Fitbit for Chickens? Time Series Data Mining Can Increase the Productivity of Poultry Farms

Conference Paper · August 2020

DOI: 10.1145/3394486.3403385

7 authors, including:


Alireza Abdoli
University of California, Riverside
13 PUBLICATIONS 66 CITATIONS
SEE PROFILE


Sara Alaee
University of Tehran
13 PUBLICATIONS 45 CITATIONS
SEE PROFILE


Amy Murillo
University of California, Riverside
26 PUBLICATIONS 142 CITATIONS
SEE PROFILE


Alec C Gerry
University of California, Riverside
113 PUBLICATIONS 1,456 CITATIONS
SEE PROFILE

Fitbit for Chickens? Time Series Data Mining Can Increase the Productivity of Poultry Farms

Alireza Abdoli[1] Sara Alaee[1] Shima Imani[1] Amy Murillo[2] Alec Gerry[2] Leslie Hickle[3] Eamonn Keogh[1]
[1] Department of Computer Science and Engineering, UC Riverside, CA
[2] Department of Entomology, UC Riverside, CA
[3] Farmsense Inc, Riverside, CA
{aabdo002, salae001, siman003, amy.murillo, alec.gerry}@ucr.edu, leslie@farmsense.io, eamonn@cs.ucr.edu

ABSTRACT
Chickens are the most important poultry species in the world. Globally, industrial-scale production systems account for most of the poultry meat and eggs produced. The welfare of these birds matters for both ethical and economic reasons. From an ethical perspective, poultry have a sufficient degree of awareness to suffer pain if their health is poor, or deprivation if poorly housed. From an economic viewpoint, consumers increasingly value poultry welfare, so better market access can be obtained by producers who demonstrate concern for their flocks. Recent advances in sensor technology has allowed the opportunity to record behavioral patterns in chickens, and several research groups have shown that such data can be exploited to enhance chicken welfare. However, classifying chicken behaviors poses several unique challenges which are not observed in the UCR archive or other classic benchmark collections. In particular, some behaviors are manifested in the *shape* of the subsequences, whereas others only in more abstract *features*. Most algorithms only work well for one such modality. In addition, our data of interest has classes that greatly differ in duration, and are only weakly labeled, again defying the assumptions of the classic benchmark datasets. In this work, we propose a general-purpose framework to robustly learn and classify from datasets exhibiting these issues. While our experience is with fowl, the lessons we have learned may be more generally applicable to real-world datasets in other domains including manufacturing and human health.

CCS CONCEPTS
• Applied computing → Computers on other domains → Agriculture. Computing methodologies → Similarity Detection.

Keywords
Time series, Classification, Similarity search, Poultry welfare.

ACM Reference format:

Alireza Abdoli, Sara Alaee, Shima Imani, Amy C. Murillo, Alec C. Gerry, Leslie Hickle and Eamonn J. Keogh. 2020. Fitbit for Chickens? Time Series Data Mining Can Increase the Productivity of Poultry Farms. In *Proceedings of ACM Conference on Knowledge Discovery and Data Mining (KDD '20), August 22–27, 2020, San Diego, CA, USA. ACM, New York, NY, USA, 9 pages.*

ACM ISBN 978-1-4503-0000-0/18/06...$15.00

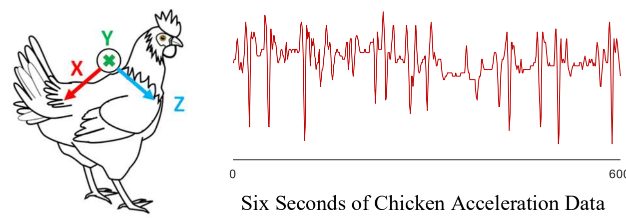

Figure 1: (*left*) A chicken wearing sensor on its back, recording acceleration data (*right*) Six seconds of chicken data (only the X-axis is shown) hints at the complexity of data in this domain.

1. INTRODUCTION
Poultry farming refers to raising domesticated birds (e.g. chickens, turkeys and ducks) to produce meat, eggs and/or feathers. Given the ever-growing world population and the demand for nutritious food, the poultry are farmed in great numbers. There are more chickens in the world than any other bird. Poultry meat and eggs contribute to human nutrition by providing high-quality protein and low levels of fat [21].

In recent years, progress in sensor technologies have enabled practical deployments in increasingly prosaic settings. Recognition of *human* activities has been approached in various ways, among which the most practical and accurate are wearable sensors that are attached to the subject [30]. Such collected sensor data is typically extracted in the form of *time series data*; which can be investigated with a host of data mining techniques to summarize the behaviors of the subject [1]. While there are hundreds of research efforts in time series classification, most studies on time series have only been evaluated on the highly contrived UCR Archive [17]. In this work we show that real-world applications introduce unexpected difficulties and challenges that have been largely "hidden" by the curators of the UCR Archive (in fairness, the most recent version of the archive explicitly recognizes and addresses the issue [17]). The two main issues we address are:

- **The shape vs. feature classification dichotomy:** Most time series classification research efforts solely rely on either *shape-based* or *feature-based* techniques [17][26]. However, as we show in this work there exist problems for which some classes are suited to just one of those paradigms, but some to the other. Thus motivated, we propose a joint shape/feature classification algorithm which outperforms both sole shape and feature-based methods in terms of classification accuracy.
- **Learning from noisy and highly skewed real-world chicken data:** Most time series mining papers in the literature assume all classes are the same length, and the data is strongly labeled (again, simply reflecting what data is available in benchmarks).

Figure 2: (*top*) A poultry farm with six chicken houses in Cullman County, AL USA hints at the scale of poultry production (*bottom*) Chickens in a poultry house.

In contrast, our domain of interest has well-defined behaviors that differ in length by almost an order of magnitude. Because of the two issues above, most algorithms in the literature are either not defined for our practical problem-at-hand, or they perform poorly.

1.1 Poultry Management and Welfare

The United States of America is the world's largest poultry meat producer, with 18 percent of global output, followed by China, Brazil and the Russian Federation. Approximately 70 percent of chickens raised for meat globally are raised in intensive farming systems. This includes the majority of chickens in the US, UK, and Europe, as well as rapidly increasing numbers in developing countries [20].

If poultry are to achieve their genetic potential for meat or egg production, they need an environment that meets their physiological requirements. This includes: (1) a suitable physical environment in terms of temperature, humidity, air movement and the surfaces on which they live; (2) adequate food and water; (3) minimal exposure to disease causing organisms; and (4) avoidance of exposure to stress resulting from the physical and social environment. The factors influencing these are determined largely by housing and management. Consumers increasingly want to be sure that all animals being raised for food are treated with respect and are properly cared for during their lives, and companies involved in raising chickens for food are becoming increasingly responsive to the public's concern. They recognize that they have an ethical obligation to make sure that the animals on their farms are well cared for. The chicken industry has come together on a specific set of expectations that will ensure that the birds they raise are taken care of with the highest standards starting at hatch [19].

1.2 Challenges in Working with Chickens

There are hundreds of studies on *quantifying* human body behaviors with sensors [5]. Such studies typically involve finding discrete well-defined classes of behaviors, and then monitoring data for future occurrences of behaviors. One example is "`step-counting`" to measure compliance with a suggested exercise routine. However, the task of studying behaviors (activity recognition) in chickens is considerably more difficult than humans for the following reasons, as outlined in Table 1 and discussed below:

Table 1: Activity recognition trends in humans vs. chickens.

Activity	Humans	Chickens
Sensor Placement	Flexible	Limited
Individual Variability	Well-Studied	Limited-Studies
Cooperativity	High	Limited to None

- In case of humans, the sensors can be easily placed on the extremities of the limbs (i.e. smart-shoes or smartwatches); However, the placement of sensors on chickens has been primarily restricted to the back of the birds due to sensor limitations and the welfare of animal. This provides only coarse information about the bird's behaviors.
- The variability of human behaviors is well-studied, and it is understood what fraction can be attributed to individual's personality, sex, body type, mood and so forth, versus variability in sensor placement [24]. More importantly, it is known how to account for this variability [25]. However, it is less clear how much variability exists in birds and how we can best account for it [1].
- Creating a dictionary for a human subject is relatively straight-forward. During an explicit training session, behaviors of interest can be "acted out" in a fixed order for a fixed duration of time. For example, one of the most studied human motion time series datasets is the UCR archives `gun-point` [18]. When recording that dataset, the actor's behaviors were cued by a metronome [17], to enable subsequent extraction of the behaviors. Chickens are clearly not as cooperative, and many hours/days of video recordings must be analyzed to create a behavior dictionary. Moreover, it is difficult, even for an experienced avian ethologist, to define precisely where a given behavior begins and ends, thus we must be able to work with "weakly labeled" data. For example, our data may be labeled as "*most of that eight seconds is comprised of dustbathing*" or "*there are about twelve pecks in that ten second snippet*".

In this work, we introduce a novel shape/feature classification algorithm, which can take such weakly labeled data together with some mild constraints "*a preening behavior probably lasts between 0.3 and 1.5 seconds*" and automatically learns to classify the behaviors.

It is important to note that we are not advocating placing sensors on every chicken in the flock. This would clearly be prohibitively expensive and time consuming. Our claim is that by studying the behaviors of a few sentinel birds, we can utilize the learned lessons to improve the welfare for the entire breed. This idea is widely understood, so we will not dwell on it further. See [27] and the references thereof.

The rest of this paper is organized as follows. In Section 2, we review related work. Section 3 discusses data acquisition practices. Section 4 elaborates on definitions and notations, while the proposed shape/feature classification method is discussed in Section 5. An extensive empirical evaluation is conducted in Section 6. We offer conclusions and future work in Section 7.

2. RELATED WORK

As noted above, the many thousands of research efforts on classification of *human* behavior do not bear directly on the task-at-hand, thus we confine our discussion to non-human animals. In recent works [11][23], sensors were used for classification of sheep behaviors; with mounted sensors on ear/collar or leg of the sheep. Similar work has been performed for domestic bovines [8] and goats [26], and for various kinds of wild animals. There has been a work for monitoring of tiny insects [15]. In addition, there has been some work on poultry behaviors using sensors [1]**Error! Reference source not found.**. However, this work is complementary to our efforts. They use only statistical features (mean, entropy, etc.) to quantify periods of general behaviors, such as sleep, stand, walk, etc. [10]. In contrast, Abdoli et al. [1] used the *shape* of the time series to annotate very specific and dynamic behaviors, such as individual instances of a single peck. However, as we will show in a later section, relying solely on *shape* will result in low accuracy, as some chicken behaviors (for example, `preening`) have highly conserved values for some *features*, but are not expressed in well conserved shapes.

3. DATA ACQUISITION AND CLEANSING

All chickens were housed and cared for in accordance with UC Riverside Institutional Animal Care and Use Protocol. Data is collected from chickens by placing the sensor on bird's back, as shown in Figure 1 (*left)*. The sensor is placed on back of the bird to allow for high-quality recording chicken behaviors, while the minimum interference and discomfort.

Figure 3: (*left*) Axivity AX3 data logger sensor used for batch processing mode (*right*) mbientlab Meta Tracker (MTR) sensor used for online real-time processing mode.

We initially collected data using the Axivity AX3 data logger sensor [2], shown in Figure 3 (*left*), weighing about 11 grams and configured with 100 Hz sampling frequency and +/- 8g sensitivity which allows for two weeks of continuous data collection with the battery fully charged. As a data logger, the Axivity AX3 saves all acquired data to an internal memory device which can only be accessed after the sensors are removed from the chickens. This requires considerable human time and effort. On the other hand, diseases and abnormalities might spread quickly throughout the chickens in the houses. Therefore, in order to provide instantaneous access to the chicken data we utilized Mbientlab MetaTracker sensors, shown in Figure 3 (*right*), which are capable of wireless real-time data transmission through the Bluetooth technology. In Figure 4 we show the logic model, in for our operations. The workflow, as outlined goes from step 1 to 7.

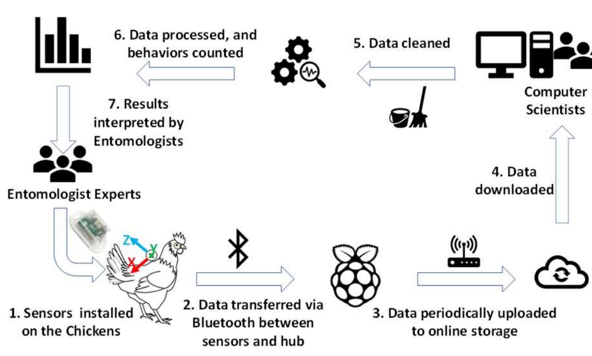

Figure 4: The workflow to study chicken behavior and welfare in online processing mode.

Note that the workflow was supposed to be linear, steps 1 to 7. However, as with most real-word deployments, as we approached step 7, the computer scientists in our group saw opportunities to improve the quality of the data, and the entomologists saw opportunities to monitor additional behaviors. Thus, we cycled through this workflow several times on new cohorts of birds before converging on a commercially deployable system. This also might be a good place to note why *entomologists*, and not directly *bird experts*, are the main customers of our system. By far, the most important stressor of chickens, and most important vectors of chicken diseases are insects and mites[1] [29][27].

4. DEFINITIONS AND NOTATIONS

We begin by providing definitions and notation to be used throughout the paper.

4.1 Definitions

Definition 1: A *time series* T is a sequence of real-valued numbers t_i: $T = [t_1, t_2, \ldots, t_n]$ where n is the length of T.

We are typically not interested in the global properties of time series, but in the similarity between local subsequences:

Definition 2: A *subsequence* $T_{i,m}$ of a time series T is a continuous subset of the values from T of length m starting at position i. $T_{i,m} = [t_i, t_{i+1}, \ldots, t_{i+m-1}]$ where $1 \leq i \leq n - m + 1$.

We will classify the time series using a combination of feature and shape measures.

Definition 3: The *nearest neighbor (NN) classifier* for a time series T is an algorithm that for each query Q (i.e. *subsequence*) finds its nearest neighbor in T and assigns Q that neighbor's class.

In case of a shape-based classifier, the nearest neighbor is defined using either the Euclidean distance (ED) or the Dynamic Time Warping (DTW) as the distance measure. The distance between a query (i.e. an unlabeled exemplar) and all the other subsequences in the time series is stored in an ordered vector called the *distance profile* (D).

Definition 4: A distance profile D is a vector of Euclidean distances between a given query and each subsequence in the time series.

To avoid extracting redundant subsequence matches in *distance profile* we must be aware of *trivial matches*:

Definition 5: Given a time series $\boldsymbol{T}$, containing a subsequence $\boldsymbol{T_{p,m}}$, if $\boldsymbol{T_{p,m}}$ scores highly on any scoring function, then $\boldsymbol{T_{j,m}}$ which $\boldsymbol{j \epsilon [-m/2, m/2]}$ will almost certainly score high on the same function. These spurious high scoring subsequences are *trivial matches*.

To avoid counting trivial matches when finding matches to a query, we discard some of the patterns using the concept of an *exclusion zone,* a standard practice [6].

While distance is the measure of similarity between two subsequences in a shape-based classifier, the feature-based classifier finds similarity based on a set of *features*.

Definition 6: A *time series* T is a sequence of real-valued numbers t_i: $T = [t_1, t_2, \ldots, t_n]$ where n is the length of T.

[1] Mites are *arachnids*, not insects. But as a practical matter they fall under the purview of entomologists.

Given time series T and function X, the *feature vector F* will be the value of X for every subsequence of T. Each feature vector F corresponds to a measurable property of the time series.

5. METHODOLOGY

The problem of time series classification has been around for decades [17]. There are two main approaches for time series classification in the literature, namely, *shape*-based classification [32][33][34] and *feature*-based classification [9]. Shape-based classification determines the best class according to a distance measure (e.g. Euclidean distance). Feature-based classification, on the other hand, finds the best class according to the set of features defined for the time series. These features measure properties of the time series (e.g. autocorrelation, complexity [16], etc.) [31].

5.1 Shape versus Feature Classification

As noted above, researchers typically use *either* shape-based or feature-based techniques for classification of time series. However, deciding which classifier is the best for the problem at hand is a difficult problem to solve. There are three obvious possibilities.

- One of the two techniques dominates for *all* problems. However, an inspection of the literature, or a little introspection convinces us otherwise. There are clearly problems for which one of the two approaches is much better on all classes [31].
- On a problem-by-problem basis, one of the techniques dominates. For example, perhaps for electrocardiograms *shape* is the best, but for electroencephalograms *feature* is the best. This seems to be the implicit assumption of most of the community [7].

However, there is a third possibility that seems to have largely escaped the attention of the community:

- On a single problem, it might be possible that one of the techniques is better for one subset of the classes, and the other technique is better for another subset.

Given this third possibility, it is clear that neither of the two techniques will dominate for some problems, but that some "combination" of both might be the best.

We have observed that the task of classifying chicken behavior from accelerometer data is unlikely to yield to a single modality of classification. The `pecking` behavior has highly conserved *shape*. Whereas the `preening` and `dustbathing` behaviors do not have a stereotypical shape, but are recognizable using several *features*, including "complexity" and "frequency" [7].

5.2 Chicken Behavior Classification

From literature reviews [28][29], and conversations with poultry experts, we expect that the following behaviors correlate with poultry health:

- **Feeding/pecking:** bringing the beak to the ground and retrieving a morsel of food.
- **Preening:** grooming of the feathers using the beak.
- **Dustbathing:** sitting or rolling in the dirt.

In particular, birds infected with ectoparasites are expected to do more preening/dustbathing [29][27]. Table 2 details the combination of shape and features we used for classifying `pecking`, `preening` and `dustbathing` behaviors. In case of the `pecking` behavior the shape works well, however, to distinguish and ward off noisy subsequences that look like pecks we further applied standard deviation and complexity features. The dustbathing behavior (likewise `preening`) does not have a well-conserved shape. So, we utilized features to classify their instances. Given the intense nature of the `preening` and `dustbathing` behaviors the "complexity"[2] feature is useful to distinguish the aforesaid behaviors from other behaviors [14]. Moreover, to further classify between `preening` and `dustbathing` instances we applied the power spectral density feature which provides discrimination to distinguish these behaviors [7]. Note that although we used the complexity feature for the `pecking`, `preening` and `dustbathing`, we learn different thresholds for each behavior class.

As can be seen in Table 2, the length for pecking behavior instances is assumed to be constant, while the length for the instances of preening and dustbathing instances can vary from 0.3 of a second to 8 seconds. The shape-based classification will not work well for `preening` and `dustbathing` and we should utilize features (i.e. complexity and power spectral density).

Table 2: Shape-Feature classification for chicken behaviors.

Behavior	Length	Shape	Feature
`Pecking`	Constant	✓ (Euclidean distance)	✓ (standard deviation, complexity)
`Preening`	Variable	X	✓ (complexity, power spectral density)
`Dustbathing`	Variable	X	✓ (complexity, power spectral density)

We are now in a position to discuss the proposed combined shape-feature algorithm.

5.3 Proposed Approach

The proposed approach starts with the following steps:

1. ***Propose a set of useful features and shapes***, allowing the possibility that different classes may best be distinguished with different subsets of shapes and features.
2. ***Calculate the shape-vectors and feature-vectors***, given the identified set.
3. ***Learn relevant thresholds for every class*** (given the user's class-dependent tolerance for false positives/false negatives)
4. ***Deploy the model to classify data***

The algorithms outlined above has two subroutines outlined in Table 3. Individual elements are motivated and explained in the following subsections.

5.3.1 Propose a set of useful features and shapes

The user will provide a set of shapes and features for every class. These suggested shapes and features can come from domain experts or visual inspection of the time series by the user. The choice of which features to use is beyond the scope of this paper. Fulcher et al. [7] presented a library of over 9,000 features that can be used to quantify time series properties including the classic features such as *min*, *max*, *standard deviation*, *periodicity* etc.

---

[2] The term "complexity" is overloaded, even in narrow context of time series features. Here we mean *complexity* feature introduced in [14][16].

5.3.2 *Calculate shape and feature-vectors*

We are given a time series, the set of classes describing the time series and the set of *shapes* and *features* describing each individual class. The classification process involves calculating shape-vector and feature-vector(s), based on the proposed set of shapes and features, for every class.

5.3.3 *Set relevant thresholds for every class*

For each class, we need a threshold that best defines our relative tolerance for the false positives vs. false negatives. The thresholds can be either manually adjusted by the user (static) or learned through a feedback loop (dynamic), assuming ground truth labels are available.

In the dynamic case, the user may inspect the results produced by multiple runs of the algorithm and choose the threshold setting corresponding to the most desired point on the ROC curve. For each class, we need a threshold that best defines our relative tolerance for the false positives vs. false negatives. The thresholds can be either manually adjusted by the user (static) or learned through a feedback loop (dynamic), assuming ground truth labels are available. In the dynamic case, the user may inspect the results produced by multiple runs of the algorithm and choose the threshold setting corresponding to the most desired point on the ROC curve.

Table 3 and Table 4 show calculation of shape and feature vectors for pecking, preening and dustbathing behaviors.

The typical shape of an instance of pecking behavior is shown in Figure 8. Note that, in principle, a single behavior could have two or more possible instantiations; just like the number four has two written versions, closed '4' and open '4', which are semantically identical. We call such behaviors a polymorphic behavior. Our algorithm allows for having multiple instances of shapes for the same class, so we can account for instances which belong to the same class but differ in shape.

Table 3: Algorithm for computing shape-based profile and feature-vectors for pecking behavior.

```
Input:  T, Train data
        Q, Query
        L, Label
Output: Pecking_Locations, the location of pecking
        instances in T
1   // Calculating distance profile, see [3]
2   [distVector,ind] ← sort(distanceProfile(T, Q))
3   // Calculate standard deviation (std) feature
4   [stdVector] ← std(T, 100)
5   [stdVector] ← applyThreshold(stdVector, L)
6   // Calculate complexity feature
7   [cplxVector] ← sum(abs(diff(T)), 100) /
    stdVector // see [14] for complexity
8   [cmplxVector] ← applyThreshold(cmplxVector, L)
9   // Combine shape-vector and feature-vectors
10  Pecking_Locations = distVector & stdVector &
    complexityVector
```

In line 2, the distance profile (Definition 4) of the shape-based query for pecking behavior is calculated and sorted in ascending order. Essentially, the subsequences most similar to the query Q are prioritized as likely candidates.

In [1] the authors attempted to classify just with shape, however, given technical limitations of time series data (noisy data and etc.) and fast-paced nature of pecking behavior (lasting for one fourth of a second) the authors had to set a conservative threshold to avoid misclassifications of real chicken pecks and noisy data that look similar to pecks. In this study, we mitigate this issue with combined shape-feature classification. Standard deviation and complexity are the two suggested features to distinguish a real peck from similar noisy data.

In line 4, we calculate the standard deviation over the time series T with a sliding window of one second (data was recorded at 100Hz). Next, in line 5, `stdVector` is passed to `applyThreshold` function which tests different thresholds given the label L and finds the best threshold for the feature vector.

Similarly, in line 7, the complexity feature is calculated over the time series with a sliding window of one second. `cmplxVector` is passed to the `applyThreshold` function to obtain the threshold yielding the best classification results for the feature complexity.

Given the distance profile (`distVector`), standard deviation vector (`stdVector`) and complexity vector (`cmplxVector`), in line 10, we combine the shape-feature vectors using logical operations so that noisy data are filtered by standard deviation and complexity features.

Table 4 shows the steps for classification of preening and dustbathing behaviors using the complexity and power spectral density features. In line 2 the complexity feature is calculated over the time series. In line 3 `cmplxVector` and class labels L are passed to `applyThreshold` so that the best threshold for the preening and dustbathing classes are found.

Table 4: Algorithm for computing feature-vectors for preening and dustbathing behaviors.

```
Input:  T, Train data
        L, Label
Output: Dustbathing_Locations, Preening_Locations, the
location of dustbathing and preening instances in T
1   // Calculate complexity feature
2   [cmplxVector] ← sum(abs(diff(T)), 100) /
    stdVector // see [14] for complexity
3   [cmplxVector] ←    applyThreshold(cmplxVector,L)
4   for each region in cmplxVector
5   // Calculate power spectral density for every region
6    [plmbVector] ← plomb(region, 100)
7    // Calculate AUC of plmbVector
8    [plmbVector] ← AUC(plmbVector)
9    Dustbathing_Locations = (plmbVector < thr_plmb)
10   Preening_Locations = (plmbVector > thr_plmb)
11  end
```

Figure 5 depicts the complexity feature vector which as demonstrated in the figure helps with distinguishing preening and dustbathing behaviors from other classes.

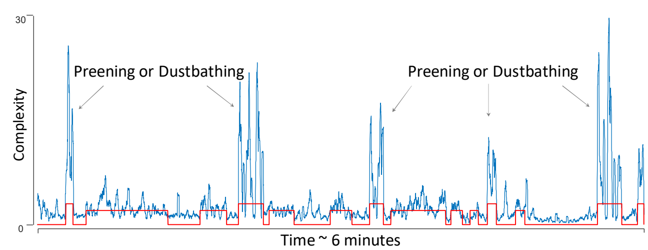

Figure 5: The complexity feature vector for distinguishing preening and dustbathing instances from other classes. Note that the feature (blue) tends to peak at location marked as positive with the vector of class labels (red).

Returning to Table 4, following the calculation of complexity feature vector, we identify the regions corresponding to either preening or dustbathing behavior.

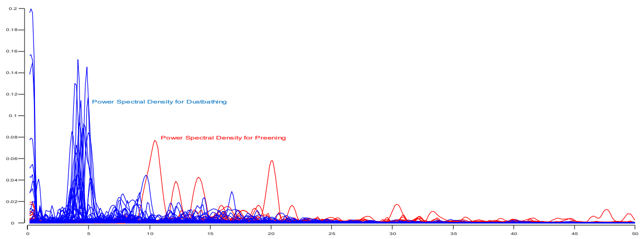

Figure 6: The power spectral density feature vector for distinguishing between preening and dustbathing instances.

In line 4, we initiate a for loop and for every distinguished region from the complexity feature we calculate the power spectral density function over that region. The main benefit of the spectral power density is to differentiate between `preening` and `dustbathing` instances. Next, we account for the area under the curve (AUC) of the power spectral density of every instance, and the value of the AUC helps to classify between preening and dustbathing instances.

Following discussion of the shape-feature based classification algorithm we will initially provide extensive evaluation results for a labeled chicken dataset. Next, we will demonstrate that the proposed classification algorithm generalizes well to unforeseen chickens for which we do not have any labels of any kind.

6. EXPERIMENTAL EVALUATION

To ensure that our experiments are reproducible, we have built a supporting website [35]; which contains all data, code and raw spreadsheets for the results.

Initially we present the results for the training/test dataset, shown in Figure 7, and then we proceed to a case study of utilizing the shape-feature based classification algorithm to recognize healthy and unhealthy chickens based on the behavior count throughout the day.

6.1 Performance Evaluation

The original dataset is split into mutually exclusive training and test datasets, as illustrated in Figure 7. For some fraction of the time series data collected, a video camera also recorded the chicken activity (~ 30 minutes).

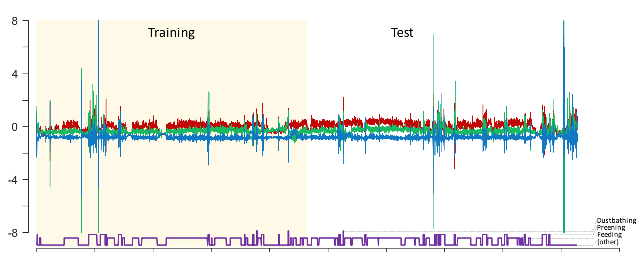

Figure 7: Three-dimensional chicken time series (the top/red time series is X-axis; middle/green time series is Y-axis and bottom/blue time series is Z-axis time series). The purple lines represent annotations of observed chicken behaviors captured on video; the height of each annotated region represents a distinct behavior.

This video recording provides ground truth to act as training data. The sensor data was carefully annotated [12], based on the video-recorded chicken activities; it must be noted that even the most careful human labeling of chicken behaviors can contain errors, especially false negatives.

As shown in Figure 7, the annotations (purple) take the form of a categorical vector that indicate that in the corresponding region one or more examples of the corresponding behavior were observed. Such data is often called "*weakly labeled*" data. In addition, there are almost certainly instances of the behavior outside the annotated regions which the annotator failed to label, perhaps because the chicken in question was occluded in the video. However, we believe that such false negatives are rare enough to be ignored, and that moreover, our algorithm, is not very sensitive to mislabeled data.

We do not know the exact number of instances of a class inside a region. As noted above, the labels are of the form "*there are about ten pecks in that six second snippet*". To address this issue, we utilize the concept of Multiple Instance Learning (MIL) [22]. MIL assumes each annotated region as a "bag" containing one or more instances of a class. If at least a single instance of a class is matched inside a bag, it is treated as a true positive. If no instances of the class are detected inside the bag, then the entire bag is treated as a false negative. In case an instance of class is mismatched inside a bag belonging to some other behavior, then it is treated as a false positive. Finally, if no mismatch occurs inside a bag of a non-relevant class, then the entire bag is treated as a true negative.

Next, we investigate the utility of our algorithm for separating distinct chicken behaviors. A healthy chicken is expected to display a set of behaviors [10]. In this work, we only considered `pecking`, `preening` and `dustbathing` behaviors. Pecking is perhaps the most familiar behavior in chickens which may be performed tens of thousands of times each day.

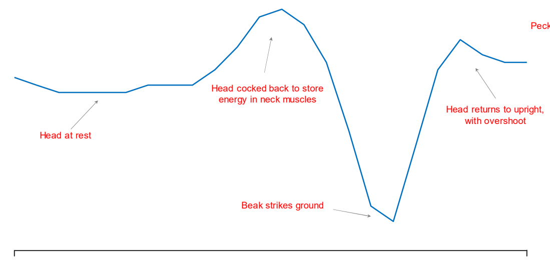

Figure 8: The shape of a chicken peck, used as the query for creating a shape-based model for classification of pecking behavior.

Figure 8 shows the typical query-template we used for creating the shape-based model for the pecking behavior. Given the query in Figure 8, we calculate the distance profile (see definition 4) for the training dataset. Then, we create a histogram of distances to separate `pecking` and non-`pecking` instances with a high degree of confidence. Then, we calculate probability vectors and classify the time series based on the probabilities.

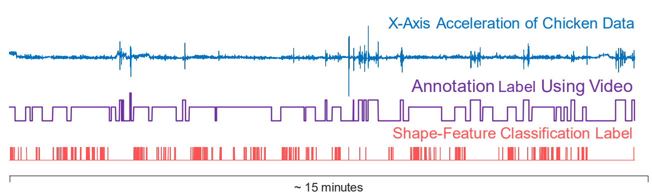

Figure 9: The shape-feature classification for the pecking behavior in pink. For clarity only the X-Axis is shown in blue with the annotation video from the video in purple.

Table 5 presents the confusion matrix for the classification of pecking behavior in the test dataset.

Table 5: Confusion matrix for pecking behavior

		Actual Class	
		Pecking	Non-Pecking
Predicted Class	Pecking	25 TP	4 FP
	Non-Pecking	2 FN	40 TN

$$\text{Precision}_{(\text{Pecking})} = \frac{TP}{TP + FP} = 25 / 29 = 0.86$$

$$\text{Recall}_{(\text{Pecking})} = \frac{TP}{TP + FN} = 25 / 27 = 0.92$$

$$\text{Accuracy}_{(\text{Pecking})} = \frac{TP + TN}{TP + FP + TN + FN} = 65 / 71 = 0.91$$

Given the results above, our classification model has 86% precision and 92% recall in matching instances of the pecking behavior. Overall, the classifier has 91% accuracy for the preening behavior, which compares very favorably to 70% default rate (i.e., guessing every observed object as the majority class).

The preening behavior does not have a well-conserved shape. Thus, we should turn our attention to features. Similarly, we calculate the feature vectors (see definition 6) for the training dataset. Next, we create a histogram of feature values to separate preening/dustbathing from non-class instances. Finally, we calculate probability vectors and classify time series based on the probabilities.

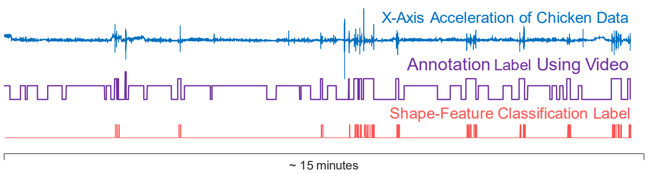

Figure 10: The shape-feature classification for the preening behavior in pink. For clarity only the X-Axis is shown in blue with the annotation video from the video in purple.

Table 6 presents the confusion matrix for the classification of preening behavior in the test dataset.

Table 6: Confusion matrix for preening behavior

		Actual Class	
		Preening	Non-Preening
Predicted Class	Preening	14 TP	0 FP
	Non-Preening	0 FN	57 TN

$$\text{Precision}_{(\text{Preening})} = 14 / 14 = 1.00$$

$$\text{Recall}_{(\text{Preening})} = 14 / 14 = 1.00$$

$$\text{Accuracy}_{(\text{Preening})} = 71 / 71 = 1.00$$

Given the results above, our classification model has a 100% precision and recall in matching instances of the preening behavior. Overall, the classifier has 100% accuracy for the preening behavior, which compares very favorably to 70% default rate.

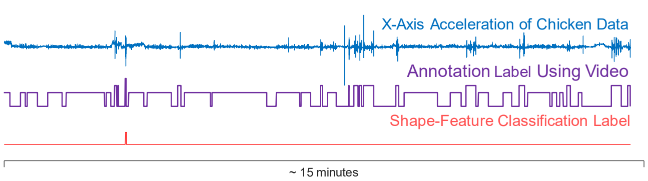

Figure 11: The shape-feature classification for the dustbathing behavior in rose. For clarity only the X-Axis is shown in blue with the annotation video from the video in purple.

Table 7 presents the confusion matrix for the classification of dustbathing behavior in the test dataset.

Table 7: Confusion matrix for dustbathing behavior

		Actual Class	
		Dustbathing	Non-Dustbathing
Predicted Class	Dustbathing	1 TP	0 FP
	Non-Dustbathing	0 FN	70 TN

$$\text{Precision}_{(\text{Preening})} = 1 / 1 = 1.00$$

$$\text{Recall}_{(\text{Preening})} = 1 / 1 = 1.00$$

$$\text{Accuracy}_{(\text{Preening})} = 71 / 71 = 1.00$$

Given these evaluation results, our model has 1.00 precision in matching dustbathing subsequences and 1.00 recall in matching relevant instances of the dustbathing behavior. Finally, the model has 100% overall accuracy in matching dustbathing subsequences compared to 99% default rate (i.e., guessing every observed object as the majority class).

6.2 A Critical Lesion Study

Our fundamental claim in this work is that using a combined shape/feature modality is superior in some domains, including the domain-at-hand. There is a very simple way to demonstrate this. We can repeat the previous experiments twice, once using *only* shape, and once using *only* features. Everything else remains exactly the same. Table 8 shows the results, which confirm the superiority of our more expressive approach.

Table 8: Shape-Feature classification for chicken behaviors.

Classification	Activity	Precision	Recall	Accuracy
Shape	Pecking	0.71	0.81	0.85
	Preening	0.91	0.71	0.93
	Dustbathing	1	1	1
Feature	Pecking	0.60	0.77	0.77
	Preening	1	1	1
	Dustbathing	1	1	1
Shape-Feature (this paper)	Pecking	0.86	0.92	0.91
	Preening	1	1	1
	Dustbathing	1	1	1

While `dustbathing` is easy to recognize with either modality, a feature-based approach has difficulty with `pecking`, and a shape based-approach has difficulty with `preening`. As the only difference in these three experiments are the allowed modalities, thus we can attribute our success to more general algorithm.

6.3 Case Study: A Day of Chicken Life

Up to now, we trained and tested the proposed shape-feature based classification algorithm on labeled data. However, as we mentioned early in this study, we believe that the lessons learned with the short (labeled) dataset can be generalized to other chickens that were not previously seen.

We start by looking into classifying a day of data for a single chicken, as shown in Figure 12. Most animals have a daily recurrent pattern of activity called a "Circadian Rhythm". We examine the existence of such pattern in Figure 12. As can be seen in the figure, there are not many activities before sunrise and after artificial lights went off at 10 pm. This makes sense as chickens are visual creatures and quickly fall to sleep with a lack of light.

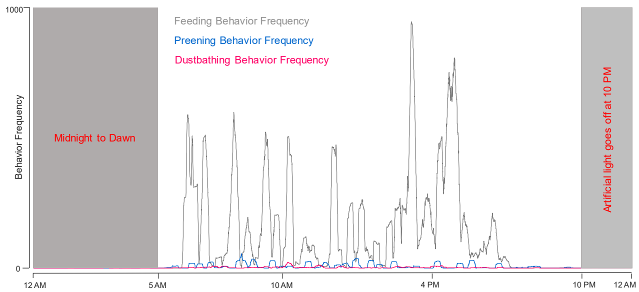

Figure 12: Shape-Feature based classification of a 24-hour of chicken data. The gray regions denote nighttime.

Next, we go one step further and we put the classification results into work to differentiate between healthy and sick chickens based on the frequency (count) of the behaviors.

6.4 Case Study: Healthy vs. Sick Chickens

Diseases are of particular importance due to the heavy economic losses they cause in poultry production **Error! Reference source not found.**. Rapid detection and diagnosis allow for decreasing the costs associated with the disease. In this section, we will use our feature-shape algorithm to recognize unhealthy chickens to combat the spread of diseases in poultry.

We look into two 24-hour days of chicken data for all the chickens. Given the field-inspection and manual verification by entomologist researchers, we know that the chickens were all healthy on the first day and all were sick (infested with ectoparasites) on the other day. The aim is that the shape-feature classification results can help to distinguish between healthy and sick chickens with an acceptable confidence.

We start by looking at the histogram for the distribution of the number of pecking behaviors for all the chickens when they were healthy (blue) and sick (red), as shown in Figure 13. As evident in the figure the peck counts for healthy and sick chickens have some overlaps.

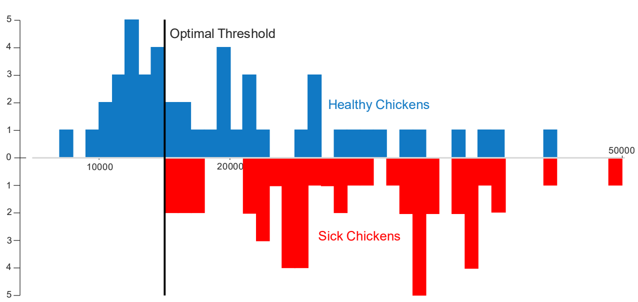

Figure 13: The histogram for distribution of the number of pecking behaviors in healthy (blue) and sick (red) chickens.

Next, we proceed to inspecting the distribution of the number of preening behaviors for all the chickens when they are healthy (blue) and sick (red). As shown in Figure 14 the distribution of preening counts for healthy chickens is skewed to the left of the histogram while the distribution of the preening counts for sick chickens is concentrated at the center and right side of the histogram. This shows that the shape-feature classification algorithm can distinguish between sick and healthy chickens based on the preening count with a good confidence.

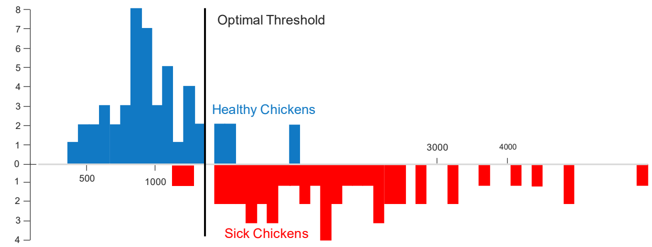

Figure 14: The histogram for distribution of the number of preening behaviors in healthy (blue) and sick (red) chickens.

Finally, in Figure 15, we look into the distribution of the number of dustbathing behaviors for healthy (blue) and sick (red) chickens. It is obvious that the distribution of dustbathing for healthy chickens is mostly accumulated at the left side of the histogram. Whereas the distribution of the dustbathing count for sick chickens is spread over at the center of the histogram.

In overall, the pecking behavior seems to provide an acceptable confidence for distinguishing healthy and sick chickens. The preening and dustbathing behaviors show great confidence towards differentiating between sick and healthy chickens.

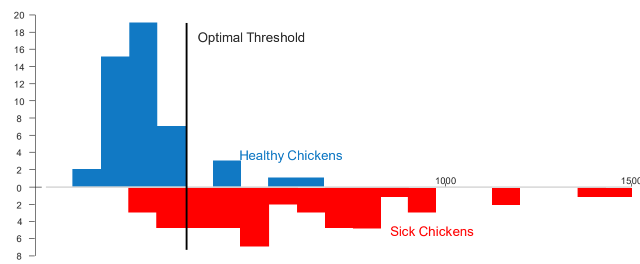

Figure 15: The histogram for distribution of the number of dustbathing behaviors in healthy (blue) and sick (red) chickens.

It is fair to say that our proposed method is expressive since the only difference between our algorithm and the other two methods (i.e. shape-based classification and feature-based classification) is the way we combined those two possibilities. Nothing else has changed. Therefore, we can attribute any success only to the increased expressiveness of the proposed method.

7. CONCLUSIONS AND FUTURE WORK

In this work, we introduced an algorithm to classify behaviors, using both shape and feature measures, in weakly labeled time series data. We demonstrated, with an extensive empirical study, that our algorithm can robustly classify real, noisy and complex datasets, based on a combination of shape and features, and tested our proposed algorithm on real-world datasets. Our ideas are currently been evaluated with a large-scale field trial involving hundreds of birds. While our study was motivated by a pressing problem in poultry welfare, it could clearly be used in other real-world data problems. For example, oil and gas production is telemetry-rich endeavor. Moreover, it features all the elements we mentioned that make our poultry problem so difficult. In particular:

- Data is often only weakly labeled. For example, the ultimate Key Performance Indicator (KPI) may only be obtained once a shift by taking a physical sample and measuring some metric. If it is bad, we get only a weak label such as "*sometime in the last eight hours something bad happened*".

- The duration of different classes can be at different scales, for example `foaming` might happen in minutes, whereas `clogging` may take hours or longer.
- Different classes may best manifest themselves in just *one* of the shape or feature modalities. For example, `foaming` is typically recognized (at least by eye) in feature space, whereas `sticking-valve` is recognized recognize in shape (as a "step pattern").

Thus, we believe that our observations and algorithms may be more generally applicable as researchers move beyond the toy problems used in the UCR archive.

8. ACKNOWLEDGMENTS

This work is supported by NSF IIS-1510741, 1631776, CPS 1544969 and gifts from FarmSense, and a Google faculty award.

9. REFERENCES

[1] A. Abdoli et al. "Time series classification to improve poultry welfare." 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2018.

[2] A. Abdoli, et al. "Time Series Classification: Lessons Learned in the (Literal) Field while Studying Chicken Behavior." 2019 IEEE International Conference on Big Data (Big Data). IEEE, 2019.

[3] A. Mueen et al., "The fastest similarity search algorithm for time series subsequences under euclidean distance", 2015.

[4] A. Murillo, A. Abdoli, R. Blatchford, et al. Parasitic mites alter chicken behaviour and negatively impact animal welfare. Sci Rep 10, 8236 (2020). https://doi.org/10.1038/s41598-020-65021-0

[5] A. Reiss et al. "Exploring and extending the boundaries of physical activity recognition." 2011 IEEE International Conference on Systems, Man, and Cybernetics. IEEE, 2011.

[6] B. Chiu, E. Keogh and S. Lonardi. "Probabilistic discovery of time series motifs". In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 493-498), 2003, ACM.

[7] B. D. Fulcher et al. "Highly comparative time-series analysis: The empirical structure of time series and their methods." Journal of the Royal Society Interface 10.83 (2013): 20130048.

[8] B. G. Smythe et al., "Behavioral responses of cattle to naturally occurring seasonal populations of horn flies (Diptera: Muscidae) under rangeland conditions," Journal of Economic Entomology, vol. 108, no. 6, pp. 2831–2836, 2015.

[9] C. Lubba et al., "catch22: CAnonical Time-series Characteristics," Data Mining and Knowledge Discovery, 33(6), 1821-1852, 2019.

[10] D. Banerjee et al., "Remote activity classification of hens using wireless body mounted sensors," in Wearable and Implantable Body Sensor Networks (BSN), 2012 Ninth International Conference on, pp. 107–112, 2012, IEEE.

[11] E. Walton et al., "Evaluation of sampling frequency, window size and sensor position for classification of sheep behaviour," Royal Society Open Science, vol. 5, no. 2, p. 171442, 2018.

[12] ELAN [Computer software]. Nijmegen: Max Planck Institute for Psycholinguistics. Retrieved from https://tla.mpi.nl/tools/tla-tools/elan/

[13] F. Palumbo, C. Gallicchio, R. Pucci, and A. Micheli, "Human activity recognition using multisensor data fusion based on reservoir computing," Journal of Ambient Intelligence and Smart Environments, 8(2), 87-107, 2016.

[14] G. Batista, E. Keogh, O. Tataw, V. de Souza. "CID: an efficient complexity-invariant distance for time series". Data Min. Knowl. Discov. 28(3): 634-669, 2014.

[15] G. E. Batista et al., "SIGKDD demo: sensors and software to allow computational entomology, an emerging application of data mining," in Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 761–764, 2011, ACM.

[16] H. A. Dau and E. Keogh, "Matrix Profile V: A Generic Technique to Incorporate Domain Knowledge into Motif Discovery", In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17), pp.125-134, 2017, ACM.

[17] H. A. Dau, A. Bagnall, K. Kamgar, C. M. Yeh, Y. Zhu, S. Gharghabi, C. A. Ratanamahatana, and E. Keogh. "The UCR time series archive." IEEE/CAA Journal of Automatica Sinica 6, no. 6 (2019): 1293-1305.

[18] H. Ding et al., "Querying and mining of time series data: experimental comparison of representations and distance measures," Proceedings of the VLDB Endowment, vol. 1, no. 2, pp. 1542–1552, 2008.

[19] http://www.fao.org/poultry-production-products/production/management-and-housing/ [Last visited: February 2020]

[20] http://www.fao.org/poultry-production-products/products-processing/en/ [Last visited: February 2020].

[21] https://www.ciwf.com/farmed-animals/chickens/ [Last visited: 31 January 2020].

[22] J. Amores. "Multiple instance classification: Review, taxonomy and comparative study." Artificial intelligence 201 (2013): 81-105.

[23] J. Barwick et al., "Categorising sheep activity using a tri-axial accelerometer," Computers and Electronics in Agriculture, vol. 145, pp. 289–297, 2018.

[24] J. Doppler et al., "Variability in foot-worn sensor placement for activity recognition," in Wearable Computers, 2009. ISWC'09. International Symposium on, pp. 143–144, 2009.

[25] M. Kreil, B. Sick, and P. Lukowicz. "Coping with variability in motion based activity recognition." Proceedings of the 3rd International Workshop on Sensor-based Activity Recognition and Interaction. 2016.

[26] M. Moreau et al., "Use of a tri-axial accelerometer for automated recording and classification of goats' grazing behaviour," Applied Animal Behaviour Science, vol. 119, no. 3–4, pp. 158–170, 2009.

[27] M. S. Dawkins, S. Donelly and T. A. Jones, "Chicken welfare is influenced more by housing conditions than by stocking density," Nature 427: 342-344, 2004.

[28] N. Li et al. "Automated techniques for monitoring the behaviour and welfare of broilers and laying hens: towards the goal of precision livestock farming." Animal, 1-9, 2019.

[29] N. S. Brown, "The effect of louse infestation, wet feathers, and relative humidity on the grooming behavior of the domestic chicken," Poultry Science, vol. 53, no. 5, pp. 1717–1719, 1974.

[30] O. D. Lara, and M. A. Labrador. "A survey on human activity recognition using wearable sensors." IEEE communications surveys & tutorials 15.3 (2012): 1192-1209.

[31] S. Alaee et al., "Features or Shape? Tackling the False Dichotomy of Time Series Classification," In Proceedings of the 2020 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2020.

[32] S. Gharghabi et al., "Matrix profile XII: MPdist: a novel time series distance measure to allow data mining in more challenging scenarios". In 2018 IEEE International Conference on Data Mining (ICDM), pp. 965-970. IEEE, 2018.

[33] S. Imani and E. Keogh. "Matrix Profile XIX: Time Series Semantic Motifs: A New Primitive for Finding Higher-Level Structure in Time Series." 2019 IEEE International Conference on Data Mining (ICDM). IEEE, 2019.

[34] S. Imani et al., Matrix Profile XIII: Time Series Snippets: A New Primitive for Time Series Data Mining. In 2018 IEEE International Conference on Big Knowledge (ICBK) (pp. 382-389). IEEE, 2018.

[35] Website: https://sites.google.com/site/chickenkdd/