



OPEN

Deep learning-based detection of bowel sound events in continuous recordings

Yusuf Çelik

Accurate identification of bowel sound events is essential for the long-term and non-invasive monitoring of gastrointestinal activity. In this study, we present a band-aware multi-band deep learning approach that explicitly accounts for the distinct acoustic characteristics of bowel sounds across different frequency ranges by jointly analyzing low- and high-frequency components. The proposed method operates on time–frequency representations extracted from continuous recordings to produce frame-level bowel sound probabilities, which are subsequently integrated into recording-level decisions through a unified decision-making strategy. Despite the pronounced class imbalance in the dataset, the proposed system achieved accuracy levels exceeding 98% at the frame level and 99% at the recording level, yielding stable and consistent outcomes through probability aggregation and temporal smoothing steps. These results indicate that the proposed approach provides a reliable and well-structured framework for bowel sound event detection and establishes a solid methodological basis for the analysis of long-duration recordings.

Keywords Bowel sounds, Sound event detection, Biomedical signal analysis

The auscultation-based evaluation of the gastrointestinal system has long been used in clinical practice. However, this traditional stethoscope-based approach is limited by the short duration of measurement and the subjective nature of interpretation. In recent years, the digital processing and artificial intelligence–assisted analysis of acoustic signals recorded from the abdominal wall, known as bowel sounds, have gained importance as a noninvasive, continuous, and objective monitoring method¹. This trend demonstrates that bowel sounds are evolving from merely audible findings into potential digital biomarkers.

Deep learning, a subfield of artificial intelligence, has revolutionized pattern recognition across multiple domains². It has shown remarkable success in areas such as speech and sound processing^{3,4}, medical imaging⁵, and biomedical signal analysis⁶. Its ability to automatically learn hierarchical representations directly from raw data makes it particularly effective for modeling complex, nonstationary acoustic phenomena like bowel sounds. This capability has positioned deep learning as a powerful tool for analyzing physiological signals in various contexts, including gastrointestinal monitoring.

In the literature, two main directions have emerged in this context:

- the detection or segmentation of bowel sound events,
- the classification of gastrointestinal motility or pathological conditions based on features extracted from these events.

The open dataset released by Ficek et al.⁷ became one of the first publicly available resources referenced in deep learning–based studies on bowel sound analysis. Owing to its annotation protocol and data-sharing framework, this dataset established a foundational benchmark for the field. Subsequent research has focused on developing real-time systems capable of operating under noisy and long-duration recording conditions. Baronetto et al.⁸ utilized the multi-channel GastroDigitalShirt system to collect 84 h of recordings and employed a pretrained EfficientNet-B2 model combined with an attention mechanism to detect the presence of bowel sounds within 10-second segments with high accuracy. In a follow-up study, Baronetto et al.⁹ extended this approach into a multiscale and noise-robust framework, validating it on highly imbalanced long-term monitoring data.

In recent years, research on bowel sound analysis has diversified across domains of hardware design, physiology, and clinical application, leading to the development of novel approaches ranging from portable systems to real-time artificial intelligence solutions. Kutsumi et al.¹⁰ developed a convolutional neural network–

Department of Computer Engineering, Munzur University, Tunceli, Türkiye, Turkey. email: celikyusuf@munzur.edu.tr

based model using data recorded solely through a smartphone microphone and achieved high concordance with motility-related metrics. Studies examining motility variations across sleep stages have demonstrated the physiological relevance of bowel sounds, particularly their connection to the gut–brain axis¹¹. Clinically oriented investigations have extended beyond motility monitoring toward disease-specific diagnostic systems. Using lightweight and embedded models, Zhang et al.¹² demonstrated the feasibility of bowel sound–based artificial intelligence systems for colorectal cancer diagnosis. Furthermore, Kalahasty et al.¹³ proposed the “You Only Listen Once” (YOLO) approach, which combines one-dimensional convolutional neural networks with transformer layers to enable real-time detection of distinct bowel sound events.

Such real-time and high-resolution models have revealed the potential of bowel sound analysis while also exposing new challenges related to data volume, annotation accuracy, and computational cost. Data scarcity and the high cost of the labeling process remain major limitations in this field. Consequently, transfer learning and self-supervised pretraining approaches have gained prominence in recent years. Yu et al.¹⁴ enhanced bowel sound recognition performance using pretrained representation learning models and demonstrated that unsupervised pretraining strategies can be effective in small-data scenarios.

From a methodological perspective, newly developed modeling architectures have made significant contributions to the advancement of the field. Matynia and Nowak¹⁵ introduced BowelRCNN, a region-based convolutional framework capable of directly estimating the temporal boundaries of bowel sound events. Similarly, Mutlu and Serbes¹⁶ employed lightweight convolutional networks on Mel-spectrogram representations to classify heart, lung, and bowel sounds in real time on resource-constrained hardware. Similarly, Lin et al.¹⁷ introduced a comprehensive AI-based auscultation framework for the automatic classification and acoustic differentiation of heart, lung, and bowel sounds. Their system demonstrated high diagnostic accuracy across multiple supervised learning models and highlighted the clinical potential of multi-organ acoustic analysis for applications such as telemedicine and continuous patient monitoring.

From a clinical standpoint, the need for noninvasive and continuous monitoring, particularly in newborn and preterm infants, further underscores the importance of bowel sound–based analytical solutions¹⁸. Such applications highlight the translational potential and real-world applicability of increasingly mature modeling approaches in clinical settings.

Previous studies have generally revealed challenges such as the noise sensitivity of bowel sound signals, inconsistencies in annotation, and class imbalance. In the literature, binary decisions of “bowel sound present/absent” are typically produced through band limitation, time–frequency transformation, and sliding-window segmentation, while sampling or weighting strategies are employed to address data imbalance¹. This situation underscores the need for reproducible and systematic investigation of bowel sound events.

This study reexamines the binary classification problem defined as the presence or absence of sound using a systematic and comprehensive approach based on the dataset published by Ficek et al.⁷. The aim is to provide a reproducible reference for this fundamental task, evaluate the impact of preprocessing and windowing choices, and investigate how different strategies for addressing class imbalance affect performance. To support this analysis, a band-aware multi-band deep learning framework is employed, in which low- and high-frequency time–frequency representations are processed through parallel temporal branches to capture the frequency-dependent acoustic characteristics of bowel sounds. The model produces frame-level event probabilities that are subsequently aggregated to obtain recording-level decisions, enabling event detection to be examined at both the frame and file levels. This unified evaluation strategy allows a structured assessment of decision mechanisms across multiple temporal resolutions and provides a solid foundation for future event-based or disease-oriented bowel sound analysis studies.

Methodology

The proposed bowel sound event detection system follows an end-to-end workflow, illustrated in Fig. 1. Continuous audio recordings are first partitioned into training and fixed hold-out test sets. The signals then undergo signal preprocessing before being transformed into dual-band wavelet-based time–frequency

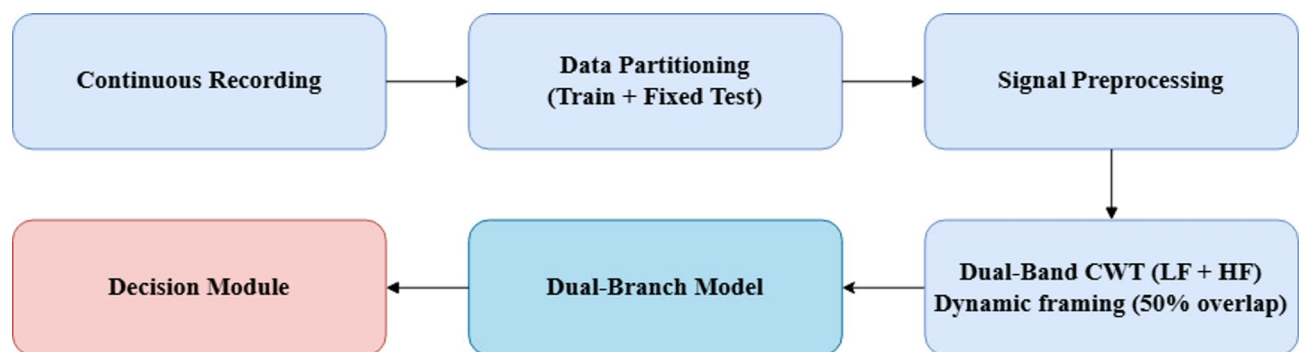


Fig. 1. End-to-end workflow of the proposed bowel sound event detection system, including data partitioning, signal preprocessing, dual-band CWT transformation, dynamic framing, dual-branch modeling, and temporal decision aggregation.

representations (Low-Frequency (LF) and High-Frequency (HF) components). Event-oriented dynamic framing with 50% overlap is applied to increase temporal resolution in annotated regions.

The resulting representations are processed by a dual-branch temporal model, where band-specific features are learned in parallel and fused through an attention-based mechanism. Frame-level probabilities are subsequently generated and combined through a temporal decision module to produce recording-level outputs. This unified pipeline enables evaluation at both frame and recording levels.

All experiments were conducted using Google Colab with NVIDIA GPU acceleration.

Dataset

In this study, the bowel sound dataset collected by Poznań University of Medical Sciences and publicly released on the Kaggle platform was used^{7,19}.

The original study was conducted in accordance with relevant guidelines and regulations and was approved by the Bioethical Committee at Poznań University of Medical Sciences, and written informed consent was obtained from all participants. The dataset was created as part of a clinical study aimed at recording gastrointestinal acoustic activity throughout the night.

Recordings were obtained using a contact microphone specifically designed for bowel sounds and a Tascam DR-10CH audio recorder (24-bit, 44.1 kHz). The microphone features a piezoelectric transducer with a low resonance frequency (≈ 1.3 kHz) and a 3D-printed housing that incorporates acoustic isolation layers.

Data were collected from nineteen volunteers, and the continuous recordings were segmented into two-second clips. Each segment was manually labeled by gastroenterology experts based on auditory and morphological characteristics. The labeling scheme consists of seven subclasses: a (single-burst sound), b (double-burst sound), bc (mixed burst–continuous sound), d (long-duration digestive sound), m (mechanical-like sound), s (continuous random sound), and nan (acoustic events not related to bowel activity).

Figure 2 illustrates the waveforms of two sample recordings from the dataset along with the annotation scheme; regions marked in different colors correspond to distinct bowel sound types as identified by experts (Table 1).

This open-access dataset provides a standardized and highly reproducible reference source for research on the automatic analysis and modeling of bowel sounds.

Preprocessing

Before being used in the modeling process, the bowel sound recordings underwent a multi-stage preprocessing pipeline designed to preserve the characteristic time–frequency structure while minimizing amplitude variability. Recordings were first converted to mono, and the DC offset was removed by subtracting the mean value. A 6th-order Butterworth band-pass filter (40–2000 Hz) was applied to retain the principal energy range of gastrointestinal sounds. This was followed by a pre-emphasis filter (coefficient = 0.97) to enhance high-frequency components. An RMS (Root Mean Square)-based gating mechanism (-40 dB threshold) suppressed low-energy silent regions, and all recordings were peak-normalized to ensure numerical stability.

Each preprocessed recording was then transformed into the time–frequency domain using the Continuous Wavelet Transform (CWT) with a Morlet wavelet. Previous studies have shown that bowel sound energy is predominantly concentrated within the 80–1700 Hz range. Based on this evidence, two complementary sub-bands were defined to capture discriminative spectral characteristics: a low-frequency band (LF, 80–400 Hz) and a high-frequency band (HF, 500–1700 Hz)²⁰. The intermediate range between 400 Hz and 500 Hz was intentionally excluded to create a clear separation between the two branches and to reduce potential overlap between low- and high-frequency representations. A resolution of 12 voices per octave was used for both bands. The resulting CWT amplitudes were log-transformed and normalized using scale-wise Z-score normalization.

Time–frequency representations were segmented into 10 ms analysis windows.

Bowel sound recordings were categorized into three classes based on expert annotations:

- event: frames containing bowel sounds,
- nan_event: acoustic events not related to bowel activity,
- none: silent or low-energy background frames.

The dataset consists of 1,606 files in total, where 1,283 files (79.9%) contain at least one bowel sound event, and 323 recordings (20.1%) are entirely event-free.

A 50% overlapping segmentation (5 ms shift) was applied only to event and nan_event regions after the dataset partitioning stage, and the training and test recordings were processed independently. This strategy increased temporal resolution in event regions while avoiding redundant data in silent segments. As a result, the total number of labeled frames was approximately doubled, providing a richer temporal sampling, particularly for complex or sustained acoustic patterns. The none segments were kept non-overlapping, leading to an overall 8.8% increase in total frame count—from 321,200 to 349,535.

In the evaluation phase, a 50% overlap was applied to event-related frames to increase the temporal sampling density of sound events and to examine the stability of the model's predictions across different temporal offsets. This procedure increases the number of event frames in the test set, providing a denser temporal representation of acoustic events near their boundaries. However, it should be emphasized that the proposed model itself does not rely on this overlap strategy. In practical deployment scenarios, the system can operate on continuous recordings using a uniform sliding-window segmentation, since the model's predictions are determined by learned acoustic patterns rather than a specific framing scheme.

For each frame, the root-mean-square (RMS) energy of the wavelet coefficients was computed independently for each CWT frequency scale, resulting in a 50-dimensional feature vector that preserves the spectral distribution

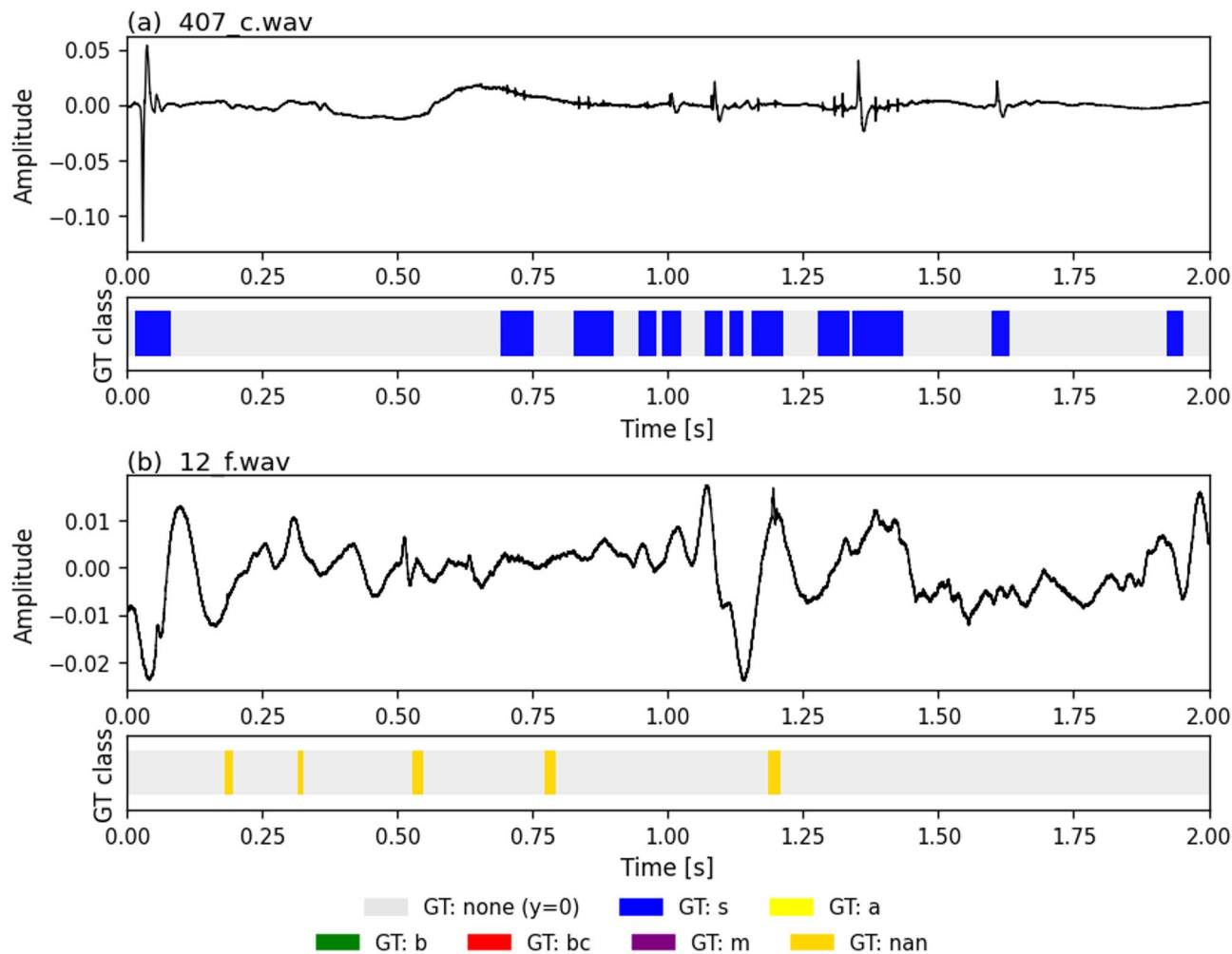


Fig. 2. Waveforms and expert annotations from two bowel sound recordings. **(a)** In *407_c.wav*, short and frequently recurring bowel sound events are observed, while **(b)** *12_f.wav* exhibits sparse and high-amplitude acoustic events. The colored bars below each waveform indicate the time intervals and class labels assigned by gastroenterology experts (GT: Ground Truth).

Category/ Class	Description	No overlap (frames)	50% overlap (frames)	Change (%)
Event	Frames labeled as bowel sound events by experts	20 936	40 348	+ 93
a	Single burst bowel sound	1 708	3 406	+ 99
b	Double burst bowel sound	980	1 932	+ 97
bc	Burst–continuous mixed sound	5 059	10 069	+ 99
d	Long-lasting digestive sound	5	8	+ 60
m	Mechanical-like sound	353	697	+ 97
o	Other acoustic event	14	30	+ 114
s	Continuous random sound	12 817	24 206	+ 89
nan_event	Frames annotated as noise (non-bowel acoustic activity)	9 121	17 050	+ 87
none	Unlabeled background or silent frames	291 143	292 137	~ 0
Total		321 200	349 535	+ 8.8

Table 1. Frame-level distribution before and after 50% overlapping segmentation.

Protocol	isevent	TRAIN (frames)	TEST (frames)	TOTAL
Kaggle-provided split (as released in the dataset)	1 (positive)	48,370	9028	57,398
	0 (negative)	232,528	59,609	292,137
Total		280,898	68,637	349,535
Random 85/15 split (15% test + 5-fold CV on remaining)	1 (positive)	48,563	8835	57,398
	0 (negative)	248,412	58,402	292,137
Total		296,975	52,560	349,535

Table 2. Frame-level isevent distribution under two partitioning protocols.

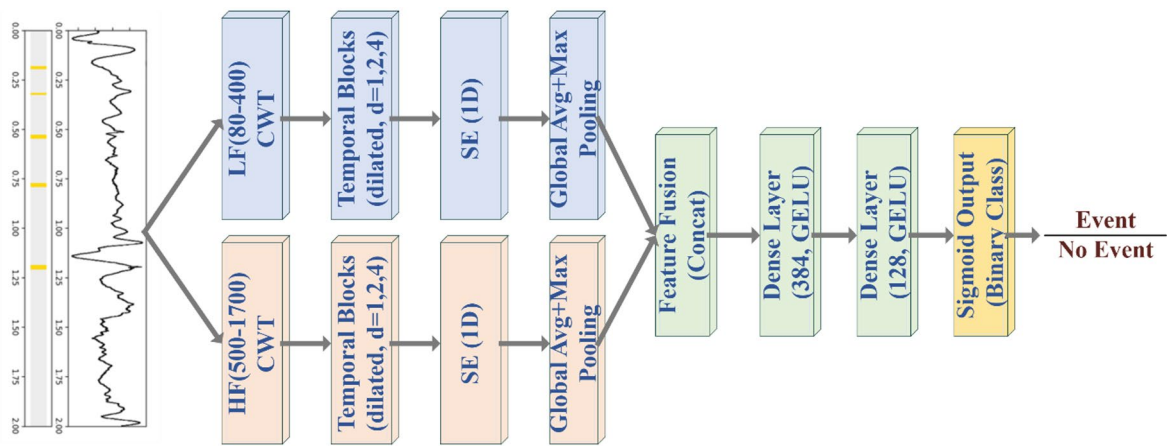


Fig. 3. Overview of the proposed Band-Aware Dual-Branch Temporal Model for bowel-sound event detection, which processes low- and high-frequency CWT bands through parallel temporal blocks followed by SE(1D) attention, pooling, and dense layers for binary classification.

across frequency bins. Consequently, each frame was represented by a 50-dimensional feature vector obtained from both LF and HF bands. These features, together with their corresponding labels and temporal indices, were stored in HDF5 (Hierarchical Data Format version 5) format.

Overall, this preprocessing pipeline yielded a stable and interpretable feature space that effectively captured the characteristic acoustic patterns of bowel sounds.

Partitioning protocols

The dataset (1,606 WAV files, totaling 349,535 frames) was evaluated under two different partitioning protocols:

- Protocol 1: followed the original train/test split provided in the Kaggle version of the dataset. The test portion corresponds to approximately 20% of the entire dataset, while the remaining training portion was further divided into five folds (5-fold cross-validation) for model training and validation.
- Protocol 2: replicated the experimental setup reported by Ficek et al. (2021), in which 15% of the data was randomly assigned to the test set, and the remaining 85% was used for 5-fold cross-validation.

The random split protocol was introduced independently of the Kaggle partition to assess the model’s overall generalizability and stability under random sampling conditions. Table 2 summarizes the frame-level is_event distribution for both evaluation protocols.

In both protocols, the data were split at the file level, and a fixed hold-out test set was strictly excluded from cross-validation, model selection, hyperparameter tuning, and threshold optimization, serving solely for final performance evaluation.

Model architecture

Following the data partitioning protocols, a dual-branch deep learning model was designed for event detection. Figure 3 illustrates the overall architecture and data flow of the proposed framework. The model operates on continuous time–frequency representations and produces frame-level event probabilities (1 = event, 0 = no event). Each input sample consists of a temporal context window of 41 consecutive frames, resulting in an input tensor of size 41 × 50. Temporal convolutions are applied along this temporal axis to capture local dependencies between adjacent frames.

The proposed model adopts a band-aware dual-branch temporal architecture designed to capture frequency-dependent characteristics of bowel sounds. Given a time–frequency representation at each frame, the input

is split into low-frequency (80–400 Hz) and high-frequency (500–1700 Hz) bands according to predefined frequency indices.

Each band is processed by an independent temporal branch. Within each branch, the input features are first projected into a latent space using a linear layer (LF: 96 channels, HF: 160 channels). This is followed by a sequence of three temporal convolution blocks employing depthwise-separable 1D convolutions with kernel size 5 and dilation factors of 1, 2, and 4, respectively. This design enables the model to capture short- and long-range temporal dependencies while maintaining computational efficiency. The proposed architecture contains 375,457 trainable parameters, reflecting the lightweight design enabled by the use of depthwise-separable temporal convolutions.

To enhance channel-wise feature selectivity, a squeeze-and-excitation (SE) module is applied after the temporal blocks in each branch. The SE module performs global temporal pooling followed by a lightweight gating mechanism that adaptively recalibrates channel responses.

The outputs of each branch are summarized using both global average pooling and global max pooling over time. These pooled representations are concatenated across bands and passed to a fully connected prediction head composed of two hidden layers (384 and 128 units) with GELU activations and dropout regularization (dropout rate = 0.25, with a reduced rate of 0.125 in the final hidden layer). The final output is a single logit per frame, which is converted into a frame-level event probability using a sigmoid function.

This architecture allows the model to jointly exploit complementary temporal dynamics across frequency bands while preserving interpretability and reproducibility.

This architecture was designed to model the distinct spectral behaviors of bowel sound energy across different frequency bands, as defined in the previous section. The low-frequency (LF) branch primarily represents longer-duration and single-burst sounds, while the high-frequency (HF) branch captures short, random, and high-energy components. Through its parallel pathways, the model simultaneously learns complementary features from both bands, while the SE(1D) attention layers dynamically weight the relative contribution of each frequency branch. As a result, both spectral and temporal variations are effectively considered, leading to improved accuracy in event detection.

Experimental setup and results

Experimental setup

The proposed model aims to discriminate frames that contain bowel sounds (events) from those consisting of background or noise segments (no events). During training, the class imbalance problem was addressed through weighted sampling, ensuring balanced representation of both classes. To further improve generalization, random masking and perturbations along the time and frequency axes (SpecAugment-style transformations) were applied, enhancing the model's robustness to various noise conditions and spectral variations.

Model predictions were generated at the frame level, followed by a post-processing stage designed to ensure temporal consistency. In this stage, a grid of median-filter kernel sizes and binary-smoothing parameters was evaluated using the validation (VAL) set. The configuration yielding the highest validation accuracy (VAL-ACC) was selected and fixed for the test data to avoid any information leakage. This process suppressed isolated errors, merged short event clusters separated by brief gaps, and enabled the model to produce consistent responses for short-duration events (typically within 20–40 ms).

The probabilities produced by the model at the frame level represent short time segments within each audio recording. To evaluate the system's performance at the recording (file) level, these frame probabilities were aggregated into a single value per file (file-level aggregation), resulting in one probability score for each recording.

In this study, the Maximum Probability Aggregation (MPA) approach, originally introduced by Kim et al.²¹ for sound event detection (SED), was adopted. In this method, the recording-level probability score s_{file} is obtained by selecting the maximum probability among all frames belonging to the corresponding recording:

$$s_{\text{file}} = \max_{i \in \mathcal{I}_{\text{file}}} p_i$$
 where p_i denotes the event probability of frame i , and $\mathcal{I}_{\text{file}}$ represents the set of frame indices within the given recording.

Accordingly, if any frame exhibits a high event probability, the corresponding recording is assumed to contain an event. During the decision stage, various threshold values (τ) were tested on the validation set, and the one yielding the highest accuracy was selected. The final class label for each recording was determined according to the following decision rule:

$$\hat{y}_{\text{file}} = \begin{cases} 1, & \text{if } s_{\text{file}} \geq \tau \\ 0, & \text{otherwise} \end{cases}$$
 The selected threshold (τ) was optimized exclusively on the validation set and

subsequently fixed for evaluation on the test set, ensuring that no information from the test data leaked into the model selection process. The validation-based grid search yielded similar optimal values for both experimental protocols, and the selected recording-level decision threshold was $\tau = 0.70$ for both partitioning scenarios.

Five-fold cross-validation was applied within the training subset for model selection, and each fold-specific model was subsequently evaluated on the same fixed hold-out test set. The evaluation metrics included Accuracy, Recall (Sensitivity), F1-score, and Average Precision (AP).

The model was trained using the AdamW optimizer with an initial learning rate of 2×10^{-4} and a weight decay 5×10^{-5} . A OneCycle learning rate scheduling policy was applied during training to improve convergence stability. The batch size was set to 16, and training was performed for a maximum of 120 epochs. Early stopping was employed with a patience of 7 epochs based on validation loss to prevent overfitting. A dropout rate of 0.25 was used in the fully connected layers. The main training hyperparameters used in the experiments are summarized in Table 3.

Parameter	Value
Optimizer	AdamW
Initial learning rate	2×10^{-4}
Weight decay	5×10^{-5}
Learning rate scheduler	OneCycleLR
Batch size	16
Maximum epochs	120
Early stopping patience	7
Dropout rate	0.25

Table 3. Training hyperparameters used in the experiments.

Level	ACC	Sensitivity	F1-score	AP	AUC
Frame-level	0.9884 ± 0.00049	0.9544 ± 0.00316	0.9548 ± 0.00147	0.9918 ± 0.00045	0.9990 ± 0.000008
File-level	0.9979 ± 0.00019	0.9965 ± 0.00053	0.9970 ± 0.00033	0.9992 ± 0.00008	0.9994 ± 0.00005

Table 4. Performance metrics on the fixed hold-out test set under the Kaggle-provided split (mean $\pm$ standard deviation).

Level	ACC	Sensitivity	F1-score	AP	AUC
Frame-level	0.9876 ± 0.00105	0.9632 ± 0.00301	0.9620 ± 0.00184	0.9944 ± 0.00017	0.9988 ± 0.00004
File-level	0.9984 ± 0.00200	0.9979 ± 0.00208	0.9989 ± 0.00110	0.9996 ± 0.00042	0.9996 ± 0.00042

Table 5. Performance metrics on the randomly sampled fixed hold-out test set (85/15 split protocol, mean $\pm$ standard deviation).

Results

The performance of the proposed model was evaluated under two different data partitioning protocols: the original Kaggle-provided split and a random 85/15 split. In both scenarios, five-fold cross-validation (5-fold CV) was applied to the training data, while the decision threshold and temporal post-processing settings determined during validation were kept fixed during the test phase. The reported mean $\pm$ standard deviation values correspond to performance variability across five independently trained models, each evaluated on the same fixed hold-out test set.

Regarding the Kaggle-provided split, the internal 5-fold cross-validation on the training data yielded an average validation accuracy of 0.9874 ± 0.001 and an F1-score of 0.9626 ± 0.002 . As summarized in Table 4, the model subsequently achieved an average frame-level accuracy of 0.9884 ± 0.00049 on the fixed hold-out test set, with a sensitivity of 0.9544 ± 0.00316 , an F1-score of 0.9548 ± 0.00147 , an average precision (AP) of 0.9918 ± 0.00045 , and an AUC (Area Under the Curve) of 0.9990 . When frame-level probabilities were aggregated into recording-level decisions using the Maximum Probability Aggregation (MPA) strategy, the model further achieved an average file-level accuracy of 0.9979 ± 0.00019 and an F1-score of 0.9970 ± 0.00033 , as also reported in Table 4.

A similar evaluation was conducted using the random 85/15 split protocol. In this scenario, the internal 5-fold cross-validation on the training data reached a consistent average validation accuracy of 0.9880 ± 0.0008 and an F1-score of 0.9626 ± 0.003 , showing stable performance across all folds. As shown in Table 5, the model subsequently obtained an average frame-level accuracy of 0.9876 ± 0.00105 on the hold-out test set, with a sensitivity of 0.9632 ± 0.00301 , an F1-score of 0.9620 ± 0.00184 , an AP of 0.9944 ± 0.00017 , and an AUC of 0.9988 ± 0.00004 . At the file level, aggregation using the same MPA strategy resulted in an average accuracy of 0.9984 ± 0.00200 , an F1-score of 0.9989 ± 0.00110 , and an AUC of 0.9996 ± 0.00042 .

In both partitioning protocols, the model achieved near-perfect accuracy, F1-score, and AUC values at both the frame and file levels, successfully detecting almost all events. Under the Kaggle split, results exhibited lower variance and greater stability, whereas the random 85/15 scenario demonstrated the model's strong generalizability under random sampling conditions. The MPA strategy applied at the file level further enhanced decision reliability by suppressing short-duration erroneous frames.

In addition to these quantitative findings, Fig. 4 presents a comparative visualization of the model's frame-level predictions and the ground-truth labels. The displayed recordings were selected from bowel sound samples containing varying numbers of events. In panels (a), (b), and (c), the model's predictions generally align well with the true event regions, with only minor discrepancies in a few frames. Panel (d) corresponds to a silent recording with no events, where the model correctly produced a "no event" decision without generating any false positives. These visual examples confirm that the model produces temporally consistent predictions and effectively suppresses background noise.

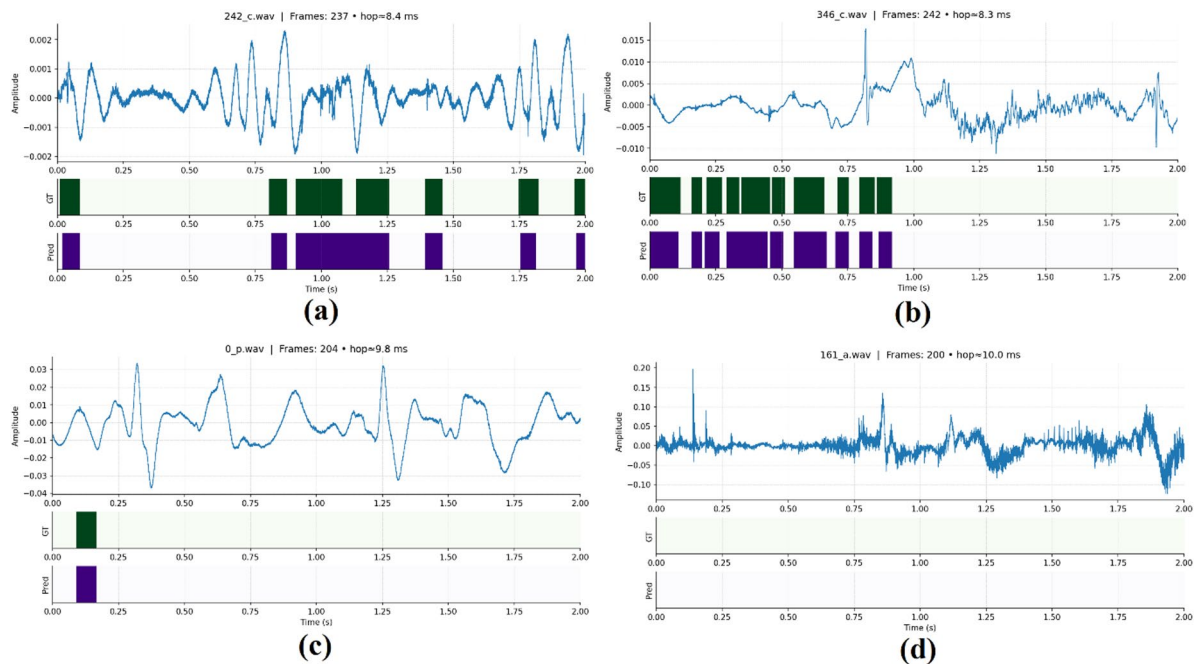


Fig. 4. Frame-level predictions of the proposed model compared with the ground truth (GT) for different bowel sound recordings. The top waveforms represent the signal amplitude, while the green (GT) and purple (Pred) bars indicate the temporal locations of bowel sound events. Transparent regions correspond to non-event frames. Panels (a–c) show recordings containing different numbers of events, where the model predictions largely overlap with the reference annotations. Panel (d) represents a silent recording with no bowel sound events, where the model correctly produced no false activations.

As a result, the proposed model achieved an average performance exceeding 99.8% in both evaluation scenarios, confirming its stable, generalizable, and highly reliable event detection capability.

Discussion

The obtained results demonstrate that the proposed system can accurately and consistently detect bowel sound events under both the Kaggle and random 85/15 data partitioning protocols. The model achieved average accuracy, F1-score, and AUC values approaching 99% at both the frame and file levels, exhibiting highly stable performance. The very low inter-fold variance indicates that the system is robust to variations caused by data sampling.

This study was conducted using the Ficek et al.⁷ dataset, which is widely used in bowel sound analysis research. The work by Ficek and colleagues introduced this dataset and presented the first baseline frame-level results. While their study was limited to frame-level analysis, the present research provides a systematic performance evaluation at both frame and file levels on the same dataset. Consequently, it has been demonstrated that bowel sound events can be reliably identified not only within short-duration frames but also across entire recordings.

Similarly, Yu et al.¹⁴ utilized the same dataset to enhance model representational power under data-scarce conditions by leveraging pre-trained self-supervised audio models, such as HuBERT and wav2vec 2.0. The authors applied low-parameter fine-tuning (LP1) to the wav2vec 2.0 model, achieving 85.34% accuracy, and demonstrated that self-supervised pretraining can significantly improve model performance when limited labeled data are available.

In a related effort, Matynia and Nowak¹⁵ proposed a region-based convolutional neural network (BowelRCNN) for temporal detection and localization of bowel sound events. Their model identified event regions within 0.2-second windows and predicted onset and offset times to perform temporal segmentation. They reported 96% accuracy, 71% F1-score, and an average Intersection-over-Union (IoU) of 0.551.

When evaluated comparatively, studies conducted on the same dataset have reported varying levels of performance. Ficek et al.⁷ established the baseline reference by achieving 97.4–97.7% accuracy and 71–77% sensitivity at the frame level using CRNN and CDNN architectures. Yu et al.¹⁴ aimed to enhance representational capability under limited-data conditions through low-parameter fine-tuning of the pre-trained wav2vec 2.0 model, yet the overall accuracy remained around 85%. Matynia and Nowak¹⁵ employed a region-based CNN (BowelRCNN) on time–frequency representations to estimate the temporal locations of events, reporting 96% accuracy and 71% F1-score.

In comparison, the present study achieved 98.8% accuracy and 95–96% sensitivity across both frame- and file-level evaluations, demonstrating a higher and more balanced overall performance than previous approaches. Despite the dataset exhibiting a clear class imbalance, with event-containing frames constituting only a small fraction of the total, the system successfully detected the minority class with high sensitivity, maintaining a

Study	Evaluation level	Main approach	ACC [%]	Sens. [%]	F1 [%]
[7] – CRNN	Frame	CNN+BiLSTM hybrid (frame-level classification)	97.7	77.3	79.9
[7] – CDNN	Frame	Pure CNN (baseline)	97.4	71.1	76.7
[14] – wav2vec 2.0 (LP1)	Segment	Pre-trained self-supervised model + low-parameter fine-tuning	85.3	–	–
[15] – BowelRCNN	Frame	Two-stage region-based CNN for temporal event localization	96.8	74.2	71.1
This study (proposed)	Frame	Probabilistic event detection using multi-band CWT representation	98.8	96.0	96.0
This study (proposed)	File	Frame-level probability fusion with Maximum Probability Aggregation (MPA)	99.8	99.7	99.8

Table 6. Comparative overview of recent deep learning approaches evaluated on the same Bowel Sounds dataset.

strong accuracy–recall balance. This outcome highlights the stability and robustness of the proposed model even under imbalanced data conditions.

Notably, the file-level evaluation represents one of the key contributions of this study. By converting frame-level probabilities into single recording-level decisions through the Maximum Probability Aggregation (MPA) method, the system effectively reduced the impact of short-term frame-level errors, achieving nearly perfect recording-level accuracy. The average file-level accuracy exceeded 99.8% in both scenarios, confirming that the proposed system can reliably analyze long-duration recordings with high decision stability and strong error tolerance.

While the current MPA-based aggregation provides a robust baseline for segment-level detection, it is inherently designed for binary presence–absence decisions. In clinical settings where more detailed descriptors of gastrointestinal activity are required, such as event frequency, duration, or temporal patterns, this approach should be considered a coarse-grained detection mechanism. Consequently, isolated artifacts or impulse-like disturbances may potentially influence recording-level decisions if they generate high frame-level probabilities. Future extensions of this framework may therefore incorporate more fine-grained event-level analyses, including onset–offset detection and event counting, which could enable the estimation of motility-related indicators such as inter-sound intervals.

Table 6 presents a comparison of various approaches previously conducted on the same dataset alongside the proposed system’s key performance metrics obtained at both the frame and file levels. Only metrics explicitly reported in the original studies were included; F1-scores were computed when derivable from reported precision and sensitivity. As shown, while earlier studies reported notable success at different evaluation levels, the proposed system demonstrates a significant improvement, particularly in terms of sensitivity and recording-level decision stability.

The high file-level accuracy indicates that the proposed approach can generate consistent decisions not only at the frame level but also when aggregating information across entire recordings. This characteristic suggests that the method may be suitable for supporting recording-level analyses in bowel sound research, particularly in scenarios requiring stable interpretation of long-duration acoustic data.

In conclusion, this study was conducted using the dataset introduced by Ficek et al.⁷ and extends the bowel sound event detection problem—previously examined only at the frame level—to include file-level decision analysis. The obtained results indicate that high accuracy and stable decision-making can be achieved within the defined experimental setting, highlighting the potential of data-driven approaches for bowel sound characterization.

Conclusion

In this study, a deep learning–based bowel sound event detection system was evaluated using the publicly available dataset introduced by Ficek et al.⁷. The primary contribution of this work is the development of a band-aware dual-branch architecture that achieves consistently high accuracy, sensitivity, and decision stability across different data partitioning scenarios. While previous studies have mainly focused on frame-level evaluation, this work additionally introduces a recording-level decision analysis framework, ensuring that high frame-level performance translates into stable and reliable event detection in continuous recordings.

All experiments were conducted using a single, widely adopted public dataset, and the proposed framework addresses the problem of binary event detection. For this reason, the discrimination of different gastrointestinal conditions or clinical states was beyond the scope of the present work.

Another limitation concerns the available metadata of the dataset. The publicly distributed Kaggle version does not provide subject identifiers linking individual audio clips to specific volunteers. Consequently, subject-wise evaluation protocols (e.g., leave-one-subject-out validation) cannot be directly implemented using the provided dataset structure. Future datasets including explicit subject metadata would enable a more comprehensive evaluation of cross-subject generalization.

Despite these limitations, the proposed architecture demonstrated stable and reliable performance across different data partitioning scenarios. Future research may focus on validating the proposed framework on larger and more diverse bowel sound datasets and establishing subject-independent evaluation protocols. In addition, extending the framework toward real-time gastrointestinal monitoring systems represents a promising direction for practical clinical applications.

Data availability

The dataset used in this study is publicly available on the Kaggle platform (Ficek et al., 2021). No additional datasets or proprietary materials were generated or used in this study.

Received: 20 November 2025; Accepted: 29 March 2026

Published online: 30 March 2026

References

1. Redij, R. et al. Practicing digital gastroenterology through phonoenterography leveraging artificial intelligence: Future perspectives using microwave systems. *Sensors* **23**, 2302. <https://doi.org/10.3390/s23042302> (2023).
2. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444. <https://doi.org/10.1038/nature14539> (2015).
3. Çelik, Y. Application of deep learning for voice command classification in Turkish language. *Bitlis Eren Univ. J. Sci.* **13**, 701–708. <https://doi.org/10.17798/bitlisfen.1477191> (2024).
4. Yamamoto, R., Song, E. & Kim, J. M. Parallel WaveGAN: a fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6199–6203 (2020). <https://doi.org/10.1109/icassp40776.2020.9053795>
5. Litjens, G. et al. A survey on deep learning in medical image analysis. *Med. Image. Anal.* **42**, 60–88. <https://doi.org/10.1016/j.media.2017.07.005> (2017).
6. Acharya, U. R. et al. A deep convolutional neural network model to classify heartbeats. *Comput. Biol. Med.* **89**, 389–396. <https://doi.org/10.1016/j.combiomed.2017.08.022> (2017).
7. Ficek, J. et al. Analysis of gastrointestinal acoustic activity using deep neural networks. *Sensors* **21**, 7602. <https://doi.org/10.3390/s21227602> (2021).
8. Baronetto, A. et al. Segment-based spotting of bowel sounds using pretrained models in continuous data streams. *IEEE J. Biomedical Health Inf.* **27**, 3164–3174. <https://doi.org/10.1109/jbhi.2023.3269910> (2023).
9. Baronetto, A. et al. Multiscale bowel sound event spotting in highly imbalanced wearable monitoring data: algorithm development and validation study. *JMIR AI* **3**, e51118. <https://doi.org/10.2196/51118> (2024).
10. Kutsumi, Y. et al. Automated bowel sound and motility analysis with CNN using a smartphone. *Sensors* **23**, 407. <https://doi.org/10.3390/s23010407> (2022).
11. Wang, G. et al. Differences in intestinal motility during different sleep stages based on long-term bowel sounds. *Biomed. Eng. Online* **22**, 105. <https://doi.org/10.1186/s12938-023-01166-z> (2023).
12. Zhang, H. et al. Towards practical colorectal cancer diagnosis: a bowel sound-based system with portable sensor and on-board lightweight AI model. *IEEE Internet Things J.* <https://doi.org/10.1109/jiot.2025.3600644> (2025).
13. Kalahasty, R. et al. A novel You Only Listen Once (YOLO) deep learning model for automatic prominent bowel sounds detection: feasibility study in healthy subjects. *Sensors* **25**, 4735. <https://doi.org/10.3390/s25154735> (2025).
14. Yu, Y. et al. Enhancing bowel sound recognition with self-attention and self-supervised pre-training. *PLOS ONE* **19**, e0311503. <https://doi.org/10.1371/journal.pone.0311503> (2024).
15. Matynia, I. & Nowak, R. BowelRCNN: region-based convolutional neural network system for bowel sound auscultation. *arXiv* (2025). <https://doi.org/10.48550/arXiv.2504.08659>
16. Mutlu, E., Hüseyin, V. & Serbes, G. FISC-Net: classification of auscultation sounds on an edge device. In 2023 Intelligent Methods, Systems, and Applications (IMSA), 174–180 (2023). <https://doi.org/10.1109/imsa58542.2023.10217510>
17. Lin, Y. S., Kapadia, A. & Ortigoza, E. B. Automatic classification and acoustic auscultation of heart, lung, and bowel sounds using artificial intelligence. *Res. Square*. <https://doi.org/10.21203/rs.3.rs-7061625/v1> (2025).
18. Mao, Y. et al. Gastrointestinal emergencies and monitoring in preterm infants: a review. *Ann. Med.* **57** <https://doi.org/10.1080/07853890.2025.2525397> (2025).
19. Nowak, R. Bowel sounds. *Kaggle* <https://doi.org/10.34740/kaggle/dsv/2789758> (2021). data set.
20. Zahra, A. et al. Benchmarking machine learning for bowel sound pattern classification. *PLoS One* **21**, e0338911. <https://doi.org/10.1371/journal.pone.0338911> (2026).
21. Kim, S. W. et al. Sound event detection based on auxiliary decoder and maximum probability aggregation for DCASE 2024 task 4. *arXiv* (2024). <https://arxiv.org/abs/2406.12721>

Acknowledgements

The author would like to thank the creators of the publicly available bowel sound dataset for providing an accessible resource that enables reproducible research.

Author contributions

Y.Ç. designed the study, implemented the methodology, performed the experiments, analyzed the results, prepared the figures and tables, and wrote the manuscript.

Funding

This research received no specific funding.

Declarations

Ethics declaration

The dataset used in this study is publicly available and was originally collected and published by Ficek et al.⁷. The original data collection was conducted in accordance with relevant guidelines and regulations and was approved by the Bioethical Committee at Poznan University of Medical Sciences. Written informed consent was obtained from all participants, as reported in the original study. The present study does not involve any new data collection from human participants. The authors only used publicly available data and had no access to identifiable personal information. Therefore, no additional ethical approval was required.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.Ç.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026