

Artificial Intelligence for Automated Operative Note Generation in Endoscopic Neurosurgery

Nguyen Thanh Huy^{1*}, Pham Quang Minh¹, Le Thi Bich², Tran Van Nam³

¹ Department of Neurosurgery and Endoscopic Surgery, Faculty of Medicine, Vietnam National University of Agriculture, Hanoi, Vietnam.

² Department of AI-Assisted Operative Documentation, Faculty of Medicine, Can Tho University of Medicine and Pharmacy, Can Tho, Vietnam.

³ Department of Automated Operative Note Generation, Faculty of Medicine, Hue University, Hue, Vietnam.

*E-mail ✉ huy.nguyen@vnua.edu.vn

Abstract

Operative notes form a vital part of patient management. However, their manual preparation is vulnerable to human mistakes, particularly under the stress of busy clinical settings. Automatically generating these notes from surgical video recordings can reduce administrative demands, increase reliability, and provide richer contextual detail. For endoscopic pituitary surgery, experts reached consensus on 27 distinct steps. In the 97 videos collected for this research, an experienced surgeon annotated the precise timing of each step. A three-stage model was designed to automatically detect the presence of any given step within a video. The first stage applied a convolutional neural network to classify each video frame in binary fashion for every step. The second stage used a discriminator to convert these frame-level results into a binary decision about whether each step was present in the entire video. The third stage utilized an accumulator to combine these video-level decisions into a multi-label classification of all steps present in the recording. Trained on 77 videos and assessed on a held-out set of 20 videos, the system achieved a weighted F1 score of 0.80. The derived classifications were inserted into a standardized clinical template and further augmented with quantitative video analytics. This investigation therefore confirms that automatic production of operative notes from surgical videos is achievable and can meaningfully support surgeons in their documentation responsibilities.

Keywords: Computer vision, Image recognition, Operation report, Step recognition, Surgical AI, Workflow analysis

Introduction

Operative notes provide the official written account of a surgical intervention, detailing information from patient demographics to the sequence of procedural actions. They play an indispensable role in ongoing patient care, clinical handover, quality auditing, scientific research, training, and legal documentation [1]. Despite their importance, these records are often deficient, imprecise, insufficiently comprehensive, difficult to read, or

produced with considerable delay [1, 2]. The primary reasons relate to human factors, including growing administrative loads on surgical teams and limited time and support [1]. For example, research in a Dublin hospital in 2021 found that 80% of operative notes were incomplete and required an average of 10 minutes to complete [2].

Surgical video recording has become increasingly routine in modern practice [3]. Examination of such recordings has revealed that they frequently contain clinically relevant information absent from conventional operative notes [4]. In one analysis of laparoscopic cholecystectomy procedures, several critical surgical steps documented in the videos were missing from the corresponding written notes; for instance, achievement of the critical view of safety went unreported in 25% of cases where it was clearly performed [4]. Surgical video

Access this article online

<https://smerpub.com/>

Received: 11 April 2026; Accepted: 29 June 2026

Copyright CC BY-NC-SA 4.0

How to cite this article: Huy N, Minh P, Bich L, Nam T. Artificial Intelligence for Automated Operative Note Generation in Endoscopic Neurosurgery. J Med Sci Interdiscip Res. 2026;6(1):279-89. <https://doi.org/10.51847/NcOfaurK4b>

analysis can therefore substantially improve both the accuracy and depth of information captured in operative notes [5]. Manual review of videos is, however, highly resource-intensive, limiting its practicality in routine clinical workflows [6, 7]. Automated approaches hold considerable promise for overcoming these limitations [6-8] and for directly assisting surgeons in note preparation [5, 6].

Surgical videos can be hierarchically decomposed into broader phases and more refined individual steps [9, 10]. Artificial neural networks (ANNs) have proven effective for the automated identification of these elements, particularly in laparoscopic cholecystectomy, where large public datasets have accelerated progress [8, 10]. Comparable results have been obtained in other procedures, such as endoscopic pituitary surgery [11, 12]. The core mechanism typically involves a convolutional neural network (CNN) that analyses individual frames to determine the depicted phase or step, allowing detection of transitions [10, 11]. Incorporating temporal dynamics—either through time-aware neural architectures or statistical modeling of sequence order—can further improve accuracy [10, 11].

Multiple investigations have highlighted the potential of phase and step recognition to enable fully automatic operative note generation [6, 8, 10, 11]. For example, in laparoscopic cholecystectomy, elevated predictive uncertainty in phase transitions by the ANN was linked to adverse intraoperative events, suggesting its utility as an alert mechanism within generated notes [13]. In contrast, operative notes for endoscopic pituitary surgery reference far more detailed steps (27 steps versus the smaller number of phases, such as 8, used in earlier studies) [13, 14]. Additionally, overlapping or concurrent steps convert the classification task from single-label to multi-label [14].

A dedicated three-stage architecture was therefore developed to manage this complexity. In the initial stage, a CNN performs per-frame binary classification for each step independently. The intermediate stage applies a discriminator that aggregates frame predictions into a binary video-level decision for each step, testing several threshold-based implementations. The final stage uses an accumulator to produce multi-label output by synthesizing the video-level decisions and enforcing clinically derived logical constraints (e.g., preventing a step that logically requires a preceding step).

These classifications are then mapped onto a pre-established clinical template for operative notes and

enriched with a visual summary showing the duration of three key phases. Accordingly, the present work offers two main contributions:

1. The first demonstration of fully automated operative note generation in endoscopic pituitary surgery based on workflow recognition.
2. A new three-stage architecture that functions as a multi-label classifier for determining the presence of surgical steps in an individual video.

Materials and Methods

Dataset

A collection of 97 endoscopic pituitary surgery videos was assembled at the National Hospital for Neurology and Neurosurgery (Queen Square, London, United Kingdom) from 30 August 2018 to 20 February 2021. The study received approval from the institutional governance committee, and all participants provided written informed consent. Videos were excluded when the procedure represented a revision surgery conducted within six months of the initial operation or when substantial portions of the recording were absent. The 97 videos exhibited a median duration of 74 minutes, with 15 recordings showing minor gaps in footage. All procedures were captured using a high-definition endoscope (Hopkins Telescope) at resolutions between 720p and 2160p, recorded at 25 frames per second (fps), and saved in mp4 format. To ensure uniformity and reduce processing demands, all videos were standardized to 720p resolution and converted to JPEG images at 1 fps. Twenty-seven surgical steps were established as essential markers for the automatic production of operative notes. These steps were determined through a Delphi consensus process involving specialist endoscopic pituitary surgeons and were delineated according to anatomical landmarks, operative maneuvers, and instrument application [14]. **Table 1** presents the correspondence between each step and the associated statements in the operative notes. For clarity in discussion, the prefix “S” (denoting “surgical step”) is appended to the step number when referencing a specific step. In this investigation, “instruments” refer specifically to S08 (drill), S11 (stealth pointer), S12 (Doppler), S18 (surgiflo), S19 (bipolar), and S20 (spongostan placement). These instruments support a “core” step and may therefore coincide with it, but they are treated as independent surgical steps for classification purposes.

Table 1. Mapping of surgical steps to corresponding statements in operation notes. Phases are defined in Ref. [14].

A step enclosed in square brackets [] within a statement denotes an extension of the primary statement by the content of the bracketed step. A reliant step denotes one that depends on the presence of another step, as indicated in the “reliant” column. Note that S09 and S10 represent distinct steps that correspond to the same operative note statement (spanning both rows).

Phase	Step	Classification	Dependency	Operative note description
01	01	Core	–	The middle and superior turbinates were gently lateralized with a Freer elevator.
	02	Core	–	The sphenoid ostium was located [03].
	03	Core	02	It was subsequently enlarged using Kerrison rongeurs.
	04	Optional	–	The septum was displaced [05] until the contralateral sphenoid ostium became visible.
	05	Optional	–	A partial posterior septectomy was then completed.
	06	Core	–	The sphenoid sinus was entered, and sphenoid septations [07] were removed to expose the sellar face. [08]
	07	Core	–	The mucosal tissue was also removed.
	08	Instrument	–	A high-speed drill was utilized to accomplish this step.
02	09	Core	–	The sella, carotid prominence, optic prominence, and optic–carotid recesses...
	10	Core	–	...were subsequently identified bilaterally [11] [12].
	11	Instrument	09/10	Their locations were verified using neuronavigation.
	12	Instrument	09/10	Verification was also performed with a Micro Doppler probe.
	13	Core	–	The sella was carefully opened using rongeurs.
	14	Core	–	A cruciate durotomy was created with a retractable scalpel.
	15	Core	–	The tumor became immediately visible upon entering the sella and was excised in a piecemeal manner using curettes and pituitary rongeurs.
	16	Core	–	The evacuated pituitary fossa was inspected, revealing descent of the diaphragm.
03	17	Core	18/19/20	Hemostasis was achieved using [18-20].
	18	Instrument	17	Surgiflo was applied.
	19	Instrument	17	Bipolar cautery was used.
	20	Instrument	–	Spongostan placement was also performed.
	21	Optional	–	A fat graft was harvested from the left lower abdominal quadrant and positioned over the defect.
	22	Optional	–	A MedPor implant was appropriately sized and inserted.
	23	Optional	–	A fascia lata graft was harvested and placed over the reconstruction.
	24	Optional	–	Evicel was applied.
	25	Optional	–	Adherus dural sealant was administered.
	26	Optional	–	Bismuth-soaked ribbon gauze was used to pack the nasal cavity and reinforce the repair.
	27	Core	–	Residual debris was removed from the nasal cavity and choana.

Two specialist surgeons independently reviewed all 97 videos and jointly assigned timestamps to each surgical step by consensus. A single step could occur multiple times in non-contiguous segments within the same video, and each occurrence received a separate timestamp. Furthermore, the 27 steps varied considerably in duration (median range of 1–11 minutes) and did not always

follow a fixed order; these characteristics needed to be addressed during classification.

A video in which a particular step has been annotated is termed a “positive video”; otherwise, it is designated a “negative video.” Likewise, a frame containing a step is a “positive frame,” while one without the step is a “negative frame.” **Figure 1** illustrates the distribution of annotations for each step across the full set of 97 videos,

revealing substantial variation in annotation counts. The overall number of annotations is closely associated with the count of positive videos, although certain steps possess multiple additional annotations. When

considering only positive videos, core steps appear in most recordings, while optional steps and instrument-related feature in considerably fewer videos.

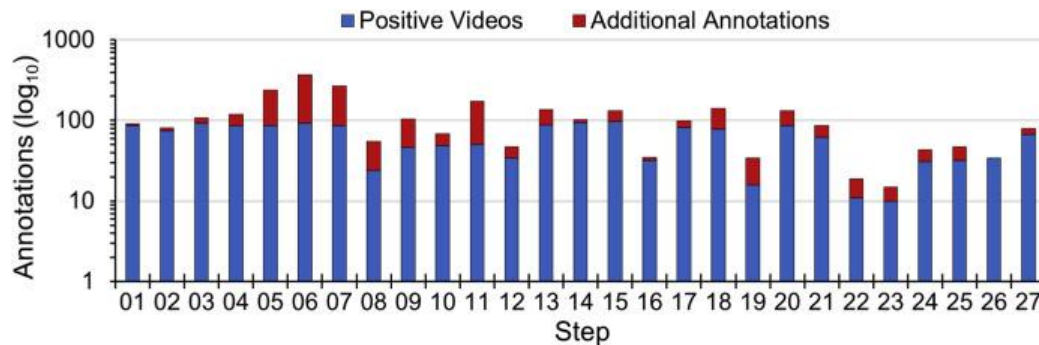


Figure 1. Bar chart displaying the number of annotations for each surgical step across all 97 videos. “Positive Videos” represents the count of videos in which a step appears at least once; adding this value to “Additional Annotations” yields the total number of annotations. The annotation scale employs a base-10 logarithmic transformation.

Twenty videos were randomly selected to form the fixed test dataset for all stages of the classification framework. For the remaining 77 videos, the composition of the training and validation sets was adjusted at each step and stage of the architecture to mitigate imbalance in the distribution of positive videos, as detailed in the following section.

Surgical step classification

The primary goal is to determine automatically which of the 27 surgical steps occur in any individual video. A

three-stage framework was therefore designed for this purpose, as shown in **Figure 2**. Training proceeded sequentially across the stages, allowing full hyperparameter optimization of Stage I on the validation data before proceeding to Stage II. We implemented the entire system in Python 3.82 [15] with PyTorch 1.8.13 [16]. We will release the source code publicly. All experiments were run on an NVIDIA Tesla V100 Tensor Core GPU possessing 32 GB of memory and CUDA 11.25 [17].

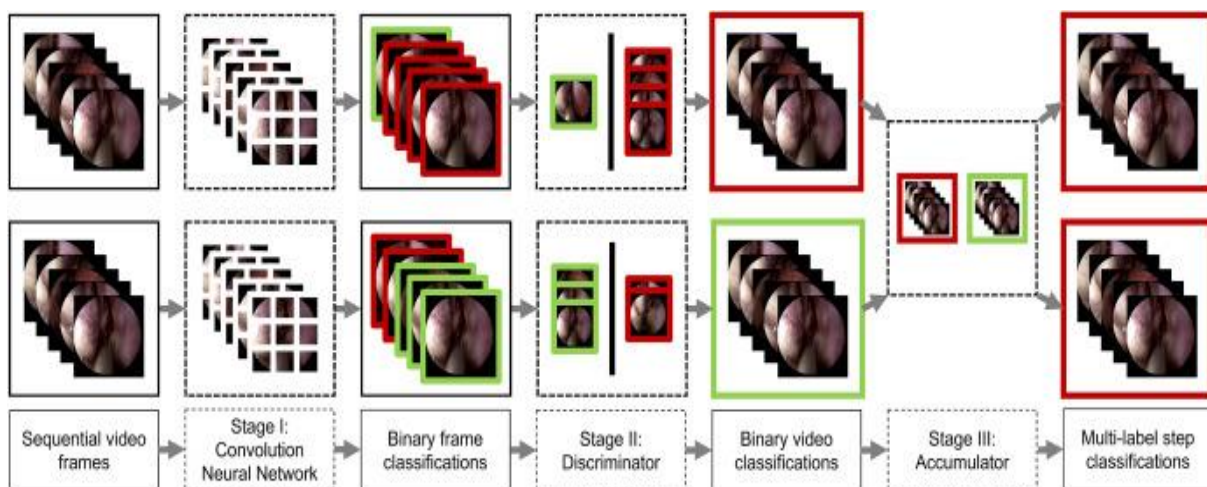


Figure 2. Schematic overview of the proposed architecture for automatic operative note generation. Stage I applies a convolutional neural network to perform binary image classification for each step on every video frame. Stage II employs a discriminator to achieve binary video classification for each step. Stage III uses an

accumulator to perform multi-label step classification for each video. Green outlines indicate positive classifications, while red outlines indicate negative classifications. This illustration depicts only two steps (S02 and S03) for simplicity; the complete system processes all 27 steps. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Stage I: Binary frame classification (CNN)

The opening stage develops a dedicated binary classifier at the frame level for each step. Earlier research on a 50-video subset of the current 97-video collection examined multi-class frame classification for seven steps and found that ResNet50 performed best [11, 18]. For any given step, ResNet50 identifies positive versus negative frames by recognizing characteristic visual elements. For example, when processing S02 (identification of the sphenoid ostium, Phase 1), the model detects specific anatomical features that distinguish it from other steps. In S17 (hemostasis, Phase 3), tissue repair methods alter the typical visual field, allowing the network to distinguish this activity.

We therefore adopted a ResNet50 model initialized with ImageNet pre-training. Its terminal layer was replaced with a linear classifier paired with softmax activation. Training used cross-entropy loss and a stochastic gradient descent optimizer with a learning rate of 0.001 and momentum of 0.9. Models were trained for 8 epochs using a batch size of 8. Training images received random resizing, cropping, and horizontal flipping, whereas validation images were left unaltered. All images underwent color normalization and were resized to 224×224 pixels to conform to ImageNet standards. The epoch achieving the best weighted F1 score on the validation set was selected, given that this metric effectively mitigates risks of inadequate precision or recall.

Because the 77 non-test videos do not all contain every step, training and validation for each step drew exclusively from videos positive for that step. These videos were randomly partitioned in an approximate 4:1 ratio for training and validation. The ResNet50 model processed both positive and negative frames drawn from

the respective sets. Positive frames for each annotated step instance included the marked frame and all following frames until the next non-instrument step. To counteract severe class imbalance, negative frames were randomly subsampled from within each positive video until their count matched that of the positive frames. No frames from videos lacking the step entirely were included in training or validation.

Preparing the Stage II discriminators required classifying every frame in the full dataset. All frames from the 97 videos were therefore color-normalized and resized to 224×224 pixels. Each frame was then passed through the highest-performing CNN model for its step, producing both the final binary label and the underlying probability score (before applying the 0.5 decision threshold) as ordered temporal sequences.

Stage II: Binary video classification (discriminator)

The subsequent stage constructs a binary classifier at the video level for each step. To this end, six distinct custom discriminators were designed and evaluated for every step. These discriminators drew inspiration from temporal smoothing techniques, which had previously enhanced outcomes in 7-step multi-class classification tasks when applied to CNN frame outputs on a 50-video subset drawn from the present 97-video collection [11]. Each discriminator computes a single “discrimination number” (δ) derived from the temporally sequenced frame classifications of a video. A video is classified as positive if this value meets or exceeds a predefined “discrimination threshold” (τ). **Table 2** details the six discriminators and the formulas used to compute their respective discrimination numbers.

Table 2. Specifications of the six discriminators developed and assessed for Stage II. A discriminator designates a video as positive when its discrimination number satisfies or surpasses the discrimination threshold (i.e., $\delta \geq \tau$). The table also lists the stride and maximum values applied for each threshold. Note that 16200 represents the highest possible discrimination threshold for integer-based discriminators, given that the longest video in the training set contains 16103 frames.

Discriminator type	Discrimination metric (δ)	Threshold (τ) – step size	Threshold (τ) – maximum
Binary count	Total number of positive frames ($\sum$ positive frames)	100	16200

Binary ratio	Positive frames divided by total frames ($\sum$ positive frames/frames)	0.01	1
Probability count	Sum of frame probabilities ($\sum$ frame probabilities)	100	16200
Probability ratio	Sum of frame probabilities divided by total frames ($\sum$ frame probabilities/frames)	0.01	1
Chain count	Longest consecutive sequence of positive frames	1	16200
Chain ratio	Longest consecutive sequence of positive frames divided by total frames (longest period of positive frames/frames)	0.001	1

For every discriminator, the threshold τ constituted a tunable hyperparameter. Its optimal value was identified through a linear grid search commencing at $\tau = 0$ for all variants, with discriminator-specific maximum values and stride intervals as specified in **Table 2**. The weighted F1 score served as the primary performance metric because it is well-suited to imbalanced data distributions. The τ value that produced the maximum weighted F1 score was adopted as optimal. No separate validation set was necessary because threshold selection occurred in a single traversal; thus, the full set of 77 non-test videos was used. For the subsequent Stage III accumulator, video-level classifications across all steps were required. Accordingly, each video was evaluated for every step using that step's optimal τ , and the resulting classifications were compiled into a step-ordered sequence.

Stage III: Multi-label steps classification (accumulator)

The final stage establishes a multi-label classifier that determines the complete set of steps present in each video. Utilizing the step-ordered classification sequence produced by Stage II, an “accumulator” module was implemented to enforce logical consistency among the predicted steps. These consistency rules, derived from clinical expert consensus, appear in the “Reliant” column of **Table 1**. As an example, a positive classification for S11 (“confirmed using neuronavigation”) is only permissible when either S09 or S10 (“The sella, carotid prominence, optic prominence, and optic-carotid recesses were then identified on both sides”) has also received a positive classification. Should the discriminator indicate S11 as positive while both S09 and S10 remain negative, the prediction for S11 is overridden to negative. The resulting refined classifications were subsequently fed into the operative notes template.

Table 3. Mean weighted F1 scores averaged across the 7 steps for each of the six discriminators, shown both prior to and following application of the accumulator. Values are reported to two significant figures accompanied by standard deviation.

Discriminator type	Discriminator (Training)	Discriminator (Testing)	Accumulator (Training)	Accumulator (Testing)
Binary count	0.72 $\pm$ 0.15	0.73 $\pm$ 0.24	0.71 $\pm$ 0.15	0.73 $\pm$ 0.23
Binary ratio	0.77 $\pm$ 0.13	0.77 $\pm$ 0.21	0.76 $\pm$ 0.13	0.77 $\pm$ 0.18
Probability count	0.72 $\pm$ 0.15	0.73 $\pm$ 0.24	0.71 $\pm$ 0.15	0.73 $\pm$ 0.23
Probability ratio	0.76 $\pm$ 0.15	0.79 $\pm$ 0.20	0.76 $\pm$ 0.12	0.80 $\pm$ 0.18
Chain count	0.60 $\pm$ 0.21	0.61 $\pm$ 0.27	0.58 $\pm$ 0.21	0.60 $\pm$ 0.26
Chain ratio	0.69 $\pm$ 0.19	0.72 $\pm$ 0.26	0.68 $\pm$ 0.18	0.72 $\pm$ 0.25

Operation notes template

The proposed template for automatically produced “smart” operative notes, shown in **Figure 3**, was constructed from three distinct resources. The first source comprised the Royal College of Surgeons in England Good Surgical Practice guidelines, which outlined the core elements required in all operative notes, including patient details, procedural description, and postoperative instructions [19]. The second source was an established

international expert consensus that specified the procedural building blocks particular to endoscopic pituitary surgery, covering phases, steps, instruments, and possible complications [14]. In the current approach, procedural content is generated automatically using step classifications obtained from the three-stage framework together with video-derived metrics, specifically the 3-phase duration summary, based on earlier multi-class models reported in Ref. [11]. The third source involved

any unfinished portions of the note, which would be addressed either through structured templates or direct manual completion by the operating team.

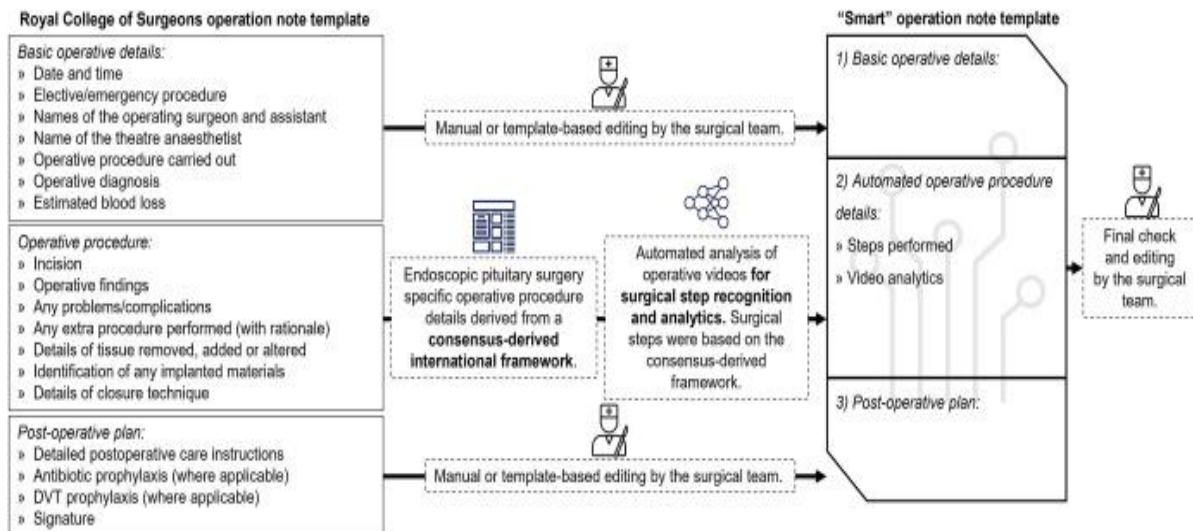


Figure 3. Diagram illustrating the intended workflow for producing the automatically generated “smart” operation notes template.

The template synthesizes information from three sources: the Royal College of Surgeons guidelines [19], an international Delphi consensus study [14], and subsequent manual adjustments by surgical staff. The surgical team remains responsible for final review and confirmation.

Results and Discussion

Stage I: CNN

Figure 4 illustrates the weighted F1 scores achieved by the complete three-stage system across all 27 steps. For the CNN component in Stage I, the initial green bar for

each step shows training performance, and the subsequent yellow bar shows validation results. Averaged across all 27 steps, the mean weighted F1 score was 0.90 ± 0.04 on the training data and 0.83 ± 0.06 on the validation data. Overall, ResNet50 demonstrated strong capability in separating positive from negative frames for most steps, with acceptable transfer between training and validation sets. A clear outlier was S19 (bipolar cautery, instrument), which recorded a notably lower validation score of 0.65 despite a training score of 0.91. This limited generalization is probably explained by the scarcity of positive examples for this step (**Figure 1**), which introduced considerable visual heterogeneity.

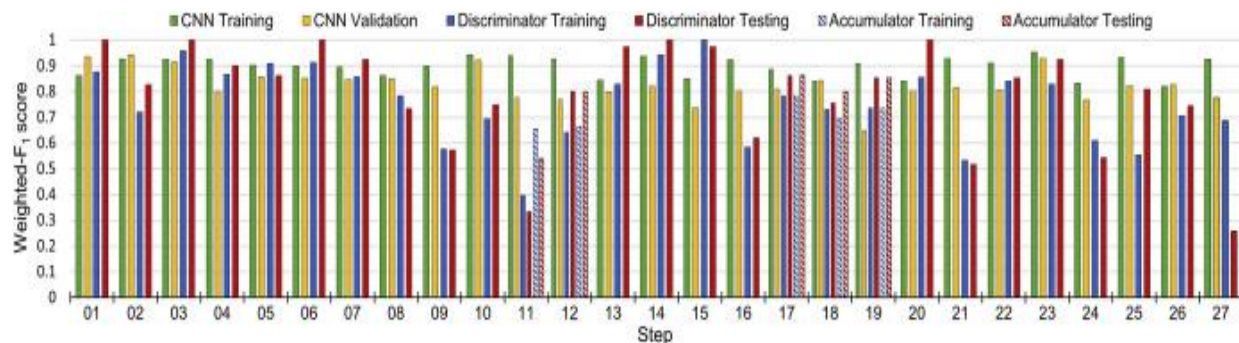


Figure 4. Bar chart depicting weighted F1 scores for the three stages across the 27 surgical steps according to the legend. Stage I refers to CNN (ResNet50) binary frame-level classifications; Stage II to Discriminator

(probability fraction) binary video-level classifications; and Stage III to Accumulator multi-label classifications after correction for interdependent reliant steps (**Table 1**).

Stage II: Discriminator

Weighted F1 scores for the six discriminators, averaged over the 27 steps, appear in the “Discriminator” column of **Table 3** for both training and testing sets. Results were generally robust. On the training data, the leading performer (binary fraction) attained 0.77, while on the testing data the top variant (probability fraction) reached 0.79. Discriminators relying on fractional values consistently outperformed those based on integer counts, likely because video durations vary widely. Chain-style discriminators produced inferior results compared with both probability- and binary-based methods, which performed similarly. The weaker showing of chain-based approaches can be attributed to intermittent “step flickering,” a phenomenon documented in earlier research [11, 13], in which correctly positive frames are sporadically misclassified as negative because of transient endoscopic obstructions. Such misclassifications disrupt continuous sequences and undermine the reliability of chain-derived discrimination numbers.

Performance transfer from training to testing data was strong, with mean scores remaining nearly equivalent, though variability was mildly elevated on the test set. This consistency indicates good generalisability across the discriminators. Step-by-step weighted F1 scores for the best-performing probability fraction discriminator are presented in **Figure 4**, with the blue bar showing training results and the red bar showing testing results. Most steps yielded strong training scores above 0.75, though notable differences existed; for instance, S15 (tumor excision, core) achieved a perfect score of 1.0, whereas S11 (stealth pointer, instrument) scored only 0.40. Performance proved robust to modest perturbations around the optimal discrimination threshold ($\tau \pm 0.05$). Additional evaluations of ensemble discriminators, generated by merging the probability fraction classifier with the other five variants via intersection ($\cap$, logical AND) or union ($\cup$, logical OR). None of these ensembles surpassed the baseline probability fraction discriminator in weighted F1 score on either training or testing data.

Stage III: Accumulator

The “Accumulator” column in **Table 3** presents the weighted F1 scores for the six discriminators following the accumulation process, averaged across the 27 steps

on both training and testing datasets. In all instances, the weighted F1 score shifted by roughly 0.01 on both datasets, with comparably minor changes in standard deviation. These findings indicate that the accumulator has only a marginal influence on overall classification. Such limited effect aligns with expectations, given the relatively small number of logical rules capable of altering discriminator outputs.

This pattern is further evident in **Figure 4**, which displays the weighted F1 scores for all 27 steps after applying the accumulator to the probability fraction discriminator outputs, enforcing dependencies for reliant steps (**Table 1**). The fifth bar (dashed blue) corresponds to training performance, and the sixth bar (dashed red) corresponds to testing performance. The most substantial shifts occurred as a reduction for S03 (anterior sphenoidotomy, core) and an improvement for S11 (stealth pointer, instrument). These directional changes mirror the performance levels of their respective reliant steps at the discrimination stage. As detailed in **Table 1**, S03 depends on S02, which exhibited weaker results, whereas S11 relies on S09 or S10, both of which demonstrated stronger performance.

Principal findings

This study introduced and evaluated a new three-stage architecture using a dataset of 97 endoscopic pituitary surgery videos. The system functions as a multi-label classifier to identify which surgical steps are present in any given recording. The architecture consists of three sequential components: (I) binary frame classification, (II) binary video classification, and (III) multi-label step classification. In Stage I, the ResNet50 CNN proved effective as a binary frame classifier. In Stage II, several original discriminators were developed, among which the binary fraction approach emerged as the most successful for video-level classification. This method designates a video as positive when the proportion of positive frames (termed the discrimination number) meets or exceeds a specified discrimination threshold. In Stage III, a purpose-built accumulator enforced clinical consistency among steps according to predefined logical constraints, with negligible influence on overall classification accuracy. The generated operative notes were additionally supplemented with a 3-phase duration chart produced via previously established techniques. An

illustrative example of an automatically created smart operative note for one video is included as supplementary material.

Comparison to literature

Established techniques for step recognition in surgical videos commonly employ CNNs for spatial feature extraction and recurrent neural networks or temporal CNN variants for capturing time-based patterns [10]. Building on this foundation, we selected ResNet50 for binary frame classification in Stage I of the present architecture, as it was previously identified as the optimal CNN in related work, achieving a 0.67 weighted F1 score for 7-step multi-class recognition on a 50-video subset of the current 97-video dataset [11]. The 0.83 weighted F1 score achieved here by ResNet50 is therefore consistent with prior findings and reflects the anticipated improvement associated with the less complex binary classification problem. Temporal modeling through neural networks was deliberately omitted in this investigation, as individual steps tend to be brief, yield limited positive examples, and require multiple consecutive frames, which would further diminish the available training data. Instead, statistical approaches to temporal sequencing—such as the temporal smoothing strategy described in Ref. [11]—were incorporated in Stage II through the custom discriminators.

Earlier research by Ref. [13] offered an initial demonstration of automatic operative note generation based on 8-phase classification in laparoscopic cholecystectomy. That work trained a spatial-temporal CNN on 52 videos and evaluated it on 15 videos, obtaining an accuracy of 0.80. This result is comparable to the 0.80 weighted F1 score reported in the present study, despite the increased complexity of handling 27 multi-label steps rather than a smaller set of single-label phases.

Strengths and limitations

A notable advantage of the present investigation lies in the substantial size of the 97-video dataset. However, as all recordings originated from a single institution, the extent to which the proposed architecture generalizes to other settings remains uncertain. Expanding the dataset with videos from multiple centers would help address this limitation.

A further strength concerns the development of several innovative discriminators and their demonstrated effectiveness for binary video classification.

Nevertheless, these models exhibit reduced performance on steps where the counts of positive and negative frames are closely balanced, owing to the threshold-based decision mechanism. Refining the discriminators to operate at a finer level of granularity could resolve this shortcoming. Additionally, initializing ResNet50 with pre-training on multi-class classification of 3 phases or 7 steps, or on instrument recognition tasks [11], may improve binary frame-level step classification and thereby boost subsequent discrimination performance. Such pre-training could potentially eliminate the need for intermediate binary classification, enabling direct progression to 27-step multi-label classification. At present, however, the relatively modest dataset size, the scarcity of examples for certain steps (e.g., S23 appearing in only 15 videos), and the brief duration of some steps (e.g., S04 averaging under 60 seconds) indicate that larger volumes of data and more advanced modeling techniques will be necessary. Once these foundations are strengthened, more elaborate discriminators may be explored for multi-step video analysis.

A third strength is the incorporation of an accumulator mechanism to maintain clinical consistency among predicted steps, which has proven effective. A constraint of the current implementation is the fixed nature of the logical rules. Using step-transition probabilities within statistical frameworks, such as hidden Markov models, could enable dynamic rule adjustment and strengthen the accumulator's contribution to multi-label step classification in Stage III.

Finally, the operative notes template benefits from being grounded in both national surgical standards and international consensus recommendations. Further enhancement could involve incorporating still images of critical steps as additional video analytics alongside the existing 3-phase duration chart. Moreover, step flickering can be leveraged to identify procedures that deviate from expected patterns, following the approach described in Ref. [13].

Clinical translation

Although model training requires approximately 12 hours, inference on a single video takes under 2 minutes, allowing operative note statements to be produced promptly after surgery. Subsequent steps include rigorous clinical validation of the generated notes. Once reliability is confirmed, integration into routine clinical workflows can proceed. At that stage, the advantages of automation—namely improved accuracy and detail in

operative notes coupled with decreased administrative workload for clinicians—will become fully apparent [4, 20]. All automatically generated notes will still require final approval by the surgical team. For surgeons who prefer not to rely solely on automated output, the predicted steps can serve as a structured prompt to guide manual documentation, thereby minimizing the risk of overlooking important elements.

To adapt this methodology to other surgical procedures, the following sequence of actions is recommended:

1. Decompose the procedure into clearly delineated surgical steps. This process is best conducted through an international Delphi consensus study [14].
2. Associate each step with a corresponding statement for the operative note, while accounting for dependencies between steps (e.g., **Table 1**). Align with recommendations from the relevant professional regulatory body [19].
3. Record a sufficient number of procedures. A minimum of 50 videos from multiple centers is suggested to promote generalisability [11].
4. Annotate each video with precise start and end timestamps for every surgical step. This annotation process is manual and should be performed by clinicians experienced in the specific procedure [12].
5. Train the three-stage architecture presented in this work on the annotated videos. Hold-out testing is recommended to verify that a predefined performance threshold is met before clinical deployment.
6. Following completion of a new procedure, input the surgical video into the trained model, which will then automatically produce the operative note.

Conclusion

This study demonstrates that automatic generation of operative notes from endoscopic pituitary surgery videos through workflow recognition is feasible and can deliver strong performance with reasonable computational efficiency. It represents the first reported instance of such automated note generation in this surgical domain and further augments the notes with supplementary video analytics. Consequently, the work establishes a foundation for subsequent efforts to translate step recognition techniques into clinical practice, where they can serve as supportive tools for clinicians and help alleviate administrative demands.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. Lefter LP, Walker SR, Dewhurst F, Turner RWL. An audit of operative notes: facts and ways to improve. *ANZ J Surg.* 2008;78(9):800–4. doi:10.1111/j.1445-2197.2008.04654.x
2. Wolfe N, Teeling SP, Ward M, McNamara M, Koshy L. Operation note transformation: the application of lean six sigma to improve the process of documenting the operation note in a private hospital setting. *Int J Environ Res Public Health.* 2021;18(22):12217. doi:10.3390/ijerph182212217
3. Mascagni P, Alapatt D, Laracca GG, Guerriero L, Spota A, Fiorillo C, et al. Multicentric validation of EndoDigest: a computer vision platform for video documentation of the critical view of safety in laparoscopic cholecystectomy. *Surg Endosc.* 2022;36(12):9139–49. doi:10.1007/s00464-022-09112-1
4. Wauben L, van Grevenstein W, Goossens R, van der Meulen F, Lange J. Operative notes do not reflect reality in laparoscopic cholecystectomy. *Br J Surg.* 2011;98(10):1432–7. doi:10.1002/bjs.7576
5. O'Connor N, Sugrue M, Melly C, McGeehan G, Bucholtz M, Crawford A, et al. It's time for a minimum synoptic operation template in patients undergoing laparoscopic cholecystectomy: a systematic review. *World J Emerg Surg.* 2022;17(1):15. doi:10.1186/s13017-022-00411-5
6. Maier-Hein L, Eisenmann M, Sarikaya D, März K, Collins T, Malpani A, et al. Surgical data science – from concepts toward clinical translation. *Med Image Anal.* 2022;76:102306. doi:10.1016/j.media.2021.102306
7. Maier-Hein L, Vedula SS, Speidel S, Navab N, Kikinis R, Park A, et al. Surgical data science for next-generation interventions. *Nat Biomed Eng.* 2017;1(9):691–6. doi:10.1038/s41551-017-0132-7
8. Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial intelligence in surgery: promises and

- perils. *Ann Surg.* 2018;268(1):70–6. doi:10.1097/sla.0000000000002693
9. Meireles OR, Rosman G, Altieri MS, Carin L, Hager G, Madani A, et al. SAGES consensus recommendations on an annotation framework for surgical video. *Surg Endosc.* 2021;35(9):4919–30. doi:10.1007/s00464-021-08578-9
 10. Garrow CR, Kowalewski KF, Li L, Wagner M, Schmidt MW, Engelhardt S, et al. Machine learning for surgical phase recognition: a systematic review. *Ann Surg.* 2021;273(4):684–93. doi:10.1097/sla.0000000000004425
 11. Das A, Bano S, Vasconcelos F, Khan D, Marcus HJ, Stoyanov D. Reducing prediction volatility in the surgical workflow recognition of endoscopic pituitary surgery. *Int J Comput Assist Radiol Surg.* 2022;17(10):1851–60. doi:10.1007/s11548-022-02599-y
 12. Khan DZ, Luengo I, Barbarisi S, Addis C, Culshaw L, Dorward NL, et al. Automated operative workflow analysis of endoscopic pituitary surgery using machine learning: development and preclinical evaluation (IDEAL stage 0). *J Neurosurg.* 2021;136(3):862–72. doi:10.3171/2021.6.jns21923
 13. Berlet M, Vogel T, Ostler D, Czempiel T, Kähler M, Brunner S, et al. Surgical reporting for laparoscopic cholecystectomy based on phase annotation by a convolutional neural network (CNN) and the phenomenon of phase flickering: a proof of concept. *Int J Comput Assist Radiol Surg.* 2022;17(7):1257–65. doi:10.1007/s11548-022-02680-6
 14. Marcus HJ, Khan DZ, Borg A. Pituitary society expert Delphi consensus: operative workflow in endoscopic transsphenoidal pituitary adenoma resection. *Pituitary.* 2021;24(6):837–47. doi:10.1007/s11102-021-01162-3
 15. Rossum GV, Drake FLJ. Python reference manual. Amsterdam: Centrum voor Wiskunde en Informatica; 1995.
 16. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, editors. *Advances in Neural Information Processing Systems*. Red Hook (NY): Curran Associates, Inc. 2019;32:8024–35.
 17. Cook S. CUDA programming: a developer's guide to parallel computing with GPUs. 1st ed. San Francisco (CA): Morgan Kaufmann Publishers, Inc; 2012.
 18. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas (NV): IEEE; 2016. p. 770–8. doi:10.1109/cvpr.2016.90
 19. Hoggett L, Wright A, Wilson J. How to write an operation note. *BMJ.* 2017;358:j355. doi:10.1136/bmj.j355
 20. Thomson D, Baldwin M, Bellini M, Silva M. Improving the quality of operative notes for laparoscopic cholecystectomy: assessing the impact of a standardized operation note proforma. *Int J Surg.* 2016;32:118–22. doi:10.1016/j.ijsu.2016.01.037