ActiveSense: A Novel Active Learning Framework for Human Activity Recognition

Farzad Shahabi[1,2], Yang Gao[1,2], and Nabil Alshurafa[1,2]
[1]Department of Computer Science, Northwestern University, Evanston, IL, USA
[2]Department of Preventive Medicine, Northwestern University, Chicago, IL, USA
{farzad.shahabi, yang.gao, nabil}@northwestern.edu

2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops) | 978-1-6654-1647-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/PerComWorkshops53856.2022.9767388

***Abstract*—One of the persistent challenges in building machine-learned models for mobile health applications of fine-grained activity is the generation of accurate annotations with well-defined start/end time labels. Large amounts of unlabeled data exist, and annotation is often labor-intensive and costly. Moreover, it is not clear whether labeling all the data is even necessary to building the most effective machine-learned model. Active learning approaches harness model uncertainty by selecting the most informative samples, reducing the time and effort in labeling unnecessary segments of the data. Model uncertainty, however, is strongly linked to classifier performance, introducing bias in sample selection and impacting model generalizability. In this paper, we propose and study the effects of a new active learning framework on the Necksense dataset which harnesses intrinsic uncertainty as well as model uncertainty by utilizing the Area Under the Margin (AUM) statistic, leading to a significant reduction in the number of samples needed to annotate. We also show that we are able to design a more generalizable model training on 0.15% (n=192 samples) of the data compared to the original model trained on 85% (n=104,681 samples) of the data.**

***Index Terms*—Active Learning, Data Map, Model Uncertainty, Machine Learning**

I. Introduction

With the advent of portable and wearable devices, such as mobile phones and smartwatches, users effortlessly create large amounts of sensory data [1], [2]. As a result, human activity recognition (HAR) systems that detect lifestyle habits like eating [3], brushing [4], and smoking [5] have been actively growing in recent years, and improving their ability to identify and recognize users' actions in a controlled setting [6]. These systems function through activity classifiers that rely on sufficient and representative amounts of accurately labeled data to generalize well in real-world settings.

Data labeling and annotation is a crucial bottleneck to machine learning systems that can lead to costly, task-intensive, error-prone labor work (e.g., labeling Electronic Health Records [7]). In [8], researchers propose a recall approach for the labeling process, which is fraught with burden and memory recollection error. More recently, researchers are using data from wearable video cameras to generate more reliable annotations [9]. However, having humans watch the video footage to label redundant and irrelevant instances containing minimal information for a machine learning (ML) system can waste a lot of time and cost .

Human-in-the-loop machine learning algorithms, such as "active learning" methods, help minimize the data annotation effort by adaptively selecting a subset of samples from the unlabeled set, based on a defined query budget in each querying iteration. These candidate samples are designed to be maximally informative to the classifier being built, which intuitively leads to improvements in classification decision boundaries if accounted for in the classification process [10]. Typically, the classifier is initialized by a "labeled batch" used for active learning model initialization to form the model classifier's initial decision boundaries.

In general, there are three types of selective sampling strategies within active learning: 1) membership query synthesis (MQS), 2) stream-based sampling, and 3) pool-based sampling. In MQS, the algorithm generates data instances from a certain underlying distribution and a labeler is then asked to annotate. In the stream-based sampling scenario, samples are generated sequentially in an online/real-time setting, and a decision needs to be made in real time whether this sample should be labeled. In pool-based sampling, annotators have access to all unlabeled samples, and can assess each before deciding which sample to label [10]. Within each of these methods, uncertainty sampling is one of the most widely adopted query strategies for selecting the most informative sample to label [11]. Under this paradigm, an uncertainty metric needs to be defined to determine data samples or segments for which the model is least certain in its decision and prioritize them for annotation and training. In [12], Thomaz *et al.* build an active learning system that reduces the number of labels needed by applying pool-based sampling strategies and requiring annotation of only 8% of the training data to achieve similar performance. They chose cluster-based query strategies (which select the most informative samples within each cluster) to make sure that sample diversity is accounted for. Huang *et al.* [13] study the cluster structure of unlabeled data and how that affects performance, and show that clustering methods impact active learning classification outcomes. In [14], Aldana *et al.* followed an information theoretic approach to create clusters with the highest entropy levels from the samples, which help the ML models learn the underlying data patterns. In [15], Ashari *et al.* build a memory-aware active learning system, to reduce response time to the query, and use a maximum entropy uncertainty metric with data clustering to measure how certain the model is about its predicted label. Among all the different querying strategies that measure sample uncertainty, maximum entropy is one of

the most commonly used approaches [10].

Nevertheless, defining uncertainty metrics that can capture the samples necessary for building a generalizable model is still a major challenge within active learning. Most approaches calculate sample uncertainty based on the informativeness determined by model uncertainty, which is often captured by variability of the model's confidence in predicting an instance without paying too much attention to the model's confidence value or intrinsic uncertainty, i.e., lower quality signals which are representative of complicated patterns in the signal itself and are often harder to learn. Simply taking model uncertainty into account is known to introduce bias in sample selection [16], which affects model generalizability. Introducing samples that are both model-uncertain and intrinsically uncertain is likely to enable faster model convergence (i.e., yield good results faster). This paper aims to assess the effect of intrinsic uncertainty on the speed of model convergence.

In [17], Swabha *et al.* introduced a model-based tool known as Data Map, which assists in diagnosing easy-to-learn and hard-to-learn samples during the training phase based on training dynamics, allowing us to classify uncertainty as a function of correctness, confidence, and variability of a model. While Data Map allows us to visualize variability and confidence of model predictions, Data Map requires assessing a Deep Neural Network in a supervised-learning fashion, across multiple epochs, and needs ground truth labels (the very labels we aim to optimize in obtaining from the annotators). Needed is a method to retrieve intrinsic uncertainty in a way that is unsupervised, prior to obtaining the ground truth.

Inspired by [18], in which Pleiss *et al.* introduce the concept of Area Under the Margin (AUM) statistics to identify mislabeled samples, we repurpose AUM to determine samples that the model is intrinsically uncertain about. Similar to Data Map, we are able to then assess easy-to-learn and hard-to-learn samples based solely on the sample itself (prior to requiring its ground truth label). Due to Data Map limitations, we adopt the power of the AUM statistic to determine which samples exhibit intrinsic uncertainty. In this paper, we propose and test a novel pool-based active learning framework that combines cluster-based maximum entropy (CME) with AUM (CME+AMU) to determine the effectiveness of the oracle in selecting samples that are both model-uncertain and intrinsically uncertain. We compare our approach to using the AUM-only and CME-only methods. We evaluate this framework on the Necksense free-living dataset [19]. The Necksense free-living dataset is a multi-sensor dataset obtained from 10 participants in a naturalistic setting, with the goal of detecting eating behavior, a known challenge in mobile health. The data used are from a neck-worn device that captures chewing sequences using proximity, ambient light, and inertial measurement unit (IMU) sensors. We make the key following contributions:

- We investigate the AUM statistic's ability to measure hard-to-learn and easy-to-learn samples and its relationship with uncertainty visualization through Data Map.
- We propose and test a novel pool-based active learning model that attempts to capture intrinsic uncertainty (measured by AUM statistic) as well as model uncertainty (measured by cluster-based uncertainty sampling). In addition, we compare it against sampling strategies that use only model uncertainty or only intrinsic uncertainty. We evaluate our proposed model against the best evaluation model obtained from the Necksense free-living dataset.

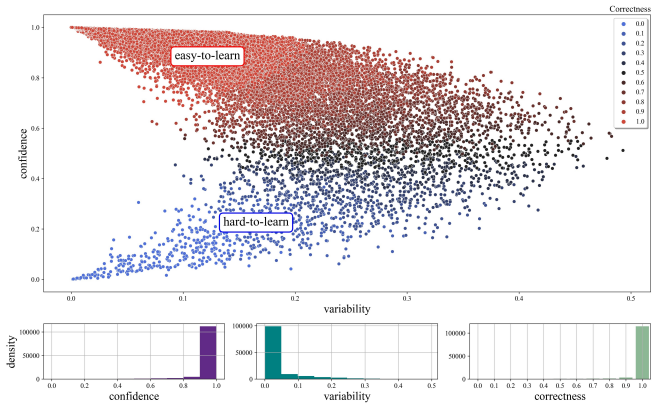

Fig. 1: Uncertainty Visualization through Data Map for the Necksense Dataset. Here we see that the easy-to-learn segments are those with high confidence, low variability, and high correctness in model prediction. Hard-to-learn segments are typically those with lower confidence and correctness in model prediction. As shown in the distribution plots, the majority of the segments are considered easy-to-learn.

II. Methodology

This section begins with a brief introduction about the Necksense dataset (section II-A) followed by data uncertainty visualization using Data Map (section II-B) with the training dynamics (i.e., confidence, variability, and correctness). Then in section II-C, we measure the logit values (pre-softmax output) obtained from the final layer neurons of the MLP network and utilize them to compute AUM values for data instances in the training phase. In section II-D, we explore how the AUM statistic is linked to the Data Map representation of the samples through visual substantiation and show examples of selected easy and hard chewing signals. In section II-E, we introduce our novel framework combining cluster-based maximum entropy and AUM.

A. Necksense Dataset

Necksense free-living dataset is obtained from 10 participants with and without obesity with varying Body Mass Index (BMI) in a naturalistic setting. The data is collected by a multi-sensor necklace, worn around the neck throughout the day, that comprises an IMU sensor, an ambient light sensor, and a proximity sensor to capture motion, leaning forward, and chewing actions occurring during an eating episode. The candidate segments from classification are obtained from the proximity sensor. Therefore, the device captures periodicity in the chewing signal when oriented towards the jaw. More than

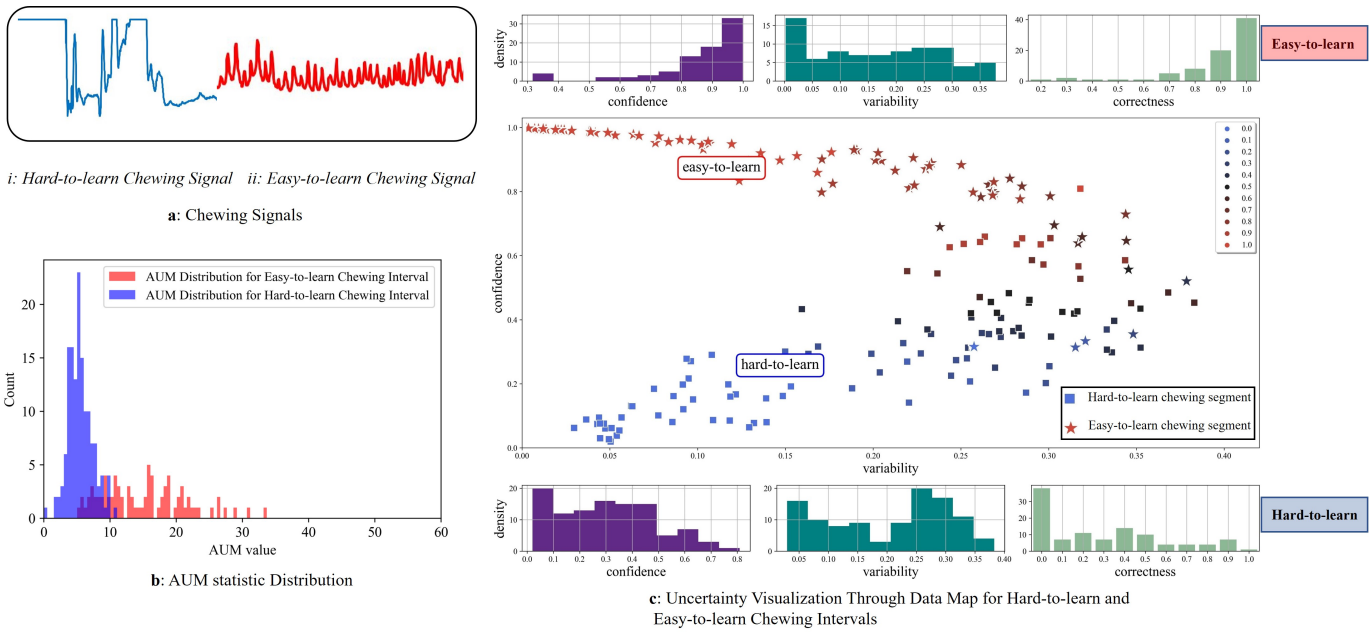

Fig. 2: Uncertainty Visualization by Data Map and AUM statistic distribution for hard and easy chewing intervals. a) Snippets from long chewing signals, one hard-to-learn (blue) and one easy-to-learn (red); b) Distribution of 144 candidate segments extracted from the hard-to-learn signal and 80 segments extracted from the easy-to-learn signal; c) The Data Map visual for the hard-to-learn signal (squares), and easy-to-learn signal (stars).

137 hours of data were acquired, which provide information about chewing, feeding gestures, and neck motion to detect eating episodes. In this study, to evaluate our proposed active learning method, we use the Necksense free-living study, comprising 123155 candidate segments (9.4% of the segments correspond to actual chews, and 90.6% correspond to non-chews).

B. Uncertainty Visualization Through Data Map

Data Map visualization generates three different statistics (confidence, variability, and correctness) obtained from the training phase of a Deep Neural Network across multiple epochs. Consider a training dataset $\mathbf{D}$ of size $\mathbf{N}$ where $\mathbf{D} = \{(\mathbf{x}, \mathbf{y}^*)_\mathbf{i}\}^{\mathbf{N}}_{\mathbf{i=1}}$, $\mathbf{x}$ is the candidate segment (or data instance), and $\mathbf{y}^*$ is the ground truth class label for the data instance. We assume the model predicts a probability distribution across all classes and data instances over $\mathbf{E}$ epochs. Confidence is the first training dynamic, defined as the average model's confidence of a data instance across different epochs, and estimated by:

$$\mu_\mathbf{i} = \frac{\mathbf{1}}{\mathbf{E}} \sum_{\mathbf{e=1}}^{\mathbf{E}} \mathbf{p}^\mathbf{e}_\theta(\mathbf{y_i}^* | \mathbf{x_i}) \tag{1}$$

where $\mathbf{p}^\mathbf{e}_\theta$ is the model's probability distribution over the instance $\mathbf{i}$ at the end of epoch number $\mathbf{e}$. Intuitively, a higher confidence value for an instance means the sample, in terms of learnability, is easier to learn from the model's perspective. Variability is another statistic defining model's confidence variation in predicting one instance with respect to ground truth over epochs estimated by the following:

$$\sigma_i = \sqrt{\frac{\sum_{e=1}^{E} (p^e_\theta(y_i{*}|x_i) - \mu_i)^2}{E}} \tag{2}$$

Correctness is another statistic measured by the number of times the model has predicted a sample correctly across different epochs. Fig 1 demonstrates how Necksense instances are scattered based on the three defined training statistics. It can be observed that the trainset is decomposed into easy-to-learn and hard-to-learn regions. A vast majority of the data instances are easy-to-learn with high confidence and low variability in the top-left of Fig 1. In the right side of Fig 1, there are instances with relatively high variability. The bottom-left of Fig 1 corresponds to the hard-to-learn region, which contains instances with low confidence and low variability. Easy-to-learn samples play an important role in model convergence and learning, while hard-to-learn instances are needed for the model's robustness and out-of-distribution generalizability [17]. It should be noted that uncertainty sampling strategies utilizing the metrics (e.g., entropy or margin in confidence) choose instances with highest variability in confidence from the data during the training phase since it implies the model's indecisiveness about these samples.

C. Area-Under-margin (AUM) for Necksense

We used a four-layer feed forward deep neural network to obtain AUM values and trained it on the Necksense candidate segments. During the training phase, the AUM statistic captures the mean differences between the logit values obtained

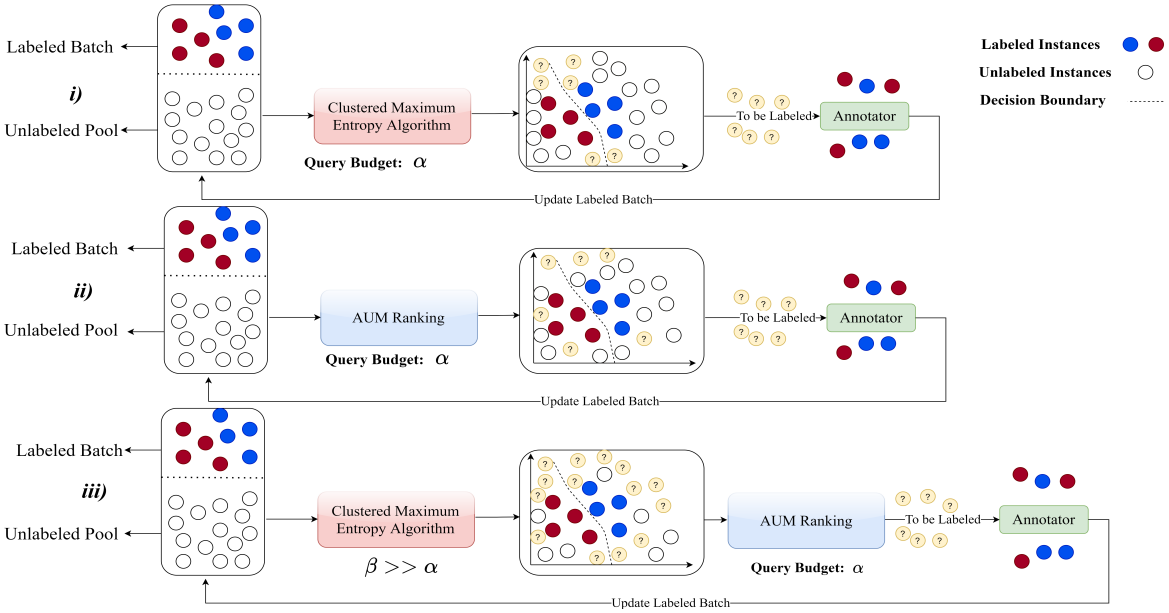

Fig. 3: Part *i* is the traditional active learning pipeline with uncertainty sampling strategy (CME). Part *ii* is the AUM-only method which selects α instances with AUM values. Part *iii* shows the proposed active learning framework (CME+AUM)

from the neurons in the final layer of the deep learning model across different epochs. Originally used for identifying mislabeled instances, we repurpose this statistic and apply it to find the candidate segments that should be labeled based on intrinsic uncertainty. We show its link to Data Map to be able to generate a measure of difficulty and improve our understanding of the segments the oracle selects for labeling. The AUM statistic is formulated as follows:

$$\mathbf{AUM}(\mathbf{x},\mathbf{y}) = \frac{\mathbf{1}}{\mathbf{E}}\sum_{\mathbf{e=1}}^{\mathbf{E}} \mathbf{M}^{(\mathbf{e})}(\mathbf{x},\mathbf{y}) \tag{3}$$

$$\mathbf{M}^{(\mathbf{e})}(\mathbf{x},\mathbf{y}) = \mathbf{z}_{\mathbf{i}}^{(\mathbf{e})}(\mathbf{x}) - \mathbf{max}_{\mathbf{i}\neq\mathbf{j}}\mathbf{z}_{\mathbf{j}}^{(\mathbf{e})}(\mathbf{x}) \tag{4}$$

To elaborate more on the previous formulation, $\mathbf{M}^{(\mathbf{e})}(\mathbf{x},\mathbf{y})$ is the margin at the end of epoch $\mathbf{e}$ which is equal to the difference between logit $\mathbf{z_i}$ of the ground truth subtracted by the largest other assigned logit $\mathbf{max_{i\neq j}z_j}$. Logit value is defined as the pre-softmax output of the final layer of the DNN (Deep Neural Network). This representation allows us to have another measure to identify hard, ambiguous, and easy samples, which can be translated as another metric for training loss (e.g., easy samples have low training loss). The AUM statistic is modified to suit the unsupervised setting formulated by: $\mathbf{AUM}(\mathbf{x},\mathbf{y}) = \frac{\mathbf{1}}{\mathbf{E}}\sum_{\mathbf{e=1}}^{\mathbf{E}} |\mathbf{M}^{(\mathbf{e})}(\mathbf{x},\mathbf{y})|$ where $\mathbf{M}^{(\mathbf{e})}(\mathbf{x},\mathbf{y})$ is defined as in equation (4). Since the ground truth for logit value $\mathbf{z_i}$ is unknown prior to querying, we take the absolute value of the $\mathbf{M}^{(\mathbf{e})}(\mathbf{x},\mathbf{y})$ in the binary classification setting and add them up across epochs.

D. Data Map and the AUM statistic for Hard and Easy Chewing Intervals

This section explores the relationship between Data Map training dynamics and the AUM statistic. Therefore, we can analyze further how visually high/low-quality chewing intervals are translated in terms of the AUM statistic distribution for the candidate's segments and compare it against uncertainty visualization using Data Map. When a model fails to predict one instance correctly, innately, this error originates from two sources. Model uncertainty is referred to as the model's limitation to classify and predict the instance correctly [20]. It is observed that model uncertainty has a strong correlation with the variability statistic. However, intrinsic uncertainty is inherent ambiguity that the sample has on its own, which human annotators can notice. Intrinsic uncertainty has a strong relationship with model confidence, highlighting that when the model is highly confident about an instance, there is a strong agreement between the annotator and the model. To the authors' best knowledge, no uncertainty metric in the literature accounts for both model uncertainty (model variability) and intrinsic uncertainty (model confidence). The AUM statistic provides the desired filtering approach combined with uncertainty sampling using cluster-based maximum entropy to represent and account for the model and intrinsic uncertainty.

We can visually evaluate the quality of the chewing intervals by looking at their raw signals. Figure 2a shows that the hard-to-learn signal on the left reaches proximity sensor saturation at some periods and does not follow a normal chewing pattern where peaks are noticeable and well-separated. Figure 2a shows that the easy-to-learn chewing signal on the right indicates a visually higher quality chewing interval where peaks in the proximity signal are distinct and exhibit periodicity in jaw

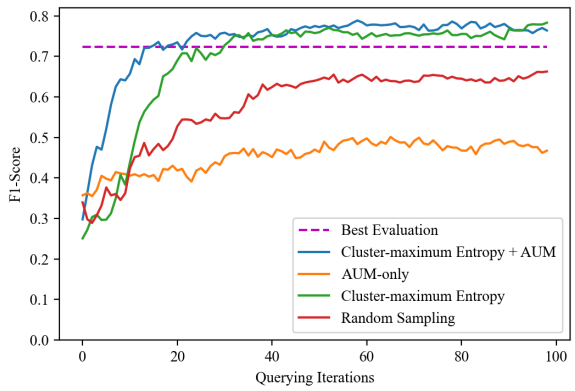

Fig. 4: ActiveSense Result on Necksense dataset

motion. Figure 2b displays the AUM values distribution for the segments of two hard-to-learn and easy-to-learn chewing intervals. The AUM model could determine that the harder-to-learn chewing segments follow a distribution with a lower mean of 5.30 compared with the easy-to-learn interval with a mean of 25.23. Figure 2c displays the uncertainty visualization for the segments of easy-to-learn and hard-to-learn chewing intervals. It can be inferred from Fig. 2b and Fig. 2c that lower values in AUM distribution are translated to lower model confidence and higher sample ambiguity, which we visually confirmed earlier.

E. AUM-based Active Learning Framework

Fig 3 shows the high-level pipeline of the proposed active learning framework. The top diagram *i* shows how a traditional active learning method e.g., cluster-based maximum entropy would work. A cluster-based active learning algorithm would first perform efficient clustering on the unlabeled batch of data to account for sampling diversity from the data clusters which arise from potentially different distributions. It should be noted that the number of data clusters might not match the number of class labels that exist in the dataset. In each querying iteration, the active learning algorithm picks a subset of samples, called the query budget α, from the unlabeled batch of data as labeling candidates nominated to be labeled and identified with respect to the existing class labels by the annotator. The labeled batch then is updated by the newly labeled instances. Samples selected by active learning algorithms (e.g., maximum entropy) tend to account for model uncertainty due to significant changes in decision boundaries in each querying iteration. This is one of the primary reasons that model uncertainty is best translated into the Data Map variability statistic. The performance of the active learning method is evaluated on an unseen labeled test set in each querying iteration. The middle diagram *ii* depicts the AUM-only approach where the algorithm picks α least AUM values from the unlabeled pool as candidates for annotation. This approach only accounts for sample ambiguity in sample selection where the model's confidence for the selected samples is relatively lower. The bottom diagram *iii* displays the framework for the proposed active learning approach. Similar active learning approaches e.g., cluster-based maximum entropy is utilized to consider model uncertainty. However, to cast a wider confidence net, we increase the querying budget to $\beta >> \alpha$. Since Data Map is unable to be employed in an unsupervised setting, we harness AUM statistic, which is originally proposed to identify mislabeled data, and modified it to suit the unsupervised setting. Therefore, an AUM filter is added to select the α samples with least AUM values (considered harder-to-learn with lower confidence).

III. Evaluation and Results

A. Evaluation

We evaluate our active learning framework on the Necksense free-living study data. We randomly select 85% of the data as the unlabeled batch and 15% of the data as the unseen labeled set in a participant independent fashion. We assess the performance of the proposed model against random sampling, AUM-only, and cluster-based maximum entropy methods. We select a labeled subset of 0.1% (set empirically) of all the instances for model initialization and a query budget of six [15] for all methods. For cluster-based maximum entropy, we cluster the unlabeled batch using k-means with the same initialization of clusters across all algorithms. Based on the Silhouette metric [21], which captures cluster separation quality, we set k=6 clusters. During each querying iteration, one sample with highest entropy is selected from each data cluster, resulting in six samples selected for annotation. In our proposed model, we set $\beta = 120$, which selects 20 samples with highest entropy from each cluster. The selected samples are filtered and ranked with respect to their AUM values. At the end, we select 6 samples with the least AUM value. In order to get AUM values, we train a feed-forward multi-layer perceptron with 4 hidden layers (100 neurons in each layer) for fast computation and low complexity in each query iteration. Stochastic gradient descent method for weight optimization is used, and the network is trained on the labeled batch and evaluated on 120 selected samples chosen by maximum entropy method for 100 epochs in each iteration. The pre-softmax values are retrieved at the end of each epoch from the final layer. For the AUM-only approach, we train the network on the labeled batch and retrieve and rank the AUM values for the unlabeled pool in each querying iteration. The random sampling method, also named passive learning, randomly selects unlabeled instances in each querying iteration and updates the labeled batch. The best evaluation shows the results which utilize the whole training set. We use the XGBoost [22] as the optimal classifier used for the Necksense dataset. The results are aggregated and shown at per-second level, where overlapping candidate segments are combined to determine whether each second is a chew or not.

B. Results

Fig 4 shows the results for CME-only, AUM-only, CME+AUM, and random sampling, and compares them with the best evaluation used in Necksense. We can see that our proposed method outperforms other approaches and reaches the best evaluation performance after 12 query iterations, which is equivalent to 0.18% of the unlabeled data. However, cluster-based maximum entropy (CME) reaches the performance of best evaluation after 30 iterations equivalent to 0.29%, performing better than random sampling and AUM-only methods after 10 iterations. The AUM-only approach shows weak performance when it comes to model convergence due to selecting and training primarily on hard-to-learn samples. Also, since the sample selection process is not diverse in the AUM-only method, the algorithm might become biased and keep sampling from limited regions of the data sub-spaces. It is noted that at some points, active learning approaches outperform the best evaluation. Therefore, it highlights that the right composition of model uncertainty and intrinsic uncertainty needs to be considered in order to ensure proper representation of the underlying data distribution to assist with model generalizability. This signifies the fact that adding trivial or redundant samples might actually lower the model's performance. Further analysis and quantification of this is needed on other datasets to confirm.

IV. Limitations and Conclusion

In this paper, we propose a new active learning framework which harnesses intrinsic uncertainty as well as model uncertainty and is tested on a data from a neck-worn sensor worn in the real-world. We show the proposed method achieves best evaluation performance when using 0.15% of the data compared to 85% of the data in the best evaluation method. We also show how the AUM statistic, originally designed to detect mislabeled data, can be repurposed to capture segments that are easy and hard to learn without the need for ground truth labels. The main limitation of the proposed method is the need to retrain the neural network during each query iteration. Future research should look into ways of reducing training time and test our proposed method on other real-world datasets, while assessing its benefits in reducing the time and cost of generating annotations.

V. Acknowledgments

This material is based upon work supported by the National Science Foundation (NSF) under award number CNS1915847. We would also like to acknowledge support by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) under award numbers K25DK113242 and R03DK127128, and National Institute of Biomedical Imaging and Bioengineering (NIBIB) under award number R21EB030305. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the National Institutes of Health.

References

[1] W. Yu, F. Liang, X. He, W. G. Hatcher, C. Lu, J. Lin, and X. Yang, "A survey on the edge computing for the internet of things," *IEEE access*, vol. 6, pp. 6900–6919, 2017.

[2] S. Zhang, Y. Li, S. Zhang, F. Shahabi, S. Xia, Y. Deng, and N. Alshurafa, "Deep learning in human activity recognition with wearable sensors: A review on advances," *arXiv preprint arXiv:2111.00418*, 2021.

[3] S. Zhang, D. Nguyen, Z. King, J. Pradeep, and N. Alshurafa, "Habits necklace: A neck-worn sensor that captures eating related behavior and more," in *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, 2018, pp. 484–487.

[4] S. Akther, N. Saleheen, M. Saha, V. Shetty, and S. Kumar, "mteeth: Identifying brushing teeth surfaces using wrist-worn inertial sensors," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 2, pp. 1–25, 2021.

[5] R. Alharbi, B. Spring, and N. Alshurafa, "Measuring smoking topography in natural settings using non-contact passive wearable sensors," in *ANNALS OF BEHAVIORAL MEDICINE*, vol. 54. OXFORD UNIV PRESS INC JOURNALS DEPT, 2001 EVANS RD, CARY, NC 27513 USA, 2020, pp. S580–S580.

[6] E. Kim, S. Helal, and D. Cook, "Human activity recognition and pattern discovery," *IEEE pervasive computing*, vol. 9, no. 1, pp. 48–53, 2009.

[7] Z. Ji, Q. Wei, A. Franklin, T. Cohen, and H. Xu, "Cost-sensitive active learning for phenotyping of electronic health records," *AMIA Summits on Translational Science Proceedings*, vol. 2019, p. 829, 2019.

[8] Y. Vaizman, K. Ellis, G. Lanckriet, and N. Weibel, "Extrasensory app: Data collection in-the-wild with rich user interface to self-report behavior," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–12.

[9] E. Thomaz, I. Essa, and G. D. Abowd, "A practical approach for recognizing eating moments with wrist-mounted inertial sensing," in *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, 2015, pp. 1029–1040.

[10] B. Settles, "Active learning literature survey," 2009.

[11] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in *SIGIR'94*. Springer, 1994, pp. 3–12.

[12] R. Adaimi and E. Thomaz, "Leveraging active learning and conditional mutual information to minimize data annotation in human activity recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 3, pp. 1–23, 2019.

[13] S.-J. Huang, R. Jin, and Z.-H. Zhou, "Active learning by querying informative and representative examples," *Advances in neural information processing systems*, vol. 23, pp. 892–900, 2010.

[14] E. Aldana-Bobadilla and A. Kuri-Morales, "A clustering method based on the maximum entropy principle," *Entropy*, vol. 17, no. 1, pp. 151–180, 2015.

[15] Z. E. Ashari, N. S. Chaytor, D. J. Cook, and H. Ghasemzadeh, "Memory-aware active learning in mobile sensing systems," *IEEE Transactions on Mobile Computing*, 2020.

[16] D. Kottke, J. Schellinger, D. Huseljic, and B. Sick, "Limitations of assessing active learning performance at runtime," *arXiv preprint arXiv:1901.10338*, 2019.

[17] S. Swayamdipta, R. Schwartz, N. Lourie, Y. Wang, H. Hajishirzi, N. A. Smith, and Y. Choi, "Dataset cartography: Mapping and diagnosing datasets with training dynamics," *arXiv preprint arXiv:2009.10795*, 2020.

[18] G. Pleiss, T. Zhang, E. R. Elenberg, and K. Q. Weinberger, "Identifying mislabeled data using the area under the margin ranking," *arXiv preprint arXiv:2001.10528*, 2020.

[19] S. Zhang, Y. Zhao, D. T. Nguyen, R. Xu, S. Sen, J. Hester, and N. Alshurafa, "Necksense: A multi-sensor necklace for detecting eating activities in free-living conditions," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 2, pp. 1–26, 2020.

[20] Y. Gal, "Uncertainty in deep learning," 2016.

[21] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[22] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.