The Challenge of Metrics in Automated Dietary Monitoring as Analysis Transitions from Small Data to Big Data

Surya Sharma
Electrical and Computer Engineering
Clemson University
Clemson, USA
eating@suryasharma.com

Adam Hoover
Electrical and Computer Engineering
Clemson University
Clemson, USA
ahoover@clemson.edu

2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 978-1-7281-6215-7/20/$31.00 ©2020 IEEE DOI: 10.1109/BIBM49941.2020.9313465

***Abstract*—Many works in the field of automated dietary monitoring (ADM) have analyzed small data sets consisting of <10 subjects and <20 meals. This is often the first step in researching new sensors or body positions for detecting consumption. Metrics tend to focus on within-meal accuracy by quantifying physiological event detection (bites, chews, swallows). As analysis shifts to larger datasets containing many days of data from everyday life and researchers build methods that can be used in everyday life, it becomes equally important to quantify the accuracy of how many meals are detected. In small data sets most meals can be detected at least partially. In larger datasets, some meals are missed and false positives occur. In this work we discuss the pros and cons of time-based metrics and episode-level metrics. We demonstrate how class imbalance affects some of the commonly used time metrics, and discuss why episode level metrics need to be reported as the field transitions from small data sets to big data sets.**

***Index Terms*—automated dietary monitoring, eating detecting, metrics, gesture recognition, m-health, wearables, obesity**

I. Introduction

Automated dietary monitoring (ADM) is a field concerned with monitoring eating and energy intake using wearable or environmental sensors. Information from these sensors can be used by nutritionists and clinicians to better understand the dietary intake of their patients or subjects. Many researchers have demonstrated wearable sensors that can detect periods of eating. For example, researchers have demonstrated eye-glasses fitted with cameras [1], motion sensors [2], or EMG sensors [3]. Others have shown necklaces [4], smartwatches [5], chest belts [6], earphones [7], and other proof-of-concepts [8], [9]. While these methods may work towards the same goal, published works in ADM have reported as many as 22 separate metrics and 45 separate outcomes [10], which makes a direct comparison of such works challenging.

One reason for this disparity is that researchers are often collecting their own data sets to evaluate a new idea. We believe that an unintended consequence of this is that evaluation methods are affected by the size of the data sets collected. We also believe that because many ADM researchers work in the fields of computer science and electrical engineering, ADM is often evaluated using metrics that have traditionally been used in classification and segmentation problems. For example, a recent survey found that decision trees (N=16), hidden Markov models (HMM, N=10), random forests (N=19), and support vector machines (SVM, N=21) were commonly used classifiers [11], while some work has used feature selection and Neural Networks (NN) [1], [12] or end-to-end deep learning [13]–[15]. Such classifiers often report metric in terms of samples classified, or values from a confusion matrix [16].

While some authors report outcomes on detecting eating episodes (aka activities, moments, events, periods, meals or snacks), a survey of N=40 works in ADM by Bell et al. showed that most researchers report metrics that are influenced by the amount of time spent in eating or evaluate the number of windows correctly classified [10]. Very few works in the literature discuss episode level recall or the prevalence of falsely detected meals, which is a well known issue plaguing the field.

We believe that as the field transitions to big data, there is a need to discuss what metrics should be published, and why. In this work we discuss the pros and cons of time and episode level metrics. We show how some metrics are affected by the imbalance of eating and non-eating time in data sets, and how this affects comparisons between works. We discuss the importance of episode level metrics as the field transitions towards big data sets and deployment. This work is informed by our experience creating and analyzing the Clemson all-day data set (CAD), the largest wrist tracking based eating activity data set currently known to us. The data set contains 4,680 total hours of data (with 250 hours eating) collected from 351 participants over 354 days. CAD is 10x - 50x larger than previous work [17], and affords us the unique opportunity to present our insights.

II. The effect of data set size

Data sets are often first created to test new sensor modalities or ideas. In our group early experiments tested if wrist motion data could be used to detect bites of food, and then eating [5],

TABLE I
CHANGE IN PRECISION AND RECALL OF MEAL DETECTION WHEN TRANSITIONING FROM CONTROLLED ENVIRONMENTS OR SMALL DATA SETS TO FREE-LIVING OR LARGE DATA SETS

Previous Work	Controlled/Small		Free Living/Large	
	Precision	Recall	Precision	Recall
Thomaz [30]	67	89	65 (-2)	79 (-10)
Mirtchouk [31]	88	87	45 (-43)	85 (-2)
Chun [22]	95	82	78 (-17)	73 (-9)
Zhang [32]	94	90	79 (-15)	77 (-13)
Kyritsis [14]	86	94	46 (-40)	63 (-31)

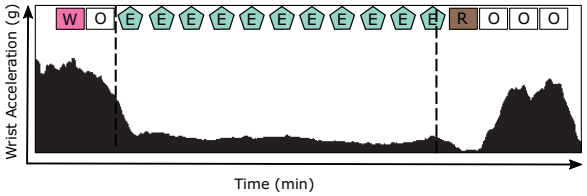

Fig. 1. A 10 minute meal from CAD [17] where the subject ate continuously. Such meals may not be indicative of meals in free-living [17]. Dashed vertical lines represent self-reported start and stop times. Colored blocks represent activity during one minute of time, E = eating, O = other, W = walking and R=resting.

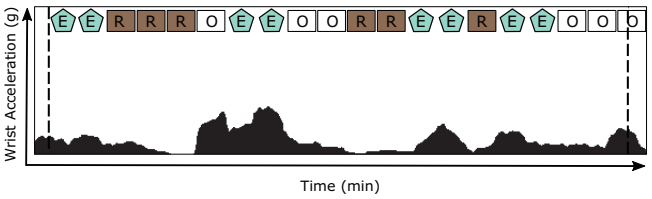

Fig. 2. Another 11 minute meal from CAD [17] where the subject stopped eating food for brief periods of time. Such meals are more common and indicative of meals in free-living [17].

[18], [19]. Other groups have tested other sensor modalities like sensors near the ear [20], [21], the neck [22], [23], or sensors mounted on eye glasses [2], [24]–[26]. These early experiments allow researchers to quickly prototype devices and test hypotheses. Data is often collected by the researchers in laboratory or semi-controlled settings, using video recordings for ground truth annotations. Researchers often use themselves as subjects, or academics and students in the area. Through these experiments and data sets, researchers are able to show evidence that a sensor modality is successful in detecting eating.

Once the feasibility of such a device has been established, researchers may attempt large scale data collections to determine the effectiveness of their ideas in a less controlled environment. Instrument validation is important to the clinical community and requires the accuracy of a tool be independently tested. Validation data sets are collected from participants wearing the sensors all day long in free-living conditions. Authors may use video cameras or self-reports for annotation. For example, Bedri et al. collected in-the-wild data at the Aware home at Georgia Institute of Technology, a sensor instrumented home specially built to support data collection [7], [27]. Similarly, Doulah et al. collected data for a field-like study by inviting participants to live in an observational facility during the day where they were free to conduct their daily business or leave for errands, as long as they returned to eat [28]. Other groups have used self-reports as researchers have shown that cameras influence human behavior, and participants wearing cameras "cannot be themselves" [29]. For example, Zhang et al. collected data for their experiments using EMG eyeglasses with the help of a diary where participants logged eating activities at 1 minute intervals [24], [26]. For collecting data for CAD, we used a button that participants press at the start and end of meals [17].

In the field of ADM there is evidence that experiments on small data sets do not generalize to the broad variability of behavior seen in individuals in everyday life. Table I reports how work has repeatedly shown that methods that work well on small or controlled data sets do not work as well on larger data sets [22], [30]–[32]. Similarly, Doulah et al. showed that eating micro-structure is different in the lab compared to eating in-the-wild [28]. Our work showed that meals in free-living conditions contain much larger amounts of secondary activities, such as walking, talking and resting, compared to meals consumed in an experimental setting [17].

To fully capture behavioral variations in the population, our group has created two large data sets - the Clemson cafeteria data set [33], [34], and the Clemson all-day data set (CAD) [17], [35]. The cafeteria data set contains 518 courses of food consumed in a public cafeteria by 271 subjects. Hand movement gestures in this data set have been annotated through video data from ceiling mounted cameras. The all-day data set contains 4,680 hours (354 days) of wrist tracking data containing 250 hours of eating (1,063 meals). Both these data sets are orders of magnitude larger than previous data sets. We recruited participants from the broader community in the cities around Clemson, SC, balancing for gender, age and ethnicity. Annotation of meals in CAD was captured through self-reported start and end times of meals, provided by participants using a button on the wrist mounted device. Analysis of CAD has shown that participants in free-living conduct a wide variety of secondary activities during eating, such as doing laundry, putting away groceries, walking a dog, playing with a niece, and attending a fair [17]. This was only possible as participants were asked to self-report periods of eating using a button on the accompanying wrist watch, and were not hindered by video cameras.

Figures 1 and 2 show examples of two meals from CAD. The graphs plot wrist motion (Y axis) vs time (X axis). Self-reported start and end times are indicated by vertical dashed lines, while colored boxes show machine detected labels for one minute of data. The meals show how individuals may stop eating, rest or walk for a brief period of time, and resume eating during what is often called a "meal". Should this time be considered as eating or not eating? Evaluating classifiers without a concrete answer to this question creates a challenge. A classifier may detect the time spent walking or resting as

TABLE II

PRECISION AND F_1 SCORE DROP AS THE SIZE OF THE NON-EATING CLASS IN A DATA SET INCREASES. WE USE A TOY EXAMPLE WHERE THE SIZE OF THE EATING CLASS REMAINS FIXED, WHILE THE SIZE OF THE NON-EATING CLASS INCREASES. WE ASSUME THAT THE INCREASE IN TN AND FP IS A LINEAR FUNCTION OF THE SIZE OF THE NON-EATING CLASS.

Ratio	Eating hrs	Non-eating hrs	TP hrs	FN hrs	FP hrs	TN hrs	Precision %	F_1 score %	Recall %	Weight	ACC_W %
1:2	100	200	90	10	50	150	0.64	0.75	0.9	2	0.83
1:4	100	400	90	10	100	300	0.47	0.62	0.9	4	0.83
1:5	100	500	90	10	125	375	0.42	0.57	0.9	5	0.83
1:10	100	1000	90	10	250	750	0.26	0.40	0.9	10	0.83
1:15	100	1500	90	10	375	1125	0.19	0.31	0.9	15	0.83
1:20	100	2000	90	10	500	1500	0.15	0.26	0.9	20	0.83

TABLE III

F_1 SCORE AND RECALL ARE UNDEFINED IF A PARTICIPANT DOES NOT CONSUME FOOD DURING THE DAY, WHILE ACC_W IS NOT.

Ratio	Eating hrs	Non-eating hrs	TP hrs	FN hrs	FP hrs	TN hrs	Precision %	F_1 score %	Recall %	Weight	ACC_W %
–	0	2000	0	0	500	1500	0	–	–	5	0.6

TABLE IV

PRECISION AND F_1 SHOW A DOWNWARD TREND AS DATA SET SIZES INCREASE. TABLE SORTED BY HOURS OF EATING DATA IN DATA SET USED. WE SHOW SELECTED WORK FROM THE FIELD OF DETECTING EATING USING WRIST MOTION TRACKING DATA.

Row	Work	Eating Hours	Total Hours	Precision	Recall	F1-Score	Dataset
1	Kyritsis 2020 [14]	1.6	35	86	94	90	FreeFIC held-out [14]
2	Thomaz 2015 [30]	2	32	67	89	76	Wild-7 [30]
3	Kyritsis 2020 [14]	5	77	88	92	90	FreeFIC [14]
4	Mirtchouk 2017 [31]	12	144	25	83	38	ACE-E [31]
5	Thomaz 2015 [30]	16	422	65	79	71	Wild-Long [30]
6	Kyritsis 2020 [14]	20	250	46	63	53	ACE-E+FL [31]
7	Mirtchouk 2017 [31]	20	254	31	87	46	ACE-E/FL [31]
8	Sharma 2020 [17]	250	4680	14	76	23	CAD [17]

non-eating, but be penalized during evaluation because the time was contained within a self-reported "meal".

The next sections discuss the lessons we learned while cleaning and analyzing CAD in terms of metrics, and the advantages and disadvantages of time and episode level metrics.

III. TIME METRICS

Time metrics are calculated by first labeling each datum of the data set as eating or non-eating. A classifier then tries to replicate these labels, yielding TP = true positives, FP = false positives, TN = true negatives and FN = false negatives. From these, metrics such as precision, recall, F_1 score, balanced accuracy ACC_B, Weighted accuracy ACC_W can be calculated. These metrics evaluate the detection performance of the classifiers. Some work has reported Cohen's Kappa κ and Jaccard index but there is ongoing debate on if and when these metrics should be used [36].

Time metrics are useful for evaluating data collected in the laboratory or data that has video evidence. These metrics evaluate what amount of time in a data set was detected as eating. We believe their adoption is the result of similarities between the problem of eating detection (segmentation of time periods during the day), and problems in image segmentation and detection. These metrics are very useful when evaluating the performance of fine-grained annotated data, which is common in small data sets.

Researchers have used many sensor modalities to measure different physiological events associated with eating, such as bites, chews or swallows. Direct comparison between these methods is not practical. By quantifying the duration of consumption, time metrics make it easier to compare the performance of different sensor modalities and different approaches to detecting eating.

However, time metrics can suffer from a number of problems. The first problem is that they are affected by the balance of eating vs non-eating data in a data set. This makes it hard to compare research in ADM conducted independently on small data sets. We learned that as the size of the data in our non-eating class increased, precision and thus F_1 score dropped, even though the classifier detected as many episodes as it did previously [17]. Table II demonstrates using a simple example. It shows how commonly used metrics such as precision and F_1 score drop as the amount of data representing non-eating activities increases. Our example starts with a data set containing 100 hours of eating and 200 hours of non-eating, a ratio of 1:2. We imagine a classifier that produces the confusion matrix values of TP = 90 h, FN = 10 h, TN = 150 h and FP = 50 h. The resulting values for various metrics are shown in row 1 of table II. Assuming that the relationship between non-eating hours of data and true negatives and false positives is linear, we show how TN and FP change if the amount of non-eating in a data set is increased from a ratio of

TABLE V

TIME METRICS DO NOT CORRELATE WITH EPISODE METRICS. WE USE A TOY EXAMPLE SHOWING BREAKFAST AT 9 AM AND LUNCH AT 12 PM (GT MEALS, CHECKERBOARD BOXES). EACH LONG BLOCK IS 15 MINUTES LONG, AND A SHORT BLOCK IS 7.5 MINUTES LONG. CLASSIFIER A DETECTS 100% OF THE BREAKFAST (TP, GRAY BLOCK) AT 9 AM, BUT MISSES ALL OF LUNCH. CLASSIFIER B DETECTS 50% OF BREAKFAST AND LUNCH. CLASSIFIER C DETECTS 50% OF BREAKFAST AND LUNCH, AND TRIGGERS A 15 MINUTE FP (WHITE BLOCK) AT 10:00 AM. CLASSIFIER D DETECTS 50% OF BREAKFAST AND LUNCH, BUT TRIGGERS TWO 7.5 MINUTE FPS AT 10:00 AND 11:00.

Source	Labeling	Time Metrics		Episode Metrics	
	09:00 10:00 11:00 12:00	TPR %	Precision %	TPR %	Precision %
meals (GT)		-	-	-	-
classifier A		50	100	50	100
classifier B		50	100	100	100
classifier C		50	50	100	66
classifier D		50	50	100	50

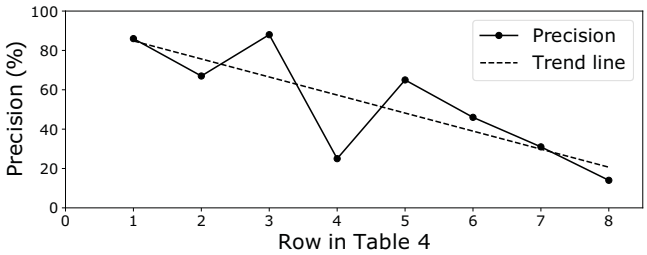

Fig. 3. Precision reported in the literature (Table IV) drops as the data set sizes increase. We use row number on the X axis instead of hours of eating data as 250 hours in the last row skews the plot unreasonably.

1:2 to 1:20 (the amount of hours spent eating in an average American day [37]). This table demonstrates the well-known importance of using weighted accuracy on imbalanced data.

A similar drop in precision and F_1 score can be seen across work in the literature as the size of data sets increases. Table IV lists works that have studied wrist motion tracking to detect eating. The table is sorted by the number of hours of eating captured in the data set. Figure 3 plots precision (y-axis) vs study (x-axis) and shows the trend line. It can be seen that precision and F_1 decrease as the data set size increases. This implies that performance differences may be an artifact of dataset size rather than classifier performance. Researchers may incorrectly conclude that a method works better or worse than another when comparing work on data sets of different sizes.

A second problem with time metrics is that some of them are undefined for data which does not contain any eating. We learned that this can be an issue when analyzing CAD, which contains data from participants who did not report eating during the day [17]. Table III shows how precision, F_1 score and recall are undefined in this case. The ACC_W measure is still useful assuming a standard weight (e.g. 20:1) can be applied.

A third problem with time metrics is that it is debatable what the ground truth class of eating should contain - all the time from the first bite/chew/swallow to the last bite/chew/swallow, or time periods for the bites/chews/swallows? A "meal" may contain multiple consecutive minutes during which no consumption is taking place. The subject may be watching media, talking with family, or doing other things during thie time, but still consider that time part of the "meal". How should this time be evaluated?

A fourth problem with time metrics is that they require video to obtain fine-grained labeling of periods of eating. The use of video cameras to annotate ground truth data increases the effort to create a data set. Previous authors have had to design custom tools and hardware specifically for annotation - for example Doulah et al. use a custom foot pedal to mark events during eating [2], [38], while other groups have used custom software (ChronoViz [39], CafeteriaView [34], PhoneView [40]). On the other hand, video cameras cannot be used in conjunction with large data sets due to their limited capacities and battery lives. Previous work has also shown that cameras affect human behavior negatively [29], and thus individuals participating in data collection may not behave as they would in naturalistic conditions, complicating the efficacy of these metrics further.

IV. EPISODE METRICS

Episode metrics quantify the number of meals/snacks detected by a classifier. They are calculated by examining overlap between periods of time labeled as eating in the ground truth with periods of time detected by a classifier. At the episode level, a TP indicates a detected meal, an FN indicates a missed meal, and a FP indicates an extraneous detection. True negatives (TNs) are undefined.

Time and episode metrics can tell a conflicting story on true positives. Table V shows an example. We show two meals, each 15 minutes long. Breakfast is consumed at 9:00 am and lunch is consumed at 12:00 pm. We show two classifiers (classifier A and classifier B). Classifier A detects the first meal completely, and overlaps breakfast 100%, while classifier B detects both meals, only overlapping 50% of both meals. This results in a time metric TPR of 50%, and a precision of 100% for both the classifiers. On the other hand, when evaluating episode metrics, we see that classifier A has a TPR of 50%, while classifier B has a TPR of 100%. This example shows how time based TPR does not correlate to the TPR of eating episodes. To an end user looking for good detection of meals, classifier B is a better classifier.

Time and episode metrics can also tell a conflicting story on false positives. Table V demonstrates via classifiers C and D. Both classifiers C and D detect 50% of breakfast and lunch

TABLE VI
EATING EPISODE METRICS ARE OFTEN NOT REPORTED IN WORK DETECTING EATING EPISODES USING WRIST TRACKING.

Work	EA	Subjects	TPR (%)	FP/TP	Dataset
Thomaz 2015 [30]	-	7	-	-	Wild-7 [30]
Thomaz 2015 [30]	-	1	-	-	Wild-Long [30]
Kyritsis 2020 [14]	6	6	-	-	FreeFIC held-out [14]
Kyritsis 2020 [14]	17	6	-	-	FreeFIC [14]
Mirtchouk 2017 [31]	31	5	94	-	ACE-E/FL [31]
Mirtchouk 2017 [31]	55	6	87	-	ACE-E [31]
Kyritsis 2020 [14]	86	11	-	-	ACE-E+FL [31]
Dong 2014 [40]	116	43	86	3.8	iPhone [40]
Sharma 2020 [17]	1,063	351	89	5.2	CAD [17]

like classifier B, but they also trigger false positive detections. Classifier C triggers one false positive that is 15 minutes long at 10 am, while classifier D triggers two false positives that are 7.5 minutes long each at 10 am and 11 am. Both classifiers C and D result in a time based TPR and precision of 50%, however classifier C has a higher precision. This example shows how time based precision does not correlate to the precision in detecting eating episodes. Episode metrics show to an end user (such as a nutritionist or clinician) that classifier C is better than classifier D.

The above examples show that episode level metrics are needed to convey the full story, and provide data that is useful to users, especially as the field transitions to real-world use and large data sets. Data sets with 1,000+ meals are more likely to show false positives and variations in eating patterns which can only be described by episode metrics.

If episode level metrics are the goal, the burden of data collection and cleaning can be reduced, as it may be sufficient to only collect annotations and labels in terms of time stamps - "When did you eat" - rather than require fine-grained annotations in the form of minute by minute logging or video camera recordings. Further, the absence of cameras may encourage more naturalistic behavior during eating.

While time metrics help evaluate classifier performance and monitor in-meal behavior such as portion size and eating rate [28], episode level metrics can help monitor between-meal behaviors and daily patterns such as intermittent fasting [41] which has been shown to help with weight loss, as it suggests easy behavioral modifications such as time blocking or not eating in a particular time window.

Table VI shows episode metric performance of recent work in the field of detecting eating episodes using wrist tracking. We report the method (work), and the data set used by the researchers. Works on smaller data sets tend not to report episode level metrics. We hypothesize that this is because in smaller data sets, all meals can be detected partially, and a low number of false positives are triggered, making the reporting of such metrics unwarranted. On the other hand, some works with larger data sets report episode TPR indicating that not all meals could be detected. Similarly, most works do not report the metrics for false positive detections. While Mirtchouk et al. note that their method resulted in false positives that were short and had lengths similar to snacks, they do not report these numbers numerically. We believe it is important to report these numbers as time metric based reporting of false positives does not capture the full performance of a method.

Table VI also introduces a new metric: false positives per true positive (FP/TP). Some previous work has reported false positives per hour (FP/hr) [10], however we believe that FP/TP provides a more intuitive understanding of the episode level performance of a classifier. It indicates the expected number of false alarms that a user could expect relative to the number of true meals detected.

Episode level metrics have limitations. They cannot quantify in-meal eating behavior fully. Methods built towards these metrics are able to provide coarse details such as the start and end of a meal, but not fine-grained details. Further, FPs in episode metrics have a higher impact than FPs in time metrics. A high ratio of FP/TP implies that the method found several large periods of time as false alarms, rather than several short periods of time resembling individual ingestion events.

V. CONCLUSION

In this paper we show how experiments on large data sets in everyday life or free-living conditions tend to be different from experiments on small or laboratory data sets. Our discussion was informed by our experience collecting and analyzing the Clemson all-day data set (CAD), which is 10x-50x larger than other data sets in the field. We believe that small data sets, while pushing the field forward, do not adequately capture eating variability or generalize well to the broad behavior of individuals in everyday life. This can be seen in figure 3 and tables I, II and IV, all of which show how precision and F_1 score drop as methods are evaluated on larger data sets or individuals eating in unconstrained conditions. We believe that big data sets will be needed to develop ADM devices for the real-world, specifically to train their classifiers. This is in addition to validating them, which may need an even *bigger* data set.

VI. DISCUSSION

As research in ADM transitions from small data sets (10 subjects, 1-10 days [30], [42]) to big data sets (250+ subjects, 500+ courses [17], [33], [34]) and from laboratory experiments to everyday life [22], [30]–[32], we believe that episode metrics will be the default for large data sets as opposed to time metrics. This is because time metrics require video for ground truthing and annotation which is a complicated

and time-consuming process. Annotation at this scale becomes impractical as data set sizes increase, and researchers may have to rely on self-reports or coarse annotation. Further, time metrics are hard to explain (and less useful) to end users or clinicians who do not come from engineering or computer science disciplines. We have observed this in colloquial discussions, where individuals or users typically speak in terms of meals per day and describe a meal by a single time, e.g. "I had a snack at 4 pm and dinner at 6:30 pm". Some previous work has agreed with this and discussed reporting eating episodes at coarse resolutions up to an hour [30].

Time metrics also do not work well in the presence of secondary activities which are commonly seen when individuals eat in-the-wild. At least 15%-20% of self-reported meals are activities like resting or walking [17]. If these periods of resting and walking are self-reported as eating, time metrics will not provide appropriate evaluation.

It should be noted that it took us years to collect and clean the Clemson all-day dataset [17], which contains 4,680 hours of data (354 days) and 1,063 self-reported meals, and the Clemson cafeteria data set [33], [34]. Both these data sets are orders of magnitude larger than previous data sets. We recognize the barrier to entry big data creates - new researchers would find it extremely hard to collect such large data sets. Expecting all researchers to collect large data sets would be a disservice to the field of ADM. In our own group, our early experiments evaluated ideas for new sensors on a small set of participants [5], [19], and only recently have we collected and analyzed a larger data set [17], [34]. For this reason through this paper we only provide a word of caution, reminding researchers not to compare the performance of methods on small data sets with those on large data sets, as there is no evidence that methods that work well on small data sets generalize well to large data sets. We hope that more established researchers in ADM attempt to work towards larger data sets and publish results demonstrating the effectiveness of their methods in the larger population.

Finally, we recognize that this paper does not consider the evaluation of energy intake [43], [44] or physiological event detection [10], [45]. However, we believe that these topics have been discussed well in the previous literature and were beyond the scope of this paper.

ACKNOWLEDGEMENT

This work was supported by the National Institutes of Health (NIH) grant #1R01HL118181-01A1.

REFERENCES

[1] J. M. Fontana, M. Farooq, and E. Sazonov, "Automatic ingestion monitor: a novel wearable device for monitoring of ingestive behavior," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 6, pp. 1772–1779, 2014.

[2] A. Doulah, T. Ghosh, D. Hossain, M. H. Imtiaz, and E. Sazonov, "Automatic Ingestion Monitor Version 2 — A Novel Wearable Device for Automatic Food Intake Detection and Passive Capture of Food Images," *IEEE Journal of Biomedical and Health Informatics*, 2020.

[3] R. Zhang, S. Bernhart, and O. Amft, "Diet eyeglasses: Recognising food chewing using emg and smart eyeglasses," in *2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. IEEE, 2016, pp. 7–12.

[4] N. Alshurafa, H. Kalantarian, M. Pourhomayoun, J. J. Liu, S. Sarin, B. Shahbazi, and M. Sarrafzadeh, "Recognition of nutrition intake using time-frequency decomposition in a wearable necklace using a piezoelectric sensor," *IEEE Sensors Journal*, vol. 15, no. 7, pp. 3909–3916, 2015.

[5] Y. Dong, A. Hoover, J. Scisco, and E. Muth, "A new method for measuring meal intake in humans via automated wrist motion tracking," *Applied Psychophysiology and Biofeedback*, vol. 37, no. 3, pp. 205–215, 2012.

[6] B. Dong, S. Biswas, R. Gernhardt, and J. Schlemminger, "A mobile food intake monitoring system based on breathing signal analysis," in *Proceedings of the 8th International Conference on Body Area Networks*, 2013, pp. 165–168.

[7] A. Bedri, R. Li, M. Haynes, R. P. Kosaraju, I. Grover, T. Prioleau, M. Y. Beh, M. Goel, T. Starner, and G. Abowd, "Earbit: using wearable sensors to detect eating episodes in unconstrained environments," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, p. 37, 2017.

[8] S. Bi, T. Wang, N. Tobias, J. Nordrum, S. Wang, G. Halvorsen, S. Sen, R. Peterson, K. Odame, K. Caine *et al.*, "Auracle: Detecting eating episodes with an ear-mounted sensor," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, pp. 1–27, 2018.

[9] K. San Chun, H. Jeong, R. Adaimi, and E. Thomaz, "Eating episode detection with jawbone-mounted inertial sensing," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 4361–4364.

[10] B. M. Bell, R. Alam, N. Alshurafa, E. Thomaz, A. S. Mondol, K. de la Haye, J. A. Stankovic, J. Lach, and D. Spruijt-Metz, "Automatic, wearable-based, in-field eating detection approaches for public health research: a scoping review," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–14, 2020.

[11] H. Heydarian, M. Adam, T. Burrows, C. Collins, and M. E. Rollo, "Assessing eating behaviour using upper limb mounted motion sensors: A systematic review," *Nutrients*, vol. 11, no. 5, p. 1168, 2019.

[12] Y. Gao, N. Zhang, H. Wang, X. Ding, X. Ye, G. Chen, and Y. Cao, "ihear food: eating detection using commodity bluetooth headsets," in *2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*. IEEE, 2016, pp. 163–172.

[13] K. Kyritsis, C. Diou, and A. Delopoulos, "Modeling wrist micromovements to measure in-meal eating behavior from inertial sensor data," *IEEE Journal of Biomedical and Health Informatics*, 2019.

[14] ——, "A data driven end-to-end approach for in-the-wild monitoring of eating behavior using smartwatches," *IEEE Journal of Biomedical and Health Informatics*, 2020.

[15] S. Sharma, "Detecting periods of eating in everyday life by tracking wrist motion—what is a meal?" 2020.

[16] S. Visa, B. Ramsay, A. L. Ralescu, and E. Van Der Knaap, "Confusion matrix-based feature selection." *MAICS*, vol. 710, pp. 120–127, 2011.

[17] S. Sharma, P. Jasper, E. Muth, and A. Hoover, "The impact of walking and resting on wrist motion for automated detection of meals," *accepted to ACM Transactions on Computing for Healthcare*, 2020.

[18] Y. Dong, A. Hoover, and E. Muth, "A device for detecting and counting bites of food taken by a person during eating," in *2009 IEEE International Conference on Bioinformatics and Biomedicine*. IEEE, 2009, pp. 265–268.

[19] Y. Dong, A. Hoover, J. Scisco, and E. Muth, "Detecting eating using a wrist mounted device during normal daily activities," in *Proceedings of the International Conference on Embedded Systems, Cyber-physical Systems, and Applications (ESCS)*. The Steering Committee of The World Congress in Computer Science, 2011, p. 1.

[20] A. Bedri, A. Verlekar, E. Thomaz, V. Avva, and T. Starner, "Detecting mastication: A wearable approach," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 2015, pp. 247–250.

[21] ——, "A wearable system for detecting eating activities with proximity sensors in the outer ear," in *Proceedings of the 2015 ACM International Symposium on Wearable Computers*, 2015, pp. 91–92.

[22] K. S. Chun, S. Bhattacharya, and E. Thomaz, "Detecting eating episodes by tracking jawbone movements with a non-contact wearable sensor,"

Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 2, no. 1, p. 4, 2018.
[23] 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. 37, no. 4, pp. 1–26, 2020.
[24] R. Zhang and O. Amft, "Free-living eating event spotting using emg-monitoring eyeglasses," in *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, 2018, pp. 128–132.
[25] A. Bedri, D. Li, R. Khurana, K. Bhuwalka, and M. Goel, "Fitbyte: Automatic diet monitoring in unconstrained situations using multimodal sensing on eyeglasses," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–12.
[26] R. Zhang and O. Amft, "Retrieval and timing performance of chewing-based eating event detection in wearable sensors," *Sensors*, vol. 20, no. 2, p. 557, 2020.
[27] J. A. Kientz, S. N. Patel, B. Jones, E. Price, E. D. Mynatt, and G. D. Abowd, "The georgia tech aware home," in *CHI'08 Extended Abstracts on Human factors in Computing Systems*, 2008, pp. 3675–3680.
[28] A. Doulah, X. Yang, J. Parton, J. A. Higgins, M. A. McCrory, and E. Sazonov, "The importance of field experiments in testing of sensors for dietary assessment and eating behavior monitoring," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 5759–5762.
[29] R. Alharbi, T. Stump, N. Vafaie, A. Pfammatter, B. Spring, and N. Alshurafa, "I can't be myself: Effects of wearable cameras on the capture of authentic behavior in the wild," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, p. 90, 2018.
[30] 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*. ACM, 2015, pp. 1029–1040.
[31] M. Mirtchouk, D. Lustig, A. Smith, I. Ching, M. Zheng, and S. Kleinberg, "Recognizing eating from body-worn sensors: Combining free-living and laboratory data," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, p. 85, 2017.
[32] R. Zhang and O. Amft, "Monitoring chewing and eating in free-living using smart eyeglasses," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 1, pp. 23–32, 2017.
[33] R. I. Ramos-Garcia, E. R. Muth, J. N. Gowdy, and A. W. Hoover, "Improving the recognition of eating gestures using intergesture sequential dependencies," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 3, pp. 825–831, 2015.
[34] Y. Shen, J. Salley, E. Muth, and A. Hoover, "Assessing the accuracy of a wrist motion tracking method for counting bites across demographic and food variables," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 3, pp. 599–606, 2017.
[35] S. Sharma, P. Jasper, E. Muth, and A. Hoover, "Automatic detection of periods of eating using wrist motion tracking," in *2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*. IEEE, 2016, pp. 362–363.
[36] Wikipedia contributors, "Cohen's kappa - limitations," 2019, [Online; accessed 6-August-2019]. [Online]. Available: https://en.wikipedia.org/wiki/Cohen%27s_kappa#Limitations
[37] United States Department of Labor, "Bureau of labor statistics data for activity: Eating and drinking," 2016. [Online]. Available: https://data.bls.gov/timeseries/TUU10101AA01013237
[38] A. Doulah, M. Farooq, X. Yang, J. Parton, M. A. McCrory, J. A. Higgins, and E. Sazonov, "Meal microstructure characterization from sensor-based food intake detection," *Frontiers in nutrition*, vol. 4, p. 31, 2017.
[39] A. Fouse, N. Weibel, E. Hutchins, and J. D. Hollan, "Chronoviz: a system for supporting navigation of time-coded data," in *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, 2011, pp. 299–304.
[40] Y. Dong, J. Scisco, M. Wilson, E. Muth, and A. Hoover, "Detecting periods of eating during free-living by tracking wrist motion," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 4, pp. 1253–1260, 2014.
[41] M. C. Klempel, C. M. Kroeger, S. Bhutani, J. F. Trepanowski, and K. A. Varady, "Intermittent fasting combined with calorie restriction is effective for weight loss and cardio-protection in obese women," *Nutrition Journal*, vol. 11, no. 1, p. 98, 2012.
[42] M. Farooq, A. Doulah, J. Parton, M. A. McCrory, J. A. Higgins, and E. Sazonov, "Validation of sensor-based food intake detection by multicamera video observation in an unconstrained environment," *Nutrients*, vol. 11, no. 3, p. 609, 2019.
[43] A. Doulah, M. A. Mccrory, J. A. Higgins, and E. Sazonov, "A systematic review of technology-driven methodologies for estimation of energy intake," *IEEE Access*, vol. 7, pp. 49 653–49 668, 2019.
[44] J. N. Salley, A. W. Hoover, M. L. Wilson, and E. R. Muth, "Comparison between human and bite-based methods of estimating caloric intake," *Journal of the Academy of Nutrition and Dietetics*, vol. 116, no. 10, pp. 1568–1577, 2016.
[45] T. Vu, F. Lin, N. Alshurafa, and W. Xu, "Wearable food intake monitoring technologies: A comprehensive review," *Computers*, vol. 6, no. 1, p. 4, 2017.

Surya Sharma received a PhD in computer engineering from Clemson University in 2020. His research interests include healthcare, wearable devices, industry 4.0, computer vision, machine learning and deep learning.

Adam Hoover received a PhD in computer science and engineering from the University of South Florida in 1996. He is currently a professor in the Department of Electrical and Computer Engineering at Clemson University. His research interests include wearable devices, mHealth tools, tracking systems, computer vision and deep learning. He is on the editorial board for the IEEE Journal of Biomedical and Health Informatics.