International Journal on Advanced Science Engineering Information Technology

Vol.15 (2025) No. 2
ISSN: 2088-5334

Comparative Analysis and Exploration of 3DResNet-18 and InceptionV3-GRU with Temporal Segment Network (TSN) Framework for Laparoscopic Surgical Video Expertise Classification

Liem Roy Marcelino [a], Samuel Batara Kelengate Munthe [a], Ronny Dominikus Munthe [a], Eko Adi Sarwoko [a], Aris Sugiharto [a], Helmie Arif Wibawa [a], Aris Puji Widodo [a], Fajar Agung Nugroho [a], Anis Farihan Binti Mat Raffei [b], Adi Wibowo [a,*]

[a] *Department of Informatics, Universitas Diponegoro, Semarang, Indonesia*
[b] *Faculty of Computing, Universiti Malaysia Pahang Al-Sultan Abdullah, Pekan, Pahang, Malaysia*
[*] *Corresponding author: bowo.adi@live.undip.ac.id*

Abstract*—Laparoscopic surgery is a widely adopted minimally invasive procedure that requires surgeons to master complex skills such as suturing, knot-tying, and needle-passing. Traditional assessment of these skills is often subjective and prone to bias, relying heavily on manual evaluation by expert surgeons, which can vary between evaluators. We applied deep learning models to automate surgical skill evaluation to address this issue and move towards a more objective and standardized assessment method. In this study, we utilized two advanced architectures—3D ResNet-18 and InceptionV3-GRU—within a Temporal Segment Network (TSN) framework to classify skill levels using the publicly available JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS) dataset. We focused on optimizing temporal sampling by adjusting the number of frames and frame intervals in the video data. Our findings show that capturing longer sequences of actions improved accuracy for suturing and needle-passing tasks while capturing more detailed motions enhanced performance for knot-tying. Our findings suggest that capturing longer sequences of actions improved accuracy for suturing and needle-passing tasks while capturing more detailed motions enhanced performance for knot-tying. The 3D ResNet-18 model achieved 100% accuracy across all tasks, significantly outperforming the InceptionV3-GRU model, which achieved 85.71% for suturing, 77.42% for knot-tying, and 100% for needle-passing. These results demonstrate the superior capability of the 3D ResNet-18 model in surgical skill classification and highlight the critical role of temporal optimization in improving performance across different surgical tasks.

***Keywords*—Laparoscopic surgery; deep learning; 3D CNN; ResNet-18; InceptionV3; GRU; TSN.**

Manuscript received 8 Jun. 2024; revised 27 Sep. 2024; accepted 24 Nov. 2024. Date of publication 30 Apr. 2025.

I. INTRODUCTION

Laparoscopic surgery, a minimally invasive technique widely used in clinical practice, allows surgeons to access the abdomen and pelvis through small incisions, known as keyhole surgery [1], [2], [3]. This approach offers advantages over open surgery, such as faster recovery times, less pain, and lower complication risks [2]. Successful laparoscopic procedures heavily rely on skills like suturing, knot-tying, and needle passing, necessitating structured training for surgeons to enhance these essential skills [4]. Currently, the assessment of laparoscopic skills tends to be subjective, usually conducted by experienced surgeons through direct observation, rating skills across novice (N), intermediate (I), and expert (E) levels [5].

The advances in deep learning have had a transformative impact across various sectors, especially in medicine [6], [7]. Deep learning's integration into laparoscopic surgery has revolutionized the way surgical skills are evaluated and enhanced. The application of computer vision, a branch of deep learning, allows for the sophisticated analysis of visual data from surgical procedures. This technology excels in interpreting complex images and videos, transforming them into actionable insights [3], [7]. In the context of laparoscopy, this means leveraging the vast amount of video data generated during surgeries. The extensive video footage of laparoscopic procedures offers a unique opportunity for deep learning

algorithms to analyze and learn from these intricate surgical techniques [8]. By processing this data, deep learning systems can identify key patterns, movements, and techniques that characterize expert surgical practice.

Convolutional Neural Networks (CNNs) is a deep learning model that excel in visual data processing, especially for 2D images, by leveraging spatial pixel correlations [9], [10]. However, they are less effective for 3D motion data like videos. The development of 3D CNNs addresses this by combining both 2D CNN to extract features from spatial dimensions and 1D CNN for temporal dimensions, making them suitable for recognizing actions or movements in video data [11], [12]. In laparoscopic surgery, this capability is particularly useful. For instance, Funke et al. [13] applied a 3D CNN with Temporal Segment Network (TSN) to the JIGSAWS dataset, achieving an impressive accuracy of 100% on suturing, 95.8% on knot-tying, and 100% on needle-passing. Similarly, Jian et al. [14] applied an attention-enhanced I3D CNN with multitask-TSN to the same dataset, achieving an impressive 100% accuracy on suturing, 97.2% on knot-tying, and 100% on needle-passing

The Temporal Segment Network (TSN) framework as demonstrated in Funke et al. [13] and Jian et al. [14] were introduced by Wang et al. which further refines the capabilities of CNNs in video classification tasks, particularly in the domain of human action recognition. TSN operates by segmenting videos into short snippets and utilizing a ConvNet to classify these snippets individually. Subsequently, predictions from these snippet-level classifiers are aggregated using a consensus function to yield the overall video classification result. Among the aggregation functions, average pooling emerges as a widely employed and effective choice.

During the training phase of TSN, snippets are sampled based on a segment-based strategy, where the video is partitioned into non-overlapping segments, and one or more consecutive frames are randomly sampled from each segment to form snippets. This process generates K snippets, typically ranging from 3 to 9 for videos of standard durations. Conversely, during testing, snippets are sampled equidistantly from the test video, often exceeding the number of snippets sampled during training. This adaptive sampling strategy ensures the robustness and generalizability of the TSN model across diverse video inputs, making it a potent tool for video classification tasks, including intricate scenarios like action recognition in laparoscopic surgery.

One main advantage of using CNN is the ability to use pre-trained model such as ResNet and Inception trained with huge image dataset like ImageNet [15]. Tran et al., [16] leveraged this with their 3D Resnet-18 model, combining pre-trained 2D spatial and 1D temporal CNNs with residual blocks. This approach reduces the number of parameters, simplifying the training process, and enabling the model to perform on par or better than existing state-of-the-art models in action recognition datasets like Sports-1M, Kinetics, UCF101, and HMDB51.

Recurrent Neural Networks (RNNs), particularly exemplified by architectures like Gated Recurrent Unit (GRU), offer a powerful approach for processing sequential data due to their ability to retain and forget information over time [17]. By integrating CNN's feature extraction capabilities with GRU's sequential data processing, researchers have addressed various challenges encountered by both networks [18]. For instance, Lu et al. [19] showcased the effectiveness of CNN-GRU models in human activity recognition on the UCI-HAR dataset, achieving a notable accuracy of 96.39%. Similarly, Ullah & Munir [20] utilized CNNs for video data feature extraction and Bi-GRUs for temporal processing, resulting in significantly improved execution speed, up to 167 times faster in frames per second. These studies underscore the adaptability and efficacy of CNN-GRU models in diverse applications involving spatiotemporal data.

In our study, we aimed to address these gaps by leveraging advanced deep learning models—3D ResNet-18 and InceptionV3-GRU—to improve laparoscopic video classification. The 3D ResNet-18 architecture combines 2D CNNs for spatial feature extraction with 1D CNNs for temporal analysis, using residual blocks to enhance training efficiency and model accuracy. In contrast, the InceptionV3-GRU model integrates a pre-trained InceptionV3 network with GRU, enabling it to process sequential data and capture long-term temporal dependencies in laparoscopic videos. Both models were evaluated using the Temporal Segment Network (TSN) framework to improve generalizability. By comparing these models, we aimed to identify the most effective approach for automated laparoscopic skill classification. Furthermore, we examined the impact of varying frame steps and frame counts within the TSN framework to optimize its performance across different video characteristics, thereby contributing to more reliable and scalable surgical skill assessments.

II. Materials and Method

This section outlines the methodology employed in the study, starting with a description of the JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS) dataset, which includes videos of laparoscopic skills performed by surgeons of varying proficiency levels. We detail the data preprocessing steps, including video extraction and image processing, followed by the data splitting strategy, ensuring balanced distributions across skill levels for effective model evaluation. The study proposes and compares two distinct models: the first model, based on 3D ResNet-18, combines a 3D Convolutional Neural Network (CNN) architecture with residual blocks, while the second model, named InceptionV3-GRU, integrates the InceptionV3 CNN pre-trained model with a GRU. Both models utilize input snippets containing 32 frames, each with a resolution of 224x224, resulting in a final input shape of 32x224x224. The output consists of three probability values corresponding to expertise levels (novice, intermediate, expert), with probabilities summing to 1 using a SoftMax classifier.

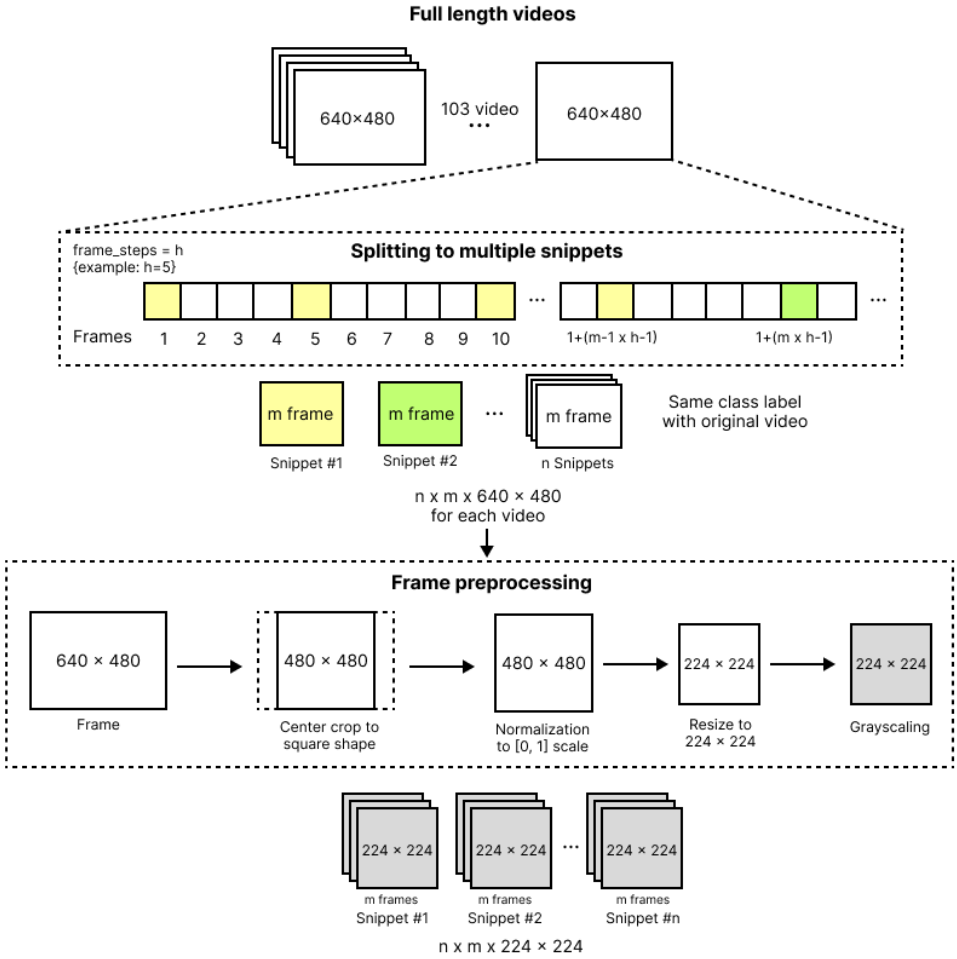

Fig. 1 Data Preprocessing Workflow

A. Data Description

The dataset used in the study originates from the JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS), which acts as a repository for surgical activity data geared towards modeling human motion. This dataset was amassed through a collaborative effort between The Johns Hopkins University (JHU) and Intuitive Surgical, Inc. (ISI), based in Sunnyvale, CA. The study made use of the Da Vinci Surgical System and involved eight surgeons whose surgical proficiency spanned from novice to expert levels. The JIGSAWS dataset comprises video recordings showcasing three distinct laparoscopic skills: suturing, knot-tying, and needle-passing. These recordings are categorized into three expertise levels—novice, intermediate, and expert—each annotated with specific skill or activity details. The videos are maintained at a resolution of 640x480, with durations varying between 26 seconds to 2 minutes and 32 seconds with 30 frames per second [1].

Table 1 describes the distribution of video counts for each laparoscopic skill across different expertise levels. There are a total of 103 videos, evenly distributed among the skills and expertise levels. Each video was recorded from two different angles: from the left side (capture1) and from the right side (capture2), resulting in a total of 206 videos in the dataset. In this study, only videos captured from the left side (capture1) will be utilized.

TABLE I
VIDEO COUNT FOR EACH SKILL AND EXPERTISE

Laparoscopic Skill	Video Count for Each Level			Total
	Novice	Intermediate	Expert	
Suturing	19	10	10	39
Knot-Tying	16	10	10	36
Needle-Passing	11	8	9	28

B. Data Preprocessing

Error! Reference source not found. illustrates the main data preprocessing workflow, which encompasses both video extraction and image processing stages. In the video preprocessing phase, frames are extracted from each video, resulting in the extraction of n snippets, each containing m frames. The parameter frame steps dictate the skipping interval for frame selection. This skipping interval determines how often frames are sampled from the original video sequence. Formula (1) defines the set of indices used with m as its frame count and h as its frame steps in the preprocessing workflow:

$$Indices = \{1, 1 + (1 \times (h-1)), 1 + (2 \times (h-1)), \ \dots, 1 + ((m-1) \times (h-1))\} \quad (1)$$

C. Data Splitting

The data will be divided into two sections: training data and testing data. Each section comprises snippets with 32 frames obtained from the video processing stage. Some videos will be manually selected for use as testing data, while the rest will be designated for training data. Consequently, frames derived from a particular video will not be utilized in multiple data types. The number snippets obtained for suturing, knot-tying, and needle-passing skills are 174, 272, and 167, respectively. The distribution of data is presented in Table 2.

TABLE II
SNIPPET COUNT IN ALL DATASETS FOR EACH SKILL AND EXPERTISE

Laparoscopic Skill	Expertise Level	Training Snippet Count	Testing Snippet Count	Total Snippet Count
Suturing	Novice	44	26	70
	Intermediate	34	19	53
	Expert	37	14	51
Knot-Tying	Novice	81	29	110
	Intermediate	68	13	81
	Expert	63	18	81
Needle-Passing	Novice	51	13	64
	Intermediate	28	16	44
	Expert	43	16	59

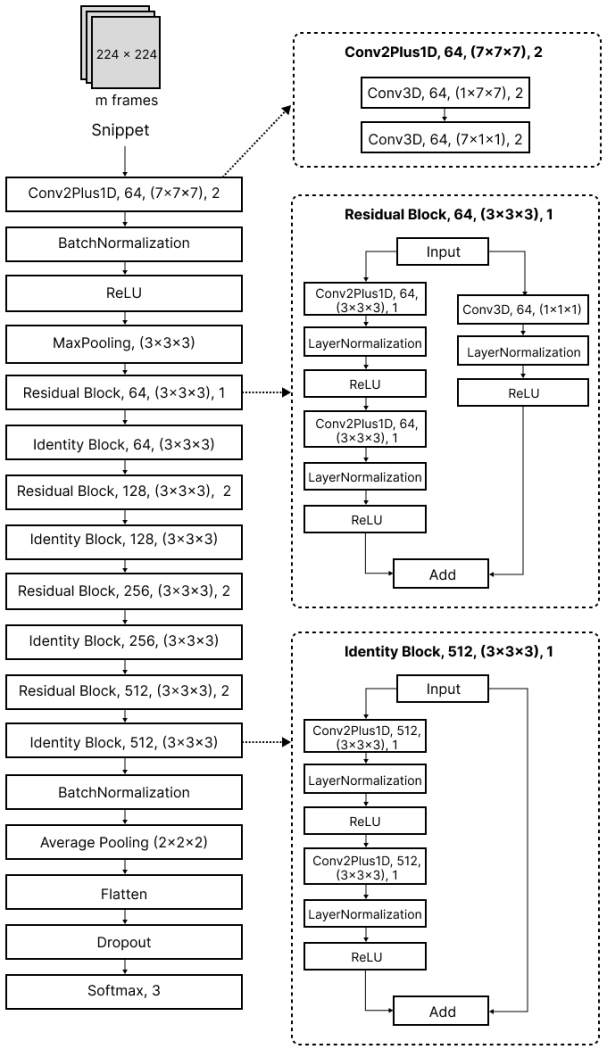

Fig. 2 3D ResNet-18 for laparoscopy surgical expertise classification

D. 3D ResNet-18

The 3D ResNet-18 architecture integrates a 3D CNN framework, merging 2D and 1D CNN. Additionally, it includes residual blocks, renowned for their advantages such as enhanced gradient flow during training and the facilitation of deeper network architectures. This architecture comprises three pivotal components: Conv2Plus1D, Residual Block, and Identity Block, along with other components such as batch normalization, ReLU activation function, pooling layers, dropout, and SoftMax classifier. The overview of the architecture is shown in **Error! Reference source not found.**. Further details regarding the key components of this architecture will be explained in the subsequent subsection.

1) Conv2DPlus1D: The Conv2DPlus1D layer, inspired by Tran et al. [16] 3D convolution, harnesses the feature extraction capabilities of CNNs. This layer integrates 2D CNN for spatial dimension processing and 1D CNN for temporal dimension processing. Utilizing an n x n x n filter, the input undergoes spatial 3D convolution (1 x n x n) to

mimic 2D CNN behavior, followed by temporal 3D convolution (n x 1 x 1) to simulate 1D CNN behavior.

2) Residual Block: Residual networks, developed by Microsoft for image recognition, are a type of CNN model specifically designed to address the issue of vanishing gradients during training. They utilize residual functions to mitigate vanishing gradient, a problem commonly encountered in deep neural networks. By implementing residual functions, the network can combine the output of a layer with the output of the preceding layer before passing through the activation function. This approach allows gradients to directly flow back to previous layers in the network, enabling better data representation learning and overcoming gradient disappearance issues. The Residual Block represents one of the implementations of the residual network, characterized by multiple Conv2DPlus1D layers followed by normalization layers and ReLU activation functions. This layer employs 2 layers of n x n x n convolution layer to produce the output, along with a 1 x 1 x 1 convolution applied to adjust the shape of the residual. Finally, the output and the residual are combined using addition to address the vanishing gradient problem [21], [22].

3) Identity Block: The Identity Block also functions as a residual layer, comprising multiple convolutional layers, normalization layers, and ReLU activations. Unlike the residual block layer, this layer does not entail conventional 3D convolution. Instead, this layer utilizes a residual with the same size as the input, removing the necessity for conventional 3D convolution to adjust the size between input and output. Hence, it is dubbed the "identity block."

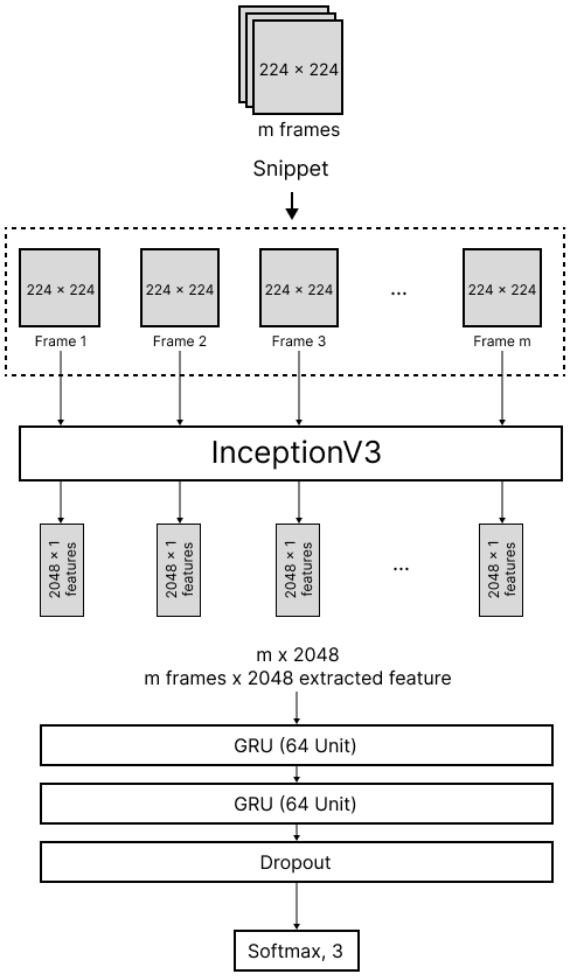

Fig. 3 GRU model architecture for laparoscopy surgical expertise classification

E. InceptionV3-GRU

The InceptionV3-GRU architecture integrates feature extraction via InceptionV3, leveraging a pre-trained CNN model to extract features from each frame in the snippet input. These extracted features are then processed using a Gated Recurrent Unit (GRU) to handle the sequential frames of the snippets. Finally, a softmax classifier is employed to determine the expertise level. Further elaboration on the key components of this architecture will be provided in the subsequent subsections.

1) InceptionV3: InceptionV3 is a deep learning model developed by Google for image recognition, representing an evolution from its predecessors, InceptionV1 and V2. It utilizes an architecture known as the "inception module", enabling the model to extract features at various spatial scales in parallel. Moreover, InceptionV3 employs a regularization technique called "batch normalization" to enhance model

accuracy and reduce training time [23]. In this architecture, the InceptionV3 pre-trained model is utilized to extract features from each frame in the snippet input, which are then processed by a GRU layer. An overview of the architecture is depicted in **Error! Reference source not found.**.

2) Gated Recurrent Unit (GRU): GRU, a variation of Recurrent Neural Network (RNN) with gated units, introduces a hidden state to capture and retain temporal dependencies across sequential data. The hidden state is updated by two gates: the update gate and the reset gate. The update gate determines the amount of information from the previous time step to retain for the current time step, while the reset gate allows GRU to decide whether to ignore or consider information from the previous time step. The hidden state stores crucial information from the previous time step deemed essential for the current time step, making GRU an effective neural network architecture for processing sequential or time-series data by retaining significant information from the previous time step and deciding whether to retain or discard it. In the proposed model, features extracted by the InceptionV3 model from the sequential frames of the snippets are processed using GRU, leveraging the advantages of GRU in capturing and utilizing temporal dependencies of each frame effectively.

F. SoftMax Classifier

The SoftMax classifier commences with a dropout layer, which serves as a regularization technique to mitigate overfitting by randomly deactivating a proportion of input units during training. This dropout mechanism introduces robustness to the model by preventing reliance on specific features and encouraging the learning of more generalizable representations. Following the dropout layer, the final output layer for both proposed architectures consist of a fully connected layer with three neurons, aligning with the number of expertise levels (novice, intermediate, expert) in the classification task. Subsequently, the SoftMax activation function is applied after the fully connected layer. The SoftMax function computes the probability distribution across all classes, assigning a probability value to each class between 0 and 1. These probabilities signify the model's confidence in each class prediction, collectively summing to 1 across all classes. Ultimately, the class with the highest probability is identified as the output of the classifier, representing the predicted expertise level for the given input data.

III. Results and Discussion

In this section, we present the results of our experiments conducted using Python 3 and TensorFlow. We trained the model using the Adam optimizer for 50 epochs, with various hyper-parameters tuned for optimal performance. Experimental results showed that performance plateaued at 50 epochs, with additional training increasing the risk of overfitting. This approach aligns with findings in video classification research [24], [25]. Early stopping was used to preserve model generalization by capturing the best-performing weights. Various hyper-parameters were tuned to ensure optimal performance of each model. All computations were performed on a P100 GPU, facilitating efficient processing and experimentation.

A. Performance Metrics

In laparoscopic surgical skill expertise classification, accuracy is crucial for assessing the model's performance. It measures the ratio of correctly classified instances (TP and TN) to the total evaluated instances, providing a comprehensive evaluation of the model's ability to correctly identify true positives (TP) and true negatives (TN) while distinguishing false positives (FP) and false negatives (FN). Accuracy serves as a valuable metric for gauging the model's effectiveness in skill classification tasks, reflecting its proficiency in accurately distinguishing between different skill levels. Accuracy is denoted by formula (2), which calculates the ratio of correctly classified instances (TP and TN) to the total evaluated instances, thus providing a comprehensive evaluation of the model's performance in laparoscopic surgical skill expertise classification.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

B. Model Improvement

The model improvement process primarily involved hyperparameter tuning, including adjustments to batch size, learning rate, and dropout rate. Specifically, we experimented with batch sizes of 4 and 8, learning rates of 0.001, 0.0001, and 0.00001, and dropout rates of 0, 0.25, and 0.5. These adjustments aimed to optimize the model's performance by finding the most suitable combination of hyperparameters compares the impact of hyperparameter configurations on the performance of 3D ResNet-18 and InceptionV3-GRU models trained for suturing, knot-tying, and needle-passing task.

Table 3 provides a detailed comparison of hyperparameter configurations and performance metrics for 3D ResNet-18 and InceptionV3-GRU models across the suturing, knot-tying, and needle-passing tasks, with the best performances highlighted in bold for each task. Notably, the 3D ResNet-18 model achieves optimal performance with a batch size of 4, a learning rate of 0.0001, and a dropout rate of 0.25, yielding a test accuracy of 100% on suturing task. In contrast, the InceptionV3-GRU model reaches its peak accuracy of 85.71% with a similar configuration but without dropout, albeit with a higher test loss. For the knot-tying task, the 3D ResNet-18 model again outperforms the InceptionV3-GRU model, achieving 100% accuracy with a batch size of 4, a learning rate of 0.001, and no dropout. The InceptionV3-GRU model achieves its highest accuracy of 77.42% with batch size of 8, learning rate of 0.001, and dropout rate of 0.25. In the needle-passing task, the 3D ResNet-18 model achieves 100% accuracy and 0.007 test loss with a batch size of 4, a learning rate of 0.0001, and a dropout rate of 0.25, while the InceptionV3-GRU model also reaches 100% accuracy but with a slightly higher test loss of 0.0547 on batch size of 8, learning rate of 0.0001, and no dropout. Across all tasks, the 3D ResNet-18 model consistently outperforms the InceptionV3-GRU model, demonstrating superior generalization, accuracy, and lower test loss across various hyperparameter configurations.

TABLE III
HYPERPARAMETER TUNING RESULTS AND BEST PERFORMING MODELS FOR 3D RESNET-18 AND INCEPTIONV3-GRU ACROSS SUTURING, KNOT-TYING, AND NEEDLE-PASSING TASKS

No	Batch Size	Learning Rate	Drop out	Suturing				Knot-Tying				Needle-Passing			
				Test Accuracy		Test Loss		Test Accuracy		Test Loss		Test Accuracy		Test Loss	
				3D ResNet-18	InceptionV3-GRU	3D ResNet-18	InceptionV3-GRU	3D ResNet-18	InceptionV3-GRU	3D ResNet-18	InceptionV3-GRU	3D ResNet-18	InceptionV3-GRU	3D ResNet-18	InceptionV3-GRU
1	4	0.001	0	100.00	83.33	0.0706	0.3843	**100.00**	70.97	**0.0008**	0.9279	96.00	92.00	0.0706	0.1573
2	4	0.001	0.25	90.48	76.19	0.1638	0.4779	96.80	70.97	0.0352	0.8821	96.00	84.00	0.1638	0.3298
3	4	0.001	0.5	97.62	71.43	1.1195	0.6832	100.00	74.19	0.0009	0.9531	60.00	92.00	1.1195	0.3316
4	4	0.0001	0	100.00	**85.71**	0.0264	**0.3633**	100.00	70.97	0.0026	1.0659	100.00	88.00	0.0264	0.3609
5	4	0.0001	0.25	**100.00**	83.33	**0.007**	0.5169	100.00	58.06	0.0065	1.0132	**100.00**	84.00	**0.007**	0.4753
6	4	0.0001	0.5	100.00	66.67	0.0052	0.8452	100.00	67.74	0.0076	1.2087	100.00	88.00	0.0052	0.2899
7	4	0.00001	0	92.86	59.52	0.0265	0.8839	96.80	45.16	0.0605	1.0727	100.00	76.00	0.0265	0.6242
8	4	0.00001	0.25	100.00	59.52	0.0342	0.9299	96.80	58.06	0.0695	0.9908	100.00	68.00	0.0342	0.9189
9	4	0.00001	0.5	95.24	64.29	0.022	0.7707	96.80	38.71	0.1139	1.1322	100.00	76.00	0.022	0.6354
10	8	0.001	0	47.62	71.43	0.1243	0.6907	100.00	70.97	0.0026	0.8944	96.00	88.00	0.1243	0.3100
11	8	0.001	0.25	80.95	71.43	0.0868	0.6851	100.00	**77.42**	0.0026	**0.8653**	96.00	88.00	0.0868	0.2615
12	8	0.001	0.5	90.48	69.05	0.153	0.7815	93.60	74.19	0.1605	0.9036	92.00	92.00	0.153	0.1307
13	8	0.0001	0	100.00	73.81	0.0357	0.8504	100.00	77.42	0.0183	1.2238	100.00	**100.00**	0.0357	**0.0547**
14	8	0.0001	0.25	100.00	69.05	0.0762	0.8511	100.00	70.97	0.0141	1.0320	96.00	96.00	0.0762	0.0769
15	8	0.0001	0.5	100.00	80.95	0.0345	0.4757	100.00	67.74	0.0106	1.1135	100.00	92.00	0.0345	0.2271
16	8	0.00001	0	100.00	57.14	0.0354	0.9225	100.00	35.48	0.0716	1.1086	100.00	80.00	0.0354	0.6822
17	8	0.00001	0.25	100.00	50.00	0.0531	1.0028	96.80	54.84	0.1017	1.0315	100.00	72.00	0.0531	0.8286
18	8	0.00001	0.5	97.62	71.43	0.0345	0.8417	96.80	45.16	0.0826	1.1389	100.00	68.00	0.0345	0.7235

C. Performance Analysis

The bar graph in **Error! Reference source not found.** presents a comparative analysis of test accuracy between two machine learning models, namely 3D ResNet-18 and InceptionV3-GRU, across three distinct surgical skills: suturing, knot-tying, and needle-passing. Notably, both models achieved perfect test accuracy of 100% for needle-passing tasks. However, for suturing and knot-tying, while 3D ResNet-18 maintained a flawless score, InceptionV3-GRU exhibited a decrease to 86% for suturing and 77% accuracy for knot-tying. Overall, the graph highlights the strong performance of both models on the test set, with 3D ResNet-18 showcasing consistent perfect accuracy across all skills and InceptionV3-GRU demonstrating decreases in suturing and knot-tying accuracy.

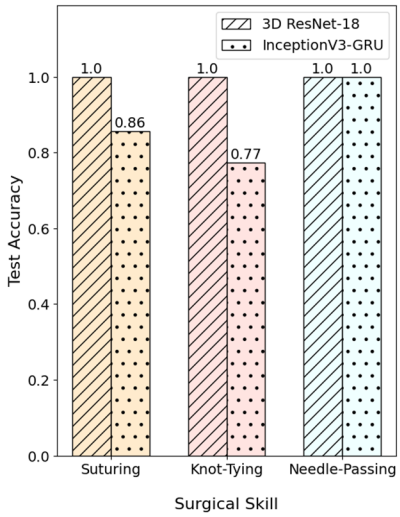

Fig. 4 Test accuracy comparison between 3D ResNet-18 and InceptionV3-GRU

The superior performance of 3DResNet-18 compared to InceptionV3-GRU in suturing and knot-tying tasks can be attributed to its ability to seamlessly integrate spatial and temporal information through 3D convolutions, enabling the model to capture the fine-grained spatiotemporal dynamics inherent in these complex surgical gestures. Unlike InceptionV3-GRU, which processes spatial and temporal features separately—potentially leading to temporal discontinuities—3DResNet-18 simultaneously analyzes multiple frames, preserving temporal continuity and contextual information. Additionally, the residual connections in 3DResNet-18 enhance gradient flow, reducing overfitting and improving generalization, particularly on smaller datasets like JIGSAWS [13], [21], [26].

TABLE IV
COMPARISON OF COMPUTATIONAL AND MEMORY COSTS BETWEEN 3D RESNET-18 AND INCEPTIONV3-GRU MODELS

Model	#Trainable Params	#Non-Trainable Params	Memory Size	Train Time/Epoch
3D Resnet-18	15,404,675	1,152	58.77 MB	5s
Inception V3-GRU	431,043	21,802,784	83.17 MB	11s

Table 4 highlights notable differences in computational demands, memory consumption, and training efficiency between the 3D ResNet-18 and InceptionV3-GRU models. Despite ResNet-18 possessing a significantly larger number of trainable parameters (15,404,675 parameters) compared to the InceptionV3-GRU (431,043 parameters), it demonstrates faster training performance, completing an epoch in 5 seconds, while InceptionV3-GRU requires 11 seconds per epoch. This slower training time for InceptionV3-GRU can be attributed to its more complex architecture, where the InceptionV3 module must first extract image features before passing them to the GRU layer for temporal sequence modeling. This multi-stage feature extraction process imposes additional computational overhead, despite the fewer parameters being updated during training. Furthermore, InceptionV3-GRU contains a substantial number of parameters (total of 22,233,827 parameters) compared to 3D ResNet-18 (total of 15,405,827 parameters) with higher

memory consumption, requiring 83.17 MB, in contrast to 3D ResNet-18's 58.77 MB. The efficiency of ResNet-18 can be attributed to its streamlined residual architecture, which enables effective training of a larger parameter set without excessive computational complexity. In contrast, the InceptionV3-GRU model prioritizes feature extraction at the cost of slower training speed and greater memory usage, reflecting its design for tasks that benefit from comprehensive pre-trained feature representation.

In addition to the initial approach, we aim to investigate the impact of frame count and frame steps on our selected architectures. Inspired by the methodology employed by Funke et al. [13], who utilized 64 frames with 10 Hz snippets (equivalent to 64 frames with 3 frame steps), our investigation will explore various combinations of frame count and frame steps. Specifically, we will evaluate configurations employing 32 and 64 frames, paired with frame steps of 3, 5, 10, and 15. Through systematic exploration of these parameters, we endeavor to elucidate their influence on the performance and efficacy of our architectures, thereby contributing valuable insights for the optimization of similar systems. However, it's worth mentioning that knot-tying data lacks the necessary length to accommodate the configuration involving 64 frames with a 15-frame step. Therefore, this configuration will be omitted from our experiment.

The experimental findings presented in Table 5 shed light on the intricate relationship between frame count, frame steps, and model performance across various surgical tasks. Firstly, the 3D ResNet-18 architecture emerges as a robust performer across different configurations, showcasing consistent high accuracy levels across suturing, knot-tying, and needle-passing tasks. Despite minor fluctuations, its resilience underscores its suitability for handling diverse data configurations effectively. This stability underscores the adaptability of the model, which is crucial in real-world surgical applications where data variability is inevitable.

In contrast, the InceptionV3-GRU model exhibits more variability in performance across different configurations. While it struggles with lower accuracy in certain setups, particularly evident in knot-tying tasks, it still demonstrates commendable performance in other configurations. This variability offers valuable insights into the nuanced impact of frame count and frame steps on model accuracy, emphasizing the importance of careful configuration selection to optimize performance. Moreover, the InceptionV3-GRU model's ability to maintain competitive accuracy levels in favorable configurations highlights its potential for adaptation to specific task requirements.

The analysis of frame count and frame steps from Table 5 also reveals intriguing patterns in model performance across surgical tasks. Notably, for 32 frames, both Suturing and Needle Passing tasks exhibit optimal performance with a frame step of 15, suggesting a preference for generalized information from a broader temporal context. However, as the frame step decreases, indicating a finer-grained temporal resolution, performance diminishes and only gradually recovers, failing to surpass the accuracy achieved with a 15 frame steps. This trend suggests that these tasks benefit from assimilating information from multiple points across time rather than relying on narrow, detailed windows. Conversely, the Knot-Tying task showcases a distinct behavior, as models perform best with lower frame steps, particularly excelling with 5 frame steps, indicating a need for finer-grained information to achieve optimal accuracy.

TABLE V
EFFECT OF FRAME COUNT AND FRAME STEPS ON MODEL PERFORMANCE ACROSS SURGICAL TASKS

Model	Snippet Configuration	Accuracy		
		Suturing	Knot-Tying	Needle-Passing
3D Resnet-18	**32 frames, frame steps=15, 5, 15**	**100.00**	**100.00**	**100.00**
3D Resnet-18	32 frames, frame steps=15	100.00	84.21	100.00
3D Resnet-18	32 frames, frame steps=10	100.00	100.00	100.00
3D Resnet-18	32 frames, frame steps=5	100.00	100.00	100.00
3D Resnet-18	32 frames, frame steps=3	100.00	100.00	100.00
3D Resnet-18	64 frames, frame steps=15	73.07	-	100.00
3D Resnet-18	64 frames, frame steps=10	88.37	75.00	100.00
3D Resnet-18	64 frames, frame steps=5	100.00	100.00	100.00
3D Resnet-18	**64 frames, frame steps=3**	**100.00**	**100.00**	**100.00**
InceptionV3-GRU	**32 frames, frame steps=15, 5, 15**	**85.71**	**77.42**	**100.00**
InceptionV3-GRU	32 frames, frame steps=15	85.71	26.32	100.00
InceptionV3-GRU	32 frames, frame steps=10	61.50	51.72	80.03
InceptionV3-GRU	32 frames, frame steps=5	70.59	77.42	92.96
InceptionV3-GRU	32 frames, frame steps=3	76.95	69.60	94.51
InceptionV3-GRU	64 frames, frame steps=15	69.23	-	80.00
InceptionV3-GRU	64 frames, frame steps=10	48.69	58.33	88.79
InceptionV3-GRU	64 frames, frame steps=5	51.68	60.38	89.64
InceptionV3-GRU	**64 frames, frame steps=3**	**73.33**	**62.13**	**92.19**
I3D ConvNet (RGB) [13]	**64 frames, frame steps=3**	**100.00**	**95.80**	**96.40**
I3D ConvNet (OF) [13]	**64 frames, frame steps=3**	**100.00**	**95.10**	**100.00**

When considering 64 frames, a consistent trend emerges across all tasks, with models performing better with lower frame steps. Surprisingly, even tasks like Suturing and Needle Passing, which benefit from generalized information, exhibit this trend. This phenomenon suggests that while higher frame steps facilitate generalization, an excess of fine-grained information inherent in higher frame counts can overwhelm the model, leading to decreased performance. This is evident when comparing the accuracy of the more generalized 32 frames, 15 frame steps configuration against the more fine-grained 64 frames, 3 frame steps set up, where the former consistently outperforms the latter across tasks. Despite these nuances, the 3D ResNet-18 model consistently outshines its counterparts, achieving perfect scores on numerous configurations across all tasks, surpassing even the performance of the I3D ConvNet by Funke et al. [13] in several instances, showcasing its robustness and adaptability in modeling surgical tasks.

TABLE VI
COMPARISON WITH OTHER MODELS

Model	Accuracy		
	Suturing	Knot-Tying	Needle-Passing
CNN-LSTM [27]	98.40	94.80	98.40
CNN [28]	100.00	92.10	100.00
I3D ConvNet (RGB) [13]	100.00	95.80	96.40
I3D ConvNet (OF) [13]	100.00	95.10	100.00
MT-TSN [14] w/o attention	100.00	97.20	100.00
MT-TSN [14] with attention	100.00	97.20	100.00
3D Resnet-18 (Ours)	100.00	100.00	100.00
InceptionV3-GRU (Ours)	85.71	77.42	100.00

In Table 6, we benchmark our models—3D ResNet-18 and InceptionV3-GRU—against existing state-of-the-art models evaluated on the JIGSAWS dataset. Previous studies, such as Funke et al. [13] with the I3D ConvNet (RGB) model, achieved 100% accuracy in suturing, 95.8% in knot-tying, and 96.40% in needle-passing while the I3D ConvNet (OF) variant reached 100% in suturing, 95.1% in knot-tying, and 100% in needle-passing. Jian et al. [14], using their multi-task TSN (MT-TSN) model, achieved up to 100% in suturing and needle-passing as well as 97.2% accuracy in knot-tying both with and without attention mechanisms. These models have set a high benchmark for video-based surgical skill assessment. Compared to these approaches, our 3D ResNet-18 model performs exceptionally well, achieving 100% accuracy across all three tasks—suturing, knot-tying, and needle-passing—representing a substantial improvement over models like CNN-LSTM [27], which achieved lower accuracy in suturing (98.4%), knot-tying (94.8%) and needle-passing (98.4%) as well as CNN [28] with lower accuracy in knot-tying (92.1%). Our InceptionV3-GRU model, while competitive with 100% accuracy in needle-passing, showed mixed results with 85.71% in suturing and 77.42% in knot-tying, suggesting potential limitations in the GRU's ability to capture intricate temporal patterns compared to the 3D ResNet-18 and I3D ConvNet.

One of the primary limitations of the JIGSAWS dataset is its focus on controlled, simulated environments, which do not fully reflect the complexities of real-world surgical procedures. While simulators provide consistency and control, they lack the realism needed to assess the nuanced challenges that trained surgeons face during live surgeries. In contrast, using actual surgical data, such as video recordings and motion data, offers more accurate skill assessments but presents difficulties in standardization, making reproducibility a challenge. To further enhance these systems, expanding datasets to include more varied and representative surgical procedures is essential. This could involve incorporating virtual reality (VR) simulations, offering a realistic and controlled environment for assessing technical skills and training on rare but significant intraoperative events, such as hemorrhage or vascular injury [29], [30]. Additionally, the integration of AI into surgery brings forth significant ethical considerations, including privacy, transparency, accountability for errors, technical robustness, bias, and discrimination. Addressing these ethical challenges and ensuring the technical reliability of AI technologies are crucial to safeguarding patient safety and fostering trust in AI-driven surgical assessment tools [31].

The integration of advanced deep learning models, such as 3D ResNet-18 and InceptionV3-GRU, into laparoscopic surgical skill assessment has significant practical implications for clinical training and real-time evaluation. These models can automatically analyze video-based data from laparoscopic procedures, offering objective, consistent assessments of critical surgical tasks, including suturing, knot-tying, and needle-passing. This method addresses the limitations of traditional assessment frameworks, which are often subjective and reliant on expert evaluation. By providing automated, data-driven feedback, such systems can enhance both the accuracy and efficiency of skill evaluation during surgical training. This trend towards automation aligns with the broader movement in surgery to reduce human error and improve patient outcomes through the application of data science and machine learning [8], [30].

IV. CONCLUSION

In this study, we evaluated the classification of laparoscopic surgical skill expertise using advanced deep learning models—specifically, 3D ResNet-18 and InceptionV3-GRU—within the Temporal Segment Network framework applied to the JIGSAWS dataset. The 3D ResNet-18 model demonstrated exceptional performance, consistently achieving 100% accuracy across suturing, knot-tying, and needle-passing tasks, outperforming the InceptionV3-GRU architecture and surpassing existing state-of-the-art models. Our analysis of frame count and frame steps revealed that while 3D ResNet-18 maintained robust performance across various configurations, optimal settings varied depending on the specific surgical task, highlighting the importance of tailored approaches in skill assessment. The integration of models like 3D ResNet-18 offers objective and consistent evaluations of critical surgical tasks, enhancing the accuracy and efficiency of skill assessment in surgical training.

Future work should involve expanding datasets to include more varied and representative surgical scenarios, potentially incorporating virtual reality simulations to provide realistic yet controlled environments. Additionally, addressing ethical considerations such as privacy, transparency, and bias is crucial for the safe integration of AI-driven assessment tools in surgical practice. Future developments in this field hold significant potential for improving surgical education and patient outcomes through the application of deep learning and artificial intelligence.

ACKNOWLEDGMENT

This work was supported by the Faculty of Sciences and Mathematics Universitas Diponegoro Indonesia 2023 and Ministry of Research Technology and Higher Education of the Republic of Indonesia numbers: 601-96/UN7.D2/PP/VI/2024.

REFERENCES

[1] Y. Gao *et al.*, "JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS): A Surgical Activity Dataset for Human Motion Modeling," *Model. Monit. Comput. Assist. Interv. – MICCAI Work.*, pp. 1–10, 2014.

[2] L. I. Basunbul, L. S. S. Alhazmi, S. A. Almughamisi, N. M. Aljuaid, H. Rizk, and R. Moshref, "Recent Technical Developments in the Field of Laparoscopic Surgery: A Literature Review," *Cureus*, vol. 14, no. 2, 2022, doi: 10.7759/cureus.22246.

[3] M. Ali, R. M. G. Pena, G. O. Ruiz, and S. Ali, "A comprehensive survey on recent deep learning-based methods applied to surgical data," 2022, doi: 10.48550/arXiv.2209.01435.

[4] F. von Bechtolsheim *et al.*, "Does practice make perfect? Laparoscopic training mainly improves motion efficiency: a prospective trial.," *Updates Surg.*, vol. 75, no. 5, pp. 1103–1115, Aug. 2023, doi:10.1007/s13304-023-01511-w.

[5] C. W. Reynolds *et al.*, "Evidence supporting performance measures of laparoscopic appendectomy through a novel surgical proficiency assessment tool and low-cost laparoscopic training system," *Surg. Endosc.*, vol. 37, no. 9, pp. 7170–7177, 2023, doi: 10.1007/s00464-023-10182-y.

[6] B. Chen *et al.*, "Trends and hotspots in research on medical images with deep learning: a bibliometric analysis from 2013 to 2023," *Front. Artif. Intell.*, vol. 6, no. November, pp. 1–14, 2023, doi:10.3389/frai.2023.1289669.

[7] I. Galić, M. Habijan, H. Leventić, and K. Romić, "Machine Learning Empowering Personalized Medicine: A Comprehensive Review of Medical Image Analysis Methods," *Electron.*, vol. 12, no. 21, 2023, doi: 10.3390/electronics12214411.

[8] K. Guo *et al.*, "Current applications of artificial intelligence-based computer vision in laparoscopic surgery," *Laparosc. Endosc. Robot. Surg.*, vol. 6, no. 3, pp. 91–96, 2023, doi: 10.1016/j.lers.2023.07.001.

[9] L. Alzubaidi *et al.*, *Review of deep learning: concepts, CNN architectures, challenges, applications, future directions*, vol. 8, no. 1. Springer International Publishing, 2021. doi: 10.1186/s40537-021-00444-8.

[10] F. M. Shiri, T. Perumal, N. Mustapha, and R. Mohamed, "A Comprehensive Overview and Comparative Analysis on Deep Learning Models: CNN, RNN, LSTM, GRU," 2023, doi:10.48550/arXiv.2305.17473.

[11] S. Tiwari, G. Jain, D. K. Shetty, M. Sudhi, J. M. Balakrishnan, and S. R. Bhatta, "A Comprehensive Review on the Application of 3D Convolutional Neural Networks in Medical Imaging," *Eng. Proc.*, vol. 59, no. 1, pp. 1–9, 2023, doi: 10.3390/engproc2023059003.

[12] J. Liu, T. Wang, A. Skidmore, Y. Sun, P. Jia, and K. Zhang, "Integrated 1D, 2D, and 3D CNNs Enable Robust and Efficient Land Cover Classification from Hyperspectral Imagery," *Remote Sens.*, vol. 15, no. 19, 2023, doi: 10.3390/rs15194797.

[13] I. Funke, S. T. Mees, J. Weitz, and S. Speidel, "Video-based surgical skill assessment using 3D convolutional neural networks.," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 14, no. 7, pp. 1217–1225, Jul. 2019, doi: 10.1007/s11548-019-01995-1.

[14] Z. Jian, W. Yue, Q. Wu, W. Li, Z. Wang, and V. Lam, "Multitask Learning for Video-based Surgical Skill Assessment," in *2020 Digital Image Computing: Techniques and Applications (DICTA)*, 2020, pp. 1–8. doi: 10.1109/DICTA51227.2020.9363408.

[15] H. E. Kim, A. C. Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification : a literature review," *BMC Med. Imaging*, pp. 1–13, 2022, doi:10.1186/s12880-022-00793-7.

[16] D. Tran, H. Wang, L. Torresani, J. Ray, Y. Lecun, and M. Paluri, "A Closer Look at Spatiotemporal Convolutions for Action Recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 6450–6459, 2018, doi: 10.1109/CVPR.2018.00675.

[17] R. Achmad, Y. Tokoro, J. Haurissa, and A. Wijanarko, "Recurrent Neural Network-Gated Recurrent Unit for Indonesia-Sentani Papua Machine Translation," *J. Inf. Syst. Informatics*, vol. 5, no. 4, pp. 1449–1460, 2023, doi: 10.51519/journalisi.v5i4.597.

[18] X. Li, "CNN-GRU model based on attention mechanism for large-scale energy storage optimization in smart grid," *Front. Energy Res.*, vol. 11, no. July, pp. 1–16, 2023, doi: 10.3389/fenrg.2023.1228256.

[19] L. Lu, C. Zhang, K. Cao, T. Deng, and Q. Yang, "A Multichannel CNN-GRU Model for Human Activity Recognition," *IEEE Access*, vol. 10, pp. 66797–66810, 2022, doi:10.1109/access.2022.3185112.

[20] H. Ullah and A. Munir, "Human Activity Recognition Using Cascaded Dual Attention CNN and Bi-Directional GRU Framework," *J. Imaging*, vol. 9, no. 7, 2023, doi: 10.3390/jimaging9070130.

[21] M. Shafiq and Z. Gu, "Deep Residual Learning for Image Recognition: A Survey," *Appl. Sci.*, vol. 12, no. 18, 2022, doi:10.3390/app12188972.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 770–778, 2016, doi:10.1109/CVPR.2016.90.

[23] J. Cao, M. Yan, Y. Jia, X. Tian, and Z. Zhang, "Application of a modified Inception-v3 model in the dynasty-based classification of ancient murals," *EURASIP J. Adv. Signal Process.*, vol. 2021, no. 1, 2021, doi: 10.1186/s13634-021-00740-8.

[24] K. Kasa, D. Burns, M. G. Goldenberg, O. Selim, C. Whyne, and M. Hardisty, "Multi-Modal Deep Learning for Assessing Surgeon Technical Skill," *Sensors*, vol. 22, no. 19, 2022, doi:10.3390/s22197328.

[25] Q.-Q. Hong, L. Yang, and B. Zeng, "RANET: A Grasp Generative Residual Attention Network for Robotic Grasping Detection," *Int. J. Control. Autom. Syst.*, vol. 20, no. 12, pp. 3996–4004, 2022, doi: 10.1007/s12555-021-0929-8.

[26] W. Zhou, J. Lu, Z. Xiong, and W. Wang, "Leveraging TCN and Transformer for effective visual-audio fusion in continuous emotion recognition," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2023, pp. 5756–5763. doi: 10.1109/cvprw59228.2023.00610.

[27] X. A. Nguyen, D. Ljuhar, M. Pacilli, R. M. Nataraja, and S. Chauhan, "Surgical skill levels: Classification and analysis using deep neural network model and motion signals.," *Comput. Methods Programs Biomed.*, vol. 177, pp. 1–8, Aug. 2019, doi:10.1016/j.cmpb.2019.05.008.

[28] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P. A. Muller, "Accurate and interpretable evaluation of surgical skills from kinematic data using fully convolutional neural networks," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 14, no. 9, pp. 1611–1617, 2019, doi: 10.1007/s11548-019-02039-4.

[29] R. Pedrett, P. Mascagni, G. Beldi, N. Padoy, and J. L. Lavanchy, "Technical skill assessment in minimally invasive surgery using artificial intelligence: a systematic review," *Surg. Endosc.*, vol. 37, no. 10, pp. 7412–7424, 2023, doi: 10.1007/s00464-023-10335-z.

[30] E. Yanik *et al.*, "Deep neural networks for the assessment of surgical skills: A systematic review," *J. Def. Model. Simul.*, vol. 19, no. 2, pp. 159–171, 2022, doi: 10.1177/15485129211034586.

[31] G. Karimian, E. Petelos, and S. M. A. A. Evers, "The ethical issues of the application of artificial intelligence in healthcare: a systematic scoping review," *AI Ethics*, vol. 2, no. 4, pp. 539–551, 2022, doi:10.1007/s43681-021-00131-7.