Uncertainty Estimation with Data Augmentation for Active Learning Tasks on Health Data

Sotirios Vavaroutas[1], Lorena Qendro[1,2], Cecilia Mascolo[1]

Abstract— **Supervised machine learning (ML) is revolutionising healthcare, but the acquisition of reliable labels for signals harvested from medical sensors is usually challenging, manual, and costly. Active learning can assist in establishing labels on-the-fly by querying the user only for the most uncertain –and thus informative– samples. However, current approaches rely on naive data selection algorithms, which still require many iterations to achieve the desired accuracy. To this aim, we introduce a novel framework that exploits data augmentation for estimating the uncertainty introduced by sensor signals.**

Our experiments on classifying medical signals show that our framework selects informative samples up to 50% more diverse. Sample diversity is a key indicator of uncertainty, and our framework can capture this diversity better than previous solutions as it picks unlabelled samples with a higher average point distance during the first queries compared to the baselines, which pick samples that are closer together. Through our experiments, we show that augmentation-based uncertainty makes better decisions, as the more informative signals are labelled first and the learner is able to train on samples with more diverse features earlier on, thus enabling the potential expansion of ML in more real-life healthcare use cases.

I. Introduction

The limited availability of good quality and labelled health data from sensors restricts the potential expansion of machine learning (ML) in real-life use cases [1]. Active learning [2] (AL) introduces a solution to reduce the cost of labelling raw signals, a task known to be heavily manual and laborious. The key idea of AL is that algorithms that independently pick the samples on which they are trained will achieve superior performance with fewer epochs. The operation of an AL system relies on the existence of an "oracle", which refers to either a human or an automatic system capable of supplying the labels for a given sample when queried. AL interactively queries the oracle to label new –previously unlabelled– signals at each training pass [3], making it ideal for clinical tasks where labelled samples are scarce. AL essentially prompts the oracle about the most uncertain points to achieve a higher accuracy with only a limited set of labelled sensor signals, yet it is not only ideal for training on small unlabelled datasets but can also achieve better accuracy with fewer epochs than training on an entire dataset.

Previous studies have employed data augmentation in AL to increase the generalisation ability of algorithms, but its use for uncertainty estimation remains largely unexplored. For instance, using augmentation for consistency estimation on each AL cycle improves the performance of the training step [4], yet there is no exploration of its use at test time for uncertainty quantification. Tree search has been used to generate augmentations [5] too, but this approach is very specific to language models. Augmentation has also been used at test-time [6] but is based on the entropy of the predictions of just one augmented set without leveraging the difference observed across various augmented sets. In the context of wellbeing, wearable stress and affect detection has been explored using Monte-Carlo Dropout to represent uncertainty [7], but not using augmentation thus restricting the use to only models that have been trained with dropout regularisation. The use of AL has been explored for activity classification [8], but without featuring a purely uncertainty-based approach for assessing the benefits and costs of labelling a given sample. Finally, augmentation has been explored on physiological signals [9], but in semi-supervised learning instead of AL.

Our contributions are centred around employing data augmentation to capture uncertainty during the active learner's sampling phase, obtain higher and more exact accuracy metrics during testing, and explore the application of our augmentation-based estimation technique to AL tasks for signal processing. Data augmentation is typically used to enhance datasets with a restricted amount of data points, but it also proves insightful for uncertainty estimation. Our experiments show that our approach queries the oracle for a more diverse set of samples, speeding up convergence to a model's ideal accuracy and directly translating to a reduced burden for the users who need to label fewer health samples.

II. Methods

This section discusses our AL framework, which uses data augmentation to capture uncertainty on biological signals.

Active Learning. When training on unlabelled data, an active learner starts by randomly asking the oracle to supply the labels for a small batch of samples and trains on them. This forms its initial labelled training set and it subsequently requests the labels for a further batch of samples, but now the queried signals are carefully chosen. The samples in the newly-labelled batch then enter the labelled pool. The learner can use the new knowledge to proceed with training in a usual supervised manner, while also deciding during the test time of its current iteration which instances to query next. This process can be repeated until a target accuracy is reached or until the oracle stops the training.

Implementation of our Active Learner. In AL, the size of the training set changes from epoch to epoch and an extra querying functionality should be implemented to choose the most uncertain points to label. In our case, we have no

[1] University of Cambridge
[2] Nokia Bell Labs, Cambridge

prior knowledge about the physiological signals, therefore the samples for the first query are sampled randomly, with the active learner then stepping in to pick the most informative signals for all subsequent queries issued to the oracle. At this stage, the novelty of our solution lies in the sample selection strategy which uses our augmentation-based uncertainty estimation technique to ensure that the samples fed to the oracle are carefully selected and as diverse as possible. See Algorithm 1 for a summary.

Test-Time Uncertainty with Augmentation. Data augmentation is commonly used to increase the number of samples for training in small datasets. However, it can also be effectively employed for uncertainty estimation because it captures data uncertainty, consistent across all samples, rather than model uncertainty, which might develop when a model is trained with inadequate data.

The first step in our framework consists of applying a set of augmentations Q to a dataset at test time when the previous training iteration of the active learner has ended. At this point, they are used to augment the pooled dataset and capture its uncertainty. This yields $|Q|$ additional datasets, each composed of the same number of samples placed in the same order. With the original non-augmented dataset, there are $|Q| + 1$ datasets with the same initial signals, each differing from the other by applying different transformations.

Our approach for capturing uncertainty has two steps: first, we ask the classifier model to predict the label for each signal in the unlabelled pool, and we use softmax to calculate the probability of the prediction being correct. Then, we repeat the same task for all the transformed versions of that signal and calculate the entropy of the various probability values; the higher the entropy, the larger the uncertainty will be.

Better Accuracy with Augmentation. In addition to using transformations for capturing uncertainty, we use the same set of transformations for test-time accuracy measurement (see Section III). Given that $|Q|$ additional test datasets are generated through this process, they are used independently to measure the model's accuracy after each epoch and at the end of the learning process. The mean of the accuracy values resulting from these datasets and the original test dataset is significantly more robust. As such, it forms the final accuracy measure considered in our findings.

Early Stopping Strategy. Since manually labelling biological signals is laborious, setting a threshold on how much the algorithm is allowed to ask is imperative. After all, a cost could be associated with each query to the (human or automatic) oracle. Given this, our framework includes an early stopping with a "patience" strategy. The user can decide on the minimum increase in accuracy over the last iterations (based on desired patience) required to invoke early stopping and there is a variable E in our algorithm that can be set to indicate the minimum increase in accuracy over its last three running iterations required to continue the training.

III. Experiments

Datasets. To simulate the human oracle labelling on both binary and multi-class classification tasks, we use the below health datasets that, although labelled, are processed so that the respective labels are kept separately and fed to the learner only when queried for each sample.

Algorithm 1: Augmentation-Based AL.

```
Input  : #Samples per query $N$, Augmentations $Q$
Output: Labelled data $D_L$, Trained model $M_L$
Data   : Unlabelled data $D_U$, Early stopping $E$
$A \leftarrow augment(D_U, Q)$ /*Apply $Q$ to $D_U$*/
while Max queries and E not reached do
    for 5 augmentations from Q do
        $H \leftarrow H \cup softmax\_probabilities(A_Q)$
    /*Get uncertainty with entropy*/
    $H' \leftarrow \arg\max_{x} \; -\sum_i P(y_i \mid x) \log P(y_i \mid x)$
    $D_Q \leftarrow \arg\max (H', N)$ /*Best points*/
    $D_N \leftarrow query(D_Q)$ /*Query user*/
    $D_L \leftarrow D_L \cup D_N$ /*Add to $D_L$*/
    $D_U \leftarrow D_U \setminus D_N$ /*Remove from $D_U$*/
    $M_L \leftarrow learner.train(D_L)$ /*Update*/
    return $M_L, D_L$
```

Epileptic Seizure Recognition [10]: This dataset consists of EEG recordings modelled as time series. It features 11 500 labelled samples from on-body sensors with 178 attributes each. The labels identify if the subject suffered from an epileptic seizure, further classifying the cases in which the subjects did have seizures into four classes depending on whether they had their eyes open or closed during the data collection process, for instance. As the boundaries amongst the four non-epileptic classes are insignificant, this dataset is mostly used for binary classification in practice [10].

Heart Disease Recognition [11]: This ECG dataset identifies heart disease in patients through 14 attributes. The presence of heart disease is further categorised in 4 classes. Although most experiments to date concentrate on distinguishing the absence of heart disease (label 0) from its presence (labels 1, 2, 3 and 4) [11], it is intriguing also to try a multi-class experiment to examine how our uncertainty estimation technique compares to the baselines.

Baselines. The baselines we compare our approach to are based on uncertainty sampling. Uncertainty sampling is the most frequently-used approach in AL [12], as it only queries the label for a sample if its classification uncertainty is high. In contrast to approaches like passive learning which relies on random sampling, an uncertainty-based strategy is optimal for interactive labelling as it makes informed decisions [4]. The most widely-used uncertainty measures are the classification uncertainty, the classification margin, and the classification entropy [13].

The classification uncertainty of an instance x is the simplest utility metric, defined as:

$$U(x) = \arg\max_{x} 1 - P(\hat{y} \mid x), \tag{1}$$

where $\hat{y}$ is the most likely prediction for that instance.

A further uncertainty measure is the classification entropy of the class probabilities, which is proportional to the average number of guesses required to find the correct class:

$$H(x) = \arg\max_x \; -\sum_i P(y_i \mid x) \log P(y_i \mid x). \qquad (2)$$

The classification margin is the difference of the probabilities of the first ($\hat{y_1}$) and second ($\hat{y_2}$) most likely predictions:

$$M(x) = \arg\min_x P(\hat{y_1} \mid x) - P(\hat{y_2} \mid x). \qquad (3)$$

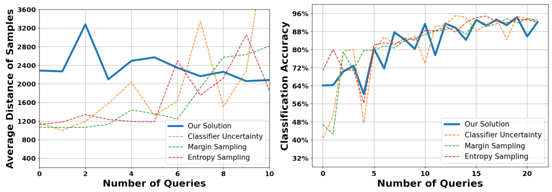
Despite these approaches calculating uncertainty, they don't use augmentation. Thus, and given how common they are for uncertainty sampling [13], they form our baselines.

Transformations. Our approach can function with any augmentations Q. In our experiments, we add Gaussian noise to the input signal, shift it forwards/backwards along the temporal dimension, randomly reverse it, and flip its polarity. In these experiments, a total of $|Q| = 11$ augmentations is available to the active learner, as each augmentation is used up to three times with different parameters. Coupled with the non-augmented version of each sample, the system has 12 options to pick from at each iteration to calculate uncertainty.

The decision to use these augmentations results from a thorough study of prior literature on EEG and ECG signals. For instance, despite many options being available to add noise to EEG signals, the fact that they have strong randomness and non-stationarity meant that we could not use solutions that add local noises like Poisson, Salt, or Pepper [14]. As such, we opted to rely on adding Gaussian noise, which does not locally affect EEG signals' features. Various studies on EEG and ECG signal processing use Gaussian noise to enlarge the size of their dataset and avoid issues like overfitting [15], [16]. Reversing the time series was also helpful for our use case where the samples are considered independently with no sliding window, as it has the same effect as if the input signal is delayed, turning the time t into $-t$ and, thus, still allowing the network to learn useful relationships. Scaling and vertical and horizontal flipping to augment the ECG signals were further identified as optimal augmentations for our task, too [17].

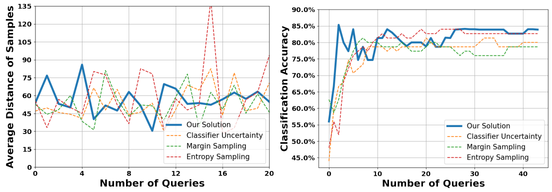
Setup. For our experiments, we use modAL 0.4.1 [18] as an AL tool and TensorFlow Keras 2.8. For our models, we rely on a CNN architecture with three 1D convolutions, max pooling, and a set of dense and flattened layers for our EEG experiments and on a Logistic Regression model for our ECG experiments. We employ a 70/30 and a 75/25 train/test split for the EEG and ECG datasets, respectively. Since in the absence of prior knowledge, the samples for the first query are chosen randomly, we set the value of initial samples to 50 for the EEG dataset and to 5 for the ECG dataset, which is for both one of the lowest possible values in comparison to the size of the respective datasets. Additionally, since the set of augmentations Q may vary depending on the task, our system calculates the entropy at each iteration using a randomly sampled set of 5 out of the $|Q| + 1$ datasets, making the solution optimised for a wider variety of biological signals.

Results. In our experiments, we focused on binary classification with the Epileptic Seizure EEG dataset and on both binary and multi-class classification with the Heart Disease ECG dataset. Although the Heart Disease ECG samples are classified into five classes, we opted to first focus on distinguishing the presence of heart disease (labels 1-4) from the absence (label 0) in a binary fashion, similar to how this dataset is mostly used in practice [11]. After this experiment, we performed multi-class classification with the ECG dataset to examine the behaviour of our proposal in this case too.



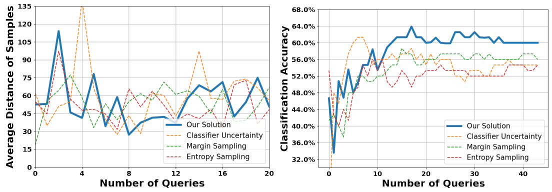
(a) Avg. Euclidean Distance (b) Validation Accuracy

Fig. 1: Epileptic Seizure Recognition Experiment



(a) Avg. Euclidean Distance (b) Validation Accuracy

Fig. 2: Heart Disease Binary Classification Experiment



(a) Avg. Euclidean Distance (b) Validation Accuracy

Fig. 3: Heart Disease Multi-Class Experiment

The results of the query-over-query validation accuracy achieved during our experiments are depicted in Figures 1b, 2b and 3b with a solid line for our proposed solution and with dashed lines for the baselines. The validation accuracy results for the baseline query strategies are derived from the original test dataset, while the results concerning the proposed solution are derived from the mean of the accuracy of the original and all the augmented test datasets.

At first glance, the major conclusion is that capturing the uncertainty of the input samples through data augmentation at each query of the learner stabilises the convergence to the maximum validation accuracy. However, going more in-depth with this analysis, we investigated the first 10 queries more closely, being the ones of the highest interest for determining the diversity of the samples. In all experiments, the accuracy does not substantially change in subsequent queries, so it is clear that the most significant queries are those coming first. In other words, with the early stopping functionality enabled, the learner would stop querying the human or automatic oracle for more labels from that point onwards. Consequently, the system must have managed to label the most informative samples before that point.

Through our study of those initial queries, we found

	Epileptic Seizure	Heart Binary	Heart Multi-Class
Our Solution	2383.0	57.4	54.2
Classifier Uncertainty	1706.5	49.7	54.9
Margin Sampling	1549.2	50.7	51.7
Entropy Sampling	1669.1	56.9	54.3

TABLE I: Avg. Distance of Samples for First 10 Queries

that the average Euclidean distance between the queried points of our solution is notably higher than the one of the baselines. Sample diversity, as represented by the average euclidean distance, is crucial in AL [19], and our framework can capture this diversity better than previous solutions by choosing distant samples in the pool.

As observed in Figure 1a, our approach picks distant samples while the baselines initially ask the user to label samples that are closer together due to their least sophisticated labelling request algorithm. Additionally, in the EEG-based epileptic seizure recognition task, for instance, the average point distance of the first 10 batches of points queried by our solution is 2 383, while the respective one is 1 706, 1 549, and 1 669 for the uncertainty, margin, and entropy sampling baselines, respectively. Consequently, our approach picks sensor signals up to 50% more diverse than the alternatives, making it clear that an augmentation-based uncertainty measure makes better decisions earlier.

Even in the ECG-based heart failure recognition task, our approach picks samples up to 15% more diverse, as indicated by the average euclidean distances recorded amongst the samples picked during the first 10 queries (see Table I). The average distance recorded for our solution is 57.4, the highest number observed for this experiment, showing that the more informative points are labelled first, and the learner can train on signals with a diverse set of features. Finally, our augmentation-based query strategy reaches a final classification accuracy that is 5% higher compared to the alternatives for the multi-class heart failure experiment while also coming imperceptibly second in terms of the average distance of the samples for the first 10 queries issued.

Of course, our multi-class heart failure experiment reaches a lower accuracy than the binary one, but this is expected as the model does not have to distinguish the samples into two classes but five: this is also the case for the baselines. On a further note, the fact that our approach chooses more diverse samples during its earlier iterations will reassure the human or automatic oracle responding to the learner's queries for labels at each iteration. Our proposal ensures that they will be labelling the most informative samples, while the baselines tend to choose less varied samples, reducing the algorithm's generalisation ability at this stage.

Concerning early stopping, setting the variable $E = 0.05$, for instance, means that the learner will stop generating queries when the increase in accuracy over its last three iterations is $\leq 5\%$. Based on our findings, this means that the total queries can be lowered by up to 75% compared to when early stopping is disabled (see Figures 1b, 2b and 3b). This can vary according to the cost of labelling, but it's evident that users will have to label fewer samples.

IV. Conclusion

AL can address the low availability of labelled health datasets by picking the samples to label during the learning process itself. Paramount to its success is the existence of an effective sampling technique that decides on the most uncertain, and thus informative, samples at each iteration.

We propose an AL framework leveraging augmentation to both capture uncertainty during its sampling phase and to get more precise accuracy metrics at test time. Augmentation is commonly used to increase the size of small datasets, but we found that it can be used for better-informed sampling in AL too. Our experiments show that our solution queries the oracle for diverse samples from its first iterations, while also having a lower variance than existing uncertainty sampling approaches. This can increase the adoption of ML in healthcare, as labelling more diverse samples directly translates to cost savings due to the fact that only a few queries would normally be answered by a clinician in the AL loop.

Acknowledgment

This work is supported by Arm, by Nokia Bell Labs, and by EPSRC grant EP/S023046/1 for the EPSRC Centre for Doctoral Training in Sensor Technologies and Applications.

References

[1] B. Li and T. S. Alstrøm, "On Uncertainty Estimation in Active Learning for Image Segmentation," *arXiv*, 2020.

[2] B. Settles and M. Craven, "An Analysis of Active Learning Strategies for Sequence Labeling Tasks," in *EMNLP Proc.*, 2008, p. 1070–1079.

[3] Y. Gal, R. Islam, and Z. Ghahramani, "Deep Bayesian Active Learning with Image Data," *arXiv*, 2017.

[4] S. Hong, H. Ha, J. Kim, and M. Choi, "Deep Active Learning with Augmentation-Based Consistency Estimation," *arXiv*, 2020.

[5] H. Quteineh, S. Samothrakis, and R. Sutcliffe, "Textual Data Augmentation for Efficient Active Learning on Tiny Datasets," Nov 2020.

[6] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren, "Test-Time Augmentation with Uncertainty Estimation for Deep Learning-Based Medical Image Segmentation," *MIDL*, Jul 2018.

[7] A. Ragav and G. K. Gudur, "Bayesian Active Learning for Wearable Stress and Affect Detection," *arXiv*, Dec 2020.

[8] J. Xu, L. Song, J. Y. Xu, G. J. Pottie, and M. van der Schaar, "Personalized Active Learning for Activity Classification Using Wireless Wearable Sensors," *JSTSP*, vol. 10, no. 5, pp. 865–876, Apr 2016.

[9] H. Yu and A. Sano, "Semi-Supervised Learning and Data Augmentation in Wearable-based Momentary Stress Detection," Feb 2022.

[10] Q. Wu and E. Fokoue, "Epileptic Seizure Recognition Data Set," *UCI Machine Learning Repository*, May 2017.

[11] A. Janosi, W. Steinbrunn, M. Pfisterer, R. Detrano, and D. Aha, "Heart Disease Data Set," *UCI Machine Learning Repository*, 1988.

[12] D. D. Lewis and J. Catlett, "Heterogeneous Uncertainty Sampling for Supervised Learning," in *ML Proc. 1994*, 1994, pp. 148–156.

[13] B. Settles, "Active Learning Literature Survey," Univ. of Wisconsin–Madison, Computer Sciences Technical Report 1648, 2009.

[14] F. Wang, S.-h. Zhong, J. Peng, J. Jiang, and Y. Liu, "Data Augmentation for EEG-Based Emotion Recognition with Deep Convolutional Neural Networks," in *MultiMedia Modeling*, 2018, pp. 82–93.

[15] M. Parvan, A. R. Ghiasi, T. Y. Rezaii, and A. Farzamnia, "Transfer Learning based Motor Imagery Classification using Convolutional Neural Networks," in *Proc. of 27th ICEE*, Apr 2019, pp. 1825–1828.

[16] L. Qendro, A. Campbell, P. Lio, and C. Mascolo, "Early Exit Ensembles for Uncertainty Quantification," in *MLHC*, 2021, pp. 181–195.

[17] S. Soltanieh, A. Etemad, and J. Hashemi, "Analysis of Augmentations for Contrastive ECG Representation Learning," *arXiv*, 2022.

[18] T. Danka and P. Horváth, "modAL: A Modular Active Learning Framework for Python," *arXiv*, May 2018.

[19] Y. Yang, Z. Ma, F. Nie, X. Chang, and A. Hauptmann, "Multi-Class Active Learning by Uncertainty Sampling with Diversity Maximization," *Intl. Journal of Computer Vision*, vol. 113, pp. 113–127, 2015.