**Endoscopic Surgical Operation and Object Detection using Custom Architecture Models**

S.Srividhya1[0009-0004-1947-1352], Jaladi Deepika1[0009-0002-4203-7103] , G.Narayanee Nimeshika1[0009-0004-9544-1016]

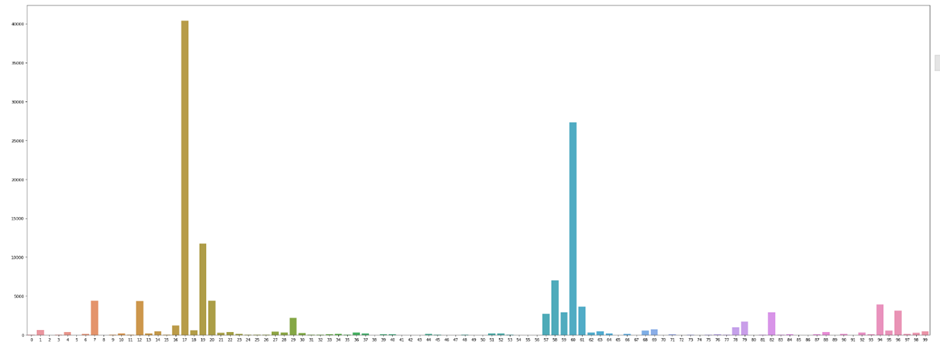
1Vellore Institute of Technology, Chennai  
 lncs@springer.com

**Abstract:** 1.2 million Americans have cholecystectomy each year. Surgeons need strong organ visualization and training to prevent injuring other organs. Advances in computer hardware, cameras, etc. could video the doctor's cholecystectomy surgery. Surgical triplets (instrument, verb, target) were annotated on 1 fps videos and 90489 frames. Each video frame's dataset includes <instrument, verb, target>. Thus, to ease cholecystectomy surgery, surgeons should be provided a deep learning model that processes real-time surgical video and displays the instrument, verb, target, and instrument on the video monitor. Thus, physicians need not address organ visibility. 5 CNN algorithms are compared in this study. Model performance is assessed using accuracy, precision, recall, F1 score, and hamming loss on the validation dataset. Custom architecture CNN, Resnet50, Vgg19, Vgg16, MobileNetv2, yolo model, F1 scores were 0.43, 0.75, 0.77, 0.8, 0.63, and 0.24 respectively. This study proposes YOLOv8-based bounding box medical picture categorization. X-rays, CT scans, and MRIs are reliably detected and classified using the suggested procedure. The advanced YOLOv8 model recognizes many items in real time. A convolutional neural network classifies the object of interest in the described approach. The solution outperforms state-of-the-art algorithms with 96.7% classification accuracy on a public medical image dataset.

**Keywords:** Cholecystectomy, CholecT45 dataset, Deep learning, Convolutional neural network, Resnet50, Vgg19, Vgg16, MobileNetv2, Yolov8

**1 INTRODUCTION**

Every year over 1.2 million people undergo this surgery in the United States alone, considering the world wide, there is a need for highly trained and experienced doctors for these patients. Cholecystectomy surgery at present is performed using a micro camera that would capture the real the video and display on the monitor and then the doctors would perform the surgery looking in the monitor. There might be times where there are less visibility /less experienced doctors who need to perform the surgery. Our proposed work would help these doctors to know the instrument being operated, verb/action being performed, target/organ which being handled Based on the results being displayed in real-time, the doctor can operate accordingly. The dataset is called Cholet 45 which is publicly available. The dataset consists of 45 videos which are used as training dataset while the 5 videos are used for validating Convolutional neural network models. Every image was captured at 1fps and they were manually annotated by the medical experts in the format of <instrument, verb, target> also called as triplet annotation. Each Image that was captured has a size of 774-pixel width and 434-pixel height. There are around 100 different categories/classes of triplets for the overall dataset, every frame of the video may belong to one or more triplet classes as the images were extracted at 1 fps, it is possible. There is also the availability of only instrument being annotated, only verb/action, only target for the entire dataset but our proposed work uses the triplet annotation instead of using the instrument, verb and target individually. There are around 6 classes for instruments,9 classes for verb and 14 classes for target for the dataset.



**Figure 1.**Data distribution of triplet classes

Figure 1 depicts the number of images that belong to each triplet class. The total number of images tallies up to 90489 Triplet Class 17 (grasper, retract, gallbladder) accounts for more 40,000 images. Our proposed work mainly focuses on the comparison of the performance of the different convolutional neural networks. In comparison with custom architecture CNN (Convolutional neural network), Resnet50, Vgg19, Vgg16, MobileNetv2, Vgg16 was able to extract features from images in better way than rest and also the validation accuracy of Vgg16 was higher than rest other models. The validation accuracy of CNN (Convolutional neural network), Resnet50, Vgg19, Vgg16, MobileNetv2, Yolo V8 are 58%, 72%, 73%, 75%, 61% and 90% respectively.

YOLO (You Only Look Once) is the first object detection model to combine object categorization and bounding box prediction into a single end-to-end differentiable network. It was developed and is kept up using the Darknet system. The first YOLO model to be created using the PyTorch framework is YOLOv5, which is substantially smaller and simpler to use. However, YOLO v6, YOLOv7 have been developed, YOLO v8 gives the highest accuracy among its predecessors. YOLO v8 also uses PyTorch Framework. YOLO v8 takes almost 60% time to train while producing outcomes with higher mean average precision. Here, the issue of prolonged training is somewhat addressed. The trade-off between training time and precision is achieved more in v8. Hence, we have used Yolo v8 in our project.

**1.1 Paper Organization:**

The rest of the paper is structured as follows: In Section 2, problem statements and related work in the area is presented. Section 3, introduces the dataset and the classification using various Algorithms such as Section 4, presents the results, graphs and conclusion drawn which includes the performance analysis. Section 5, includes future work and Section 6 and 7, concludes the paper by acknowledging and citing the references respectively.

**2 WORK**

**2.1 Problem Statement:**

To define surgical operations as triplets of "instruments, verb, and target" where "instrument" indicates the tool used, "verb" to the action taken, and "target" to the underlying anatomy or things acted upon. to find the coordinates of the bounding box, and then employ Yolov8 to determine the object.

**2.2 Related Works:**

**[1]** This study proposed a laparoscopic cholecystectomy (LC) video recognition AI. Two export surgeons divided and cut videos from five hospitals. The export surgeons marked videos with significant bleeding, gallbladder perforation, major bile leakage, incidental finding, and complexity degree (1-5). The dataset was split between 80% training and 20% validation to train the model. Surgical phase recognition averaged 89%. The model was 92% accurate for complexity level 1 and 81% for level 5.

**[2]** Time series information improved operative phase detection in laparoscopic cholecystectomy. The author wanted to improve CNN models for laparoscopic cholecystectomy surgery phase detection. Six surgical stages were created from 115 laparoscopic cholecystectomy recordings.3fps photos were used. Three medical specialists annotated and labelled the dataset. EfficientNet, Sharpness-Aware Minimization optimizer optimized learning parameter accuracy, precision, and recall to 0.97, 0.85, and 0.86.

**[3]** Used a CNN architecture called EndoNet for the surgical phase recognition in laparoscopic cholecystectomy. The overall accuracy achieved was 82% and it is the first paper that was published on multiple recognition tasks on laparoscopic cholecystectomy videos. Authors used Cholec80 and Endovis dataset for the purpose of training the recognition model. A LSTM model was developed by the authors to estimate the surgical phase that resulted in 96% of accuracy.

**[4]** Due to low visibility sometimes doctors interpret the wrong perception during surgery and thus the chance of adverse effect increase .Author used a different deep learning model to provide a real time assistance to doctors which would identify the safe zone to operate and caution zone to operate such as dissection, liver, gallbladder, and hepatocystic triangle in real time .The model was trained using laparoscopic cholecystectomy dataset ,the annotations of the dataset was done by surgeons .The F1 score achieved by the author for safe zone was 0.70 (±0.28) and for no zone it was 0.83 (±0.31)

**[5]** During laparoscopic cholecystectomy there is always a chance of an error that can be made by surgeon and thus bile duct injury is possible which is usually around 0.2-1.5%. To address this real world problem, authors have developed a artificial intelligence that would guide the doctors during surgery to avoid bile duct injury. Authors have used YOLOv3 model to detect the cystic duct, common bile duct, lower edge of the left medial liver segment, and Rouviere's sulcus. Authors used 41 videos to train the model The mean average precision obtained was 0.710.

**[6]** The majority of object detection models in medical images are designed to detect a singular object. This paper proposes a deep learning solution for the object detection problem in three-dimensional medical images, i.e., the localization and classification of multiple structures. For supervised learning methods, large annotated datasets are typically difficult to acquire. Consequently, we developed a combined Cycle Generative Adversarial Network (CycleGAN) and You Only Look Once (YOLO) method for data augmentation from one modality to another using CycleGAN and organ detection from generated images using YOLO.

**[7]** Due to a lack of adherence to the social distance and mask-wearing regulations in crowded places like hospitals, schools, and malls where people must congregate, the impact of Covid 19 cases is growing globally. Although the authorities have taken several preventative measures to ensure that masks are used, it can be difficult to inspect masks in crowded situations. Visual inspections can miss people who are not wearing masks, which is a crucial role in the spread of the disease. The goal of this work is to develop a mask inspection system based on artificial intelligence (AI) using the YOLO V7 deep learning technique to guard against the Covid-19 epidemic in crowded public spaces.

**3 PRINCIPLE AND METHODOLOGY**

In this paper, we have used custom architecture CNN (convolutional neural network), Resnet50, Vgg19, Vgg16, and MobileNetv2 models to recognize the triplet class<instrument, verb, target>(Multi label classification) by analyzing the video frames. The block diagram Fig-2 represents the architecture of the custom-made CNN Model. The custom-made model comprises 3 convolutional layers accompanied by max pooling layer and dropouts. Dense layers to predict weather the triplet class.

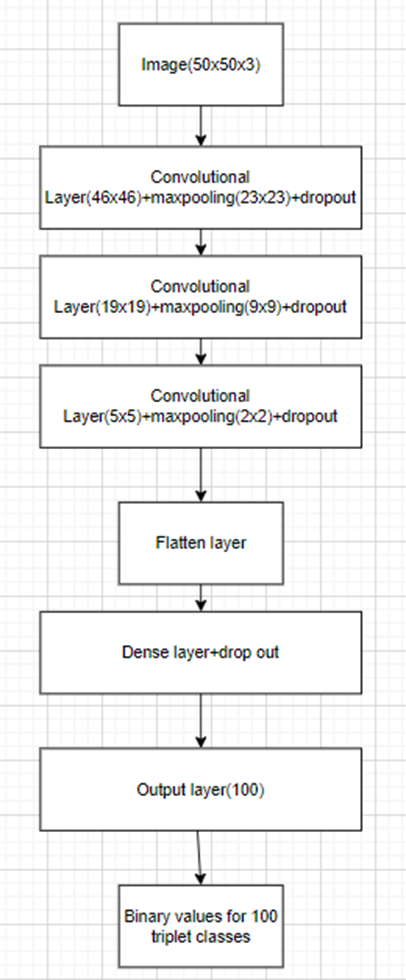
****

Fig -2 architecture of custom-made model

**3.1 Transfer learning**

Transfer learning uses a model built on a large dataset to extract/recognize features from the current dataset. Since transfer learning models have already been trained on a large dataset, we can use the model's weights by keeping the early/top layer non-trainable when training for our dataset. Keeping transfer learning model top layers non-trainable saves time and computation. This paper uses Resnet-50, Vgg19, Vgg16, MobileNetv2, and YoloV8.

**3.1.1 Resnet-50**

Resnet-50 features 50 deep layers of transfer learning. Millions of ImageNet photos trained the model. Resnet-34 had 34 layers. The Resnet architecture is founded on two principles: (i) the number of filters in each layer is the same regardless of the size of the output feature map, and (ii) if the feature map is halved, it has double the filters to preserve layer time complexity. Resnet-50 has 48 convolutional layers and 1 max and 1 average pooling layer.

**3.1.2 VGG19**

Vgg19 includes 19 deep layers of transfer learning. The model was trained on millions of ImageNet photos. 16 convolutional layers, 5max pooling layers, and 3 fully connected layers make up Vgg19.Vgg19 was the transfer learning model that followed Alex Net, which improved the convolutional neural network. In the Vgg19 convolutional layer, the kernel is (3\*3) and stride is 1 pixel. Max pooling layer has (2\*2) window with stride value 2.

**3.1.3 VGG16**

16-convolutional layer transfer learning model Vgg16. Convolutional layer kernels have several filters (3\*3). The first two convolutional layers have 64 filters, and the feature map from the second layer is max pooled.3rd and 4th convolutional layers have 128 filters, and the feature map from the 4th layer is max pooled. The feature map from the 7th convolutional layer has a max pooling layer and 256 filters. Convolutional layers 8–13 have 512 filters and a max pooling layer on feature maps from the 13th layer.

**3.1.4 MobileNetV2**

It is a transfer learning convolutional neural network that performs well on modern mobile devices. There are two types of blocks used in mobile netv2, the first one is the residual block with stride as 1 and another block with stride as 2 for downsizing the feature maps. There are 3 types of layers used in both of these blocks. The first layer uses a filter of (1\*1) filter with Relu6 activation. The second layer uses a depth wise convolution while the third layer uses a convolutional layer without any non-linearity.

**3.1.5 YoloV8**

YOLOv8's bounding box medical image classification model recognizes and classifies objects. The cutting-edge YOLOv8 neural network predicts multiple bounding boxes and class probabilities. Darknet-53 is YOLOv8's feature extractor and object detection layers. Using a huge medical image dataset, the YOLOv8 model detects tumors, lesions, and fractures. The identified object classifies photos using a CNN. The CNN is trained on medical images labelled by class. CNNs classify object features. Workflow of the suggested model. YOLOv8 extracts bounding box coordinates from medical images. Bounding box coordinates cut out the detected object. CNN classifies cropped images. CNN categorizes. The model produces the detected object's class label and bounding box coordinates from the input image.

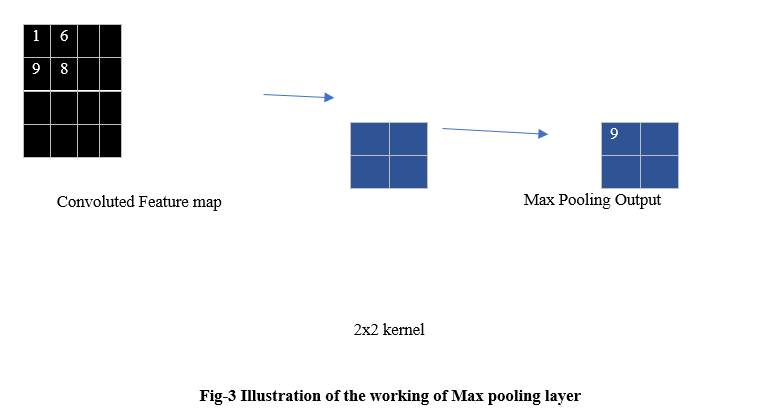
**3.2 Working principle of Convolutional neural network**

After collecting the dataset, every image is normalized and loaded into the variable. Forward propagation passes normalized images through the convolutional neural network in smaller batches. The convolutional layers extracted features and created a feature map from the images, while the pooling layers extracted the critical features. The final layers, called dense layers, determine the image's triplet class and calculate loss and accuracy. The last layer of the multilabel classification comprises 100 decision nodes with sigmoid activation functions that output prediction values between 0 and 1 for all 100 classes. Backpropagation is used to adjust inner layer parameters like weights to reduce loss and improve accuracy in the following training session based on the projected value.

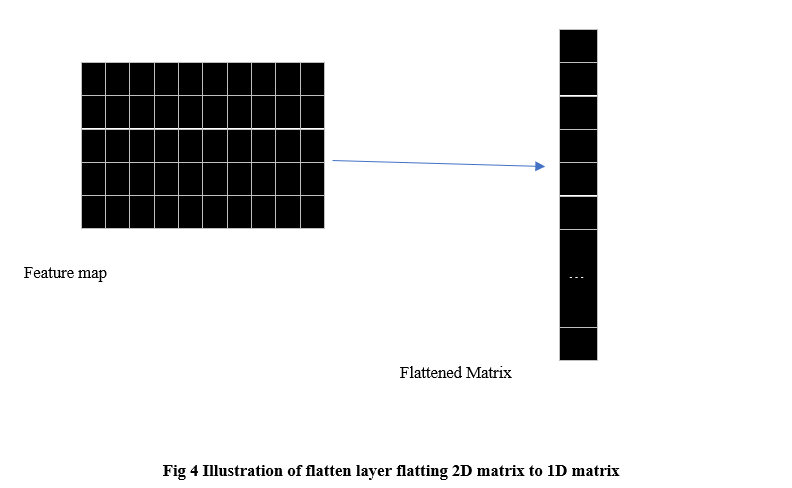
**3.2.1 CNN Architecture**

*Convolution layer:* This layer extracts essential features from pictures using kernel matrix/weighted matrix to build feature map. A hyperparameter-set convolutional layer generates a feature map. Backpropagation updates kernel matrix/weight matrix values for image feature extraction.

*Max pooling layer:* After extracting features from convolution layer, feature map illustrated in Fig 3 is sent through the max pooling layer, where a filter matrix is applied across the entire feature map to retrieve the maximum values.



*Flatten layer:*This is the layer that converts the two-dimensional matrix into single dimension so that the input can be feed to the dense layer



*Dense layer:* It is a type of layer where neurons represented in Fig 5 from the current layer get input from the previous layers. The output of each neuron depends on activation function used weights, input and bias.

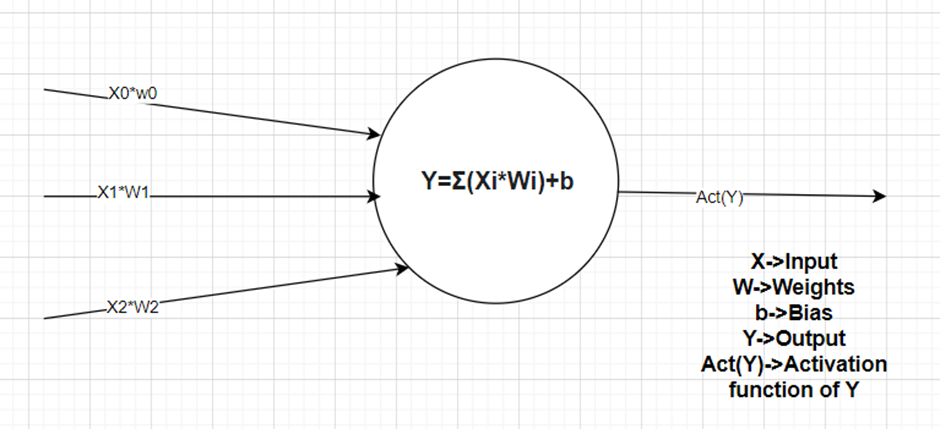


Fig 5 Illustration of Neuron of Dense layer

**4 RESULTS AND CONCLUSION**

The training dataset features 90489 frames taken from 45 endoscopic surgical recordings, and the validation dataset has 1319 frames from 5 videos. 100 multilabel triplet classes exist. F1score, accuracy, recall, precision, and Hamming loss assess model performance.

**4.1 Custom architecture model**

Fig 6 shows validation loss versus epochs. 90,489 pictures and 1319 validation images trained the model. Loss and validation loss for each epoch were plotted.

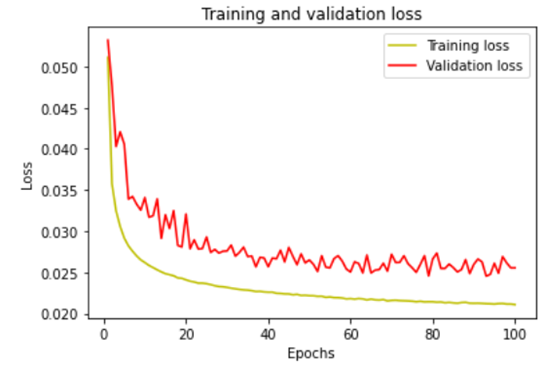
****

Fig 6 Loss vs. epochs graph for custom architecture model after training 100 epochs

Fig 7 shows the accuracy and validation accuracy of the custom architecture model trained on 90,489 photos and validated on 1319 images. Accuracy and validation accuracy for each epoch were plotted.

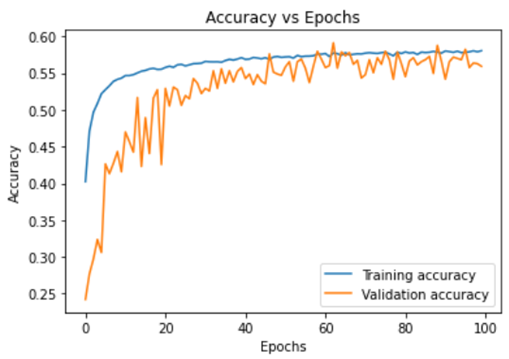


Fig 7 Accuracy vs. epochs graph for custom architecture model after training 100 epochs

The F1 score, recall, and precision obtained by the custom architecture mod 0.43, 0.44, and 3,0.44,0.45 respectively.

**4.2 Resnet-50**

Fig. 8 shows validation loss versus epochs. 90,489 pictures and 1319 validation images trained the model.Loss and validation loss for each epoch were plotted.

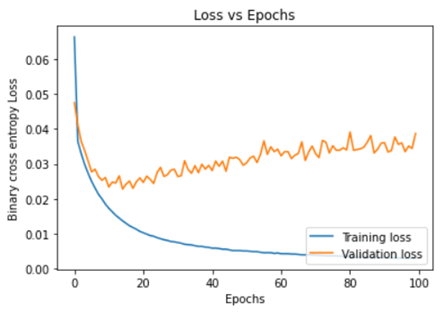


Fig 8 Loss vs epochs graph for Resnet 50 model after training 100 epochs

Fig 9 shows the Resnet50 model's training and validation accuracy. Accuracy and validation accuracy for each epoch were plotted.

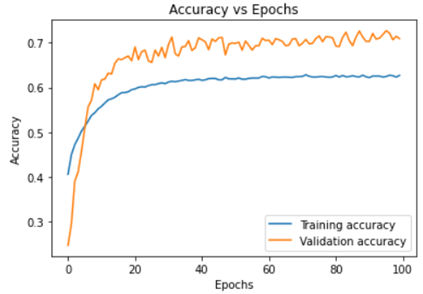


Fig 9 Accuracy vs epochs graph for Resnet50 model after training 100 epochs

The F1 score, recall and precision (avg=samples) obtained by the Resnet50 model are 0.75, 0.85, 0.71 for the validation dataset.

**4.3 VGG19**

Fig. 10 shows validation loss versus epochs. 90,489 pictures and 1319 validation images trained the model. Loss and validation loss for each epoch were plotted.

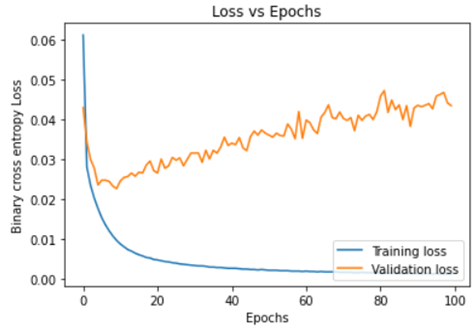


Fig 10 Loss vs epochs graph for VGG19 model after training 100 epochs

Fig 11 shows the VGG19 model's training and validation accuracy. Accuracy and validation accuracy for each epoch were plotted.

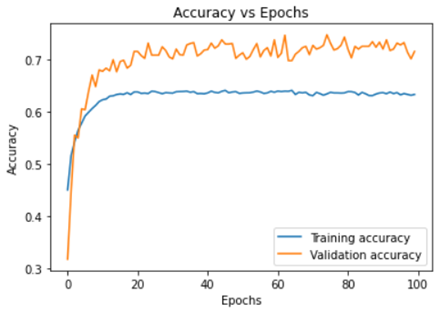


Fig 11 Accuracy vs epochs graph for VGG19 model after training 100 epochs

The F1 Score, recall and precision obtained by VGG19 model are 0.77,0.86,0.73 for the validation dataset.

**4.4 VGG16**

Fig 12 shows validation loss versus epochs. 90,489 pictures and 1319 validation images trained the model. Loss and validation loss for each epoch were plotted.

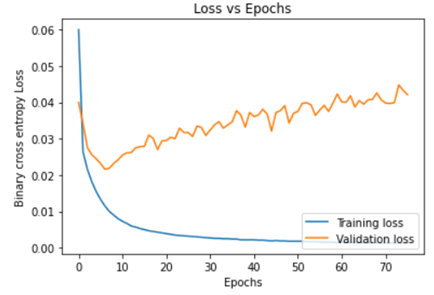


Fig 12 Loss vs epochs graph for VGG16 model after training 75 epochs

The VGG16 model was trained on 90,489 images and validated on 1319 photos. Fig. 13 shows its accuracy and validation accuracy. Accuracy and validation accuracy for each epoch were plotted.

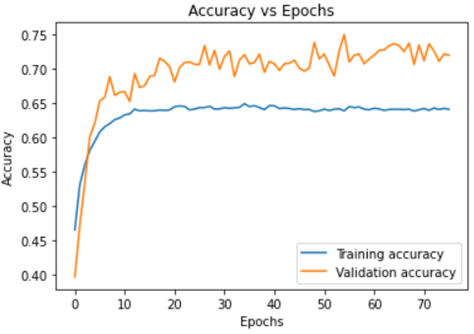


Fig 13 Accuracy vs. epochs graph for the VGG16 model after training 75 epochs

The F1 score, recall and precision (avg=samples) obtained by the VGG16 model were 0.8, 0.89, and 0.760.76 for the validation dataset.

**4.5 MobilenetV2**

Fig 14 shows validation loss versus epochs. 90,489 pictures and 1319 validation images trained the model. Loss and validation loss for each epoch were plotted.

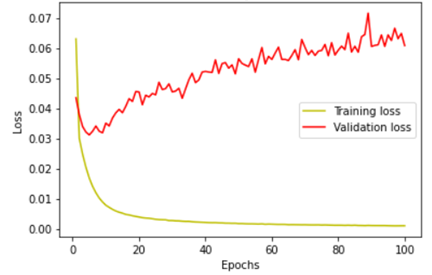


Fig 14 Loss vs epochs graph for MobilenetV2 model after training 100 epochs

The MobilenetV2 model was trained on 90,489 photos and validated on 1319 images. Fig. 15 shows its accuracy and validation accuracy. Accuracy and validation accuracy for each epoch were plotted.

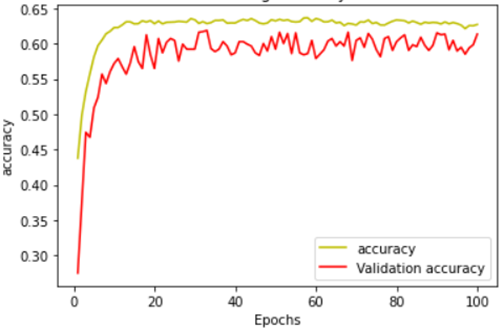


Fig 15 Accuracy vs. epochs graph for MobilenetV2 model after training 100 epochs

The F1 score, recall, and precision obtained by the MobilenetV2 model are 0.63, 0.69, and 0.62 respectively.

**4.6 Yolo V8**

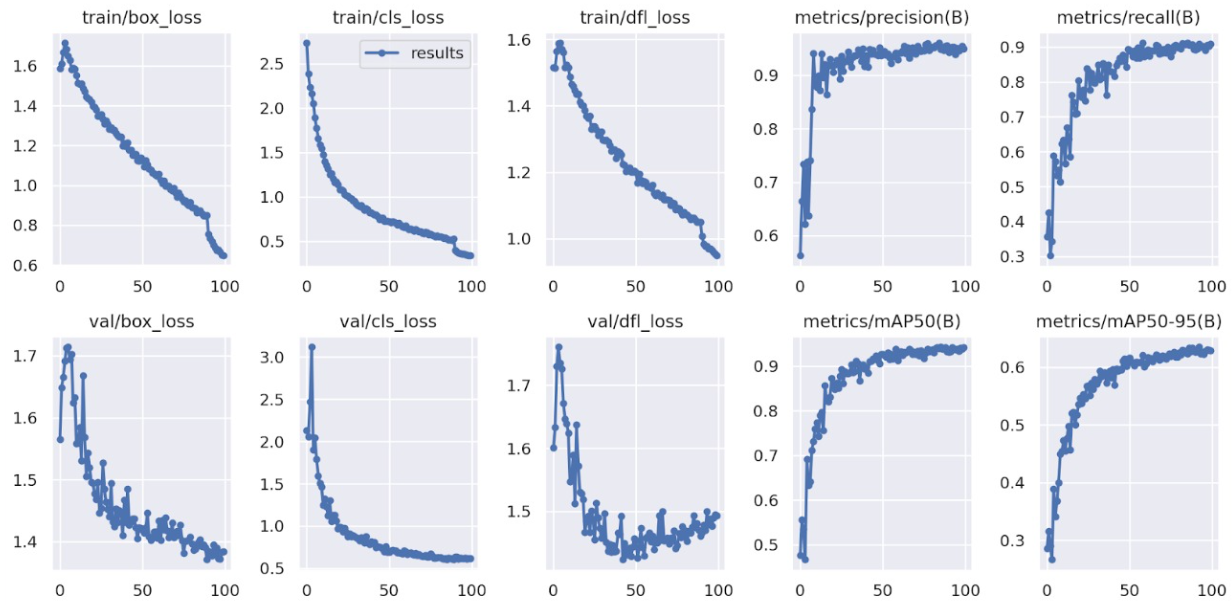
****

Fig 16 represents all the graphs for the trained 100 epochs

The metrics such as recall, precision and loss have been shown in the Fig 16 for the trained 100 epochs. The F1 score for this model is determined as 0.24 with 0.16 and 0.51 as recall and precision value respectively.

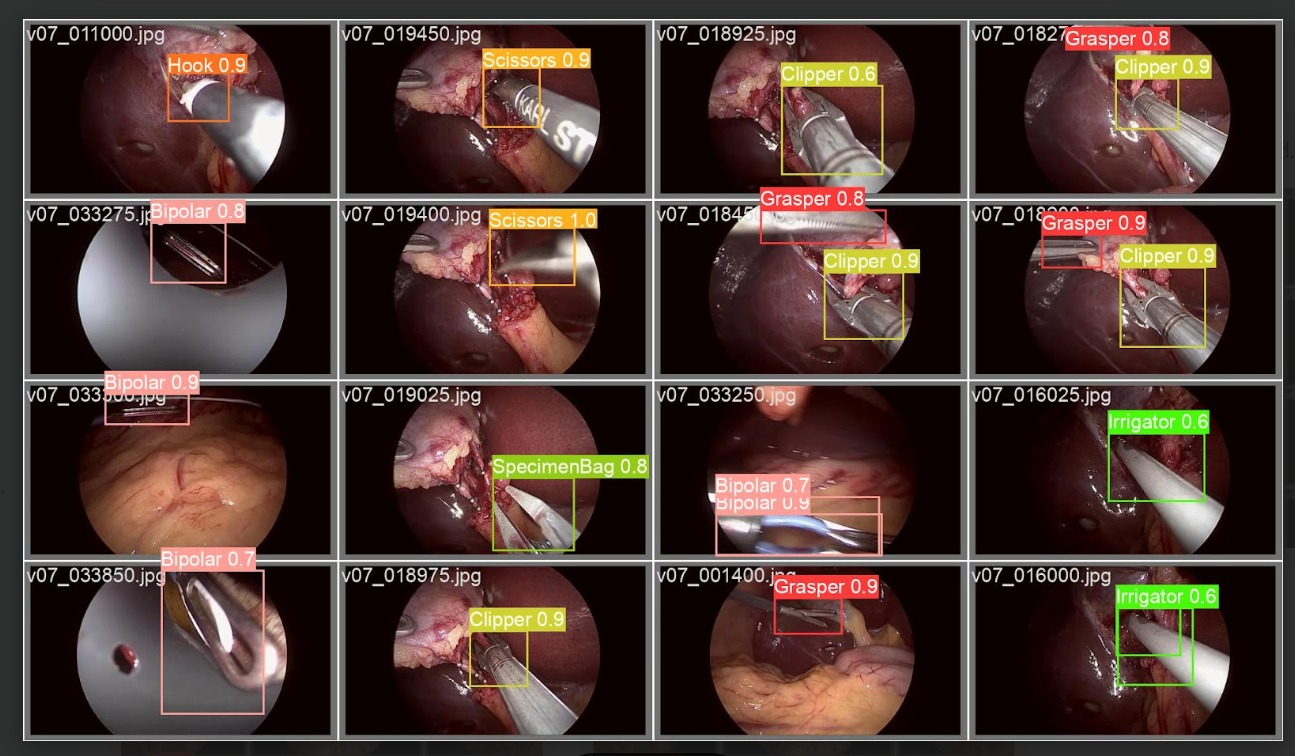


Fig 17

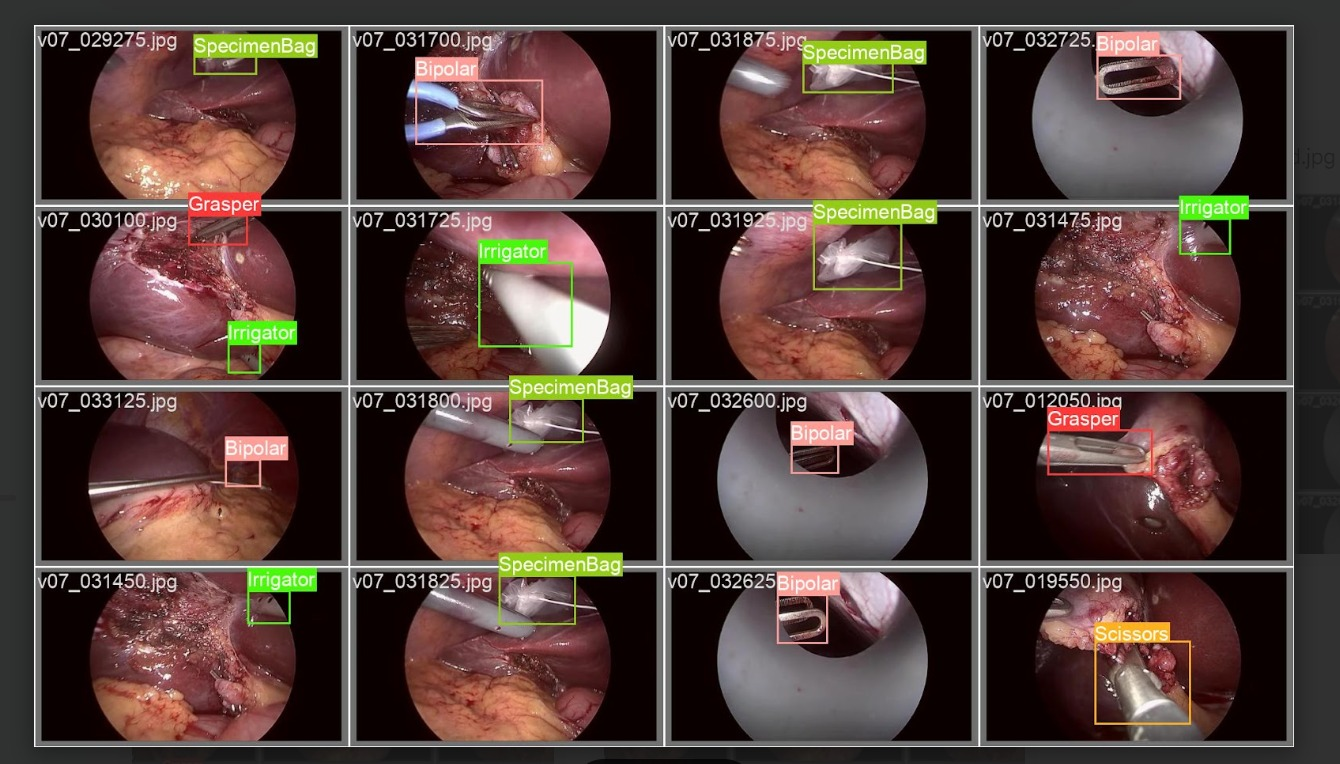


Fig 18

Fig 17 and 18 show the images with the identified instruments with labels such as hook, irrigator, specimen bag etc.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Metrics** | **Custom architecture model** | **Resnet50** | **VGG19** | **VGG16** | **MobilenetV2** | **YoloV8** |
| F1 score | 0.43 | 0.75 | 0.77 | 0.8 | 0.63 | 0.24 |
| Recall | 0.44 | 0.85 | 0.86 | 0.89 | 0.69 | 0.16 |
| Precision | 0.45 | 0.71 | 0.73 | 0.76 | 0.62 | 0.51 |
| Exact match ratio | 0.36 | 0.545 | 0.5815 | 0.597 | 0.466 | 0.323 |
| Hamming loss | 0.0089 | 0.006 | 0.0054 | 0.005 | 0.007 | 0.0034 |

Table 1 Illustration of the performance of the CNN model on the validation dataset

The above table depicts the overall performance of the CNN models on the validation dataset. The conclusion we can make from looking at the values obtained, VGG16 outperformed the rest of the other models.

**5 FUTURE WORK**

This article compared CNN models on the cholecT50 dataset. There are many picture pre-processing methods to highlight/enhance characteristics and compare the performance of different algorithms to determine the triplet class. As a highly imbalanced multilabel dataset, future work can focus on class weights/a strategy to balance a class or a custom architectural loss function to reduce class imbalance.

*Multi-organ detection:* YOLO model can detect various organs. Modifying the model's loss function to support several classes allows this.

*Improved Accuracy:* Researchers can try multiple architectures and tweaks to improve organ detection model performance.

*Fine-grained detection:* YOLO currently recognizes organs as one object. However, a finer-grained technique can detect blood vessels and tumors in organs.

*Data augmentation:* Enlarging and diversifying the training dataset can improve model performance. Researchers can generate new training data using rotation, translation, and scaling.

*Transfer learning:* pre-trained models can initialize the YOLO model's weights, speeding up training and improving performance. Pre-training can enhance model accuracy.

3D organ detection: YOLO models currently analyze 2D images. These models can be extended to accommodate 3D pictures from CT or MRI scans, which could aid medical diagnosis.

*Real-time organ detection:* YOLO model can be optimized for medical applications like robotic surgery and endoscopy.

**6 ACKNOWLEDGEMENT**

We wish to express our sincere thanks and deep sense of gratitude to our project guide, Dr. Vallidevi K., for her consistent encouragement and valuable guidance offered to us in a pleasant manner throughout the course of the project work. We are extremely grateful to Dr. Ganesan R., Dean of the School of Computer Science Engineering, VIT Chennai, for extending the facilities of the school towards our project and for his unstinting support. We express our thanks to our Head of Department, Dr. Maheswari R., for her support throughout the course of this project. We also take this opportunity to thank all the faculty of the school for their support and the wisdom they imparted to us throughout the course. We thank our parents, family, and friends for bearing with us throughout the course of our project and for the opportunity they provided us to undergo this course at such a prestigious institution.

**7 REFERENCES**

[1] Maryam Hammami; Denis Friboulet; Razmig Kechichian, Data augmentation for multiorgan detection in medical images, 14 December,2020

[2] Maryam Hammami , Denis Friboulet and Razmig Kechichian, Cycle Gan-Based Data Augmentation For Multi-Organ Detection In Ct Images Via Yolo

[3] Yan Wang, Qiyuan Sun, Guodong Sun, Lin Gu, Zhenzhong Liu, Object Detection of Surgical Instruments Based on YOLOv4, 03-05 July, 2021

[4] Ade Iriani Sapitri, Siti Nurmaini, Muhammad Naufal Rachmatullah, Bambang Tutuko, Annisa Darmawahyuni, Firdaus Firdaus, Dian Palupi Rini, Anggun Islami, Deep learning based real time detection for cardiac objects with fetal ultrasound video

[5] Yunus EGl, YOLO V7 and Computer Vision-Based Mask-Wearing Warning System for Congested Public Areas

[6] Golany, T., Aides, A., Freedman, D., Rabani, N., Liu, Y., Rivlin, E., & Reissman, P., 2022, Artificial intelligence for phase recognition in complex laparoscopic cholecystectomy. Surgical Endoscopy.

[7] Shinozuka, K. I., Turuda, S., Fujinaga, A., Nakanuma, H., Kawamura, M., Matsunobu, Y., & Tokuyasu, T., 2022, Artificial intelligence software available for medical devices: Surgical phase recognition in laparoscopic cholecystectomy. Surgical Endoscopy.

[8] Twinanda, A. P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., & Padoy, N., 2016, EndoNet: a deep architecture for recognition tasks on laparoscopic videos. IEEE Transactions on Medical Imaging.

[9] Madani, A., Namazi, B., Altieri, M. S., Hashimoto, D. A., Rivera, A. M., Pucher, P. H., & Alseidi, A., 2022, Artificial intelligence for intraoperative guidance: using semantic segmentation to identify surgical anatomy during laparoscopic cholecystectomy. Annals of surgery.

[10] Tokuyasu, T., Iwashita, Y., Matsunobu, Y., Kamiyama, T., Ishikake, M., Sakaguchi, S., & Inomata, M., 2021, Development of an artificial intelligence system using deep learning to indicate anatomical landmarks during laparoscopic cholecystectomy. Surgical endoscopy.