Master Project Report

Giacomo Mossio (5585864)

Bowel Sound Pattern Spotting Fine Tuning using Active Learning

Date: July 7th, 2024

Supervisors: Prof. Dr. Oliver Amft, Annalisa Baronetto

Abstract

This project explores the application of Active Learning (AL) for fine-tuning bowel sound pattern spotting models. The data was previously collected using the GastroDigitalShirt [1], an innovative device that continuously monitors digestion acoustics. The primary objective is to reduce the amount of labeled data required for training without compromising model performance. Various different AL techniques were applied and some insights were found. Our results demonstrate that AL can decrease the labeled data requirement by up to 90%, significantly reducing the labeling workload. Future work includes experimenting with different models, extending the approach to new datasets, and integrating crowdsourcing for more efficient data labeling.

Contents

1 Introduction **1**

2 Related Work **1**

3 Methods **2**
3.1 Dataset and Model 2
3.2 Resources and Evaluation Metrics 2
3.3 Active Learning Process 2
3.4 Implementation 4

4 Study **4**

5 Results **4**
5.1 Overview of Experiments 4
5.2 Best Result 5
5.3 Heuristics Comparison 7
5.4 Analysis on the Segments selected by AL 8

6 Discussion **9**
6.1 Why Active Learning works 9
6.2 Insights 11
6.3 Limitations 11

7 Conclusion **12**
7.1 Key Contributions 12
7.2 Future Works 12

Appendix **I**

Code Listings I

References II

List of Figures II

List of Tables II

1 Introduction

Digestive disorders are prevalent and often require invasive and extensive evaluation to diagnose accurately. Consequently, there is a growing interest in non-invasive diagnostic methods that can offer continuous monitoring, such as the analysis of bowel sounds (BS). These methods can significantly enhance patient comfort and provide continuous data for better diagnosis and monitoring.

Devices like the GastroDigitalShirt [1] have been developed to continuously monitor bowel sounds through multiple microphones installed on the t-shirt. However, the vast amount of audio data collected by such devices necessitates automated methods for analysis. Deep neural networks have shown great promise in modeling complex patterns in audio data. Specifically, EfficientUNet [2], pretrained on AudioSet [3], has demonstrated superior performance in bowel sound recognition tasks according to A. Baronetto's experiments [4].

Training such models typically requires extensive labeled datasets, which can be labor-intensive and costly to obtain. For instance, the dataset collected at the IES Lab contains 136 hours of labeled data with 11,482 BS events, and labeling each hour took 8 to 12 hours of work [4].

Even if the model has already been trained with the whole labelled dataset resulting in good performances, this project tries to find how we can keep similar performances with less labelled data.

In order to reduce the amount of needed labeled data, this project aims to design and test an Active Learning (AL) pipeline for BS spotting. AL is a semi-supervised approach that selects the most informative samples for labeling, thereby reducing the amount of labeled data required for training without compromising model performance.

2 Related Work

As said in the introduction, this whole project is built on top of the existing work done by A. Baronetto [4].

The first step of the project involved a comprehensive review of the existing literature on Active Learning (AL) and bowel sound (BS) recognition.

In particular, the paper "Bayesian Active Learning for Production, a Systematic Study and a Reusable Library" [5] presented the library used for the implementation of AL in this project. Other influential works included studies on active learning for sound event classification and detection [6, 7, 8], as well as methods for efficient audio annotation and classification with a large amount of unlabeled data [9].

These studies collectively informed the design and implementation of the AL pipeline for BS spotting, emphasizing the importance of reducing labeling effort while maintaining high model performance.

3 Methods

3.1 Dataset and Model

Dataset:

The dataset used in this project was previously collected [1] from a group of 27 participants, consisting of 18 healthy individuals and 9 patients diagnosed with Inflammatory Bowel Disease. The total duration of the labelled collected audio data amounts to 136 hours. For analysis, the data was segmented into non-overlapping 10-second segments. Each audio segment was then transformed into Mel Spectrograms with a temporal resolution of 25 milliseconds. The dataset exhibited a significant data imbalance, with bowel sounds (BS) accounting for only 0.89% of the total audio data.

EfficientUnet Model:

The model employed for bowel sound recognition is the EfficientUnet [2], which integrates the U-Net architecture [10] with EfficientNet [11] as the encoder. This model was pretrained on the AudioSet dataset [3]. The audio segments are tranformed into log Mel-spectrograms of shape 128 × 1056, that being the input of our model. The output generated by the model is a sequence of numbers between 0 and 1, to which is applied a softmax and a argmax so that they become a binary detection mask that indicates the presence of 1, Bowel Sounds (BS) or 0, Non-Bowel Sounds (NBS).

3.2 Resources and Evaluation Metrics

For the computational tasks, it was mainly utilized the Intelligent Embedded Systems Laboratory cluster equipped with 8 Nvidia A100 80GB GPUs, enabling efficient training and evaluation of the models. The performance of the models was evaluated using standard metrics such as F1-score, precision, recall and computational efficiency (training time and resource utilization).

3.3 Active Learning Process

The Active Learning process is the core of this project and here follows an illustration of it in Figure 1.

As depicted in the image, the process involves the following steps:

- **Initial Training:** Starting with a small subset of the labeled data, the EfficientUNet model is initially trained.
- **Uncertainty Sampling:** The trained model is then used to predict on the unlabeled data. Samples with the highest uncertainty, where the model is least confident, are selected for labeling.

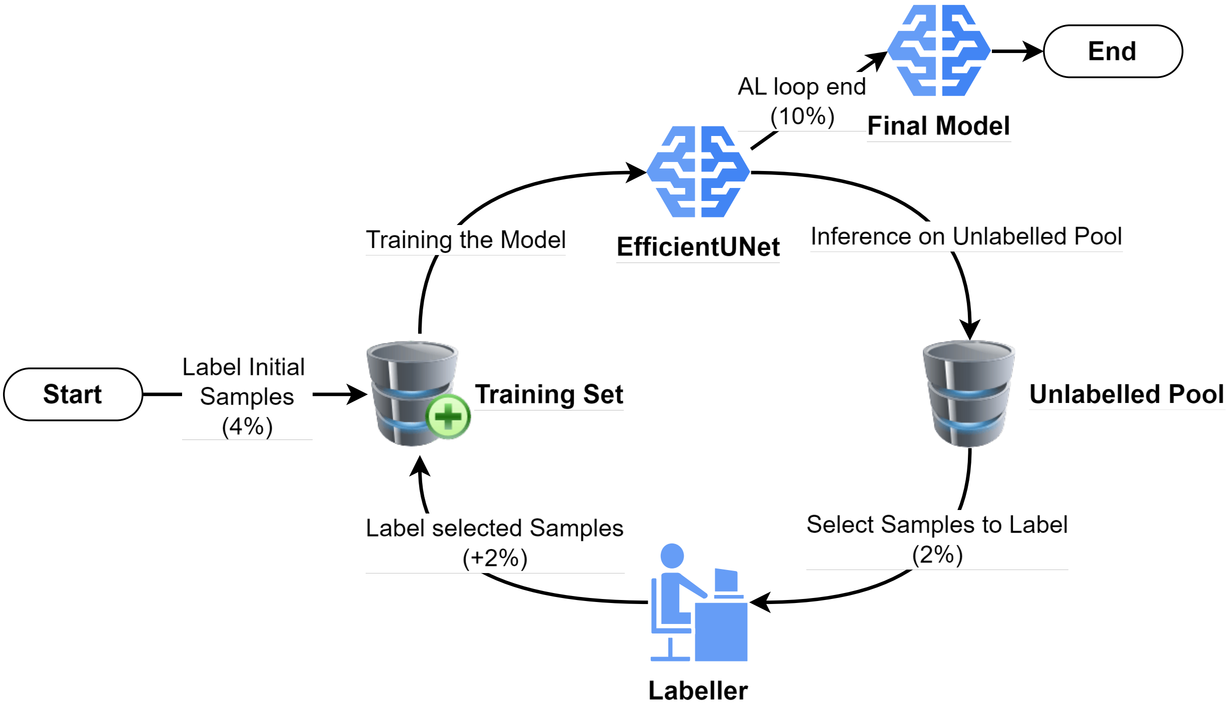

Figure 1: Implementation of Active Learning Process

- **Labelling:** The top x% uncertain samples are then labeled either by experts or unmasked in this case, enriching the training dataset.
- **Model Retraining:** The model is retrained with the newly labeled data, iterating the process to enhance performance.
- **AL loop end:** When we are satisfied with the model performances, or we run out of resources, we terminate the loop and use our final model for inference

More in detail, the sample selection process involved the following steps:

- The current model predicts on the unlabeled dataset.
- Each prediction can be seen as a measure of uncertainty because the model outputs a number ranging from 0 (NBS) to 1 (BS).
- A heuristic, such as entropy, margin, or BALD, calculates a measure of uncertainty for each of the 10-second samples.
- Considering the results of the heuristic, the top x% samples are selected for labeling.

This iterative AL process aims to significantly reduce the labeling workload while maintaining high model performance.

Initial sampling is usually random but in this case was heuristic-based. That was possible thanks to the pretrained nature of the model. This contributed to a small further reduction of needed labelled data because the initial subset was already specifically chosen to maximize the model's learning efficiency.

3.4 Implementation

The implementation of the project built on top of the existing GastroDigitalShirt [4] codebase, using Python with PyTorch Lightning [12] and integrating Bayesian Active Learning techniques via the BaaL library [5].

Modifications and Customizations:

- **Data Loader:** Customized to handle active learning queries, ensuring that the dataset is dynamically updated with newly labeled samples.
- **Trainer:** Adapted to use BaaL's active learning cycle, which integrates the uncertainty sampling and retraining process.
- **Model:** The EfficientUnet model was kept unchanged to maintain consistency in performance evaluation.
- **Pipeline:** The overall pipeline was adapted to include the Active Learning loop, integrating initial training, uncertainty sampling, new data annotation, and model retraining.

The integration of these components ensured that the active learning process was incorporated into the existing framework, allowing for efficient handling of the iterative training and labeling cycles.

4 Study

This project builds on the existing GastroDigitalShirt [1] work that involved 27 participants, including 18 healthy individuals and 9 patients with Inflammatory Bowel Disease (IBD). Participants, aged 21 to 69, wore a smart T-Shirt equipped with eight embedded microphones to record bowel sounds of fasting, meal and postprandial phases. The recordings captured both bowel sounds and other environmental noises while participants engaged in minimal activity to avoid abnormal bowel motility.

5 Results

5.1 Overview of Experiments

The active learning process was optimized by experimenting with various parameters:

- **AL Start Percentage:** The initial percentage of labeled data, tested with values of 2%, 4%, 10%, and 20%.

- **AL Step Percentage:** The percentage of data added in each active learning iteration, varied among 1%, 2%, 3%, 5%, and 10%.
- **AL Number of Steps:** The number of active learning steps performed, tested with values of 3, 4, 5, 8, and 10.
- **Heuristic:** The method used to select the most informative samples, including Random, Entropy, Margin, and BALD.
- **Training Epochs per step:** The number of training cycles per active learning step, experimented with values of 3, 5, 10, and 20.
- **Final Epochs:** The number of epochs used for the final training, tested 10, 15 and 25
- **Training Strategies:**
  - **Training from Scratch:** Retraining the model from the beginning at each iteration to avoid bias and ensure fresh learning.
  - **Using Checkpoints:** Resuming training from the last saved model state to save time and benefit from previously learned features.

Additionally, several model hyperparameters, such as learning rate, loss function, optimizer, and warm-up steps, were modified but those hyperparameters did not improve the results so it was decided to keep the original ones for a consistent baseline comparison. These EfficientUnet hyperparameters included a learning rate of 0.0001, a batch size of 32, the ADAM optimizer, and the DCEL(Dice Cross Entropy Loss) loss function.

Unfortunately, assessing performances of each single configuration is time-consuming and resources-consuming due to the use of Leave One Out Cross Validation (LOOCV), where a model is trained 27 times using each participant except one, and then tested across all configurations. Testing each configuration thoroughly would have required a substantial amount of time, so only the most promising ones were fully tested.

5.2 Best Result

The performance of the active learning approach was evaluated LOOCV F1 Scores. The performance of the best configuration found can be seen in Figure 2. The blue boxplot on the left represents the F1 scores achieved with active learning using 10% of the labeled dataset, while the orange boxplot on the right shows the baseline performance using 100% of the labeled dataset. The optimal configuration for AL involved:

- AL Start Percentage: 4%
- AL Step Percentage: 2%
- AL Steps: 4

- Heuristic: Entropy
- Training Strategy: Training from scratch
- Training Epochs per Step: 10
- Final Epoch Count: 25

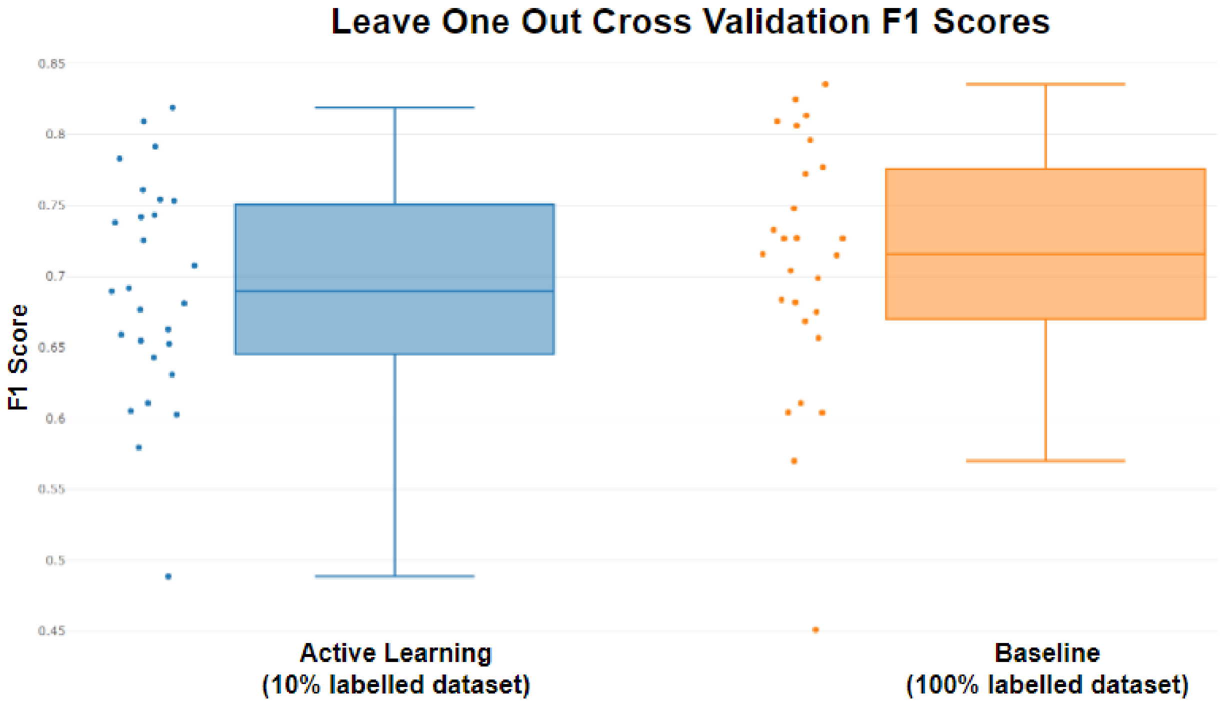

Figure 2: Leave One Out Cross Validation F1 Scores for Active Learning (10% labeled dataset) versus Baseline (100% labeled dataset).

The results demonstrate that the active learning approach, even with only 10% of the labeled dataset, achieved competitive F1 scores compared to the baseline model trained with the full dataset. This indicates the efficiency of the active learning process in reducing the labeling workload while maintaining high model performance.

In Figure 3, we observe a comparison of avergae F1 scores between active learning and the baseline method over successive steps. The lines represent the average F1 scores of the 27 LOOCV iterations. The steps were used to make a fair comparison, instead of time because other processes might have been running on the GPUs at the time of the experiments. Active learning, depicted by the blue line, demonstrates a faster convergence compared to baseline. This result is impressive considering that AL consists of many iterations were the model is trained from scratch every time. That shows that the amount of data with which the model is trained has a big impact on the amount of time the model needs for training.

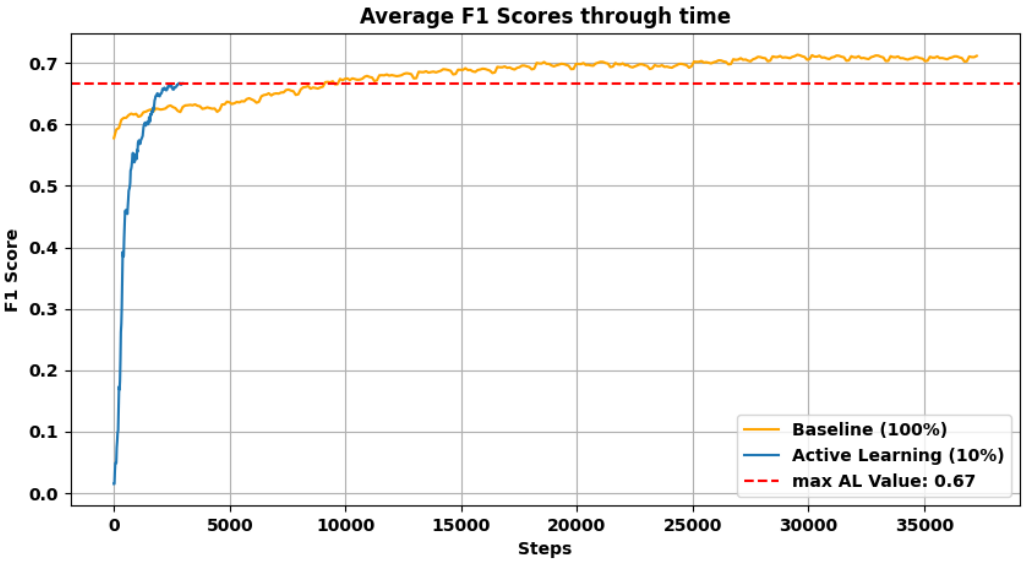

Figure 3: Comparison of Baseline (100% labeled dataset) and Active Learning (10% labeled dataset) average F1 scores through time

5.3 Heuristics Comparison

To optimize the selection of informative samples during active learning, various heuristics were compared. These heuristics determine how samples are chosen to maximize model performance with minimal labeling effort. The following heuristics were evaluated:

- **Random:** Selects samples randomly without considering their informativeness.
- **Entropy:** Prioritizes samples that exhibit the highest uncertainty in model predictions.
- **Margin:** Focuses on samples where the model's prediction is most uncertain between the top two classes.
- **BALD (Bayesian Active Learning by Disagreement):** Uses a Bayesian approach to select samples that maximize the mutual information between the model's parameters and the sample's label.

Figure 4 displays the average F1 scores over successive steps for each heuristic. It's noteworthy that heuristics based on uncertainty significantly impact active learning effectiveness compared to random selection. Interestingly, in this study, the performance difference among heuristics that prioritize uncertainty, like Entropy and Margin, is not substantial. This suggests that in this specific case simpler heuristics, like Entropy, can effectively enhance active learning outcomes without the added complexity of Bayesian approaches.

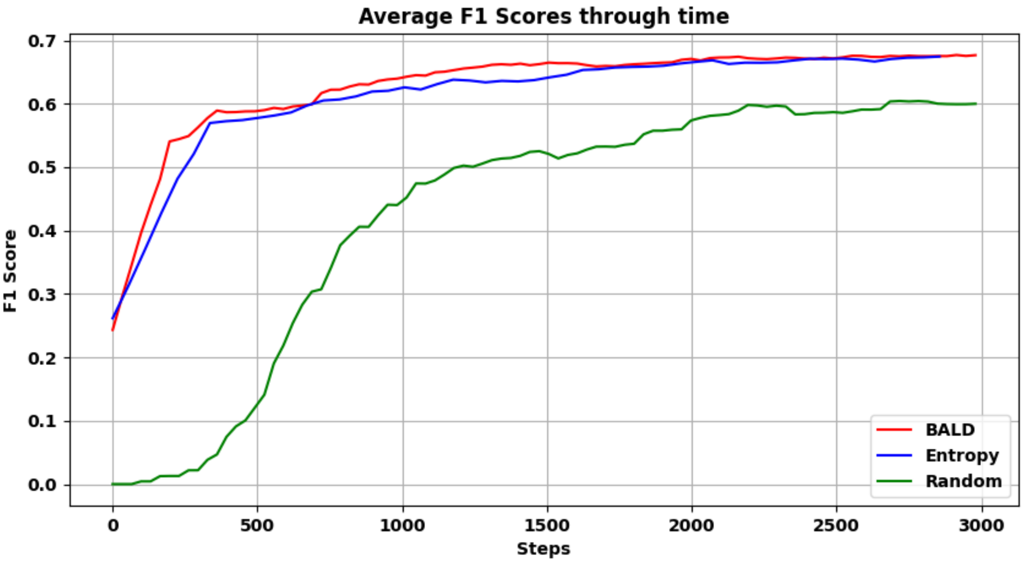

Figure 4: Comparison of different Heuristics in Active Learning: Average F1 Scores through Time

These findings underscore the practicality of employing straightforward yet effective heuristics such as Entropy in active learning scenarios. They provide valuable insights into optimizing sample selection strategies to achieve competitive model performance with reduced labeling effort.

5.4 Analysis on the Segments selected by AL

In previous experiments, we demonstrated the effectiveness of active learning (AL) in significantly reducing the amount of labeled data required compared to baseline methods. However, AL is typically employed beforehand to select the most informative samples for labeling.

As part of this project, an AL process was implemented to identify and label segments in audio data considered most crucial for model training. This process outputs a CSV file listing segments by file number, track number, and start/end times of 10-second segments.

To understand better AL's efficacy, an analysis was conducted on the segments selected for labeling. Figure 5 illustrates that the 10-second audio segments chosen by AL exhibit a significantly higher Bowel Sounds (BS) Ratio compared to the entire dataset. The graph is segmented by participant, with the combined sum of blue lines approximately five times greater than that of the red lines. This discrepancy indicates that AL prioritizes segments containing a substantial amount of BS, facilitating quicker learning and adaptation of the model in identifying these features.

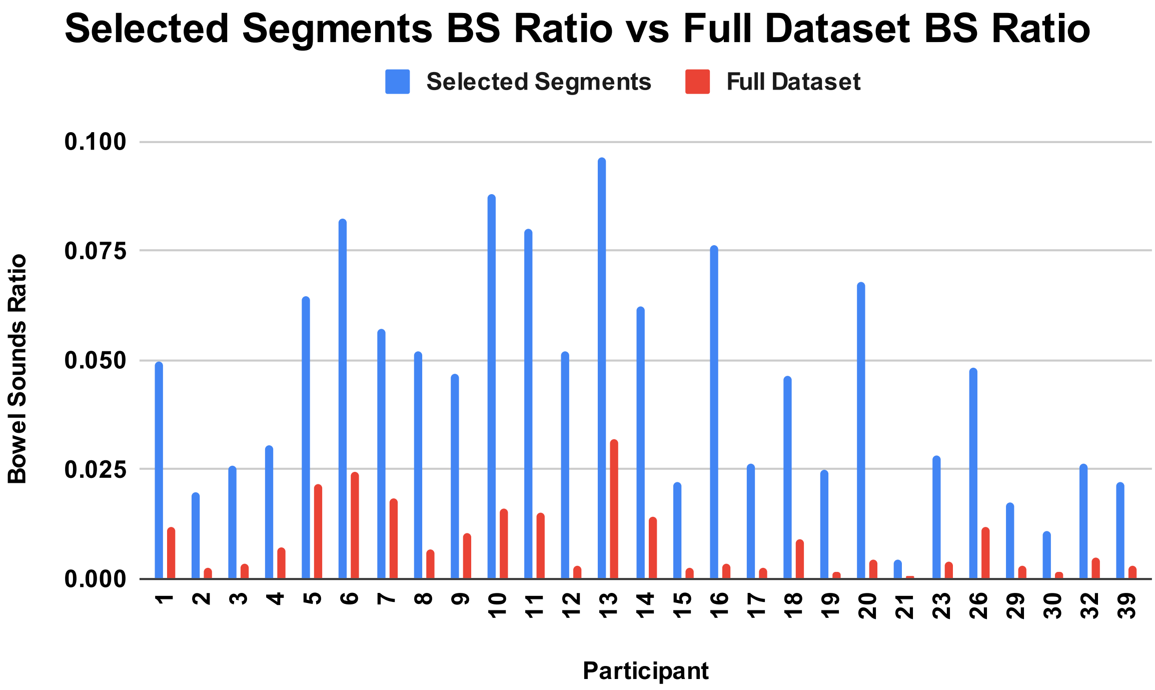

Figure 5: Comparison of Bowel Sounds Ratio in Segments Selected by Active Learning (Blue) versus Entire Dataset (Red)

In scenarios with a highly unbalanced dataset like ours, where certain features are sparse, AL proves exceptionally effective in focusing labeling efforts on critical samples.

Furthermore it is noteworthy, as showed in Figure 6, that samples recommended by the active learning (AL) process are more frequently selected from the postprandial digestive phase compared to the meal and fasting digestive phases. This observation suggests higher uncertainty in predictions during the postprandial phase. This could be attributed to a greater variety of sounds captured by the microphones during this phase. Alternatively, it may indicate a higher prevalence of bowel sounds (BS) in general during this phase, prompting AL to prioritize its selection.

6 Discussion

6.1 Why Active Learning works

The results presented in this study highlight the effectiveness of the active learning (AL) approach in reducing the amount of labeled data. The key factors

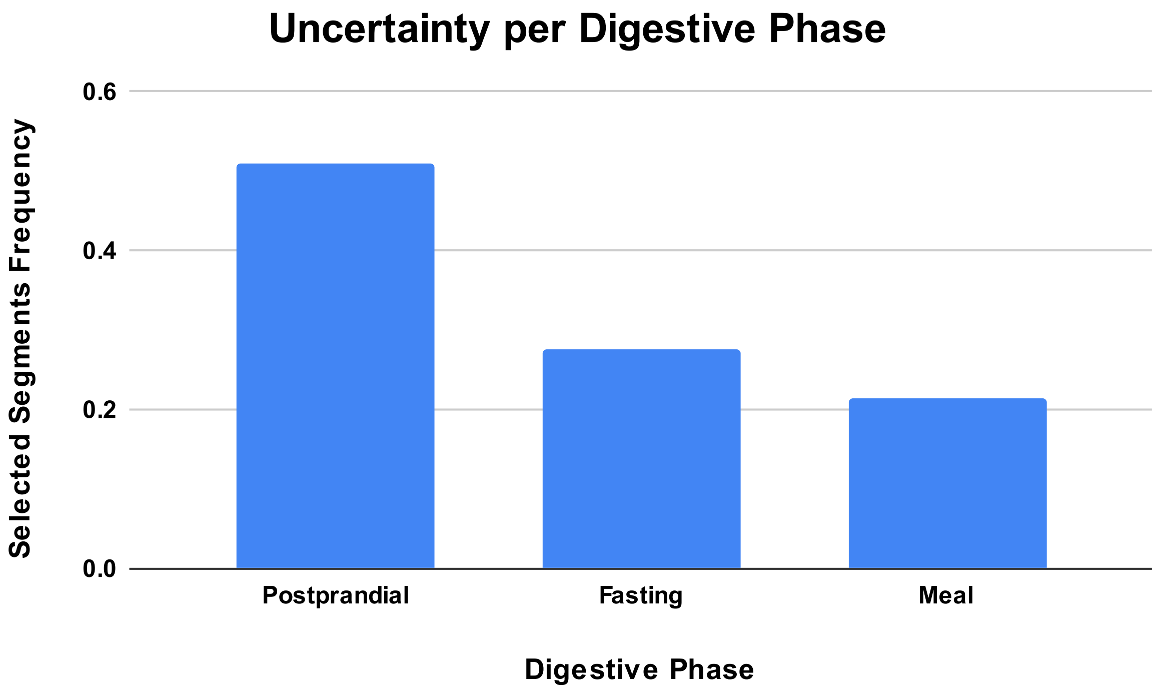

Figure 6: Comparison of Sample Selection Across Digestive Phases by Active Learning

contributing to the success of this method are:

- **Unbalanced Dataset with High Noise:** The dataset contains a substantial amount of noise and imbalanced data distribution, which makes it challenging to train models effectively using traditional methods. Active learning helps focus labeling efforts on the most informative samples, enhancing model performance.
- **High Labeling Costs:** Due to the extensive time and resources required for labeling data, on average 8 to 12 hours for labelling 1 hour of audio data [4], active learning proves advantageous by minimizing the amount of data needed while still achieving competitive performance.
- **Heuristics Based on Uncertainty:** Heuristics that prioritize samples with higher prediction uncertainty ensure that the model learns from the most challenging and informative data points, accelerating the training process.
- **Selection of Segments with Many Bowel Sounds:** The ability of the method to select segments rich in bowel sounds (BS) demonstrates its efficiency in identifying critical features within the dataset, leading to speed up in the learning process.

6.2 Insights

Several insights emerged from the experiments, similar to what discussed in the paper [8]. The following active learning best practices can be applied in similar contexts:

- **Underfitting vs. Overfitting:** The experiments indicate that underfitting has a more harmful impact on the process than overfitting. Therefore, it is preferable to avoid early stopping to ensure the model learns adequately from the data.
- **Initial Pool Impact:** The size and composition of the initial pool of labeled data become less significant after a few iterations, as the active learning process quickly focuses on the most informative samples.
- **Sample Size per Iteration:** Keeping the number of new samples added in each iteration small is beneficial. This approach allows the model to iteratively refine its understanding without being overwhelmed by excessive new data.
- **Heuristic Simplicity:** The study reveals that the choice of heuristic is not crucial; simpler heuristics, such as Entropy, perform sufficiently well, suggesting that complexity in heuristic selection may be unnecessary.
- **Training Strategy Convergence:** As the model stabilizes, the performance difference between training strategies using checkpoints and training from scratch diminishes, indicating that either strategy can be effectively used once the model has matured.

6.3 Limitations

While the active learning approach demonstrates significant advantages, several limitations must be acknowledged:

- **Interactive Nature:** The necessity for repeated training and data labeling phases can be time-consuming and resource-intensive, potentially limiting the method's scalability.
- **Labeling Consistency:** Variability in human labeling can introduce inconsistencies, affecting the overall reliability of the labeled dataset.
- **Initial Model Performance:** The effectiveness of the active learning process heavily depends on the initial performance of the model. If the initial model is not sufficiently competent, the subsequent active learning iterations may not yield optimal results. This is very important because we have to pick a model to use only basing our choice on the initial performances and is not always guaranteed that good initial performances lead to good performances with more data.

7 Conclusion

7.1 Key Contributions

- **Active Learning Efficiency:** The results show that active learning can reduce the required amount of labeled data by up to 90%. This significant reduction highlights the method's ability to focus on the most informative samples.
- **Time Savings:** On average, it takes 8 to 12 hours to label 1 hour of data. With the dataset containing 136 hours of labeled data, active learning would have saved approximately 1224 hours of data labeling, equating to 153 days of 8-hour labeling sessions.
- **Key Concept:** Not all samples contribute equally to model performance. Samples with the highest uncertainty provide the most information gain, making them critical for effective training.
- **Operational Mechanism:** AL algorithms identify and select the most informative and uncertain samples from the dataset, optimizing the learning process.
- **Optimal Usage Scenarios:** AL is particularly effective in datasets with significant noise or redundant samples, where traditional training methods may struggle.

7.2 Future Works

Possible future directions of this study are:

- **Experiment with Different Models:** Exploring alternative models for active learning could further enhance its effectiveness.
- **Extend to New Data:** Labeling new data, already available, suggested by AL, might further improve model performance.
- **Implement Crowdsourcing:** Leveraging crowdsourcing for efficient labeling of new data could streamline the data annotation process and reduce costs. [13]

In conclusion, our study confirms that active learning is a powerful approach for optimizing model performance with minimal labeled data, particularly in challenging datasets with high noise levels and labeling costs. By focusing on the most informative samples, AL significantly reduces the labeling workload while maintaining high performance. Future research should continue to explore new models, extend the approach to new datasets, and integrate crowdsourcing for even more efficient labeling processes.

Appendix

Code Listings

References

[1] A. Baronetto, L. S. Graf, S. Fischer, M. F. Neurath, and O. Amft, "Gastrodigitalshirt: A smart shirt for digestion acoustics monitoring," in *ISWC '20: Proceedings of the 2020 International Symposium on Wearable Computers.* ACM, Sep. 2020, pp. 17–21.

[2] B. Baheti, S. Innani, S. Gajre, and S. Talbar, "Eff-unet: A novel architecture for semantic segmentation in unstructured environment," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 1473–1481.

[3] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[4] A. Baronetto, L. S. Graf, S. Fischer, M. F. Neurath, and O. Amft, "Multiscale bowel sound event spotting in highly imbalanced wearable monitoring data," *under review*, 2020.

[5] P. Atighehchian, F. Branchaud-Charron, and A. Lacoste, "Bayesian active learning for production, a systematic study and a reusable library," *arXiv preprint arXiv:2006.09916*, 2020.

[6] S. Shishkin, D. Hollosi, S. Doclo, and S. Goetze, "Active learning for sound event classification using monte-carlo dropout and pann embeddings," 2021. [Online]. Available: https://publica.fraunhofer.de/handle/publica/417264

[7] S. Zhao, T. Heittola, and T. Virtanen, "Active learning for sound event detection," 2020. [Online]. Available: https://arxiv.org/abs/2002.05033

[8] N. Beck, D. Sivasubramanian, A. Dani, G. Ramakrishnan, and R. Iyer, "Effective evaluation of deep active learning on image classification tasks," 2021. [Online]. Available: https://arxiv.org/abs/2106.15324

[9] Y. Wang, A. E. Mendez Mendez, M. Cartwright, and J. P. Bello, "Active learning for efficient audio annotation and classification with a large amount of unlabeled data," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 880–884.

[10] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” 2015. [Online]. Available: https://arxiv.org/abs/1505.04597

[11] M. Tan and Q. V. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” 2020. [Online]. Available: https://arxiv.org/abs/1905.11946

[12] Lightning-AI, “Pytorch lightning,” https://github.com/Lightning-AI/pytorch-lightning, 2021.

[13] S. Hantke, Z. Zhang, and B. Schuller, “Towards intelligent crowdsourcing for audio data annotation: Integrating active learning in the real world,” in *Proceedings INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association.* ISCA, 2017, pp. 3951–3955.

List of Figures

1 Implementation of Active Learning Process 3
2 Leave One Out Cross Validation F1 Scores for Active Learning (10% labeled dataset) versus Baseline (100% labeled dataset). . 6
3 Comparison of Baseline (100% labeled dataset) and Active Learning (10% labeled dataset) average F1 scores through time 7
4 Comparison of different Heuristics in Active Learning: Average F1 Scores through Time 8
5 Comparison of Bowel Sounds Ratio in Segments Selected by Active Learning (Blue) versus Entire Dataset (Red) 9
6 Comparison of Sample Selection Across Digestive Phases by Active Learning 10

List of Tables