Journal of Algorithms and Computation

journal homepage: http://jac.ut.ac.ir

Proposing a novel deep method for detection and localization of anatomical landmarks from the endoscopic video frames

S. Ayyoubi Nezhad[*1], G. Tajeddin[†1], T. Khatibi [‡1], M. Sohrabi [§2]

[1] School of Industrial and Systems Engineering, Tarbiat Modares University (TMU), Tehran, Iran

[2] Gastrointestinal and liver diseases research center, Iran University of Medical Sciences (IUMS), Tehran, Iran

ABSTRACT

Early detection of gastrointestinal cancer remains a major challenge, particularly in identifying cancerous regions at their initial stages. Anatomical landmarks are crucial for guiding physicians during endoscopic screenings, with accurate localization enhancing diagnostic precision. This study proposes a deep learning approach using convolutional neural networks (CNNs) to detect and localize anatomical landmarks in endoscopic video frames from 40 patients at Firoozgar Hospital, Tehran. Pre-processed frames were annotated with bounding boxes to highlight regions of interest. The CNN model achieved 97.0% accuracy for landmark detection and classification and an MSE of 0.004 for bounding box regression, showing promise for assisting early diagnosis.

Keywords: Machine learning, Computer vision, Object Detection, Medical Image Analysis, Symptoms localization.

ARTICLE INFO

Article history:
Research paper
Received 09, November 2024
Accepted 05, December 2024
Available online 20, December 2024

AMS subject classification: 92F08.

[*] shima.ayyoubinezhad@modares.ac.ir
[†] Corresponding author: G. Tajeddin, Email: g.tajeddin@modares.ac.ir .
[‡] Toktam.khatibi@modares.ac.ir
[§] sohrab_r@yahoo.com

1 Introduction

According to the World Health Organization (WHO) *Progress Monitor 2022*, more than 75% of the 20.4 million premature deaths that occurred between the ages of 30 and 70 in 2019 were caused by non-communicable diseases [1]. Of every ten premature deaths attributed to non-communicable diseases, four result from cardiovascular diseases, and three are due to cancer. Cancer ranks as the first or second leading cause of death before age 70 in 112 of 183 countries and the third or fourth cause in another 23 countries [1].

In 2020, 19.3 million new cancer cases and 10 million cancer-related deaths were reported globally. Notably, 50% of new cases and 58.3% of deaths occurred in women and men in Asian countries, which host approximately 59.5% of the world's population. Unlike other regions, Asia (58.3%) and Africa (7.2%) recorded higher cancer-related deaths than new cases [2].

Early cancer diagnosis significantly improves survival rates by enabling timely and appropriate medical intervention [3]. To address the challenges associated with early detection, computer-aided diagnostic (CAD) systems have been developed [4]. These systems utilize artificial intelligence (AI) and deep learning to analyze imaging data without human intervention, thereby supporting physicians in their decision-making [4-6]. CAD systems employ a variety of algorithms and techniques to tackle tasks such as segmentation [7], classification [8], localization [9], lesion detection [10], and surgical instrument tracking [11]. These methods aim to overcome the challenges of gastrointestinal disease detection [12].

Several advanced techniques have been proposed in this domain. Zhang et al. introduced convolutional neural network (CNN)-based object detection methods, including Faster Recurrent Convolutional Neural Networks (RCNN) and single-shot multi-box detectors (SSD), to identify seven classes of endoscopic images from the EDA2019 challenge dataset. They compared the performance of these models to demonstrate their capabilities [13, 14].

Another study proposed a Mask RCNN model for detecting and segmenting gastric cancer lesions in endoscopic video frames. The method achieved a sensitivity of 96.0% per image and an average Dice score of 71.0% for gastric cancer region segmentation. Their approach used a CNN to extract feature maps, followed by a region proposal network to identify regions of interest (ROI). The bounding box and lesion probability were derived from a fully connected layer, while the mask branch segmented lesions within the bounding box [7].

Caroppo et al. developed an unsupervised deep learning model for registering ROIs in wireless capsule endoscopy, designing a novel loss function to optimize model convergence [15]. Another study trained a series of deep learning models on 1,300 colonoscopy images to segment polyps [16].

Hoang et al. proposed combining a Residual Neural Network with Faster RCNN for symptom localization in endoscopic images. They introduced a novel data augmentation method to enhance their results, using Faster RCNN with ResNet-50 to generate bounding boxes around polyps [17]. Similarly, Hong et al. designed an ensemble learning model that used Mask RCNN for both polyp detection and segmentation in endoscopic images [18].

In another study, researchers introduced a plug-in module that concurrently detects and tracks polyps by combining strategies to extract spatiotemporal information for enhanced learning [19]. A modified Mask RCNN model was also proposed to classify gastrointestinal diseases and segment ulcer regions. This approach employed a pre-trained ResNet-101 model for feature extraction [20].

The effectiveness of deep convolutional neural networks (DNNs) in fields such as image analysis [15], video processing [21], and graph analysis [22] has inspired their widespread adoption. Their flexibility in architecture design, end-to-end feature extraction, and ability to achieve specific outcomes make DNNs the preferred choice for addressing complex challenges.

In this study, we harnessed the potential of deep learning methods to detect and localize anatomical landmarks of the upper gastrointestinal tract in endoscopic video frames, aiming to assist physicians during endoscopic procedures. The anatomical landmarks examined include the Z-line, esophagus, pylorus, and antrum.

This paper is structured as follows: Section 2 describes the dataset preparation process and the methodology used in this study. Section 3 outlines the analysis of performance metrics and evaluates the proposed method’s effectiveness. Section 4 concludes the paper and suggests directions for future research.

2 Methodology

In the following subsections, the main steps of the proposed method are explained.

2.1 Ethics statement and dataset

In terms of medical ethics, this study adheres to the principles outlined in the Declaration of Helsinki. Physicians provided the participating patients with detailed explanations about the study's objectives and procedures, ensuring informed consent was obtained. Patients experiencing stomach pain and referred to the endoscopy department of Firoozgar Hospital were recruited for the study. Endoscopic videos were collected from 40 patients.

2.1.1 Data Description

The endoscopic videos were recorded at a frame rate of 30 frames per second (fps). Images were extracted from these videos using the Python OpenCV library. Since the lengths of the endoscopic videos varied, the number of frames extracted from each video also differed. Following the extraction process, only frames containing anatomical landmarks—specifically the esophagus, Z-line, antrum, and pylorus—were selected for further analysis.

Each image had a resolution of 576 × 768 pixels. The distribution of images across the different classes is presented in Figure 1.

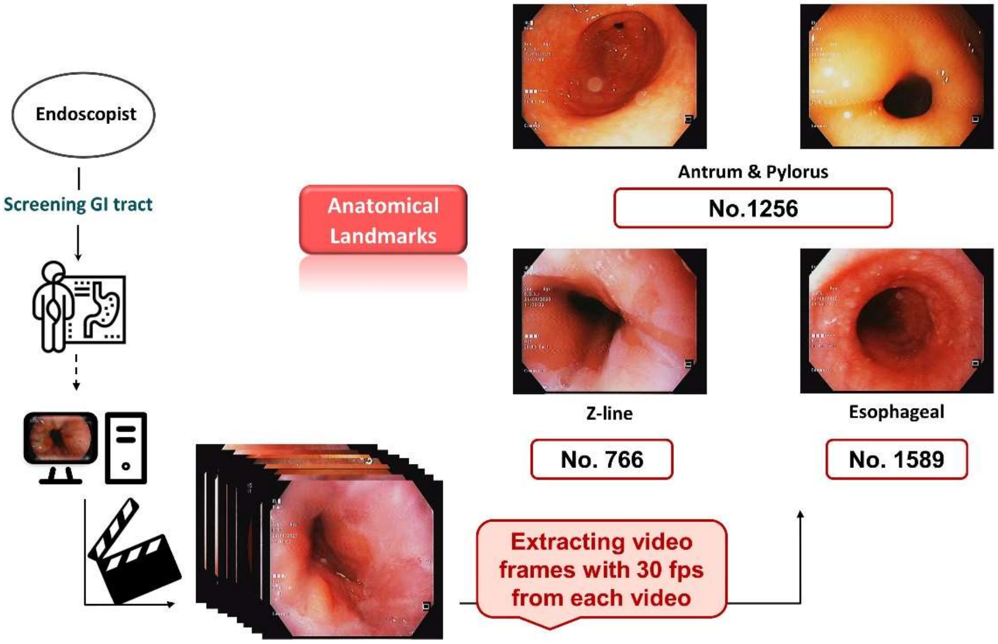


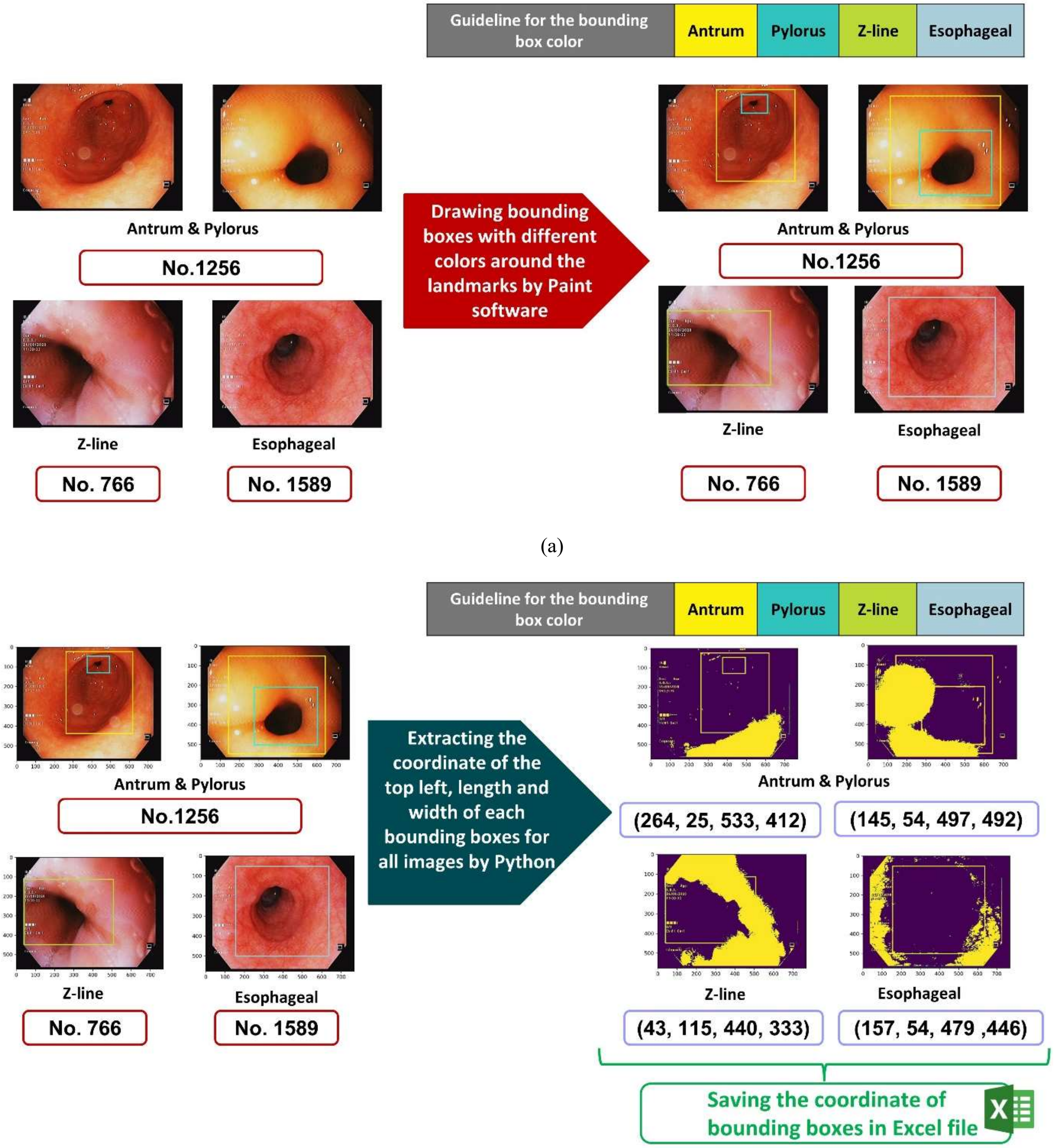
Figure 1 Data description.

As shown in Figure 1, a total of 1,256 images belong to the antrum and pylorus class, 1,589 images are classified as esophageal, and 766 images are categorized as the Z-line class.

2.1.2 Data Preparation

To prepare the data, regions corresponding to the anatomical landmarks described in Section 2.1.1 were annotated using bounding boxes under expert supervision. To enhance accuracy, the bounding boxes were drawn in Paint software using colors that contrasted sharply with the surrounding tissue.

In the subsequent step, the coordinates of the top-left corner, along with the bounding box's width and height, were identified using the Python OpenCV library and recorded in an Excel file for further processing. The detailed steps for data preparation are illustrated in Figure 2.

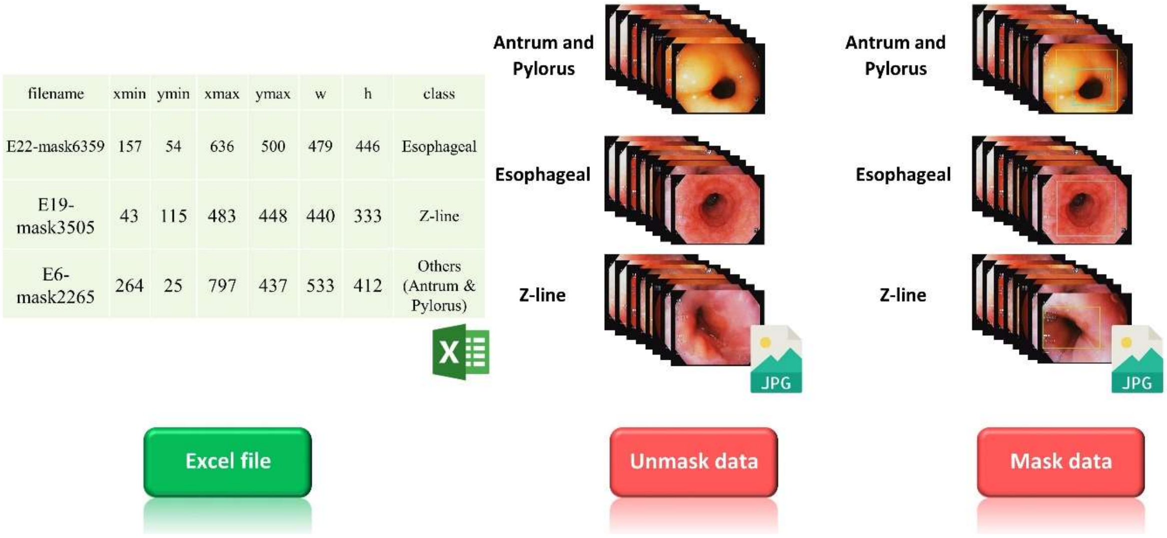


(a)

(b)

Figure 2 Steps of Data Preparation: (a) Drawing bounding boxes around anatomical landmarks, (b) Extracting the coordinates of each bounding box

Figure 2 illustrates the data preparation steps, including drawing bounding boxes using Paint software and extracting their coordinates with the OpenCV library in Python. Figure 3 presents the dataset.



filename	xmin	ymin	xmax	ymax	w	h	class
E22-mask6359	157	54	636	500	479	446	Esophageal
E19-mask3505	43	115	483	448	440	333	Z-line
E6-mask2265	264	25	797	437	533	412	Others (Antrum & Pylorus)

Figure 3 The dataset that is collected and labeled

Figure 3 depicts the dataset, which includes unmasked images, masked images with bounding boxes highlighting anatomical landmarks, and an Excel file containing the bounding box coordinates and corresponding labels for each image.

2.2 Methods

The main steps of the proposed method for detecting and localizing anatomical landmarks in endoscopic video frames are illustrated in Figure 4. The details of the methodology are described in the following subsections:

2.2.1 Dataset Preprocessing

Initially, the unmasked images were resized to 96×128 pixels. The bounding box coordinates were subsequently rescaled and normalized to match the resized images.

2.2.2 The Proposed Method

Figure 4 outlines the primary steps of the proposed method along with the architecture of the CNN model.

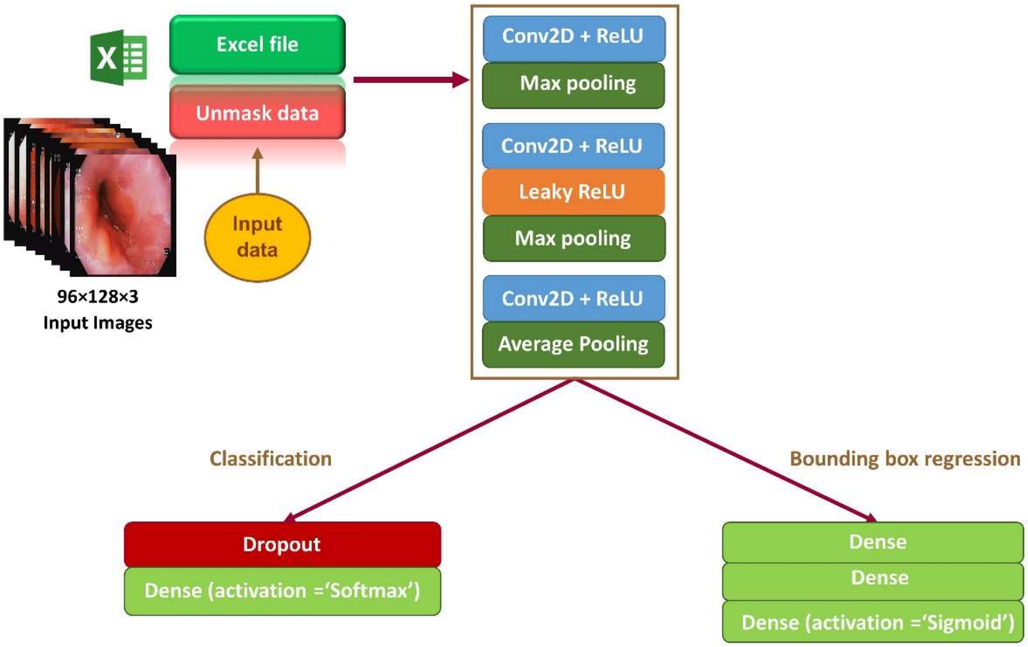


Figure 4 The primary steps of the proposed method for detecting and localizing anatomical landmarks in endoscopic video frames

As illustrated in Figure 4, the proposed method generates two distinct outputs: one for regressing the coordinates of the bounding boxes and the other for classifying the images. The resized unmasked images, along with the rescaled bounding box coordinates, are fed into the CNN for bounding box regression and anatomical landmark classification in endoscopic video frames.

Prior to designing the CNN, the input data is split into training and validation sets in an 80:20 ratio .The detailed architecture of the CNN is presented in Figure 5.

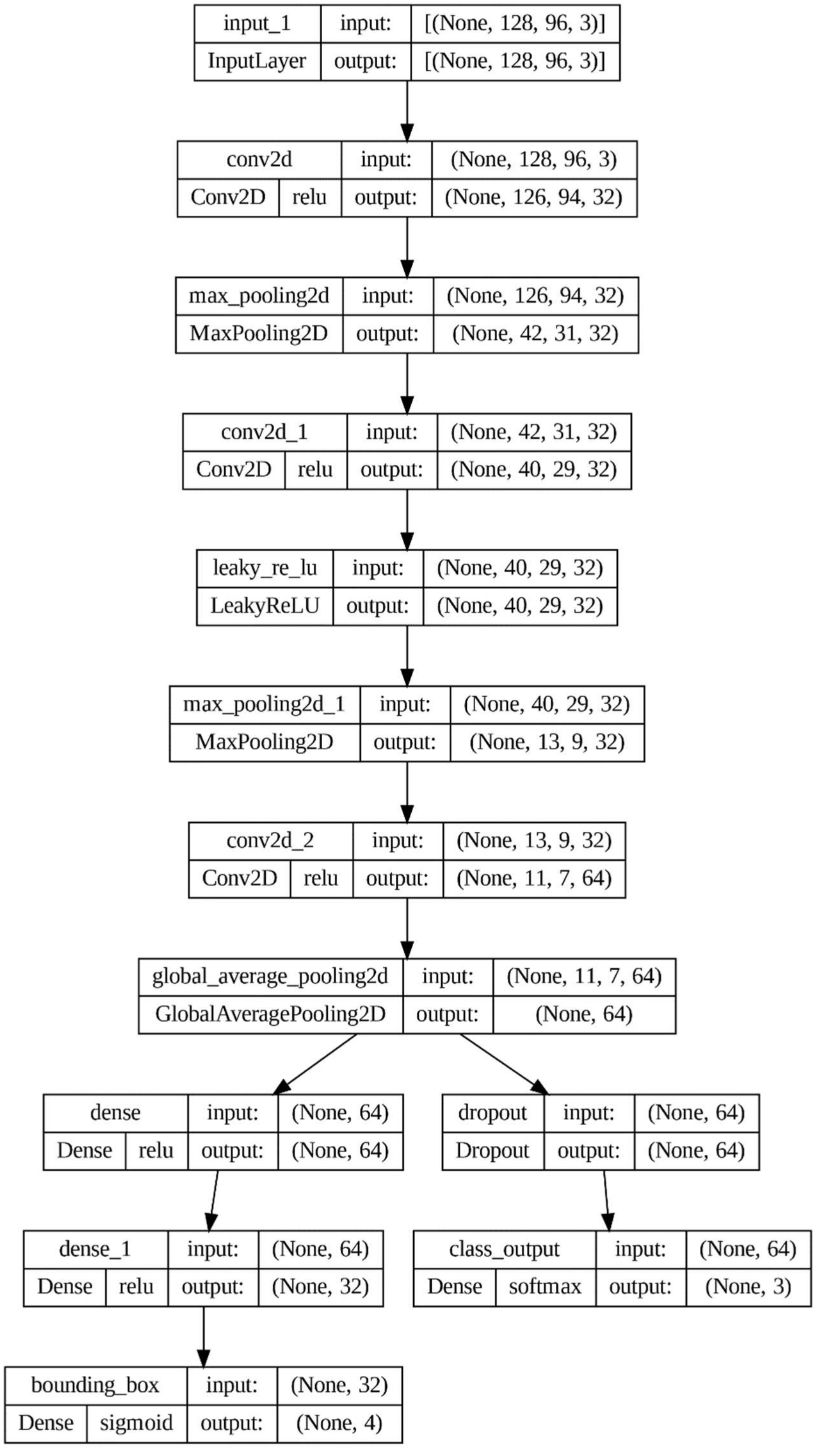

input_1 | input: | [(None, 128, 96, 3)]
InputLayer | output: | [(None, 128, 96, 3)]
conv2d | input: | (None, 128, 96, 3)
Conv2D | relu | output: | (None, 126, 94, 32)
max_pooling2d | input: | (None, 126, 94, 32)
MaxPooling2D | output: | (None, 42, 31, 32)
conv2d_1 | input: | (None, 42, 31, 32)
Conv2D | relu | output: | (None, 40, 29, 32)
leaky_re_lu | input: | (None, 40, 29, 32)
LeakyReLU | output: | (None, 40, 29, 32)
max_pooling2d_1 | input: | (None, 40, 29, 32)
MaxPooling2D | output: | (None, 13, 9, 32)
conv2d_2 | input: | (None, 13, 9, 32)
Conv2D | relu | output: | (None, 11, 7, 64)
global_average_pooling2d | input: | (None, 11, 7, 64)
GlobalAveragePooling2D | output: | (None, 64)
dense | input: | (None, 64)
Dense | relu | output: | (None, 64)
dropout | input: | (None, 64)
Dropout | output: | (None, 64)
dense_1 | input: | (None, 64)
Dense | relu | output: | (None, 32)
class_output | input: | (None, 64)
Dense | softmax | output: | (None, 3)
bounding_box | input: | (None, 32)
Dense | sigmoid | output: | (None, 4)

Figure 5 The CNN architecture designed for the proposed method to detect and localize anatomical landmarks in endoscopic video frames

As illustrated in Figure 5, the CNN architecture comprises multiple layers, each serving a specific function. At the core of the convolutional block is the convolutional layer, a crucial component of the CNN architecture, which performs both linear and non-linear operations. Convolution, a type of linear operation, is instrumental in extracting features from input images [23]. The mathematical representation of the convolution operation is provided in Eq. (1).

$$\mathrm{s(t)} = (x * w)(t) \tag{1}$$

In Eq. (1), the output, commonly referred to as the feature map, is computed as the summation of the element-wise product between the first argument, known as the input, and the second argument, referred to as the kernel. To address challenges such as vanishing and exploding gradients, the CNN applies a non-linear activation function after the linear operation.

The most widely used activation function is ReLU (Rectified Linear Unit), which is defined as $f(x) = max(0, x)$.

Another activation function, Leaky ReLU, overcomes the limitations of ReLU by allowing small gradients for negative input values. The Leaky ReLU is expressed mathematically in Eq. (2).

$$Irelu(x) = \begin{cases} \alpha x & if\ x \leq 0 \\ x & if\ x > 0 \end{cases} \tag{2}$$

The next component of the convolutional block is the pooling layer, which modifies the output by performing a typical down-sampling operation on the feature maps. This reduces the number of trainable parameters, thereby lowering computational complexity. In both the convolutional and pooling layers, hyperparameters such as kernel size, stride, and padding play a crucial role in determining the performance of the model [24].

There are various types of pooling layers, including max pooling, average pooling, mixed pooling, and others [25]. Among these, max pooling is the most commonly used, as it selects the maximum value from each patch of the feature map. The operation performed by max pooling is mathematically represented in Eq. (3).

$$P_{jm} = {}_{k=1}^{r}\max\left(x_{j(m-1)n+k}\right) \tag{3}$$

The Softmax activation function is employed in the final dense layer to classify the inputs into specific categories. The Softmax function transforms the outputs into a probability distribution over classes, as represented in Eq. (4):

$$Softmax(y_i) = \frac{y_i}{\sum_j y_j} \tag{4}$$

Another commonly used activation function, the Sigmoid, generates outputs within the range [0,1], making it suitable for tasks such as bounding box regression. The Sigmoid function is mathematically expressed in Eq. (5):

$$Sigmoid(x) = \frac{1}{1 + e^{-x}} \tag{5}$$

The CNN model is trained for 100 epochs using the Adam optimizer [26], with a learning rate of 0.0001 and a batch size of 8. The ReLU activation function is utilized in all layers except the last ones [27]. The final layer for the classification head applies the Softmax activation function, while the coordinate regression head uses the Sigmoid activation function.

3 Results and Discussion

This section presents the results of the proposed method, evaluated based on performance metrics for classification and bounding box coordinate regression. Table 1 summarizes the performance measures of the proposed method for classifying anatomical landmarks.

Table 1: The performance metrics of the proposed method for detecting and localizing anatomical landmarks in endoscopic video frames

Micro-F1-Score	**Micro-Recall**	**Micro-Precision**	**Macro-F1-Score**	**Macro-Recall**	**Macro-Precision**	**AUC**	**Accuracy**	**Method**
97.00	97.00	97.00	95.00	95.00	96.00	99.00	**97.00**	The proposed method

As presented in Table 1, the proposed method achieved an overall accuracy of 97%. To further evaluate its classification performance across different classes of anatomical landmarks, the results are detailed in Table 2.

Table 2: Macro performance metrics of the proposed method for detecting and localizing anatomical landmarks in endoscopic video frames

F1-Score	**Recall**	**Precision**	**AUC**	**Accuracy**	**Anatomical landmarks**
95.00	96.00	94.00	99.61	**96.00**	Esophageal
91.00	88.00	95.00	99.62	**87.50**	Z-line
100.00	100.00	99.00	99.99	**100.00**	Other (Antrum & Pylorus)

An evaluation of the classification performance for each class in Table 2 reveals that the "Other" class, which includes the antrum and pylorus landmarks, is classified with high accuracy and robust performance metrics. However, a small number of z-line and esophageal landmarks were misclassified.

To evaluate the effectiveness of the coordinate bounding box regression, the regression metrics were analyzed and are presented in Figure 6.

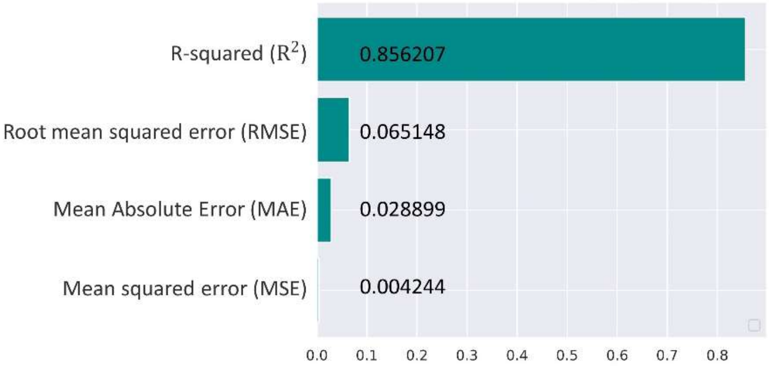


Figure 6 Evaluation of the performance metrics for bounding box regression

The bar chart in Figure 6 illustrates the evaluation metrics for the proposed model, including R-squared (R^2), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Squared Error (MSE). The R^2 value of 0.856 indicates a strong correlation between the predicted and actual values, demonstrating the model's high explanatory power. The RMSE, MAE, and MSE values—0.065, 0.028, and 0.004, respectively—reflect the model's ability to achieve precise and accurate predictions. These metrics collectively highlight the effectiveness of the proposed CNN model in detecting and localizing anatomical landmarks from endoscopic frames with minimal error.

Figure 7 illustrates the ROC curve for each anatomical landmark, showcasing the performance of the proposed method in detecting and localizing anatomical landmarks from endoscopic video frames.

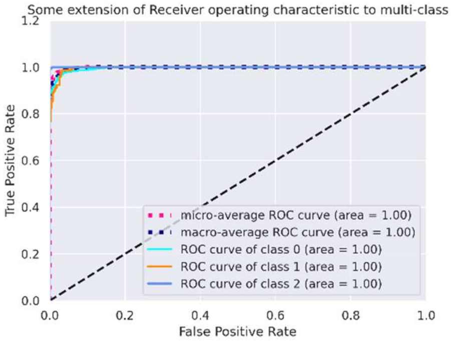


Figure 7 Visualization of the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC)

As depicted in Figure 7, the AUC values indicate the effectiveness of the proposed method. Figure 8 presents the accuracy and loss curves per epoch for both the training and validation datasets.

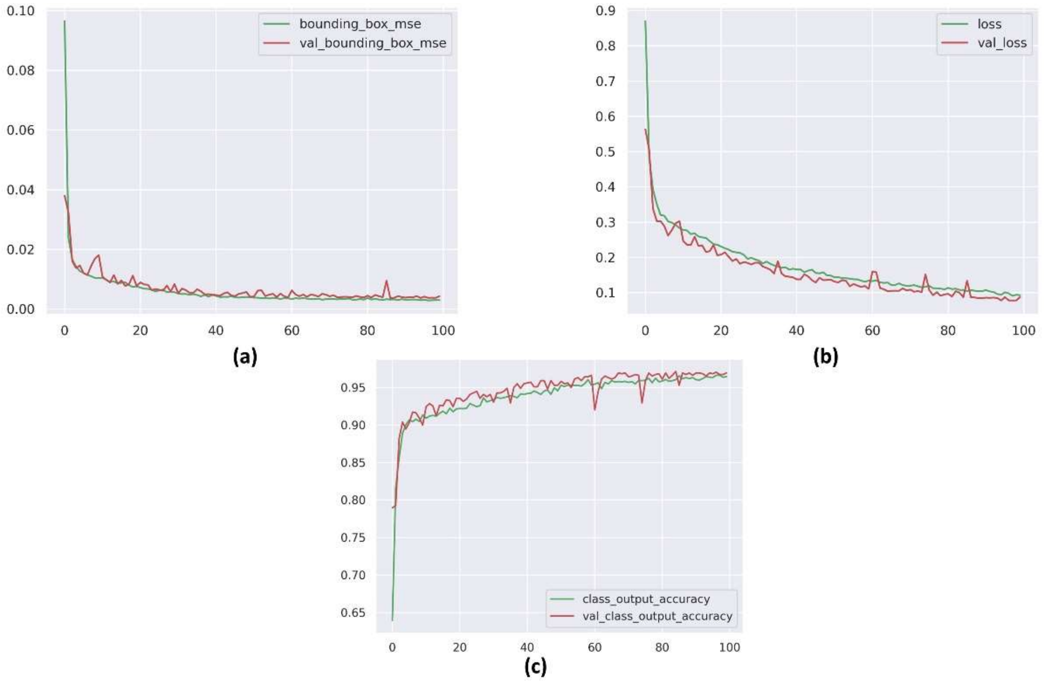


Figure 8 Accuracy and Loss per Epoch for Detection and Localization Evaluation on Training and Validation Data: (a) MSE of Bounding Box Regression, (b) Loss Function per Epoch, (c) Accuracy per Epoch

As shown in Figure 8, the proposed method demonstrates stable performance without overfitting. The loss function and accuracy curves are smooth, with minimal fluctuations, indicating that the model converges effectively and that learning progresses steadily towards an optimal solution.

To evaluate the performance of the bounding box regression module in the proposed method, we compare its results with a similar study focusing on bounding box regression for endoscopic video frames. The comparison is summarized in Table 3.

Our method demonstrates significant improvements in accuracy, macro-precision, and mean squared error (MSE) compared to previous studies. For example, while Wan et al. [28] achieved a MAP of 55.8% in polyp detection, our method achieved an accuracy of 97.0% and a low MSE of 0.004 in the classification and bounding box regression of anatomical landmarks. Furthermore, our method outperforms Gao et al. [8] in sensitivity (97.0% vs. 83.0%) and macro-precision (96.0% vs. 79.1%), offering a more robust solution for detection and localization tasks in endoscopic images.

Table 3: Comparison of the Proposed Method with Related Deep Learning Models for Detection and Localization in Medical Imaging

Comparison Focus	Performance measures	Data	Problem	Technique	Year	Author
Detection of polyps	Mean Average Precision (MAP) =55.8%	Colonoscopy images	Polyp detection	Faster RCNN + function of the location of the object	2020	Wan et al. [28]
Detection of surgical instruments	Average precision (AP) = 79.1%	Endoscopic images	Surgical instruments	Multilevel feature-aggregated deep convolutional neural network (MLFA-Net)	2020	Chu et al. [29]
Segmentation of cancer regions	Sensitivity = 96.0%	Endoscopic images	Segmentation of early gastric cancer regions	Modified mask (RCNN)	2020	Shibata et al. [7]
ROI registration in wireless capsule endoscopy	Precision = 62.8%	Wireless capsule endoscopy	Register the ROIs	The unsupervised deep learning model	2021	Liao et al. [30]
Early cancer detection	Sensitivity = 83.0%	Endoscopic images	Early squamous cell cancer detection	Yolact (you only look at coefficient)	2023	Gao et al. [8]
Classification and bounding box regression for anatomical landmarks	MSE = 0.004 Accuracy = 97.0% Macro-precision = 96.0%	Endoscopic images	Classification and bounding box regression of the anatomical landmarks	CNN	2023	Our proposed method

Table 4 presents the processing time details of the proposed method, computed using Google Colab. To ensure efficient results, the maximum RAM was upgraded to 25.45 GB, and the maximum disk space was 107.72 GB. The GPU models available in Google Colab include NVIDIA K80, P100, P4, T4, and V100 GPUs. The preprocessing, dataset preparation, and implementation of the proposed method were carried out using Python libraries such as Scikit-learn, TensorFlow, Keras, and OpenCV.

Table 4: The processing time for each step of implementing the proposed methods

Processing time (Sec.)	**Processing steps**
0.57	Rescaling the coordinate of the bounding box in an Excel file
39.00	Importing the images of the esophageal class into the Google Colab and preparing them to feed into the CNN model
13.00	Importing the images of the z-line class into the Google colab and preparing them to feed into the CNN model
53.00	Importing the images of the antrum and pylorus class into the Google colab and preparing them to feed into the CNN model
3159.17	Training the proposed method on the training data
2.62	Applying the proposed method to predict the validation data

The primary objective of the proposed method is to develop novel models that leverage the advantages of CNNs for detecting and localizing anatomical landmarks in endoscopic video frames.

4 Conclusion and future works

Cancer is one of the leading causes of death globally, ranking as the first or second cause in 60% of countries worldwide. Early detection remains a significant challenge, as suspicious lesions often progress to malignancy before being identified. Accurate early diagnosis or prognosis of such lesions plays a crucial role in enabling physicians to prescribe appropriate treatments. In recent years, convolutional neural network (CNN) models have shown great promise in addressing these challenges, owing to their capabilities in end-to-end feature extraction and detection.

Anatomical landmarks are critical regions that guide physicians during endoscopic screenings, aiding in the identification of abnormalities. Accurate localization of these landmarks using bounding boxes can be an essential step toward supporting physicians during procedures.

In this study, we proposed a CNN-based model designed to detect and localize anatomical landmarks from endoscopic video frames. The dataset was collected from 40 patients referred to the endoscopy department of Firoozgar Hospital with complaints of stomach pain. Endoscopic frames were extracted from the videos, and the landmarks were labeled with ground truth annotations provided by an experienced endoscopist.

The proposed model is composed of two outputs: one predicts the coordinates of the bounding box around the lesions (regression), and the other classifies the detected anatomical landmarks. Feature maps are extracted using CNN layers, with one output classifying the input images and the other performing bounding box regression.

The evaluation of the model's performance demonstrated favorable accuracy in identifying anatomical landmark regions. These results suggest that the proposed method has the potential to serve as an effective assistance tool in endoscopic screenings, helping physicians identify critical landmarks and improving diagnostic outcomes.

However, the study has certain limitations. The data used for training and evaluation were collected from a single endoscopy department, which may limit the generalizability of the findings. Additionally, the study focused solely on anatomical landmarks and did not include other types of abnormalities, such as gastric cancer lesions.

For future work, we plan to expand the dataset to include images from multiple medical centers to improve the robustness and generalizability of the model. Furthermore, we aim to extend the proposed approach by incorporating images of gastric cancer lesions, enabling the model to assist not only in anatomical landmark detection but also in the diagnosis of pathological abnormalities. Integration with other advanced machine learning techniques and further optimization of the model’s architecture will also be explored to enhance its performance and clinical applicability.

References

[1] F. Bray, M. Laversanne, E. Weiderpass, and I. Soerjomataram, The ever-increasing importance of cancer as a leading cause of premature death worldwide, (in eng), *Cancer,* vol. 127, no. 16, pp. 3029-3030, Aug 15 2021, doi: 10.1002/cncr.33587.

[2] H. Sung *et al.*, Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, *CA: A Cancer Journal for Clinicians,* 2021, doi: 10.3322/caac.21660.

[3] D. Crosby *et al.*, Early detection of cancer, (in eng), *Science,* vol. 375, no. 6586, p. eaay9040, Mar 18 2022, doi: 10.1126/science.aay9040.

[4] K. Sumiyama, T. Futakuchi, S. Kamba, H. Matsui, and N. Tamai, Artificial intelligence in endoscopy: Present and future perspectives, *Digestive Endoscopy,* vol. 33, no. 2, pp. 218-230, 2021, doi: https://doi.org/10.1111/den.13837.

[5] J. Wu, J. Chen, and J. Cai, Application of Artificial Intelligence in Gastrointestinal Endoscopy, *Journal of Clinical Gastroenterology,* vol. 55, no. 2, pp. 110-120, 2021, doi: 10.1097/mcg.0000000000001423.

[6] X. Pang, Z. Zhao, and Y. Weng, The Role and Impact of Deep Learning Methods in Computer-Aided Diagnosis Using Gastrointestinal Endoscopy, *Diagnostics,* vol. 11, no. 4, p. 694, 2021. [Online]. Available: https://www.mdpi.com/2075-4418/11/4/694.

[7] T. Shibata, A. Teramoto, H. Yamada, N. Ohmiya, K. Saito, and H. Fujita, Automated Detection and Segmentation of Early Gastric Cancer from Endoscopic Images Using Mask R-CNN, *Applied Sciences,* vol. 10, no. 11, p. 3842, 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/11/3842.

[8] X. W. Gao, S. Taylor, W. Pang, R. Hui, X. Lu, and B. Braden, Fusion of colour contrasted images for early detection of oesophageal squamous cell dysplasia from endoscopic videos in real time, *Information Fusion,* vol. 92, pp. 64-79, 2023/04/01/ 2023, doi: https://doi.org/10.1016/j.inffus.2022.11.023.

[9] S. Mazumdar, S. Sinha, S. Jha, and B. Jagtap, Computer-aided automated diminutive colonic polyp detection in colonoscopy by using deep machine learning system; first indigenous algorithm developed in India, *Indian Journal of Gastroenterology,* vol. 42, no. 2, pp. 226-232, 2023/04/01 2023, doi: 10.1007/s12664-022-01331-7.

[10] N. Ghatwary, M. Zolgharni, F. Janan, and X. Ye, Learning Spatiotemporal Features for Esophageal Abnormality Detection From Endoscopic Videos, *IEEE Journal of Biomedical and Health Informatics,* vol. 25, no. 1, pp. 131-142, 2021, doi: 10.1109/JBHI.2020.2995193.

[11] S. M. Cho, Y.-G. Kim, J. Jeong, I. Kim, H.-j. Lee, and N. Kim, Automatic tip detection of surgical instruments in biportal endoscopic spine surgery, *Computers in Biology and Medicine,* vol. 133, p. 104384, 2021/06/01/ 2021, doi: https://doi.org/10.1016/j.compbiomed.2021.104384.

[12] K. Yoshiok, K. Tanioka, S. Hiwa, and T. Hiroyasu, Deep-learning models in medical image analysis: Detection of esophagitis from the Kvasir Dataset, *arXiv preprint arXiv:2301.02390,* 2023, doi: https://doi.org/10.48550/arXiv.2301.02390.

[13] S. Ren, K. He, R. Girshick, and J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, *Advances in neural information processing systems,* vol. 28, 2015, doi: 10.1109/TPAMI.2016.2577031.

[14] C. Zhang, N. Zhang, D. Wang, Y. Cao, and B. Liu, Artifact Detection in Endoscopic Video with Deep Convolutional Neural Networks, in *2020 Second International Conference on Transdisciplinary AI (TransAI)*, 21-23 Sept. 2020 2020, pp. 1-8, doi: 10.1109/TransAI49837.2020.00007.

[15] A. Caroppo, A. Leone, and P. Siciliano, Deep transfer learning approaches for bleeding detection in endoscopy images, *Computerized Medical Imaging and Graphics,* Article vol. 88, 2021, Art no. 101852, doi: 10.1016/j.compmedimag.2020.101852.

[16] S. Chen, G. Urban, and P. Baldi, Weakly Supervised Polyp Segmentation in Colonoscopy Images Using Deep Neural Networks, (in English), *Journal of Imaging,* vol. 8, no. 5, p. 121, 2022-08-17 2022, doi: https://doi.org/10.3390/jimaging8050121.

[17] T.-H. Hoang, H.-D. Nguyen, V.-A. Nguyen, T.-A. Nguyen, V.-T. Nguyen, and M.-T. Tran, Enhancing Endoscopic Image Classification with Symptom Localization and Data Augmentation,

presented at the Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 2019. [Online]. Available: https://doi.org/10.1145/3343031.3356073.

[18] A. Hong, G. Lee, H. Lee, J. Seo, and D. Yeo, Deep learning model generalization with ensemble in endoscopic images, *EndoCV@ ISBI,* pp. 80-89, 2021, doi: https://ceur-ws.org/Vol-2886/paper8.pdf.

[19] T. Yu *et al.*, An end-to-end tracking method for polyp detectors in colonoscopy videos, *Artificial Intelligence in Medicine,* vol. 131, p. 102363, 2022/09/01/ 2022, doi: https://doi.org/10.1016/j.artmed.2022.102363.

[20] M. A. Khan *et al.*, Gastrointestinal diseases segmentation and classification based on duo-deep architectures, *Pattern Recognition Letters,* vol. 131, pp. 193-204, 2020/03/01/ 2020, doi: https://doi.org/10.1016/j.patrec.2019.12.024.

[21] Y. Horiuchi *et al.*, Performance of a computer-aided diagnosis system in diagnosing early gastric cancer using magnifying endoscopy videos with narrow-band imaging (with videos), *Gastrointestinal Endoscopy,* vol. 92, no. 4, pp. 856-865. e1, 2020, doi: https://doi.org/10.1016/j.gie.2020.04.079.

[22] D. Ahmedt-Aristizabal, M. A. Armin, S. Denman, C. Fookes, and L. Petersson, Graph-Based Deep Learning for Medical Diagnosis and Analysis: Past, Present and Future, *Sensors,* vol. 21, no. 14, p. 4758, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/14/4758.

[23] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016, doi: http://www.deeplearningbook.org.

[24] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, Convolutional neural networks: an overview and application in radiology, *Insights into Imaging,* vol. 9, no. 4, pp. 611-629, 2018/08/01 2018, doi: 10.1007/s13244-018-0639-9.

[25] H. Gholamalinejad and H. Khosravi, *Pooling Methods in Deep Neural Networks, a Review*. 2020, doi: https://doi.org/10.48550/arXiv.2009.07485.

[26] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980,* 2014, doi: https://doi.org/10.48550/arXiv.1412.6980.

[27] V. Nair and G. E. Hinton, Rectified linear units improve restricted boltzmann machines, in *Icml*, 2010, doi: https://www.bibsonomy.org/bibtex/2a0b7deb2839b69a52fcfcc15b7277a2a/georgheyer.

[28] J. Wan, T. Chen, B. Chen, Y. Yu, Y. Sheng, and X. Ma, A Polyp Detection Method Based on FBnet, *Computers, Materials \& Continua,* vol. 63, no. 3, pp. 1263--1272, 2020. [Online]. Available: http://www.techscience.com/cmc/v63n3/38874.

[29] Y. Chu *et al.*, Multi-level feature aggregation network for instrument identification of endoscopic images, *Physics in Medicine & Biology,* vol. 65, no. 16, p. 165004, 2020/08/10 2020, doi: 10.1088/1361-6560/ab8dda.

[30] C. Liao, C. Wang, J. Bai, L. Lan, and X. Wu, Deep learning for registration of region of interest in consecutive wireless capsule endoscopy frames, *Computer Methods and Programs in Biomedicine,* vol. 208, p. 106189, 2021/09/01/ 2021, doi: https://doi.org/10.1016/j.cmpb.2021.106189.