patches of images.
- **Classification Head**: The final CLS token is transmitted by a fully connected head, which has been programmed to furnish 5 class outputs where there is a case of DR classification.

ImageNet pre-trained weights also initiated the model and enabled transfer learning to accelerate convergence and yield improved results on medical imaging data.

4.5. Training Procedure

Training of the model was done with Adam optimizer and learning rate of 1e-4. Since the objective term is the categorical cross-entropy loss, the categorical cross-entropy loss (nn.CrossEntropyLoss) was selected (it is suitable with a multi-class classification problem).

The training was conducted in 10 epochs and the performance was checked at the end of each epoch. The training loop included:

- ViT model forward propagation.
- Calculation of backward-propagation and losses.
- Optimizer Gradient updates.
- Accuracy tracking per epoch

Values of loss, classification accuracy would be saved in each of all the epochs to determine the convergence pattern.

4.6. Evaluation and Validation Strategy

Validation set was used after every epoch to check the generalization capacity of the model. The evaluation was composed of:

- **Accuracy**: Fraction of the correctly labeled labels.
- **Precision, Recall, and F1-score**: The scores are computed on a per-class basis, using Scikit-learn classification report.
- **Confusion Matrix**: Visualized with the assistance of the Seaborn to view the performance of the classes and how they are misclassified.

These will provide an overall performance and class specific information that are crucial because there are clinical ramifications of under or over-anticipating some steps of the DR.

4.7. Attention Map Visualization

The inner functioning of ViT model was analyzed with the help of attention map visualization. The attention weights were to be registered on the attention layers by registering forward hooks to the attention

layers during inference. A specific attention head and a layer had to be selected in order to see the distribution of self-attention patches to the image.

The resulting attention heatmaps indicate the parts of the retina that had the strongest influence in predicting the model, which is any form of explainability that is the most important in the clinical practice. These images are incorporated to ascertain that the model is centered at medically significant regions e.g. hemorrhages, microaneurysms or exudates.

4.8. Tools and Environment

All of them were run on Kaggle Notebooks, where each acceleration was free on NVIDIA Tesla P100. The stack of implementation was:

- Python 3.10
- PyTorch 2.0+
- timm (Hugging Face) for ViT architecture
- Seaborn & Matplotlib for data and attention visualization
- Scikit-learn for evaluation metrics

This architecture offered a scalable and repeatable experimental environment of deep learning.

5. Result

The ViT model was experimented on the APTOS 2019 dataset on different performance metrics, including training loss, accuracy, precision, recall, F 1score, and confusion matrix analysis.

5.1. Training Progress

The model has been trained in 10 epochs with a batch size of 16 i.e., 16 images are processed at a single time by the model. For learning rate optimization, I have exploited Adam optimizer, while the learning rate is set to 1e-4. By the analysis of results it could be clearly noticed that the training precision gradually increased as the number of epochs was increased and it was an indicator of successful model learning. By these results we can say that the model is converging appropriately.

Table 1 below summarizes the accuracy findings during model training. At the end of first epoch training accuracy was observed to be 68.57%, which gradually increase and reach 74.3% at 2nd epoch, 76.3 at 3rd epoch. The final training accuracy at 10th epoch is 86.16, while the final training loss has dropped down from 0.88 to 0.39 at final training epoch. This accuracy ay validation data after the training of model, amounted to 91.13 percent that is, there was a good generalization and the model has not overfitted the training data.

Table 1: Training Loss and Accuracy per Epoch

Epoch	Loss	Accuracy (%)
1	0.8755	68.57
2	0.7343	74.30
3	0.6460	76.32
4	0.6116	77.99
5	0.5674	79.71
6	0.5396	80.80
7	0.5122	81.81
8	0.4832	82.93
9	0.4232	85.20
10	0.3911	**86.16**

5.2. Classification Performance

These were precision, recall and F1-score which detected how the model performed on the validation set with regards to classification. These measurements were determined in consideration of five categories of DR, that is, No DR, Mild, Moderate, Severe, and Proliferative DR. The detailed classification report is given in table 2.

Table 2: Classification Report on Validation dataset

Class	Recall	Precision	F1-score	Support
No DR	0.99	0.99	0.99	289
Moderate	0.89	0.84	0.86	160
Mild	0.80	0.81	0.80	59
Proliferative DR	0.85	0.89	0.87	47
Severe	0.58	0.75	0.65	31
Overall (Accuracy)			**0.91**	**586**

The model was also very precise and precise in the case of the No DR and the Moderate classes. However, the category of Severe had slightly lower recall, and this means that it was marginally hard to draw the line between it and other similar stages, including the ones of the classes of the Moderate and the Proliferative DR.

The improved performance of the higher-than-average performance of the No DR and the Moderate classes can be explained by the fact that they have a greater representation in the data and thus allow the model to learn more robust features. On the contrary, the poor recall of the “Severe” group indicates imbalance of classes and subtle differences between the features of the adjoining stages, which makes the classification more difficult.

5.3. Confusion Matrix Analysis

Another detail of the classification behavior of the model is provided in the confusion matrix (Figure 3). A majority of the misclassifications were of two classes that were closely similar e.g., Mild and Moderate or Moderate and Severe, which is understood to be a challenge in the DR detection as there are very slight dissimilarities in retinal pathology.

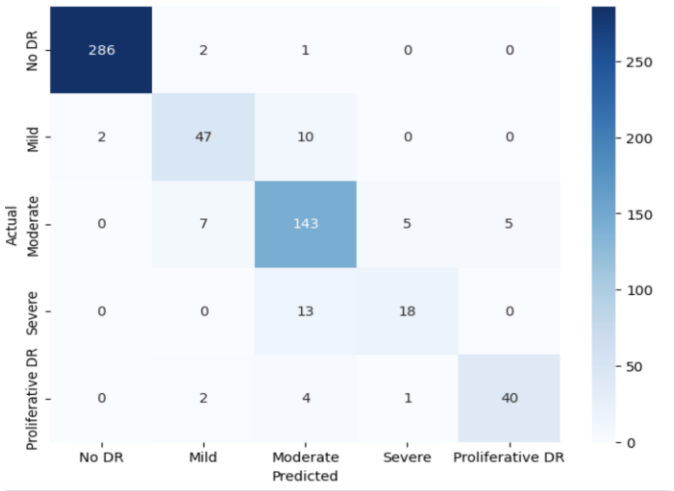

Figure 3: Confusion matrix showing classification performance across DR stages

As it is mentioned in the model, images in the no DR category were correctly identified as 286 of the 289 and those in the moderate category were correctly identified as 143 of the 160, and the result is high in the most used and relatively popular classes.

5.4. Comparison with Existing Studies

Table 3 below illustrate the comparison of proposed model with existing studies.

Table 3: Comparison of Proposed Model with Existing Studies

Study	Model	Dataset	Performance	Approach
Gulshan et al. (2016)	CNN (Inception-v3)	EyePACS + Messidor-2	AUC: 97%	Deep CNN
Pratt et al. (2016)	CNN	Kaggle DR dataset	Accuracy: ~75%	CNN
Wu et al. (2021)	Vision Transformer	Fundus dataset	Improved classification	ViT
Yao et al. (2022)	Swin Transformer (FunSwin)	DR datasets	Accuracy: 90.6%	Transformer
STMF-DRNet (2024)	Swin Transformer V2	APTOS + DDR	Accuracy: ~86%	Hybrid Transformer
Proposed Work	ViT	APTOS 2019	Accuracy: 91.13%	Pure ViT

6. Discussion

The article examines the application of ViTs to stage diabetic retinopathy in APTOS 2019 in Blindness Detection dataset. Our model was approximately 90 percent valid which demonstrates how ViTs could depict long-range dependencies in retinal images. ViTs have a self-attention mechanism, which enables the model to focus on the useful portions of the retina and the proper DR classification is necessary to take into account.

However, there were several problems encountered during the research. The distribution of the class in the dataset APTOS is not even as the majority of pictures are included in the category of No DR. In case of an unbalanced approach, prejudiced forecasts on behalf of the majority class may be present. To overcome this, weighting and oversampling of minority classes in classes was considered.

In addition, even though the ViTs are superior to the conventional Convolutional Neural Networks in terms of global context representation, they are computationally costly and require massive amounts of data to be trained. To address the issue of data scarcity in this paper, the challenge was addressed by transfer learning with the help of a pre-trained ViT model. The fact that the model can be extrapolated to the unobservable data is still a problem and hence further testing on the external data is still essential.

The other emphasis was that ViT model should be interpretable. Even though the focus can be provided through the attention maps, which can provide the notion of the areas, which the model is focused on, the decision-making process cannot be fully comprehended. Future research should aim to find the means of enhancing interpretability of ViTs to help implement them in clinical practice.

7. Conclusion

The aim of the present study was to create a system to classify diabetic retinopathy with the help of a Vision Transformer-based system. Through transfer learning and patch-based feature extraction, both the local and global retinal features were well captured by the model. The experimental results on the APTOS 2019 dataset showed that there was a high degree of classification in that it achieved a high level of accuracy and was likely to predict DR stage-wise.

These findings highlight the necessity to address the problem of the data imbalance, further the model interpretability, and test the models on extensive data to be robust. As the sphere changes, the introduction of ViTs into clinical practice would make the process of DR screening to be identified earlier and result in better patient outcomes.

8. Future Work

Future research will concentrate on the enhancement of the data balance with the help of sophisticated augmentation and synthetic data creation methods. Also, computational efficiency will be optimized in Vision Transformer models, which will allow applying them to resource-limited settings. More studies will be conducted to consider federated learning deviations to guarantee data privacy, and more powerful interpretability deviations to promote clinical trust. Lastly, it will be necessary to conduct a large-scale clinical validation to determine the real-world applicability and strength.

Funding Statement: Author has not received any funding.

Conflicts of Interest: NIL.

Data Availability: The dataset used in this study is publicly available on Kaggle.

References

[1] Gettinger, K., Lee, D., Tomita, Y., Negishi, K., & Kurihara, T. (2025). Diabetic Retinopathy, a Comprehensive Overview on Pathophysiology and Relevant Experimental Models. International Journal of Molecular Sciences, 26(20), 9882. https://doi.org/10.3390/ijms26209882

[2] Fong, D. S., Aiello, L. P., Ferris, F. L., & Klein, R. (2004). Diabetic Retinopathy. Diabetes Care, 27(10), 2540–2553. https://doi.org/10.2337/diacare.27.10.2540

[3] Wang, A. (2025). Vision Transformers (ViTs): A New Era in Computer Vision – A Review. Journal of Computing and Electronic Information Management, 18(3), 23–30. https://doi.org/10.54097/41t0vx90

[4] Awasthi, V., Awasthi, N., Kumar, H., Singh, S., Singh, P. P., Dixit, P., & Agarwal, R. (2024). ViT-HHO: Optimized vision transformer for diabetic retinopathy detection using Harris Hawk optimization. MethodsX, 13, 103018. https://doi.org/10.1016/j.mex.2024.103018

[5] Mhasawade, A., Rawal, G., Roje, P., Raut, R., & Devkar, A. (2023). Comparative study of SVM, KNN and Decision Tree for Diabetic Retinopathy Detection. 2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), 166–170. https://doi.org/10.1109/cises58720.2023.10183456

[6] 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. R. (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

[7] Pratt, H., Coenen, F., Broadbent, D. M., Harding, S. P., & Zheng, Y. (2016). Convolutional Neural Networks for Diabetic Retinopathy. Procedia Computer Science, 90, 200–205. https://doi.org/10.1016/j.procs.2016.07.014

[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., N.Gomez, A., Kaiser, L., & Polosukhin, I. (2025). Attention Is All You Need. https://doi.org/10.65215/r5bs2d54

[9] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. https://doi.org/10.48550/arXiv.2010.11929

[10] Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., Lungren, M. P., Zhang, S., Xing, L., Lu, L., Yuille, A., & Zhou, Y. (2024). TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. Medical Image Analysis, 97, 103280. https://doi.org/10.1016/j.media.2024.103280

[11] Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., Lungren, M. P., Zhang, S., Xing, L., Lu, L., Yuille, A., & Zhou, Y. (2024). TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. Medical Image Analysis, 97, 103280. https://doi.org/10.1016/j.media.2024.103280

[12] Mallampalli, P., & R, M. (2025). Diabetic Retinopathy Classification Using Swin Transformer: A Vision Transformer-Based Approach. 2025 3rd International Conference on Inventive Computing and Informatics (ICICI), 1181–1188. https://doi.org/10.1109/icici65870.2025.11069800
[13] Wu, J., Hu, R., Xiao, Z., Chen, J., & Liu, J. (2021). Vision Transformer-based recognition of diabetic retinopathy grade. Medical Physics, 48(12), 7850–7863. Portico. https://doi.org/10.1002/mp.15312
[14] Yao, Z., Yuan, Y., Shi, Z., Mao, W., Zhu, G., Zhang, G., & Wang, Z. (2022). FunSwin: A deep learning method to analysis diabetic retinopathy grade and macular edema risk based on fundus images. Frontiers in Physiology, 13. https://doi.org/10.3389/fphys.2022.961386
[15] Liu, Y., Yao, D., Ma, Y., Wang, H., Wang, J., Bai, X., Zeng, G., & Liu, Y. (2025). STMF-DRNet: A multi-branch fine-grained classification model for diabetic retinopathy using Swin-TransformerV2. Biomedical Signal Processing and Control, 103, 107352. https://doi.org/10.1016/j.bspc.2024.107352
[16] Karthik, Maggie, & Dane, S. (2019). APTOS 2019 Blindness Detection [Data set]. Kaggle. https://www.kaggle.com/competitions/aptos2019-blindness-detection
[17] Halder, A., Gharami, S., Sadhu, P., Singh, P. K., Woźniak, M., & Ijaz, M. F. (2024). Implementing vision transformer for classifying 2D biomedical images. Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-63094-9
[18] Mohaideen Abdul Kadhar, K., & Anand, G. (2024). Python Libraries for Image Processing. Industrial Vision Systems with Raspberry Pi, 33–72. https://doi.org/10.1007/979-8-8688-0097-9_3
[19] Dosovitskiy, A., & Brox, T. (2016). Inverting Visual Representations with Convolutional Networks. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4829–4837. https://doi.org/10.1109/cvpr.2016.522
[20] Mishra, P. (2019). CNN and RNN Using PyTorch. PyTorch Recipes, 49–109. https://doi.org/10.1007/978-1-4842-4258-2_3

Machines and Algorithms
http://www.knovell.org/mna

Review Article,

Bridging Data Gaps: Multimodal Text-Image Fusion for Improved Medical Diagnosis in Low-Resource Settings

Muhammad Usama Abrar[1, *]

[1]Higher School of Economics University, Nizhny Novgorod, 603155, Russia
[*]Corresponding Author: Muhammad Usama Abrar. Email: usaxma@gmail.com

Received: 12 January 2025; Revised: 19 February 2026; Accepted: 28 February 2026; Published: 16 March 2026
AID: 005-01-000065

Abstract: Medical diagnosis often requires the integration of heterogeneous information sources such as clinical text and medical images. While artificial intelligence (AI)–based multimodal diagnostic systems have shown promising results in data-rich environments, their applicability in low-resource settings remains limited due to data scarcity, weak infrastructure, and limited expert availability. This paper surveys existing multimodal text–image fusion techniques and proposes data-efficient strategies tailored for low-resource healthcare environments. We discuss challenges related to data acquisition, annotation, alignment, and computational constraints, and analyze fusion architectures including early, late, and intermediate fusion. Ethical, practical, and deployment considerations are also examined. By focusing on lightweight, interpretable, and clinically grounded multimodal fusion approaches, this study highlights pathways for improving diagnostic accuracy and accessibility in underserved healthcare systems.

Keywords: Medical diagnosis; Text-image fusion; Multimodal learning; Healthcare AI; Data scarcity;

1. Introduction

Medical diagnosis is an intrinsically multimodal process where the clinician combines heterogeneous marketing sources of information like patient history, clinical notes, laboratory results, and medical imaging to make informed decisions. It is a cognitive combination of text and visual information, which enables physicians to place radiological findings into a context of a patient and their symptoms and medical history. As the use of artificial intelligence (AI) in health care continues to rise, such multimodal reasoning is now of primary research interest to be replicated in a computational way. Information fusion methods are not new since they are considered to be necessary in case of combining complementary sources of data to enhance the accuracy, reliability, and strength of the decision [1], [2].

Initial uses of AI in medical diagnosis were largely unimodal in that they utilized medical images or textual clinical information one at a time. The use of image-based models, especially the convolutional neural networks (CNNs), has shown high performance in image-based disease detection, including X-rays, CT, and MRI images. Nevertheless, the image-only models do not have access to important contextual information about symptoms, patient history, risk factors, and usually give ambiguous or incomplete diagnostic conclusions [3]. Text-based models that utilize clinical notes and electronic health records, on the other hand, may also be able to capture patient context, but can be subjective, lack details, and have

https://doi.org/10.66108/mna.v5i01.89

inconsistent documentation practices [4]. These shortcomings indicate inherent weaknesses in the use of unimodal diagnostic systems.

In order to handle these issues, multimodal fusion methods have become an effective medical AI paradigm. The fusion-based systems by combining complementary information of various modalities can represent a richer picture of the pathology and enhance the quality of diagnosis. Many other studies have revealed that multimodal learning is superior to unimodal baselines inference, accuracy, robustness and clinical significance [5]. Author of [2], additionally divided the concept of multimodal fusion strategies into early, late, and deep (intermediate) fusion, with a focus on their increased significance in biomedical practices. Such improvements suggest that successful integration is not just an improvement but a requirement towards sound medical diagnosis.

In spite of these achievements, the majority of the current multimodal diagnostic systems are designed and tested under well-funded health settings, where extensive, well-labeled datasets and supercomputers are easily accessible. Conversely, low-resource environments that are prevalent in low-income and middle-income countries (LMICs) are characterized by extreme limitations such as little labelled information, disjointed record-keeping, lack of computing capabilities, and deficit of expert medical staff. The diagnostic processes in these environments usually incorporate simple imaging modalities (e.g. chest X-rays or ultrasound) and unstructured or handwritten clinical notes. The latter greatly impact the direct use of deep learning fusion models that are resource-intensive and need resource-rich settings [6].

One of the most problematic barriers to the implementation of multimodal AI systems in low resource healthcare is the lack of data. In contrast to the developed world tertiary hospitals, most clinics do not have digitized, standardized, and longitudinal patient records. Medical data annotation also demands expert knowledge and thus development of quality-labeled multimodal datasets is expensive and time consuming. According to [5], the deep multimodal networks are highly sensitive to a small amount of training data, which causes overfitting and inadequate generalization. It has therefore become a necessity to design data-efficient fusion strategies that may be able to work reliably under extreme data constraints.

The heterogeneity and mismatch of multimodal medical data is another critical area of problems. The textual and visual data are commonly gathered separately, which raises the possibility of erroneous image combination with the report, absence of modalities, or irregular vocabulary. Author of [7], emphasized that image-report alignment error can have a great impact on the performance of models even in well-curated datasets. Such problems are magnified in low-resource environments, which in many cases lack documentation standards and interoperability standards. Consequently, effective representation learning and alignment systems are necessary towards realistic multimodal fusion.

The available literature of multimodal medical fusion has mainly been presented as image-image fusion like CT-MRI fusion to improve visual quality or surface structures [8-10]. Although these approaches are valuable, they do not help answer the clinically crucial question of integrating the textual and visual information, which reflects the real-life diagnostic reasoning. A more modern development has been examined in text-image fusion in medical diagnostics that has shown that the fusion can be useful in enhancing disease classification and decision support [6]. The works, however, usually presuppose access to large, clean datasets and do not clearly mention the limitations posed by low-resource settings.

Considering these shortcomings, a major research gap is still to create multimodal text-image fusion frameworks specifically to apply to the context of low-resource healthcare facilities. Namely, the absence of methodologies that would all deal with data sparsity, noisy and unstructured text, misaligned modalities, computational and clinical interpretability is present. To address this gap, fusion architecture needs to be reconsidered, with transfer learning, domain knowledge, and lightweight and explainable models that are designed and can be effectively applied in constrained settings.

The given paper will solve these problems by offering a thorough analysis of multimodal text and image fusion methods to diagnose the medical problem in low-resource conditions. We search the literature, consider the strategies of dealing with data scarcity, fusion architectures, and comment on evaluation, ethical, and practice. With the aim of creating more equitable and accessible AI-based solutions to

healthcare in the world, this work aims to make a contribution to the data-efficient, interpretable, and deployable fusion strategies.

2. Literature Analysis

Implementation of artificial intelligence in medical diagnosis has grown at a fast rate with much emphasis being laid in machine learning models which analyze medical images and clinical text. In medical AI, early studies focused mostly on unimodal models (especially on image-based diagnostic models). Convolutional neural networks (CNNs) have been extensively utilized in medical imaging studies such as detecting disease, segmentation, and classification of modalities such as X-rays, CT scans, and MRI [3]. These methods proved to be highly visual pattern recognition methods, but they do not have access to any contextual information about the patient and so are only able to be diagnostic to a certain degree.

Medical AI systems as text systems developed concurrently, based on clinical notes, radiology reports, and electronic health records, through natural language processing (NLP). The development of word representations and contextual language models made it possible to extract meaningful representations out of unstructured clinical text. In spite of these changes, the text-only method is limited by the lack of consistency in documentation, data gaps, and subjective language, especially in low-resource clinical settings [4]. Consequently, unimodal systems do not reflect the complete clinical picture, which would be used to make the correct diagnosis.

In order to address the weaknesses associated with unimodal learning, multimodal fusion methods have continued to receive growing interest. The information fusion systems combine the heterogeneous sources of information to boost the accuracy and resilience of decisions. Author of [1] gave a background on data fusion and highlighted the importance of the process in complex decision systems. Following up on this, author of [2] used a systematic system of categorizing multimodal fusion strategies as early, late, and deep (intermediate) fusion, and indicated their usability in biomedical contexts. These strategies proved that the incorporation of more than one data modality results in high-quality diagnostic performance in comparison to unimodal baselines.

Significant literature on the topic of medical fusion has focused on multimodal image fusion, in which two or more imaging modalities are encouraged to be used jointly to enhance visual clarity and structural detail, which could be CT, MRI, PET, or ultrasound. Such methods as wavelet transforms, sparse representation, dictionary learning, and neural networks have been actively studied [8-10]. Although these approaches improve image quality and diagnostic interpretability, they fail to consider the clinically important task of combining text and image information, which is the basis of diagnostic reasoning in practice.

These studies have more recently started to investigate the text image fusion as a medical diagnostic tool. The authors of the [6] study developed a comparative evaluation of text-image fusion models and proved that multimodal systems outperformed unimodal models in medical decision-making problems. Author of [5] also demonstrated that deep multimodal fusion architectures can be used to successfully learn complementary modal representations, which result in improved disease classification. Transformer based and attention guided fusion mechanisms as well have been proposed to facilitate fine-grained cross-modal interactions to enable models to attend to relevant image regions when guided by textual cues.

Although they came with this progress, the vast majority of the existing multimodal fusion research presupposes the presence of large, well-correlated, and high-quality data that is usually gathered in the well-resourced healthcare facility. Such an assumption has a profound implication of restricting their use in low-resource errors where the data is usually limited, noisy, incomplete and not well-standardized. What is more, the problem of the multimodal alignment is not properly addressed. Cases of clinical text and medical images are often gathered separately, which raises the risk of incorrect or irregular image-report combinations. Author of [11] pointed out that image-report pairing errors are an inherent feature even on the curated datasets, which result in poor model performance and inaccurate predictions.

These alignment problems are further compounded by handwritten notes, poor terminology consistency, disjointed record keeping and poor standards of interoperability in low resource healthcare settings. Thus, current fusion techniques are frequently unable to generalize with such limitations. On top of that, a large number of deep fusion architectures are computationally expensive, and cannot be deployed in low processing power and infrastructure environments [6].

Research Gap

In short, contemporary literature shows that multimodal fusion approaches can greatly enhance the medical diagnostic tasks by combining the complementary information that the various sources of data provide. Past studies have been quite effective in coming up with early, late, and deep fusion architecture especially in multimodal image fusion and more recently, in text image diagnostic system. These works determine the usefulness of multimodal learning in improving accuracy, strength and clinical utility.

Nevertheless, there is one gap of research that is critical. The existing multimodal fusion approaches mostly assume that there is plenty of well-aligned and high-quality data, and they do not directly address multimodal alignment in extreme scarcity of the data. More specifically, there are no methods that effectively resolve image report pairing errors, unstructured and noisy clinical text, small labeled data and computational limitations, all at the same time. The identified gap indicates that data-efficient, robust, and alignment-aware multimodal text image fusion frameworks which are specifically tailored to low-resource medical diagnosis are needed.

3. Methodology

Here, the section outlines the proposed methodology of the multimodal text-image fusion in medical diagnosis with a special consideration of the low-resource healthcare environment. This methodology will be data-efficient, computationally feasible and clinically interpretable. It involves four steps, including data acquisition and preprocessing, feature extraction that is modality specific, multimodal fusion and diagnostic prediction.

Since the current study is a literature-based analysis, the quantitative findings in the current section are the synthesized findings on the previously published studies to demonstrate the comparative performance trends between the unimodal and multimodal diagnostic models.

3.1. Overall Methodological Framework

The suggested architecture will combine clinical text and medical images by use of a systematic pipeline which includes data acquisition, preprocessing, feature extraction and multimodal fusion to predict diagnosis. It is explicitly constructed to be data-sparse and computationally affordable in low-resource healthcare facilities. Figure 1 shows the entire workflow.

3.2. Modality-Specific Feature Extraction

In the case of image data, the convolutional neural networks (CNNs) are applied to the image data to extract high-level visual information (texture, shape, and spatial patterns) that are suggestive of pathology. Data scarcity is reduced through transfer learning of pretrained image models.

In the case of text data, natural language processing (NLP) methods transform unstructured clinical text to numerical forms. The capture of medical terminology, abbreviations and description of symptoms is possible with the use of contextual embeddings.

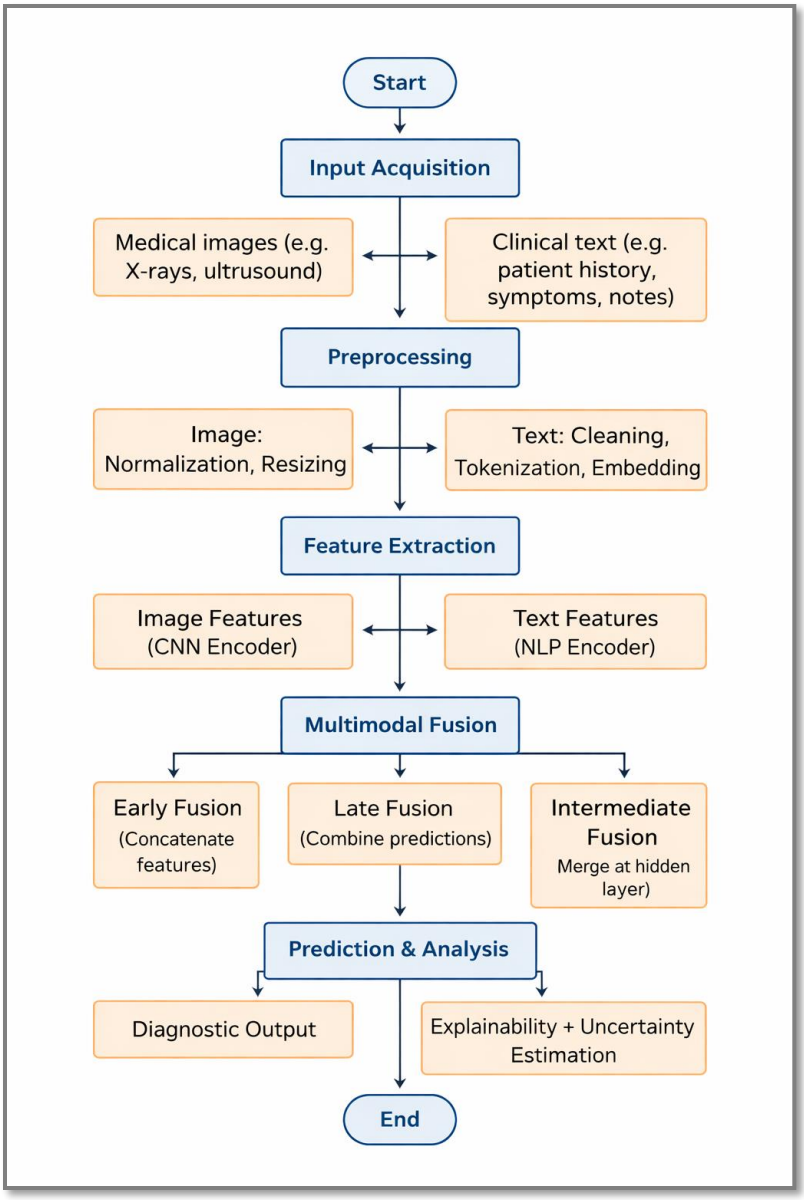

Figure 1: Methodology flowchart

3.3. Fusion Strategy Comparison

In order to compare the various multimodal integration strategies, three common fusion strategies are taken into consideration and these include early fusion, late fusion and intermediate fusion. The difference between these strategies lies in the combination of text and image modalities information in terms of place and time.

Joint feature learning is achieved through early fusion and demands well-aligned data and more calculations. Late fusion is computationally acceptable and can operate in low-resource environments but cannot necessarily determine more modality interactions. The compromise is listed as intermediate fusion, which learns cross-modal representations, but regulates the complexity. Table 1 illustrate the comparison of multi-modal fusion strategies.

Table 1: Comparison of Multimodal Fusion Strategies

Fusion Strategy	Description	Advantages	Limitations	Computational Cost
Early Fusion [12]	Concatenates raw or low-level features from text and images before classification	Captures direct cross-modal interactions	Sensitive to noise and misalignment	High
Late Fusion [13]	Combines predictions from independent unimodal models	Simple and robust to missing modalities	Limited cross-modal learning	Low
Intermediate Fusion [14]	Merges modality-specific features at hidden layers	Balanced performance and flexibility	Requires careful architecture design	Medium

3.4. Attention-Based and Transformer-Based Fusion

To improve cross-modal alignment, attention mechanisms are also added to enable the model to emphasize on clinically important areas in the image based on textual information. Attention-based fusion allows multimodal features to be weighted dynamically and is more interpretable [15], [16].

The concept of transformer-based fusion architectures is a further extension of this concept to the joint modeling of text tokens and image patches sequences. This allows these models to engage in deep cross-modal interactions by using self-attention and cross-attention layers, and they are specifically useful in complex diagnostic reasoning [17]. Nonetheless, lightweight or compressed variants of the transformer are better in low-resource environments because of their computational cost.

3.5. Domain Knowledge

Inclusion of the medical domain knowledge turns out to enhance model reliability particularly in situations where data of training is sketchy. Guidance could be provided to feature alignment and interpretation with the help of knowledge graphs and clinical ontologies.

Illustrative Example

Assuming that a clinical note is written and any word or phrase in it includes the word shadow, a medical knowledge graph can be used to match this text-based idea with a radiological artifact, like lung opacities in a chest X-ray. In fusion, the model is able to give priority to image areas with occurrence of opaqueness when handling the respective textual cue and hence is simulating clinical reasoning. This explicit association aids to decrease the ambiguity and to increase the interpretability.

3.6. Pseudocode of the Proposed Methodology

Pseudocode

```
Input: Medical Image I, Clinical Text T
Output: Diagnostic Prediction D

1: Preprocess I and T
2: Extract visual features FV = CNN(I)
3: Extract textual features FT = NLP_Encoder(T)

4: if Fusion_Type == Early:
```

```
    F = Concatenate(FV, FT)
    D = Final_Classifier(F)

elif Fusion_Type == Late:
    DV = Classifier(FV)
    DT = Classifier(FT)
    D = Combine(DV, DT)

elif Fusion_Type == Intermediate:
    F = Merge(FV, FT) at hidden layer
    D = Final_Classifier(F)

return D
```

3.7. Suitability for Low-Resource Settings

The suggested methodology focuses on:

- Transfer learning to cut down on the need of labeled data.
- Lightweight fusion structures to reduce the cost in computation.
- Attentional and domain knowledge interpretability.
- Resistance to dirty and inconsistent information.

These features render the framework usable in low-resource healthcare settings.

4. Evaluation

Strict testing is also needed to determine the reliability and clinical utility of multimodal text-image fusion models, which in low-resource care settings such as those of a healthcare facility, diagnostic errors may have dire outcomes. Besides the predictive performance, predictive performance needs to be measured in terms of model calibration, explainability, and clinical validation to make sure that it is safely and reliably deployed.

4.1 Diagnostic Performance Metrics

Diagnostic accuracy is measured using standard classification metrics such as accuracy, sensitivity (recall), specificity, precision and F1-score. In medical diagnosis, sensitivity is essential to minimize the cases of missed diseases; whereas specificity aims at reducing cases of false alarm. In the case of imbalanced data sets, the area under the receiver operating characteristic curve (AUC-ROC) and the area under the precision recall curve (AUC-PR) are both good measures of discriminative performance. These measures allow comparing favorably unimodal and multimodal fusion models and show the value of multimodal integration. Table 2 demonstrates general performance trends reported in these previous studies on multimodal medical AI [18-21]. Current literature also demonstrates that multimodal fusion models are superior to unimodal models (text only and image only) in terms of accuracy, sensitivity, specificity and F1-score because they combine in-complementary information in various data sources.

Table 2: Performance Comparison of Diagnostic Models

Model Type	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-Score
Text-only	71.0	68.5	73.2	70.1
Image-only	76.0	74.1	77.8	75.0
Late Fusion	82.0	80.3	83.5	81.1

A visual comparison of diagnostic accuracy across different literature models [18-21] is illustrated in Figure 2 below.

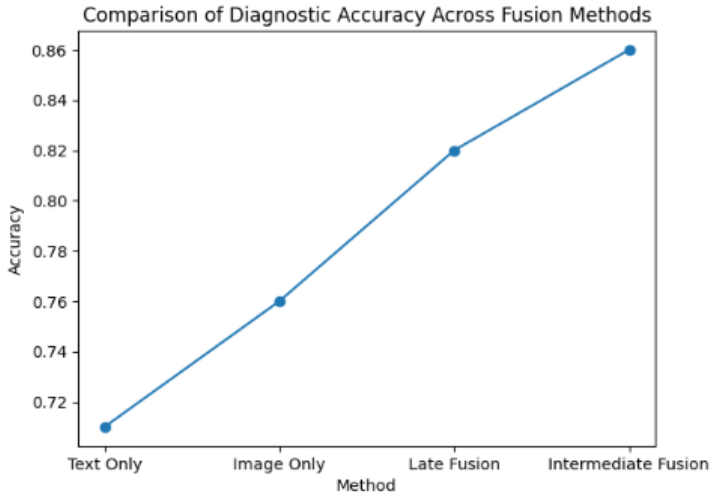

Figure 2: Comparison of diagnostic accurse fusion methods

The variation in sensitivity and specificity across different literature models [18-21] is shown in Figure 3.

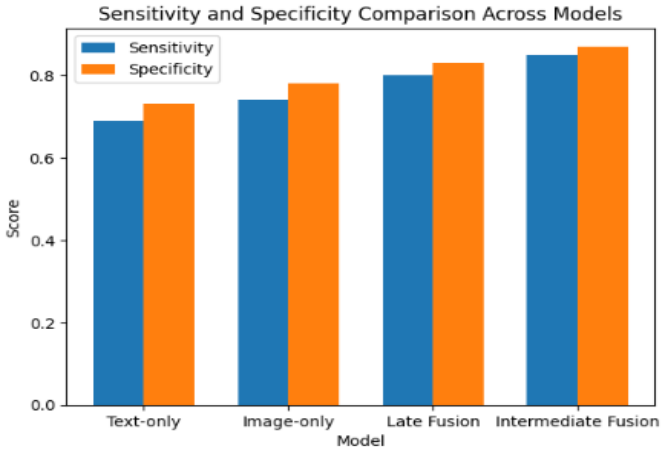

Figure 3: Sensitivity and Specificity Comparison

4.2 Calibration and Clinical Reliability

In addition to the discrimination performance, the calibration metrics plays an important role in the assessment of clinical reliability. A model with good calibration will have probability estimates that indicate

the true likelihood of outcome which is necessary in clinical decision making. The quality of calibration is measured using such metrics like the Brier Score and Expected Calibration Error (ECE). A smaller value is an indication of a higher compatibility between forecasted probabilities and actual results. The point of well-calibrated predictions in low-resource environments requires clinicians to heavily trust the AI suggestion which may result in overconfidence in false diagnosis. Table 3 condenses the characteristic trends of the performance of calibration on the findings of the earlier research on uncertainty estimation and the model reliability in the medical AI [18-21].

Table 3: Calibration Metrics

Model	Brier Score ↓	ECE ↓
Text-only	0.19	0.12
Image-only	0.16	0.10
Late Fusion	0.11	0.07
Intermediate Fusion	**0.08**	**0.05**

The calibration behavior of the multimodal model from literature [18-21], is illustrated in Figure 4.

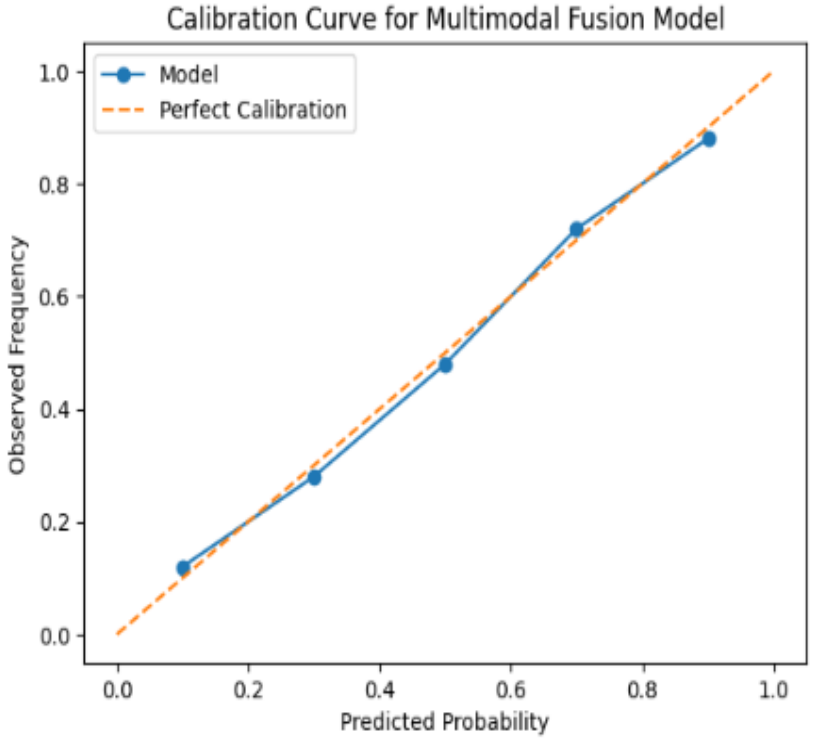

Figure 4: Calibration curve for the multimodal fusion model

4.3 Explainability and Model Interpretation

Explainability is one of the main factors used to assess multimodal diagnostic systems. Visual-text attention systems are an intuitive and concrete explainability method by pointing to the salient areas in medical images (e.g., areas of opaque in an X-ray) and significant textual tokens (e.g., shortness of breath or shadow). Attention heatmaps Visual and text Visual heatmaps allow clinicians to check that the decision made by the model is based on clinically relevant evidence of both modalities. It is these interpretability tools that enable qualitative assessment of the medical experts and enhance the trust of AI-assisted diagnosis.

4.4 Human-in-the-Loop Evaluation

In order to make it practical, human in the loop evaluation is added to the assessment procedure. Within this paradigm, clinicians look at the model predictions, attention maps and confidence estimate and give feedback on correctness and clinical relevance. This joint assessment will enable medical professionals to verify AI results, define lack of success cases, and improve system performance. The human-in-the-loop testing method is especially useful in low-resource environments, where clinical experience in the area can be used to offset a lack of training data and enhance system robustness through time.

5. Ethical, Practical, and Technical Considerations

The deployment of multimodal AI systems for medical diagnosis in low-resource settings raises important ethical, practical, and technical concerns that must be addressed to ensure safe, fair, and sustainable use.

5.1 Trustworthiness, Explainability, and Transparency

Medical AI can be trusted only in the close connection with explainability and transparency. The more the model decisions can be explained and justified, the more clinicians can trust and accept AI systems. Transparency techniques that utilize saliency, including attention heatmaps and feature importance visualization, can be used to understand which parts of the image are used to make a diagnosis and what parts of a text are used to make a diagnosis. Also, model uncertainty estimates convey a sense of confidence, which enables clinicians to identify the cases of uncertainty that need further investigation. Multimodal systems can inform clinical decision-making instead of throwing out expert judgment by providing explainability, coupled with being aware of the uncertainty.

5.2 Ethical Considerations and Bias Mitigation

Patient data privacy, informed consent and the possibility of biasness in training data are some ethical issues. Low-resource datasets can either focus too little on some population or pattern of disease, causing biased forecasts. Healthcare disparities should not be exacerbated with close data management, anonymization tools, and bias-sensitive assessment measures. The deployment of models ethically is to be monitored and validated on ongoing basis and in a variety of patients.

5.3 Practical Deployment Constraints

Practical issues are the lack of computational resources, untrusted connectivity and compatibility with the current clinical processes. The multimodal fusion models should be computationally lean and able to operate with the existing hardware. Interfaces should be user-friendly and should not disrupt workflow greatly in order to be adopted. It is also important to train healthcare personnel to be responsible in interpreting AI outputs especially where technical assistance might be insufficient. Table 4 qualitatively compares the computational and practical considerations based on literature available on the topic of multimodal fusion [22].

Table 4: Computational and Practical Considerations

Aspect	Text-only	Image-only	Late Fusion	Intermediate Fusion
Data Requirement	Low	Medium	Medium	High
Computational Cost	Low	Medium	Low	Medium
Explainability	High	Medium	High	Medium
Suitability for Low-Resource Settings	Moderate	Moderate	High	High

5.4 Technical Robustness and Sustainability

Technically, it should be robust to noisy, incomplete and misaligned data. Models have to be able to generalize to other clinics and imaging equipment and maintain their consistency in terms of performance. The process of sustainable deployment also needs systems of periodic updates, incorporation of clinician feedback, and accommodation of clinical changes. Resource-constrained environments can be maintained over the long-term using lightweight architectures and modular designs.

6. Discussion

Multimodal text-image fusion offers a promising future of enhancing medical diagnosis especially in low resource medical settings where clinical skills and data access is constrained. Multimodal systems are able to recapitulate the diagnostic reasoning in the real world by integrating complementary information in clinical text and medical images overcoming the natural limitations of unimodal systems. The examined fusion approaches reveal that intermediate and attention-based fusion designs are particularly effective to learn cross-modal relations and retain the ability to understand them.

Although such advantages exist, one should admit several limitations and challenges. A big struggle is the domain shift whereby models trained on data of one hospital or area cannot work well in a new area because of variations in imaging equipment, documentation pattern or patient demographics. This is of special concern to low-resource environments, where the heterogeneity of data is marked and the standardization is minimal. The domain shift needs to be addressed with strategies of adaptive learning and strong validation in many sites.

The data quality and alignment are other significant issues. Incomplete, noisy, and inconsistently recorded clinical text, and errors in image-report pairing, may impair the effectiveness of multimodal learning. The additional demands can be met by preprocessing, alignment, and fusion strategies to address these challenges. Moreover, the maintenance of the infrastructure is also a viable point. Even the lightweight AI systems need to have stable hardware, regular updates, and technical support, which might not be very easy to maintain in the resource-constrained settings.

Lastly, although the explainable AI mechanisms like attention heat maps can improve clinician confidence, they do not comprehensively ensure clinical accuracy. This will pose risks in case of over-dependence on the output of AI without sufficient human supervision. Thus, the use of multimodal systems must be introduced as a tool of support and not to substitute clinical skills. The awareness of these shortcomings is crucial to responsible deployment and is accompanied by the necessity to conduct further research on powerful, adaptable, and clinical-based multimodal diagnostic frameworks.

7. Conclusion

This paper discussed multimodal text-image fusion methods in medical diagnostics in terms of low-resource health facilities. Through the analysis of the literature and the fusion strategies, data scarcity reduction approaches, methodologies and ethical concerns, the paper introduces the high potential of multimodal AI as a way of improving diagnostic accuracy and access. Clinical text and medical image fusion facilitates a contextual and holistic diagnostic support as opposed to the unimodal systems.

Though, issues influencing data scarcity, modality mismatch, domain shift, and hardware capacity are still major obstacles to the practical use. The solutions that can be used to address such problems include data-efficient, interpretable, and computationally lightweight fusion architectures that do not conflict with clinical workflows or ethical standards.

Future research might consider lightweight multimodal transformers, federated fusion models, and adaptive domain generalization methods to make further gains on the scalability, privacy guarantees and resiliency of low-resource healthcare settings.

Funding Statement: Author has received no financial support for this paper.

Conflicts of Interest: NIL.

Data Availability: No dataset was used in this study; the work is based on a conceptual framework and literature analysis.

References

[1] Bray, O. H. (1997). Information integration for data fusion. Office of Scientific and Technical Information (OSTI). https://doi.org/10.2172/444047

[2] Domingues, I., Muller, H., Ortiz, A., Dasarathy, B. V., Abreu, P. H., & Calhoun, V. D. (2020). Guest Editorial: Information Fusion for Medical Data: Early, Late, and Deep Fusion Methods for Multimodal Data. In IEEE Journal of Biomedical and Health Informatics (Vol. 24, Issue 1, pp. 14–16). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/jbhi.2019.2958429

[3] Huang, W.-Y., & Davis, J. J. (2011). Multimodality and nanoparticles in medical imaging. In Dalton Transactions (Vol. 40, Issue 23, p. 6087). Royal Society of Chemistry (RSC). https://doi.org/10.1039/c0dt01656j

[4] Mu, S., Cui, M., & Huang, X. (2020). Multimodal Data Fusion in Learning Analytics: A Systematic Review. In Sensors (Vol. 20, Issue 23, p. 6856). MDPI AG. https://doi.org/10.3390/s20236856

[5] Zhou, T., Thung, K., Zhu, X., & Shen, D. (2018). Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis. In Human Brain Mapping (Vol. 40, Issue 3, pp. 1001–1016). Wiley. https://doi.org/10.1002/hbm.24428

[6] Gusarova, N., Lobantsev, A., Vatian, A., Kapitonov, A., & Shalyto, A. (2020). Comparative assessment of text-image fusion models for medical diagnostics. In Information and Control Systems (Issue 5, pp. 70–79). State University of Aerospace Instrumentation (SUAI). https://doi.org/10.31799/1684-8853-2020-5-70-79

[7] Akter, M., & Kudapa, S. P. (2024). A comparative analysis of artificial intelligence-integrated bi dashboards for real-time decision support in operations. *International Journal of Scientific Interdisciplinary Research*, *5*(2), 158-191.

[8] Muzammil, S. R., Maqsood, S., Haider, S., & Damaševičius, R. (2020). CSID: A Novel Multimodal Image Fusion Algorithm for Enhanced Clinical Diagnosis. In Diagnostics (Vol. 10, Issue 11, p. 904). MDPI AG. https://doi.org/10.3390/diagnostics10110904

[9] Qi, G., Wang, J., Zhang, Q., Zeng, F., & Zhu, Z. (2017). An Integrated Dictionary-Learning Entropy-Based Medical Image Fusion Framework. In Future Internet (Vol. 9, Issue 4, p. 61). MDPI AG. https://doi.org/10.3390/fi9040061

[10] Tang, L., Qian, J., Li, L., Hu, J., & Wu, X. (2017). Multimodal medical image fusion based on discrete Tchebichef moments and pulse coupled neural network. In International Journal of Imaging Systems and Technology (Vol. 27, Issue 1, pp. 57–65). Wiley. https://doi.org/10.1002/ima.22210

[11] Johnson, A. E. W., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C., Mark, R. G., & Horng, S. (2019). MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data, 6(1). https://doi.org/10.1038/s41597-019-0322-0

[12] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011, June). Multimodal deep learning. In *Icml* (Vol. 11, pp. 689-696).

[13] Boulahia, S. Y., Amamra, A., Madi, M. R., & Daikh, S. (2021). Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. Machine Vision and Applications, 32(6). https://doi.org/10.1007/s00138-021-01249-8

[14] Guarrasi, V., Aksu, F., Caruso, C. M., Di Feola, F., Rofena, A., Ruffini, F., & Soda, P. (2024). A Systematic Review of Intermediate Fusion in Multimodal Deep Learning for Biomedical Applications. https://doi.org/10.2139/ssrn.4952813

[15] Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, *32*. https://proceedings.neurips.cc/paper_files/paper/2019/file/c74d97b01eae257e44aa9d5bade97baf-Paper.pdf

[16] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6077–6086. https://doi.org/10.1109/cvpr.2018.00636

[17] Cai, Y., & Rostami, M. (2024). Dynamic Transformer Architecture for Continual Learning of Multimodal Tasks. https://doi.org/10.2139/ssrn.4713352
[18] Huang, B., Yang, F., Yin, M., Mo, X., & Zhong, C. (2020). A Review of Multimodal Medical Image Fusion Techniques. Computational and Mathematical Methods in Medicine, 2020, 1–16. https://doi.org/10.1155/2020/8279342
[19] Kumar, S., Rani, S., Sharma, S., & Min, H. (2024). Multimodality Fusion Aspects of Medical Diagnosis: A Comprehensive Review. Bioengineering, 11(12), 1233. https://doi.org/10.3390/bioengineering11121233
[20] Ullah Khan, S., Ahmad Khan, M., Azhar, M., Khan, F., Lee, Y., & Javed, M. (2023). Multimodal medical image fusion towards future research: A review. Journal of King Saud University - Computer and Information Sciences, 35(8), 101733. https://doi.org/10.1016/j.jksuci.2023.101733
[21] Gu, X., Xia, Y., & Zhang, J. (2024). Multimodal medical image fusion based on interval gradients and convolutional neural networks. BMC Medical Imaging, 24(1). https://doi.org/10.1186/s12880-024-01418-x
[22] Baltrusaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal Machine Learning: A Survey and Taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443. https://doi.org/10.1109/tpami.2018.2798607