Deep Multimodal Fusion for Surgical Feedback Classification

Rafal Kocielnik RAFALKO@CALTECH.EDU
California Institute of Technology, USA

Elyssa Y. Wong and **Timothy N. Chu** EYWONG@USC.EDU, TNCHU@USC.EDU
University of Southern California, USA

Lydia Lin LJLIN@USC.EDU
University of Southern California & California Institute of Technology, USA

De-An Huang DEAHUANG@NVIDIA.COM
NVIDIA, USA

Jiayun Wang and **Anima Anandkumar** PETERW@CALTECH.EDU, ANIMA@CALTECH.EDU
California Institute of Technology, USA

Andrew J. Hung ANDREW.HUNG@CSHS.ORG
Cedars-Sinai Medical Center, USA

Abstract

Quantification of real-time informal feedback delivered by an experienced surgeon to a trainee during surgery is important for skill improvements in surgical training. Such feedback in the live operating room is inherently multimodal, consisting of verbal conversations (e.g., questions and answers) as well as non-verbal elements (e.g., through visual cues like pointing to anatomic elements). In this work, we leverage a clinically-validated five-category classification of surgical feedback: *"Anatomic"*, *"Technical"*, *"Procedural"*, *"Praise"* and *"Visual Aid"*. We then develop a multi-label machine learning model to classify these five categories of surgical feedback from inputs of text, audio, and video modalities. The ultimate goal of our work is to help automate the annotation of real-time contextual surgical feedback at scale. Our automated classification of surgical feedback achieves AUCs ranging from 71.5 to 77.6 with the fusion improving performance by 3.1%. We also show that high-quality manual transcriptions of feedback audio from experts improve AUCs to between 76.5 and 96.2, which demonstrates a clear path toward future improvements. Empirically, we find that the *Staged* training strategy, with first pre-training each modality separately and then training them jointly, is more effective than training different modalities altogether. We also present intuitive findings on the importance of modalities for different feedback categories. This work offers an important first look at the feasibility of automated classification of real-world live surgical feedback based on text, audio, and video modalities.

Keywords: Surgical feedback, Multimodality, Robot-Assisted Surgery, Deep Learning

1. Introduction

Importance: Real-time informal verbal feedback in surgical settings is pivotal not just for immediate correction and guidance but also for long-term proficiency and mastery (Agha et al., 2015). The quality of such feedback has been demonstrated to significantly influence intraoperative performance (Bonrath et al., 2015), profoundly impact surgical skill acquisition (Ma et al., 2022) as well as trainee's sense of autonomy (Haglund et al., 2021). It also has broader implications for the overall surgical training paradigm. Despite the inherent challenges posed by the unstructured and personalized nature of surgical feedback, it's undeniable that a systematic approach to understanding it is the linchpin to refining and enhancing surgical training.

Challenges: However, quantifying and conducting a systematic analysis of the properties of real-world surgical feedback presents notable challenges. We, therefore, adopt a recent clinically validated classification system for surgical feedback that has been shown to offer high reliability and generalizability as

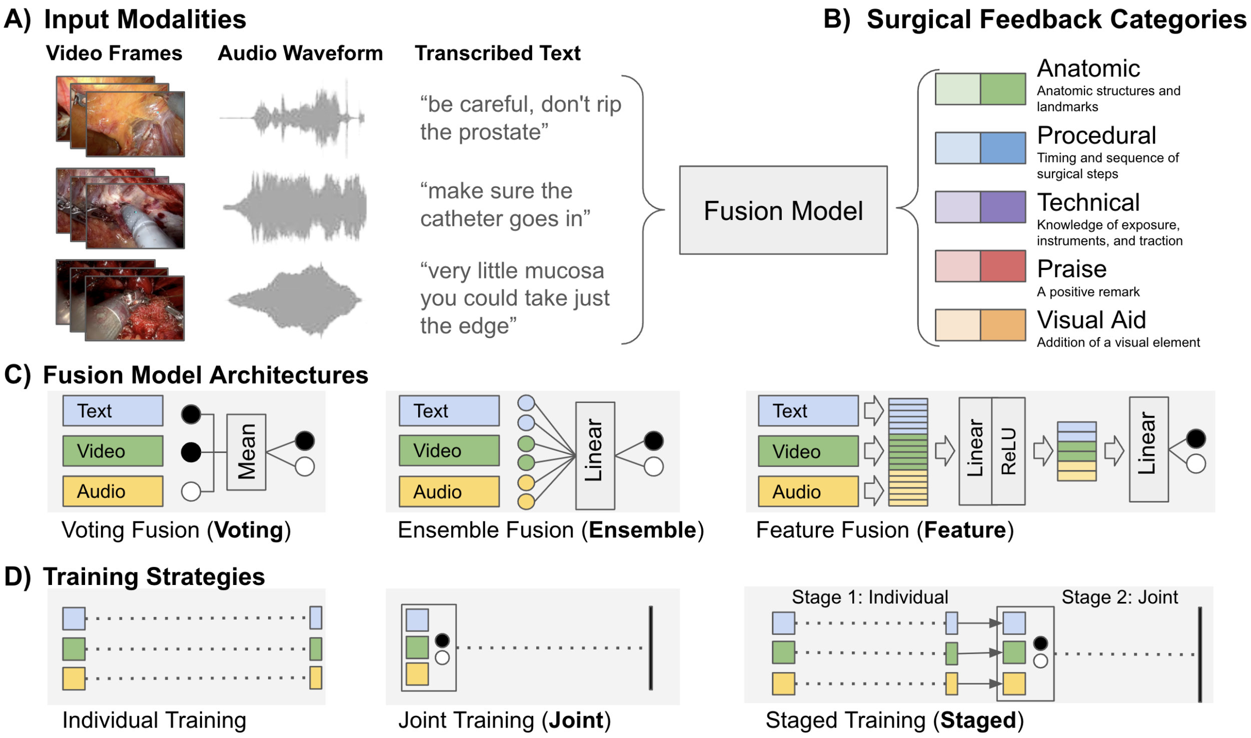

A) Input Modalities
Video Frames
Audio Waveform
Transcribed Text
"be careful, don't rip the prostate"
"make sure the catheter goes in"
"very little mucosa you could take just the edge"
Fusion Model
B) Surgical Feedback Categories
Anatomic
Anatomic structures and landmarks
Procedural
Timing and sequence of surgical steps
Technical
Knowledge of exposure, instruments, and traction
Praise
A positive remark
Visual Aid
Addition of a visual element
C) Fusion Model Architectures
Text
Video
Audio
Mean
Linear
ReLU
Voting Fusion (Voting)
Ensemble Fusion (Ensemble)
Feature Fusion (Feature)
D) Training Strategies
Stage 1: Individual
Stage 2: Joint
Individual Training
Joint Training (Joint)
Staged Training (Staged)

Figure 1: Overview of the work. Multimodal inputs consist of text, audio, and video (**A**) and 5 binary multi-label classification outputs adapted from a clinically validated framework introduced in Wong et al. (2023) (**B**). We explore model architectures (**C**) and training strategies (**D**) for improving the performance of surgical feedback classification using multimodal fusion.

well as practical utility (Wong et al., 2023). However, their system requires manual annotations of surgical feedback, which is time and resource-demanding. This is primarily due to the necessity for expertise in comprehending both the surgical context and the feedback's intent (Agha et al., 2015). Furthermore, feedback delivery in the live operating room is inherently multimodal and adds to the complexity. The delivery encompasses verbal conversations, non-verbal appraisals, and visual cues.

Approach: We explore automated intraoperative surgical feedback classification with machine learning techniques in this pilot study. Specifically, we leverage multi-modal inputs composed of text, audio, and video (Fig. 1-A) in order to perform binary multi-label classification of surgical feedback into 5 components (Fig. 1-B). In our experiments we systematically vary 2 dimensions: 1) complexity of the fusion model architecture (Fig. 1-C) and 2) training strategy (Fig. 1-D). We arrive at an optimal *Staged Fusion* approach which starts with independent training of each modality and continues with training modalities jointly. This approach helps mitigate the dominance of one modality that can suppress extracting information from other modalities.

Findings: We summarize our findings as follows:

- We achieve Areas under the ROC Curve (AUCs) varying from 71.5 to 77.6 with automated surgical feedback classification (Table 3).
- We further show that manual transcription of specialized surgical feedback by experts, though costly, further improves AUCs to between 76.5 and 96.2, indicating a path to further improvements.
- We find that the training process is more important for fusion effectiveness (3.1% gain) than model architecture (1.1%) in ablation studies.
- We confirm our intuition that video modality is most important for the classification of *"Visual Aid"* feedback, while emotion extracted from audio is important for the detection of *"Praise"*.

Feedback	Description
Anatomic	Familiarity with anatomic structures and landmarks.
Procedural	Pertains to timing and sequence of surgical steps.
Technical	Performance of discreet task with appropriate knowledge of exposure, instruments, traction, etc.
Praise	A complementary remark
Visual Aid	Addition of visual element to direct trainee's attention or focus

Table 1: Categories of surgical feedback adapted from recent clinically validated classification system introduced in Wong et al. (2023)

.

Contributions: Our main contributions include:

- To the best of our knowledge, we are the first to explore the feasibility of the automated classification of live surgical feedback.
- We systematically explore model architectures and training strategies for multi-modal fusion in a novel context of real-world live surgical feedback. The emphasis on training strategy distinguishes our approach as significantly novel, given that most prior work focused on exploring model architectures.

2. Background and Related Work

Feedback in Robot-Assisted Surgery . Wong et al. (2023) first report on the development of a manual classification system for verbal feedback during robot-assisted surgery. This work also demonstrates the reliability, generalizability, and utility of this manual classification system. It specifically shows that using the proposed feedback categorization it is possible to detect significant differences in feedback type frequency and subsequent trainee reactions based on surgeon experience level and the surgical task being performed. For example, technical feedback with a visual component was associated with an increased rate of trainee behavioral change or verbal acknowledgment responses. Hence we adopt this classification system as it offers a tangible link between feedback and subsequent trainee behavior.

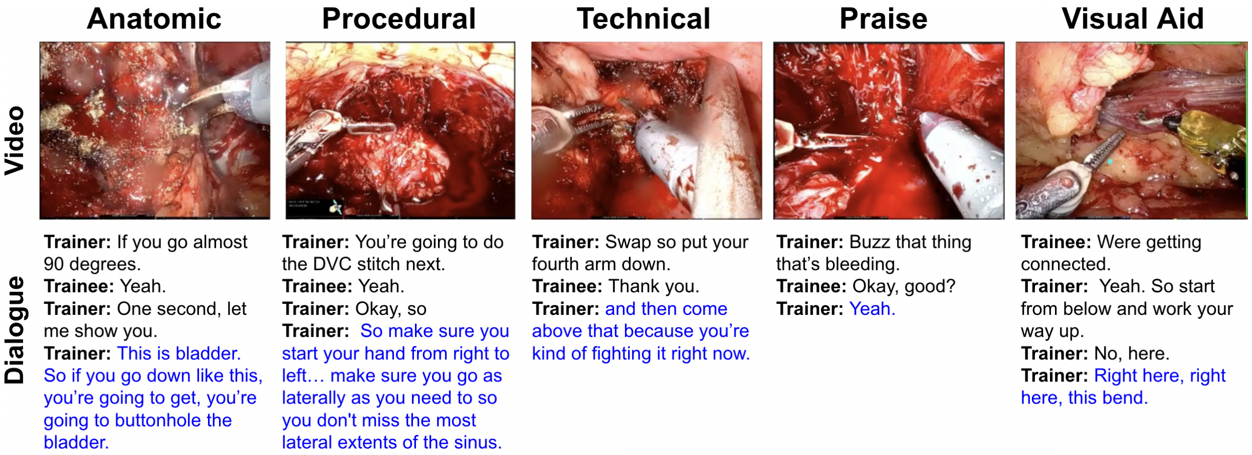
To the best of our knowledge, there exists no prior work on automated surgical feedback classification. Our work pioneers predicting real-time verbal feedback for robotic-assisted surgery with multi-modal sensory inputs.

Deep Learning for Multi-Modality Data . Prior work mostly focused on fusing visual modalities but not the importance of training strategies. Boulahia et al. (2021) explore *early*, *intermediate*, and *late* fusion for general activity recognition. Their method focuses on visual channels and is not directly applicable to surgical feedback which includes text and audio modalities. We borrow the late fusion concept from their work, but expand on aspects of model complexity and training strategy. Li et al. (2020) align text and image modalities for image captioning task. This fusion approach aims to generate output in one modality based on input from other modalities, which is substantially different than our task. Walsman et al. (2019) focus on the fusion of visual channels for the scene and goal representation in robotic vision. Their work applies fusion in 3D simulated setting with clear and distinct objects, which are not present in our context. In the medical domain, Narazani et al. (2022) explore fusion for PET and MRI visual modalities. Their work, again focuses on visual channels only and reports no gains from the proposed fusion approaches. In contrast, our research systematically investigates a range of multi-modal fusion techniques and training strategies.

3. Methods

3.1. Data Acquisition

We used a dataset of real-life feedback delivered by trainers to trainees during live robot-assisted surgery cases from Wong et al. (2023). Trainers were defined as those providing feedback and trainees were those receiving feedback while actively operating on the surgeon console. This feedback has been recorded using wireless microphones worn by the surgeons and video capturing the surgeon's point-of-view (i.e., endoscope camera view). Video and audio were recorded synchronously with an external recorder. All surgeries were performed using da Vinci Xi surgical robotic system (DiMaio et al., 2011). The feedback instances were timestamped and manually tran-



Anatomic — **Trainer:** If you go almost 90 degrees. **Trainee:** Yeah. **Trainer:** One second, let me show you. **Trainer:** This is bladder. So if you go down like this, you're going to get, you're going to buttonhole the bladder.

Procedural — **Trainer:** You're going to do the DVC stitch next. **Trainee:** Yeah. **Trainer:** Okay, so **Trainer:** So make sure you start your hand from right to left… make sure you go as laterally as you need to so you don't miss the most lateral extents of the sinus.

Technical — **Trainer:** Swap so put your fourth arm down. **Trainee:** Thank you. **Trainer:** and then come above that because you're kind of fighting it right now.

Praise — **Trainer:** Buzz that thing that's bleeding. **Trainee:** Okay, good? **Trainer:** Yeah.

Visual Aid — **Trainee:** Were getting connected. **Trainer:** Yeah. So start from below and work your way up. **Trainer:** No, here. **Trainer:** Right here, right here, this bend.

Figure 2: Examples of video (frame) along with the dialogue between trainee (feedback recipient) and attending surgeon (feedback provider) from different surgical cases in our dataset.

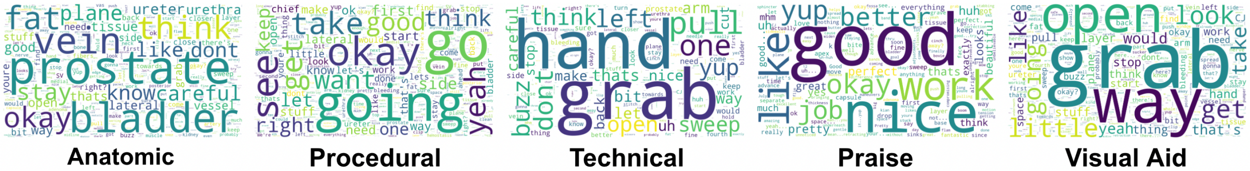


Figure 3: Most frequent words used in the delivery of each type of feedback visualized via word clouds. The larger the word, the more frequently it has been used in this category of feedback. For example, *Anatomic* feedback includes words related to physical structures like "prostate" and "bladder", while *Technical* feedback includes words describing the use of instruments like "grab" and "hand".

Component	Count	Count/Case	Word count
Anatomic	1104	35.6 ± 23.0	11.0 ± 7.7
Procedural	817	26.4 ± 14.9	9.8 ± 8.0
Technical	3223	104.0 ± 67.7	8.1 ± 6.8
Praise	262	9.0 ± 8.4	3.6 ± 3.7
Visual Aid	303	11.7 ± 10.2	10.5 ± 6.7
Any feedback	3912	126.2 ± 72.0	8.1 ± 6.8

Table 2: Statistics per surgical feedback category including total instances, instances per individual surgical case as well as mean word count of transcribed feedback text. **Any feedback** refers to feedback of any type. One feedback might have multiple labels.

scribed from audio recordings. Feedback instance has been defined as trainers' utterances meant to alter or approve trainee behavior. The dataset contains 3912 individual instances of feedback as shown in Table 2.

3.2. Surgical Feedback Categorization

Two medical students were involved in feedback identification and transcription. Manual transcription included only utterances from the attending surgeon providing feedback. Any utterances by trainees or unrelated conversations were not transcribed.

This feedback has been categorized using surgical feedback quantification framework introduced by Wong et al. (2023). This categorization scheme has been shown to offer high reliability and generalizability as well as practical utility in the clinical setting. The five feedback dimensions from this framework along with their definitions are presented in Table 1. The categories are non-exclusive. Further details of the annotation can be found in Wong et al. (2023).

Examples of aligned video frames and audio transcriptions are shown in Fig. 2. Dialogue is very important in feedback categorization, whereas video offers supplementary sources, but similar feedback can be delivered in different visual contexts. Fig. 3 shows the most frequent words for each feedback category as word clouds where the larger the word, the more frequently it appears in the underlying feedback instances. *Anatomic* type of feedback most frequently includes words related to physical structures such as “prostate”, “bladder”, and “vein”. At the same time, *Technical* feedback frequently includes words such as “grab” and “hand”, “pull’ referring to the use of instruments.

3.3. Speaker Diarization and Automated Speech Recognition

In addition to manual transcription, we performed *Automated Speech Recognition (ASR)* using pre-trained Whisper medium model introduced in Radford et al. (2022). This model was pre-trained on 680k hours of labeled English-only speech data specifically for speech recognition. Speech data was annotated using large-scale weak supervision. Given, the interactive dialogue-like structure of the exchanges around and leading to feedback (see Fig. 2), we further applied speaker diarization, the concept of partitioning speech from different speakers in a single audio clip, using *Pyannote* (Bredin and Laurent, 2021; Bredin et al., 2020). This was done to provide more context about feedback such as the speaker and the conversations before and after the feedback delivery. Speaker diarization was paired with the ASR to transcribe each separate segment of audio.

3.4. Individual-Modality-Input Models

We leverage pre-trained transformer models to extract information from each individual modality.

Text: We fine-tune *BERT* base model with 110M parameters introduced by Devlin et al. (2018). The model has been pre-trained on general-knowledge text including BooksCorpus and English Wikipedia. We also experiment with specialized text models pre-trained on biomedical datasets including *BioBert* (Lee et al., 2020) and *BioClinicalBert* (Alsentzer et al., 2019). However, no noticeable improvement in performance has been observed, which we attribute to the relatively casual and conversational nature of the feedback with only occasional use of specialized vocabulary.

Audio: We fine-tune *Wave2Vec* base model with 95M parameters introduced by Baevski et al. (2020). We specifically use a model pre-trained on emotion recognition tasks from *“SUPERB”* dataset (Yang et al., 2021). This model extracts features related to the emotion in the delivery of feedback from audio and is different than text transcription.

Video: We fine-tune *VideoMAE* base model with 86M trainable parameters introduced by Tong et al. (2022). This model is an extension of Masked Auto Encoders introduced by He et al. (2022) from images to video. We use a model pre-trained on Kinetics-400 dataset (Kay et al., 2017) containing video clips of 400 human action classes.

3.5. Model Architectures of Multi-Modality Fusion

We explore different variants of late fusion (Fig. 1-C) varying the model complexity from a simple majority vote to feature fusion with additional layers.

Voting Fusion (Best Voting): In this architecture, each modality model predicts the label independently (e.g., whether feedback component is *“Anatomic”* or *“Non-anatomic”* based on video only). The prediction given by the majority of models (i.e., at least 2 out of 3 models), is used as the final label for the fusion model. We further explore voting fusion via max of model predictions (i.e., at least 1 model predicts a positive label). We report the best of these voting approaches in our results.

Ensemble Fusion (Ensemble): In this architecture, each model returns a size 2 vector representation of the modality. These reduced representations are combined via a linear 6x2 layer which weights each modality and returns the probability of each class (e.g., *“Anatomic”* or *“Non-Anatomic”*) as the final fusion output. Compared to Best Voting approach, the Ensemble architecture can learn the optimal weighting for combining the representations from each individual modality.

Feature Fusion (Feature): In this architecture, we extract much richer representations from each modality in the form of 256-dimension vector. This can help capture more detailed information, but may also add complexity to the learning process. The representations are concatenated into one 756-dim vector and passed via 2 fully-connected linear layers that reduce the dimensions to 96 and finally 2 in a funnel fashion. This sequential architecture is augmented with ReLu activation and additional dropout in between. The additional steps can help the model calculate intermediate fusion features.

Model	Anatomic	Procedural	Technical	Praise	Vis. Aid	Mean %
Text (Manual)[1]	$81.5_{3.3}$	$69.3_{3.6}$	$74.3_{1.9}$	$95.2_{2.4}$	$78.4_{3.1}$	
Text (ASR)[2]	$70.3_{3.2}$	$65.7_{4.7}$	$66.5_{4.0}$	$76.2_{8.5}{}^{\dagger}$	$66.7_{6.8}$	
Audio (Emotion)	$67.3_{0.3}$	$61.8_{2.3}$	$67.2_{2.8}$	$67.3_{6.2}{}^{\dagger}$	$61.2_{5.5}$	
Video	$65.7_{2.1}$	$64.0_{2.8}$	$66.0_{0.5}$	$57.0_{2.2}$	$73.0_{6.4}{}^{\ddagger}$	
[1]Fusion Using Manual Transcription						
Best Voting	$79.7_{2.0}$ ↓2.2%	$72.0_{2.2}$ ↑3.8%	$74.2_{5.0}$ ↓0.2%	$76.9_{4.3}$ ↓19.3%	$78.4_{1.3}$ ↓0.0%	↓3.6%
Joint-Ensemble	$81.7_{3.3}$ ↑0.2%	$72.3^{*}_{0.8}$ ↑4.3%	$74.7_{4.4}$ ↑0.4%	$95.5_{1.1}$ ↑0.3%	$82.2^{*}_{1.7}$ ↑4.9%	↑2.0%
Staged-Ensemble	$\underline{86.0^{*}_{2.6}}$ ↑5.5%	$\underline{76.5^{*}_{2.3}}$ ↑10.3%	$78.8^{*}_{3.8}$ ↑6.1%	$\underline{96.2^{*}_{1.9}}$ ↑1.0%	$\underline{86.1^{*}_{1.4}}$ ↑9.8%	↑6.5%
Joint-Feature	$81.8_{1.5}$ ↑0.4%	$72.2_{5.6}$ ↑4.1%	$76.2^{*}_{0.8}$ ↑2.5%	$95.5_{1.5}$ ↑0.3%	$80.6_{2.5}$ ↑2.8%	↑2.0%
Staged-Feature	$\underline{86.0^{*}_{1.8}}$ ↑5.5%	$76.3^{*}_{2.8}$ ↑10.1%	$\underline{80.3^{*}_{4.9}}$ ↑8.1%	$95.9_{1.0}$ ↑0.7%	$85.8^{*}_{1.7}$ ↑9.4%	↑6.8%
[2]Fusion Using Automated Transcription (ASR) and Speaker Diarization						
Best Voting	$69.2_{0.3}$ ↓1.7%	$63.8_{1.9}$ ↓2.8%	$68.5_{2.7}$ ↑1.2%	$70.5_{3.4}$ ↓7.5%	$70.5_{3.6}$ ↓3.4%	↓2.8%
Joint-Ensemble	$70.5_{0.9}$ ↑0.2%	$65.8_{1.3}$ ↑0.3%	$68.5_{1.8}$ ↑1.4%	$75.2_{1.8}$ ↓1.2%	$76.5^{*}_{3.9}$ ↑4.9%	↑1.1%
Staged-Ensemble	$\underline{71.7^{*}_{3.3}}$ ↑1.9%	$\underline{71.5^{*}_{1.7}}$ ↑8.9%	$69.2_{5.4}$ ↑2.2%	$\underline{76.8_{8.2}}$ ↑0.9%	$74.0_{3.7}$ ↑1.5%	↑3.1%
Joint-Feature	$68.3_{2.8}$ ↓2.8%	$66.3_{1.5}$ ↑1.0%	$66.5_{1.0}$ ↓1.7%	$75.6_{2.7}$ ↓0.8%	$76.0_{8.5}$ ↑4.1%	↑0.0%
Staged-Feature	$70.5_{2.5}$ ↑0.2%	$66.7_{3.0}$ ↑1.5%	$\underline{72.2^{*}_{2.6}}$ ↑6.7%	$76.2_{7.4}$ ↑0.0%	$\underline{77.6^{*}_{5.8}}$ ↑6.4%	↑3.0%

Table 3: Feedback classification results based on Manual Transcription - *Text (Manual)* and Automated Speech Recognition - *Text (ASR)*. **Mean %** refers to the average gain of the model taking multimodality over the best performing single modality input. The subscripts are the standard deviation of different runs. * indicates a statistically significant gain compared to the best individual modality model at $p<0.05$. Note that for ***Praise***, due to the information contained in particular modalities, is expected that †*Text* input only leads to high classification performance while video only leads to relatively low performance. Similarly for ***Visual Aid*** due to reliance on visual pointing, the ‡*Video* modality is expected to perform particularly well. See Fig. 4 for details.

3.6. Training Strategies of Multi-Modality Fusion

We explore 3 training strategies as depicted in Fig. 1-D: 1) Individual training of each modality, 2) Joint training (J) of all modalities, 3) Staged training (S), which starts with individual training followed by further joint training.

Individual Training: Each modality model is trained independently for the same number of epochs. Each modality also makes an independent prediction about the final label. This setup offers a simple no-fusion baseline. We further use the independently trained models with the voting fusion model (Best Voting) to offer the basic fusion baseline.

Joint Training (Joint): The individual modality models are combined under one fusion architecture (Ensemble or Feature) and trained jointly for the whole duration of the training. This approach allows the fusion model to learn how to extract relevant information from each modality simultaneously and possibly also learn the differences between modalities relevant to the task.

Staged Training (Staged): The models for each modality are first pre-trained independently on the same task for half of the training time (Stage 1 *"Individual"*). Then the pre-trained models are combined under the fusion model (*Ensemble* or *Feature*) and trained further jointly for the remainder of the training time (Stage 2 *"Joint"*). The first stage helps each model extract relevant information from its modality without interference from other modalities. Extraction of such information from less predictive modalities can otherwise be suppressed.

3.7. Evaluation Schemes and Setups

We obtain baselines for each individual modality by fine-tuning models for the same number of epochs and reporting the top AUC score obtained on the test set. For all our experiments, we use label-balanced datasets for each feedback dimension obtained via random downsampling of the majority class. We use an 80%/20% random train/test split and perform each experiment 3 times with a controlled random seed and report mean AUC as well as standard deviation. Dimension-specific label balancing leads to variable dataset size for each dimension, specifically: *Anatomic* (N=2208), *Procedural*(N=1634), *Technical*(N=1378), *Praise* (N=524), *Visual Aid* (N=606).

We further compare the performance of the fusion approaches to the best-performing individual modality model using McNemar's non-parametric statistical test as suggested in Dietterich (1998) and further adapted to the settings involving expensive deep learning setups by Vanwinckelen and Blockeel (2012). We use a Python implementation of McNemar's test provided in Raschka (2018).

3.8. Data Processing and Model Training

We trim a 10-second video with audio information when human-annotated feedback appears. This includes 5 seconds before (to capture context) and 5 seconds after (to capture delivery) the feedback onset. We preprocess the video by downsampling the resolution to 320 × 250 and extracting 16 randomly uniformly sampled frames. We preprocess the audio by downsampling it to 16kHz mono. We train all the fusion models for a total of 20 epochs with the same initial learning rate (LR) of $5\text{e} - 6$, Adam optimizer, and a scheduler that reduces LR when an AUC has stopped improving for 2 epochs (patience) with a reduction factor of 0.5. We use a batch size of 2 with a gradient accumulation of 10.

4. Results

4.1. Feedback Classification Results

Table 3 summarizes the results of the classification of feedback components. The top rows report AUCs for individual modalities. We include 2 versions of transcribed text from audio. *Text(Manual)* - costly manual transcription by human experts, *Text(ASR)* - automated transcription from audio using ASR and Speaker Diarization as described in §3.3. Text modality itself is highly predictive for each component. In the subsequent two sections of the table we report AUCs for multi-modal fusion approaches relying on high quality, but costly manual transcriptions - *"Fusion Using Manual Transcription* and same fusion approaches relying on automated transcription from audio - *"Fusion Using Automated Transcriptions (ASR) and Speaker Diarization"*.

In each row we report the AUCs for different fusion approaches. The Best Voting is a majority vote fusion baseline. The following rows report results for joint (Joint) and staged (Staged) training approaches of the same architecture *Ensemble Fusion* (Ensemble) model. The last two rows report the results of joint (Joint) and staged (Staged) training for *Feature Fusion* (Feature) model. Next to each AUC score for fusion approaches, we report relative gain or loss with respect to the highest AUC from individual modalities. We also underscore the highest AUC per each feedback component across all models.

Varying, but Consistent Gains from Fusion: The top AUC for automated classification of *"Praise"* is high at 76.8, and fusion provides the least gain of 0.9% for this component. This is likely because praise can be delivered in different visual contexts and hence video does not provide much more information. The AUC is also high, at 77.6, for *"Visual Aid"*. In this case, the gain from incorporating video is substantial at 6.4%. This is due to the visually observable pointing associated with this feedback component. For *"Anatomic"*, *"Procedural"*, and *"Technical"* the top AUCs are 71.7, 71.5, and 72.2 respectively. The best fusion approach provides a noticeable gain for these components of between 1.9% and 8.9%.

Staged Training Outperforms Other Approaches: Looking at the results of staged training (*Staged-Ensemble* and *Staged-Feature*) in comparison to other fusion approaches, the staging outperforms them all. We observe a mean gain of 3.0% to 6.8% across fusion relying on automated and manual transcription respectively. This is compared to smaller gains of just 2.0% or even no gain for the joint training with the exact same model architectures (*Joint-Ensemble* and *Joint-Feature*). At the same time, the simple majority vote fusion (*Best Voting*) leads to AUC loss in 4 out of 5 dimensions over the best individual modality.

Given relatively high standard deviations we examine the statistical significance of the observed gains as described in §3.7. Statistically significant gains are marked with "*" in Table 3. For fusion relying on manual transcriptions, the best-performing fusion architecture offers statistically significant gains over the best-performing single modality model for all feedback components: *Anatomic* (M=4.5, p<0.01), *Procedural* (M=7.2, p<0.01), *Technical* (M=6.0, p<0.01), *Praise* (M=1.0, p<0.05), and *Visual Aid* (M=7.7, p<0.01); where M denotes the best mean absolute AUC gain.

For fusion relying on automated transcription, we observe statistically significant gains from the best-performing architecture for 4 out of 5 feedback components: *Anatomic* (M=1.4, p<0.05), *Procedural* (M=5.7, p<0.01), *Technical* (M=5.0, p<0.01), *Praise* (M=0.6, p=0.58, n.s.), *Visual Aid* (M=4.6, p<0.05).

Intuitive Value of Individual Modalities: We further note intuitive patterns in the predictive value of individual modalities per feedback components. For *"Praise"*, text and audio modalities alone achieve relatively higher AUCs of 76.2 and 67.3 respectively compared to video with AUC of 57.0. This is intuitive as this component captures the feedback delivery and can contain emotional undertones, but praise can be delivered in any visual context. On the other hand, for *"Visual Aid"* video alone achieves a high AUC of 73.0, while audio alone achieves a much lower AUC of 61.2. This is again intuitive as this component captures the surgeon using a visual aid in the form of a cursor or surgical instrument as a pointer (see Fig. 4-A).

Manual v.s. Auto Transcription: Table 3 shows that the experiments leveraging fusion using manual expert-provided transcriptions offer higher AUCs ranging from 76.5 to 96.2. This represents a 15.3% average improvement over fusion relying on fully automated ASR transcription and speaker diarization. These results offer a likely upper bound for classification performance as well as show the potential of improving the processing of audio. However, obtaining manual expert transcriptions of live surgical feedback is costly. While there exist services for medical text transcription (Princeton-Transcription, 2023), they are still costly and may not offer the same quality for specialized surgical domains.

4.2. Additional Analysis

To better understand the value of fusion, we manually inspect several examples of disagreements between simple majority vote fusion, the true label, and the prediction from the best-performing fusion model. In Fig. 4 we show two illustrative examples of such disagreements. In A) we show an example of *"Visual Aid"* component classification, where neither text nor audio provides the correct label. Looking at the video, it is clear that the trainer is using a visual pointer, but pointing can also happen using instruments. In example B) for classification of *"Praise"* the feedback text by itself is insufficient to correctly determine if the feedback is intended to be positive. The inclusion of delivery aspects from audio is important in that case.

At scale, we quantified the impact on Precision and Recall. In the case of fusion relying on manual transcription, for *"Visual Aid"* the best fusion model improved Precision by 15.5% and Recall by 0.6% compared to the Best Voting (see Appendix A). It also improved Precision by 10.5% and Recall by 9.2% compared to the best single modality. A similar impact can be observed for *"Praise"* with improvement in Precision by 21.6% and Recall by 35.1% compared to baseline Best Voting and improvements in Precision by 1.9% and Recall by 0.7% compared to the best single modality. Similar trends can be seen in fusion relying on automated transcription. For *"Visual Aid"* there is an improvement in Precision by 10.0% and in Recall by 13.2% when using the best fusion approach compared to Best Voting. Similarly, compared to the best single modality, fusion improves Precision by 6.7% and Recall by 8.1%. Enhanced Precision indicates fewer false positives, affirming the relevance of the identified feedback instances of a particular type. Improved Recall signifies fewer missed authentic feedback cases of that type. Improvement in both Precision and Recall underscores not only an increase in the accurate detection of specific feedback types but also a broader and more reliable capture of feedback components.

5. Discussion

We thoroughly evaluate different multi-modality fusion architectures and training strategies across 3 data splits and 5 surgical feedback classification dimensions. The low effectiveness of a simple majority vote gives us insights into the manner in which the modalities need to be combined. It seems important to have a number of trainable parameters in order to learn how to combine the information across modalities. We gain evidence for this via examination of disagreements between *Best Voting* and the staged fusion setting, which shows that improvements are based on both precision and recall scores.

Further increase in the number of trainable parameters does not translate to improvements. It is likely that only a limited complexity is needed to relate the

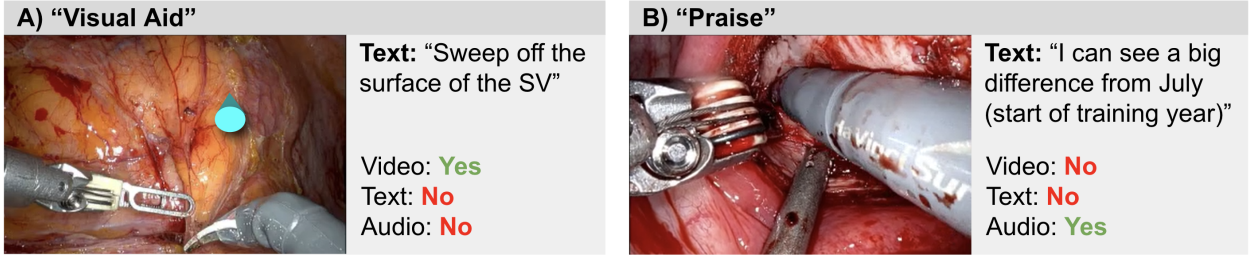

A) "Visual Aid"
Text: "Sweep off the surface of the SV"
Video: Yes
Text: No
Audio: No
B) "Praise"
Text: "I can see a big difference from July (start of training year)"
Video: No
Text: No
Audio: Yes

Figure 4: Examples of video clips where a single modality does not contain enough information to make a correct classification. **A**) The verbal feedback itself is not enough to determine whether a trainer is pointing to anything specific. It is necessary to look at video modality and the use of a pointer (teal cone) to make a correct determination. Please note that we enlarged the pointer for visual clarity. **B**) The text itself is too ambiguous to determine whether the feedback is positive or not. Additional information from the tone of voice provides the necessary distinction.

modalities effectively. We did not freeze any of the individual model weights and these models are themselves complex.

We introduce staged training to address the issue of dominance of the text modality over other modalities. We observe that this approach led to the highest gain across classifications of all the feedback components irrespective of the fusion model complexity. This shows the importance of considering the training process itself for fusion, while most of the prior work focused on model architectures.

We further note that the automated multimodal classification we introduced in this work is based on a clinically validated manual system from Wong et al. (2023). As such, these classification dimensions have been shown to be generalizable across 6 types of surgical procedures. They have also been shown to predict significant differences in surgeon experience level, the surgical task being performed, as well as the likelihood of behavioral adjustment observed among trainees (a measure of feedback effectiveness). This further shows the practical real-world utility of the automation of this classification system through the novel deep multimodal fusion approach we proposed.

We note several future directions. First, the quantification of surgical feedback is an important first step towards generating the optimal feedback automatically using retrieval or generative models in the future (Laca et al., 2022). Second, in this study, we experimented with both manual and automated transcription of feedback from audio. We show that manual transcription, which requires substantial effort, offers better performance. Further experiments should try to improve the performance of automated transcription (Moore, 2015). Finally, our individual models are all transformer architectures capable of unsupervised pre-training, which could improve the overall performance even further.

6. Conclusion

We present the first work to explore an automated classification of components of real-world informal live surgical feedback using a clinically validated classification scheme. We show that it is feasible to classify components of such feedback with promising AUCs varying from 71.5 up to 77.6. Secondly, we show that this feedback is indeed inherently multimodal and fusion can meaningfully improve AUC by as much as 8.9%. Third, we show that the multimodal fusion through staged training is more effective than the fusion model architecture itself. This work provides important insights into the importance of training strategy for effective multi-modal fusion. We open up opportunities for quantification of surgical feedback at scale from text, audio, and video recordings, which can lead to improvements in surgical training and outcomes.

Acknowledgments

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under Award Number R01CA251579 and R01CA273031. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Riaz A Agha, Alexander J Fowler, and Nick Sevdalis. The role of non-technical skills in surgery. *Annals of medicine and surgery*, 4(4):422–427, 2015.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

Esther M Bonrath, Nicolas J Dedy, Lauren E Gordon, and Teodor P Grantcharov. Comprehensive surgical coaching enhances surgical skill in the operating room. *Annals of surgery*, 262(2):205–212, 2015.

Said Yacine Boulahia, Abdenour Amamra, Mohamed Ridha Madi, and Said Daikh. Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. *Machine Vision and Applications*, 32(6):121, 2021.

Hervé Bredin and Antoine Laurent. End-to-end speaker segmentation for overlap-aware resegmentation. In *Proc. Interspeech 2021*, Brno, Czech Republic, August 2021.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. pyannote.audio: neural building blocks for speaker diarization. In *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, May 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Thomas G Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.

Simon DiMaio, Mike Hanuschik, and Usha Kreaden. The da vinci surgical system. *Surgical robotics: systems applications and visions*, pages 199–217, 2011.

Michael M Haglund, Andrew B Cutler, Alexander Suarez, Rajeev Dharmapurikar, Shivanand P Lad, and Katherine E McDaniel. The surgical autonomy program: a pilot study of social learning theory applied to competency-based neurosurgical education. *Neurosurgery*, 88(4):E345–E350, 2021.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Jasper A Laca, Rafal Kocielnik, Jessica H Nguyen, Jonathan You, Ryan Tsang, Elyssa Y Wong, Andrew Shtulman, Anima Anandkumar, and Andrew J Hung. Using real-time feedback to improve surgical performance on a robotic tissue dissection task. *European Urology Open Science*, 46:15–21, 2022.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020.

Runzhuo Ma, Ryan S Lee, Jessica H Nguyen, Andrew Cowan, Taseen F Haque, Jonathan You, Sidney I Roberts, Steven Cen, Anthony Jarc, Inderbir S Gill, et al. Tailored feedback based on clinically relevant performance metrics expedites the acquisition of robotic suturing skills—an unblinded pilot randomized controlled trial. *The Journal of Urology*, 208(2):414–424, 2022.

Robert J Moore. Automated transcription and conversation analysis. *Research on Language and Social Interaction*, 48(3):253–270, 2015.

Marla Narazani, Ignacio Sarasua, Sebastian Pölsterl, Aldana Lizarraga, Igor Yakushev, and Christian Wachinger. Is a pet all you need? a multimodal study for alzheimer's disease using 3d cnns. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part I*, pages 66–76. Springer, 2022.

Princeton-Transcription. Princeton transcription — intelligent transcription services. https://princetontranscription.com/, 09 2023. (Accessed on 09/05/2023).

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

Sebastian Raschka. Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. *The Journal of Open Source Software*, 3(24), April 2018. doi: 10.21105/joss.00638. URL https://joss.theoj.org/papers/10.21105/joss.00638.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pretraining. *arXiv preprint arXiv:2203.12602*, 2022.

Gitte Vanwinckelen and Hendrik Blockeel. On estimating model accuracy with repeated cross-validation. In *Proceedings of the 21st Belgian-Dutch Conference on Machine Learning*, pages 39–44, 2012.

Aaron Walsman, Yonatan Bisk, Saadia Gabriel, Dipendra Misra, Yoav Artzi, Yejin Choi, and Dieter Fox. Early fusion for goal directed robotic vision. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1025–1031. IEEE, 2019.

Elyssa Y. Wong, Cherine H. Yang, Istabraq S. Dalieh, Daniela C. Sotelo, Jasper A. Laca, Runzhuo Ma, Timothy N. Chu, Rafal Kocielnik, Mitchell G. Goldenberg, Jamal A. Nabhani, Steven Cen, and Andrew J. Hung. Deconstructing and quantifying live surgical feedback in the operating room. In *American Urological Association Annual Conferenc)*. AUA, 2023.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*, 2021.

Appendix A. Additional Classification Metrics

In Table 4 we present additional F1-binary, Precision, and Recall metrics from *Visual Aid* feedback component classification. Tables 5, 6, 7, and 8 contain additional metrics for *Praise*, *Anatomic*, *Procedural*, and *Technical* components respectively.

Model	F1-binary	Precision	Recall
Text (Manual)	78.05	78.94	77.59
Text (ASR)	65.50	67.89	64.48
Audio (Emotion)	60.83	62.21	60.11
Video	73.19	72.64	73.77
Fusion using Manual Transcription			
Best Voting	79.58	76.56	84.16
Joint-Ensemble	81.90	83.55	80.33
Staged-Ensemble	85.84	87.27	84.70
Joint-Feature	79.81	83.87	76.50
Staged-Feature	86.10	84.45	97.98
Fusion using ASR Transcription			
Best Voting	70.47	70.48	70.49
Joint-Ensemble	76.42	76.93	75.96
Staged-Ensemble	72.86	77.34	69.40
Joint-Feature	73.20	81.16	67.21
Staged-Feature	78.11	77.51	79.78

Table 4: Additional metrics for **Visual Aid** component - F1 binary, Precision and Recall.

Model	F1-binary	Precision	Recall
Text (Manual)	94.95	95.16	94.90
Text (ASR)	73.98	79.11	70.10
Audio (Emotion)	64.31	66.40	62.41
Video	54.35	59.88	52.20
Fusion using Manual Transcription			
Best Voting	74.87	79.72	70.70
Joint-Ensemble	94.56	95.66	93.62
Staged-Ensemble	96.18	96.94	95.54
Joint-Feature	94.90	95.68	94.26
Staged-Feature	96.16	96.89	95.54
Fusion using ASR Transcription			
Best Voting	71.09	69.27	73.24
Joint-Ensemble	75.25	73.82	78.41
Staged-Ensemble	75.42	79.29	72.04
Joint-Feature	77.49	71.51	84.68
Staged-Feature	74.27	80.01	69.45

Table 5: Additional metrics for **Praise** component - F1 binary, Precision and Recall.

Model	F1-binary	Precision	Recall
Text (Manual)	80.94	84.03	78.33
Text (ASR)	67.88	73.00	63.67
Audio (Emotion)	68.88	65.95	73.00
Video	66.19	65.61	67.00
Fusion using Manual Transcription			
Best Voting	79.75	79.51	80.00
Joint-Ensemble	81.42	82.62	80.33
Staged-Ensemble	85.64	87.49	84.33
Joint-Feature	81.60	82.60	80.67
Staged-Feature	85.68	87.53	84.00
Fusion using ASR Transcription			
Best Voting	69.92	68.25	71.67
Joint-Ensemble	70.03	71.32	69.00
Staged-Ensemble	71.65	71.72	71.63
Joint-Feature	67.58	69.98	66.00
Staged-Feature	70.24	71.54	69.33

Table 6: Additional metrics for **Anatomic** component - F1 binary, Precision and Recall.

Model	F1-binary	Precision	Recall
Text (Manual)	70.39	68.02	73.33
Text (ASR)	65.71	65.76	65.67
Audio (Emotion)	64.89	60.29	70.67
Video	63.88	64.10	63.67
Fusion using Manual Transcription			
Best Voting	72.04	71.83	72.33
Joint-Ensemble	72.80	71.90	74.33
Staged-Ensemble	77.16	75.19	79.67
Joint-Feature	73.43	70.06	77.33
Staged-Feature	76.75	75.66	78.33
Fusion using ASR Transcription			
Best Voting	64.56	63.23	66.00
Joint-Ensemble	64.69	67.29	63.00
Staged-Ensemble	72.89	69.59	76.67
Joint-Feature	64.77	68.39	62.67
Staged-Feature	67.39	66.00	69.33

Table 7: Additional metrics for **Procedural** component - F1 binary, Precision and Recall.

Model	F1-binary	Precision	Recall
Text (Manual)	74.48	74.26	75.00
Text (ASR)	63.98	66.86	65.00
Audio (Emotion)	68.11	67.97	69.00
Video	64.99	67.13	63.67
Fusion using Manual Transcription			
Best Voting	73.84	75.15	72.67
Joint-Ensemble	74.85	74.18	75.67
Staged-Ensemble	80.06	76.18	84.67
Joint-Feature	74.28	80.88	69.00
Staged-Feature	80.67	79.77	81.67
Fusion using ASR Transcription			
Best Voting	64.64	71.01	63.00
Joint-Ensemble	65.48	69.70	67.00
Staged-Ensemble	70.34	69.01	70.00
Joint-Feature	65.15	66.52	67.67
Staged-Feature	68.50	75.20	68.00

Table 8: Additional metrics for **Technical** component - F1 binary, Precision and Recall.