


RESEARCH

Open Access



The impact of ensemble learning on surgical tools classification during laparoscopic cholecystectomy

Jaafar Jaafari^{1*} , Samira Douzi^{2*}, Khadija Douzi^{1*} and Badr Hssina^{1*}

*Correspondence:

jaafarjaafari@etu.fstm.ac.ma;

s.douzi@um5r.ac.ma;

khadija.douzi@fstm.ac.ma;

badr.hssina@fstm.ac.ma

¹ FSTM, University Hassan II,

Casablanca, Morocco

² FMPR, University

Mohammed V, Rabat,

Morocco

Abstract

Laparoscopic surgery also known as minimally invasive surgery (MIS), is a type of surgical procedure that allows a surgeon to examine the organs inside of the abdomen without having to make large incisions in the skin. It unifies the competence and skills of highly trained surgeons with the power and precision of machines. Furthermore, surgical instruments are inserted through the abdomen with the help of a laparoscope, which is a tube with a high-intensity light and a high-resolution camera at the end. In addition, recorded videos from this type of surgery have become a steadily more important information source. However, MIS videos are often very long, thereby, navigating through these videos is time and effort consuming. The automatic identification of tool presence in laparoscopic videos leads to detecting what tools are used at each time in surgery and helps in the automatic recognition of surgical workflow. The aim of this paper is to predict surgical tools from laparoscopic videos using three states of the arts CNNs, namely: VGG19, Inception v-4, and NASNet-A. In addition, an ensemble learning method is proposed, combining the three CNNs, to solve the tool presence detection problem as a multi-label classification problem. The proposed methods are evaluated on a dataset of 80 cholecystectomy videos (Cholec80 dataset). The results present an improvement of approximately 6.19% and a mean average precision of 97.84% when the ensemble learning method is applied.

Keywords: Laparoscopic surgery, Computer vision, Convolutional neural network, Ensemble learning, Transfer learning

Introduction

Improving recovery for patients through reducing surgical trauma is the key objective of minimally invasive surgery. In traditional open surgery, the surgeon makes one large cut to see the part of your body that they're operating on, causing pain, organ dysfunction, catabolism, fluid/salt retention, and sleep disturbances [1]. Thus, when the laparoscopic revolution began in the early 1990s, surgeons were immediately struck by how much better their patients looked after laparoscopy compared to open cholecystectomy. Patients undergoing MIS benefits from myriad advantages, including smaller incisions, less pain, minimal to no scars, low risk of infection, less blood loss, lower rate of complication, and shorter hospital stay. Minimally invasive surgery

refers to any surgical procedure that is performed through tiny incisions instead of a large opening.

Furthermore, there are three major types of MIS:

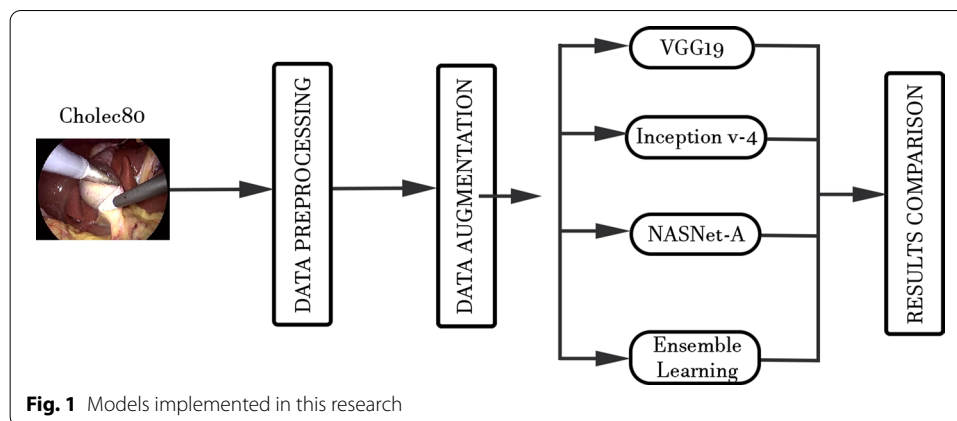
- Endoscopy: The surgeon uses the endoscope itself to do the procedure. The endoscope goes in through the body's natural openings, without the surgeon making any cuts.
- Robot-assisted surgery (robotic surgery): The surgeon makes several small cuts to guide the endoscope and robotic tools into the body. From there, the surgeon controls the surgery while sitting nearby a computer console.
- Laparoscopy: The surgeon uses a laparoscope, which is a thin tube fixed with a light and camera and several other small surgical instruments.

During laparoscopic surgery, the surgeon makes several small incisions in the skin, typically a few millimeters long. Furthermore, a laparoscope (a long thin tube fixed with a camera and light) is inserted into one of the incisions. Moreover, images from the endoscope are shown on monitors in the operating room so surgeons can get a clear and magnified view of the surgical area. Finally, special surgical tools or instruments are inserted through other small incisions. The surgeon uses these to explore, remove, or repair a problem inside the body.

On the other hand, it is well known that junior surgeons are not allowed to operate on real patients. Therefore, surgery videos are a precious tool to learn procedures [2]. Additionally, those videos are widely used in the teaching field, also, there are plenty of online platforms that offers learning surgery procedures, including YouTube. In addition, the video recorded during the surgery offers limitless possibilities. In some countries, it is mandatory to store the surgery as evidence [3]. Moreover, human error and negative events can go unnoticed during the surgery, that's why Surgical Quality Assessment (SQA) using video recorded surgery is essential. SQA is an internal verification approach, to evaluate the surgeons via the surgery videos, and give a review based on some standardized rating checklist. However, MIS videos tend to last several hours. Thus, navigation and searching through these videos are cumbersome and time-effort consuming.

In this work, we tackle these problems by investigating the detection and the classification of surgical tools in laparoscopic surgery. To achieve this goal, we propose numerous models (Fig. 1) based on convolutional neural networks, namely: VGGNET [4], Inception v-4 [5], and NASNet-A [6]. In addition, we combine these three state-of-the-art neural networks to study and investigate the impact of ensemble learning on deep learning models applied to laparoscopic surgery. Therefore, after training each of the three networks, we use average ensembling technique on the models to avoid over fitting. Furthermore, the multilabel classification methods for surgical tool presence detection were tested on the benchmark dataset Cholec80.

The rest of this paper is organized as follows: in “[Related works](#)” section, we present the related work; in “[Background and methods](#)” section we present the material and methods; in “[Experiments and results](#)” section, we provide the experiments and



results; and in “[Conclusion and future works](#)” section we draw our conclusions and future works.

Related works

Recently, deep learning models have been performing very well and reached exceptional results in different medical image analysis tasks. In particular, convolutional neural networks (CNN) have become the standard approach to handle this kind of problem [7–11]. Furthermore, in the literature, numerous deep learning approaches have been proposed for automated laparoscopic surgical tools classification and segmentation. Pan Shi [12] proposed an attention-guided convolutional neural network (CNN) for frame-by-frame detection of surgical tools in MIS videos, which comprises a coarse (CDM) and a refined (RDM) detection module. The proposed method was tested on two public datasets (EndoVis Challenge and ATLAS Dione). Sheng Wang [13] implemented a deep learning based multilabel classification method for surgical tool presence detection in laparoscopic videos. Moreover, it combines two state-of-the-art deep neural networks (VGGNet and Google Net). This method was tested on M2CAI surgical tool presence detection challenge. Kletz [14] evaluated the achievable performance of segmenting surgical instruments from their background by using a region-based fully convolutional network, for instance-aware instrument segmentation as well as instrument recognition. Jalal [15] designed a deep learning pipeline, namely, a convolutional neural network (CNN) and a nonlinear autoregressive network with exogenous input (NARX), designed to predict surgical phases from laparoscopic videos. A convolutional neural network (CNN) is used to perform the tool classification task by automatically learning visual features from laparoscopic videos. The output of the CNN, which represents binary usage signals of surgical tools, is provided to a NARX neural network that performs multistep-ahead predictions of surgical phases. Wang [16] created a framework for one-stage object detection based on a sample adaptive process controlled by reinforcement learning. Zhang [17] assembled a Modulated Anchoring Network for the detection of laparoscopic surgery tools based on Faster R-CNN. This approach was tested on the m2cai16-tool-locations dataset. Namazi [18] made a multilabel classifier, called LapTool-Net to detect the presence of surgical tools in each frame of a laparoscopic video, based

on a Recurrent Convolutional Neural Network (RCNN) architecture to simultaneously extract the spatiotemporal features. To overcome the high imbalance and avoid overfitting caused by the lack of variety in the training data, a high down-sampling rate is chosen based on the more frequent combinations. Chittajallu [19] presented a self-supervised method for instrument segmentation using the kinematic model of the robot as a source of information. The authors implemented a Fully Convolutional Neural network (FCN) on VIVO dataset obtained from a robotized endoscopy system, resulting an average precision of 91%. Kletz [20] developed a Real-Time Instrument Segmentation tool in robot-assisted minimally invasive surgery (RMIS) using multiresolution Feature Fusion (MFF) block and a light-weight CNN to identify the surgical tool. The dataset used by the author is MICCAI 2017. Shvets [21] presented deep learning-based solution for the robotic instrument semantic segmentation using U-NET, TerausNet, and LinkNet. This approach was tested on MICCAI 2017 Endoscopic Vision SubChallenge: Robotic Instrument Segmentation dataset. Kanakatte [22] proposed a deep neural network that combines the advantage of spatial representation using CNN and temporal information using LSTM. It is an instance segmentation algorithm, which segments and localizes the surgical tool using a spatiotemporal deep network, using a pre-trained ResNet and VGGNet. Lshirbaji [23] proposed a deep learning-based approach to detect surgical tools in laparoscopic images using a convolutional neural network (VGG16) in combination with two long short-term memory (LSTM) models. Furthermore, a pre-trained CNN model was trained to learn visual features from images. Then, LSTM was employed to include temporal information through a video clip of neighbour frames. Finally, the second LSTM was utilized to model temporal dependencies across the whole surgical video. Experimental evaluation has been conducted with the Cholec80 dataset to validate our approach, resulting an average precision of 91.04%.

Background and methods

Machine learning and deep learning

Machine learning (ML) is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Moreover, ML focuses on the development of a self-adaptive algorithm that gets increasingly better analysis and patterns with experience or with newly added data.

On the other hand, deep learning (DL) is a subfield of machine learning that attempt to simulate the behavior of the human brain in processing data and creating patterns using artificial neural networks. Furthermore, it utilizes a hierarchical level of artificial neural networks to carry out the process of machine learning. The artificial neural networks are built like the human brain, with neuron nodes connected together like the web. While traditional programs build analysis of data in a linear way, the hierarchical function of deep learning systems enables machines to process data with a nonlinear approach. Deep learning drives many artificial intelligence (AI) applications and services that improve automation, performing analytical and physical tasks without human intervention.

The main differences between ML and DL are [24]:

- Feature extraction (Fig. 2): A machine learning workflow starts with relevant features being manually extracted from images. The features are then used to create a model that classifies the objects in the image. However, with a deep learning workflow, relevant features are automatically extracted from images.
- Data size matter: Deep learning algorithms scale with data which means that they often continue to improve as the size of your data rises. On the other hand, most shallow learning (machine learning) methods stop improving the accuracy at a certain level of performance when you add more examples and training data to the network.
- Input data: Machine learning algorithms almost always require structured data, while deep learning networks rely on layers of ANN (artificial neural networks).

In deep learning, a computer model learns to perform classification tasks directly from images, text, or sound. Deep learning models can achieve state-of-the-art accuracy, sometimes exceeding human-level performance. Models are trained by using a large set of labeled data and neural network architectures that contain many layers.

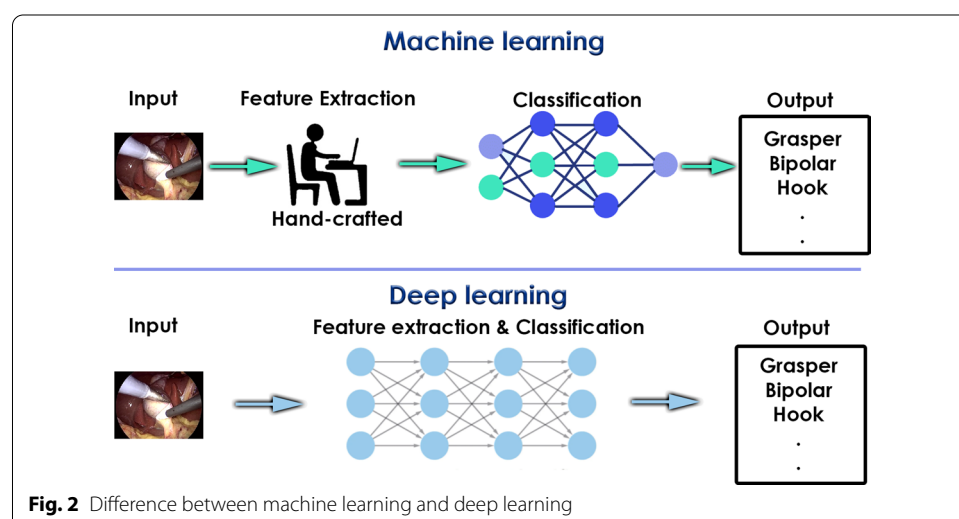
While deep learning was first theorized in the 1980s, there are two reasons it has only recently become useful:

- Deep learning algorithm requires a large amount of labeled data [25].
- Deep learning algorithm requires substantial computing power [26].

One of the most popular types of deep neural networks is known as convolutional neural networks (CNN or ConvNet).

Computer vision and convolutional neural networks

The internet is comprised of text and images. It is relatively straightforward to index and search text, but to index and search images, algorithms need to know what the



images contain. To get the most out of image data, we need computers to “see” an image and understand the content.

Computer vision is a field of study focused on the problem of helping computers to see. Furthermore, it is one of the most powerful and compelling types of AI. It attempts to replicate parts of the complexity of the human vision system and enables computers to identify and process objects in images and video in the same way that humans do. Thanks to advances in deep learning and neural networks and the availability of datasets, the field has been able to take great leaps in recent years and has been able to surpass humans in some tasks related to detecting and labeling objects. Along with a tremendous amount of visual data (more than 3 billion images are shared online every day), the computing power required to analyze the data is now accessible. Computer vision has succeeded in myriad high-level problems, like: medical imaging, surveillance, motion capture, fingerprint recognition and biometrics, automotive safety, optical character recognition (OCR), and much more.

The convolutional neural network (CNN) [27] is most commonly used algorithm to analyze visual imagery. CNN is one of the most popular deep neural networks. They are made up of neurons that have learnable weights and biases. Furthermore, they consist of an input layer, hidden layers and an output layer. In all feed-forward neural networks, any middle layer is called hidden, on the other hand, in CNN, at least one hidden layer must perform convolution. In addition, CNNs have two major components:

- The Hidden layer/Feature extraction part: In this part, the network will perform a series of convolutions and pooling operations during which the features are detected.
- The Classification part: Here, the fully connected layers will serve as a classifier on top of these extracted features by assigning a probability for the predicted object on the image.

Moreover, we use three main types of layers to build ConvNet architectures: Convolutional Layer, Pooling Layer, and Fully-Connected Layer (Fig. 3):

- Input layer: Convolutional Neural Networks take advantage of the fact that the input consists of images and they constrain the architecture in a more sensible way. In particular, unlike a regular Neural Network, the layers of a ConvNet have neurons

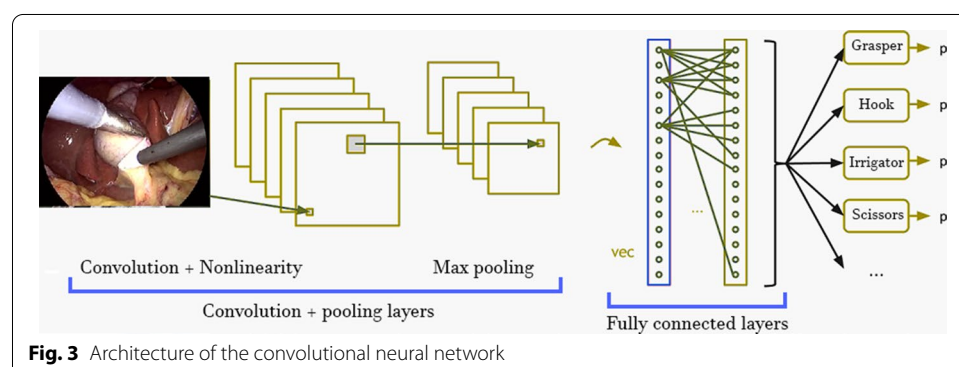


Fig. 3 Architecture of the convolutional neural network

arranged in 3 dimensions: width, height, and depth, where depth is generally the number of color channels used (RGB = 3, CYMK = 4, HSV = 3).

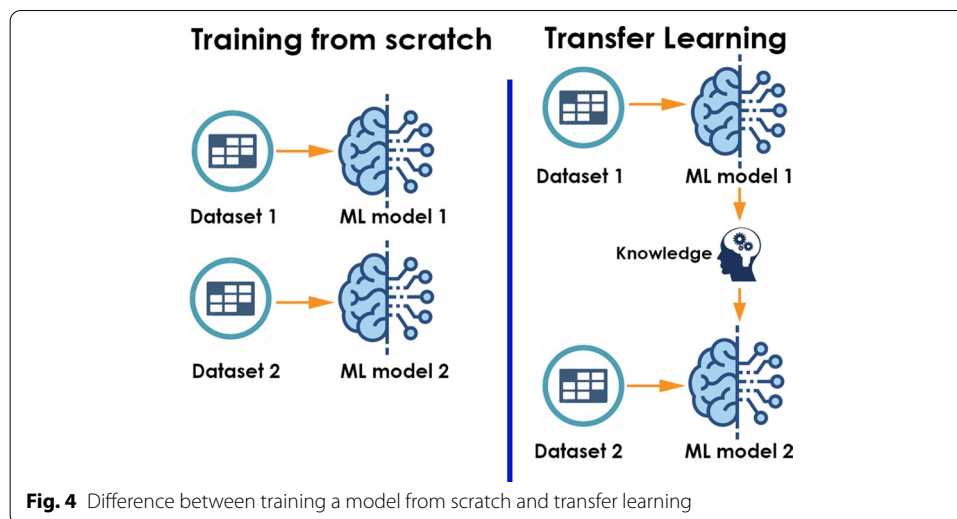
- **Convolutional layer:** is the core building block of a CNN. It generates a feature map, also called an activation map (a highlight of the relevant features of the image) using a feature detector. A feature detector, also known as a kernel or a filter, moves across the receptive field of the image, checking if the feature is present. Furthermore, the first layers detect basic features such as horizontal and vertical edges, while the next layer extracts more complex features. This process is known as convolution. The feature detector is a two-dimensional (2-D) array of weights, which represents part of the image. While they can vary in size, the filter size is typically a 3×3 matrix. The most common activation function used in this layer is ReLU.
- **Pooling layer:** conducts dimensionality reduction, by reducing the number of parameters in the input. The kernel applies an aggregation function to the values within the receptive field, populating the output array. There are two main types of pooling: Max pooling and average pooling. Max pooling selects the pixel with the maximum value to send to the output array, while average pooling calculates the average value within the receptive field to send to the output array.
- **Fully connected layer:** in the fully connected layer, each node in the output layer connects directly to a node in the previous layer. This layer performs the task of classification based on the features extracted through the previous layers and their different filters. The classification layer outputs a set of confidence scores using generally a softmax and sigmoid activation functions that specify how likely the images belong to a “class”.

Thus, CNNs eliminate the need for manual feature extraction to identify features used to classify images. The CNN works by extracting features directly from images. This automated feature extraction makes deep learning models highly accurate for computer vision tasks such as object classification.

Training a convolutional neural network

In order to train a convolutional neural network to perform object classification, three common ways are used (Fig. 4):

- **Training from scratch:** refers to building a deep network such as CNN in order to learn the features and model. In addition, to train it from scratch, a very large labeled data set is needed. This is a less common approach because with the large amount of data and rate of learning, these networks typically take days or weeks to train.
- **Transfer learning [28]:** is a concept where you transfer the weights of an already trained model (pre-trained) to another problem set on a different dataset. Furthermore, it refers to exploiting a pre-trained model as feature extractor, and taking advantage of features learned by a model trained on a larger dataset in the same domain. This is done by removing the last fully-connected layer, then instantiating a fresh fully-connected layer with an output corresponding to the number of our problem classes. The pre-trained model is “frozen” and only the weights of the classifier get updated during training. In this case, the convolutional base extracted all the fea-



tures associated with each image, and you just trained a classifier that determines the image class given that set of extracted features.

- **Fine-tuning:** In this approach, we not only replace and retrain the classifier on top of the ConvNet on the new dataset, but also we fine-tune the weights of the pretrained network by continuing the backpropagation. It is possible to fine-tune all the layers of the ConvNet, or it's possible to keep some of the earlier layers fixed (due to overfitting concerns) and only fine-tune some higher-level portion of the network. This is motivated by the observation that the earlier features of a ConvNet contain more generic features (e.g. edge detectors or color blob detectors) that should be useful to many tasks, but later layers of the ConvNet becomes progressively more specific to the details of the classes contained in the original dataset.

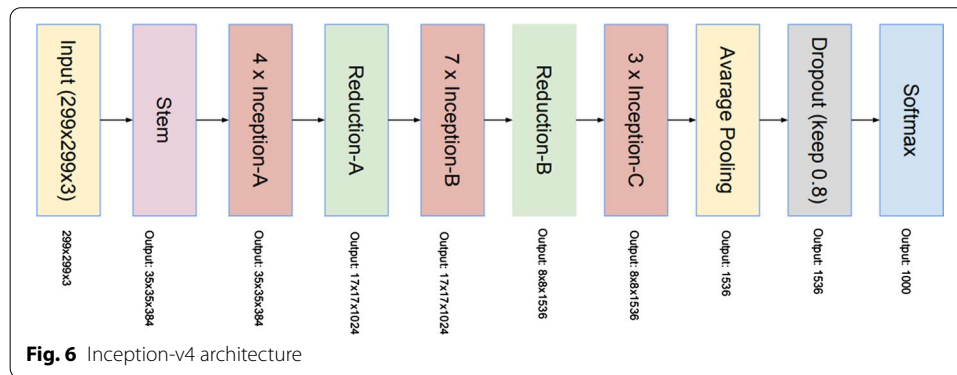
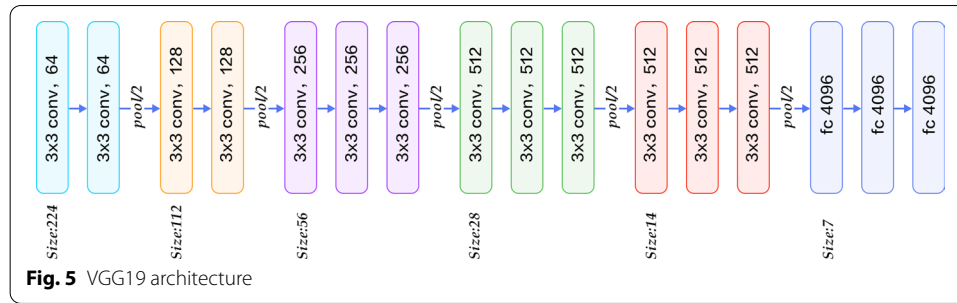
In addition, using transfer learning and fine-tuning, we can exploit the robust of discriminative filters learned by state-of-the-art networks on challenging datasets (such as ImageNet [29] or COCO [30]), and use these networks to recognize objects the model was never trained on.

In this research, we fine-tuned three state-of-the-arts algorithms, namely: VGG19, Inception v-4, and NASNet-A.

VGG19

VGG was a breakthrough in the world of Convolutional Neural Networks after the AlexNet [31]. Furthermore, it was created by Visual Geometry Group at Oxford. The main contribution of VGG is to show that classification/localisation accuracy can be improved by increasing the depth of CNN instead of using small receptive fields in the layers. Moreover, VGG19 (Fig. 5) is a convolutional neural network that is 19 layers deep, consisting of:

- A fixed input size of 224×224 . For a RGB images, the input network is changed to $224 \times 224 \times 3$, as 224 is the height and the weight, and the 3 is the RGB channels.



- A kernel of 3×3 with a stride size of 1 pixel.
- A max pooling of 2×2 pixel windows with a stride of 2.
- A Rectified linear unit (ReLU) as non-linearity function (the previous models used tanh or sigmoid).
- Three fully connected layers, where the first two layers have a size of 4096, and the last layer have a 1000 channels, which is the number of classes in the ImageNet dataset.
- The final layer contains a Softmax function.

Inception v-4

Inception-v4 (Fig. 6) is a convolutional neural network architecture that builds on previous iterations of the Inception family by simplifying the architecture and using more inception modules than Inception-v3.

Inception module was firstly introduced in Inception-v1 and GoogLeNet. The input goes through 1×1 , 3×3 , and 5×5 conv, as well as max pooling simultaneously and concatenated together as output. Thus, we don't need to think of which filter size should be used at each layer.

The inception-v4 introduced also a Batch normalization (BN), where ReLU is used as activation function to address the saturation problem and the resulting vanishing gradients. In addition, 5×5 conv was replaced by two 3×3 convs for parameter reduction while maintaining the receptive field size. Furthermore, a factorization concept is also introduced in convolution layer to reduce the dimensionality, so as to reduce the overfitting problem.

NASNet-A

Equipped with abundance of computing power and engineering genius, Google introduced NASNet (Neural Search Architecture Network), which framed the problem of finding the best CNN architecture as a Reinforcement Learning problem.

Reinforcement Learning is a type of machine learning technique that enables an agent to learn the best actions possible in a virtual environment in order to attain their goals using feedback from its own actions and experiences. Furthermore, the idea was to search the best combination of parameters of the given search space of filter sizes, output channels, strides, number of layers, etc. In this Reinforcement Learning setting, the reward after each search action was the accuracy for the searched architecture on the given dataset.

In NASNet, though the overall architecture is predefined in Fig. 7, the blocks or cells are not predefined by authors. Instead, they are searched by reinforcement learning search method. Moreover, the number of motif repetitions N and the number of initial convolutional filters are as free parameters, and used for scaling.

Specifically, these cells are called Normal Cell and Reduction Cell, where normal cell is a convolutional cell that return a feature map of the same dimension, and the reduction cell is a convolutional cell that return a feature map where the feature map height and width is reduced by a factor of two.

NASNet achieved state-of-the-art result in the ImageNet competition. However, the computation power required for NASNet was so big that only a handful of companies were able to make use of the same methodology.

In order to form the NASNet-A [6] (Fig. 8), the processing time took over 4 days using 500 GPUs resulting in 2000 GPU-hours.

Dataset

The healthcare industry has benefited greatly from deep learning capabilities ever since the digitization of hospital records and images. Furthermore, recent advances in artificial intelligence and availability of surgical datasets hold potential to enhance surgical training, improve surgeon performance and ultimately surgical outcomes. This is particularly true for minimally invasive surgery (MIS).

Gallbladder removal, or cholecystectomy, is performed more than 750,000 times annually in the US alone, mainly for benign gallstone disease which affects 10–15% of adults. The vast majority of gallbladder resections are done in relatively universal and predefined steps.

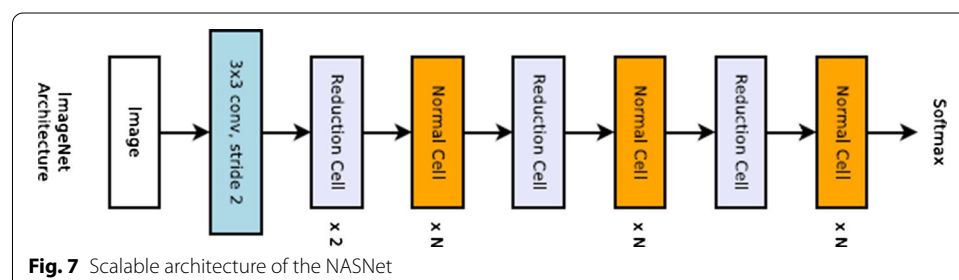
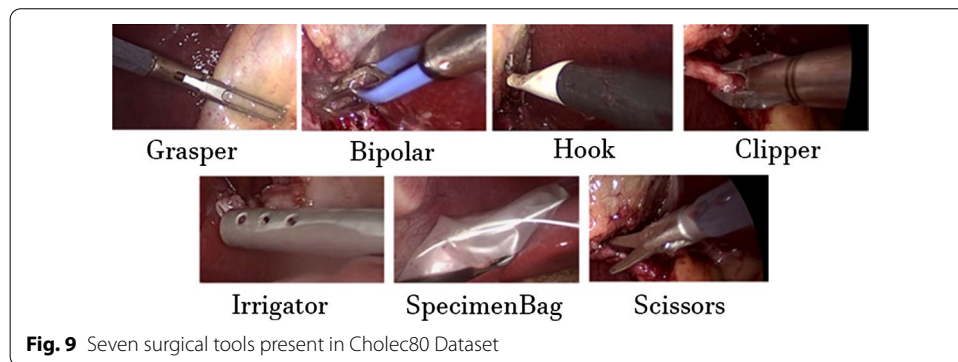
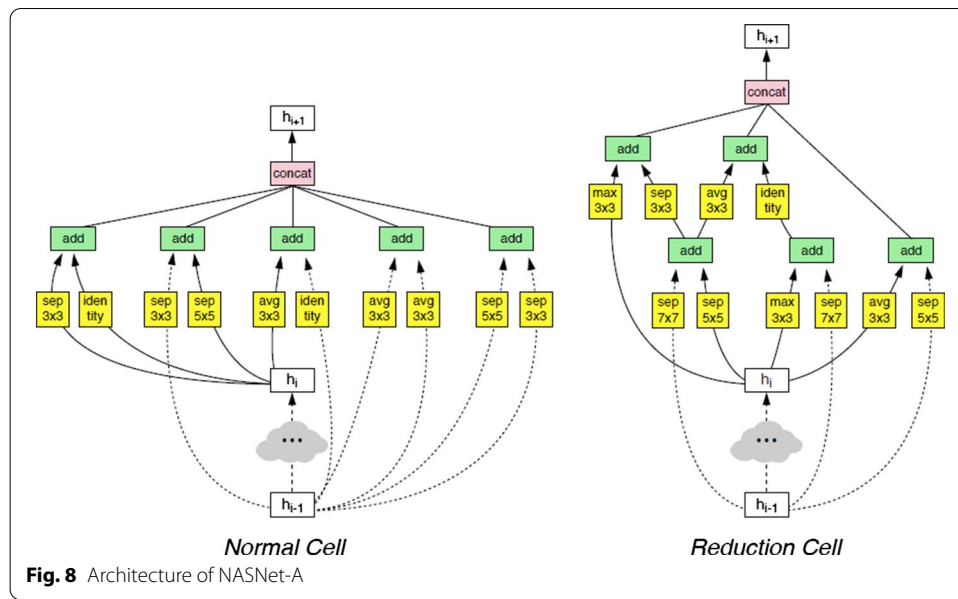


Fig. 7 Scalable architecture of the NASNet



In this work, we use a cholecystectomy surgery video dataset, called Cholec80. Cholec80 [32] is a cholecystectomy surgeries videos dataset containing 80 videos, performed by 13 surgeons. These videos are encoded with a frame rate of 25 fps (frames per second) and a video resolution of 1920×1080 pixels. Video length is varied between 12 min 19 s (minimum) and 1 h 39 min 55 s (maximum), with 38 min and 26 s on average and more than 51 h of surgery in total. Additionally, cholec80 is fully annotated with frame-level surgical tool labels for binary surgical tools detection.

In cholec80, seven tools were used and annotated (Fig. 9): specimen bag, bipolar, scissors, clipper, hook, grasper, and irrigator. Each tool is labeled as present if half of it appears in the image. One binary label is provided per image and per tool as an annotation (Multilabel classification).

Experiments and results

In this section, we present the experiments carried out using the presented dataset and the models described in the previous section. First, we describe how we pre-processed the dataset. Next, we present the common setup for the experiments, and a comparison

between the models using the three proposed neural network and the ensemble learning approach. Finally, we discuss and compare the outcomes with the state-of-the-art methods.

Data pre-processing

All videos in this study are processed in the same way. In the beginning, videos are processed using FFmpeg 3.0 and all video streams are encoded using 25 frames per second (FPS).

Removing irrelevant frames

Since videos are not edited by any professional, they have several empties and irrelevant scenes, principally at the beginning and the end of the videos. Furthermore, these frames are noisy and computationally expensive. Therefore, we cut these non-relevant frames using a background detection model. The latter was trained to identify unimportant segments that were captured outside the body. Next, these frames are used to recognize the real start and end of the surgery in the original video and cut it down. This step and the final verified video files are automatically processed and stored in local computer.

Splitting videos and resizing frames

The new dataset is now clean and does not contain irrelevant frames. Therefore, we split the preprocessed videos into 1 frame per second images, to match the tool presences annotations.

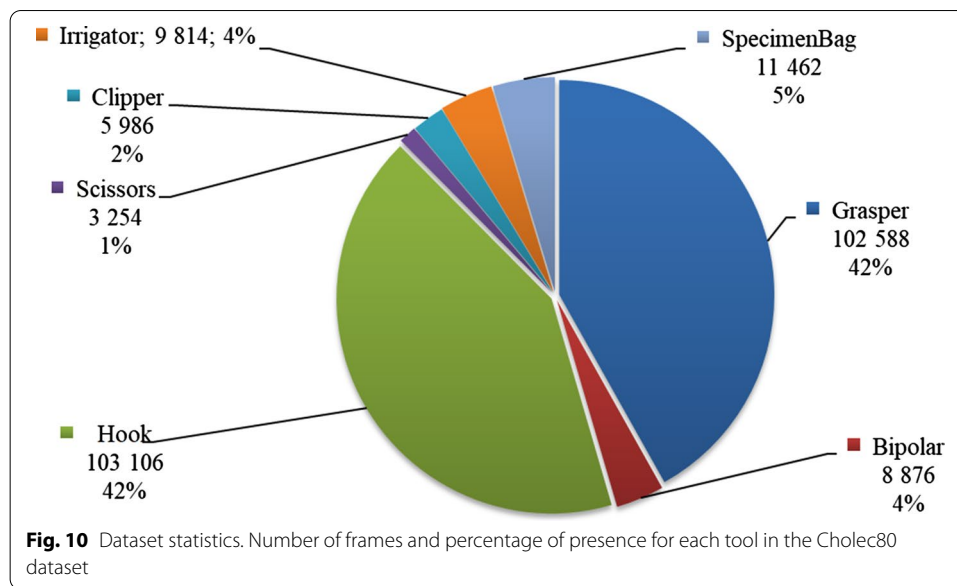
Finally, and given that neural networks receive inputs of the same size, all images need to be resized to a fixed size before inputting them to the CNNs. Moreover, the image dimension is often reduced in order to fit a reasonably sized batch in GPU memory. Thus, for each proposed neural network, the input images are resized to fit their input shape. Therefore, all images are resized to 224×224 pixels for VGG19, 299×299 pixels for the inception v4, and 331×331 pixels for the NASNet-A.

Data augmentation

In Cholecystectomy surgery, some tools are used more frequently than others. Consequently, Cholec80 video frames belonging to those tools outnumber the video frames belonging to the other tools, leading to unbalanced data (Fig. 10). This issue affects the generalization of the model and reduces the CNN efficiency to classify the different tools.

To overcome this problem, image augmentation techniques are used to increase the size of the minority classes. Images are augmented by affine transformations and blurring. We consider those transformations that preserve tool presence, namely:

- Rotation: Minority class images are rotated at an angle of zero, 40, 85, 125, 250, and 300.
- Mirroring: Mirror the image along the x-axis and y-axis.
- Shearing: The images were shifted at 40° in the counter-clockwise direction.
- Padding: Padding 5px on each border, using the reflect mode, which pad with the reflection of image without repeating the last value on the edge.



Experimental setup

To verify the impact of the ensemble learning approach, we carried out the experiments considering two scenarios:

- Scenario 1: The models VGG19, Inception v4, and NASNet-A are fine-tuned and used as a classifier.
- Scenario 2: The model combines the three neural networks using ensemble learning.

As noted earlier, fine-tuning a neural network means that the representations learnt by the previous network are used to extract the meaningful features of the new dataset images and the activation maps generated from the last convolutional layer are fed to the newly constructed fully connected network which acts as the classifier.

Thus, for each model, we use its original first blocks as feature extractors.

For both scenarios, we freeze all the layer except the last ones, in order to back-propagate the gradient and update the weights of the last layers by setting. In order to unfreeze the last blocks, we set these layers as ‘trainable = True’.

Next, we remove the fully connected layer at the end of the pretrained networks (Softmax layer), that contains 1000 classes which is the number of ImageNet dataset labels, and create a new freshly initialized sigmoid layer that is compatible with our multilabel classification task, with seven classes (surgical tools number), and append it to our model. This will predict a probability of class membership for the seven labels and assign a value between 0 and 1.

Finally, we start training our model, so that the SGD epochs their weights can be fine-tuned for the new task. Furthermore, we used a smaller learning rate to train the network because we expect that the pretrained weights are quite good already as compared to randomly initialized weights.

The first frozen early blocks learn very generic features and lets the network capture common features like edges and curves. In addition, this step ensures that any

previous robust features learned by the CNN are not destroyed. On the other hand, training the unfrozen blocks and the fully connected layer allows the model to learn specific and uncommon features from our dataset.

As we mention in the section earlier, we use Cholec80 for performance evaluation. 60 videos are assigned to the training set, while 20 videos are assigned to the test set. For validation purposes, 10 videos from the training set are used.

For all the models we use Adam optimizer with an initial learning rate of 0.001, that we decay by a factor of 10 after 10K iterations.

Since different tools can be present at the same time, the tool detection is Multilabel and Multiclass task, therefore, the final activation function is a Sigmoid function, calculated as :

$$S(x) = \frac{1}{1 + e^{-x}} \quad (1)$$

where e = Euler's number

And binary Cross-Entropy is used as Loss function, it is calculated as:

$$H_p(q) = -\frac{1}{N} \sum_{i=1}^N y_i \cdot \log(p(y_i)) + (1 - y_i) \cdot \log(1 - p(y_i)) \quad (2)$$

where y is the label (1 for the surgical tool is present and 0 for the surgical tool is not present) and $p(y)$ is the predicted probability of the surgical bein present for all the surgical tools.

All procedures were implemented on Intel® Core™ i7-9700K processor, 16GB memory, and NVIDIA Geforce GTX 2080.

The performance of the proposed approaches on the classification task is measured by the average precision (AP). It is a measure that combines recall and precision for a particular surgical tool. The AP is calculated as:

$$AP = \frac{1}{k} \sum_{Recall_i} Precision(Recall_i) \quad (3)$$

where

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

and

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

and k is the number of the points interpolated in the Precision-Recall curve,

TP = True Positives, FP = False Positives and FN = False Negatives.

The performance of the each model is calculated by the mean Average Precision (mAP), which is the mean of the average precision scores for each surgical tool.

It is calculated as :

Table 1 Frame-level presence detection average precision (AP) per-class and mean average precision (mAP) for each model

Tool	VGG19	Inception v-4	NasNet-A	Ensemble learning
Grasper	97.89	96.54	97.32	97.70
Bipolar	96.72	94.33	97.11	98.14
Hook	99.83	99.70	99.89	99.91
Scissors	87.59	80.84	90.06	94.54
Clipper	97.65	93.67	98.54	99.51
Irrigator	96.10	92.08	95.91	97.79
SpecimenBag	95.21	93.94	96.35	97.29
Average (mAP)	95.85	93.01	96.45	97.84

Table 2 Areas under the ROC curves per-class and the average mA_z for each model

Tool	VGG19	Inception v-4	NasNet-A	Ensemble learning
Grasper	96.45	95.16	96.31	96.88
Bipolar	99.67	99.17	99.53	99.81
Hook	99.90	99.79	99.91	99.93
Scissors	98.76	98.11	99.18	99.64
Clipper	99.82	99.37	99.89	99.95
Irrigator	99.41	98.74	99.41	99.85
SpecimenBag	99.62	99.39	99.56	99.81
Average (mA_z)	99.09	98.58	99.11	99.41

$$mAP = \frac{1}{7} \sum_{i=1}^7 AP_i \quad (6)$$

where 7 is the number of surgical instruments.

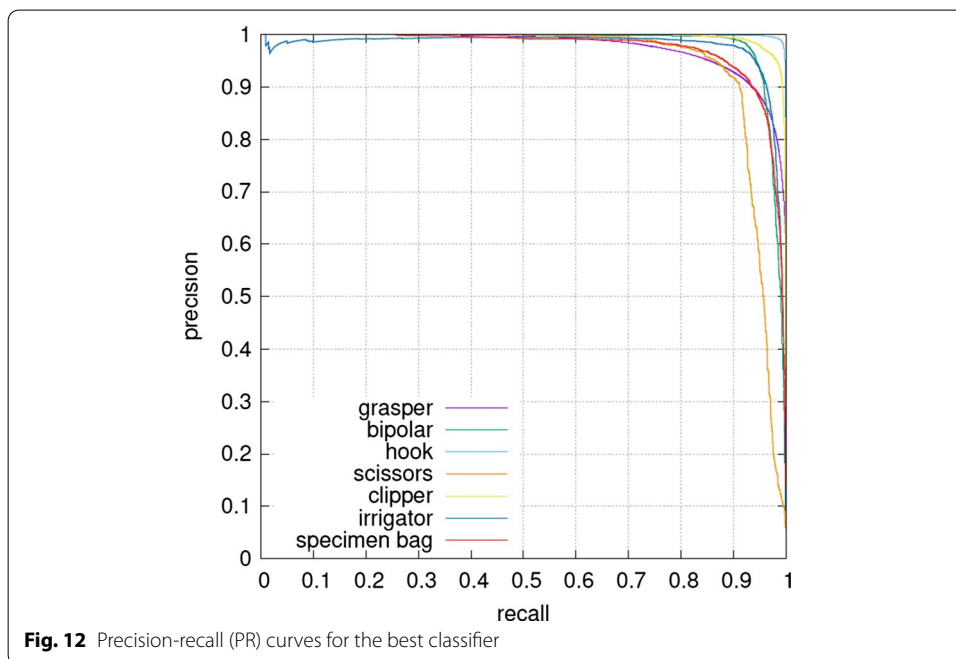
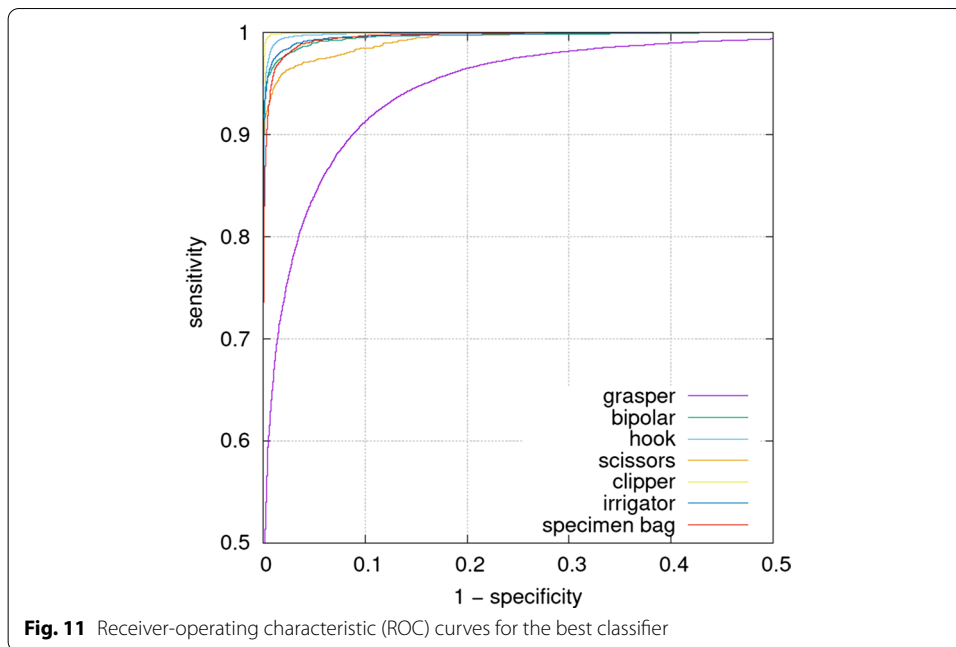
Another curve used in this research is the area under the ROC. The area under the ROC curve represents the degree or measure of separability which tells how much the model is capable of distinguishing between classes. Furthermore, an Area under the ROC curve shows the relationship between clinical sensitivity and specificity for every possible cut-off. Moreover, the ROC curve is a graph with: The x-axis showing $1 - \text{specificity}$, while the y-axis showing sensitivity.

Results and discussion

In this section, our main goal is to compare the performance of the deep learning models for both scenarios described earlier.

As we mention earlier, we implemented multiple Convolutional Neural Networks for surgical tool classification. These networks are trained on Cholec80, which is a cholecystectomy surgery videos dataset containing 80 videos. Each frame is annotated with a single or multiple surgical tools. Thus, accuracy is calculated by comparing the ground truth annotations with the predicted labels.

The classification performance for all the models using the augmented data is reported in Tables 1 and 2. It specifies the average precision and the Areas under the ROC curves



of the trained models for classifying the seven surgical tools. Additionally, Figs. 11, 12 present the Precision-Recall curve and the ROC (Receiver Operating Characteristic) curves of the best ensemble.

As expected, among the three neural networks, NasNet-A achieved the best mean average precision, with an accuracy of 97.32%, 97.11%, 99.89%, 90.06%, 98.54%, 95.91%, and 96.35% in classifying the Grasper, Bipolar, Hook, Scissors, Clipper, Irrigator, and the SpecimenBag respectively, and a mean average precision of 96.45%.

Unexpectedly, the VGG19 architecture performed quite good and outperformed the Inception v-4, even though it's older and less sophisticated. Furthermore, VGG19 achieved a mean average precision of 95.85% and outperformed the NASNet-A in classifying multiple tools.

Despite the fact that Inception v-4 is more recent and has more layers than VGG19, it obtained the lowest average precisions, therefore, the lowest mean average precision with 93.01%, less than VGG19 and NASNet-A.

It can be seen that the ensemble learning network can effectively improve the detection accuracy.

Therefore, the ensemble learning approach obtained the best results among all the proposed models in classifying all the surgical tools. Moreover, the latter achieved an accuracy of 97.70%, 98.14%, 99.91%, 94.54%, 99.51%, 97.79%, and 97.29% in classifying the Grasper, Bipolar, Hook, Scissors, Clipper, Irrigator, and the SpecimenBag respectively, and a mean average precision of 97.84%.

The average precision scores for all the surgical instruments surpassed 94%. Furthermore, the hook obtained the best detection precision among all the surgical tools, with an AP of 99.91%. Moreover, one potential explanation for this high value is that the hooker had a good visibility and high discrimination, making it easily recognizable from other instruments. Additionally, the clipper obtained the second highest average precision. One possible explanation is that it has a unique shape, making it easily distinguishable from other tools.

On the other hand, the grasper, bipolar, irrigator, and the specimen bag attained an average precision above 97%, this means that they occasionally misclassified.

The grasper and irrigator usually have a similar appearance to some other surgical tools and their shape is irregular, this may explain the reason of their misclassification.

Whereas the specimen bag, which is the third most represented class in cholec80, has a universal shape along with infrequent and irregular presence.

However, the scissors obtained the lowest AP with 94.54%, one possible reason is that this tool is less represented class in cholec80, even after augmentation. This problem requires an exploration and explanation in future work.

As a result, the ensemble learning approach yields significantly all the other models, leading to an improved average prediction performance and the reduction in the variance component of prediction errors made by the contributing models. In addition, the ensemble learning minimizes the main causes of error in learning models which are noise, bias, and variance.

Meanwhile, according to Table 2, the NASNet-A and the VGG19 get the best scores. Moreover, VGG19 outperforms NASNet-A in classifying the majority of the surgical tools. However, it falls gravely in classifying scissors, which is the less represented class in the Cholec80 dataset. Thus, NASNet-A obtained the best average score with 99.11%, against 99.09% for the VGG19. These scores are surprising owing to the unsophisticated architecture of the VGG19.

On the other hand, Inception v-4 gets the worst results with an average score of 98.58%.

Again, the ensemble learning approach outperforms all the models in classifying all the surgical tools. Furthermore, it obtains the scores: 96.88%, 99.81%, 99.93%, 99.64%,

99.95%, 99.85%, and 99.81% in classifying the Grasper, Bipolar, Hook, Scissors, Clipper, Irrigator, and the SpecimenBag respectively, and a mean average precision of 99.41%. Thus, this value means that the proposed classifier is an excellent classifier and it is able to detect more numbers of True positives and True negatives than False negatives and False positives.

To evaluate the performance of the best proposed model, we compared the results with those of previous studies. Table 3 outlines and compares our model with the best performing related models. The average precision computed for all classes and the mean average precision for each model are used to differentiate our model from the others.

It can be seen that our approach outperformed considerably the compared methods, in classifying all the surgical tools.

Compared to other Network, our method's accuracy was improved by 6.19%, which was largely attributed to the introduction of the ensemble learning.

In addition, the augmentation techniques used in the preprocessing phase increased the AP of the less represented class, especially the scissors, which obtained an accuracy of 4% more than best model classifying scissor.

In summary, all of these results demonstrate that our approach had an impressive capability in classifying surgical tools, and surpassed all the state-of-the-art detection algorithms mentioned above by a large margin.

Conclusion and future works

In this paper, we presented a study to analyze the impact of ensemble learning on the frame-by-frame detection of surgical tools in MIS videos. Furthermore, we implemented three convolutional neural networks, namely: VGG19, Inception v-4, and NASNet-A. First, we preprocessed the Cholec80 dataset videos and overcame the unbalanced data problem by using multiple data augmentation techniques. Next, we combined these three state-of-the-art CNNs in an ensemble learning model, then we compare it to a fine-tuned version of each one of them.

The experimental results show that the ensemble learning method yields significantly the other models in classifying surgical tools. In addition, the proposed method outperformed all the compared state-of-the-art algorithms on surgical tools detection in minimally invasive surgery videos, with an improvement of 6.19% in terms of mAP score in frame-level tool detection.

Table 3 Comparison of the average precision (AP) for all tools with other models

Tool	M.Sahu [33]	EndoNet [32]	Amy.J [34]	Jo [35]	Lin [36]	Shi [12]	Our model
Grasper	73.9	84.8	87.2	92.1	85.41	89.88	97.70
Bipolar	40.8	86.9	75.1	82.3	90.36	90.52	98.14
Hook	95.1	95.6	95.3	85.9	90.84	99.33	99.91
Scissors	26.2	58.6	70.8	81.2	90.58	90.78	94.54
Clipper	35.3	80.1	88.4	85.3	90.05	90.19	99.51
Irrigator	33.2	74.4	73.5	82.9	87.42	89.62	97.79
SpecimenBag	76.6	86.8	82.1	83.2	89.98	91.25	97.29
Average (mAP)	54.44	81.0	81.8	84.7	89.23	91.65	97.84

For future work, we expect that we improve the accuracy and tackle the real-time nature of the object detection model to implement it in the on-site online learning, assisted direction of surgery, and the surgical quality assessment.

Acknowledgements

Not applicable.

Author contributions

All authors read and approved the final manuscript.

Funding

Not applicable. This research received no specific grant from any funding agency.

Availability of data and materials

Not applicable. For any collaboration, please contact the authors.

Declarations

Ethics approval and consent to participate

The author confirms the sole responsibility for this manuscript. The author read and approved the final manuscript.

Consent for publication

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Competing interests

The authors declare that they have no competing interests.

Received: 13 September 2021 Accepted: 6 April 2022

Published online: 28 April 2022

References

1. Carli F, et al. Editorial I: Measuring the outcome of surgical procedures: what are the challenges? *Br J Anaesth*. 2001;87(4):531–3.
2. Mota P, Carvalho N, Carvalho-Dias E, Joao Costa M, Correia- Pinto J, Lima E. Video-based surgical learning: improving trainee education and preparation for surgery. *J Surg Educ*. 2018;75(3):828–35. <https://doi.org/10.1016/j.j Surg.2017.09.027>.
3. Henken KR, Jansen FW, Klein J, Stassen LPS, Dankelman J, van den Dobbelen JJ. Implications of the law on video recording in clinical practice. *Surg Endosc*. 2012;26:29092916. <https://doi.org/10.1007/s00464-012-2284-6>.
4. Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. [arXiv:1409.1556](https://arxiv.org/abs/1409.1556).
5. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, inception-ResNet and the impact of residual connections on learning. In: *AAAI conference on artificial intelligence*. 2016.
6. Zoph B, Vasudevan V, Shlens J, Le QV. Learning transferable architectures for scalable image recognition. In: *IEEE/CVF conference on computer vision and pattern recognition*. 2018. p. 8697–710. <https://doi.org/10.1109/CVPR.2018.00907>.
7. Li L, Huang H, Jin X. AE-CNN classification of pulmonary tuberculosis based on CT images. In: *2018 9th international conference on information technology in medicine and education (ITME)*; 2018. <https://doi.org/10.1109/itme.2018.00020>.
8. Xiao Z, Huang R, Ding Y, Lan T, Dong R, Qin Z, Zhang X, Wang W. A deep learning-based segmentation method for brain tumor in MR images. In: *2016 IEEE 6th international conference on computational advances in bio and medical sciences (ICCBMS)*; 2016. <https://doi.org/10.1109/iccabs.2016.7802771>.
9. Joshi S, Gore S. Ischemic stroke lesion segmentation by analyzing MRI images using dilated and transposed convolutions in convolutional neural networks. In: *2018 fourth international conference on computing communication control and automation (ICCCAA)*; 2018. <https://doi.org/10.1109/iccubea.2018.8697545>.
10. Ye J, Luo Y, Zhu C, Liu F, Zhang Y. Breast cancer image classification on WSI with spatial correlations. In: *ICASSP 2019—2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*; 2019. <https://doi.org/10.1109/icassp.2019.8682560>.
11. Kiruthika M, Swapna TR, Santhosh Kumar C, Peeyush KP. Artery and vein classification for hypertensive retinopathy. In: *2019 3rd international conference on trends in electronics and informatics*.
12. Shi P, Zhao Z, Hu S, Chang F. Real-time surgical tool detection in minimally invasive surgery based on attention-guided convolutional neural network. *IEEE Access*. 2020;8:228853–62. <https://doi.org/10.1109/ACCESS.2020.3046258>.

13. Wang S, Raju A, Huang J. Deep learning based multi-label classification for surgical tool presence detection in laparoscopic videos. In: 2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017); 2017. p. 620–3. <https://doi.org/10.1109/ISBI.2017.7950597>.
14. Kletz S, Schoeffmann K, Benois-Pineau J, Husslein H. Identifying surgical instruments in laparoscopy using deep learning instance segmentation. In: International conference on content-based multimedia indexing (CBMI). 2019. p. 1–6. <https://doi.org/10.1109/CBMI.2019.8877379>.
15. Jalal Nour, Alshirbaji Tamer, Möller Knut. Predicting surgical phases using CNN-NARX neural network. *Curr Dir Biomed Eng*. 2019;5:405–7. <https://doi.org/10.1515/cdbme-2019-0102>.
16. Wang G, Wang S. Surgical tools detection based on training sample adaptation in laparoscopic videos. *IEEE Access*. 2020;8:181723–32. <https://doi.org/10.1109/ACCESS.2020.3028910>.
17. Zhang B, Wang S, Dong L, Chen P. Surgical tools detection based on modulated anchoring network in laparoscopic videos. *IEEE Access*. 2020;8:23748–58. <https://doi.org/10.1109/ACCESS.2020.2969885>.
18. Namazi B, et al. LapTool-Net: a contextual detector of surgical tools in laparoscopic videos based on recurrent convolutional neural networks; 2019. [arXiv:1905.08983](https://arxiv.org/abs/1905.08983).
19. Chittajallu DR, Dong B, Tunison P, Collins R, Wells K, Fleshman J, Sankaranarayanan G, Schwaizberg S, Cavuoto L, Enquobahrie A. XAI-CBIR: explainable AI system for content based retrieval of video frames from minimally invasive surgery videos. In: 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019); 2019. <https://doi.org/10.1109/isbi.2019.8759428>.
20. Kletz S, et al. identifying surgical instruments in laparoscopy using deep learning instance segmentation. In: 2019 international conference on content-based multimedia indexing (CBMI); 2019. p. 1–6.
21. Shvets A, Rakhlin A, Kalinin A, Iglovikov V. Automatic instrument segmentation in robot-assisted surgery using deep learning. In: 2018 17th IEEE international conference on machine learning and applications (ICMLA); 2018. p. 624–8. <https://doi.org/10.1109/ICMLA.2018.00100>.
22. Kanakatte A, Ramaswamy A, Gubbi J, Ghose A, Purushothaman B. Surgical tool segmentation and localization using spatio-temporal deep network. In: 2020 42nd annual international conference of the IEEE engineering in medicine & biology society (EMBC); 2020. p. 1658–61. <https://doi.org/10.1109/EMBC44109.2020.9176676>.
23. Ishirbaji TA, et al. A convolutional neural network with a two-stage LSTM model for tool presence detection in laparoscopic videos. *Curr Dir Biomed Eng*. 2020. <https://doi.org/10.1515/cdbme-2020-0002>.
24. Janiesch C, Zschech P, Heinrich K. Machine learning and deep learning. *Electron Mark*. 2021. <https://doi.org/10.1007/s12525-021-00475-2>.
25. Najafabadi MM, Villanustre F, Khoshgoftaar TM, et al. Deep learning applications and challenges in big data analytics. *J Big Data*. 2015;2:1. <https://doi.org/10.1186/s40537-014-0007-7>.
26. Thompson NC, Greenewald KH, Lee K, Manso GF. The computational limits of deep learning; 2020. [arXiv:2007.05558](https://arxiv.org/abs/2007.05558).
27. Schmidhuber Jürgen. Deep learning in neural networks: an overview. *Neural Netw*. 2015;61:85–117. <https://doi.org/10.1016/j.neunet.2014.09.003>.
28. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. *J Big Data*. 2016;3:9. <https://doi.org/10.1186/s40537-016-0043-6>.
29. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition; 2009. p. 248–55.
30. Lin TY, et al. Microsoft COCO: common objects in context. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, et al., editors. *Computer vision—ECCV 2014*. ECCV 2014, vol. 8693. Lecture notes in computer science. Cham: Springer; 2014. https://doi.org/10.1007/978-3-319-10602-1_48.
31. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: *Proceedings of the 25th international conference on neural information processing systems (NIPS'12)*, Vol. 1. Red Hook: Curran Associates Inc.; 2012. p. 1097–105.
32. Twinanda AP, Shehata S, Mutter D, Marescaux J, de Mathelin M, Padoy N. EndoNet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE Trans Med Imaging*. 2017;36(1):86–97. <https://doi.org/10.1109/tmi.2016.2593957>.
33. Sahu M, Mukhopadhyay A, Szengel A, Zachow S. Tool and phase recognition using contextual CNN features; 2016. [arXiv:1610.08854](https://arxiv.org/abs/1610.08854).
34. Jin A, Yeung S, Jopling J, Krause J, Azagury D, Milstein A, Fei-Fei L. Tool detection and operative skill assessment in surgical videos using region-based convolutional Neural Networks. In: 2018 IEEE winter conference on applications of computer vision (WACV). 2018.
35. Jo K, Choi Y, Choi J, Chung JW. Robust real-time detection of laparoscopic instruments in robot surgery using convolutional neural networks with motion vector prediction. *Appl Sci*. 2019;9:2865.
36. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. *IEEE Trans Pattern Anal Mach Intell*. 2020;42(2):318–27. <https://doi.org/10.1109/TPAMI.2018.2858826>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.