Flakë Bajraktari*, Lina Hauser, and Peter P. Pott

# Adaptive Ensemble Learning for Robust Surgical Phase Recognition

https://doi.org/10.1515/cdbme-2025-0232

**Abstract:** Automatic recognition of surgical phases plays a critical role in enabling intelligent, context-aware support systems during operative procedures. The inherent variability of surgical techniques and intraoperative conditions makes precise surgical phase recognition (SPR) a challenging task. This study explores ensemble learning as a strategy to improve phase recognition performance using a learnable fusion approach. A selection of state-of-the-art deep learning models was trained and tuned to capture complementary aspects of the task. This resulted in 14 base models with varied backbones and parameter settings. To aggregate the output probabilities of the base models, a lightweight fully connected network, referred to as StackingNet, was designed as a meta-model capable of learning to generate final predictions from their outputs. This approach outperformed the best individual base model within the respective ensemble in 14 out of 15 ensemble configurations, achieving a maximum F1-score improvement of 3.3 %. These results demonstrate that learnable ensemble fusion can significantly enhance accuracy of surgical phase recognition, highlighting its potential in the development of intelligent surgical assistance systems.

**Keywords:** Surgical Phase Recognition, Ensemble Learning, Deep Learning.

## 1 Introduction

The integration of artificial intelligence (AI) into surgical environments is reshaping the landscape of intraoperative assisting systems. Among various AI-driven applications, surgical phase recognition (SPR) has emerged as a critical component for enabling real-time guidance, automating procedural documentation, and enhancing overall workflow efficiency, ultimately leading to improved patient safety [1–3].

Deep learning has enabled significant progress in SPR, supported by benchmark datasets such as Cholec80 [3], which contains 80 laparoscopic cholecystectomy videos, each segmented into seven annotated phases. Standard approaches typically involve the use of convolutional neural networks (CNNs), such as ResNet [4], to extract spatial features from surgical video frames, followed by temporal models to capture phase transitions over time [3, 5, 6]. Architectures such as Recurrent Neural Networks (RNNs) [7, 8], Temporal Convolutional Networks (TCNs) [9], and Transformers [5] have all contributed to advances in this area. Despite these developments, SPR systems still face limitations. Small, imbalanced datasets and variability in surgical technique and environments hinder generalizability [10].

To adress these challenges, multi-view approaches have been explored that integrate complementary visual modalities such as laparoscopic and in-room camera data [11]. Beyond multimodal setups, ensemble learning has gained traction in various medical image analysis tasks, including cancer classification [12] and surgical tool detection [13, 14], where combining diverse models further improves accuracy and generalizability. However, these approaches rely on statistical fusion techniques, such as averaging, to aggregate base model outputs into final predictions. These methods fail to capture complex, non-linear relationships – a limitation that becomes evident, for example, in tasks with high transitional ambiguity, where visual features may appear unclear near phase boundaries.

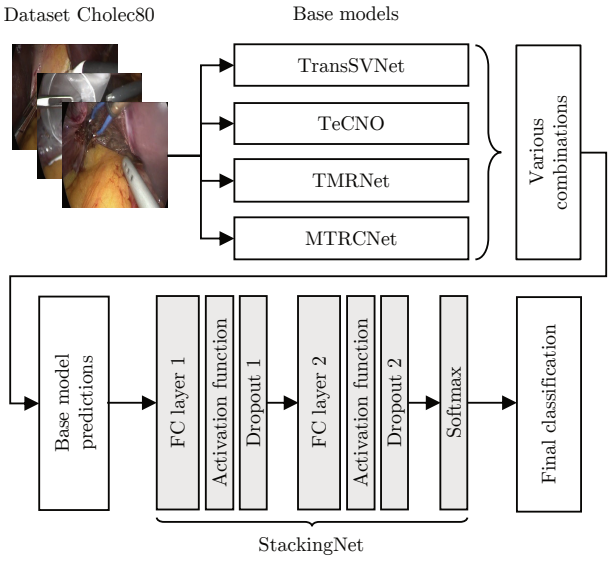


**Fig. 1:** Schematic overview of the proposed method, illustrating the *StackingNet* architecture as a learnable meta-model for fusing base model outputs into final predictions.

***Corresponding author: Flakë Bajraktari,** University of Stuttgart, Institute of Medical Device Technology, Pfaffenwaldring 9, Stuttgart, Germany, e-mail: flake.bajraktari@imt.uni-stuttgart.de
**Lina Hauser, Peter P. Pott,** University of Stuttgart, Institute of Medical Device Technology, Pfaffenwaldring 9, Stuttgart, Germany

While ensemble learning has been successfully applied in other areas of medical image analysis, its potential in SPR remains largely unexplored. This work addresses that gap by proposing an ensemble framework tailored for SPR. A fully connected artificial neural network is introduced as a learnable meta-model to fuse the outputs of multiple deep learning models across different ensemble combinations. The goal is to evaluate whether such a learnable fusion strategy can enhance SPR performance beyond what individual models can achieve.

# 2 Materials and Methods

A schematic overview of the proposed method is shown in Figure 1. This study is based on the Cholec80 dataset [3], a widely adopted benchmark for SPR. The videos were downsampled to 1 fps, and all frames were resized to $250 \times 250$ pixels. Additionally, standard data augmentation techniques were applied, including cropping, horizontal flipping, and color jittering. The dataset was split into 32 training videos, 8 validation videos, and 40 test videos. Training was performed on a workstation equipped with an Intel® Core™ i9-12900K processor and an NVIDIA GeForce RTX 3080 Ti GPU running Ubuntu 20.04.01. Model development was carried out using *PyTorch*, and hyperparameter tuning was managed using *Weights & Biases* [15]. Performance was evaluated using F1-score without applying tolerance windows around phase boundaries.

## 2.1 Base Models and Ensemble Design

The ensemble was built using four state-of-the-art temporal deep learning architectures for SPR: a Temporal Convolutional Network (*TeCNO* [9]), a Transformer-based model (*Trans-SVNet* [5]), and two LSTM-based models (*MTRCNet* [7] and *TMRNet* [8]). Where publicly available pre-trained models were not compatible with the 32:8:40 data split, models were retrained from scratch. In all cases, the epoch achieving the highest validation accuracy was selected for evaluation. To ensure statistical robustness, each model was trained in three independent runs, and the mean performance was reported. To enhance ensemble diversity, additional model variants were created by modifying architectural parameters such as the number of TCN layers, feature maps and the choice of feature extractor backbone – specifically *ResNet50* [4] and *ResNeSt50* [16]. This resulted in a pool of 14 distinct base models, summarized in Table 1.

To systematically assess the impact of ensemble learning on SPR, multiple configurations were defined to target specific research questions regarding model diversity. These configurations are listed below, with the corresponding model identifiers and architectures detailed in Table 1.

- **Architecture-based (A1–A4):** Ensembles composed of model variants sharing the same architecture.
  **A1**: 1–4 (*Trans-SVNet*),
  **A2**: 5–10 (*TeCNO*),
  **A3**: 11–12 (*TMRNet*),
  **A4**: 13–14 (*MTRCNet*).
- **Backbone-based (B1–B2):** Ensembles grouped by feature extractor backbone.
  **B1**: 1, 2, 5–7, 11, 13 (*ResNet50*),
  **B2**: 3, 4, 8–10, 12, 14 (*ResNeSt50*).
- **Performance-based (C1–C4):** Ensembles of top-ranked models based on F1-score.
  **C1**: 2, 6, 12, 14 (best from each architecture),
  **C2**: 2, 6, 11, 13 (best *ResNet50* models),
  **C3**: 3, 9, 12, 14 (best *ResNeSt50* models),
  **C4**: 2, 3, 6, 9, 11–14 (best models from each architecture and each backbone).
- **Architectural diversity (D1–D4):** Homogeneous (**D2**) vs. heterogeneous ensembles (**D1, D3, D4**).
  **D1**: 1–10 (*Trans-SVNet* + *TeCNO*),
  **D2**: 11–14 (*TMRNet* + *MTRCNet*),
  **D3**: 1–4, 11–12 (*Trans-SVNet* + *TMRNet*)
  **D4**: 5–10, 13–14 (*TeCNO* + *MTRCNet*).
- **Full ensemble (E):** 1–14 (all base models).

For each ensemble configuration, the best-performing base model in the respective ensemble served as a reference, and the improvement achieved by the ensemble was quantified relative to this baseline.

## 2.2 Meta-model StackingNet

Based on previous work by Yildiz et al. [17], Su et al. [18], and Cao et al. [19], a fully connected (FC) neural network was implemented as the meta-model in this study. The architecture, referred to as *StackingNet*, consists of an input layer, two hidden layers and an output layer for classification of seven classes, which is shown in Figure 1.

The input layer maps the ensemble feature vector to the first hidden layer. Each hidden layer applies a linear transformation, followed by a non-linear activation and dropout to reduce overfitting. The hidden layers use a fixed number of neurons, defined by the *hidden units* parameter, with configurable dropout rates. The output layer produces class scores, which are passed through a softmax function to yield probability distributions over seven classes. Hyperparameters were optimized for each ensemble based on commonly reported value ranges in the literature.

# 3 Results

The individual base models achieved F1-scores ranging from 79.40 % to 83.59 %. Table 1 presents the performance of all 14 models, along with their respective architectural configurations. The highest individual F1-score was achieved by *TMRNet* with a *ResNeSt50* backbone.

**Tab. 1:** Individual performance of each base model along with their respective architectural configurations.

| Nr. | Model | Backbone | Layer | Maps | F1-score |
|---|---|---|---|---|---|
| 1 | TransSVNet | ResNet50 | 8 | 32 | $80.91 \pm 7.24$ |
| 2 | | | 9 | 64 | $83.06 \pm 7.33$ |
| 3 | | ResNeSt50 | 8 | 32 | $82.83 \pm 7.70$ |
| 4 | | | 9 | 64 | $82.78 \pm 7.79$ |
| 5 | TeCNO | ResNet50 | 8 | 32 | $79.40 \pm 7.34$ |
| 6 | | | 9 | 64 | $82.37 \pm 7.05$ |
| 7 | | | 10 | 64 | $81.26 \pm 6.70$ |
| 8 | | ResNeSt50 | 8 | 32 | $79.40 \pm 9.47$ |
| 9 | | | 9 | 64 | $82.31 \pm 7.71$ |
| 10 | | | 10 | 64 | $82.21 \pm 7.62$ |
| 11 | TMRNet | ResNet50 | - | - | $82.25 \pm 6.37$ |
| 12 | | ResNeSt50 | - | - | $83.59 \pm 5.61$ |
| 13 | MTRCNet | ResNet50 | - | - | $78.17 \pm 8.35$ |
| 14 | | ResNeSt50 | - | - | $79.23 \pm 7.55$ |

For each of the 15 ensembles (A–E), a dedicated hyperparameter sweep was conducted to identify the optimal configuration. For every ensemble, the run with the highest validation accuracy was selected. The corresponding optimal hyperparameters are summarized in Table 2. Using these optimized settings, the ensembles were subsequently trained.

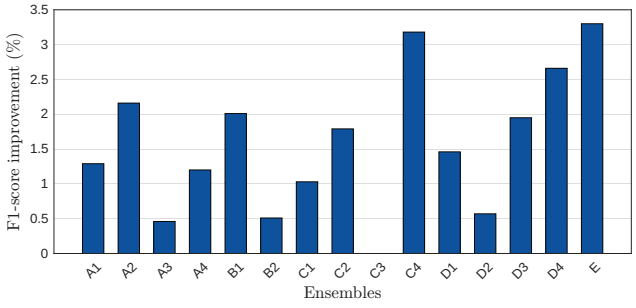


**Fig. 2:** F1-score improvement of each ensemble using *StackingNet* as meta-model.

The evaluation of the ensembles is presented in Figure 2 as F1-score improvements in % compared to the best-performing base model within the respective ensemble. All ensembles demonstrated an improvement in F1-score, with the exception of combination C3, which showed a decrease of 0.85 %. The highest gains were observed in combinations C4 and E with 3.18 % and 3.3 % respectively.

# 4 Discussion and Outlook

The evaluation of ensemble configurations reveals key insights into the influence of architectural diversity and model selection on overall performance. Nearly all ensembles achieved performance gains over their strongest base model, with the most notable improvements observed in configurations that combined models across different architectures and backbones.

Architecture-based ensembles (A1–A4) generally benefited from stacking, although smaller ensembles such as A3 and A4, each composed of only two models, showed more modest improvements. In the backbone-based group, B1 with *ResNet50* achieved substantially higher gains than B2 with *ResNeSt50*. This suggests that not only the architecture itself but also the synergy between backbone and task-specific design influences ensemble effectiveness. In the performance-based group (C1–C4), notable improvements were observed in configurations combining top-performing models from all architectures and backbones. In contrast, C3 showed a slight performance drop despite including strong individual models, suggesting that diversity in model behavior is as critical as individual performance. The architectural diversity group (D1–D4) supports this: heterogeneous ensembles (D1, D3, D4) outperformed the homogeneous configuration D2, highlighting the advantage of mixing architectures.

Hyperparameter optimization revealed that optimal *StackingNet* settings varied considerably across ensembles – especially in learning rate, batch size, and dropout rates. Larger ensembles often required stronger regularization and more training epochs, pointing to a greater risk of overfitting. The variation in activation functions and optimizers further reflects how ensemble composition influenced the learning dynamics of the meta-model.

These findings suggest that ensemble performance depends not only on base model quality, but also on combination strategies, diversity, and meta-model tuning. Balancing strong models with architectural variety and including a broader set of base models proved beneficial, even when some models performed weaker individually. This implies that the meta-model leverages each model's unique contribution, and that larger ensembles can improve robustness and generalizability. Future work could explore more advanced meta-model and broader evaluations across datasets and clinical contexts.

**Tab. 2:** Optimal hyperparameters identified for the different ensembles using *StackingNet*.

| Hyperparameter | Ensemble | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **A1** | **A2** | **A3** | **A4** | **B1** | **B2** | **C1** | **C2** | **C3** | **C4** | **D1** | **D2** | **D3** | **D4** | **E** |
| $acc_{val}$ | 92.69 | 92.51 | 98.75 | 90.71 | 93.44 | 96.56 | 98.11 | 95.29 | 97.49 | 98.04 | 92.86 | 98.58 | 98.14 | 93.20 | 97.97 |
| *Batch size* | 512 | 256 | 128 | 512 | 128 | 64 | 256 | 8 | 256 | 1024 | 16 | 512 | 16 | 64 | 64 |
| *Learning rate* | 2.5e-5 | 2.5e-5 | 1e-4 | 7.5e-5 | 2.5e-5 | 7.5e-6 | 5e-6 | 5e-5 | 5e-6 | 7.5e-6 | 2.5e-4 | 7.5e-5 | 5e-6 | 1e-5 | 7.5e-3 |
| *Epochs* | 36 | 12 | 23 | 10 | 8 | 7 | 91 | 100 | 96 | 74 | 72 | 8 | 6 | 10 | 2 |
| *Dropout 1* | 0.6 | 0.2 | 0.1 | 0.9 | 0.7 | 0.3 | 0.8 | 0.7 | 0 | 0.9 | 0.7 | 0.2 | 0.7 | 0.9 | 0.5 |
| *Dropout 2* | 0.3 | 0.7 | 0.9 | 0.1 | 0 | 0.6 | 0.7 | 0.8 | 0.2 | 0.8 | 0 | 0.6 | 0.8 | 0.5 | 0 |
| *Hidden units* | 256 | 256 | 512 | 256 | 128 | 256 | 256 | 256 | 64 | 256 | 64 | 512 | 128 | 512 | 256 |
| *Optimizer* | adam | adam | adam | adam | adam | rmsprop | adam | sgd | rmsprop | adam | sgd | adam | adam | adam | sgd |
| *Activation function* | relu | tanh | relu | tanh | tanh | tanh | relu | tanh | tanh | tanh | tanh | tanh | tanh | relu | tanh |

**Author Statement**

Research funding: The author state no funding involved. Conflict of interest: Authors state no conflict of interest.

# References

[1] 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. International Journal of Computer Assisted Radiology and Surgery 2022;17:1991–1999.

[2] Huaulmé A, Jannin P, Reche F, Faucheron JL, Moreau-Gaudry A, Voros S. Offline identification of surgical deviations in laparoscopic rectopexy. Artificial Intelligence in Medicine 2020;104:101837.

[3] Twinanda A, Shehata S, Mutter D, Marescaux J, De Mathelin M, Padoy N. EndoNet: a deep architecture for recognition tasks on laparoscopic videos. IEEE Transactions on Medical Imaging 2016;36.

[4] 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). 2016 770–778.

[5] Gao X, Jin Y, Long Y, Dou Q, Heng PA. Trans-SVNet: Accurate Phase Recognition from Surgical Videos via Hybrid Embedding Aggregation Transformer. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. 2021 593–603.

[6] Zhang B, Abbing J, Ghanem A, Fer D, Barker J, Abukhalil R, et al. Towards accurate surgical workflow recognition with convolutional networks and transformers. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization 2021;10:1–8.

[7] Jin Y, Li H, Dou Q, Chen H, Qin J, Fu CW, et al. Multi-task recurrent convolutional network with correlation loss for surgical video analysis. Medical Image Analysis 2020;59:101572.

[8] Jin Y, Long Y, Chen C, Zhao Z, Dou Q, Heng PA. Temporal Memory Relation Network for Workflow Recognition From Surgical Video. IEEE Transactions on Medical Imaging 2021; 40:1911–1923.

[9] Czempiel T, Paschali M, Keicher M, Simson W, Feussner H, Kim ST, et al. TeCNO: Surgical Phase Recognition with Multi-stage Temporal Convolutional Networks. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, volume 12263, 343–352. 2020.

[10] Kirtac K, Aydin N, Lavanchy JL, Beldi G, Smit M, Woods MS, et al. Surgical Phase Recognition: From Public Datasets to Real-World Data. Applied Sciences 2022;12:8746.

[11] Bajraktari F, Pott PP. Multi-view surgical phase recognition during laparoscopic cholecystectomy. Current Directions in Biomedical Engineering 2024;10:45–48.

[12] Mohammed M, Mwambi H, Mboya IB, Elbashir MK, Omolo B. A stacking ensemble deep learning approach to cancer type classification based on TCGA data. Scientific Reports 2021; 11:15626.

[13] Wang S, Raju A, Huang J. Deep learning based multi-label classification for surgical tool presence detection in laparoscopic videos. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). 2017 620–623.

[14] Jaafari J, Douzi S, Douzi K, Hssina B. The impact of ensemble learning on surgical tools classification during laparoscopic cholecystectomy. Journal of Big Data 2022;9:49.

[15] Biewald L. Experiment tracking with weights and biases 2020. https://www.wandb.com/. [Online] (Accessed: 2025-02-23).

[16] Zhang H, Wu C, Zhang Z, Zhu Y, Lin H, Zhang Z, et al. ResNeSt: Split-Attention Networks 2020. [Online] (Accessed: 2025-02-22).

[17] Yildiz G, Ulu A, Dızdaroğlu B, Yildiz D. Hybrid Image Improving and CNN (HIICNN) Stacking Ensemble Method for Traffic Sign Recognition. IEEE Access 2023;11:69536–69552.

[18] Su Q, Wang F, Chen D, Chen G, Li C, Wei L. Deep convolutional neural networks with ensemble learning and transfer learning for automated detection of gastrointestinal diseases. Computers in Biology and Medicine 2022;150:106054.

[19] Cao Z, Pan X, Yang Y, Huang Y, Shen HB. The lncLocator: a subcellular localization predictor for long non-coding RNAs based on a stacked ensemble classifier. Bioinformatics 2018; 34:2185–2194.