International Journal of Contemporary Issues in Social Sciences
Volume 3, Issue 1, 2024 ISSN(P):2959-3808 | 2959-2461

CONTEXTUAL PARADIGMS IN LAPAROSCOPY: SURGICAL TOOLS DETECTION & SHAPING THE FUTURE OF SURGICAL PRECISION

Maryam Hassan*[1], Dr. Fouzia Jabeen[2], Dr. Fouzia[3]

[*1] Computer Science Department Shaheed Benazir Bhutto Women University Peshawar, Pakistan
[2] HOD, Department of Computer Science Shaheed Benazir Bhutto Women University Peshawar, Pakistan;
[3]Lecturer, Department of Computer Science Shaheed Benazir Bhutto Women University Peshawar, Pakistan.

[*1] mariamh0796@gmail.com, [2]fouzia.jabeen@sbbwu.edu.pk, [3]fouzia.idrees@sbbwu.edu.pk
Corresponding Author: *

Received: 10 November, 2023 **Revised:** 20 December, 2023 **Accepted:** 25 December, 2023 **Published:** 10 January, 2024

ABSTRACT

Context-aware surgery is an emerging field within Human-Computer Interaction (HCI) that aims to enhance surgical procedures by integrating contextual information, such as surgical phases and tools recognition, into the surgical environment. This critical review examines the current state of research in context-aware laparoscopic surgery, exploring key studies, technological advancements, evaluating its potential benefits, challenges, and future directions, providing an in-depth analysis of their contributions and limitations. By critically assessing the current landscape, this review sheds light on the opportunities and obstacles in realizing the full potential of integrating context aware system in laparoscopic surgeries & serves as a foundation for researchers and practitioners interested in advancing the field, fostering innovation, and improving surgical outcomes through the integration of contextual information.

Keywords: context-aware laparoscopic surgeries, HCI in laparoscopic surgery, deep learning image recognition & classification

1. INTRODUCTION

Laparoscopic surgery, also known as minimally invasive surgery, is a technique where small incisions are made and specialized tools are used to perform surgical procedures, often with the assistance of a camera for visualization. Context-aware laparoscopic surgery in Human-Computer Interaction (HCI) refers to the integration of technology and interactive systems that take into consideration the specific context of laparoscopic surgery to enhance the surgical workflow, decision-making, and overall patient outcomes [15].

Context-aware laparoscopic surgeries involve integrating real-time contextual information, data, and technology into minimally invasive surgical procedures. This information can include preoperative imaging, patient vitals, anatomical structures, instrument positions, and more. By presenting this information in a comprehensive and intuitive manner, surgeons can make informed decisions and adjustments during the surgery [15].

Context-aware surgical tools detection in laparoscopic surgeries aims to revolutionize minimally invasive procedures by offering real-time tracking and awareness of surgical instruments. This technology enhances patient safety by preventing accidents, optimizing precision through instrument positioning data, and reducing human errors in instrument handling. It also streamlines surgical workflows, supports training and education, and provides valuable data for analytics and improvement. Ultimately, these advancements improve patient outcomes, potentially reduce healthcare costs, and may minimize radiation exposure in laparoscopic surgeries. Additionally, this technology can integrate seamlessly with robotic surgical systems, further enhancing precision and efficiency in the operating room, making it a promising and transformative development in modern surgery [15].

This review endeavours to analyze the body of research conducted over the past five years pertaining to context- aware systems which are designed to detect surgical tools in surgeries with the aim to facilitate surgeons during surgeries,

facilitate education purposes, enhance the accuracy of the procedures, reduce the workload and stress on surgeon during surgery, keeping track of patient vitals. The examination of technological progress elucidated in these studies shall furnish valuable discernments concerning the advantageous attributes, implementation modalities, inherent constraints, utilization of deep learning frameworks, levels of precision and accuracy achieved, and prospective trajectories of these systems.

The primary objective of this review paper is to provide a comprehensive analysis of contemporary and recent methodologies employed in the detection of context-aware surgical tools during laparoscopic surgeries. This endeavor aims to furnish researchers with profound insights into this domain. Furthermore, this review offers a valuable resource for interested readers, enabling them to derive inspiration for potential research directions. Additionally, individuals may contribute to the betterment of existing systems and the enhancement of their performance, or they may aspire to elevate the field to new heights, thereby making a positive impact on society.

The document is organized as follows. Section no II explores the recent trends and advancements done in the field of surgical tools detection in context-aware laparoscopic surgeries. Section III examines the evaluation of the technologies on the basis of core metrices. Section IV elaborates the improvement suggestions, potential significance and real-world applications. Section V addresses the future directions and what can be contributed for the improvement of current technologies. Section VI comprises the conclusion, this section outlines each paper's contributions, objectives, methodologies, and outcomes.

2. OVERVIEW OF THE RECENT TRENDS

This section provides a comprehensive review of the technological advancements that have been achieved in the domain of detecting surgical instruments within the context of context-aware laparoscopic surgeries. These advancements encompass a range of sophisticated methodologies, including Deep Learning for Context Detection, a Spatial-Temporal Framework based on Deep Learning, the application of Attention Networks, the utilization of a Convolutional Neural Network (CNN)-based model for Surgical Instrument Recognition, the incorporation of CNN Networks for the purpose of Gauze Detection and Segmentation, the deployment of Transformer Architecture for Surgical Tool Detection in Laparoscopic Videos, the utilization of the P-CSEM Module to enhance the detection of surgical tools, and the classification of surgical tools employing a CNN-based approach with Transfer Learning. Each one of them are discussed in detail below:

(1) Contextual detector of surgical tools in laparoscopic videos using deep learning

Babak et al. proposes an automated surgical tool detection system, LapTool-Net, aimed at identifying surgical instruments in laparoscopic surgeries. This system eases the burden on surgeons, enhances accuracy, and aids in training. LapTool-Net utilizes a decision policy based on a Recurrent Convolutional Neural Network (RCNN) to predict tools in real-time, eliminating the need for handcrafted rules. Trained on publicly available datasets (M2CAI16 and Cholec80), LapTool-Net achieved impressive accuracy, with online and offline modes reaching exact match accuracies of 80.95% and 81.84% for M2CAI16, and 85.77% and 91.92% for Cholec80, respectively [1].

(2) A deep learning spatial-temporal framework for detecting surgical tools in laparoscopic videos

Tamer et al. propose a computer-assisted system for the detection of surgical tools in laparoscopic videos, aiming to address the challenges posed by factors like camera movement, smoke, tissue variability, and blood. They advocate for the use of deep learning, particularly convolutional neural networks (CNNs), to enhance tool recognition. Their approach involves a spatial-temporal solution employing a cascade of two LSTM models: the first, LSTM-clip, model's temporal dependencies in short video sequences, while the second, LSTM-video, extends this modeling to the entire surgical video. By explicitly considering temporal information, they seek to improve tool detection accuracy. The system is evaluated on the Cholec80 dataset, and performance is assessed using the average precision (AP) metric, comparing favorably to state-of-the-art methods. This proposed system aims to contribute to the advancement of surgical practice by facilitating more precise tool detection

in laparoscopic surgery, with potential implications for enhanced patient care [2].

(3) Attention Networks for Improving Surgical Tool Classification in Laparoscopic Videos

Herag et al. introduces attention modules, namely Squeeze and Excitation (SE) and Convolutional Block Attention Module (CBAM), as enhancements to convolutional neural networks (CNNs) to improve tool classification performance. The research employs the ResNet50 CNN architecture with pre-trained weights as a base model and evaluates its performance using the Chlolec80 dataset and the average precision (AP) metric. Additionally, it employs Gradient-weighted Class Activation Mapping (Grad-CAM) for prediction visualization. Previous studies are cited to support the notion that attention modules can enhance the performance of base models. In summary, the study aims to assess the impact of attention modules on laparoscopic tool classification, contributing to the growing field of AI in surgical applications [3].

(4) Development and Validation of a Model for Laparoscopic Colorectal Surgical Instrument Recognition Using Convolutional Neural Network–Based Instance Segmentation and Videos of Laparoscopic Procedures

Daichi et al. have engineered an instance segmentation model, capable of concurrently identifying eight distinct categories of surgical instruments that find frequent application in laparoscopic colorectal surgeries. This study, which primarily seeks to enhance quality, has yielded promising outcomes, showcasing the model's proficiency in achieving accurate recognition. Notably, this high level of recognition accuracy has persisted even when confronted with an expanded array of surgical instrument categories. It is imperative to underscore that the ability to discern surgical instruments holds pivotal significance as a foundational technological advancement across diverse domains within surgical research and development. The proposed model possesses versatile utility. It extends beyond its immediate application for automatic monitoring of surgical progress within the context of this study. Rather, its potential reaches into the realm of future endeavours concerning computer-assisted surgery and the overarching aspiration of surgical automation. By underpinning these facets of surgical innovation, our model stands as a potent tool poised to catalyse transformative developments in the field of surgery [4].

(5) Gauze Detection and Segmentation in Minimally Invasive Surgery Video Using Convolutional Neural Networks

Guillermo et al. proposes the development and use of machine learning models for the detection and localizationof surgical gauze in endoscopic video signals during laparoscopic surgery. The primary goal is to automate the detection of gauze, which is a crucial task in the operating room. Detecting gauze in laparoscopic surgery can help automate various tasks, such as assisting autonomous robotic systems and segmenting laparoscopic videos into different surgical phases. The proposal includes the creation of a dataset for training machine learning models to detect and locate surgical gauze, and mentions the use of several computer vision machine learning architectures like YOLOv3, convolutional backbones, and U-Net for this purpose. The ultimate aim is to improve the efficiency and safety of laparoscopic surgeries by automating the detection of surgical gauze and providing information about the stage of the operation [5].

(6) LapFormer: surgical tool detection in laparoscopic surgical video using transformer architecture

Satoshi Kondo proposes a method called LapFormer for the detection of surgical tools in laparoscopic surgery videos. What makes LapFormer novel is its utilization of a Transformer architecture, a type of neural network architecture known for its success in natural language processing, to analyze inter-frame correlations in videos. This approach deviates from the traditional use of recurrent neural networks in this context. The authors evaluate LapFormer on a dataset called Cholec80, containing 80 videos of cholecystectomy

surgeries, and find that it outperforms conventional methods such as single-frame analysis with convolutional neural networks or multiple frame analysis with recurrent neural networks by significant margins in terms of macro-F1 score. The ultimate goal of this research is to enable computer-assisted surgery by analyzing the surgical workflow, which can support clinicians during surgery and in post-operation phases. Such analysis has various applications, including real-time warning, decision-making support, resource management in operating rooms, surgical report documentation, video database indexing, surgeon training, and skill assessment. This proposal seeks to enhance the accuracy and effectiveness of surgical tool detection, a crucial component of this broader surgical workflow analysis endeavor [6].

(7) P-CSEM: An Attention Module for Improved Laparoscopic Surgical Tool Detection

Herag et al. proposes the development and evaluation of an attention module called P-CSEM (Proposed Computer-aided Surgical Equipment Module) aimed at improving feature refinement and classification performance in the context of surgical tool recognition. The proposed model is designed to enhance the accuracy of detecting surgical tools from data captured using a laparoscopic camera. This tool recognition is essential for various applications, including surgical phase recognition, skills assessment, protocol extraction, and the management of operating rooms. The proposed P-CSEM module is integrated into the ResNet50 network architecture, which is pretrained on the ImageNet database. The authors compare the performance of P-CSEM with other attention modules like squeeze and excitation (SE) and convolutional block attention module (CBAM)using the Cholec80 database [7].

(8) Towards more efficient CNN-based surgical tools classification using transfer learning

Jaafar et al. presents a proposal which revolves around addressing the challenges associated with learning minimally invasive surgery (MIS) procedures through the use of deep learning and advanced neural networks. Minimally invasive surgery offers numerous benefits, including smaller incisions, reduced pain, lower risk of complications, shorter hospital stays, and improved cosmetic outcomes compared to traditional open surgery. However, learning MIS procedures is hindered by the absence of depth perception when viewing the operative area on a 2-D monitor and limited training opportunities due to patient safety concerns. To tackle these issues, the proposal suggests recording surgical videos and utilizing them for retrospective analysis, postoperative Surgical Quality Assessment (SQA), and training purposes. The main innovation proposed here is the development of a deep learning-based system for the classification of surgical tools in MIS videos, with the goal of creating a searchable database. This system involves preprocessing video frames, data augmentation to handle imbalanced data, and training a deep learning model. The proposed approach aims to make it easier for novice surgeons and postoperative controllers to navigate and access specific segments of MIS videos, ultimately enhancing the learning and assessment process in minimally invasive surgery. The proposal suggests evaluating the effectiveness of this approach in comparison to other methods [8].

Proposed Solutions	*Model Used*	*Objectives*	*Proposed Methodology*
LapTool-Net [1]	The model used in this research was LapTool-Net.	This research mainly focuses on improving surgical education and reducing the cost of analyzing the streaming videos of surgery by providing an automated solution for tools detection.	The researchers presented a novel approach called LapTool-Net which focuses on utilizing correlations between different tools and their context of usage. This approach employs a Recurrent Convolutional Neural Network (RCNN) to create a decision policy for a multilabel classifier.
Deep Learning Spatial-Temporal Framework [2]	Hhierarchical organised neural architecture consisting of a convolutional neural network (CNN)consisting of two LSTM models to explicitly detect surgical tools depending upon their space and time usage in laparoscopic surgeries, using images and videos.	This research focuses on the importance of image-based surgical tools presence detection for the purpose of fostering of intelligent application in future/modern operating rooms.	The proposed approach uses a hierarchical neural architecture with a CNN for spatial features from laparoscopic images. It employs two LSTM models, LSTM-clip for short video clips and LSTM-video for entire surgical videos, to capture temporal dependencies.
Attention Networks for Surgical Tools Classification [3]	CNN architecture of ResNet50 is used as base model along with two attention modules, Squeeze and Excitation & Convolutional Block Attention Module.	To provide a robust and high-performing application of surgical tools detection & classification in laparoscopic videos using attention networks.	Three models were evaluated to see how attention modules impact surgical tool classification in laparoscopic videos. The first model, called the base model, used ResNet50 architecture and replaced the Softmax layer with sigmoid activation. The second model added the SE attention module, and the third model added the CBAM attention module to the base model.
Model for Laparoscopic Colorectal Surgical Instrument Recognition [4]	The authors presented "Mask R-CNN" architecture for instance segmentation, and the backbone network employed in this model is "ResNet-50."	This research aims to study that Is it possible to create an automated surgical instrument recognition system that boasts broad versatility across various instrument types and delivers pinpoint accuracy at the pixel level?	The methodology aims to improve procedures ethically. Surgical videos become annotated static images, with eight instruments manually labeled. A CNN performs precise object recognition, aided by data augmentation. Code is on GitHub, and training uses specialized hardware, covering data collection, annotation, instance segmentation, and model evaluation to achieve study goals.

Gauze Detection and Segmentation in MIS [5]	The model used in this research for medical instruments detection, specifically for gauze segmentation, is YOLOv3 (You Only Look Once version 3).	The objective of this paper is to present a model that is designed to automate the detection of surgical gauze within endoscopic video signals.	The proposed method encompasses a multifaceted approach involving data preparation, object detection, coarse and semantic segmentation of gauze within laparoscopic videos, and the deployment of diverse CNN architectures. The overarching goal is the development of automated models for the precise detection of surgical gauze in medical videos, with direct applications in surgical assistance and video analysis.
LapFormer [6]	The model used in this research is called "LapFormer." LapFormer is a deep learning model designed for detecting the presence of surgical tools in laparoscopic surgery videos.	The objectives of the authors of this paper is to design a model for detecting the presence of surgical tools in laparoscopic videos	The proposed methodology presents spatial analysis using a Convolutional Neural Network (CNN) module with additional temporal analysis using a Transformer module. This Transformer module incorporates correlations between input feature vectors to obtain the final feature vector. It highlights that the use of a Transformer architecture for analysing laparoscopic surgery videos is novel and, to the authors' knowledge, has not been proposed before. The experiments conducted demonstrate that LapFormer outperforms previous deep learning models in the task of surgery tool detection.
P-CSEM [7]	The authors proposed a deep learning framework called the "P-CSEM," embedded with a custom attention module	It is proposed for refining spatial features in the context of surgical tool classification in laparoscopic surgery videos.	The P-CSEM approach utilizes convolutional neural networks (CNNs) combined with P-CSEM attention modules at different architectural levels to enhance feature refinement. The model's performance was evaluated using the Cholec80 database, a publicly available dataset, and achieved a mean average precision of 93.14%. Visualizations demonstrated the model's ability

			to focus more on features relevant to surgical tools. This proposed approach showcases the advantages of integrating attention modules into surgical tool classification models, enhancing robustness and precision in detection.
CNN-Based Surgical Tools Classification Using Transfer Learning [8]	The authors employed a fine-tuned convolutional neural network (CNN) to address the automated detection of surgical tools during surgical procedures	The objective of this paper is to address the issue of imbalanced data within the publicly accessible Cholec80 laparoscopy video dataset through the application of various data augmentation methodologies.	This study addresses data imbalances in the Cholec80 dataset through data augmentation and fine-tunes an InceptionResnetV2 model for surgical tool classification in laparoscopic surgery videos. The Cholec80 dataset consists of 80 laparoscopic cholecystectomy surgery videos performed by 13 surgeons, with a resolution of 1920x1080 pixels and a frame rate of 25 fps. Video durations range from 12 minutes to 1 hour 39 minutes, totalling over 51 hours of footage. The dataset is annotated for binary detection of surgical tools, including seven types.

Table 1 provides an overview of the research works under study

3. CRITICAL ANALYSIS OF KEY FACTOR

In this section of the research paper, a comprehensive critical analysis is conducted, focusing on the foundational aspects of the study. The examination primarily revolves around the data set employed for training and validation, meticulously evaluating its scope, quality, and representativeness. Additionally, the scrutiny extends to the accuracy measurements deployed within the study, emphasizing their reliability and relevance in assessing model performance. A comparative analysis is also intricately detailed, juxtaposing both common and distinctive elements inherent in the examined data sets, thereby elucidating nuanced insights into their respective characteristics. This holistic evaluation serves as a cornerstone for substantiating the study's methodologies, enriching the understanding of the chosen models, and elucidating the robustness of the findings derived from the research.

i. Data Used for Training & Validation: Frequency distribution and proportion of the respondents showing their willingness to act

The overall system described in this study, by Babak et al., is based on the M2CAI161 tool detection challenge dataset, a subset of the Cholec80 dataset. They opted for the smaller dataset to accentuate the improvements resulting from the primary contributions of this paper. The dataset comprises 15 videos of cholecystectomy procedures, a surgical operation for gallbladder removal. Each video contains labels for seven tools every 25 frames, namely Bipolar, Clipper, Grasper, Hook, Irrigator, Scissors, and Specimen bags. Among these videos, ten were allocated for training, and five for validation. It is noteworthy that the type and shape of all seven tools remain consistent between the training and validation sets. Given that our study utilized the publicly available Cholec80

dataset for training and testing our deep-learning model, it is important to highlight that Institutional Review Board (IRB) approval was not required for this study. The training methodology incorporates simultaneous tool and combination detection, with distinct losses and optimized layers, while post-processing involves an RNN-based approach to align RCNN predictions with long-term tool ordering, enhancing classification accuracy. This approach facilitates efficient training and robust performance in both online and offline modes [1].

This study by Tamer et al. utilized the Cholec80 dataset, curated by A. P. Twinanda et al. [9], which comprises endoscopic videos capturing 80 cholecystectomy procedures conducted by 13 surgeons at the University Hospital of Strasbourg. These videos were recorded at a rate of 25 frames per second (fps) and featured continuous labeling of surgical phases, as well as tool labeling at 1 Hz. Consequently, each frame labeled with a surgical tool is flanked by 48 unlabeled frames. The surgical procedures involved the use of seven distinct tools: Grasper, Bipolar, Hook, Clipper, Scissors, Irrigator, and Specimen bag. The videos exhibited a median duration of 2095 seconds, with a range spanning from 739 seconds to 5993 seconds, characterized by quartiles at 1641 and 2882 seconds. The evaluation employed a six-fold Monte Carlo cross-validation (MCCV) methodology. In each fold, 40 randomly chosen full videos were allocated for training the CNN, LSTM-clip, and LSTM-video models, while the remaining40 videos were reserved for testing and validation of model predictions. Two experiments were conducted, utilizing different CNN architectures, namely VGG16 and ResNet-50, as foundations for advanced RNNimplementations. These experiments resulted in four distinct approaches: VGG-LC, VGG-LC-LV, ResNet-LC, and ResNet-LC-LV, reflecting the use of VGG-16 or ResNet-50 model outcomes (visual features) in LSTM models. Each approach underwent six separate training iterations, corresponding to the six MCCV folds, using the respective training data [2].

Herag et al., used the Cholec80 database, as outlined in reference [9], comprises a compilation of 80 cholecystectomy procedures meticulously recorded at the University Hospital of Strasbourg. These surgical procedures were captured at a consistent frame rate of 25 frames per second (fps), while tool annotations were recorded at a frequency of 1 fps. The dataset encompasses seven distinct surgical tools, including Grasper, Hook, Bipolar, Scissors, Clipper, Irrigator, and Specimen Bag. Importantly, the annotation criteria stipulated that a minimum of half of the tool tip should be observable within each video frame for annotation purposes [3].

In the research of Daichi et al., a dataset of laparoscopic colorectal surgical videos spanning from April 1, 2009, to December 31, 2021, was utilized. These intraoperative videos were converted to the MP4 format, maintaining a display resolution of 1280 × 720 pixels and a frame rate of 30 frames per second. From this dataset, a total of 38,628 static images capturing surgical instruments for recognition were randomly selected and incorporated into the annotation dataset. Recognition tasks focused on eight types of surgical instruments, including surgical shears, spatula-type electrodes, atraumatic universal forceps, dissection forceps, endoscopic clip appliers, staplers, grasping forceps, and a suction/irrigation system. The annotation process involved 29 non-physician individuals, supervised by two board-certified surgeons (D.K. and H.H.), who manually assigned annotation labels pixel by pixel. This meticulous process was facilitated using digital drawing pens and equipment such as Wacom Cintiq Pro, Wacom Pro Pen 2, Microsoft Surface Pro 7, and Microsoft Surface Pen [4].

The research conducted by Guillermo et al. was done on the dataset comprising of 42 video files: 30 with gauze (33 minutes) and 12 without gauze (13 minutes). Recorded with a STORZ TELECAM One-Chip Camera Head, PAL color system, f = 25–50 mm (2×) focal Zoom Lens, and 752x582 pixel CCD sensor. Notably, these videos do not depict actual patient surgeries. Instead, diverse laparoscopic scenarios were recreated in a simulator using on-live animal organs sourced from a certified slaughterhouse. The scenarios encompass gauze presence, absence, tool presence, clean and stained gauze, and multiple gauzes. Human- generated masks facilitated the classification of 100x100-pixel fragments, resulting in 168,126 fragments (80,918 with gauze, 87,208 without), vital for our research, and included in the dataset [5].

Satoshi Kondo utilized the Cholec80 dataset, which comprises 80 cholecystectomy surgery videos performed by13 surgeons. These videos were

recorded at 25 fps and included annotations for tool presence (at 1 fps) and phase annotations (at 25 fps). Only the tool presence annotations were used in the experiments. The evaluation employed fourfold dataset cross-validation, dividing the dataset into four groups with 20 videos each. Three groups were used for training and validation, while the fourth served as the test set. Within each fold, the 60 videos for training and validation were further split into 55 training videos and 5 validation videos. The training dataset was constructed using techniques like Label Powerset and under-sampling, with under-sampling applied exclusively to the training data. On average, the dataset sizes for training, validation, and testing in each fold were 15,909, 11,407, and 46,045, respectively [6].

Herag et al. chose Cholec80 database which is publicly accessible, serves as the primary resource for training and evaluating models. It includes 80 cholecystectomy procedures recorded at the University Hospital of Strasbourg at 25 Hz with tool annotations at 1 FPS. Seven surgical tools (grasper, hook, bipolar, scissors, clipper, irrigator, specimen bag) are featured. The database contains 184,498 image frames from the 80 videos. Figure 5 illustratesthese surgical tool tips within the Cholec80 dataset [7].

Jaafar et al. used the Cholec80 dataset which comprises 80 cholecystectomy surgery videos by 13 surgeons. Videos are in 1920x1080 resolution at 25 fps, with durations ranging from 12 minutes to 1 hour, 39 minutes, totaling 51+ hours. Each video is annotated for surgical tools. The dataset includes seven tools (e.g., specimen bag, scissors) with varying angles, resolutions, and occasional focus issues. Tool presence is marked if at least half is visible. Each image has binary labels for tool recognition, making it a valuable multi-label classification dataset. [8]

ii. Accuracy & Evaluation Metrics

In the study by Babak et al. the experiments were conducted using a Nvidia TITAN XP GPU [1].

LapTool-Net results on M2CAI dataset: It is readily apparent that the inclusion of temporal features within the RCNN model has yielded notable enhancements in both exact match accuracy and F1-macro scores, with improvements of 3.15% and 7.52%, respectively. Furthermore, the introduction of the LP decision model results in a further augmentation of the F1-macro metric, boasting an increase of 2.94% [1].

LapTool-Net results on Cholec80 dataset: The Cholec80 dataset contains 32 tool combinations, with 20 of them accounting for over 99.5% of video time. This diversity results from its larger size compared to the M2CAI dataset. However, the five additional super classes in Cholec80 represent less than 0.4% of frames. To ensure uniformity, we selected 1500 samples for each of the 20 primary tool combinations from a total of 30,000 frames. We used the same model as the M2CAI dataset for feature extraction, decision policy, post- processing, and training. This led to improved accuracy and F1-scores. Notably, the CNN's F1-macro score on the balanced Cholec80 dataset surpassed that of M2CAI by a substantial 9.19% [1].

In the study by Tamer et al., The six-fold experimental validation revealed substantial enhancements in overall classification performance through the adoption of LSTM-clip and LSTM-video methods. Specifically, ResNet- LC-LV and VGG-LC-LV exhibited mean mAP values of 94.74% and 91.64%, surpassing the mAP values of the established VGG-16 and ResNet-50 models at 89.17% and 92.00%, respectively. LSTM-clip played a crucial role in refining CNN model predictions across nearly all tool categories by modeling temporal information from adjacent frames. Likewise, leveraging temporal information throughout the entire video sequence led to significant improvements for all tools, with the exception of the grasper and hook [2].

In the research by Herag et al., the outcomes reveal that the incorporation of attention modules led to a notable enhancement in performance, with an average increase of approximately 2% compared to the base CNN model. This indicates that attention-based networks have the capability to selectively emphasize informative features while disregarding non-essential ones in the context of tool classification. It is evident from the results that, except for the Grasper class, the base ResNet50 architecture underperformed in comparison to models utilizing attention modules. This exceptional performance of the Grasper class can be attributed to its high representation in the training dataset. A similar pattern can be observed for the Hook class, where all three models exhibit nearly identical performance, contrasting with the improvements

seen in other classes. Notably, both attention modules, SE and CBAM, achieved comparable levels of performance [3].

The mean Average Precision for ResNet50 was 86.35%, ResNet50-SE was 88.38% & ResNet50-CBAM was 88.40%. [3]

In the study by Daichi et al., the assessment metric employed was the average precision, determined by calculating the area under the precision-recall curve. This metric effectively captures true-positive, false-positive, and false- negative instances, providing an aggregated measure of performance. The mean average precision was computed for eight distinct types of surgical instruments. The validation method employed was five-fold cross-validation, utilizing a dataset consisting of 337 laparoscopic colorectal surgical videos. Notably, the annotation process involved manual pixel-by-pixel annotation of 81,760 labels across 38,628 static images, forming the annotation dataset. The results indicate mean average precisions of 90.9% for three instruments, 90.3% for four instruments, 91.6% for six instruments, and 91.8% for eight instruments in the context of instance segmentation for surgical instruments [4].

Guillermo et al. conducted an assessment encompassing both quantitative and qualitative outcomes derived from models trained for three distinct tasks. To gauge the performance of YOLOv3, they employed the mean average precision (mAP) metric, complemented by additional metrics including Precision, Recall, F1 Score,

and Frames Per Second (FPS). The DarkNet-53 classifier yielded a precision of 94.34%, recall of 76.00%, F1 Score of 84.18%, mAP of 74.62%, and FPS of 34.94%. For the results pertaining to Gauze coarse segmentation, the authors explored the suitability of InceptionV3, MobileNetV2, and ResNet-50 models. Among these, ResNet-50 emerged as the top performer, achieving an average precision of 94.09%, followed by MobileNetV2 with 84.23%, and InceptionV3 with 75.67% [5].

Satoshi Kondo's [6] evaluation encompasses the rigorous scrutiny of precision, recall, and the F1 score, esteemed metrics deployed to meticulously gauge the efficacy of varied methodologies for the detection of surgical instruments. The discernible outcomes delineate the performance prowess of the proposed approach vis-à-vis established conventions, specifically ResNet50, ResNet50-LSTM, and ResNet50-GRU, predicated on the trifecta of precision, recall, and the F1 score. The scrutiny unfolds tool-wise, where each instrument undergoes meticulous evaluation, culminating in the derivation of macro-level averages, aggregating across all instruments for a holistic perspective. The authors embarked on an insightful comparative analysis, juxtaposing their innovative model against the efficacy of the time-honored ResNet50, ResNet50-LSTM, and ResNet50-GRU paradigms. The narrative underscores the salient observation that the proposed methodology, particularly the LapFormer-2 variant, engenders notable enhancements in detection acumen when appraised in terms of recall and F1 score across the entire gamut of surgical instruments. For instance, a conspicuous augmentation of 20.3, 17.3, and 17.3 units in macro-F1 is conspicuously elucidated when benchmarking against ResNet50, ResNet50-LSTM, and ResNet50-GRU, respectively. Furthermore, the discourse discerns that certain instruments, namely 'Scissors,' 'Clipper,' and 'Irrigator,' evince substantial escalations in detection proficiency with the incorporation of the proposed method, juxtaposed against the conventional methodologies. Nevertheless, a note of sagacity is sounded, acknowledging a scope for further refinement in the realm of 'Scissors,' as it trails behind its counterparts, registering the lowest F1 score within the purview of the proposed methodology [6].

The P-CSEM attention module, presented by Herag et al., demonstrated superior performance in all tool categories except for the irrigator class, as evident in the AP and mAP results presented in Figure 6[7]. Notably, the P- CSEM model showed a significant improvement, with a 7.5% increase in AP for the grasper tool and a 4% increase for the scissors tool compared to the base model. When integrated with P-CSEM, the model achieved a mAP of 93.14%, surpassing the SE and CBAM models, which scored 91.38% and 91.57%, respectively. This integration led to a 1.76% improvement over the base model. To assess model performance, the mean of the average precision (mAP) was calculated across all tools in the testing set. Additionally, a network explain-ability analysis using gradient-weighted class activation mapping (Grad-CAM) was conducted, following the methodology described in [10], to provide insights into model interpretability and visualization of class activation. Equation (1)

defines the computation of the evaluation metric for average precision (APc), where c represents the class number, and various thresholds were used to determine recall and precision values for each class [7].

In the study by Jafaar et al., the model's performance was evaluated using the average precision (AP) metric, a standard measure for object detection accuracy. Precision-Recall curves were employed, particularly suitable for imbalanced datasets. The mean Average Precision (mAP) was calculated to assess the overall model performance across all surgical tools. This mAP accounts for the varying precision-recall trade-offs for individual tools. The Average Precision for Grasper, Bipolar, Hook, Scissors, Clipper, Irrigator, Specimen Bag is 96.58, 95.04, 99.68, 95.11, 93.71, & 94.92 respectively. The mean Average Precision is 93.75. Jaffar et al. compared their work with previous studies by M.Sahu [11], EndoNet [9], AmyJ [12], Jo[13] , & Kanakatte [14]. Their model outperformed other architectures across most surgical tools, demonstrating significant improvements, particularly in the case of the scissors tool, which is traditionally underrepresented. This success can be attributed to the multiple data augmentation techniques employed and the deep architecture of InceptionResnet-v2 [8].

Model	Dataset	Evaluation Metrics	mAP (mean Average Precision)	F1-Score	F1-macro	F1-micro	Recall
LapTool-Net [1]	Dataset based on M2CAI161 tool detection challenge dataset, a subset of Cholec 80	Accuracy, F1-macro, & F1-micro I used for individual results of M2CAI and Cholec80 data set. For model's evaluation, precision, recall & F1 score is calculated	80.55%	83.50%	84.89% for M2CAI & 89.17 for Cholec80.	89.79% for M2CAI & 91.21% for Cholec80.	87.22%
Deep Learning Spatial-Temporal Framework [2]	Cholec 80 dataset comprising videos of 80 cholecystectomy procedures conducted by 13 surgeons at University of Hospital of Strasbourg	Average Precision is used to evaluate the performance of the model.	94.95%	Not mentioned	Not- Applicable	Not-Applicable	Not mentioned

Attention Networks for Surgical Tools Classification [3]	Cholec 80 dataset comprising compilation of 80 cholecystectomy procedures meticulously recorded at University of Hospital of Strasbourg	Average precision was calculated to judge the performance of the presented methods	88.38% for ResNet50-SE & 88.40& for ResNet50-CBAM	Not mentioned	Not- Applicable	Not-Applicable	Not mentioned
Model for Laparoscopic Colorectal Surgical Instrument Recognition [4]	Dataset of laparoscopic colorectal surgical videos, recorded from 1 April, 2009 till 31 Dec, 2021	Mean Average Precision was used as an evaluation metric for tool classification	90.9% for 3 instruments, 90.3% for 4 instruments, 91.6% for 6 instruments & 91.8 % for 8 instruments.	Not mentioned	Not mentioned	Not mentioned	Not mentioned
Gauze Detection and Segmentation in MIS [5]	Dataset comprising of 42 videos, 30 with gauze (33 mins) and 12 without gauze (13 mins) recorded with a STOR2TELECOM. Videos are related to scenarios of laparoscopic surgical procedures simulated on non-live animal organs.	Inception over Union (IoU) which was 0.85, Precision, Recall, & F1 Score, are used for evaluating the performance of ResNet-50, MobileNetV2& InceptionV3	ResNet-50, MobileNetV2, & InceptionV3 got 94.09 %, 84.23%, & 75.67 of precision respectively	ResNet-50, MobileNetV2, & InceptionV3 got 92.67 %, 80.28%, & 81.82 F1 score respectively	Not mentioned	Not mentioned	ResNe7t-50, MobileNet V2, & Inception V3 got 91.31%, 76.68%, & 89.08& of recall respectively
	cholecystectomy surgery videos performed by 13 surgeons.	as evaluation metrics for the performance of the models	got 75.3 % and 76.6% of precision respectively	got 68.1 % and 71.5 % F1 scores respectively			2 got 68% and 71.1% recall respectively
LapFormer [6]	Cholec 80 dataset was used comprising of 80	Precision, recall, & F1 scores are used	Lapformer1 and Lapformer2	Lapformer1 and Lapformer2	Not mentioned	Not mentioned	Lapformer 1 and Lapformer
P-CSEM [7]	Cholec 80 database consisting of 184,498 image frames form 80 videos	Mean Average Precision was used to evaluate ResNet50+P-CSEM, ResNet50 + SE, & ResNet50+ CBAM.	Precision with 3, 4, & 16 attention blocks are 93.14%, 91.01%, & 91.50% (these are the best performances of the models)	Not mentioned	Not mentioned	Not mentioned	Not mentioned
CNN-Based Surgical Tools Classification Using Transfer Learning [8]	Cholec 80 dataset comprising of 80 cholecystectomy surgery videos	Mean Average Precision	93.75%	Not mentioned	Not mentioned	Not mentioned	Not mentioned

Table 2 provides the comparison on the basis of each model's basic metrics

Comparison on the basis of common & unique characteristics:

This section compares all of the above-mentioned models on the basis of their common & unique characteristics, discuss their industrial use, & some improvement suggestions.

Common Points:

All the models are applied to the field of medical imaging, specifically in the context of laparoscopic or minimally invasive surgery, and aim to improve the recognition and understanding of surgical instruments or tools.Most of the models utilize publicly available datasets, such as the Cholec80 dataset, for training and evaluation, making their research findings accessible for the broader medical imaging community. The models use common performance evaluation metrics such as precision, recall, F1-score, mean average precision (mAP), and average precision (AP) to assess the effectiveness of their models. These models represent innovations in deep learning techniques, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), Transformers, and attention modules. All models have potential real- world applications in surgical training, computer-assisted surgery, surgical quality assessment, and patient safety.

Unique points:

Babak et al. [1] the automated surgical tool detection system based on RCNN and decision policies, which can provide real-time support and quality assessment during surgery. Tamer et al. [2] use of LSTM models to considertemporal information for tool recognition, improving skill assessment and surgeon training. Herag et al. [3] modelintroduces and evaluates attention modules (Squeeze and Excitation - SE, and Convolutional Block Attention Module - CBAM) to enhance the ResNet50 CNN. It aims to improve tool classification performance, achieving a significant 2% performance boost. The study conducts a comparative analysis of base ResNet50, ResNet50-SE, and ResNet50-CBAM models, providing valuable insights into attention module effectiveness. Additionally, it offers class-specific analysis, guiding improvements for specific surgical tools, contributing to medical image analysis. Daichi et al. [4] the instance segmentation model for precise tool recognition in surgical videos, facilitating a detailed understanding of surgical procedures and quality assessment. Guillermo et al. [5] introduction of the attention module P-CSEM to enhance feature refinement and classification performance, specifically for tool recognition in laparoscopic surgery. Satoshi Kondo [6] utilization of Transformer architecture to analyze inter-frame correlations in videos, demonstrating superior performance in terms of macro-F1 score. Herag et al. [7] it introduces the P-CSEM attention module to enhance surgical tool detection, with superior performance. Their study compares P-CSEM with other modules, yielding insights into their effectiveness. P- CSEM integration improves mAP by 1.76%. Their use of Grad-CAM enhances network interpretability. In summary, Herag et al.'s research aligns with common trends in medical image analysis and surgical tool recognition but offers unique value with P-CSEM, comparative analysis, and Grad-CAM for improved surgical tool recognition. Jaafar et al. [8] development of a deep learning-based system for tool classification and creatinga searchable database for MIS videos, facilitating training and assessment.

Each model brings a unique approach or innovation to the field of medical imaging and minimally invasive surgery, making them suitable for specific scenarios and applications within the healthcare industry.

4. IMPROVEMENT SUGGESTIONS, POTENTIAL SIGNIFICANCE & REAL-WORLD APPLICATION FOR EACH MODEL

1. **Improvement suggestions:**

a. Improvement Suggestions for Babak et al. [1]:

Expand Dataset Diversity to enhance the generalizability of the LapTool-Net system, consider training on a more diverse set of surgical videos. While the M2CAI16 and Cholec80 datasets were used, additional datasets from different surgical procedures and institutions could help the system perform better in a wider range of scenarios. Consider Real-world Challenges, Surgical

environments often present real-world challenges such as variations in lighting, camera angles, and patient anatomy. The system should be further tested and improved under such conditions to ensure its robustness in clinical settings. Ethical Considerations, Although Institutional Review Board (IRB) approval was not required for this study, it is important to continually assess the ethical implications of deploying automated surgical tools in real medical settings. Ethical discussions and compliance should be addressed. User Interface, consider developing a user- friendly interface for surgeons and medical professionals to interact with the system effectively during surgery. User feedback and usability studies can help refine the interface. Real-time Processing Optimization, explore ways to optimize the real-time processing capability of LapTool-Net. Reducing latency and improving the system's response time can be critical in surgical procedures.

ii. Improvement Suggestions for Tamer et al. [2]:

Further Model Refinement, continue to fine-tune the LSTM models and CNN architectures. Experiment with different hyperparameters and architectures to achieve even higher accuracy in tool detection. Real-time Application, Investigate the feasibility of making the system real-time or near-real-time. This would be highly beneficial for assisting surgeons during laparoscopic procedures. Incorporate Other Modalities, consider incorporating other modalities, such as audio or haptic feedback, to provide additional information to enhance tool detection and recognition. This multi-modal approach could improve system performance. Evaluate on a Broader Dataset, while the Cholec80 dataset is a valuable resource, testing the system on additional datasets that cover a wider range of surgical procedures and conditions can help assess its performance in diverse clinical scenarios. Interoperability, ensure that the system is compatible with existing surgical equipment and technologies. Consider how it can integrate seamlessly with other tools used in the operating room. Clinical Validation, collaborate with medical professionals and institutions to conduct clinical validation studies to assess the real-world applicability and impact of the system on surgical practice. These studies can provide valuable insights and feedback.

iii. Improvement Suggestions for Herag Et Al. [3]:

Diversity in Evaluation Datasets, while the Cholec80 dataset is a valuable resource, it would be beneficial for Herag et al. to evaluate their approach

on additional datasets or surgical scenarios to assess the generalizability of their attention modules. Diverse datasets can help confirm the robustness of the model. Comparative Analysis, Herag et al. could provide a more in-depth comparative analysis of their attention modules with other approaches used in similar studies, such as Babak et al. and Tamer et al. This would help in understanding the strengths and weaknesses of different techniques. Real-world Application, investigate the feasibility of implementing the attention module-enhanced model in real surgical settings. The transition from research to practical use in healthcare settings is crucial and may involve challenges not addressed in the research. Interpretability, in addition to Grad-CAM, consider other interpretability techniques to provide deeper insights into how the attention modules are influencing the model's predictions. Interpretability is important in medical applications for building trust with healthcare professionals. Human-in-the-Loop, consider incorporating human feedback or expert annotations to refine the attention modules. Human feedback can help identify challenging cases and improve the model's performance.

iv. Improvement Suggestions for Daichi et al. [4]:

Interoperability and Integration, consider exploring how the instance segmentation model can be integrated into existing surgical equipment or systems used in laparoscopic surgeries. A seamless integration could enhance its practical utility. Real-time Application, investigate the potential for making the model work in real-time during laparoscopic surgeries. The ability to provide instant feedback to surgeons would be highly valuable. Diverse Dataset, expanding the dataset to include a wider range of surgical procedures and conditions can help evaluate the model's performance under diverse clinical scenarios. This would enhance its generalizability. Ethical and Regulatory Compliance, address ethical considerations and regulatory compliance, especially if the model is intended for clinical use. Ensure that it adheres to

privacy and patient data protection standards. Usability testing, assess the usability of the model by collaborating with medical professionals to gather feedback. This feedback can help refine the model and its user interface to better suit the needs of surgeons and other medical practitioners. Performance Comparison, consider comparing the performance of the instance segmentation model to other state- of-the-art methods and models for surgical instrument recognition. This can provide a benchmark for its capabilities.

v. Improvement Suggestions for Guillermo et al. [5]:

Real-world Data, consider expanding the dataset to include real surgical videos or videos from clinical laparoscopic surgeries. While simulated scenarios are useful for initial testing, real-world data can better represent the challenges faced in actual clinical settings. Ethical Considerations, when working with medical data, even in a simulated context, it's important to address ethical considerations and privacy concerns. Compliance with ethical standards should be a priority. Real-time Processing, investigate the possibility of making the model work in real- time during actual laparoscopic surgeries. Real-time detection and localization of surgical gauze can have a significant impact on patient safety and the efficiency of surgeries. Comparison to Existing Methods, compare the performance of the proposed models to existing methods or models for gauze detection.

Benchmarking against established solutions can help assess the innovation's impact. Generalization, ensure that the models can generalize well to different laparoscopic scenarios and surgical environments. This is important for the practical applicability of the technology. Usability for Surgeons, collaborate with medical professionals to gather feedback on the usability and user interface of the system. The system should provide information that is easy for surgeons to interpret and act upon.

vi. Improvement Suggestions for Satoshi Kondo [6]:

Real-world Data, while the Cholec80 dataset is valuable for initial experiments, it's essential to expand the dataset to include real surgical videos from clinical laparoscopic surgeries. This would better represent the complexities and variations encountered in actual clinical settings. Ethical Considerations, when dealing with medical data, it's crucial to address ethical considerations and ensure compliance with privacy and patient data protection standards, even in a simulated context. Comparison to Existing Methods, to further validate the effectiveness of LapFormer, consider comparing its performance to existing state-of-the-art methods or models for surgical tool detection. This would help demonstrate its superiority and innovation. User Interface, develop a user-friendly interface that surgeons and medical professionals can easily interact with during surgery. Usability testing and feedback from end-users can help refine the user interface. Real-time Application, investigate the possibility of making the model real-time or near-real-time, allowing for the detection of surgical tools during live laparoscopic surgeries. Real- time support can have significant implications for patient safety and surgical efficiency. Generalization, ensure that the Transformer-based approach generalizes well to different laparoscopic scenarios and surgical environments, making it applicable in various clinical settings.

vii. Improvement Suggestions for Herag et al. [7]:

Enhanced Model Interpretability, while the study mentions the use of Grad-CAM for network explain ability analysis, it would be beneficial to provide more details on the insights gained through this analysis and how they can be used to improve the model or guide clinical decision-making. Diverse Datasets, while Cholec80 is a widely used dataset, consider expanding the evaluation to include other datasets representing different surgical scenarios, conditions, and camera variations. This would ensure the model's robustness across various clinical settings. Real-time Application, investigate the feasibility of making the model work in real-time during live laparoscopic surgeries. Real-time support is essential for assisting surgeons during the procedure. User Interface, develop a user-friendly interface for surgeons and medical professionals to interact with the model effectively during surgery. Usability studies can help refine the interface for practical use. Comparative Analysis, provide a comparative analysis of the P-CSEM attention module against other state-of-the-art attention

mechanisms and models for surgical tool recognition. This can highlight its innovation and superiority. Ethical Considerations, address ethical considerations related to the use of the model in a clinical setting, including privacy, consent, and patient data protection.

viii. Improvement Suggestions for Jaafar et al. [8]:

Ethical Considerations, given the use of surgical videos, it's essential to address ethical considerations, such as patient privacy and consent, and ensure compliance with ethical standards for data usage. Usability and Interface, consider developing a user-friendly interface for the searchable database that makes it intuitive for novice surgeons and postoperative controllers to navigate and access specific segments of MIS videos. Real- time Application, investigate the feasibility of making the system real-time for live assistance during MIS procedures, offering real-time support and assessment. Diverse Dataset, while the Cholec80 dataset is valuable, consider expanding the dataset to include data from various sources and clinical settings to ensure the model's robustness. Validation in Clinical Settings, consider conducting tests and validations in clinical settings to ensure the effectiveness of the proposed system in real-world scenarios.

II. Potential Significance in various Medical Applications:

Babak et al. [1]: The automated surgical tool detection system using RCNN is valuable for tool recognition and can work best for surgical quality assessment and real-time support. Tamer et al. [2]: The use of LSTM models to consider temporal information is a valuable approach for tool recognition. It can work best for improving skills assessment, surgeon training, and surgical quality assessment. Herag et al. [3]: Herag et al.'s model is applied in laparoscopic surgery and surgical tool recognition, aiming to enhance accuracy and efficiency. It supports surgical phase recognition, skills assessment, and protocol extraction. Utilizing the Cholec80 dataset and AP metric, it advances AI-assisted surgical applications. Grad-CAM aids postoperative analysis and quality assessment, improving surgical outcomes and patient care in laparoscopic surgery and surgical AI settings. Daichi et al. [4]: The instance segmentation model for surgical instruments is valuable for precise tool recognition, which can improve the overall understanding of surgical procedures. It is highly valuable for surgical quality assessment, patient safety, and innovation in surgical techniques. Guillermo et al.

[5]: The attention module P-CSEM enhances feature refinement and classification performance, which can benefit surgical tool recognition. It can work best for resource management in operating rooms and surgical quality assessment. Satoshi Kondo [6]: The utilization of a Transformer architecture is innovative and outperforms traditional methods in terms of macro-F1 score. It can work best for real-time analysis, telemedicine, and assisting surgeons during procedures. Herag et al. [7]: The P-CSEM attention module developed by Herag et al. can find potential medical applications in improving surgical tool recognition, enhancing the accuracy of laparoscopic surgeries, and assisting in various surgical applications, including surgical phase recognition and skills assessment. Additionally, the use of Grad-CAM for interpretability can aid in postoperative analysis and quality assessment, making it valuable for improving surgical outcomes and patient safety. Jaafar et al. [8]: The deep learning-based system for tool classification and creating a searchable database can be highly useful for surgical training, surgeon education, and postoperative surgical quality assessment. It can work best for training and education purposes.

III. Real-World Applications:

Deep learning models offer multifaceted advantages in medicine. They enhance surgical training by improving tool and phase recognition, aiding in workflow analysis. In computer-assisted surgery, they provide real-time support, enhancing precision. Quality assessment ensures procedure compliance. They optimize resource allocation, reducing costs in operating rooms. Searchable video databases streamline content management. They support continuous surgeon skill improvement, ensuring patient safety. In telemedicine, they enable real-time remote guidance. These models drive research and innovation. Some may transition into clinical applications, a typical outcome in the medical field. Real-World application suggestions for each the model is described below:

LapTool-Net [1] model can be used in actual operating rooms to provide real-time recognition of surgical tools during laparoscopic surgeries, assisting surgeons in ensuring the safety and efficiency of procedures. Tamer et al.'s model [2] is well-suited for the field of surgeon training and skills assessment. It can offer real- time feedback during laparoscopic procedures, enhancing the training process and helping surgeons improve their skills. Herag et al.'s model [3] is designed for application in real laparoscopic surgeries. It aids in the recognition of surgical tools, tracks surgical phases, and assesses surgical quality in real-time, ultimately enhancing patient safety and surgical outcomes. Daichi et al.'s instance segmentation model [4] is valuable for use in surgical equipment to improve the understanding of surgical procedures. It contributes to patient safety and drives innovation in surgical techniques by providing precise tool recognition. Guillermo et al.'s model [5] is suitable for deployment in operating rooms, where it can perform real-time detection of surgical gauze, ensuring patient safety and the overall efficiency of surgical procedures. Satoshi Kondo's model [6] can be applied in real-time during laparoscopic surgeries, offering surgical tool analysis, telemedicine support, and assisting surgeons by providing live insights and feedback. The P-CSEM attention module [7] developed by Herag et al. has potential medical applications, including improving surgical tool recognition, enhancing the accuracy of laparoscopic surgeries, assisting in surgical phase recognition, and skills assessment. Jaafar et al.'s deep learning-based system [8] for tool classification and searchable database creation can find applications in surgical training, surgeon education, and postoperative surgical quality assessment in real-world clinical settings.

5. FUTURE DIRECTIONS

Future directions for LapToolNet [1] include improving accuracy, real-time performance, adaptability to diverse procedures, interpretability, robustness, scaling to larger datasets, and integration into clinical practice [1].The deep learning framework [2] for tool detection in laparoscopic videos by Tamer et al. can be enhanced through advanced neural network architectures, data augmentation, and integrated end-to-end frameworks for better spatiotemporal features[2]. Herag et al.'s study on attention modules [3] should focus on optimizing attention mechanisms, ensemble methods, transfer learning, real-time applications, imbalanced data handling, tool localization, and integration into clinical settings[3]. Daichi et al.'s instance segmentation model [4] should undergo external data testing, evaluation with larger video sets, real-time implementation, and collaboration with surgeons for improved accuracy[4]. Guillermo et al.'s gauze detection model [5] should prioritize creating a comprehensive dataset, addressing diverse gauze shapes, exploring different model architectures, and integrating it into broader laparoscopic surgery segmentation frameworks [5]. Satoshi Kondo's surgical tool detection model [6] should explore larger query sizes, adaptability to various scenarios, and inclusion of other surgical workflow aspects like action recognition [6]. Nour et al.'s P-CSEM attention module model [7] needs enhancements in dataset size, class imbalance, model explain-ability, inclusion of temporal information, real-time capabilities, and optimization of model architecture [7]. Jaafar et al.'s model for unbalanced data in laparoscopy video classification [8] should investigate alternative neural network architectures, test on different surgical datasets, and optimize for computational efficiency while maintaining accuracy [8].

6. CONCLUSION

In conclusion, the studies conducted by Babak et al. [1], Tamer et al. [2], Herag et al. [3], Daichi et al. [4], Guillermo et al. [5], Satoshi Kondo [6], Nour et al. [7], and Jaafar et al. [8] have collectively advanced the field of surgical instrument detection and recognition in laparoscopic videos. These studies have introduced innovative models and methodologies that offer promising solutions to various challenges in the domain of computer-assisted surgery.

LapToolNet, presented by Babak et al., demonstrated remarkable contextual understanding and outperformed previous approaches on the M2CAI dataset, offering a valuable tool for tool detection without requiring specialized expertise. Tamer et al.'s CNN model with temporal information showcased superior tool recognition capabilities and opens the door for further research into tool localization. Herag et al.'s investigation

into attention modules highlighted their potential to enhance tool classification, particularly the promising performance of the CBAM module. Daichi et al.'s instance segmentation model for surgical instruments provides a foundation for automated surgical progress monitoring and computer-assisted surgery.

Guillermo et al.'s work on gauze detection introduced a new dataset and efficient CNN models suitable for real- time applications, with potential for broader scene segmentation frameworks. Satoshi Kondo's LapFormer, based on Transformer architecture, significantly improved tool detection accuracy and presents opportunities for further exploration in surgical workflow analysis. Nour et al.'s P-CSEM attention module proved effective in enhancing spatial and channel characteristics in surgical tool recognition. Jaafar et al.'s model effectively addressed imbalanced data and offers promise for computer-assisted instruction and surgical video indexing.

These diverse studies collectively contribute to the advancement of technology in laparoscopic surgery, holding potential for improving patient outcomes, enhancing surgeon performance, and ultimately benefiting the field of healthcare. Future research should continue to build upon these foundations, exploring novel approaches, expanding dataset diversity, and further integrating these models into clinical practice.

The studies by Babak et al. [1], Tamer et al. [2], Daichi et al. [3], Guillermo et al. [4], Satoshi Kondo [6], and Herag et al. collectively illustrate the remarkable potential of AI and deep learning techniques in the field of surgical tool recognition and medical applications. These models contribute to enhancing tool recognition, skills assessment, surgical quality assessment, real-time support, resource management, and surgical training. Moreover, Herag et al.'s [7] innovative P-CSEM attention module and the application of Grad-CAM for interpretability extend the impact of these models to laparoscopic surgery, surgical phase recognition, and patient safety. These advancements pave the way for the continued development and innovation of AI-driven solutions in surgical settings, ultimately improving surgical outcomes and patient care. Jaafar et al.'s [8] system, on the other hand, primarily focuses on surgical training and education, underlining the versatile utility of these models across various aspects of the medical field.

The studies on surgical tool recognition and AI applications in the field of laparoscopic surgery offer valuable insights and potential for enhancement. These studies collectively highlight the importance of expanding dataset diversity, addressing real-world challenges, and considering ethical implications. User- friendly interfaces and real-time processing optimization are essential for practical use. Further model refinement, incorporation of other modalities, and interoperability with existing surgical equipment can significantly improve tool detection systems. Clinical validation and comparative analyses with state-of-the- art methods are crucial for assessing real-world applicability and innovation. Model interpretability, human- in-the-loop feedback, usability for surgeons, and generalization across various clinical settings are key factors for advancing these AI applications. By addressing these aspects, these studies can contribute to the evolution of AI-assisted surgical tools, ultimately enhancing surgical outcomes, patient safety, and the quality of healthcare in the field of laparoscopic surgery.

REFERENCES

B. Namazi, G. Sankaranarayanan, and V. Devarajan, “A contextual detector of surgical tools in laparoscopic videos using deep learning,” Surg. Endosc., vol. 36, no. 1, pp. 679–688, 2022, doi: 10.1007/s00464-021-08336-x.

T. Abdulbaki Alshirbaji, N. A. Jalal, P. D. Docherty, T. Neumuth, and K. Möller, “A deep learning spatial-temporal framework for detecting surgical tools in laparoscopic videos,” Biomed. Signal Process. Control, vol. 68, no. January, 2021, doi: 10.1016/j.bspc.2021.102801.

H. Arabian, F. A. Dalla, N. A. Jalal, T. A. Alshirbaji, and K. Moeller, “Attention Networks for Improving Surgical Tool Classification in Laparoscopic Videos,” Curr. Dir. Biomed. Eng., vol. 8, no. 2,pp. 676–679, 2022, doi: 10.1515/cdbme-2022-1172.

D. Kitaguchi et al., "Development and Validation of a Model for Laparoscopic Colorectal Surgical Instrument Recognition Using Convolutional Neural Network-Based Instance Segmentation and Videos of Laparoscopic Procedures," JAMA Network Open, vol. 5, no. 8. p. E2226265, 2022. doi: 10.1001/jamanetworkopen.2022.26265.

G. Sánchez-Brizuela et al., "Gauze Detection and Segmentation in Minimally Invasive Surgery Video Using Convolutional Neural Networks," Sensors, vol. 22, no. 14, pp. 1–16, 2022, doi: 10.3390/s22145180.

S. Kondo, "LapFormer: surgical tool detection in laparoscopic surgical video using transformer architecture," Comput. Methods Biomech. Biomed. Eng. Imaging Vis., vol. 9, no. 3, pp. 302–307, 2021, doi: 10.1080/21681163.2020.1835550.

H. Arabian, T. Abdulbaki Alshirbaji, N. A. Jalal, S. Krueger-Ziolek, and K. Moeller, "P-CSEM: An Attention Module for Improved Laparoscopic Surgical Tool Detection," Sensors (Basel)., vol. 23, no. 16, pp. 1–13, 2023, doi: 10.3390/s23167257.

J. Jaafari, S. Douzi, K. Douzi, and B. Hssina, "Towards more efficient CNN-based surgical tools classification using transfer learning," J. Big Data, vol. 8, no. 1, 2021, doi: 10.1186/s40537-021-00509- 8.

A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. De Mathelin, and N. Padoy, "EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos," IEEE Trans. Med. Imaging, vol. 36, no. 1, pp. 86–97, 2017, doi: 10.1109/TMI.2016.2593957.

N. A. Jalal, H. Arabian, T. A. Alshirbaji, P. D. Docherty, T. Neumuth, and K. Moeller, "Analysing attention convolutional neural network for surgical tool localisation: a feasibility study," Curr. Dir. Biomed. Eng., vol. 8, no. 2, pp. 548–551, 2022, doi: 10.1515/cdbme-2022-1140.

M. Sahu, A. Mukhopadhyay, A. Szengel, and S. Zachow, "Tool and Phase recognition using contextual CNN features," arvix, p. 4, 2016, [Online]. Available: http://arxiv.org/abs/1610.08854

A. Jin et al., "Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks," Proc. - 2018 IEEE Winter Conf. Appl. Comput. Vision, WACV 2018, vol. 2018-Janua, pp. 691–699, 2018, doi: 10.1109/WACV.2018.00081.

K. Jo, Y. Choi, J. Choi, and J. W. Chung, "Robust real-time detection of laparoscopic instruments in robot surgery using convolutional neural networks with motion vector prediction," Appl. Sci., vol. 9, no. 14, 2019, doi: 10.3390/app9142865.

A. Kanakatte, A. Ramaswamy, J. Gubbi, A. Ghose, and B. Purushothaman, "Surgical tool segmentation and localization using spatio-temporal deep network," Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS, vol. 2020-July, pp. 1658–1661, 2020, doi: 10.1109/EMBC44109.2020.9176676.

P. Zaphiris and C. S. Ang, Human Computer Interaction : Concepts , Methodologies , Tools , and Applications Human Computer Interaction : Concepts , Methodologies , Tools , and Applications (4 Volumes), 2015th ed., no. January 2008. Hershye PA,USA: IGI Global, 2015, 2014. doi: 10.4018/978.