This project has received funding from
The European Union's Horizon 2020 research and innovation programme under
Grant agreement No 101017562

Finding Endometriosis using Machine Learning
FEMaLe

Call/Topic: Digital transformation in Health and Care

Type of action: RIA

Date: 28-09-2023

DELIVERABLE NUMBER	D7.3
DELIVERABLE TITLE	Deep learning algorithm to automatically determinate the division plane
RESPONSIBLE AUTHOR	RTU

GRANT AGREEMENT No.	101017562
DOCUMENT TYPE	Other
WORKPACKAGE N. | TITLE	WP7 | VISUAL: augmented reality to improve laparoscopic surgery
LEAD CONTRACTOR	SURGAR
AUTHOR(S)	Saman Noorzadeh, Julie Desternes, Julien Peyras, Nicolas Bourdel
PLANNED DELIVERY DATE	September 30th, 2023
ACTUAL DELIVERY DATE	September 28th, 2023
DISSEMINATION LEVEL	Public
STATUS	Reviewed and quality checked
VERSION	Final version (1.5)
REVIEWED BY	FEMaLE PMO

Document history

Version	Date [1]	Comment	Author	Status [2]
1.1	07-08-2023	First draft created	RTU	Drafted
1.2	30-08-2023	Second draft created, including recommendations from WP participants.	SURG	Drafted
1.3	12-09-2023	Third draft prepared for FEMaLe Review Panel.	SURG	Drafted
1.4	25-09-2023	Final draft created, based on FEMaLe Review Panel feedback.	SURG	Completed
1.5	27-09-2023	Final version ready for submission, quality checked by FEMaLe PMO.	AU	Validated

[1] As per the project’s cloud storage or per email date if applicable.

[2] Drafted, completed or validated as per the project's cloud storage or per email date if applicable.

TABLE OF CONTENTS

Disclaimer

The content of this deliverable does not reflect the official opinion of the European Union. Responsibility for the information and views expressed herein lies entirely with the author(s).

All FEMaLe Consortium members are also committed to publish accurate and up to date information and take the greatest care to do so. However, the FEMaLe Consortium members cannot accept liability for any inaccuracies or omissions, nor do they accept liability for any direct, indirect, special, consequential, or other losses or damages of any kind arising out of the use of this information.

Copyright notice

Acknowledgement

The project 'Finding Endometriosis using Machine Learning' (FEMaLe) has received funding from the European Union Horizon 2020 programme under grant number 101000640.

Citation

Be so kind as to cite this work as:

Finding Endometriosis using Machine Learning, 2023: Deep learning algorithm to automatically determinate the division plane – under the supervision of the Project's Coordinator.

Legislation

Legislation H2020 Framework Programme – Regulation (EU) No 1291/2013 of the European Parliament and of the Council of 11 December 2013 establishing Horizon 2020 - The Framework Programme for Research and Innovation (2014-2020) (OJ 347, 20.12.2013, p. 104).

H2020 Specific Programme – Council Decision 2013/743/EU of 3 December 2013 establishing the Specific Programme Implementing Horizon 2020 - The Framework Programme for Research and Innovation (2014-2020) (OJ L 347, 20.12.2013, p. 965).

Rules for Participation (RfP) – Regulation (EU) No 1290/2013 of the European Parliament and of the Council of 11 of December 2013 laying down the rules for the participation and dissemination in Horizon 2020 – the Framework Programme for Research and Innovation (2014-2020) (OJ L 347, 20.12.2013, p.81).

Financial Regulation (FR) – Regulation (EC, Euratom) No 966/2012 of the European Parliament and of the Council of 25 October 2012 on the financial rules applicable to the general budget of the European Union (OJ L 298, 26.10.2012, p.1).

Rules of Application (RAP) – Commission Regulation (EC, Euratom) No 1268/2012 of 29 October 2012 on the rules of application of l Regulation (EC, Euratom) No 966/2012 of the European Parliament and of the Council on the financial rules applicable to the general budget of the Union (OJ L 298, 26.10.2012, p.1).

1. Introduction

WP7 seeks to develop an augmented reality (AR) solution designed to assist surgeons during laparoscopic procedures for endometriosis. Its primary purpose is to provide a visual representation of the division plane, a task traditionally requiring years of experience. Beyond benefiting expert surgeons, this innovative tool aims to empower junior surgeons, ultimately enhancing the quality of endometriosis surgeries. In essence, the AR tool strives to enhance surgical outcomes for both experienced and junior practitioners, all in the pursuit of improved patient care. Additionally, this tool has the potential to establish a standardised approach to these operations on a global scale. Achieving this objective entails training surgeons through visualisation platforms, a strategy recommended by such software solutions.

Such a tool is only achievable through the state-of-the-art techniques of deep learning, in which a model should be learned from the labelled laparoscopic frames and finally the machine can detect the zone on which the surgeon should operate.

1.1 Background and Context

Diagnosis and treatment of endometriosis frequently involve laparoscopic surgery, a minimally invasive procedure performed through small abdominal incisions. However, the complexity of endometriosis, with its varied locations and presentations, poses unique challenges for surgeons.

One of the key challenges in laparoscopic surgeries for endometriosis is the precise identification and treatment of affected areas within the abdominal cavity. Surgeons depend on the accurate identification of division planes to precisely target and treat endometriotic lesions while minimising harm to surrounding healthy tissue.

The division plane is a three-dimensional (3D) surgical plane within which the surgeon performs their procedures. Studying such a plane necessitates the precise annotation of data in a 3D context (while laparoscopic cameras typically provide a 2D view of the surgical field, making it difficult for surgeons to accurately perceive depth), the thorough examination of various lesions following the initial incision, and a thorough assessment of intricate excisions during the surgical procedure. Given the complexity of this task, we adopt a gradual approach, commencing with the comprehensive examination of the first incision boundaries. This initial step, however, is not as easy as it sounds. As indicated in the previously presented report, D7.2, there exists a significant degree of variability among surgeons with respect to defining these incision boundaries. Establishing a consistent and universally accepted standard for consensus in this regard is of great importance [1]. Such a standard not only serves to benefit surgeons by enhancing surgical precision but also contributes to the potential for automation within this domain.

1.2 The Challenge of Incision Boundary Detection

Laparoscopy is considered the gold standard surgical treatment for endometriosis. Surgical treatment aims to destroy or remove all visible endometriotic lesions and to repair the damage to organs and other sites caused by endometriosis, restoring the normal anatomy. Laparoscopic excision of peritoneal deposits of endometriosis may be accomplished utilising different techniques including sharp dissection, electro-excision, Argon Neutral Plasma Energy, laser energy, ultrasound scalpel or helium thermal coagulator [2]. Laparoscopic surgery in this regard is limited to highly qualified surgeons.

Yet, surgeons lack a common method for performing surgeries, and there is considerable variation in where they make their initial incisions.

1.3 The Promise of Deep Learning

The advent of deep learning and convolutional neural networks (CNNs) has opened up new possibilities for automating and enhancing the accuracy in the medical field. Deep learning models have demonstrated remarkable capabilities in semantic segmentation tasks, making them a promising tool for this critical medical application in the field of laparoscopic surgeries [3].

Semantic segmentation is a computer vision task where the goal is to classify each pixel in an image into a predefined class or category. Semantic segmentation assigns a label to each individual pixel, creating a detailed pixel-wise segmentation map. This technique is commonly used in tasks such as image segmentation, autonomous driving, and medical image analysis.

In the context of detecting incision zones in endometriosis laparoscopic surgeries, deep learning and semantic segmentation play a crucial role. Once the deep neural network is trained, the model can analyse each laparoscopic image, and generate segmentation maps where it suggests an incision boundary.

1.4 Objectives and Contributions

This study aims to leverage the power of deep learning to address the challenges associated with incision boundary detection as mentioned above. By developing and evaluating a DeepLabV3-based model, we seek to improve the precision and safety of endometriosis laparoscopic surgeries. Finally, the trained model can then be used in the AR software to help the surgeons in their operations.

2. Trustworthy AI

In developing our incision boundary segmentation algorithm in endometriosis laparoscopic surgeries, we have placed a strong emphasis on trustworthy AI principles. Having previously addressed these principles in the deliverables of WP7, we now reaffirm our commitment to upholding these standards in the current task within this work package. This commitment underscores our dedication to trustworthy AI as these principles remain fundamental to our approach in tackling the objectives of this WP of the project.

Trustworthy AI includes a set of ethical and technical guidelines aimed at ensuring transparency, fairness, accountability, and robustness in artificial intelligence systems. In the context of this deliverable:

- transparency is achieved through clear documentation of our model architecture and training procedures, allowing for a full understanding of the methods employed.
- Fairness is upheld by consistently evaluating our model's performance across diverse patient populations to prevent bias and ensure equitable outcomes.
- Accountability is maintained through meticulous data annotation and model validation, where expert and junior oversight followed by a consensus, ensures the highest standards of accuracy. As we explain later this had required hours of discussions among surgeons to create a high-quality ground truth. The algorithm is evaluated according to such ground truth and based on several quantitative metrics.
- Robustness of our model is prioritised by continuously improving the model’s ability to perform accurately in a variety of data collected from centres around the world.
- Human agency and oversight is considered: The endometriosis operation suggestion device will empower surgeons and allow them to make informed decisions. At the same time, the ground truth decisions upon which the model is built are made by surgeons, and it is the surgeon who finally takes the decision about operation after the machine suggests (we stick to the word 'suggest' for the machine to show that the main surgeon finally takes the decision).

3. Methodology

The whole pipeline that we have applied to construct the algorithm is briefly mentioned as follows:

- **Data Collection**: Laparoscopic videos of surgical procedures are collected as the input data. The beginning of each operation is only extracted before the excision starts. That is because of the fact that incision zones which are the areas of interest are still intact for detection.
- **Annotation:** Medical experts annotate the images to label the pixels within the images that correspond to the incision zones. This annotated dataset is used to train the deep learning model. The whole pipeline of annotation is described in D7.2 in detail.
- **Deep Learning Model On Semantic Segmentation:** A deep learning architecture, here DeepLabV3, is trained on the annotated dataset. During training, the model learns to recognize the specific visual patterns associated with incision zones by adjusting its internal parameters.
- **Model Evaluation:** The model is evaluated based on several metrics like Intersection over Union, specificity, sensitivity, etc.
- **Visual-Based Evaluation:** Some images are then shown to the surgeon, to get the first feedback of the performance of the model. We will explain this later.
- **Clinical Application:** The segmentation maps generated by the model can be used in real-time during surgeries to assist surgeons, when implemented into the AR software, in identifying and navigating incision zones accurately. This enhances surgical precision, reduces the risk of damaging nearby tissues, and contributes to improved patient outcomes.

3.1. Dataset

The up-to-date number of surgeries and short video sequences extracted out of surgeries are shown in Fig.1 and Table.1.

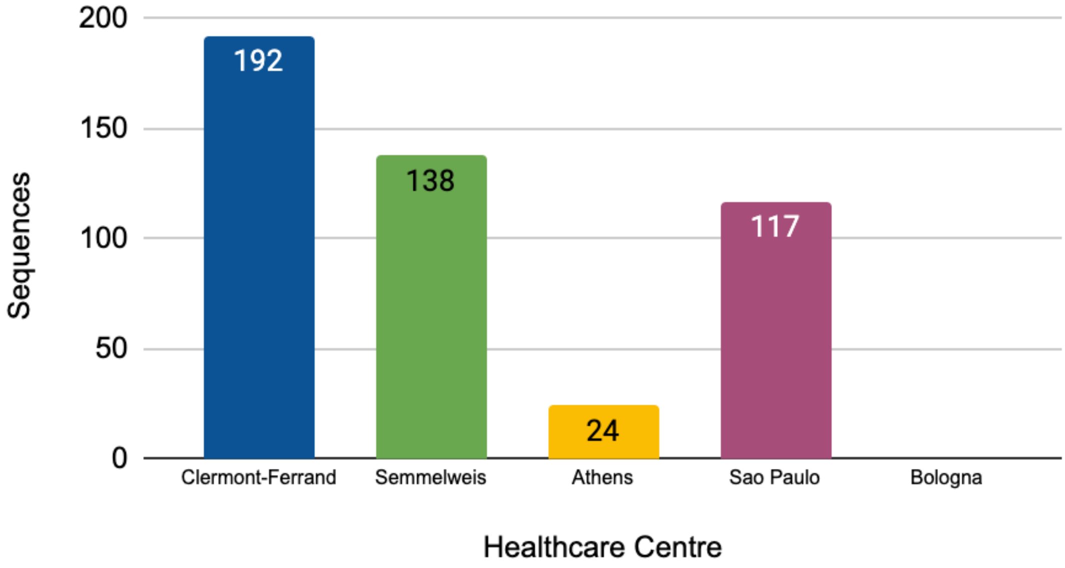


Fig.1. The Number of short video sequences for each data provider healthcare centre.

Healthcare Centre	Surgeries	Videos	Sequences
Clermont-Ferrand	72	115	192
Semmelweis	62	64	138
Athens	6	10	24
Sao Paulo	63	55	117
Bologna	1	4	8
Total	**204**	**248**	**479**

Table1. The number of surgeries, videos and short sequences collected and processed.

3.2. Pre-processing and Annotations

As mentioned, each video is cut into several interesting sequences. Then the interesting frames are marked by the annotators. Then the whole annotator team annotated the incision boundaries. Although the details can be found in D7.2, a brief summary is reminded here so that the next steps can be comprehensible in this report.

According to the ontology that is created through this study, the annotations are done as polygons in two different classes: The *Treat Zone*, and the *Check Zone*. As shown in Fig.2 as an example, the treat zone is the part that the surgeon has to treat independently of the surgical operation (whether it is coagulation, excision, etc.).

The check zone is an adequate safety margin around the resect zone, in order to guarantee the complete removal of the endometriotic lesion and any perilesional fibrosis. The check zone also allows the surgeon to indicate any areas which, in the analysed frame, cannot be classified as certainly healthy or certainly a site of endometriosis: areas which, therefore, must at least be carefully analysed by the surgeon in the real world.

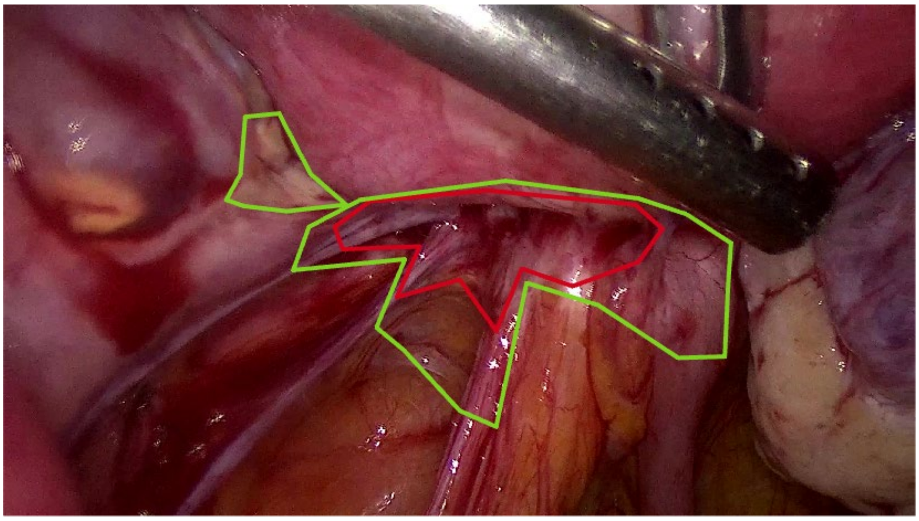


Fig.2. The *treat* and *check* zones are marked in red and green, respectively.

The annotators' team consists of 2 expert and 2 junior surgeons. And every two weeks they follow a procedure in which the juniors annotate 100 frames, one expert annotates approximately 50 frames. After these annotations are done, all the 4 annotators annotate 15 frames blindly. After completion, a discussion session takes place where the annotations from all annotators are shared with everyone present. The discussion per frame is then done to reach a consensus on the annotations. After the session, the consensus is also annotated separately. To date, we had about 55 person-hour of discussions for reaching a consensus in 9 cycles.

All these data and the annotations are stored to be served as the database and be fed into the machine. The up-to-date statistical status of the annotated dataset is shown in Table.2. We have about 8K zones annotated on the whole dataset made by all the annotators.

TOTAL

1150 Frames

CLASS	FIGURES
Total	**7983**
To Treat	3693
To Check	4290

Table.2. The number of annotated frames and the created annotations (FIGURES) per class.

Finally, the number of annotated images for each of the annotators is shown in Fig.3. It should be noted that each frame can be annotated by several annotators.

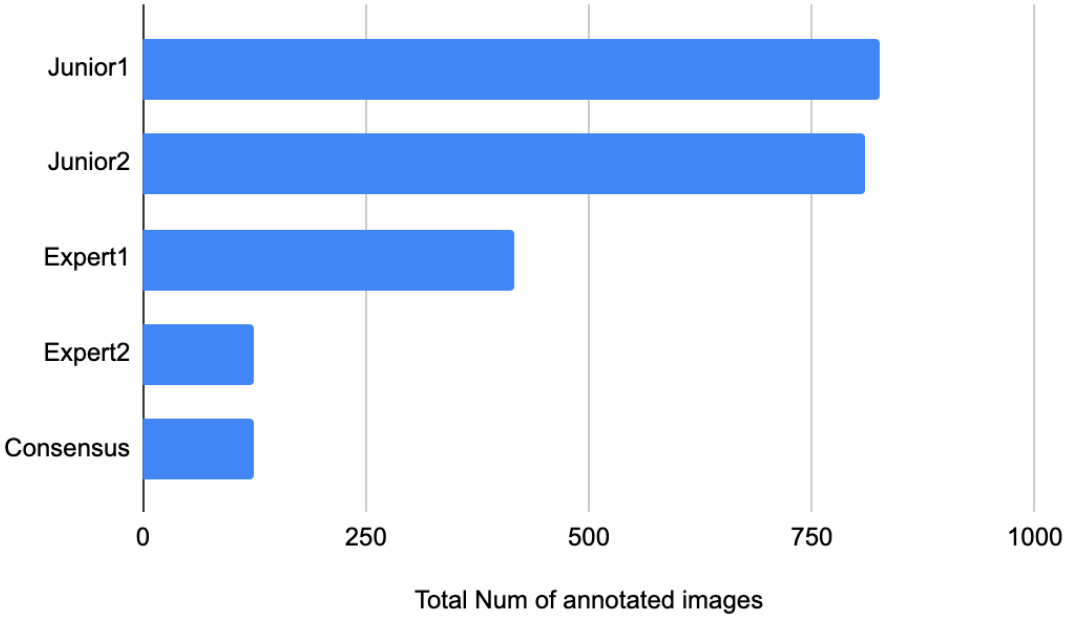


Fig.3. The number of annotated images by each of the annotators.

3.3.DeepLabV3 architecture

In computer vision, deep learning algorithms operate by constructing complex hierarchical representations of input data. This involves the use of deep neural networks with multiple layers that progressively extract features of increasing abstraction. Convolutional Neural Networks (CNNs) are a prime example of deep learning architectures that have revolutionised computer vision. CNNs are capable of automatically learning and discerning features such as edges, textures, and patterns directly from raw image data. DeepLabV3 [4][5], used here, is a deep learning model specifically designed for semantic segmentation tasks in computer vision. It's a convolutional neural network (CNN) architecture known for its exceptional performance in pixel-level image classification. DeepLabV3 achieves this by employing atrous convolution (also known as dilated convolution) and an encoder-decoder structure. Key features of DeepLabV3 include:

- Atrous/Dilated Convolution: This type of convolution allows the model to capture information from a broader receptive field, as graphically shown in Fig.4 without significantly increasing computational complexity. It helps maintain fine-grained details during the segmentation process.

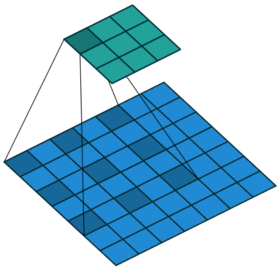
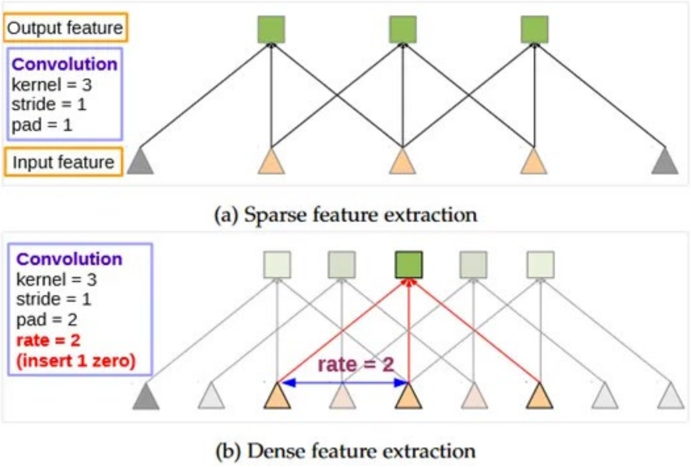


Figure 4. Atrous Convolution. Source [6].

- Multi-Scale Context: DeepLabV3 incorporates multi-scale contextual information by using multiple atrous convolution rates. This enables the model to consider both local and global context, making it robust in handling objects of various sizes within images. This is graphically represented in Fig.5.



(a) Sparse feature extraction

(b) Dense feature extraction

Figure 5. 1D Atrous Convolution for dense feature extraction. Source [5].

- Encoder-Decoder Architecture: The model consists of an encoder network that extracts features and a decoder network that upsamples and refines the segmentation maps. This architecture helps improve the accuracy of segmentation. See Fig. 6.

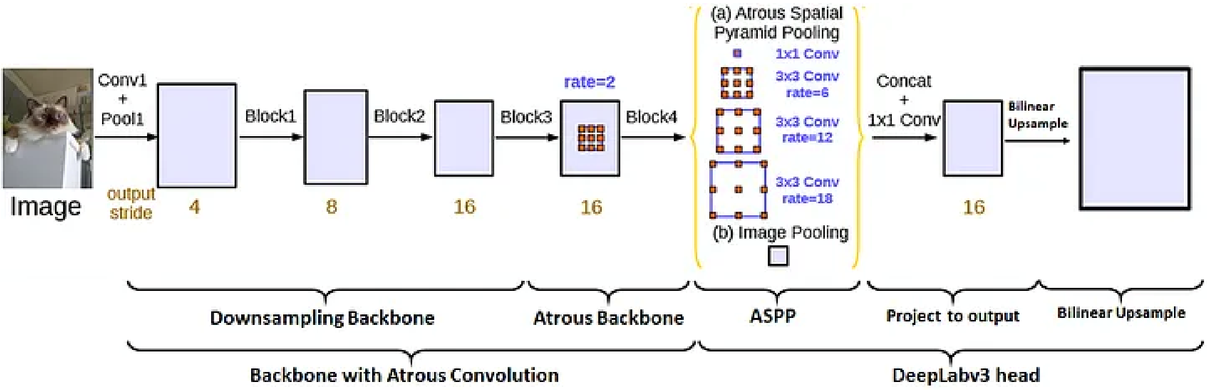


Figure 6. DeepLabv3 Architecture with labelled blocks. Source [7].

- Use of Pre-trained Backbones: DeepLabV3 often uses pre-trained CNN backbones, such as ResNet or MobileNet, to benefit from transfer learning. This means the model leverages knowledge learned from large datasets to improve performance on specific segmentation tasks.

3.4.Training

We used the images of all cycles except cycles 8, and 9 (which are put aside as test dataset), for training and validation. We created 4 different datasets each annotated by a different individual/group. These datasets are created so that we can finally evaluate the quality of each of them and evaluate the efficiency of our annotation pipeline. The datasets are as follows:

- **based on consensus annotations:** The annotations marked after each discussion session.
- **based on junior annotations**: Both junior annotations are merged per image to create a database for the junior annotations. This means that in this merged dataset only the regions which are annotated by both juniors as ‘treat’ are considered ‘treat’. All the other annotations are considered as ‘check’ zones.
- **based on expert annotations**: The annotations done by expert 1.
- **based on all annotations**: Finally, we considered a combination of all the above dataset. The consensus annotation is considered for frames annotated by both junior and expert annotators.

According to classification of data, the datasets do not have the same size. Fig.7 shows the size of each dataset.

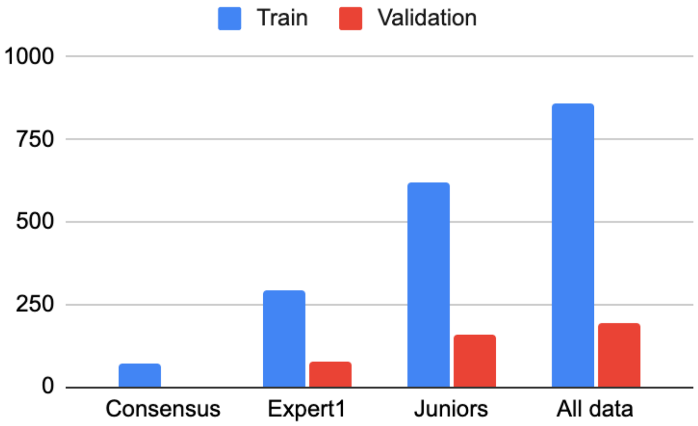


Fig.7. The number of images in each of the four defined datasets.

Training Procedures and Parameters:

We used both Mobilenet and ResNet as backbones of the DeepLabV3 architecture. Since the results of ResNet50 were considerably better than that of MobileNet we only report the results of our DeepLabV3 ResNet50 model trained on the datasets. This means that we will analyse the training results using this model and also run inference using the same model.

We do a series of data augmentation to prevent the DeepLabV3 ResNet50 model from overfitting on the dataset. This includes HorizontalFlip, RandomBrightnessContrast, Rotate, ImageCompression. Augmentation is not applied to the validation dataset. However, both the training and validation images undergo normalisation using the ImageNet normalisation values. This practice is adopted due to the utilisation of pretrained models, which are subsequently fine-tuned. Cross entropy is used as the loss function with focal loss. The number of epochs is 300, with a batch size of 6, and Adam optimizer. Fig.8 shows the training results on loss and accuracy when trained on all data.

We train 4 models with the same parameters but different train-validation dataset. These datasets are the ones described just above.

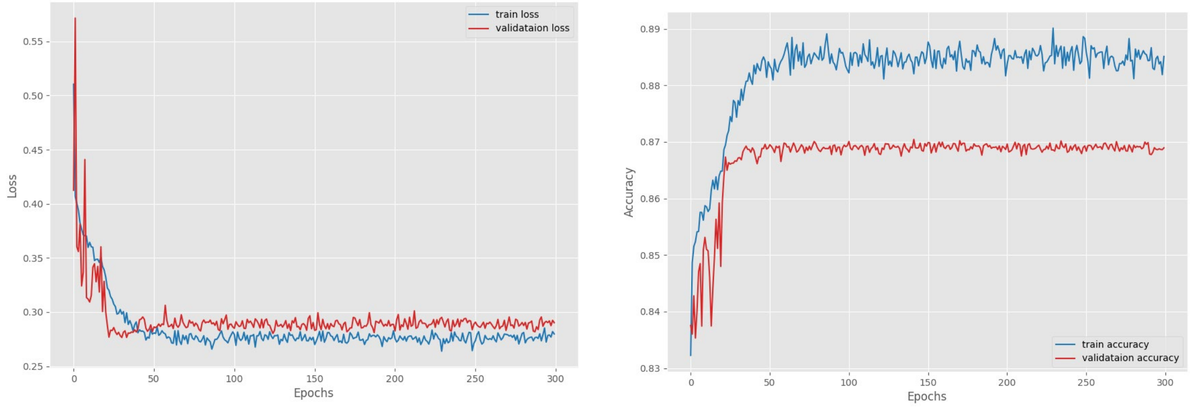


Fig.8. Accuracy and loss function trained on 'all data' set.

3.5.Evaluation Metrics

3.5.1. IoU (Intersection over Union)

This metric is used to evaluate the model's performance on the test set. IoU score is a common evaluation metric used in tasks like image segmentation to measure the accuracy of predictions made by machine learning models. It quantifies the degree of overlap between the predicted region and the ground truth region in the context of bounding boxes or segmented objects. The IOU score is calculated as the ratio of the area of intersection between the predicted and ground truth regions to the area of their union. Mathematically, it can be expressed as:

$$IOU = \frac{Area\_of\_Intersection}{Area\_of\_Union}$$

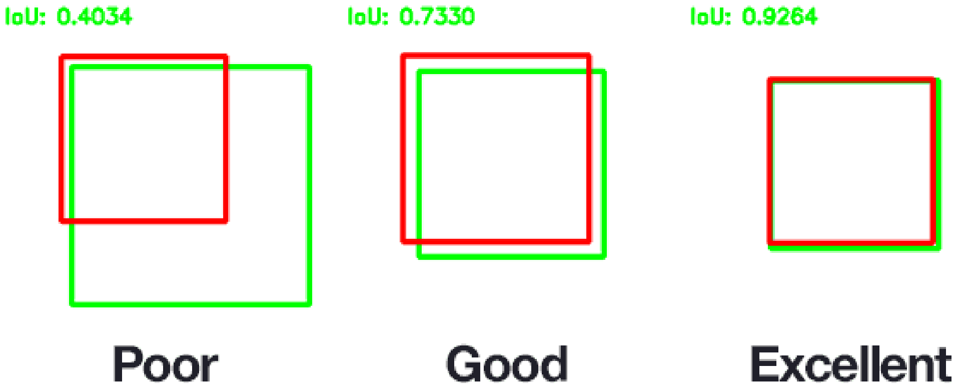


Fig.9 IOU metric

As Fig.9 shows, the IoU score ranges from 0 to 1, where:

- A score of 0 means no overlap between the predicted and ground truth regions.

- A score of 1 indicates a perfect match, where the predicted region precisely matches the ground truth region.

3.5.2. Sensitivity and Specificity and ROC Curve

Sensitivity (True Positive Rate) measures the model's ability to correctly identify all positive instances in the dataset, i.e., how well it captures the relevant regions. Specificity measures the model's ability to correctly identify all negative instances, i.e., how well it excludes irrelevant regions from the segmentation. These two metrics are complementary to that of IoU while measures like F-score are not reported since it gives almost the same results of IoU. Here's a simple explanation of the ROC curve:

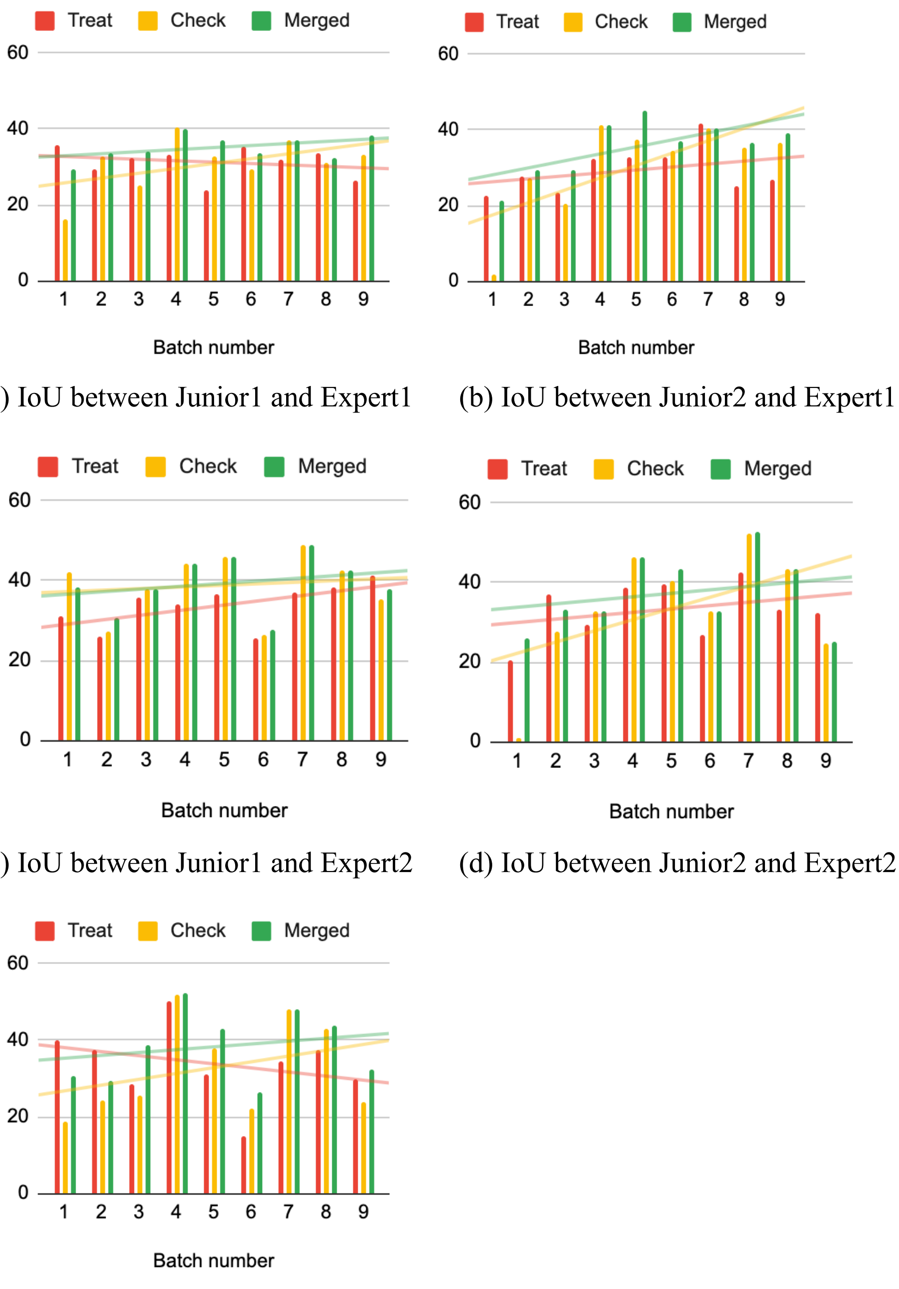
- True Positive Rate (TPR) or Sensitivity: This is the percentage of actual positive instances (true positives) that the model correctly predicts as positive. In other words, it measures how well the model identifies the positive class.
  - TPR = (True Positives) / (True Positives + False Negatives)
- False Positive Rate (FPR): This is the percentage of actual negative instances (true negatives) that the model incorrectly predicts as positive. In other words, it measures how often the model makes false alarms for the negative class.
  - FPR = (False Positives) / (False Positives + True Negatives)

The ROC curve is created by plotting sensitivity on the y-axis and 1-specificity (FRP) on the x-axis. Each point on the curve represents a different test (here we have 4 different models trained on 4 different datasets). A perfect model would lie on the top-left corner of the plot, meaning it achieves a high sensitivity (high TPR) and also a high specificity (low FPR). The diagonal dashed line from the bottom-left to the top-right represents random guessing (no predictive power). The area under the ROC curve (AUC-ROC) is also commonly used to quantify the overall performance of the model. A higher AUC-ROC indicates a better model, with an AUC of 1 indicating a perfect model, and an AUC of 0.5 indicating a model that performs no better than random chance.

4. Data Analysis and Results

4.1. Data Analysis

We begin by reviewing the data's validity and the annotation pipeline. As depicted in Fig.10, the IoU calculations between pairs of annotators reveal a notable trend: the agreement among annotators is consistently improving across almost every pair. While initially, we anticipated this improvement to primarily benefit junior surgeons, it's evident that even the expert annotators are converging in their assessments. It's essential to emphasise that this analysis is solely based on observing the trendlines.



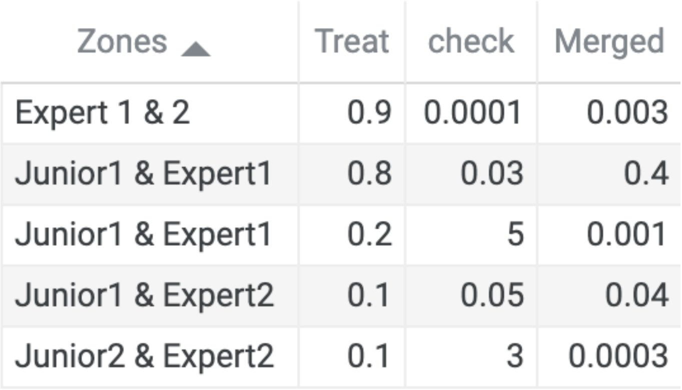
(a) IoU between Junior1 and Expert1 (b) IoU between Junior2 and Expert1

(c) IoU between Junior1 and Expert2 (d) IoU between Junior2 and Expert2

(e) IoU between Expert1 and Expert2

Fig.10. The agreement between annotators in every cycle (called batch here)

In addition to computing the trendlines, we took an extra step to determine whether the enhancement in agreements between pairwise annotators holds statistical significance. For this reason, we have used the Mann-Whitney U test since the data is not following a normal distribution. Since the number of images in every batch is only 15, we combined each 2 batches to increase the number of data populations. The Mann-Whitney U test shown in Table.3 shows that we had significant improvement in whether the check or the merged zones between annotators agreement from the first batches (1 and 2) to the last ones (8 and 9).



Zones ▲	Treat	check	Merged
Expert 1 & 2	0.9	0.0001	0.003
Junior1 & Expert1	0.8	0.03	0.4
Junior1 & Expert1	0.2	5	0.001
Junior1 & Expert2	0.1	0.05	0.04
Junior2 & Expert2	0.1	3	0.0003

Table.3 The P-values of the Mann-Whitney U test showing the improvement of agreement from batches 1&2 and batches 9&10.

This analysis verifies that the annotators are reaching a consensus through the defined annotation pipeline and that the quality of the final data can be trusted.

4.2.Performance Metrics Results

We use the images of batches 8 and 9 and their consensus annotations as ground truth. This test set is the same for all the four models. The results of the segmentation of algorithms on both treat and check zones and their averages are shown in Fig.11 for every one of 4 models. It is obvious that when the number of data increases the accuracy of the model also increases. To have an evaluation of the machine vs human performance, the result of the annotators in batches 8 and 9 are also depicted in Fig.12 This means that the annotators' annotations IoU with the consensus on the same test set are calculated so that it can be compared by the algorithm results. Although the machine is not performing as a real surgeon, the results on such preliminary data seems to be promising. With the growth of the dataset, it can be predicted that the algorithm results can soon reach that of the surgeons.

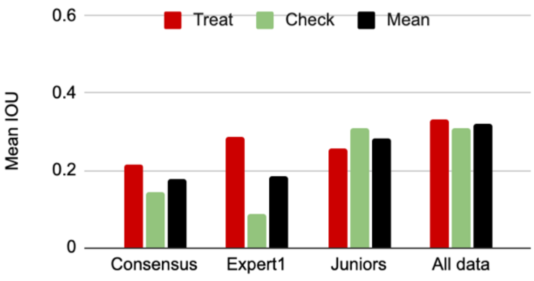


Fig.11. Test IoU of the 4 models.

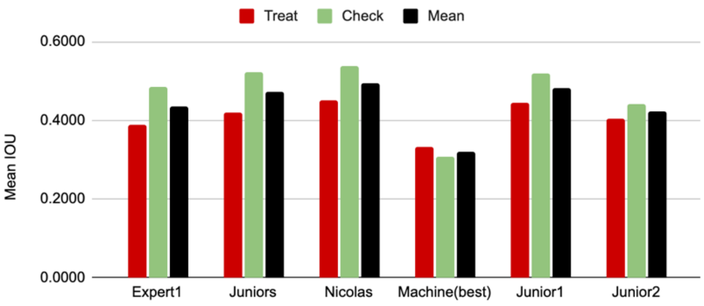


Fig.12. Test IoU of 4 annotators compared with machine.

In Fig.13 (below), the ROC curve is plotted for all the models. Fig.13(b) is a zoom on Fig.13(a). (a) shows that the AUC is above 0.5 (all models fall above diagonal line) so the algorithm is performant. The next important note is that the specificity is always very high. This is the realisation of our expectations. We do not want the machine to have a lot of false positives giving false signals to the surgeon. On the class treat, the model trained on expert1 annotations has the best sensitivity, while the specificity is not as high as other methods. So, it gives us the best true positive rate. We also see that, considering both treat and check, the model trained on consensus is a good balance between sensitivity and specificity.

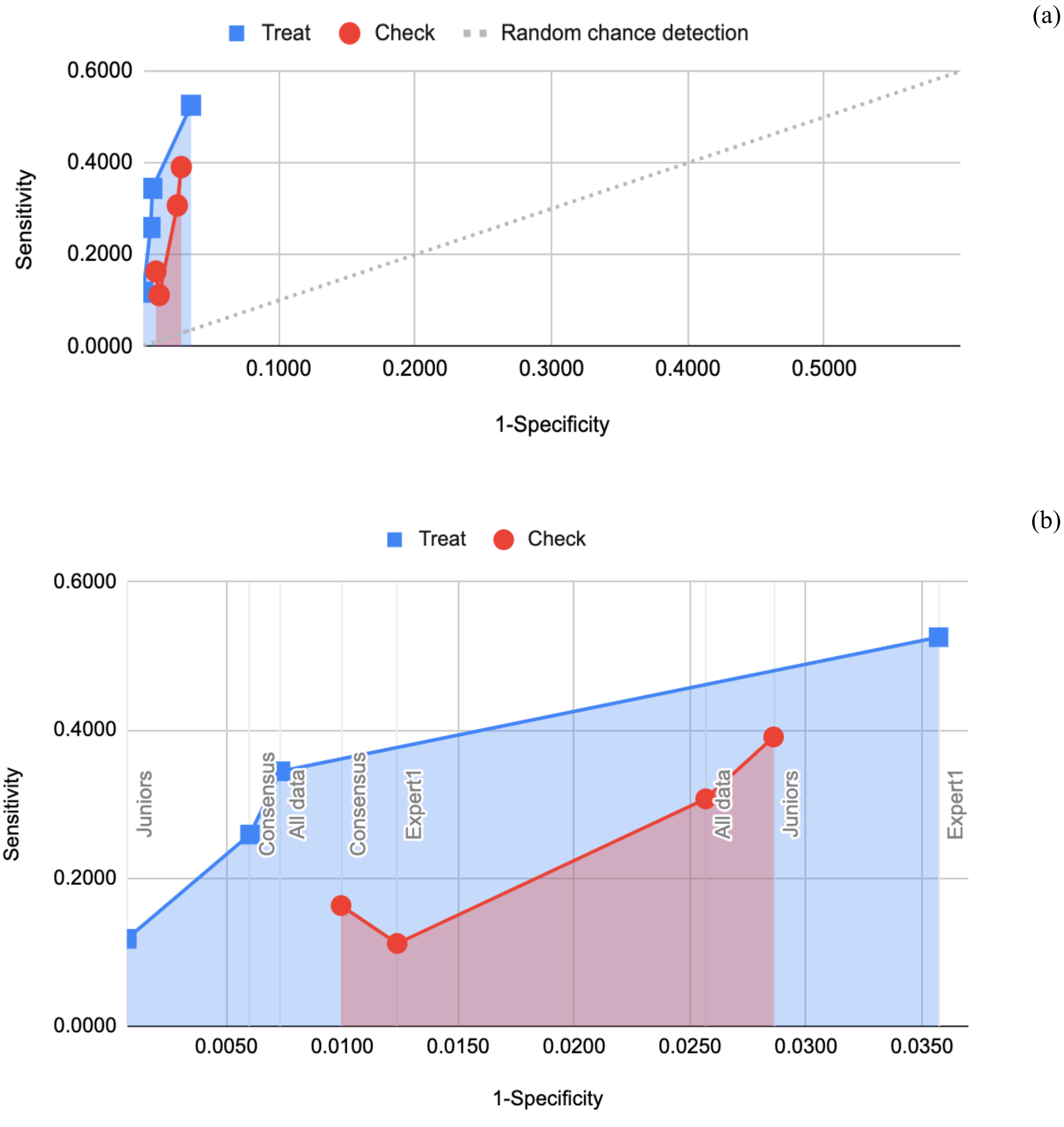
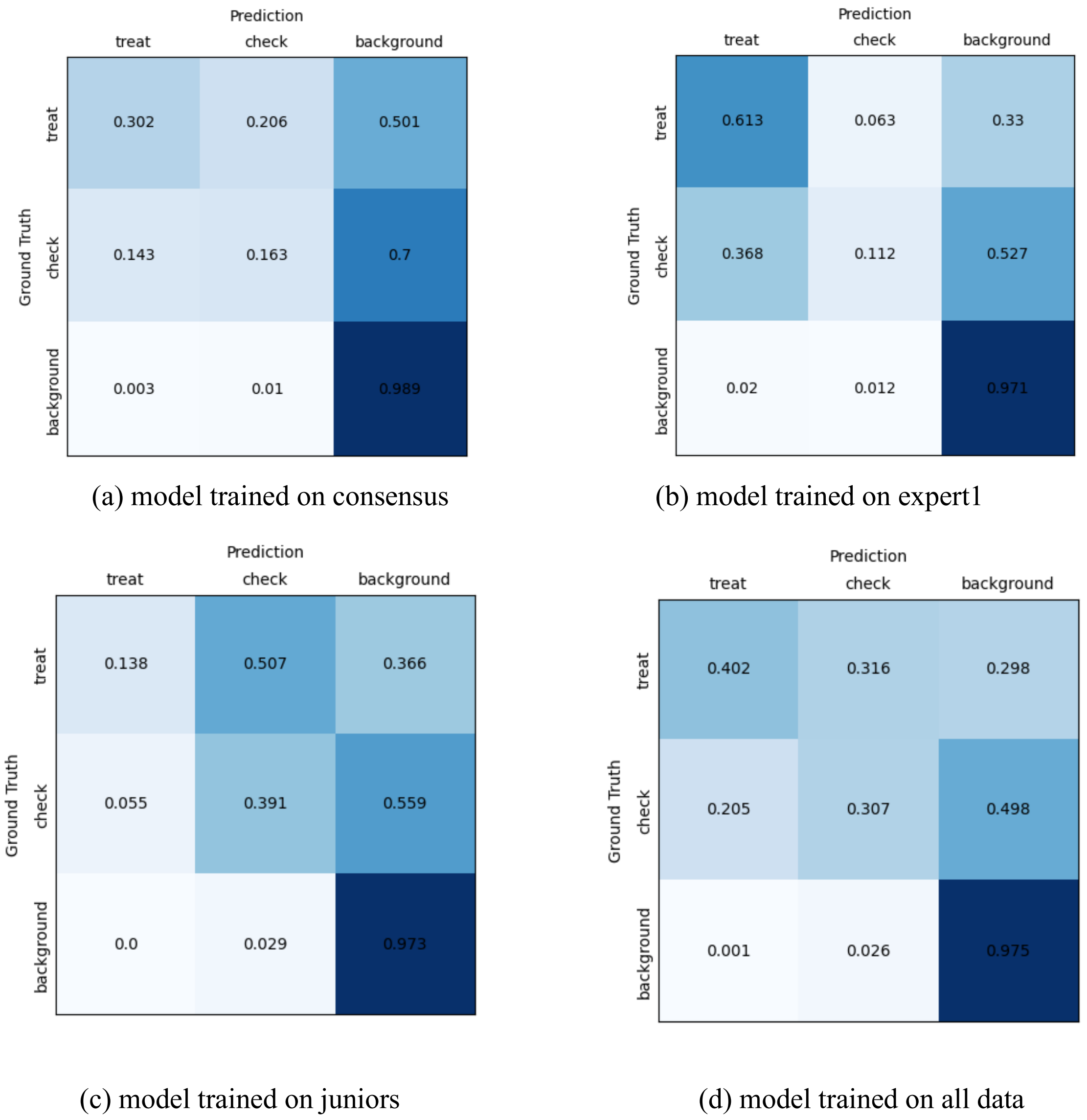


Fig.13. The ROC curve, based on all trained models. (b) is the zoom of (a).

Finally, the confusion matrices are calculated for all the trained models. Fig.14. shows these matrices. As expected, the results show that the background is not detected wrongly most of the time (3rd row of the matrices), meaning a low false positive rate. In more than 60% of the time the expert model prediction was accurate in treat zone. Examining the diagonal elements of the matrix, where an ideal model would have values of 1, it becomes apparent that the model based on all data performs the best. In this context, it's observed that the 'check' and 'treat' zones may occasionally be mistakenly interchanged, but there's a relatively low rate of background predictions when the actual class is 'treat' or 'zone'. Regarding the first row of the junior matrix, although 'treat' is predicted with some inaccuracies, it's worth noting that it is incorrectly classified as a 'check zone'. This outcome, while not ideal, still serves as a valuable warning for the surgeon. Notably, the rate of missed detections (i.e., predictions classified as background) is relatively low, at only 36%.



(a) model trained on consensus

(b) model trained on expert1

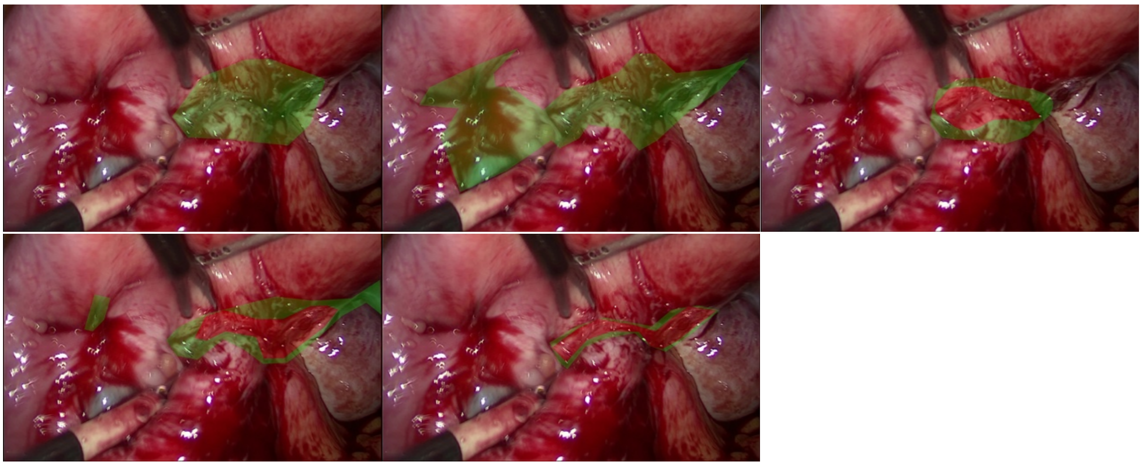
(c) model trained on juniors

(d) model trained on all data

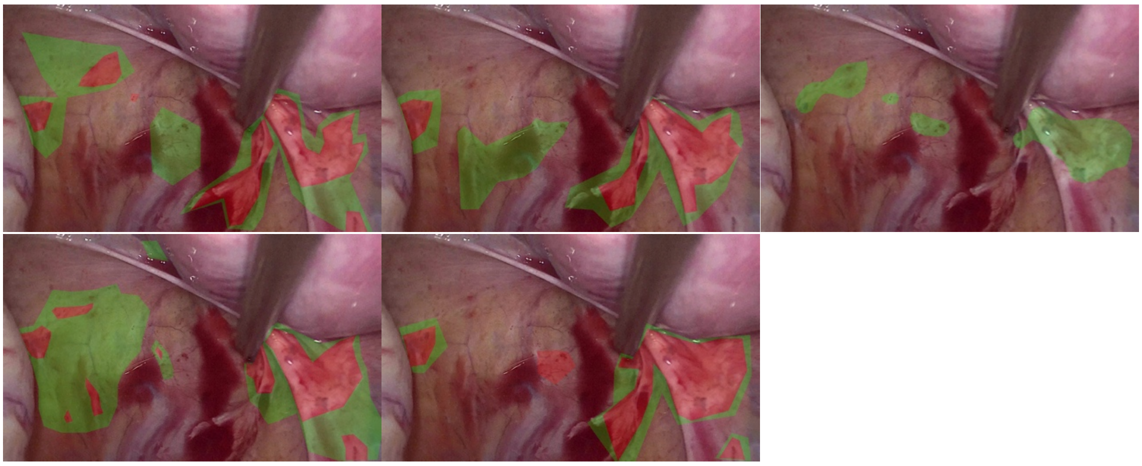
Fig.14. The Confusion matrices

4.3. Visual Results

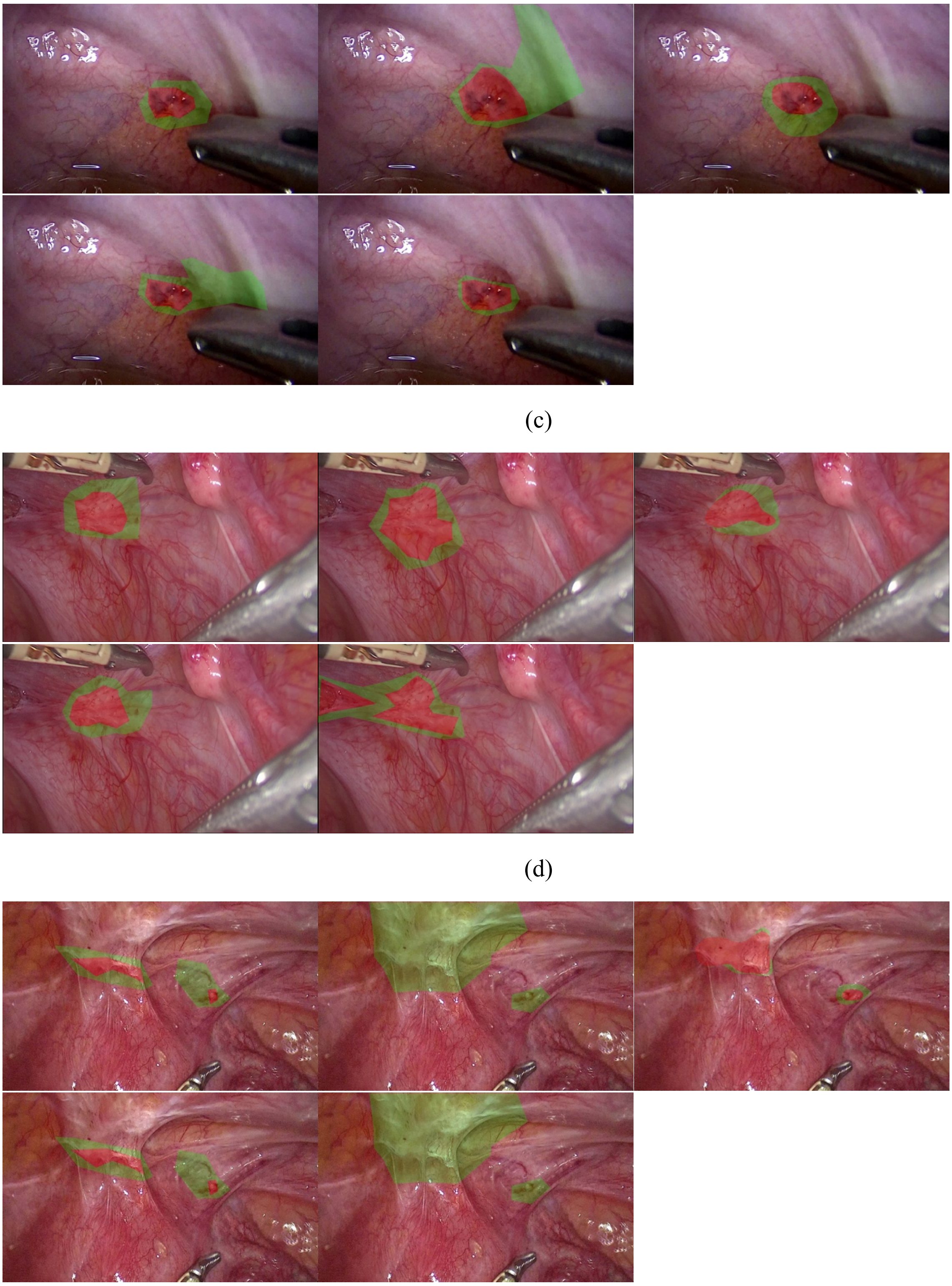
Finally, some of the detections are depicted (Fig.13) to be evaluated visually by the interested audience. The results had attracted the attention of annotator surgeons.



(a)



(b)



(c)

(d)

(e)

Fig.13. The visual results of algorithm prediction. Top row from left to right: exper1, expert2, algorithm. Bottom row from left to right: junior1, juniort2.

5. Conclusions and Discussions

Our preliminary results of applying a Deep Learning-based semantic segmentation system in finding endometriosis incision zones are very promising, even with a limited amount of training data.

More images are necessary to train the deep learning system to improve the results obtained.

5.1. Discussion on Results

Algorithm vs Experts

The results of our study indicate that while machine-generated annotations exhibit a certain level of accuracy and reliability, they do not match the level of precision achieved by human experts in the annotation process. With a limited training set we obtained an IoU of 0,33 for the Treat Zone and 0,3083 for the Check Zone (average: 0,3208). So, with the growth of the dataset it can be predicted that also machine results can improve considerably. In the context of our research, it is evident that algorithm annotations, while promising, fall short of the quality that can be achieved by experts. This divergence in performance can be attributed to several factors:

- Complexity of incision boundary segmentation: Human experts possess the ability to interpret and incorporate subtle cues that may be challenging for machines to capture accurately.
- Varied and Unpredictable Data: Real-world data can present various challenges. Human annotators can adapt and make informed decisions based on their understanding, whereas machine algorithms may struggle to handle such variability effectively.
- Subjectivity and Context Sensitivity: Some segmentation tasks may involve subjective judgement calls or context-specific, which can be particularly challenging for machines lacking a comprehensive understanding of the world.
- Training Data Quantity: The performance of machine annotations is also influenced by the quantity and diversity of the training data used to develop the algorithm. Limited or biased training data can result in suboptimal performance.

We believe that the increase of the dataset size and adding more number and variety of annotations to the algorithm as input can solve almost all of the mentioned problems.

It is worth noting that while machine-generated detections offer scalability and efficiency advantages, they are not a complete substitute for human expertise yet, especially in tasks that require high precision and nuanced understanding. Therefore, in incision zone segmentation where accuracy and reliability are paramount, the involvement of human experts supervision remains essential to ensure the highest quality results. This is in fact one of the main facts of trustworthy AI. Machines can assist in automating such tasks and reducing human workload, but experts are indispensable for tasks demanding a higher level of domain knowledge and precision. We do not believe that the machines can replace experts at this stage, but they can be used under their supervision to improve the patient impact and the quality of care.

Quality versus Quantity of Data

The results of our study reveal intriguing insights into the performance the models trained on different input data sources. Notably, the model trained on consensus-labelled images, while having a considerably smaller dataset compared to the other models, exhibited competitive segmentation performance.

Firstly, it's essential to acknowledge that the number of consensus-labelled images in this model is notably fewer than in the other models. Despite this significant disparity in dataset size, the segmentation score achieved by the consensus-labelled data model is strikingly similar to that of the model trained on data annotated by a single expert (while the quantity of training dataset is more than 5 times less).

These results underscore the remarkable quality and effectiveness of consensus-labelled data. The fact that the consensus-labelled model can achieve comparable results with only a fraction of the data volume indicates that the collaborative annotation process, involving multiple experts, has yielded high-quality annotations. This lends support to the notion that consensus-driven labelling can enhance the reliability and accuracy of the training dataset, compensating for the limited quantity of data.

In summary, our findings suggest that while consensus-labelled data may be limited in quantity, its quality and the collaborative effort that goes into its annotation can be invaluable for training robust segmentation models. The ability of the consensus-labelled model to rival the performance of models trained on larger datasets underscores the significance of quality annotations and highlights the potential for optimising data collection strategies in semantic segmentation tasks. This is in line with trustworthy AI concepts to have the best quality data. The highest score achieved when trained on all data is totally normal, arising from the fact that deep learning techniques performance are highly correlated with the amount of data.

5.2.Limitations and Future Directions

5.2.1. Limitations

Firstly, in literature there is no systematic taxonomic guideline for the endometriosis surgeries, which may present different phenotypic aspects. Only through the adoption of a shared language and collaboration between specialists, an adequate training of the machine is possible.

Secondly, Annotating this kind of images is extremely time-consuming for an expert. A major difficulty with AI training is its requirement for a massive amount of manually annotated data. This requires attention, time and expertise, making our existing datasets extremely valuable. That is why we did our best to compensate for the quantity with the quality of the data.

Afterwards, The lack of standardisation in the type of treatment. As also reported by the ESHRE 2022 guidelines [8]: 'due to the heterogeneity of patient populations, surgical approaches, preferences, and techniques, the GDG decided not to make any conclusions or recommendations on the techniques to be applied for treatment of deep endometriosis'. For this reason, our aim was not to suggest the operational technique to treat a lesion, but to aid the surgeon in recognizing a lesion and its treatment boundaries.

5.2.2. Perspectives

Future research may lead to improving the algorithm and assess the possibility to integrate the software in a real-life laparoscopic setting to help the surgeon to improve surgical performance.

Another possibility is the future annotation of the type of the lesion, so that the algorithm results can be more accurately analysed by the type of the lesion.

Another future work is to include all data in the dataset for the algorithm but to weight the data based on their annotators: more weights on experts' annotations and less on juniors'.

References

[1] C. B. Wykes, T. J. Clark, and K. S. Khan, "REVIEW: Accuracy of laparoscopy in the diagnosis of endometriosis: A systematic quantitative review," *BJOG: An International Journal of Obstetrics &amp; Gynaecology*, vol. 111, no. 11, pp. 1204–1212, Oct. 2004, doi: 10.1111/j.1471-0528.2004.00433.x.

[2] C. Bafort, Y. Beebeejaun, C. Tomassetti, J. Bosteels, and J. M. Duffy, "Laparoscopic surgery for endometriosis," *Cochrane Database of Systematic Reviews*, vol. 2020, no. 10, Oct. 2020, doi: 10.1002/14651858.cd011031.pub3.

[3] A. A. Pozdeev, N. A. Obukhova, and A. A. Motyko, "Anatomical Landmarks Detection for Laparoscopic Surgery Based on Deep Learning Technology," in *2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus)*, Jan. 2021. Accessed: Sep. 18, 2023. [Online]. Available: http://dx.doi.org/10.1109/elconrus51938.2021.9396093

[4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, Apr. 2018, doi: 10.1109/tpami.2017.2699184.

[5] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," *arXiv.org*, Jun. 17, 2017. https://arxiv.org/abs/1706.05587

[6] vdumoulin, "GitHub - vdumoulin/conv_arithmetic: A technical report on convolution arithmetic in the context of deep learning," *GitHub*. https://github.com/vdumoulin/conv_arithmetic (accessed Sep. 19, 2023).

[7] I. Berrios, "DeepLabV3 - Building blocks for Robust Segmentation," *Medium*, Jun. 05, 2023. Accessed: Sep. 19, 2023. [Online]. Available: https://medium.com/@itberrios6/deeplabv3-c0c8c93d25a4

[8] "Endometriosis guideline." https://www.eshre.eu/Guideline/Endometriosis (accessed Sep. 20, 2023).