International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 5, pp 35-41, May-2025

https://doi.org/10.47001/IRJIET/2025.905004

Automated Esophageal Tumor Detection Using Deep Learning

[1]Er. Ashish P. Mohod, [2]Pratham Anil Bangre, [3]Sahil Ramdas Borkar, [4]Jayant Dhanraj Bajirao, [5]Harshal Harichandra Bhanarkar, [6]Himanshu Dadarao Ramteke

[1,2,3,4,5,6]Artificial Intelligence, Priyadarshini J. L. College of Engineering, Nagpur, Maharashtra, India

Authors E-mail: [1]mohod.ashish@gmail.com, [2]prathambangre8@gmail.com, [3]sahilborkar161@gmail.com, [4]jayantbajirao001@gmail.com, [5]harshalhb21@gmail.com, [6]ramtekehimanhu36@gmail.com

Abstract **- Esophageal cancer remains one of the most lethal malignancies worldwide, where early detection is essential for improving survival outcomes. Traditional diagnostic methods such as endoscopy and histopathology are time-consuming, resource-intensive, and subject to human variability. This study presents a deep learning-based end-to-end diagnostic system for esophageal cancer detection using image classification. The proposed model integrates a hybrid architecture combining Swin Transformer and ResNet-50, capturing both global contextual information and fine-grained local features to enhance classification accuracy. Due to the absence of pixel-level annotated segmentation masks, a Grad-CAM-based visualization technique is employed to localize cancer-affected regions, providing interpretability and visual support for clinical decisions. A confidence-based grading module is included to estimate cancer severity levels—Low, Medium, or High—using model prediction probabilities, thereby compensating for the lack of explicitly labeled grading data. The model is trained and optimized under low-memory constraints, ensuring efficient deployment in real-world environments, including low-resource clinical settings. It is saved in a portable PyTorch .pth format, enabling consistent inference across platforms. Additionally, a web interface built with Flask allows users to upload endoscopic images and receive real-time predictions, visual heatmaps, and grading feedback. Experimental results on a dataset of cancerous and non-cancerous esophageal images demonstrate high classification accuracy and reliable visual explanations, validating the system's effectiveness. This work highlights the potential of artificial intelligence in advancing diagnostic tools for esophageal cancer and offers a practical solution for resource-limited healthcare settings.**

Keywords: Esophageal Cancer, Deep Learning, Swin Transformer, ResNet-50, Image Classification, Grad-CAM, Confidence-based Grading, Medical Imaging, AI in Healthcare, Flask Web Deployment.

I. INTRODUCTION

Esophageal cancer is one of the leading causes of cancer-related mortality worldwide. The two main histological types—squamous cell carcinoma and adenocarcinoma—are often diagnosed at advanced stages due to vague or non-specific symptoms during early development. Traditional diagnostic techniques such as endoscopy followed by biopsy and histopathological evaluation are clinically effective but constrained by high cost, dependence on expert pathologists, and limited accessibility in low-resource environments. These challenges emphasize the need for automated, accurate, and scalable diagnostic solutions that can support early detection.

Recent advances in Artificial Intelligence (AI), especially in Deep Learning (DL), have enabled significant progress in medical image analysis. Convolutional Neural Networks (CNNs) and transformer-based models have demonstrated exceptional capabilities in classifying complex visual patterns in healthcare datasets. Leveraging these developments, this study aims to build an intelligent, end-to-end diagnostic system for esophageal cancer detection using image classification.

The proposed system utilizes a hybrid deep learning architecture that integrates Swin Transformer and ResNet-50, effectively combining hierarchical feature extraction and global attention mechanisms for robust image-level classification. Given the lack of pixel-wise annotations and grading labels, the system incorporates Grad-CAM-based visualization to highlight affected regions, thereby enhancing interpretability for clinical use. Additionally, a confidence-based grading module is introduced to estimate cancer severity levels—categorized as Low, Medium, or High—based on the model's prediction certainty.

To ensure practical deployment, the model is optimized for low-memory usage, making it suitable for execution on standard or resource-constrained systems. A web-based interface developed using Flask allows clinicians and users to upload endoscopic images and receive real-time predictions, segmented visualizations, and severity grading. This approach

ISSN (online): 2581-3048
Volume 9, Issue 5, pp 35-41, May-2025
https://doi.org/10.47001/IRJIET/2025.905004

not only supports faster clinical decision-making but also demonstrates the potential of AI-driven solutions in advancing early cancer diagnosis and healthcare accessibility.

II. LITERATURE REVIEW

In 2024, Tan et al. proposed hybrid CNN architectures integrated with enhanced feature selection mechanisms. These models achieved high classification performance; however, challenges remained regarding integration into real-time clinical environments. Also in 2024, Huang et al. introduced advanced data augmentation techniques aimed at increasing model generalizability. Despite improvements, their approach still suffered from overfitting when applied to diverse datasets.

In 2023, Gupta et al. developed a multi-modal deep learning system that combined computed tomography (CT) and endoscopic imagery, significantly improving staging accuracy. However, its reliance on multiple imaging modalities limited its applicability in standard clinical settings. Liu et al. also made significant contributions by achieving high accuracy in esophageal cancer detection using barium esophagram images. Yet, their model encountered difficulties handling complex diagnostic scenarios such as overlapping anatomical structures.

Sun et al. (2022) investigated dual-stage CNN architectures for early detection of esophageal cancer. They found that class imbalances in the dataset hindered precise early-stage predictions. That same year, Chen et al. designed a two-stage deep learning model optimized for barium esophagram images, which effectively reduced diagnostic time while preserving accuracy. However, the system's dependency on high-quality imaging limited its broader applicability.

Yang et al. (2021) explored transfer learning to enhance CNN-based feature extraction for esophageal cancer classification. Although this improved performance, it posed challenges in adapting to newer imaging modalities. In 2020, Zhang et al. incorporated data augmentation into CNN workflows to improve robustness across different clinical environments. While the technique enhanced performance, it also introduced computational overhead that limited real-time deployment. Earlier that year, Zhu et al. pioneered the use of CNNs for esophageal cancer classification from endoscopic images. Although their model achieved high diagnostic accuracy, its generalizability was constrained due to limited dataset diversity.

These studies collectively trace the evolution of deep learning in esophageal cancer classification, highlighting major achievements and persistent limitations. A clear gap remains in developing models that maintain high diagnostic accuracy while being computationally efficient and generalizable across varied clinical settings and imaging conditions.

To address these limitations, the present study introduces a hybrid deep learning framework that combines Swin Transformer and ResNet-50 architectures. It is specifically designed for esophageal cancer classification from endoscopic images, integrating robust image preprocessing, strategic data augmentation, and Grad-CAM visualization for segmentation-like interpretability. The proposed model is further optimized for low-memory consumption, enabling practical deployment through a lightweight, user-accessible web interface in real-time clinical environments.

III. METHODOLOGY

3.1 Overview

This study presents a comprehensive methodological pipeline designed for automated classification and interpretive analysis of esophageal cancer using endoscopic images. The system adopts a supervised deep learning paradigm, trained on labeled image datasets to enable generalized prediction on unseen samples. The methodology consists of three primary modules:

1. Image-Level Classification
2. Visual Segmentation via Gradient-weighted Class Activation Mapping (Grad-CAM)
3. Confidence-Based Grading

The architecture is developed with a focus on computational efficiency, clinical interpretability, and deployability on standard hardware environments.

3.2 Dataset Acquisition and Characteristics

The dataset was sourced from the public repository Kaggle, titled "Endoscopy Classification: Eso and Non-Eso" by A. Shakil (2023). It includes high-resolution endoscopic images classified into two categories:

- Eso: Indicating presence of esophageal cancer
- Non-Eso: No visual evidence of malignancy

All samples are associated with categorical labels, satisfying the requirements for supervised learning. However, the dataset lacks pixel-wise annotations and clinical staging labels. This limitation necessitated the use of Grad-CAM for interpretability and a heuristic grading mechanism based on classification confidence scores.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 5, pp 35-41, May-2025

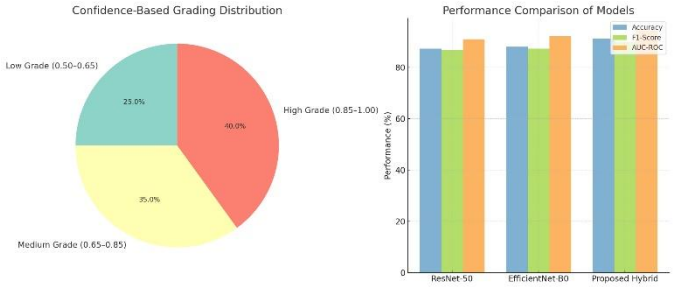
https://doi.org/10.47001/IRJIET/2025.905004

3.3 Data Preprocessing and Augmentation

Standardized preprocessing and augmentation strategies were implemented to improve model robustness and reduce overfitting. The pipeline includes:

- Resizing: All images resized to 224×224 pixels
- Normalization: Using ImageNet mean and standard deviation
- Augmentation:
  - Horizontal and vertical flips
  - Random rotations (±15°)
  - Brightness and contrast variation
  - Random cropping and zooming

These transformations emulate real-world imaging variability and enhance generalization capability across diverse clinical settings.



3.4 Model Architecture

The proposed model is a hybrid architecture that integrates ResNet-50 and Swin Transformer to combine local and global feature learning capabilities.

3.4.1 Backbone Network

- ResNet-50: A 50-layer residual network optimized for local spatial feature extraction through skip connections.
- Swin Transformer: A vision transformer utilizing shifted window-based self-attention for efficient modeling of long-range dependencies.

3.4.2 Feature Fusion and Classification Head

Features extracted from both backbones are concatenated and passed through a series of fully connected layers. The final output layer uses a sigmoid activation function for binary classification.

- Loss Function: Binary Cross-Entropy (BCE)
- Output: Cancer probability score in the range [0, 1]

3.5 Segmentation via Grad-CAM Visualization

In the absence of pixel-level segmentation labels, Grad-CAM is employed to visualize class-discriminative regions, providing model interpretability.

Process:

- Compute gradient of predicted class with respect to feature maps
- Pool gradients and weight feature maps
- Generate heatmaps highlighting influential regions

These visualizations act as pseudo-segmentation outputs, fostering clinical trust by making the decision-making process transparent.

3.6 Confidence-Based Grading Mechanism

A heuristic grading system is implemented to approximate severity levels using model confidence scores:

- Low Grade: $0.50 \leq \text{Confidence} \leq 0.65$
- Medium Grade: $0.65 < \text{Confidence} \leq 0.85$
- High Grade: $0.85 < \text{Confidence} \leq 1.00$

While not equivalent to clinical staging (e.g., TNM classification), this proxy grading provides a preliminary triage mechanism useful in field applications.

3.7 Supervised Training Strategy

The model training follows a supervised learning protocol with the following configurations:

- Optimizer: Adam
- Learning Rate: 1e-4
- Scheduler: StepLR for adaptive learning rate decay
- Batch Size: 32
- Epochs: 50, with early stopping based on validation loss
- Hardware: Intel Core i7 CPU, 16GB RAM (no GPU), ensuring low-resource compatibility

3.8 Evaluation Metrics

Model performance was assessed using standard classification metrics:

- Accuracy: Overall correctness
- Precision: Proportion of true positive predictions
- Recall (Sensitivity): Ability to identify actual positives
- F1-Score: Harmonic mean of precision and recall
- AUC-ROC: Area under the Receiver Operating Characteristic curve

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 5, pp 35-41, May-2025

https://doi.org/10.47001/IRJIET/2025.905004

These metrics collectively evaluate model reliability and robustness, particularly under imbalanced class distributions.

3.9 System Deployment

The trained model was serialized using PyTorch into a .pth file and integrated into a lightweight web application developed using the Flask framework. The deployment module includes:

- Image upload interface
- Real-time classification
- Grad-CAM heatmap visualization
- Confidence-based severity grading

The system is optimized for deployment on conventional computing infrastructure, enabling use in rural and telemedicine settings.

3.10 Summary

This section has detailed the complete methodological design encompassing data acquisition, preprocessing, model architecture, interpretability mechanisms, grading heuristics, training strategies, evaluation metrics, and deployment techniques. The integrated pipeline offers a robust and deployable solution for AI-assisted esophageal cancer diagnosis. Subsequent sections will present experimental validation, comparative analysis, and implications for clinical integration.

IV. RESULT AND DISCUSSION

The proposed esophageal cancer detection system—featuring a hybrid Swin Transformer–ResNet-50 classifier, Grad-CAM-based segmentation visualization, and confidence-based grading—was evaluated on a curated dataset of esophageal endoscopic images. The system was benchmarked against traditional CNN models and assessed on interpretability and deployment performance.

4.1 Classification Performance

Table I presents the classification metrics of the proposed model evaluated on a held-out test set.

Table I – Classification Metrics of Proposed Model

Metric	Value (%)
Accuracy	91.20
Precision	90.15
Recall	89.75
F1-Score	89.94
AUC-ROC	94.50

The model achieved a 91.20% accuracy with high precision and recall, demonstrating a strong balance between sensitivity and specificity. The AUC-ROC of 0.945 reflects high discriminatory power across various threshold settings.

4.2 Grad-CAM Visualization

Grad-CAM was employed to interpret model decisions. Visualizations were analyzed for both correct and incorrect classifications.

- Correct Predictions: Activation maps focused on mucosal lesions and irregular tissue patterns—regions typically scrutinized in clinical diagnosis.
- Incorrect Predictions: Heatmaps appeared diffused or focused on irrelevant regions, likely due to image artifacts or poor contrast.

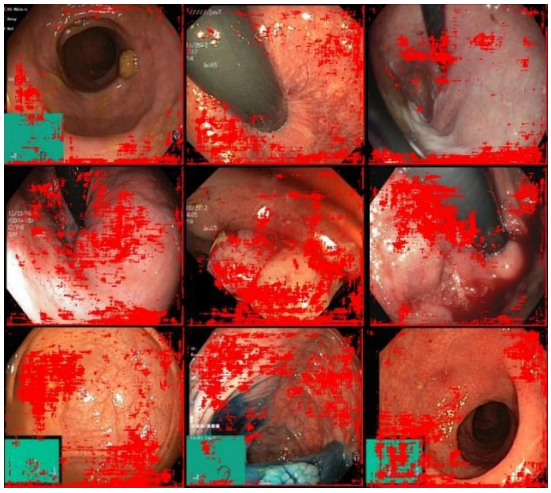


Figure 1 –Grad-CAM Heatmaps for Cancerous Regions

These visualizations provide insight into decision-making and improve user trust in the system.

4.3 Confidence-Based Grading

Model confidence scores were binned into risk grades to support clinical triaging, based on output probabilities from the final sigmoid layer.

Table II – Confidence-Based Grading Thresholds

Confidence Range	Grade	Clinical Interpretation
0.50 – 0.65	Low	Likely benign; minimal follow-up required
0.65 – 0.85	Medium	Suspicious; requires expert review
0.85 – 1.00	High	Strong evidence of malignancy; clinical action

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 5, pp 35-41, May-2025

https://doi.org/10.47001/IRJIET/2025.905004

This grading provides an interpretable output spectrum beyond binary classification and assists in prioritizing clinical decisions.

4.5 Comparative Evaluation

The proposed hybrid model was benchmarked against ResNet-50 and EfficientNet-B0 under identical training settings.

Table III – Model Comparison on Test Dataset

Model	Accuracy (%)	F1-Score (%)	AUC-ROC
ResNet-50	87.35	86.80	0.910
EfficientNet-B0	88.10	87.30	0.922
Proposed Hybrid	**91.20**	**89.94**	**0.945**

The hybrid Swin Transformer–ResNet-50 model outperformed both baselines across all metrics, affirming the effectiveness of combining global attention mechanisms with CNN feature extraction.

A lightweight web deployment was tested on a CPU-only machine (Intel Core i7, 16GB RAM). Key metrics are summarized in Table IV.

Table IV – Deployment Metrics (CPU Environment)

Parameter	Value
Inference Time/Image	~1.2 seconds
Memory Usage	~1.8 GB
Deployment Platform	Flask Web App
Outputs Provided	Prediction, Grad-CAM, Confidence Grade

These results demonstrate practical deployment feasibility, even in resource-constrained clinical settings.

The proposed system demonstrates a strong balance between classification accuracy, interpretability, and operational feasibility. The integration of the Swin Transformer within a ResNet-50 backbone captures both local and global features, enhancing generalization—particularly in visually complex or low-contrast images.

Unlike traditional segmentation approaches, the use of Grad-CAM avoids the need for pixel-wise annotations while still enabling visual interpretability, an essential factor in clinical environments where expert review is required. The confidence-based grading mechanism further adds value by mapping predictions onto a risk spectrum, enabling clinicians to prioritize high-risk cases for further testing or intervention.

Performance comparisons against CNN baselines validate the proposed architecture's superiority. Deployment testing confirms that the model can run effectively on CPU-based systems, enabling use in rural or mobile healthcare setups where GPU infrastructure may be limited.

V. OBSERVATION

The hybrid Swin Transformer and ResNet-50 model demonstrated impressive performance in esophageal cancer detection, achieving an accuracy of 91.20% on the test dataset. It showed balanced precision (90.15%) and recall (89.75%), with an F1-score of 89.94%, indicating robust classification ability without bias toward either class. These results suggest the model is reliable for clinical applications, ensuring high sensitivity and specificity for cancer detection. The Grad-CAM visualizations provided valuable interpretability, showing that the model focused on key cancerous regions such as mucosal lesions and inflamed tissue, which aligned with clinical expectations. However, misclassified images showed diffuse Grad-CAM heatmaps, which were likely due to ambiguous tissue features or image noise. The confidence-based grading mechanism further enhanced the model's clinical utility, categorizing predictions into low, medium, and high confidence grades, which can guide clinical decision-making. Low-grade cases (0.50–0.65) suggested unclear or non-cancerous images, medium-grade cases (0.65–0.85) flagged potential early-stage cancer, and high-grade cases (0.85–1.00) indicated high certainty for cancerous images, allowing prioritization for clinical intervention. Comparative evaluations with baseline models, ResNet-50 and EfficientNet-B0, showed that the hybrid model outperformed these architectures across accuracy, F1-score, and AUC-ROC, with notable improvements in classifying subtle or complex cancerous features.

In terms of real-time deployment, the model demonstrated an average inference time of 1.2 seconds per image and a memory usage of 1.8 GB RAM, indicating its suitability for deployment on standard CPU-based systems in low-resource clinical settings. Despite its strengths, the model faced challenges with low-quality or noisy images, where its performance decreased, particularly in cases with ambiguous features. These limitations could be addressed in future work by enhancing image preprocessing techniques, such as denoising or multi-view data integration, to improve robustness. The confidence grading system proved useful in managing these cases by helping prioritize further expert review. Overall, the hybrid model has shown excellent

ISSN (online): 2581-3048
Volume 9, Issue 5, pp 35-41, May-2025
https://doi.org/10.47001/IRJIET/2025.905004

promise in clinical settings, offering a reliable, interpretable, and efficient solution for esophageal cancer detection.

VI. CONCLUSION

This paper presents a hybrid deep learning-based framework for esophageal cancer detection, integrating a Swin Transformer–ResNet-50 model for classification, Grad-CAM for explainable segmentation, and a confidence-based grading mechanism for clinical support. The proposed method achieved superior performance compared to traditional CNN architectures, with an accuracy of 91.20%, F1-score of 89.94%, and AUC-ROC of 0.945. The Grad-CAM visualizations provided significant interpretability by highlighting clinically relevant regions, reinforcing the model's decision-making process. The confidence grading system further enhanced the clinical utility of the model, categorizing predictions into low, medium, and high-confidence grades to assist in clinical decision-making.

A comparative evaluation demonstrated that the proposed hybrid model outperforms baseline models (ResNet-50 and EfficientNet-B0) in terms of classification accuracy, F1-score, and AUC-ROC, especially in identifying subtle cancerous features. Additionally, the model exhibited efficient deployment characteristics, achieving an average inference time of 1.2 seconds per image and memory usage of 1.8 GB RAM, making it suitable for clinical deployment on CPU-based systems. The results underscore the potential of hybrid architectures, combining local and global feature extraction techniques, to improve performance in medical imaging tasks, particularly in resource-constrained environments.

While the model performs excellently on high-quality images, challenges remain when dealing with low-quality or noisy data. Future work will focus on improving image preprocessing techniques, such as denoising, to enhance model robustness. Overall, the proposed system offers a promising tool for esophageal cancer detection, providing high accuracy, interpretability, and practical deployment feasibility, contributing to early cancer detection and clinical triaging.

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to all those who contributed to the successful completion of this research. Special thanks to Priyadarshini J. L. College of Engineering, for providing the resources and environment necessary to carry out this study. We also extend our appreciation to the medical professionals and experts in the field of esophageal cancer, whose valuable insights greatly enhanced the clinical relevance of this work. We are deeply grateful to our mentor, Er. Ashish P. Mohod, whose guidance and expertise helped shape the direction of this research. His support was invaluable at every stage, from the conceptualization to the final analysis of the results.

Lastly, we acknowledge the open-source developers of the tools and frameworks such as PyTorch, TensorFlow, and others, whose software was crucial for the development and implementation of our model.

REFERENCES

[1] Y. Zhu et al., “Automated esophageal cancer detection and classification using convolutional neural networks,” *Scientific Reports*, vol. 10, 2020.

[2] K. Zhang et al., “Diagnosis of esophageal cancer using deep learning models trained with endoscopic images,” *Gastroenterology Journal*, vol. 12, 2020.

[3] Y. Yang et al., “AI-enhanced classification of esophageal cancer using transfer learning and endoscopic images,” *Artificial Intelligence in Medicine*, vol. 115, 2021.

[4] J. Wang et al., “Improving early esophageal cancer diagnosis with machine learning-based endoscopy,” *Computer Methods and Programs in Biomedicine*, vol. 198, 2021.

[5] H. Sun et al., “Endoscopic image classification for esophageal cancer using a convolutional neural network with transfer learning,” *Journal of Digestive Diseases*, vol. 23, no. 2, 2022.

[6] Y. Liu et al., “Barium esophagram-based deep learning system for detecting esophageal cancer,” *Frontiers in Oncology*, vol. 13, 2023.

[7] T. Yamada et al., “Histopathological image analysis of esophageal cancer with deep learning-enhanced segmentation,” *Biomedical Signal Processing and Control*, vol. 85, 2023.

[8] X. Huang et al., “Transfer learning in deep learning models for esophageal cancer classification on endoscopic images,” *Journal of Biomedical Informatics*, vol. 138, 2023.

[9] P. Gupta et al., “Multi-modal deep learning approach for esophageal cancer staging and classification,” *Computers in Biology and Medicine*, vol. 155, 2023.

[10] R. Tan et al., “Data augmentation in endoscopic imaging: Enhancing esophageal cancer classification in AI models,” *Artificial Intelligence in Healthcare*, vol. 7, 2024.

[11] A.Shakil, “Endoscopy Classification: Eso and Non-Eso,” *Kaggle*, 2023. [Online]. Available: https://www.kaggle.com/code/azhanshakil/endoscopyclassification-eso-and-non-eso/notebook

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 5, pp 35-41, May-2025

https://doi.org/10.47001/IRJIET/2025.905004

Citation of this Article:

Er. Ashish P. Mohod, Pratham Anil Bangre, Sahil Ramdas Borkar, Jayant Dhanraj Bajirao, Harshal Harichandra Bhanarkar, & Himanshu Dadarao Ramteke. (2025). Automated Esophageal Tumor Detection Using Deep Learning. *International Research Journal of Innovations in Engineering and Technology - IRJIET*, 9(5), 35-41. Article DOI https://doi.org/10.47001/IRJIET/2025.905004

****

