Transfer Learning Based Deep Neural Network for Detecting Artefacts in Endoscopic Images

Original Scientific Paper

Kirthika Natarajan
School of Engineering, Avinashilingam Institute for Home Science and Higher Education for Women, Varapalayam, Coimbatore, Tamilnadu 641 108, India.
prof.kirthika@gmail.com

Sargunam Balusamy
School of Engineering, Avinashilingam Institute for Home Science and Higher Education for Women, Varapalayam, Coimbatore, Tamilnadu 641 108, India.
sargunamb@gmail.com

Abstract *– Endoscopy is typically used to visualize various parts of the digestive tract. The technique is well suited to detect abnormalities like cancer/polyp, taking sample tissue called a biopsy, or cauterizing a bleeding vessel. During the procedure, video/ images are generated. It is affected by eight different artefacts: saturation, specularity, blood, blur, bubbles, contrast, instrument and miscellaneous artefacts like floating debris, chromatic aberration etc. The frames affected by artefacts are mostly discarded as the clinician could extract no valuable information from them. It affects post-processing steps. Based on the transfer learning approach, three state-of-the-art deep learning models, namely YOLOv3, YOLOv4 and Faster R-CNN, were trained with images from EAD public datasets and a custom dataset of endoscopic images of Indian patients annotated for artefacts mentioned above. The training set of images are data augmented and used to train all the three-artefact detectors. The predictions of the artefact detectors are combined to form an ensemble model whose results outperformed well compared to existing literature works by obtaining a mAP score of 0.561 and an IoU score of 0.682. The inference time of 80.4ms was recorded, which stands out best in the literature.*

Keywords*: Deep learning, Artefacts, Endoscopy, Transfer learning*

1. INTRODUCTION

Endoscopy is a non-surgical technique that encompasses inserting a thin and long flexible tube called an endoscope down through the throat to inspect a person's gastrointestinal tract. The flexible tube is attached with a light and a camera. A gastroenterologist uses an endoscope to diagnose and treat common ailments in the digestive tract, collect tissue samples called a biopsy, pass special tools through the endoscope to treat bleeding vessels, remove a foreign body or a polyp etc.

Recent technologies allow doctors to switch between imaging modalities like narrow-band imaging (NBI), fluorescence light and white light to detect abnormalities better. These technologies improve the visualization of the mucosal surface and microvascular pattern. The internal organ is viewed on a television monitor connected outside during the procedure. Also, the complete process is recorded. The clinician can review the recorded video for planning further treatment, report preparation, discussion with a senior clinician and follow-up etc.

Artefacts [1] are the artificial effect found in most endoscopy images but are not present in the imaged organ. The presence may be due to mishandling miniaturized components, hand movements, natural causes etc. These artefacts affect the video quality and increase procedure time. In the recorded video, most of the frames are affected by artefacts. Hence, the most affected frames are discarded, which reduces the quality of the video during post-processing, thus directly affecting the quality of treatment and diagnosis. Also, these artefacts may obscure features/ characteristics relevant to an abnormality like cancer. They also increase false detection rates in Computer-Aided Diagnostic (CAD) systems. Thus, an efficient method to detect these artefacts prior may help the doctors to speed up the procedure with improved accuracy. It can be accomplished by deploying artificial intelligence. Figure 1 displays endoscopic images affected by artefacts.

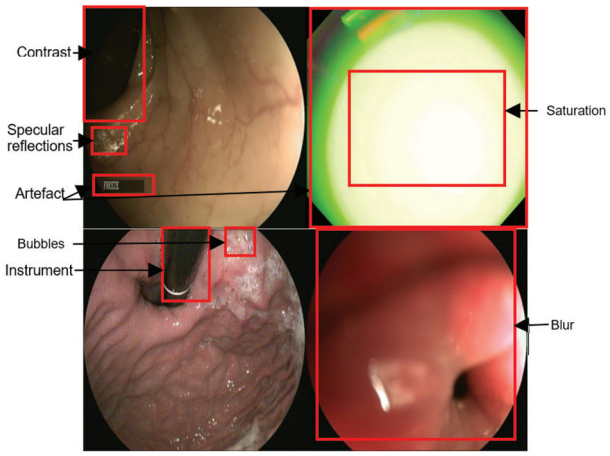


Fig. 1. Endoscopic images affected by artefacts

Figure 1. highlights only a few artefacts to maintain clarity. Encouragingly, Deep Learning (DL) algorithms, a subset of AI, have the power to extract features from annotated images without human intervention. Congregating the study toward endoscopy, we deploy pre-trained DL-based object detection models for detecting endoscopic artefacts in this research.

The motivation behind the research is as follows. Firstly, the clinician cannot adequately examine the underlying tissue due to the presence of various artefacts, which increases procedure time. Secondly, the procedure is recorded, and specific regions are imaged for further examination and report preparation. Most of the frames are affected by endoscopic imaging artefacts, due to which the affected frames are discarded as no helpful information can be extracted from them.

Third, the artefact varies in size and location in an endoscopic image. The size of artefact-like specular reflection is tiny, occurring in groups. On the other hand, other artefacts like saturation and contrast cover a large area. This challenges the object detection algorithm to detect objects of various scales. Interestingly more than one artefact occurs in most of the frames. Thus, locating tiny to significant artefacts in a single frame adds complexity to the existing problem.

Fourth, deep learning models need many labelled data to train themselves. Especially in the medical imaging sector, the amount of labelled data for every abnormality is significantly less. In such a scenario, Transfer Learning (TL) lends its helping hand. It is important to note that choosing the suitable model for every application is a tedious and trial and error process. It is an unsolved research issue to date. This research considers the unique property of every algorithm in the literature and selects the one that meets the accuracy and inference time balance. For this research purpose, several algorithms are trained; namely, You Only Look Once (YOLO)v3 [2], YOLOv4[3] and RetinaNet [4], Faster Region-based -Convolutional Neural Network (R-CNN) [5] with various backbones. The final model is chosen after estimating the performance of all trained models.

To train any DL-based object detection algorithm massive dataset is essential. A few datasets are available for research / academic purposes to study artefacts in endoscopic images. They are the Kvasir-Instrument dataset [6], Computer Vision Centre (CVC)-ClinicSpec [7], Cholec80 dataset [8] etc. Most of these datasets hold annotations for a single artefact only. But in real-time endoscopic images are severely affected by various other artefacts also. To serve the purpose of multiple artefact detection: The endoscopic Artefact Detection (EAD) dataset [9][10] is available. The datasets hold annotations for common artefacts like saturation, specular reflections, blur, blood, bubbles, instrument, contrast and miscellaneous imaging artefacts.

Authors reported that the EAD datasets suffer from a class imbalance problem [11][12]. A standard solution accepted across the globe is to use the data augmentation technique [13]. It is also vital that all the data augmentation techniques cannot be adopted to all medical images. Carefully choosing the method is essential. After trivial analysis considering the availability of data, pre-trained models, hardware requirements etc., This research paper seeks to analyze the performance of three different object detection models using a custom dataset.

The specific contribution of this research article is as follows:

- A new dataset has been curated with clinician assistance to add more images to the dataset with patients of Indian origin to combat data requisite for DL algorithms.
- We have trained 3 DL-based object detection models, namely YOLOv3, YOLOv4 and Faster R-CNN, for multi-class artefact detection.
- All the three trained artefact detectors are combined to form an ensemble model for improved performance.
- The research results prove superior performance over literature outcomes, and the results are compared with recently reported literature works.

The outline of the research article is as follows. Section 2 explores literature works related to multiple endoscopic artefact detection. Section 3 addresses the details of curation of the new dataset, annotation protocols and the details of the public dataset. Section 4 gives a comprehensive report on methodology, transfer learning approach, training of various models and design of proposed ensemble model architecture, followed by a detailed description of the results obtained. The last section reports the conclusion and presents the future scope of this research findings.

2. RELATED LITERATURE WORKS

Deep learning algorithms have shown exceptional performance in every branch of the health care industry in the past decade. In recent years, deploying DL algorithms in detecting multiple artefacts gained im-

portance after the release of EAD datasets. This section concisely presents the literature works relevant to the field of multiclass endoscopic artefact detection.

Pengyi Zhang et al. (2019) [14] proposed a modified version of Mask R-CNN called Mask Aided R-CNN. Initially, a basic Mask R-CNN is trained for the segmentation task. The trained Mask R-CNN is used to predict instant masks for training samples from the detection set. The masks are predicted only for ground truth bounding boxes. The predicted masks are termed soft-pixel level labels which are added to the segmentation set to retrain the network. This strategy proves to be the best in the detection task.

Yan-Yi Zhang and Di Xie (2019) [15] proposed a cascaded R-CNN-based model and trained the model by gradually increasing the Intersection over Union (IoU) threshold. The model was initially pre-trained using Microsoft Common Objects in Context (MSCOCO) dataset [16] and later retrained with EAD datasets.

Hoang Manh Hung et al. (2020) [17] presented a DL-based cascaded R-CNN with ResNeXt-101 backbone followed by Feature Pyramid Network (FPN). This combination improved the feature extraction capability of the network and recall rate. To differentiate the object from the background, the authors added Deformable Convolution (DCN) to the network, improving the performance.

Hongyu Hu and Yuanfan Guo (2020) [18] designed a cascaded R-CNN-based architecture with ResNeXt as backbone and FPN to extract features. The author adopted multi-scale detection techniques to scale images from 512x512 to 1024x1024 randomly. Flipping images horizontally was employed to expand the dataset size. Soft-Non-Max Suppression (NMS) was adopted, which avoids unnecessary ignoring of objects.

Zhimiao Yu and Yuanfan Guo (2020) [19] used a cascaded R-CNN-based model with ResNet101 as the backbone with FPN. The network used an ImageNet pre-trained backbone. The author adopted data augmentation, soft-NMS, cosine decay strategy for learning rate schedule, cross-entropy loss and smoothL1loss for classification and regression.

Xiaohong Gao and Barbara Braden (2020) [20] presented a DL network based on RetinaNet. The author incorporated a real-time instance segmentation task into RetinaNet to cater for the need for object detection and instance segmentation.

Anand Subramanian and Koushik Srivatsan (2020) [21] experimented RetinaNet with ResNet101 feature extractor for artefact detection. The authors used image correlation-based trackers to reduce inference time, improving the network performance.

This section summarized recent works of literature in the domain of endoscopic artefact detection. All the researchers used EAD datasets to train the algorithms. The authors selected a state-of-the-art object detection model and trained the model with various backbone and learning strategies, deployed augmentation techniques, and cascaded the structures to produce efficient results.

3. DATASETS

EAD2019 is a dataset that covers seven major artefacts like specularity, saturation, contrast, blur, bubbles, instruments and miscellaneous imaging artefacts. The dataset contains 2147 images. Figure 2 shows sample images from the EAD2019 dataset. EAD2020 comprises 2531 images and covers eight imaging artefacts, including blood and all seven artefacts covered by the EAD2019 dataset. Figure 3 illustrates some images from the EAD2020 dataset. Expert clinicians suggested all the artefacts mentioned above.

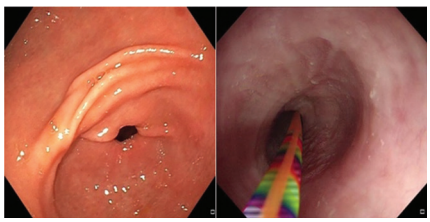


Fig. 2. Sample images from EAD2019 dataset

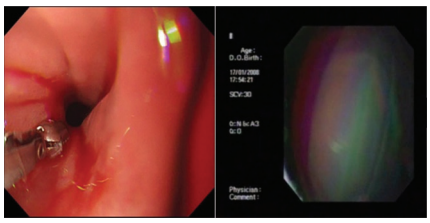


Fig. 3. Sample images from EAD2020 dataset

Both the datasets are multi-patient, i.e., the images are from 6 distinctive centres globally. It is a multi-organ dataset covering the oesophagus, stomach, liver, colon and bladder. Also, it is a multi-tissue and multi-modality dataset (white light, fluorescence light, and Narrow band imaging). Videos collected from these centres were imaged using standard endoscopes manufactured by Karl Storz, Olympus and Bio spec. The images hosted in the dataset do not contain any patient information.

Initially, senior clinicians annotated the images and later, the experienced post-doctoral fellows. Finally, the senior clinicians validated the images. All the images were annotated (bounding box) for artefact detection using python, Qt, and an Open-CV-based in-house tool. The dataset contains images and a binary mask for semantic segmentation.

3.1 CUSTOM DATASET

The public dataset contains images of patients from western countries. The images with artefacts like blur, instrument and saturation were not much found. There-

fore 2400 endoscopic images of Indian patients were collected, which includes more images on saturation, blur and instrument to combat class imbalance problems in the public dataset. Initially, images were annotated in the presence of a senior clinician to gain expertise. Later annotations were done individually and finally validated by the clinician in the ratio of 1:10(no. of images). Figure 4 shows random sample images from the custom dataset.

Fig. 4. Images from custom dataset

3.2 ANNOTATION PROTOCOLS

Annotation protocols from the EAD dataset [22] were followed for annotating the custom dataset. Images from EAD and custom dataset were used to train the endoscopic artefact detector. Thus, uniform annotation protocols were used to maintain homogeneity across all three datasets. Figure 5 portrays images of a custom dataset labelled for eight commonly occurring artefacts. Artefacts like instruments, saturation, blur and contrast cover a larger area when compared to the artefact called specularity. Specular reflections cover a small region; for precise delineation, most specular reflections are marked with a separate bounding box.

One single bounding box was used in some cases where the reflections are found in a series fashion. In the curated custom dataset, care was taken that no patient information was visible. It is apparent from the images exhibited that more than one artefact is said to be present in almost all the frames.

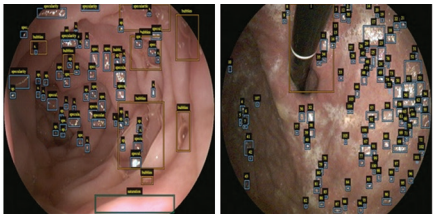


Fig. 5. Annotated images from custom dataset

3.3 ANNOTATION SOFTWARE

All annotations were done using VGG (Visual Geometry Group) Image Annotator (VIA) [23], an open-source annotation tool offered by Oxford University, United Kingdom.

4. METHODOLOGY

Deep learning models like YOLOv3, YOLOv4 and Faster R-CNN were chosen. The former two were selected based on their faster inference and faster R-CNN for its accuracy. All the models were trained and tested with images from EAD and custom datasets. This training was thoroughly carried out with Google Co-laboratory [24] single Graphics Processing Unit (GPU) environment. After training, all three models were evaluated, tuned, retrained and combined for an ensemble model. The ensemble model is tested with images from the test set. To attain the best performance transfer learning approach was chosen.

4.1 TRANSFER LEARNING

Transfer learning is a technique to train DL models on massive datasets like MSCOCO, ImageNet [25] etc. Later, for specific applications, these models can be retrained. During retraining, the model with already stored knowledge learns the features of new applications at a faster rate with reasonable accuracy. TL helps reduce the training time, hardware cost and the required high-dimensional dataset. In the present decade, many pre-trained models are available in the model zoo [26][27] for research purposes. Figure 6 shows a simple TL model. The following sub-sections discuss the three deep learning models used in this study.

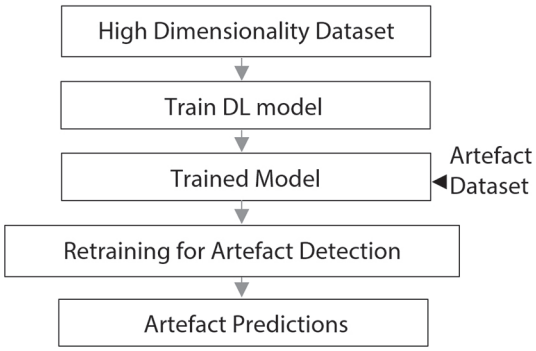


Fig. 6. Outline of transfer learning-based detection model

4.2 YOLOV3

YOLO first came into existence in the year 2016. Out of all updated versions of YOLO, YOLOv3 was tagged for its best performance in terms of speed. YOLOv3 looks at the complete image once and divides it into small grids. In each grid, bounding boxes will be drawn if there is a meaningful object. The predictions and their similarity with the predefined classes are calculated. When the score is high, it is considered a positive detection. YOLOv3 uses Darknet53 as the backbone to extract features. YOLOv3 also finds a good balance between detecting various sizes of objects, from tiny to large. This characteristic is beneficial in the case of detecting artefacts like specular reflections, which are small and artefacts like contrast and instrument, which are prominent.

4.3 YOLOV4

YOLOv4 is the fourth principal member added to the YOLO family in 2020 by Alexey Bochkovskiy. It has many special features. One or more of the features can be combined and utilized for applications to obtain state-of-the-art results. They are grouped under two heads: Bag of Specials (BoS) and Bag of Freebies (BoF). BoF helps to improve model accuracy without compromising the model inference time. On the other hand, BoS aid in improving accuracy at the cost of inference time. Thus, the researcher must select the best strategies for the best results.

4.4 FASTER R-CNN

Faster R-CNN is a two-stage object detector from the R-CNN family. It uses the Region proposal Network (RPN) to improve its performance.

In Faster R-CNN, the input image is passed into the ConvNet, which returns the feature maps, and RPN is applied to the feature maps to get object proposals. Using the Region of Interest (ROI) pooling layer, all the proposals are brought down to the same size. Finally, they are sent to a fully connected layer to classify and predict the classes of the objects in the bounding boxes.

5. EXPERIMENTAL ANALYSIS

This section presents details of datasets, training and testing of models, evaluation criteria and results obtained.

5.1 DATASET

EAD datasets embrace endoscopic images annotated for various artefacts like saturation, specularity, blood, bubbles, contrast, blur, instruments and miscellaneous artefacts. In total, 2147 images from the EAD2019 dataset and 2531 images from the EAD2020 dataset were used in this research. Apart from the existing public dataset, the newly curated dataset with 2400 annotated images was used. Thus approximately 7000 images were pooled to form the training and test set.

5.2 TRAINING AND TESTING

The proposed research work is written using python. The training of the artefact detection model was done on a google co-laboratory single GPU environment. Initially, all the images were pooled and manually split into train and test with 80% and 20% split-up. Later train set was divided into train and validation sets. Finally, 70% of the images were allocated for training, 10% of total images for validation and 20% of total images for testing. Three models, namely YOLOv3, YOLOv4 and Faster R-CNN, were trained using the train set's augmented images. The training strategy followed for each model is discussed in the sections below.

It is well known that more images are required to train deep learning-based algorithms. Images available may not be sufficient to make the detector robust; hence data augmentation was adopted.

5.3 DATA AUGMENTATION

Augmentation is a technology that magnifies the dataset by slightly modifying the existing images [28]. The augmentation technique must be carefully chosen. It may affect the performance of the detector if not appropriately selected. For our study, we have adopted rotation at various angles, namely 0^o,90^o,180^o,270^o and flipping. Hence the existing data expanded eight times. Then augmented dataset was used to train all three algorithms. Along with manual augmentation techniques, this research adopts run time augmentation techniques like a mosaic, varying hue, saturation, brightness and other augmentation techniques offered by the network to make the detector more robust.

5.4 TRAINING OF YOLOV3

YOLOv3 was cloned from Darknet [29]. The augmented dataset was used to train YOLOv3 for artefact detection. The algorithm has various runtime augmentation techniques like varying hue, saturation and exposure; it was also considered during training. Various parameters set during training are as follows, learning rate = 0.001, batch size=64, maximum batches = 16,000, image size=416x416. By setting all the initializations, the training started with pre-trained weights. The training lasted until the network reached a minimum average loss. Approximately after 55,000 iterations, the average loss curve was found to be smoothed. Once the average loss no longer reduces, the iterations can stop. On the other hand, the iterations can be stopped when the loss reaches 0.05, provided the dataset is small, and 3.0 if the dataset is bigger [29]. The average loss did not improve after 55,000 iterations. The training was stopped at 70,000 iterations. Various weight files extracted during training are tested for their performance in terms of mAP and IoU. Weights file extracted at 65,000th iteration gave its best results.

5.5 TRAINING OF YOLOV4

The basic network architecture was cloned from [30]. YOLOv4 was customized with the following features: CSPdarknet53 as the backbone, PanNet for aggregating the features and YOLOv3 head for final predictions. Special features from BoF and BoS like mosaic augmentation, Mish Activation function, NMS, optimized anchors etc., were handpicked. Pre-trained weights were opted to reduce training time.

The other important initial hyper-parameters set for training are as follows, image size= 512x512, batch size=64, momentum=0.949, decay=0.0005 and learning rate=0.013. Few run-time data augmentation techniques like varying hue, saturation, exposure, cut-mix and mosaic were deployed. With the above set parameters, training lasted till 85,000 iterations. Until 35,000

iterations, handpicked features like cut-mix and mosaic augmentation were employed. The loss did not converge as expected. Hence on trial-and-error bases, cut-mix augmentation was removed, which yielded results as expected. The training was stopped at 95,000 iterations. Weights files extracted during the training process were examined for best results, and it was decided to use the 76,000th weight file, which gave a good balance between mAP, IoU and inference time as well.

5.6 TRAINING OF FASTER R-CNN

Faster R-CNN was adopted from detectron2 [31], built by Facebook AI Research (FAIR). It holds a model zoo consisting of trained model files for faster implementations and several baselines for research. The training parameters set to train Faster R-CNN is as follows: Image size=512x512, backbone = Resnet50, learning rate =0.1, ROI head size= 256 x 256 and batch size = 4. Faster R-CNN training started with the initialized training parameters. Model checkpoints were set to every 1000th iteration. The training lasted for 80,000 iterations. The trained weight file extracted at the 74,000th iteration gave good accuracy and reasonable inference time.

All three models are trained with pre-trained weight. The need to choose the TL approach in this research is to reduce the training time. The images available for training is also limited; hence it was preferred to use TL rather than training from scratch. The impact of TL has been proven by exhibiting accurate results in lesser iterations and with a limited size of the training dataset.

5.7 PROPOSED ENSEMBLE MODEL

The term 'ensemble' means collective or collaborative. Ensemble learning model combines the predictions of multiple object detection models to improve the overall performance.

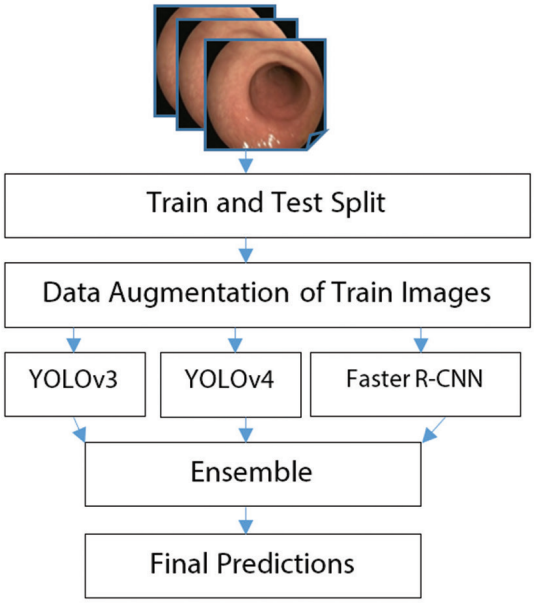


Fig. 7. Ensemble endoscopic artefact detection model

Ensemble models are classified into three types: affirmative, unanimous, and consensus. The proposed ensemble model combines the benefits of both single-stage and two-stage object detectors. The trained and tested model files of YOLOv3, YOLOv4 and Faster R-CNN are blended together for predictions. A test image is passed into the model. All three trained models predict every artefact present in the image. Based on the ensemble method chosen, final predictions will be generated. Out of all three methods result of the consensus, the model proved exemplary. Figure 7 depicts the proposed ensemble model.

5.8 RESULTS

This section discusses the performance of the proposed ensemble model against different literature results compared based on standard performance parameters like IoU, mAP and Inference time.

5.8.1 Mean average precision (mAP)

Average Precision (AP) can be calculated by intersecting the precision-recall (PR) curve and coordinate axis at recall values, say $r_1, r_2, ..r_n$. Equation (1) is used to calculate the AP score.

$$AP = \sum_n (r_{n+1} - r_n) P_{interp}(r_{n+1}) \quad (1)$$

where p_{interp}= max p(r) and mAP can be calculated by taking the mean of every AP using Equation (2) over all artefacts i. N=8, is the total number of classes.

$$mAP = \frac{1}{N} \sum_i AP_i \quad (2)$$

5.8.2 Intersection over union (IoU)

IoU must be calculated using the formula in (3). IoU is a ratio between the intersection of ground truth(A) and predicted bounding boxes(B) and the union of ground truth(A) and predicted bounding boxes(B).

$$IoU = \frac{A \cup B}{A \cap B} \quad (3)$$

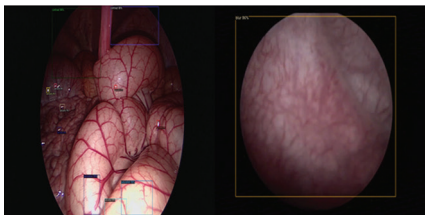
5.8.3 Score_d

Score_d is a weighted score of IoU and mAP as given in Eq. (4).

$$Score_d = 0.6 * mAP + 0.4 * IoU \quad (4)$$

The ensemble artefact detection model combines predictions of all three base learners and produces a final prediction based on the ensemble methods chosen. In the affirmative method, all the models can predict objects in the image. Even if one model proposes a bounding box for an object, it will be considered for the final predictions of the ensemble model. In a unanimous approach, all the models can predict bounding boxes around the objects in the image. If all the three models predict the same instance, then that instance is considered for final predictions provided if the IoU

is greater than 0.5. Finally, in the consensus approach, a bounding box around an instance will be considered if most models generate the same box. Fig. 8 illustrates the results of the ensemble model with a consensus approach.



(a) (b)

Fig. 8(a) &(b): Artefact detection by ensemble model

In Fig. 8(a), three different locations are affected by an artefact called saturation. Each of them was detected with 86%,99% and 98% accuracy. Similarly, artefact contrast was detected with 86% and 96% accuracy. Specular reflections are scattered around the image; each of them was predicted with 70%-98% of accuracy. In Fig. 8(b), the artefact blur was detected with 86% accuracy. Hence the prediction accuracy proves the robustness of the detector.

5.9 COMPARATIVE ANALYSIS

This section presents the comparative analysis of the multi-class endoscopic artefact detection model with literature results. This study compares research outcomes based on performance evaluators of the detectors in terms of mAP, IoU, Score_d and inference time. The average precision obtained by the model in detecting every artefact is tabulated in Table 1.

Table 1. Class-wise average precision scores

Class	YOLOv3-Spatial Pyramid Pooling[32]	Proposed model
Specular reflections	34.7	48.12
Saturation	55.7	56.46
Miscellaneous Artefact	48.0	44.91
Blur	7.5	51.31
Contrast	72.1	36.74
Bubbles	55.9	51.61
Blood	-	58.39
Instrument	-	100.00

From the table, it is evident that the model has a balanced performance over predicting all artefacts. The combination of EAD and custom datasets to counterbalance the class imbalance problem has turned prolific. Specular reflections and a few miscellaneous artefacts are tiny, but saturation covers comparatively a bigger area. But all three artefacts have a common attribute of having bright pixel areas. Yet the trained model is capable of differentiating them well. Contrast has different characteristics of having dark pixel areas.

Blur has an attribute of un-sharpness or having a poor spatial resolution. Artefacts like bubbles, blood and instrument have different attributes like well-defined boundaries for instruments and colour features for blood and bubbles. Often artefacts like bubbles, miscellaneous artefacts and specular reflections overlap, yet in most cases, the model predicts the artefacts well.

The common metric used to evaluate the performance of every detection model is the mAP and IoU. The model is evaluated by having a threshold value of 0.5. The results are deliberated in Table 2. Score_d is a metric exerted from the EAD challenge [22]. The metric is used to compare the performance of the proposed model with literature results.

Table 2. Comparative analysis

Author	mAP	IoU	Score_d
Yan-Yi Zhang and Di Xie [15]	-	-	0.3429
Xiaohong Gao and Barbara Braden [20]	-	-	0.2205
Pengyi Zhang et al. [14]	0.3117	0.4051	0.361
Anand Subramanian and Koushik Srivatsan [21]	0.2151	-	-
Proposed method	0.561	0. 682	0.6094

Almost all researchers concentrated on accuracy, but inference time is equally essential when it comes to real-time implementation of the modules in CAD and semi-automated/ fully automated robotic systems. Thus, this ensemble model with trained base learners produced an impactful research output in detecting multiple endoscopic artefacts. Often authors try to balance inference time and accuracy. An inference time of 80.4ms was observed during the testing of the proposed ensemble model. Most of the authors concentrate on prediction accuracy. Thus, there are not plenty of results to analyze the work based on inference time. This result can be a benchmark for researchers in this area.

6. CONCLUSION AND FUTURE SCOPE

For this study, three different state-of-the-art object detection models: YOLOv3, YOLOv4, and Faster R-CNN, were trained using an augmented train set comprising 56,000 images. The test set contains 1400 images. With all the three base learners trained on EAD and custom datasets, a new ensemble model has been designed, which combines the predictions of all the models. The final proposed model is evaluated against performance criteria like mAP, IoU and inference time. It was observed that the ensemble model under the consensus method with three base learners mentioned above was said to perform well against literature results proposed in [14][15][17–21] with mAP=0.561 and IoU score of 0.62 and inference time of 80.4ms.

This work can be expanded by incorporating a restoration algorithm for every possible artefact. The endoscopic artefact detector could be re-trained to detect

all possible clinical abnormalities in the GI tract. Thus, it could become an all-in-one detection and restoration system, which could aid clinician with better viewability of internal organs, reduces procedure time, improve the prediction accuracy of the CAD system and aid associated post-processing steps. The custom dataset curated can be expanded by adding more images to create a large benchmark dataset.

7. STATEMENT OF ETHICS

This research involves collecting endoscopic images to prepare a custom dataset for which ethical clearance was obtained from the institution's human ethics committee.

8. ACKNOWLEDGEMENTS

This research work is supported by DST-CURIE-AI Facility, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore.

9. REFERENCES:

[1]. Wikipedia, Artifact (Error), https://en.wikipedia.org/wiki/Artifact_(error) (accessed: 2022)

[2]. J. Redmon, A. Farhadi, "YOLOv3: An Incremental Improvement", arXiv:1804.02767, 2018.

[3]. A. Bochkovskiy, C. Y. Wang, H. Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection", arXiv:2004.10934, 2020, pp.1-17.

[4]. T. Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, "Focal Loss for Dense Object Detection", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 42, No. 2, 2020, pp. 318-327.

[5]. S. Ren, K. He, R. Girshick, J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", Proceedings of the 28th International Conference on Neural Information Processing Systems, Montreal, Canada, Vol. 1, 7-12 December 2015, pp. 91-99.

[6]. Simula Research Laboratory, Kvasir-Instrument dataset, https://datasets.simula.no/kvasir-instrument/ (accessed: 2022)

[7]. CVC-ClinicSpec dataset: http://www.cvc.uab.es/CVC-Colon/index.php/cvc-clinicspec/ (accessed: 2022)

[8]. Research Group CAMMA, Cholec80 datasets, http://camma.u-strasbg.fr/datasets (accessed: 2022)

[9]. S. Ali et al. "Endoscopy Artefact Detection (EAD) Dataset", Mendeley Data, 2019, V2.

[10]. S. Ali et al. "Endoscopy Artefact Detection (EAD) Dataset (includes updated 2020 version)", Mendeley Data, V4, 2021.

[11]. I. Oksuz, J. R. Clough, A. P. King, J. A. Schnabel, "Artefact Detection in Video endoscopy Using RetinaNet Architecture and Focal Loss Function", Proceedings of the Challenge on Endoscopy Artefacts Detection: Multi-class Artefact Detection in Video Endoscopy, Venice, Italy, 8 April 2019.

[12]. X. Wang, C. Wang, "Detect Artefacts of various Sizes on the Right Scale for each Class in Video Endoscopy", Proceedings of the Challenge on Endoscopy Artefacts Detection: Multi-class Artefact Detection in Video Endoscopy, Venice, Italy, 8 April 2019.

[13]. F. López, "Class Imbalance: Random Sampling and Data Augmentation with Imbalanced-Learn", https://towardsdatascience.com/class-imbalance-random-sampling-and-data-augmentation-with-imbalanced-learn-63f3a92ef04a (accessed: 2022)

[14]. P. Zhang, X. Li, Y. Zhong, "Ensemble Mask Aided R-CNN", Proceedings of the Challenge on Endoscopy Artefacts Detection: Multi-class Artefact Detection in Video Endoscopy, Venice, Italy, 8 April 2019.

[15]. Y. Y. Zhang, D. Xie, "Detection and Segmentation of Multi-class Artefacts in Endoscopy", Journal of Zhejiang University, Science-B, Vol. 20, No. 12, 2019, pp. 1014-1020.

[16]. Microsoft Common Object in Context dataset (MSCOCO) dataset, https://cocodataset.org (accessed: 2022)

[17]. H. M. Hung, P. T. D. Thinh, H. J. Yang, S. H. Kim, G. S. Lee, "Artefact Detection and Segmentation using Cascaded R-CNN & U-Net", Proceedings of the 2nd International Workshop and Challenge on Computer Vision in Endoscopy, Iowa City, USA, 3 April 2020, pp. 47-50.

[18]. H. Hu, Y. Guo, "Endoscopic Artefact Detection in MM Detection", Proceedings of the 2nd International Workshop and Challenge on Computer Vi-

sion in Endoscopy, Iowa City, IA, USA, 3 April 2020, pp. 78-79.

[19]. Z. Yu, Y. Guo, "Endoscopic Artefact Detection using Cascaded R-CNN Based Model", Proceedings of the 2nd International Workshop and Challenge on Computer Vision in Endoscopy, Iowa City, IA, USA, 3 April 2020, pp. 42-46.

[20]. X. Gao, B. Braden,"Artefact Detection and Segmentation Based on Deep Learning System", Proceedings of the 2nd International Workshop and Challenge on Computer Vision in Endoscopy, Iowa City, IA, USA, 3 April 2020, pp. 80-81.

[21]. A. Subramanian, K. Srinivasan, "Exploring Deep Learning-based Approaches for Endoscopic Artefact Detection and Segmentation", Proceedings of the 2nd International Workshop and Challenge on Computer Vision in Endoscopy, Iowa City, IA, USA, 3 April 2020, pp.51-56.

[22]. S. Ali et al. "Endoscopy Artifact Detection (EAD 2019) Challenge Dataset", 2019.

[23]. VGG Image Annotator (VIA), https: //www.robots.ox.ac.uk/~vgg/software/via/via.html (accessed: 2022)

[24]. Google Co-laboratory, https://colab.research.google.com (accessed: 2022)

[25]. ImageNet Dataset, https://www.image-net.org/ (accessed: 2022)

[26]. Facebook AI research, Detectron2 model zoo: https://github.com/facebookresearch/detectron2/blob/main/MODEL_ZOO.md (accessed: 2022)

[27]. Darknet, YOLOv4 model zoo: https://github.com/AlexeyAB/darknet/wiki/YOLOv4-model-zoo (accessed: 2022)

[28]. C. Shorten, T. M. Khoshgoftaar, "A Survey on Image Augmentation for Deep Learning," Journal of Big Data, Vol. 6, 2019. pp. 1-48.

[29]. Darknet, YOLOV3, https://pjreddie.com/darknet/yolo (accessed: 2022)

[30]. Darknet, YOLOv4, https://github.com/AlexeyAB/darknet (accessed: 2022)

[31]. Facebook AI research, Detectron2, https://github.com/facebookresearch/detectron2 (accessed: 2022)

[32]. S. Ali, F. Zhou, A. Bailey, B. Braden, J.E. East, X. Lu, J. Rittscher, "A Deep Learning Framework for Quality Assessment and Restoration in Video Endoscopy", Medical Image Analysis, Vol. 68,2021, pp.1-25.